LAPPEENRANNAN-LAHDEN TEKNILLINEN YLIOPISTO LUT

School of Engineering Science

Tuotantotalous

Jukka-Matti Turtiainen

MENTAL MODEL FOR EXPLORATORY DATA ANANALYSIS APPLICATIONS FOR STRUCTURED PROBLEM-SOLVING

Työn tarkastajat:	Apulaisprofessori	Lea Hannola
	Tutkijatohtori	Kirsi Kokkonen

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Degree Programme in Industrial Engineering and Management

Jukka-Matti Turtiainen

Mental Model for Exploratory Data Analysis Applications for Structured Problem-Solving
Master's thesis
2019
129 pages, 58 figures, 3 tables and 2 appendices
Examiners: Associate Professor Lea Hannola. and Post-Doctoral Researcher Kirsi Kokkonen

Keywords: Exploratory Data Analysis, Structured Problem-Solving, Lean Six Sigma, DMAIC, PDCA, SDCA

The ability to solve complex problems has been identified as one of the most important skills for 2020s. Exploratory Data Analysis (EDA) can be used as part of structured problem-solving process. EDA has been described more as an art than a science.

The research objective of the master's thesis was to develop a mental model of EDA that better describes the way it can be applied to initiate the investigation of a performance issue. The thesis presents an overview of EDA history in a literature review, from which a preliminary mental model was developed. That mental model was presented for review and critique to a team of Lean Six Sigma Master Black Belt content matter experts, who offered suggestions in face-to-face, individual interviews for further improvement of the model.

The updated mental model was presented individually to a different group of content matter experts for critique. The thesis demonstrates a practical application of the proposed EDA model through a case study, where it was applied to define the problem in a Lean Six Sigma Black Belt project. The proposed EDA mental model was well accepted by these Master Black Belt experts. Additional research topics were suggested to make the mental model more comprehensive. The proposed mental model integrates the foundational scientific thinking logic in the process of conducting Exploratory Data Analysis which brings EDA closer to science.

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Tuotantotalouden koulutusohjelma

Jukka-Matti Turtiainen

Ajatusmalli tutkivan data-analyysin sovellutuksiin osana järjestelmällistä ongelmanratkaisua

Diplomityö
2019
129 sivua, 58 kuvaa, 3 taulukkoa ja 2 liitettä
Tarkastajat: Apulaisprofessori Lea Hannola ja Tutkijatohtori Kirsi Kokkonen

Hakusanat: Tutkiva data-analyysi, järjestelmällinen ongelmanratkaisu
Keywords: Exploratory Data Analysis, Structured Problem-Solving, Lean Six Sigma, DMAIC, PDCA, SDCA

Monimutkaisten ongelmien ratkaisukykyä pidetään yhtenä tärkeimpänä taitona 2020-luvulla. Tutkivaa data-analyysiä voidaan käyttää osana järjestelmällistä ongelmanratkaisuprosessia. Tutkivan data-analyysin on kuvattu olevan enemmän taidetta kuin tarkkaa tiedettä.

Työn tutkimustavoitteena oli kehittää ajatusmalli tutkivasta data-analyysistä, joka kuvaa paremmin miten sitä voidaan soveltaa tutkittaessa suorituskykyongelmia. Tämä tutkielma esittää yleiskuvan tutkivan data-analyysin historiasta kirjallisuuskatsauksen muodossa, jonka perusteella kehitettiin ensimmäinen ajatusmalli tutkivasta data-analyysistä. Ajatusmalli esitettiin alan asiantuntijoille, Lean Six Sigma Master Blackbelteille, jotka ehdottivat parannuksia malliin henkilökohtaisissa haastatteluissa.

Ehdotusten pohjalta ajatusmallia kehitettiin edelleen, ja esiteltiin eri ryhmälle alan asiantuntijoita kritiikkiä varten. Diplomityö demonstroi ehdotetun ajatusmallin käyttöä reaalimaailmassa tapaustutkimuksen kautta, jossa sitä sovelletaan Lean Six Sigma Black Belt -projektissa. Asiantuntijat suhtautuivat positiivisesti ehdotettuun ajatusmalliin, ja he ehdottivat lisäyksiä malliin, jotta siitä tulisi entistä täydellisempi. Esitetty ajatusmalli integroi tieteellisen lähestymistavan tutkivan data-analyysin suorittamiseen, mikä tuo sen lähemmäksi tiedettä.

ACKNOWLEDGEMENTS

Seed for the topic of this thesis were planted in 2014 in Summer Academy organized by European Students of Industrial Engineering and Management, where I met Dr. Gregory H. Watson for the first time. The story of Nokia Mobile Phones and how it paid millions of dollars on the intellectual property of Six Sigma materials and training in 1995 was especially intriguing for me. That event started a journey, which has enabled me to conduct the research presented in this thesis.

Thank you Kaisa, Peter, Tuomas and Kristian for the challenge you gave me to customize a Green Belt training to fit the needs of the company and on the support on the way. Without that environment, I would have ended up writing my thesis on a different topic. Additionally, thank you all Lean Six Sigma Master Black Belts from different companies on the possibilities to discuss on the topic and conduct the interviews with some of you! Furthermore, I have to thank you my professor Lea Hannola on the patient support, which enabled the thesis to find its way to a great result!

Daniel Kahneman had a thinking buddy called Amos Tversky, I have had Mikko Rajala. We have spent hundreds of hours on discussing topics around the field of this thesis. The challenge we have had, has been how to teach Lean Six Sigma to Industrial Engineering and Management Students all around Europe. Greg has provided us tons of material, but not all of that can be taught in a week... Thank you Greg for being on the journey with us. Those experiences have developed my thinking to the level where it currently is.

I was asked the following question in Porto in 2017 during a Green Belt course: ”When I go to a factory, how do I know which data should I analyze or collect?” Without those kind of questions and situations, I would not be where I am today. Thank you everyone!

Helsinki, 4.12.2019
Jukka-Matti Turtiainen

TABLE OF CONTENTS

FIGURES

TABLES

ABBREVIATIONS

BB	Lean Six Sigma Black Belt
DMAIC	Define – Measure – Analyze – Improve – Control Project management method
EDA	Exploratory Data Analysis
GB	Lean Six Sigma Green Belt
LSS	Lean Six Sigma
MBB	Lean Six Sigma Master Black Belt
MSA	Measurement System Analysis
PDCA	Plan – Do – Check – Act Cycle
SAP	Enterprise software used to manage business and operational transactions
SDCA	Standardize – Do – Check – Act Cycle
SME	Subject Matter Expert

1. INTRODUCTION

1.1. Motivation and Objective

Complex problem-solving skills and the combination of Analytical Thinking and Innovation are listed as two of the 10 most needed skills for 2020 as listed in the most trending skills by the World Economic Forum (WEF, 2018, p. 12). Exploratory Data Analysis is a methodology for solving complex problems, but its theory has developed inadequately in academic literature. In the operational world of organizations, this methodology is frequently used. This qualifies it as more an art than a science. (de Mast, 2009, p. 367)

The author first heard about the subject Exploratory Data Analysis (EDA) in a 2014 lecture by Dr. Gregory H. Watson. Since that time, the author has been studying, applying and teaching the methods of EDA all around Europe and in several countries in Africa. The company sponsoring this master's thesis challenged the author to develop and teach a Lean Six Sigma Green Belt course, and include in that course a module describing and applying EDA. This requirement to develop that module initiated the research conducted in this thesis.

As the theory of EDA has been incompletely developed over its past 40 years of academic literature, a more structured thinking process or mental model is necessary to advance thinking about it so the methodology can be delivered in a teachable state. Kenneth Craik developed the concept of mental models in his book "*The Nature of Explanation.*" He observed that it should be possible to create models about thinking in a similar way to development of models about physical processes. (Craik, 1967, p. 99) Models developed about abstractions or thinking processes are called mental models. The objective of the master's thesis is to develop a mental model for Exploratory Data Analysis. In the future this mental model can be used as a guiding principle to explain how EDA should be performed and as a basis for training.

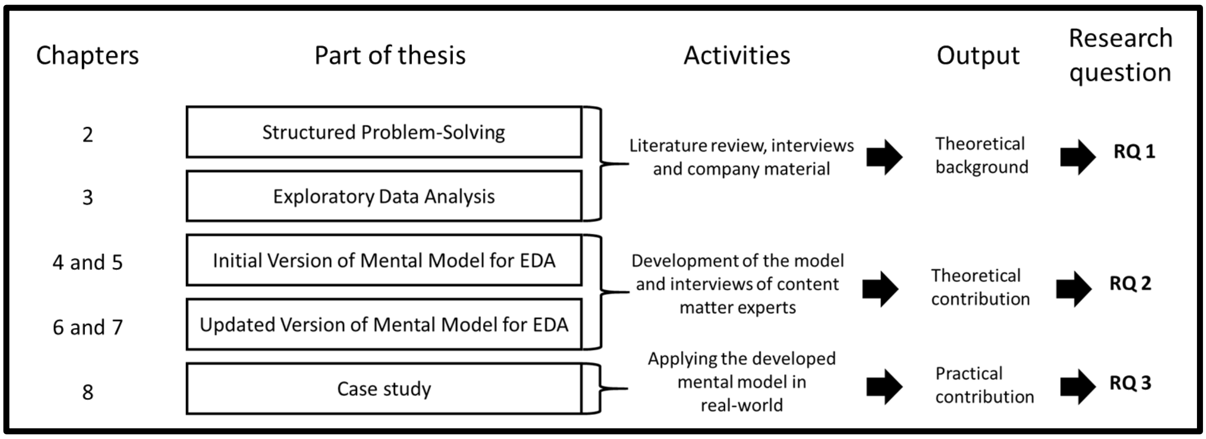
1.2. Research Questions, Methodology and Scope

RQ 1: What kind of mental models are there of Exploratory Data Analysis for structured problem-solving?

RQ 2: What type of mental model could be developed to support Exploratory Data Analysis applications for structured problem-solving?

RQ 3: How can the mental model developed for Exploratory Data Analysis be used in a real-world case?

This thesis develops a mental model for Exploratory Data Analysis in the context of structured problem-solving. The research process used in the thesis is illustrated in Figure 1. The first research question is addressed in chapters two and three using a literature search and follow-up interviews into the structured problem-solving and Exploratory Data Analysis were subsequently used to refine the EDA concept based on its pragmatic application by subject matter experts.



Chapters	Part of thesis	Activities	Output	Research question
2	Structured Problem-Solving	Literature review, interviews and company material	Theoretical background	RQ 1
3	Exploratory Data Analysis			
4 and 5	Initial Version of Mental Model for EDA	Development of the model and interviews of content matter experts	Theoretical contribution	RQ 2
6 and 7	Updated Version of Mental Model for EDA			
8	Case study	Applying the developed mental model in real-world	Practical contribution	RQ 3

Figure 1. Research Process

The second research question is addressed in chapters four through seven. Chapters four and five develop the first iteration of the mental model for EDA. Based on the literature search, a mental model for Exploratory Data Analysis is proposed in chapter four. Chapter five presents the results from interviews of qualified Lean Six Sigma (LSS) Master Black Belt (MBB) Subject Matter Experts (SMEs) who offered critiques of the proposed model and recommendations for its further advancement. Chapter six incorporates the feedback of these

SMEs n a revised mental model. Chapter seven presents the discussion of the revised model as results of interviews of a different pool of content matter experts.

The third research question is addressed in chapter eight using a case example of an Exploratory Data Analysis to support a LSS project that was conducted in an anonymous company that will be referred to as “Company X” in order to protect the privacy of their confidential information and proprietary methods. Nevertheless, the revised mental model can be demonstrated using real-world data from this project case study. Chapter nine presents findings and conclusions of this research.

This thesis focuses on creating a mental model for Exploratory Data Analysis. It does not focus on the applications of the quantitative tools presented. The scope includes application of quantitative tools that are used in structured problem-solving and are applicable to the EDA application. The only qualitative tools presented are three diagram, fishbone diagram and mind map, which are discussed in a sub-chapter 3.4. Ishikawa Analysis.

2. STRUCTURED PROBLEM-SOLVING

In this chapter a literature research and follow-up interviews are conducted to give an overview of different methods and mental models used in structured problem-solving. The chapter provides context for the usage of the exploratory data analysis applications in structured problem-solving.

2.1. The Scientific Method

The scientific method follows a three-step process, which is illustrated in Figure 2. It starts by creating a hypothesis. The hypothesis explains what phenomenon is expected to occur and it acts as the starting point of a scientific investigation. Then, an experiment is conducted to test how the phenomenon occurred in order to observe the results and confirm or disprove the hypothesis. Based on the results, the hypothesis might be revised, and a new experiment conducted. In other words, a scientific method can be applied in a sequential or iterative manner. (Hoerl & Snee, 2002, p. 43)

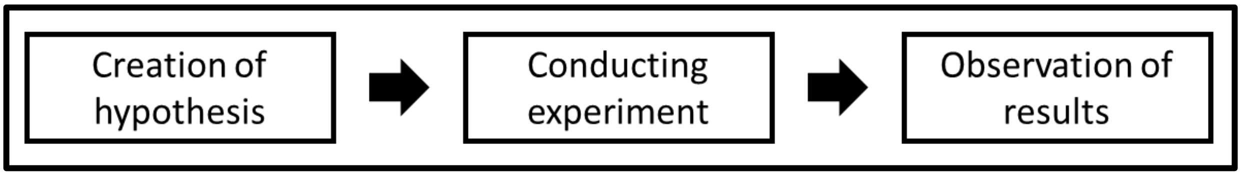


Figure 2. General Process of the Scientific Method. Adapted from Hoerl and Snee (2002, p. 43)

Walter Shewhart (1939) introduced his Cycle for Scientific Inquiry in his book *“Statistical Method From the Viewpoint of Quality Control”*. It is known better as the Shewhart Cycle. It has three steps: specification, production and inspection. Relationship of those steps to the scientific method is illustrated in Figure 3. Hypothesis is created in the Specification step. An experiment is conducted in the Production step and the hypothesis is tested in the Inspection step. The process is presented as a cycle, because the hypothesis or specifications might need to be changed as more knowledge about the process is acquired as part of the inquiry. (Watson 2018, p. 40)

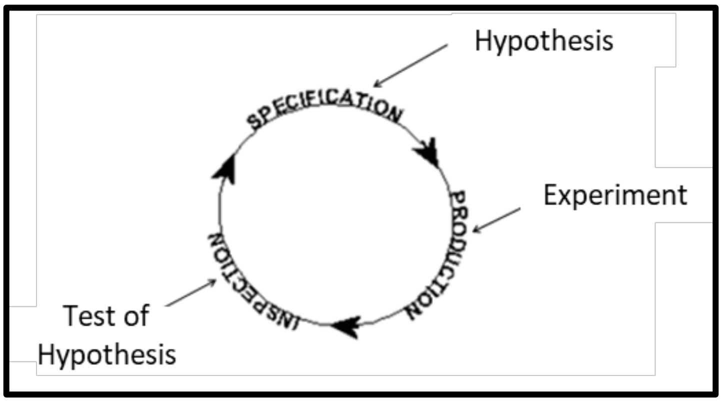


Figure 3. Shewhart Cycle (Shewhart, 1939)

2.2. Issue Statement

The issue statement can be an initiator of a problem-solving process. It is an interim step in the development of a narrative about the problem and the formal definition of a problem which will become the starting point of a more structured inquiry. Watson (2019a, p. 13) defines an issue in the following way: *"a concern that arises as a difference between customer expectation and their observations or perceptions with respect to these expectations."*

Once an issue has been noticed, it should be identified by defining what is out of control in the process, which performance indicator needs to be addressed and what are the current and desired states of performance. According to Watson (2019a, p. 13), before defining a problem statement, the following factors should be evaluated:

- *"Who does it affect? Where is it taking place?*
- *What would be the outcome if the issue is not corrected?*
- *When does this need to be fixed (sense of urgency)?*
- *Why is it important for this to be fixed?"*

2.3. Problem Statement

Defining a problem clearly is the first step in the problem-solving process. Without a clear problems statement set by project leadership, it is hard to detect quantitatively whether the problem was solved or not. (Anderson-Cook et al., 2012, p. 112) The problem statement helps the problem-solving team manage their resources and focus on a specific problem rather than general issues. Specific problems are easier to address than general ones. (Sheely et al. 2002)

A problem statement should be clear and communicate accurately the focus area for further investigation. According to Watson (2019a, p. 17), good problem statement should answer three questions:

1. *What measure needs to be changed?* – *Performance measure*
2. *How the measure needs to be changed?* – *Direction of change of performance*
3. *What is the scope of the problem? Where?* – *Location where further investigation*
   needs to focus

2.4. Plan-Do-Check-Act

The Plan-Do-Check-Act (PDCA) cycle is also known as the Shewhart cycle or Deming Cycle. Figure 4 presents the PDCA cycle in a graphical format. Walter A. Shewhart originated this model for an application of statistical testing in 1939 in his book *Statistical Method from the Viewpoint of Quality Control.* (Shewhart 1939). The PDCA model was derived from W. Edwards Deming's Japanese lectures and proposed as a generic approach for continual improvement of working processes by a Japanese research committee in the late 1950s. (Watson & DeYong 2010)

The approach starts by planning what one intends to do including desired results and methods. The Do step is where the plan is executed. The Check step is to make sure that the plan is executed, and the desired results have been achieved. The Act step is there to take actions based on observations in the Check step. If the intended results have not been gained, it might lead to creation of an updated plan, which initiates the next PDCA cycle. (Hoerl & Snee, 2002, p. 42)

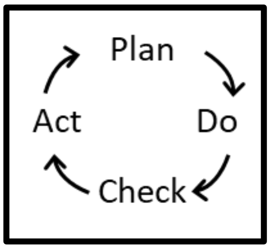


Figure 4. PDCA Cycle (Hoerl & Snee, 2002, p. 42)

2.5. Process for Developing Statistical Control

The objective of productive processes is to create a stable output that is matched to the demand for performance from its customers. Surprises are not tolerated in the process. Thus, the normal operation can be characterized as a Standardize – Do – Check – Act (SDCA) cycle. A SDCA cycle, as opposed to the PDCA cycle, emphasizes the development and maintenance of a controlled state of performance where all known problem conditions have been eliminated. This cycle is aimed at identifying and eliminating observable and assignable causes of variation (e.g., variation due to changes in material, supplier process, or personnel competence). When normal operations are interrupted by an issue, then the PDCA cycle will be invoked to escalate the attention on improvement and effect a remedy as rapidly as possible. Figure 5 illustrates the process for developing statistical control as specified by Shewhart. Watson (2019b) explained these steps in the following way:

1. *"Determine what questions need to be addressed that disclose what is "critical-to-quality" in a process operation that influences the customer perception of performance and how that factor should be appropriately measured in a manner that permits action to be taken which will control its performance."*
2. *"Define the data structure from high-level output measures to the operational measures that indicate that operations are being conducted normally. What can go wrong and where can sources of variation change the process (e.g., different materials, operators, machines, or shifts)."*

3.. *"Analyze the data according to its operational history and determine the boundary conditions representing statistical control limits for a properly performing process."*

4. *"Evaluate the process performance and determine what contingent actions need to be taken as countermeasures when the data indicates that unusual conditions of variation are present."*
5. *"Define the data collection plan – what information needs to be collected; where it is to be gathered and the procedure for collecting the data; how much data is to be collected and how often the data should be collected; how the data will be stored and in what format"*

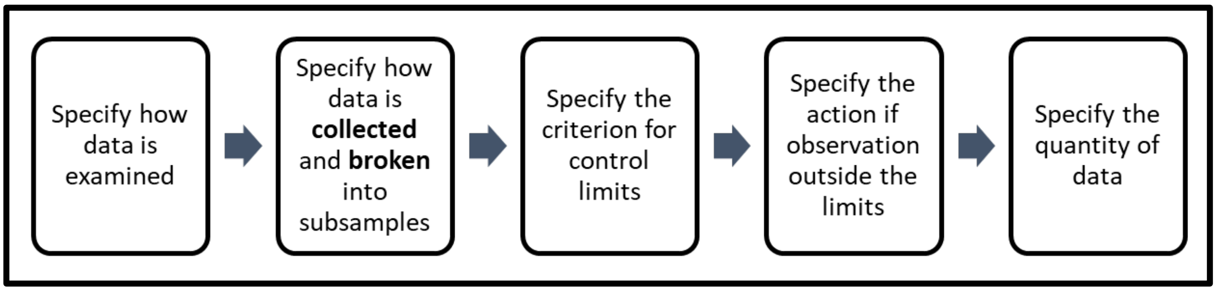

Specify how data is examined
Specify how data is **collected** and **broken** into subsamples
Specify the criterion for control limits
Specify the action if observation outside the limits
Specify the quantity of data

Figure 5. Process for Developing Statistical Control. Adapted from Shewhart (1939)

2.6. Lean Six Sigma

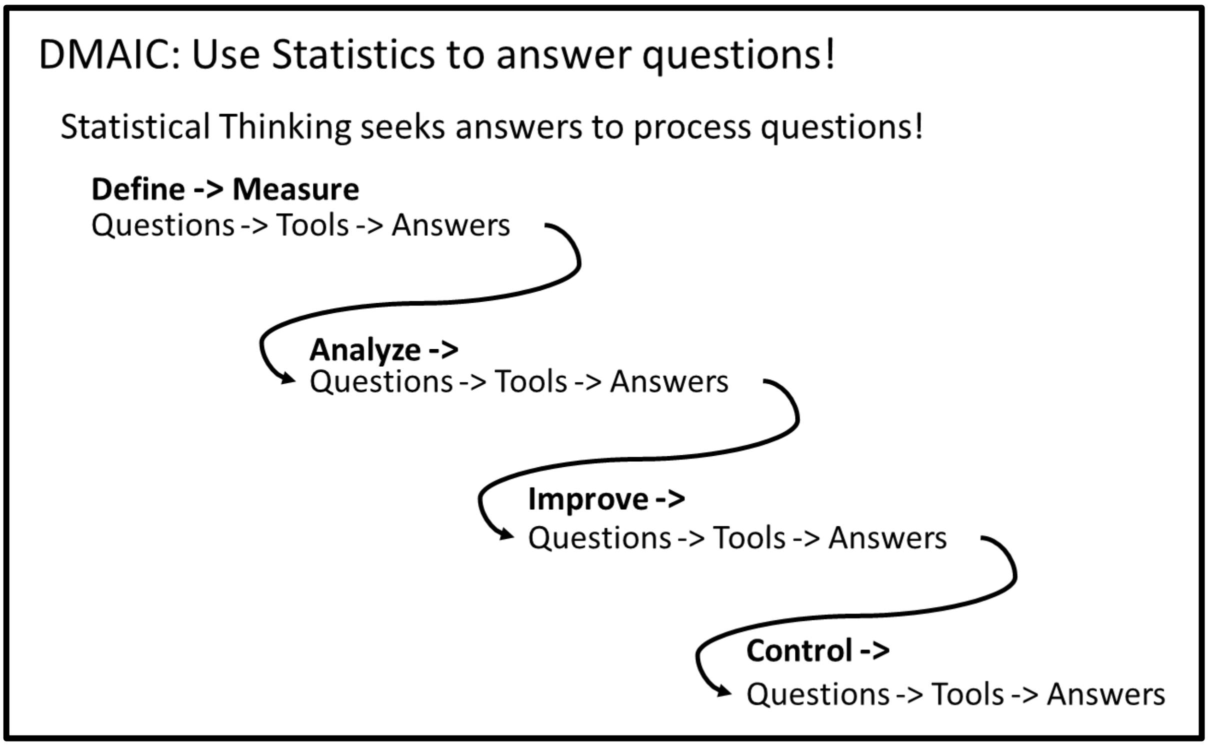
Starting development of Six Sigma was initiated in 1985 at Motorola, based upon an executive's observation that: "Our quality stinks." Six Sigma was developed in Motorola's Six Sigma Research Institute from where it was deployed to different companies in 1990's including ABB transformers, Allied Signal, Texas Instruments, Nokia Mobile Phones and General Electric. The Term Lean Six Sigma was created in 2002 when Michael L. George published book titled *Lean Six Sigma: Combining Six Sigma Quality with Lean Speed*. (Watson, 2019b)

Lean Six Sigma can be understood from four different perspectives: The first perspective is a philosophy of management. Improvement projects are selected based on a view where the need for improvement is evaluated based on the customer's experience compared to the expectations. The second perspective is an analysis methodology that applies the scientific method. That is called DMAIC, which is used to execute the change projects. The third perspective is a statistical process measurement methodology, where the average performance of a process is compared to the customer's requirement for performance in number of standard deviations. The fourth perspective of Lean Six Sigma is an organizational culture, where operations are run with data and unwanted variation and waste are reduced. (Watson, 2015a, p. 121)

2.7. Define, Measure, Analyze, Improve, Control

DMAIC stands for Define, Measure, Analyze, Improve and Control. It is a project management approach, which uses statistical thinking to find answers to process questions. Similar logic takes place in each phase of DMAIC, with questions, tools and answers applied in each phase.

The logic of the methodology is illustrated in Figure 6. The nature of the questions changes according to the different phases of DMAIC, (Watson, 2015b, p. 22)



DMAIC: Use Statistics to answer questions!

Statistical Thinking seeks answers to process questions!

Define -> Measure
Questions -> Tools -> Answers

Analyze ->
Questions -> Tools -> Answers

Improve ->
Questions -> Tools -> Answers

Control ->
Questions -> Tools -> Answers

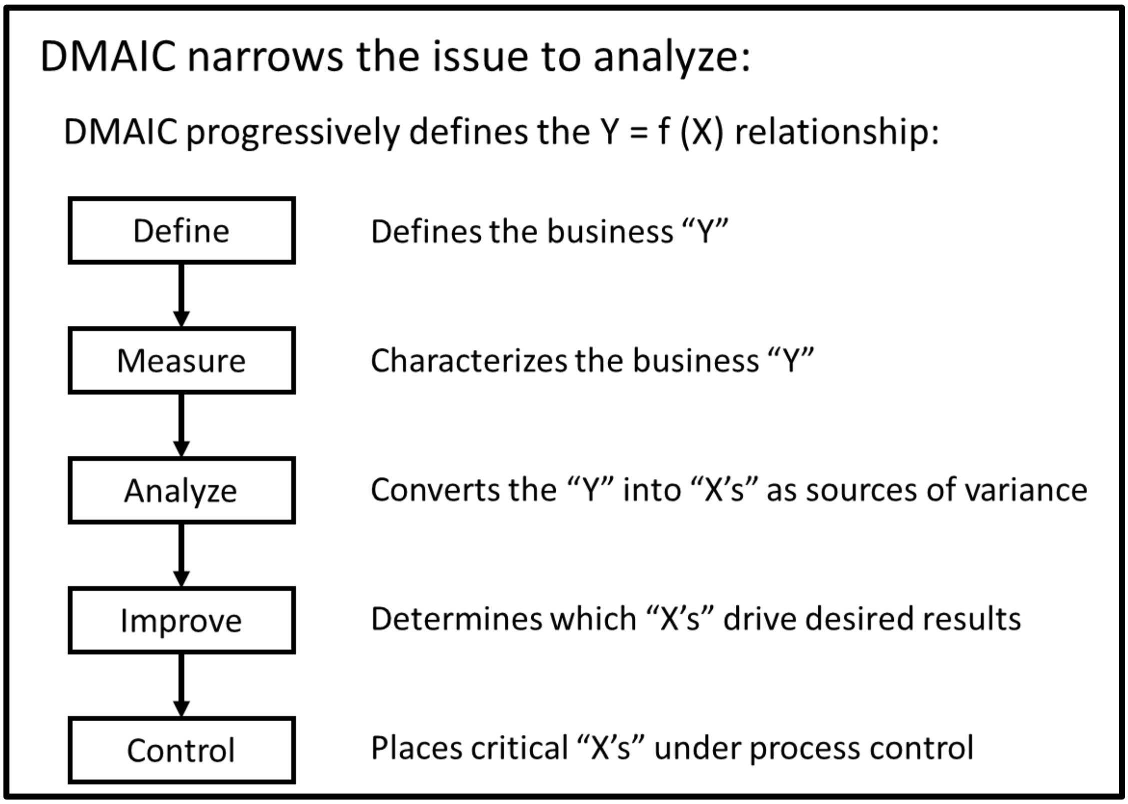
Figure 6. Statistical Thinking and DMAIC adapted from Watson (2015b, p. 22)

According to Watson (2015f, p. 7), Exploratory Data Analysis is performed initially in the Define phase where if focuses on the overall performance results (identified as the Y measure in the problem analysis), but it can be applied as an investigation process across the DMA phases. However, the procedure for using this analysis of the usage of the tools changes after Define phase to focus on investigating the detailed process analytical factors (identified as X's in the problem analysis) to conduct a "drilldown" analysis to discover the initial or so-called "root" causes that have created the performance problem. (Watson, 2015b, pp. 16-24)

2.8. Transfer Function in Lean Six Sigma

In Lean Six Sigma, the function Y = f (x) is the basic equation used for analysis as a transfer function where process inputs (X's) create the process output (Y). It presents the logical breakdown of high-level measures used at business level into logical sub-groups of operational measures. The approach helps a process team to learn about the process and to focus on what to improve. X’s at the operational level are specific activities affecting the Y – the process outcomes. Changing or controlling X’s better will improve the Y. (Watson, 2015d, pp. 22-36)

Finding the right variables to measure as X’s for improvement can be accomplished in a structured manner using DMAIC. Figure 7 defines the focus for different phases of DMAIC and the Y = f(x) relationship. Analysis starts with the Y measure in Define; then characterizing it in Measure; while in the Analyze phase, the goal is to discover those X’s which contribute the most as sources of variation to the Y. In the Improve phase experiments are made by changing the X’s in order to determine which ones are the most important contributors to improvement of the process. In Control phase these critical X’s are controlled in order for the process to consistently deliver desired results. (Watson, 2015d, pp. 26-33)

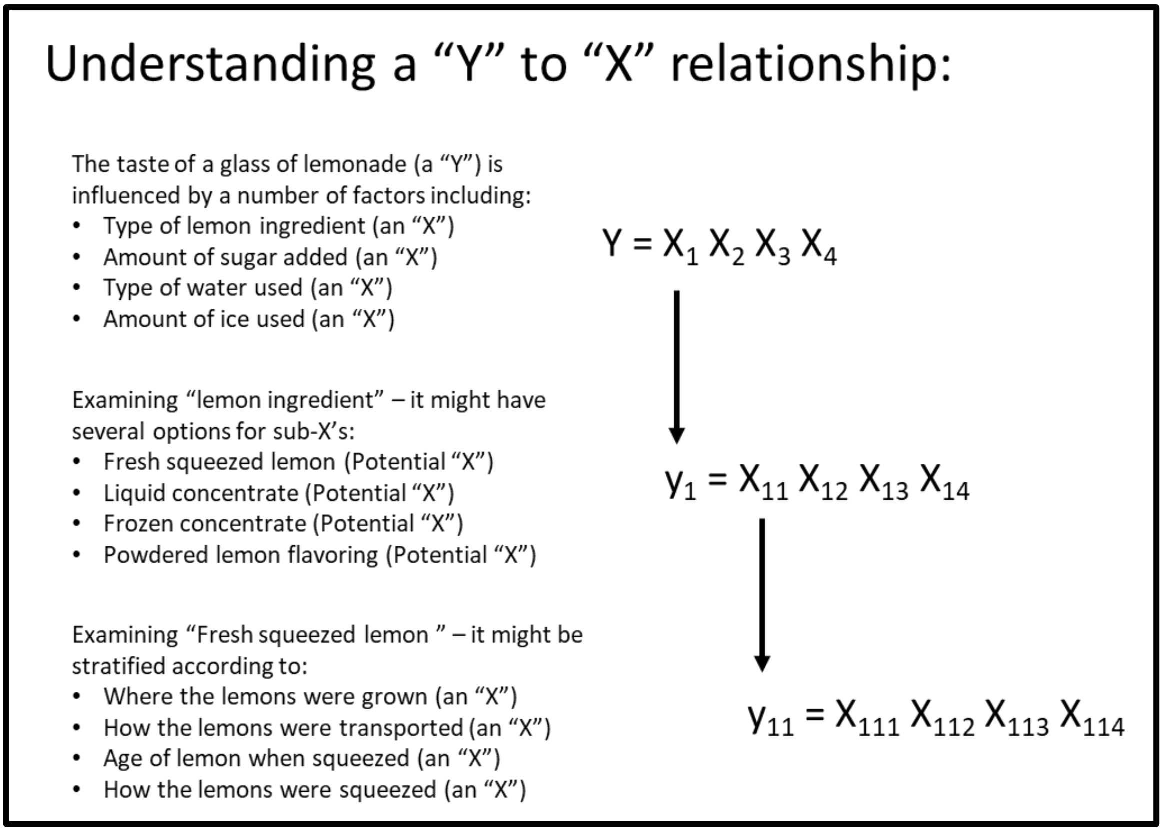

DMAIC narrows the issue to analyze:

DMAIC progressively defines the Y = f (X) relationship:

Define
Defines the business “Y”
Measure
Characterizes the business “Y”
Analyze
Converts the “Y” into “X’s” as sources of variance
Improve
Determines which “X’s” drive desired results
Control
Places critical “X’s” under process control

Figure 7. Linkage of Y = f(x) and DMAIC adapted from Watson (2015c, p. 41)

Building the Y = f(x) function might be challenging sometimes, as digging deeper into process details, the nature of the X’s has a tendency to change. (Watson, 2019c) For example, Figure 8 illustrates the decomposition of Y into detailed X’s by providing an example of lemonade. At the first level of X's, each X represents one ingredient in the lemonade recipe. On the second level of sub-X's, each ingredient has functions (e.g., X's). On the third level of X's, the nature of X's changes to reflect what occurs on the process level of making specific ingredients for lemonade.



Understanding a "Y" to "X" relationship:

The taste of a glass of lemonade (a "Y") is influenced by a number of factors including:

- Type of lemon ingredient (an "X")
- Amount of sugar added (an "X")
- Type of water used (an "X")
- Amount of ice used (an "X")

$Y = X_1\ X_2\ X_3\ X_4$

Examining "lemon ingredient" – it might have several options for sub-X's:

- Fresh squeezed lemon (Potential "X")
- Liquid concentrate (Potential "X")
- Frozen concentrate (Potential "X")
- Powdered lemon flavoring (Potential "X")

$y_1 = X_{11}\ X_{12}\ X_{13}\ X_{14}$

Examining "Fresh squeezed lemon " – it might be stratified according to:

- Where the lemons were grown (an "X")
- How the lemons were transported (an "X")
- Age of lemon when squeezed (an "X")
- How the lemons were squeezed (an "X")

$y_{11} = X_{111}\ X_{112}\ X_{113}\ X_{114}$

Figure 8. Decomposition of Y to X's (Watson, 2015e, p.16)

3. EXPLORATORY DATA ANALYSIS

This chapter discusses Exploratory Data Analysis, where the main research methodology is a literature review. The chapter has five subchapters where EDA is presented from different perspectives starting from introduction, continuing with history. Subchapter three presents the objectives of EDA and the last two describe some methods and tools that are possible to apply during the EDA process.

3.1. Introduction to Exploratory Data Analysis

Exploratory Data Analysis is an important, but under-valued part of structured problem-solving projects. It is a study process that combines statistical tools and process knowledge to develop insights and formulate hypotheses regarding improvement opportunities based on empirical data. In Exploratory Data Analysis, information is presented in a graphical format to utilize pattern recognition capabilities of human brain. (de Mast & Kemper, 2009)

Traditionally, statistical tools have been used for confirmatory data analysis, which is known as hypothesis testing. According to Simpson (2009, p.376): *"many, if not most, times projects require some combination of experiment and observation. Thus, analysts must concurrently use EDA and CDA, not EDA or CDA."* In other words, hypotheses generated by Exploratory Data Analysis can later be tested by Confirmatory Data Analysis. Analysts conducting Exploratory Data Analysis are working together with content matter experts to analyze data and generate ideas and hypotheses. (Vining 2009, p.381)

3.2. Developmental History of Exploratory Data Analysis

3.2.1. Foundations of Exploratory Data Analysis by Dr. John Tukey

According to Tukey, the basic problem when analyzing any data set is how to make it easily and effectively handleable by our minds. An effective way to do that is with pictures. Tukey (1977, p.5) comments that: "***The greatest value of a picture*** *is when it forces us to notice what we never expected to see.*" When talking about statistical techniques, pictures are graphs. He gives two additional guidelines for EDA:

1. *"Anything that makes a simpler description possible, makes the description more easily handleable.*
2. *"Anything that looks below the previously described surface makes the description more effective."*

Following these guidelines, the analyst is trying to simplify the description and to describe what is happening at subsequently deeper layers of the problem.

Tukey points out that there can be many ways to approach the same problem using the same data to find the answer to a question. Some approaches are better than others for different situations but determining which one is better for any specific situation might require the use of many data sets to develop a conclusion about the best approach. According to Tukey (1977, p.8): "*To unlock the analysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act.*"

The book Exploratory Data Analysis by Tukey was published in 1977 in a period where there were not many computers available for workers to conduct problem inquiries. The techniques and cases presented in this book by Tukey are from the point of view that the analyses must be calculated by hand using pencil and paper. Tukey is open and honest that he believes the future will have better techniques and faster analysis when computers are able to do them. Additionally, Tukey comments on the tools for EDA (1977, p.17): "*We do not guarantee to introduce you to the 'best' tools, particularly since we are not sure that there can be unique best."*

The book Exploratory Data Analysis introduces many statistical methods two of which were categorized as "main-highway methods", regression and analysis of variance. They are used by many analysts, but at the time of publishing the book, they were used more for confirmatory rather than exploratory data analysis. According to Tukey, their most important uses are often in exploration. From his point of view (Tukey 1977, p.7): "*Exploratory and confirmatory can -- and should -- proceed side by side."*

Exploratory Data Analysis can be described as detective work. Tools and methods are needed by a detective investigating a crime as well as by an analyst conducting an EDA investigation

– in both cases they increase understanding of the situation and help to narrow the conclusions to a more practical subset. The goal of Exploratory Data Analysis is to uncover indicators, which can then be studied further by either digging deeper with EDA or conducting a confirmatory data analysis of the most likely focus area for improvement. When planning any analysis, such as confirmatory data analysis, the exploratory mindset should be applied, because it exposes a deeper inquiry which might yield interesting results. Tukey comments on this subject (1977, p.19)*: "Restricting one's self to the planned analysis -- failing to accompany it with exploration -- loses sight of the most interesting results too frequently to be comfortable."*

3.2.2. Compaq Approach to Using Statistics in Business-Decision Making

In 1990, Corporate Quality Office of Compaq Computer hired an external consulting firm to create a body of knowledge of statistical methods to support internal quality professionals. Based on that work, a comprehensive approach to integration of statistical methods into a structured problem-solving method was developed. The method and training materials were shared among Compaq's business partners and suppliers as well as to all quality professionals inside the company. As part of the development of the body of knowledge, an 8-step process for data driven business-making was identified – the Compaq equivalent of DMAIC. This body of knowledge was delivered in two courses titled Data 1 and Data 2. The first course included seven statistical tools while the second course had eleven. For each statistical tool, the actions necessary for applying it for problem investigation were identified. (Compaq, 1991)

Logic of the 8-step process used by Compaq is illustrated in Figure 9. The first step defines the business problem. Once the problem is known, a relevant statistical test can be selected. After that, the test conditions are defined. Step four collects data. In step five, the statistical analysis or test is performed. Next, the analysis and test results are summarized and then applied to the business decision. Finally, the results are recorded and published.

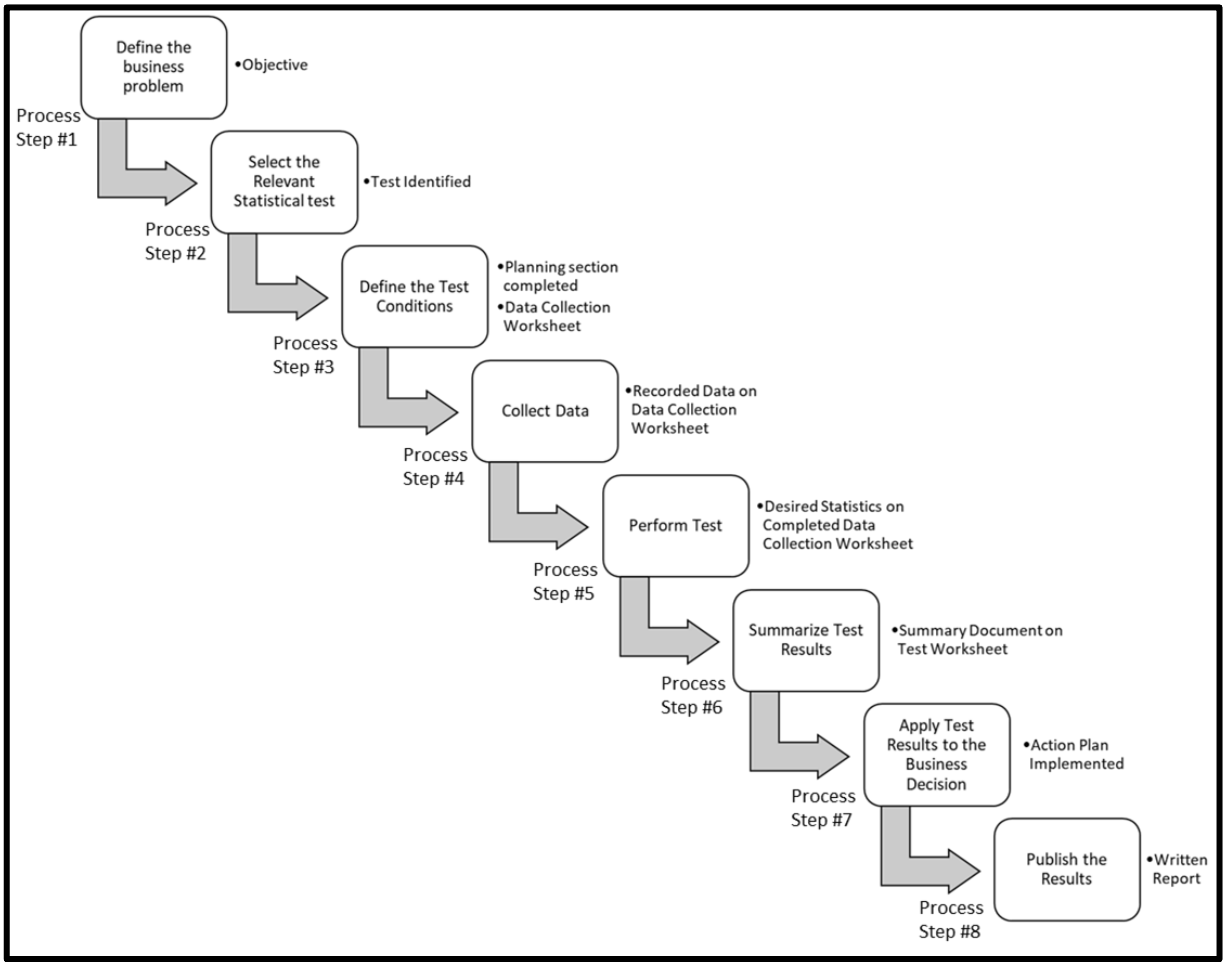

Define the business problem
•Objective
Process Step #1
Select the Relevant Statistical test
•Test Identified
Process Step #2
Define the Test Conditions
•Planning section completed
•Data Collection Worksheet
Process Step #3
Collect Data
•Recorded Data on Data Collection Worksheet
Process Step #4
Perform Test
•Desired Statistics on Completed Data Collection Worksheet
Process Step #5
Summarize Test Results
•Summary Document on Test Worksheet
Process Step #6
Apply Test Results to the Business Decision
•Action Plan Implemented
Process Step #7
Publish the Results
•Written Report
Process Step #8

Figure 9. 8-step Process for Structured Problem-Solving by Compaq. Adapted from Compaq. (1991)

3.2.3. The Principle of Reverse Loading

Harry (1997) describes how engineers typically approach analysis tasks in his book: *The Vision of Six Sigma: A Roadmap for Breakthrough*. The first question they have is: “Where can I get some data to accomplish the task?” After that, they will just use the available data, because many times the data collection points, and data items are fixed in the process. This creates a situation where the type of data, and therefore its typical analysis, has become predefined. Once an engineer has this data in hand, they will select the statistical tool based on the data, which has also been restricted by the choice of the measurement system. The statistical tool selected for the analysis therefore might not be able to provide some information about the situation from a different viewpoint than the original object of the analysis. The principle of reverse loading is a method that is intended to overcome the issue of analyzing data only to conclude that it cannot be used to answer the original question. (Harry, 1997, section 21.3)

The logic of principle of reverse loading is illustrated in Figure 10. The first part is planning, which is performed by answering five different questions and answers and this also acts as a roadmap for the analysis. The first question is: “What do you want to know?” The answer to the first question is needed to address the second question: “How do you want to see what it is that you need to know?” The answer to that question is needed to select the tool which is the third question: “What type of tool will generate what it is you need to see?” The needed data for collection is defined based upon the tool that will be used, while the last question: “where can you get the required data?” seeks to identify the location of the data collection point. Each of these questions has a single word to summarize the activity of that level. When starting from the last question, the following words are used: Location, Data, Tool, See and Know. From the practical perspective, the engineer has defined the location, then the data is collected, as after that a tool is used for analysis and the results are evaluated, which leads to knowledge. When compared to an analysis based only on the currently available data, these questions might reveal a very different set of observations by applying the principle of reverse loading. (Harry, 1997, section 21.3)

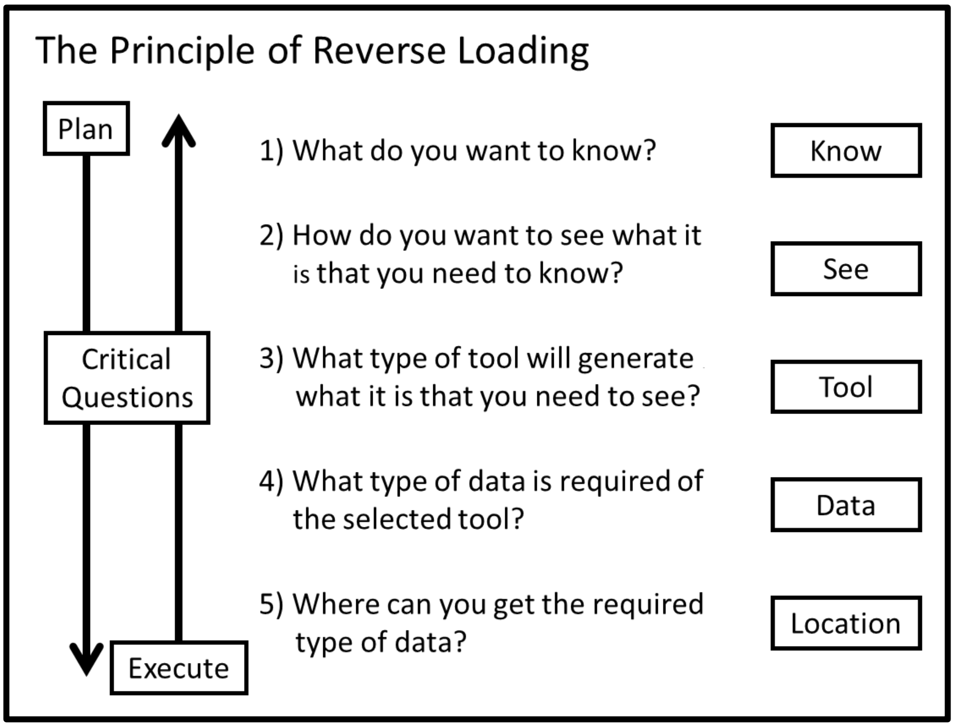

The Principle of Reverse Loading
Plan
Critical Questions
Execute
1) What do you want to know?
2) How do you want to see what it is that you need to know?
3) What type of tool will generate what it is that you need to see?
4) What type of data is required of the selected tool?
5) Where can you get the required type of data?
Know
See
Tool
Data
Location

Figure 10. The Principle of Reverse Loading. Adapted from Harry. (1997, section 21.14.)

3.2.4. Logic of Data Analysis in Shainin System

Dorian Shainin developed the Shainin System, which is best applicable in high and medium volume production, where data generation is fast, and the process consumes relatively low cost. The Shainin System is based on the logic that problems tend to be defined by a dominant cause of variation influencing the process output measure. The Pareto principle is used as basis for evaluation of this presumption. A process called Progressive Search is used to find the dominant causes. Progressive Search analyzes families of variation, defined so that each family contains similar types of events. For example, different locations or machines can be families – a rational sub-group of distinctive events that are influenced by the same causal factors. The process of Progressive Search investigates variation of dominant causes by dividing the remaining inputs into mutually exclusive families. The search is successful when all factors except for one family has been eliminated as the source of the cause. (Steiner et al., 2008)

When data analysis is conducted in a Progressive Search then "leveraging" can be used to gain comparative information about the process. The principle of leveraging is applied to the extreme values of the historical performance. These are the process performance conditions that deliver the extreme high and low levels of results when compared against each other. The "Best of the best" (BOB) condition is the best performance of the process while the "worst of the worst" (WOW) is the worst operating condition of the process. In order to do this analysis well, the team doing the analysis needs to have a deep understanding of the process practicalities. Only obtaining statistical knowledge is not enough; empirical knowledge is needed to construct the data families and understand what happens under different conditions of process operations. (Steiner et al., 2008)

The process diagnostic journey described by the Shainin System is illustrated in Figure 11. After defining the project, emphasis is placed on establishing an effective measurement system, because problem-solving relies on data that is both relevant and reliable. The next step in this diagnostic journey is to generate clues through Progressive Search. Output from a Progressive Search includes a variable or set of variables (step four) which could be contributing to the dominant cause. Step five conducts a statistically designed experiment to confirm that the list of variables related to the dominant cause are significant in creating the observed experimental effect. In step 6, results of the experiment are analyzed, and a decision is made whether or not the dominant cause has been discovered. This dominant cause is called a "Red X" in the Shainin System. If the effect of cause by the variable or variables cannot be demonstrated, the algorithm returns back to step three, where additional clues regarding the potential "Red X" must be generated. (Steiner et al., 2008)

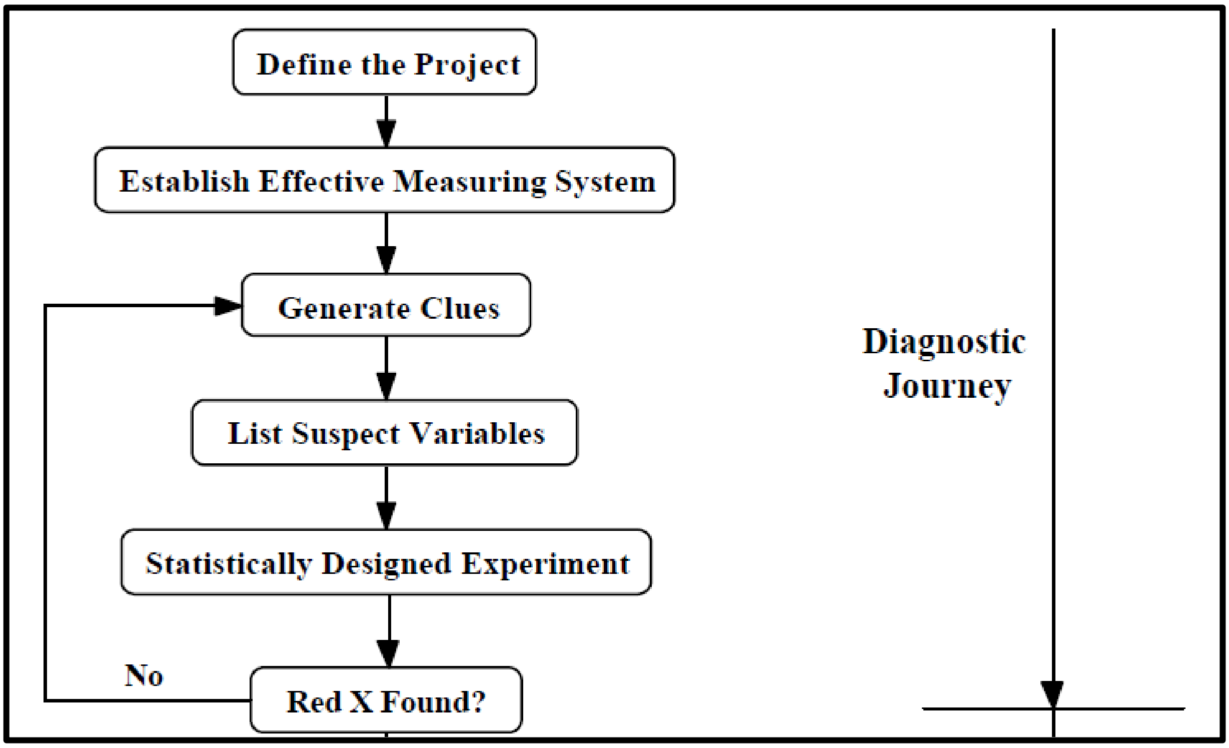


Figure 11. Diagnostic Journey of Progressive Search (Steiner et al., 2008, p.8)

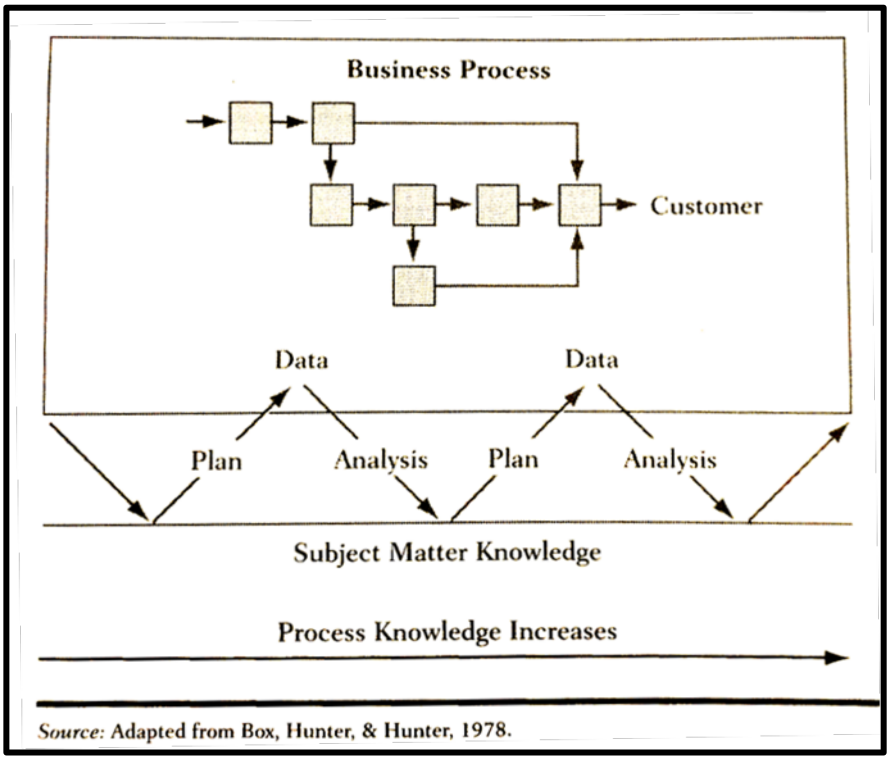
3.2.5. Mental Models of Statistical Thinking

Over the years, several thinking models have been proposed by different authors on how to approach statistical problems. Table 1 summarizes some of the proposed models by three authors. All of them basically represent the same path with different names and different levels of detail. (Notz, 2012, pp. 194-195)

Table 1. Different Mental Models for Statistical Thinking. Adapted from Notz. (2012, pp. 194-195)

Author	**Steps**	**What happens in the step?**
De Veaux et al. (2009)	1. Think	What do we know? What do we hope to learn? Are the assumptions and conditions satisfied?
	2. Show	The mechanics of calculating results
	3. Tell	Reporting the findings
Moore (2010)	1. State	What is the practical question, in the context of the real-world setting?
	2. Plan	What specific statistical operations does this problem call for?
	3. Solve	Make the graphs and carry out the calculations needed for this problem.
	4. Conclude	Give your practical conclusion in the setting of the real-world problem.
Peck et al. (2008)	1. Understanding the problem	
	2. Deciding what to measure and how to measure it	
	3. Data collection	
	4. Data summarization and preliminary analysis	
	5. Formal data analysis	
	6. Interpretation of results	

Hoerl and Snee present a mental model of Statistical Thinking that was adapted from the process proposed by Box in their book: *Statistical Thinking* (Hoerl & Snee, 2002, p. 42) The Hoerl-Snee model is presented in Figure 12. The model is a v-shaped process, which represents activities involving both subject matter knowledge of the process and data gathering from the business process. Subject matter knowledge is a good starting point for analysis, because it is the base on which the whole process of investigation is built. Subject matter knowledge should be acquired through the combination of academic study and practical experiential knowledge. (Hoerl & Snee, 2002, p. 41)



Source: Adapted from Box, Hunter, & Hunter, 1978.

Figure 12. Statistical Thinking Model adapted by Hoerl and Snee (2002, p. 42)

The objective of this Statistical Thinking Model is to improve business processes by gaining deeper process knowledge and applying it to manage the process better. This is accomplished by creating hypotheses about causal linkages within the process and testing them on observational data sets. Analysis is done by statistical methods by analyzing the stated hypothesis and the observed process variation. Hoerl and Snee (2002, p. 42) comment on their process: "*Creativity is sparked by the data. We are forced to rethink or revise our hypothesis in such a way that it could explain the observed data.*" Revising a hypothesis leads usually to the situation where more data and additional analysis are needed. In order to develop a satisfying solution, the processes of planning data collection and analysis will typically need to be repeated several times (Hoerl & Snee 2002, p. 42)

3.2.6. Model of Exploratory Data Analysis by Mast and Kemper

According to Mast and Kemper (2009, p. 366) the role of EDA is to screen data for clues, which can generate ideas that may be used to develop hypotheses. In problem-solving, EDA can be applied when some data is collected, but hypotheses have not yet been defined about the problem's causes. De Mast and Kemper developed a three-step model for EDA. The model is shown in Figure 13. Three steps in the process of EDA are (de Mast & Kemper 2009, p. 369):

1. *Display the data*
2. *Identify salient features (which stands out from what was expected)*
3. *Interpret salient features (which could have caused the unexpected observations)*

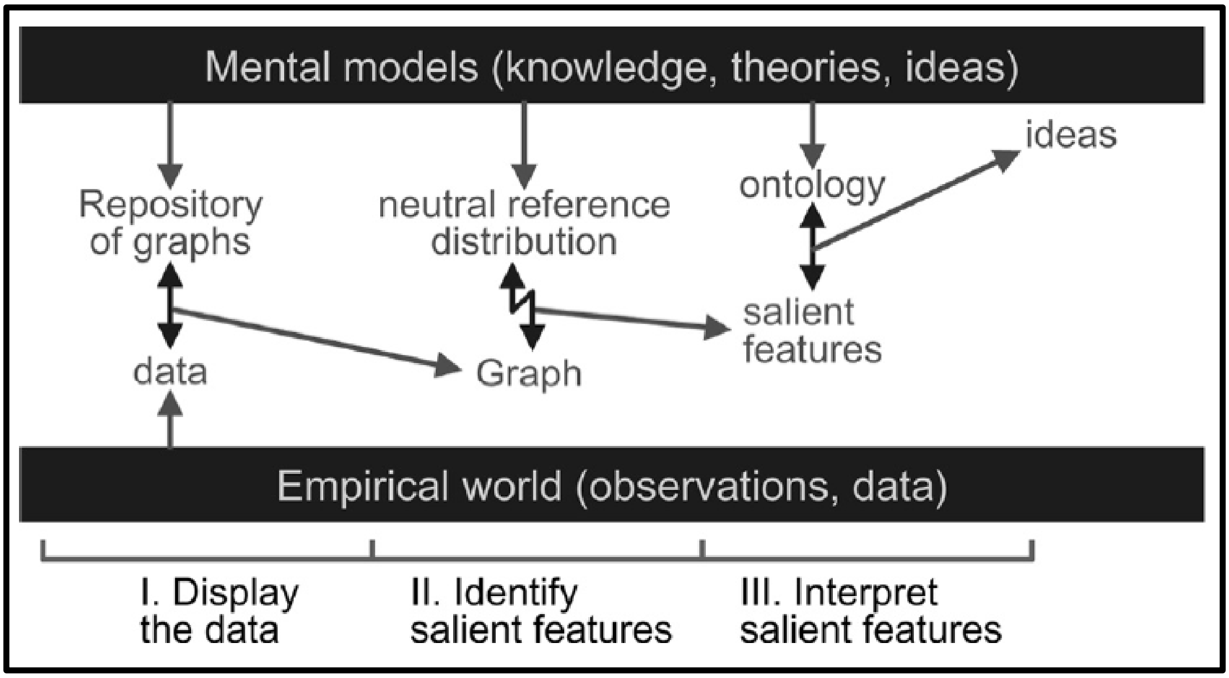

Mental models (knowledge, theories, ideas)
Repository of graphs
neutral reference distribution
ontology
ideas
salient features
data
Graph
Empirical world (observations, data)
I. Display the data
II. Identify salient features
III. Interpret salient features

Figure 13. Model of Exploratory Data Analysis by de Mast and Kemper (2009, p. 369)

The first step in the process is to display data. De Mast suggests multiple statistical tools to be used for displaying the data. Tools and reasons for using them are shown in Table 2. The emphasis of statistical tools suitable for EDA is on the graphical tools, because the aim is to optimize the pattern recognition potential of the human brain. Working memory has a limited capacity, which means that many times raw data is too complex, while summary tables display aggregate statistics, which hides the distribution of data. If graphs are used to present the distribution of data, different patterns can be detected rapidly just by using the human eye. (de Mast & Kemper, 2009)

Table 2. Tools used in Exploratory Data Analysis by de Mast and Kemper. (2009, p. 369)

STATISTICAL TOOL	WHY IT IS USED
Histograms	Distribution of one-dimensional data
Time series plots and boxplots per time unit	Data distribution over time
Boxplots and dot plots per stratum	Distribution within and across strata in the dataset
Principal components analysis and scatter plots	Data distribution on a plane

The second step in the process of EDA is to identify salient features. In this step, the analyst has an idea how data from the observed process should appear, in other words, the reference distribution. If the data distribution illustrated in the graph or graphs is unexpected, it can be caused by a salient feature which characterizes the most important source of variation. (de Mast & Kemper, 2009, pp. 369-370)

The third step in this EDA process interprets these salient features. In this step, the reason for the causes of patterns in this data, are theorized and described. In the end, these performance theories can be turned into hypotheses, and tested by confirmatory data analysis. The discussion to interpret the results should take place with people familiar with the process, for example, operators of the process under analysis. The term “ontology” used in the model means to create an understanding of the real-world concept, or concepts, which could have caused the pattern in the data – what actually is or exists. An explanation for a pattern shift in data can be, for example, the presence of two different data distributions in the process which is caused by an undisciplined application of dominant process factors (e.g., in injection molding the random choice of one of two molds which have distinctly different cavity measurements). Reasoning by analogy is one strategy to improve understanding of the characteristics regarding the salient features. A systematic description of the problem may be created by using the questions: who, where, what, when, why, and how? This sequence of questions is referred as the “5W’s and 1H” method which is applied by Japanese problem-solving teams. (Watson, 2019c; de Mast & Kemper, 2009, pp. 370-371)

According to Watson (2019c), this set of questions is sometimes referred to as the Kipling Method, due to a poem that appeared in Rudyard Kipling's book: "Just So Stories." (Kipling, 1902, pp. 29-30):

> *"I keep six honest serving-men*
>
> *(They taught me all I knew);*
>
> *Their names are What and Why and When*
>
> *And How and Where and Who."*

Sometimes there is a situation where a salient feature has been identified, but the analyst and the team have not been able to find a cause that is satisfying. The typical reason for this is a lack of process knowledge, in other words, the team's ontology is not detailed enough. A way forward in such situations is to investigate the phenomenon by observing and studying it in more detail. De Mast and Kemper emphasized the importance of blending statistical expertise and content expertise in such investigations. (de Mast & Kemper, 2009)

3.2.7. Process of Exploratory Data Analysis by Watson

The most recent approach to EDA was developed by Watson. He has linked different tools into a sequential process to find the meaningful rational sub-groups within the scope of the problem. His approach to EDA is presented in Figure 14. Each statistical method sequentially increases knowledge about the process. These methods will be covered in more detail in section 3.5. of this thesis.

If one of the sub-groups analyzed, for example, by Analysis of Variance shows a difference, the next EDA process can be initiated at that process step by looking at the time series of its data flow using an Individuals Chart. This is the means by which a drill-down using the EDA methods is guided. In the other words, the process stops when problems are identified and then divided further into rational sub-groups for the subsequent analysis. Analyses of these rational sub-groups guides the search for those factors, which can be controlled and changed to solve the problem situation. (Watson, 2018a, p. 163)

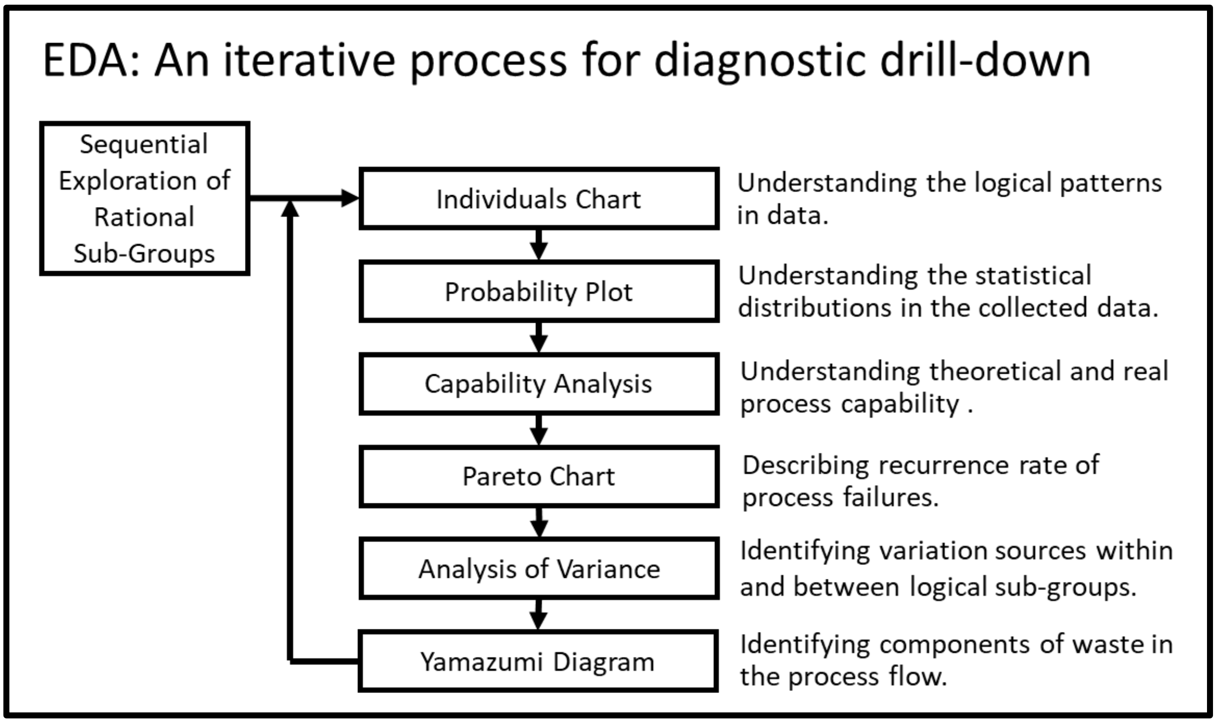

EDA: An iterative process for diagnostic drill-down
Sequential Exploration of Rational Sub-Groups
Individuals Chart
Understanding the logical patterns in data.
Probability Plot
Understanding the statistical distributions in the collected data.
Capability Analysis
Understanding theoretical and real process capability .
Pareto Chart
Describing recurrence rate of process failures.
Analysis of Variance
Identifying variation sources within and between logical sub-groups.
Yamazumi Diagram
Identifying components of waste in the process flow.

Figure 14. Iterative Process of Exploratory Data Analysis by Watson (2018a, p. 121)

Watson's approach to Exploratory Data Analysis applies this iterative, sequential approach to improve the problem definition and to locate the source of process variation or loss. This analysis is initially performed in the Define phase of a DMAIC improvement project to establish the focal point of an inquiry and its boundary conditions. The goal of this EDA process is to understand what kind of logical relationships exist in the data with respect to the physical world. (Watson, 2018a, p. 168)

3.3. Analytical Objectives of EDA

Exploratory Data Analysis can have many objectives ranging from hypothesis generation to improving problem definitions. The most profound reason for doing EDA was stated by Tukey (1977, p. 5): "*A basic problem about any body of data is to make it more easily and effectively handleable by mind"* That is done by creating graphs, and the best type of them is described by Tukey: "*forces us to notice what we never expected to see"* Thus, the first objective of EDA is to illustrate information in a data set in a way that is easily understandable and in a format that is intuitively clear to those who need to interpret the information and make decisions based upon that data.

One way to illustrate something new from a data set or process is to delve deeper and discover patterns that had not been previously recognized. According to Tukey (1977, p.5) "*Anything that looks below the previously described surface makes the description more effective."* In other words, the second objective of EDA is to look deeper into a problem than typically occurs with standard data reporting.

Another perspective of Exploratory Data Analysis is to understand why it is used. In the context of this thesis, the focus is on structured problem-solving, which means that EDA is performed when attempting to solve a problem. Watson states that (2018a, p. 63): "*The purpose of Exploratory Data Analysis (is) to rapidly investigate the performance of a process and determine quickly where the improvement effort should be focused for further action."* Therefore, a third objective of EDA is to guide the development of process improvement efforts.

3.3.1. Difference between Enumerative and Analytical Views

Deming defined two perspectives for data analysis that describe the utility of statistical applications: enumerative and analytical. These statistical applications represent different points of view regarding data analysis. In an enumerative study the purpose is to describe the overall performance of a population. This type of study does not answer the question: "Why," but address questions like "How many" or "What" type?" It is used to illustrate overall performance and determine risk of non-performance across the population of historically observed results. On the other hand, an analytical study seeks to improve future performance by discovering patterns in the historical data which reveal the process structure and its cause-and-effect-system which influence variance in the outcome results. (Watson, 2018a, pp. 32-37; Deming, 1975)

Examples of these two different types of data strategies:

- Enumerative analysis answering the question: ”What is a fair price to pay for raw material based on the historical distribution of its measured quality?”
- Analytical analysis answering the question: ”What are the right settings for this manufacturing equipment to achieve a desired quality level?”

One way to distinguish between analytical and enumerative problems is to approach them from the data perspective. How much data there is available for analysis regarding the phenomenon under analysis? If there is a 100% sample of the data then the study is enumerative and the limiting factor will be the inability to perceive the relationships between the individual data items for patterns and correlations. Compared to an analytical treatment of the data, the enumerative treatment is inconclusive with respect to causation no matter how much data can be used in the analysis. Such summary data will never yield profound knowledge about the way processes vary over time. (Deming, 1975)

According to Deming (1975, p. 148), there are some limitations of statistical inference.: "*All results are conditional on:*

A. "the frame whence came the units for test

B. the method of investigation (the questionnaire or the test-method and how it was used)

C. the people that carry out the interviews or measurements

D. *In addition, the results of an analytic study are conditional also on certain environmental states. The exact environmental conditions for any experiment will never be seen again.*"

Information extractable from analytical studies cannot be complete as it does not include all of the process factors and environmental conditions. Thus, analytical studies need to be supported with detailed knowledge of the subject-matter and operational conditions. (Deming, 1975)

3.3.2. Normality of Data

While the academic approach to statistical methods relies upon proof based on the normal distribution, applied statistics is not so neatly defined. Data from real world processes is not usually normally distributed. (Cogollo-Florez et al., 2017) For example, time series data that is used in analytical studies typically will have a long tail as it seeks to minimize or maximize performance as its desired state. Although many statistical tools require normal data to operate according to their mathematical theory, the use of mathematical transformations to normalize the data has an unwanted side effect that is not helpful for interpretation of real-world data. Transformations will tend to distort the graphical view of the data once the transformation is reversed back to the original state of the data. This distortion makes it difficult for those who use the data on a routine basis to understand and interpret it properly. Since the purpose of Exploratory Data Analysis is to increase the knowledge of the process performance, a practical approach to developing this understanding is more significant than maintaining statistical purity in the interpretation. (Watson, 2018a)

Shewhart (1939, p. 92) stated two rules for presentation of data that apply to this EDA treatment of non-normal data:
"Rule 1: Original data should be presented in a way that will preserve the evidence in the original data for all the predictions assumed to be useful

Rule 2: Any summary of a distribution of numbers in terms of symmetric functions should not give an objective degree of belief in any one of the inferences or predictions to be made therefrom that would cause human action significantly different from what this action would be if the original distribution had been taken as a basis for evidence."

Shewhart's two arguments support this approach that data should not be transformed, when it is presented to people working in the process. Actions made based on non-original data (e.g., transformed data) can be different from those based on the original data. The reason for this distinction is that information is more difficult to recognize when the data is transformed as the shape has been changed and many workers may not be able to recognize it so they can read and interpret it as they do on a daily basis. (Watson, 2018a, p. 21)

3.3.3. Best of the Best (BOB) and Worst of the Worst (WOW)

Sometimes the approach to problem-solving is to find out an inferior product and start asking what is wrong about it and what has caused that. The Shainin System supports this approach by comparing the best parts and the worst aspects of the alternatives. Abrahimian (2009, p.21) states that: "*Don't ask "What is wrong with the worst bond pads?", ask "What is different between the best and worst?"*""

Figure 15 illustrates an example of Wiper Sweep Angle data distribution and how Best of the Best and Worst of the Worst data points are located in a graph where the location of the performance indicator is nominal is best. Examples such of performance measures are: on-time delivery, technical tolerance limits and standard cost. (Spiridonova & Watson, 2018, p.29)

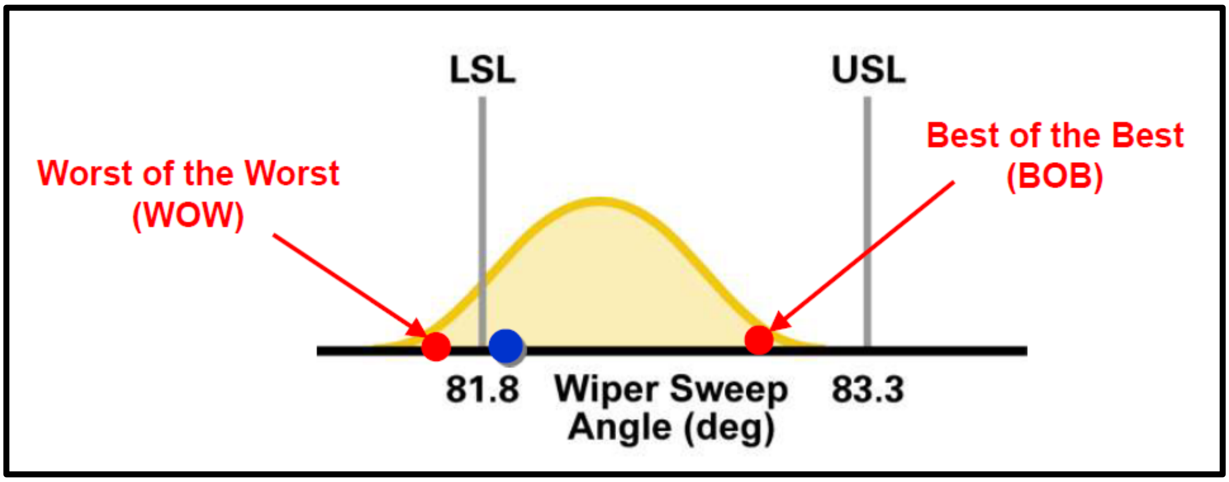


Figure 15. Comparing Extremes – BOB and WOW (Abrahimian, 2009, p.9)

Figure 16 demonstrates where BOB's and WOW's are located in distribution when the rule of performance is smaller is better. In other words, the smaller the value, the better. Examples of such performance measures are: cycle time, defects, accidents, failure rates and transaction cost. (Spiridonova & Watson, 2018, p.30)

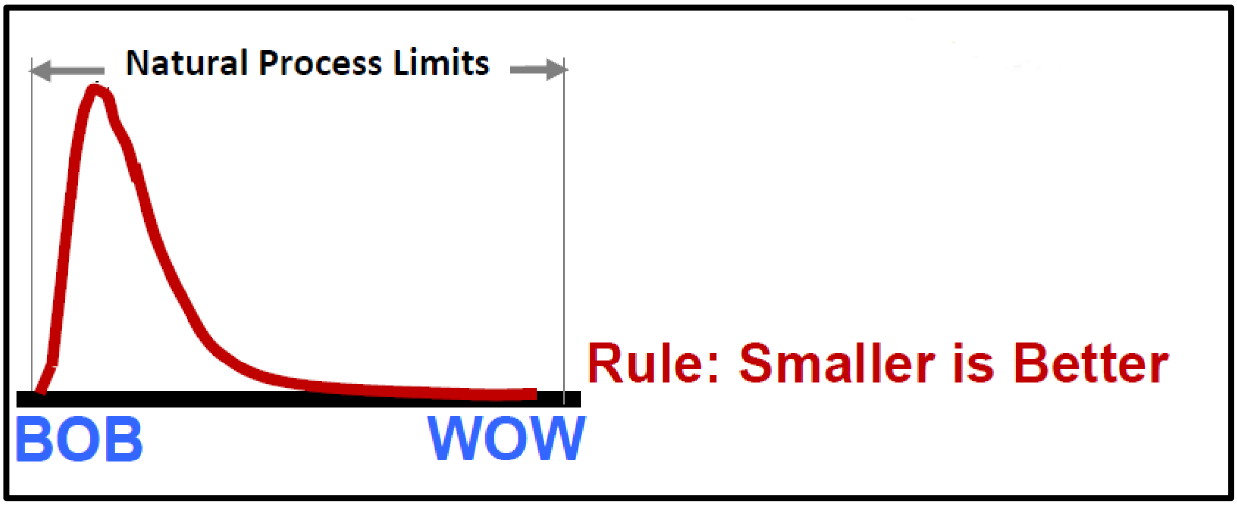


Figure 16. Position of BOB's and WOW's when Smaller is Better (Spiridonova & Watson, 2018, p.30)

Figure 17 demonstrates the location of in distribution BOB's and WOW's when the rule of performance is bigger is better. In other words, the bigger the value, the better. Examples of such performance measures are: productivity, satisfaction, motivation, market share, profit and revenue. (Spiridonova & Watson, 2018, p.31)

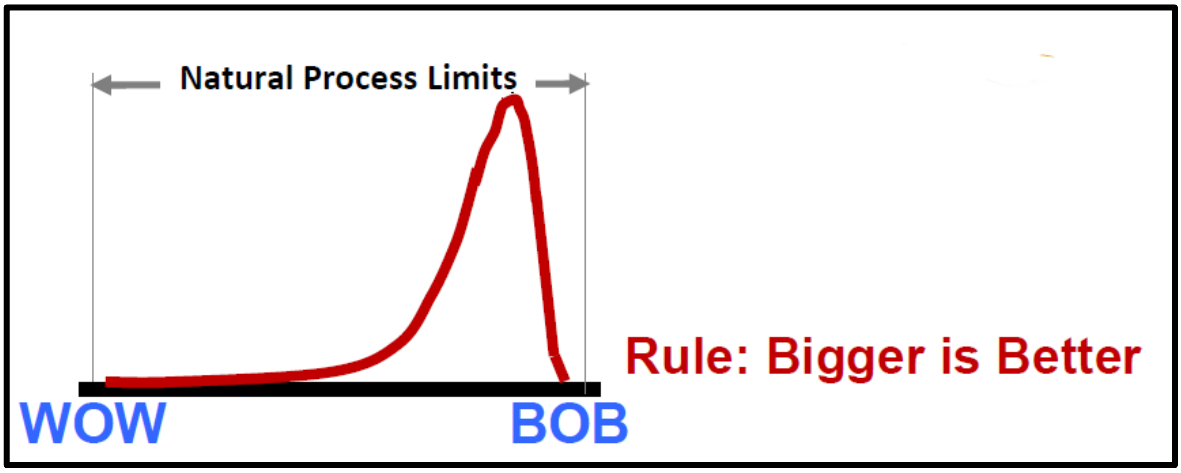


Figure 17. Position of BOB's and WOW's when Bigger is Better (Spiridonova & Watson, 2018, p.31)

3.3.4. Rational Sub-Groups

A rational subgroup is a subset of data defined by a specific factor such as a stratifying factor (machine, operator, supplier, material lot number, etc.) or a time period. The items included within a sub-group possess common characteristics that are unique (e.g., produced under similar conditions or with similar materials or by the same machine) while different sub-groups are created by separating these factors. Rational subgrouping separates factors that generate special cause variation (variation between these subgroups) which is caused by specific, identifiable factors. Variation observed within these subgroups is due to their common causes. Creating a Y = f(x) function of a process is in practice rational sub-grouping as the elements effecting the process are divided into rational sub-groups. Graphically, the relationship between rational sub-groups can be presented with a tree diagram, a fishbone diagram or a mind map. (Watson & Spiridonova, 2019)

One cornerstone of the Shainin System is Progressive Search, which means investigating something called families of variation. They are defined so that each family contains similar type of events that act at the same location or in the same time span, for example, one shift can be a family or a specific machine, specific supplier or batch. The family of a specific supplier can be divided into mutually exclusive sub-families, for example, each batch can be sub-family of a supplier family. Rational sub-groups and families of variation are the same concept. (Steiner et al., 2008, pp. 8-10)

3.3.5. Variation

Variation can be categorized in a few different ways. One way to categorize it is as common cause variation and special causes of variation. This division of variation is a basis for statistical process control, where creation of a stable process is the desired outcome. A stable process has only common cause variation. In practice it means that all data points of the process output measures are within control limits of the process control chart (these limits are plus or minus three standard deviations around the mean). If there are data points outside these control limits, then this extreme variation is a sign of special cause of variation, thereby indicating that something unusual is taking place in the process. Table 3 lists differences between these two types of variation. (Hoerl & Snee, 2002, p. 45)

Table 3. Differences between Common- and Special-Cause Variation (Hoerl & Snee, 2002, pp. 45)

Special-Cause Variation	**Common-Cause Variation**
Temporary and unpredictable	Always present
Few sources but each has a large effect	Numerous sources but each has a small effect
Often related to a specific event	Part of the normal behavior of the process
Process is unstable	Process is stable

Processes having special cause of variation are said to be unstable. If the factor causing the change in the process can be identified, it can be referred as assignable cause. An example of an assignable cause can be from stock markets, an unexpected change of interest rate by central bank. In manufacturing, it can be, for example, a change of supplier. If a process is unstable, the first thing to improve is to remove these special causes. Typically, removing the causes that are assignable to these observed special causes will improve the process significantly, because they have the biggest effect on the performance. (Hoerl & Snee, 2002, pp. 44-45)

3.4. Ishikawa Analysis as part of Exploratory Data Analysis

In Japan engineers were developing quality methods under the Quality Control (QC) Research Committee, part of the Japanese Union of Scientists and Engineers (JUSE). Value Engineering and Functional Analysis methods were developed in in the 1940s by Lawrence D. Miles, who was working at General Electric in the United States. These methods were introduced in Japan in the mid-1950s. The QC Research Committee was researching how the methods could be applied in quality improvement and was looking an answer to the question: "*How to develop a logical decomposition of the situation that frames the problem to address?*". The fishbone diagram was one of the methods developed by Kaoru Ishikawa to address the following problem: "*How to graphically depict a systems breakdown of the combined functions of a product and the process by which it was produced.*" (Watson & Spiridonova, 2019)

There are several tools which can be used for representing the breakdown of the functions of products and the processes. Fishbone diagram, mind map and tree diagrams can be used to do that. To honor Kaoru Ishikawa and his work usage of those tools are called as Ishikawa Analysis, because in the context of Exploratory Data Analysis all of the mentioned tools can be used for the same purpose.

3.4.1. Fishbone Diagram

The objective of using the fishbone diagram is to identify rational sub-groups. Aristotle was the first author to discuss the question of identifying components of a problem. Watson and Spiridonova (2019, p. 4) interpret Aristotle's book Categories in the following way: "*Here Aristotle first proposed that initiation of any scientific investigation should be to decompose or breakdown the issue addressed into its component elements"* DMAIC uses similar logic to drilldown to the detailed level sufficient to find the X's which can be controlled in order to improve the performance of the process.

Fishbone diagram can be seen as logical decomposition of the Y = f(x) relationship, since it defines the purpose, problem or process output. Figure 18 represents the classical fishbone diagram. It has seven M's, which are standard rational sub-groups identifying factors or functional elements representing sources of potential causes of problems for the process output. 7th M, money, was suggested by Watson and Spiridonova to represent the financial resources category. (Watson & Spiridonova, 2019, p. 19)

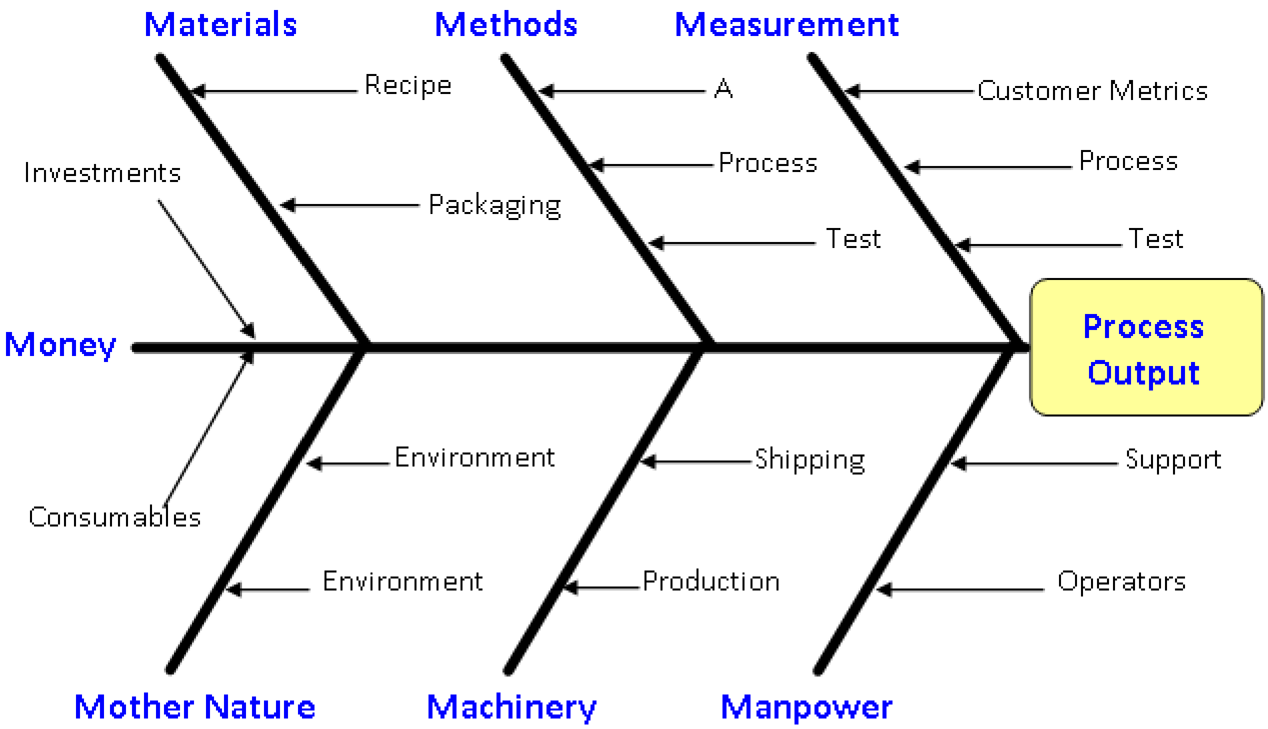


Figure 18. Classical Fishbone Diagram (Watson & Spiridonova, 2019, p. 19)

3.4.2. Mind Map

A mind map can be used instead of the fishbone diagram to present the breakdown of rational subgroups. Nowadays, there are many software packages able to do that, compared to the time when it was developed with only paper and pencil available. An example of a fishbone structured as a Mind Map is presented in Figure 19. (Watson & Spiridonova, 2019, p. 21)

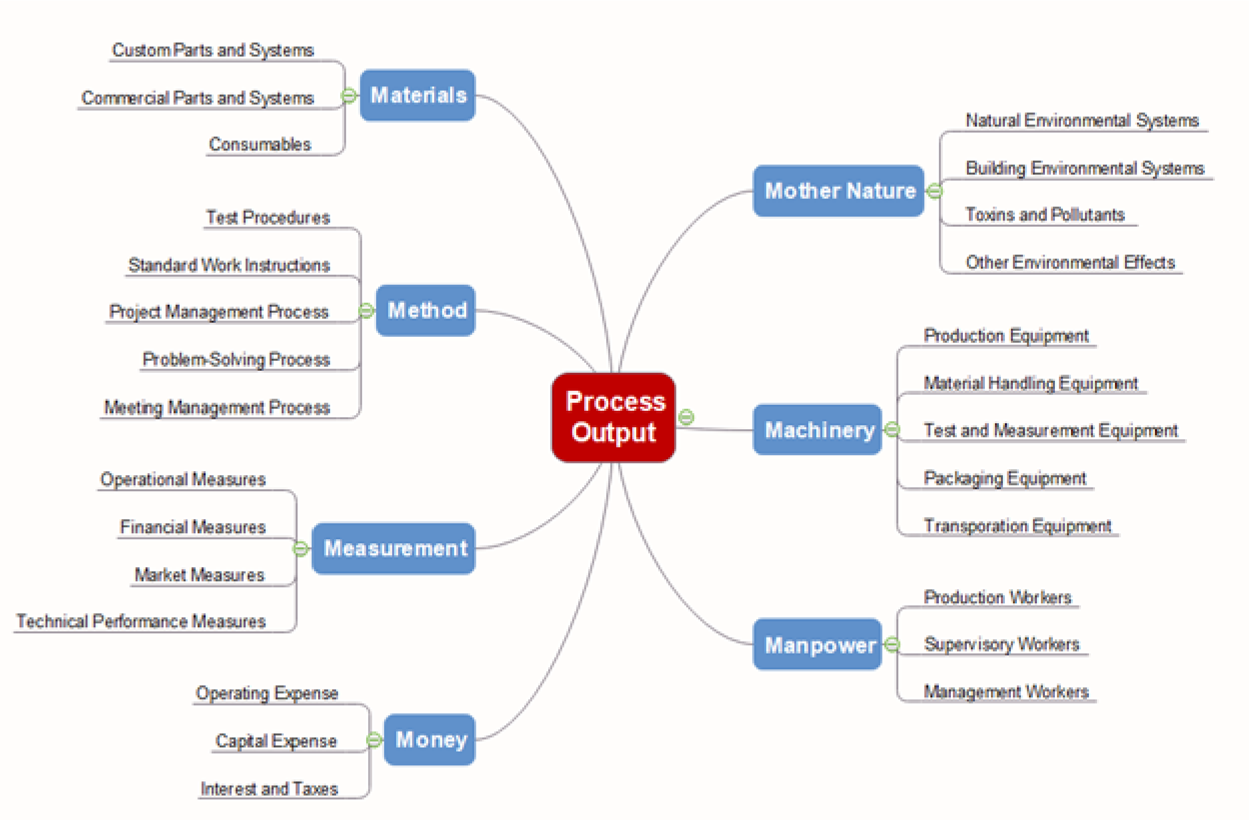


Figure 19. Using Mind Map with 7M logic (Watson & Spiridonova, 2019, p. 21)

The Mind Map was originally created by Tony Buzan, and it has been used for developing mental models, creative notetaking, and documentation of systems. In the center, there is the idea or, for example, the process output. It is then broken down into logical sub-groups, which can continue to a very detailed level. (Watson & Spiridonova, 2019, pp. 20-21)

3.4.3. Tree Diagram

Tree Diagram can also be used to break down the logical subgroups. It helps to develop thinking step by step to finer level of detail. It can serve as a communication tool, to explain to others the details. It can be used for process analysis, root cause analysis as well as for making a plan to achieve an objective. (ASQ, 2019)

Figure 20 presents a tree diagram, which is constructed by the information from the Fishbone and Mind Map by Watson and Spiridonova (2019, p.21). The most common forms of tree diagrams are from left to right and top to down. The example presents a left to right form, where the highest level of abstraction is placed on the left side of the diagram and the diagram starts to grow when going right. Machinery and Manpower are different subgroups and when going deeper into subgroups there are, for example, different types of equipment. (Sheehy et al., 2002)

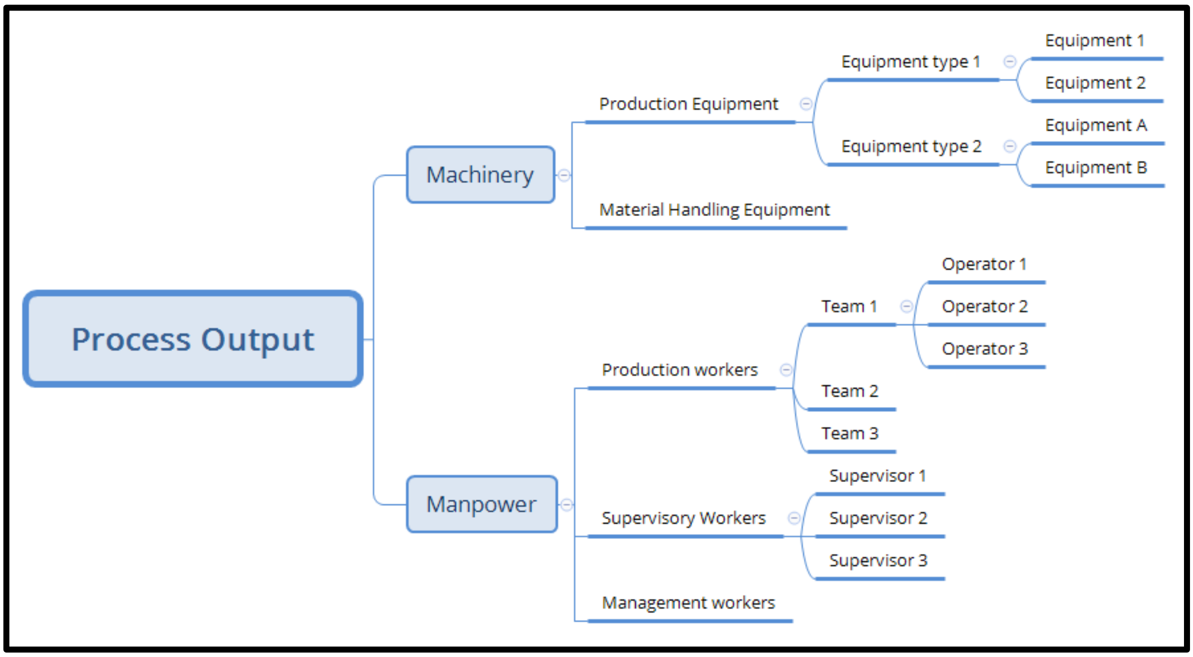


Figure 20. Tree Diagram constructed based on the information from Watson and Spiridonova (2019, p. 21)

3.5. Statistical Tools Applied in Exploratory Data Analysis

3.5.1. Individuals Control Chart

Individuals Control Chart is similar type of chart as run chart or time series chart. It gives an analytical view on the data, since it plots the data in a time-series order. It plots individual data points as well as control limits for the data, which makes it unique compared to other charts. Control limits are calculated to be + / - 3 standard deviations from the mean. Control limits can be used the observe, if there has been special cause of variation. (Watson, 2018a, p. 98)

Figure 21 presents an example of Individuals Control Chart. Data set is about pH and the mean of it is 5.985. Upper control limit is 6.390 and lower control limit is 5.579. When interpreting the data, it can be stated the there is a lot of variation and something is likely to affect the process. Good starting point would be point 8, which is above upper control limit. Point 20 is very close to lower control limit and variation does not look completely random throughout the data.

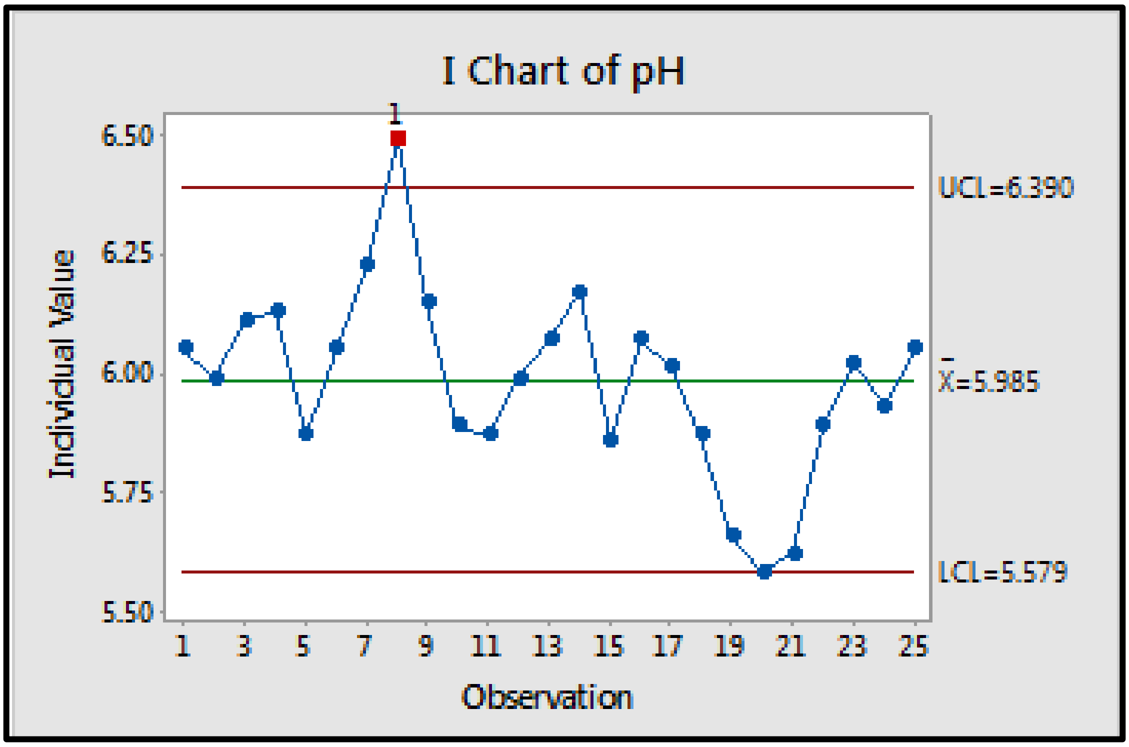


Figure 21. Individuals Control Chart (Minitab, 2019a)

Individuals Control Chart with stages is similar type of chart as Individuals Control Chart with a difference that the control limits are calculated separately for each stage. Figure 22 presents an I chart with stages, which is created with similar data as figure x, but a different that stages are added for every four data points. Four data points could represent one sub-group, which could for example be hour. If it would be hour, it would mean that a measurement was conducted every 15 min. When analyzing data with this assumption, it can be seen that the first hour process was stable compared to the other hours. On the second hour pH level was increasing all the time and during hour 5, it decreased all the time. During last four observations the pH level was as stable as during the first hour.

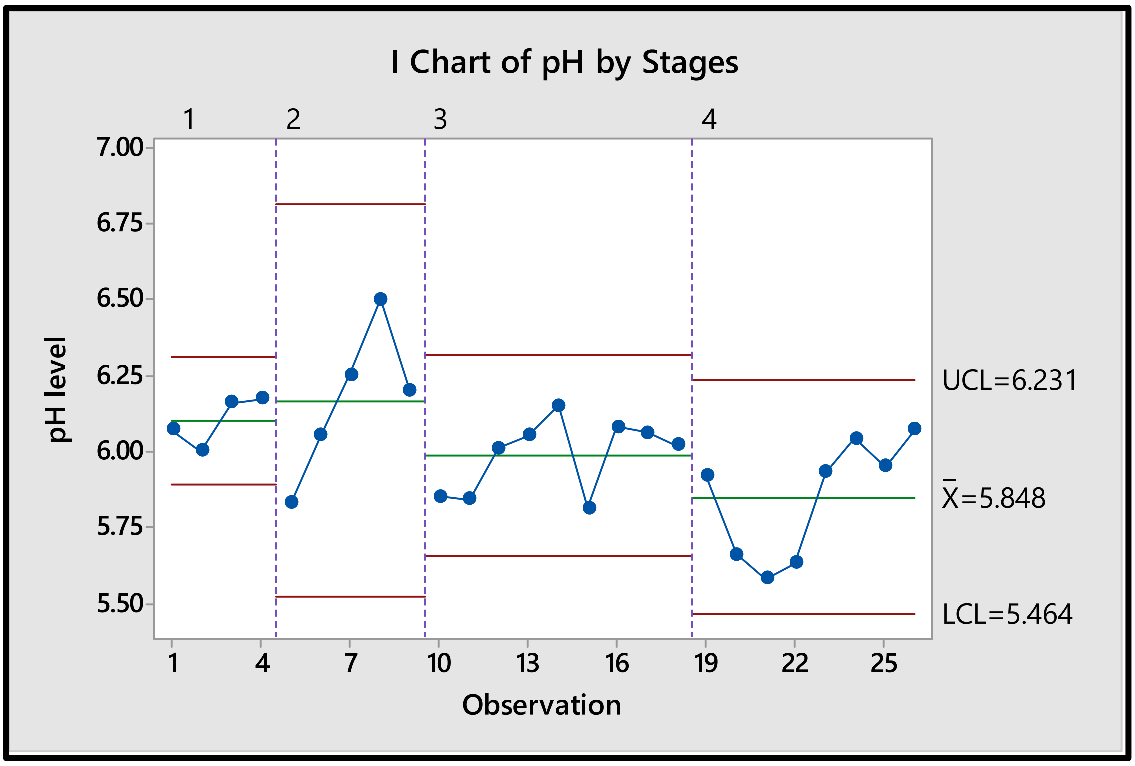
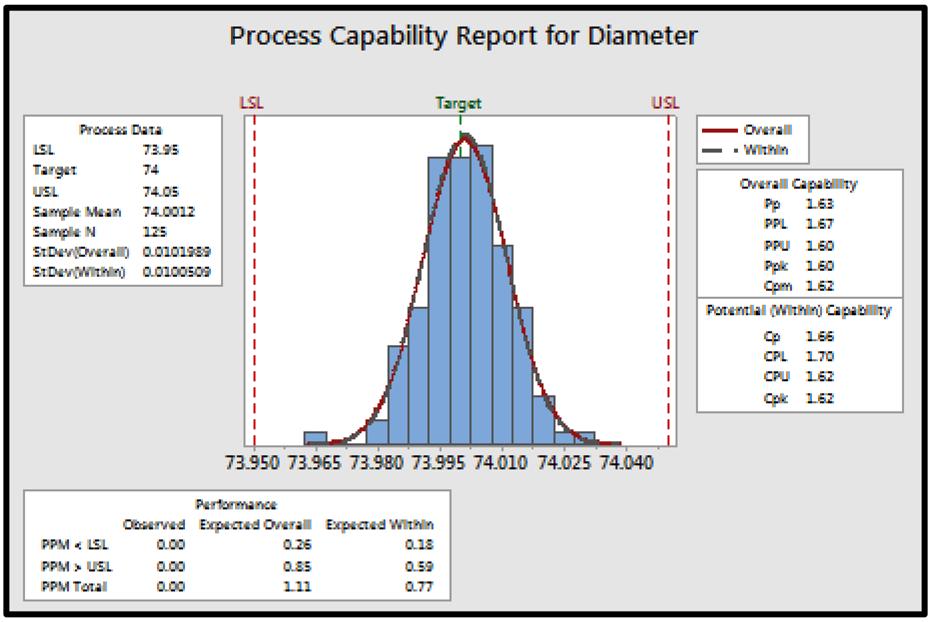


Figure 22. Individuals Control Chart with Stages, adapted based on the data from figure 21

3.5.2. Process Capability Analysis

Process Capability Analysis gives an enumerative view of the data, since it shows all the data compared to the customer requirements. Specification limits are set by the customer requirement and they can be thought of as the Voice of the Customer. The data collected from the process and analyzed with process capability analysis can be seen as the Voice of the Process. When these two voices are compared, the performance of the process can be seen. As rule of the thumb, the further the specification limits from the data values, the lower the risk that the process is no meeting customer requirements and being more capable. (Watson, 2018a, pp. 117-125)

Figure 23 presents a report of Process Capability Analysis. In the example a diameter of a part has been analyzed. The lower specification limit has been 73.95 and the upper specification limit 75.05. When focusing on the graphical analysis of the data, it seems to fit well between the specification limits and is well centered. It seems to be quite well normally distributed, because the bars are following quite closely the line drawn above them, which represents a shape of normally distributed data with the variation the data has. There are numerical values also in the process capability report, but they are out of scope in this thesis, since the focus is only on graphical interpretation of the tools.



Process Data	
LSL	73.95
Target	74
USL	74.05
Sample Mean	74.0012
Sample N	125
StDev(Overall)	0.0101989
StDev(Within)	0.0100509

Overall Capability	
Pp	1.63
PPL	1.67
PPU	1.60
Ppk	1.60
Cpm	1.62

Potential (Within) Capability	
Cp	1.66
CPL	1.70
CPU	1.62
Cpk	1.62

Performance			
	Observed	Expected Overall	Expected Within
PPM < LSL	0.00	0.26	0.18
PPM > USL	0.00	0.85	0.59
PPM Total	0.00	1.11	0.77

Figure 23. Process Capability Analysis Report (Minitab, 2019b)

3.5.3. Probability Plot

Probability Plots can be used to determine if a data set is normally distributed. Human eye can be used to determine if a data set is normally distributed or not, because probability plot is constructed so that the line it plots is straight in the case where the data is normally distributed. Additionally, there are confidence intervals around the line, to help to determine the normality of the data. In the other words, if all the data points are within the confidence intervals then that data is normally distributed with the confidence level chosen for the analysis (typically this will be based upon a 95% confidence interval). (Brook 2014, p. 125)

Figure 24 presents an example probability plot of wait times. The first observation is that the data is not normally distributed. There seems to be different kind of operating modes, where the most common waiting time is 3 seconds. Four times there has not been any waiting time and observations between 7 and 9. seem to follow function, because a straight line could be drawn between them.

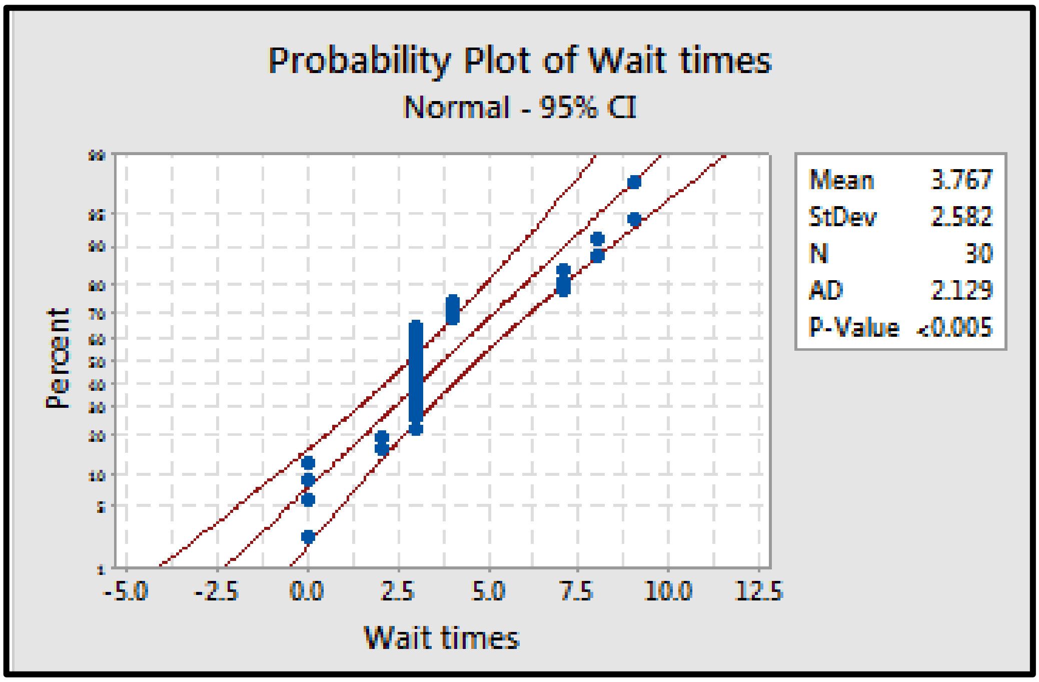


Figure 24. Probability Plot (Minitab, 2019c)

3.5.4. Pareto Chart

A suitable tool to determine the most common types of categories from a data set is Pareto chart. It plots the counts of individual categories stated in the data as well as cumulative frequency of them. The counts are plotted as bars and frequency as a line in the same graph. Additionally, the percentages of each category compared to the overall can be shown by a Pareto Chart. (Watson, 2018a, pp. 131-134)

Figure 25 represents a Pareto Chart of complaints. It can be seen that three complaint types covered 80% of all of them. Room was the topic of complaint 104 times representing 45.4% of all the complaints. Appliances were complained 45 times representing 19.7% of all complaints, summing up to 65.1% with room complaints. Third most common type of complaint was cleanliness with 43 complaints representing 18.8% of all the complaints. All together the top-3 complaint types represented 83.8%. With this information a problem-solving project could focus on the most common type of complaints and solving the root causes of those complaints.

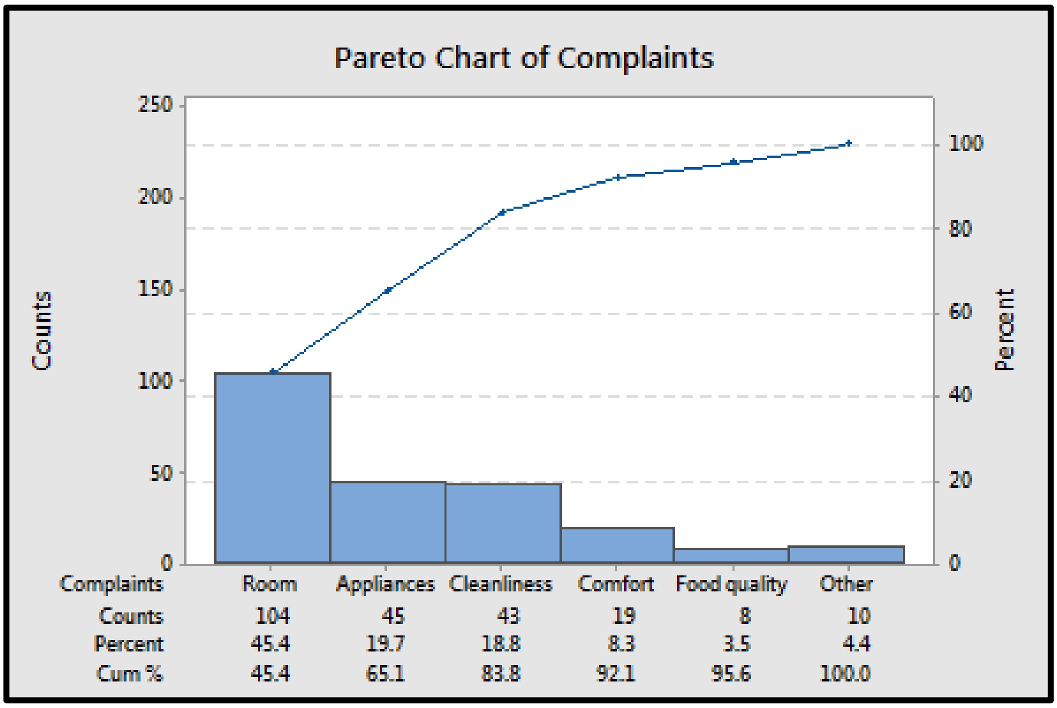


Figure 25. Pareto Chart (Minitab, 2019d)

3.5.5. Boxplots

The box plot was developed by John Tukey to display data in a visual way. The shape of the box plot reflects, to some extent, the distribution of that data set. Half of the data points are inside the box, and 25 % of the data in the whiskers, which are the lines on top and below the box. The whiskers can be 1.5 times the length of the box. If there are data points beyond that region, they are drawn as asterisks and called as outliers. The horizontal line in the box presents the value of mode. (Watson, 2015f, p. 42)

The size of the boxplot illustrates the amount of variation the data set has. It can be used to compare different data sets to each other. Figure 26 presents a situation where the height of some plants are compared based on the fertilizers used. There are three different subgroups in the presented situation, which are GrowFast, none and SuperPlant. When analyzing the box plots, there can be seen a difference between situations when fertilizers were not used (none) compared to SuperPlant. The shape of the box is practically the same, but the upper whisker is longer for SuperPlant, and the mode is higher. It means that in some cases SuperPlant has increased the height of the plants, but not every time. For GrowFast, the situation is different, because the lowest values presented by lower whisker is about the same as median for none. In other words, GrowFast has increased the height of the plants every time and the height of 75% of the plants has been >21 compared to >15 for none and >16 for GrowFast.

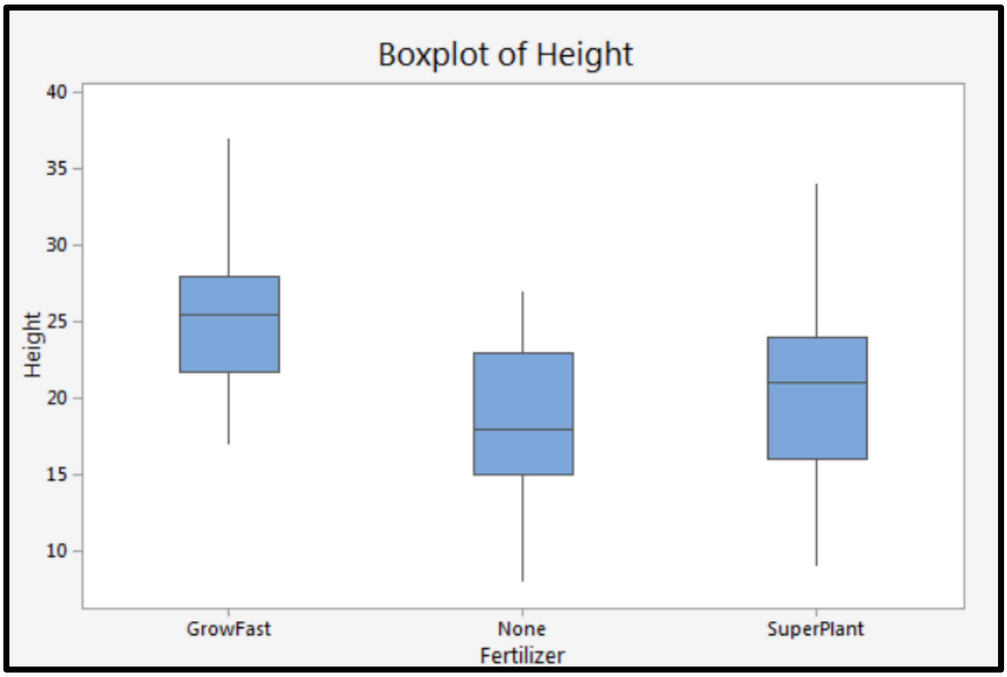
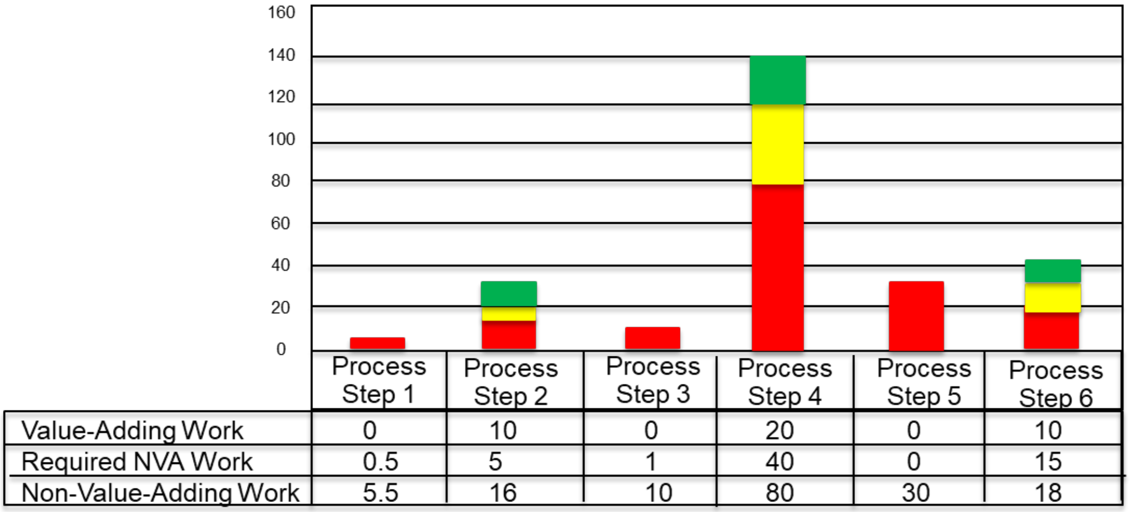


Figure 26. Boxplot (Minitab, 2019e)

3.5.6. Yamazumi Diagram

Yamazumi diagram or chart is a stacked bar chart, showing the cycle time of a process step divided into three categories: VA (value-adding work), NVA (work without added value) and NW (necessary work / required non-value-adding work). Cycle times are calculated in the production flow from the point when the operator starts the work to the point the process step is completed. Yamazumi diagrams are used to understand if a production line is well balanced and indicate how much waste is found at the process step where the bottleneck is located. (Sabadka et al., 2017)

Example of the Yamazumi diagram is shown in Figure 25. In the example case, process step 4 is clearly the bottleneck – the step in which the cycle time is the longest. It has value-adding work of 20 units and required NVA work 40, which is much more than cycle of any other process step. Additionally, process steps 1, 3 and 5 have only NVA work, which raises a question if they are really needed.



	Process Step 1	Process Step 2	Process Step 3	Process Step 4	Process Step 5	Process Step 6
Value-Adding Work	0	10	0	20	0	10
Required NVA Work	0.5	5	1	40	0	15
Non-Value-Adding Work	5.5	16	10	80	30	18

Figure 27. Yamazumi Diagram (Spiridonova & Watson, 2018, p. 51)

4. PRELIMINARY VERSION OF AN EXPLORATORY DATA ANALYSIS MENTAL MODEL

Based on the literature research and experience of the author, the initial Mental Model for Exploratory Data Analysis was developed. The development of this model is discussed in two phases. In the first phase, the first version of the model is presented to selected content matter experts, consisting of five Lean Six Sigma Master Blacks. Their comments are then used to develop the second version of the model. This chapter describes the first version of the proposed model.

4.1. Mental Model for Exploratory Data Analysis

The initial Mental Model has four steps shown in Figure 28. Each of the EDA model steps is linked to the PDCA cycle and the different phases of DMAIC, which are embedded into an EDA journey. Each step will be described in its own sub-chapter where activities accomplished in the step are explained in detail. "Check with the Sponsor" is the point when the results of the analysis are shared with the person who is responsible for the business process performance under investigation. At that point, a decision is made to determine if the analysis yields to enough information to proceed to the Improve phase of DMAIC.

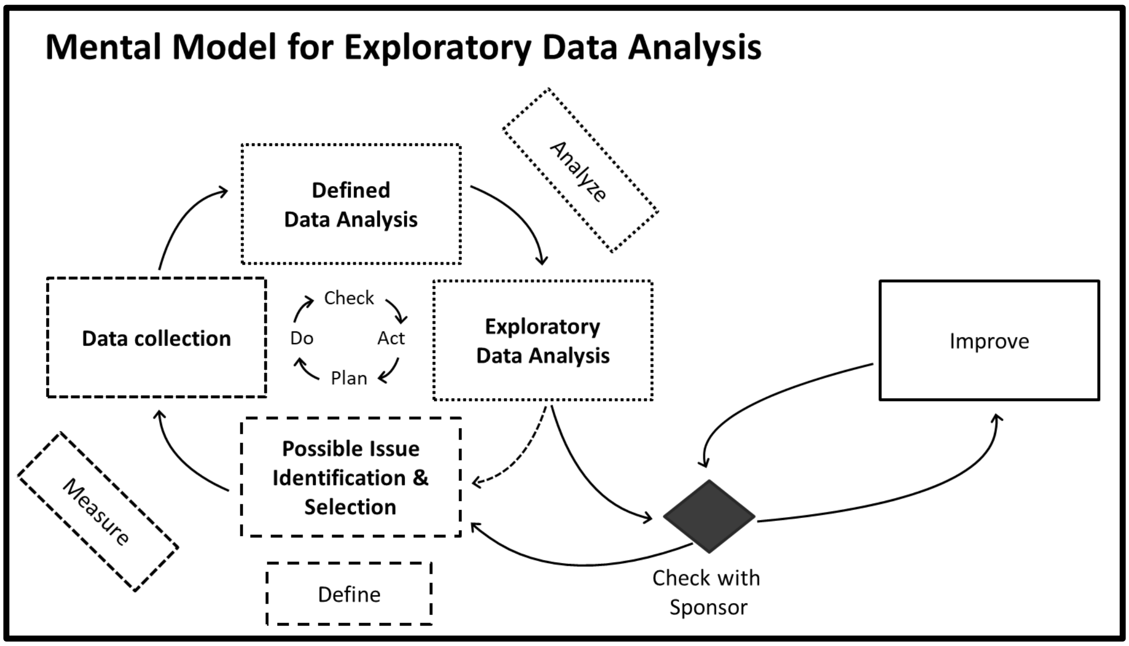
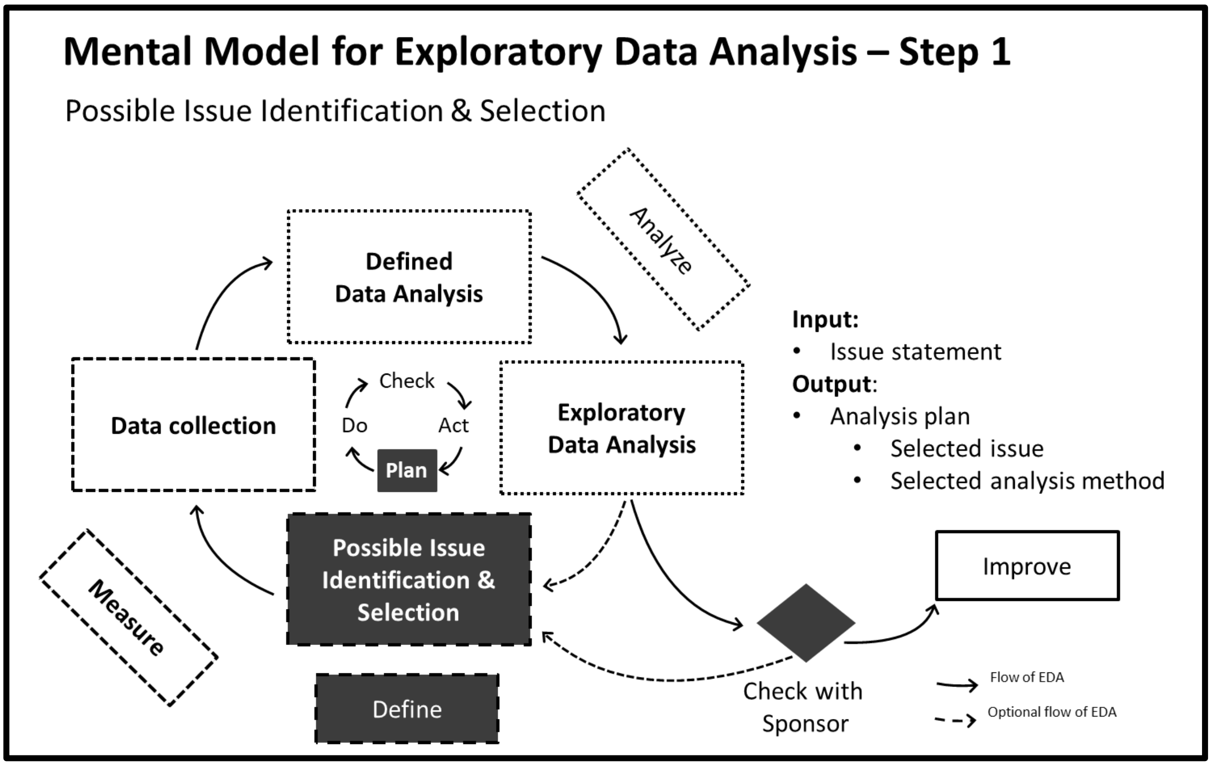

Mental Model for Exploratory Data Analysis
Analyze
Defined Data Analysis
Data collection
Check
Do
Act
Plan
Exploratory Data Analysis
Improve
Measure
Possible Issue Identification & Selection
Define
Check with Sponsor

Figure 28. Mental Model of Exploratory Data Analysis

The mental model combines ideas from many different authors. PDCA logic is generic, which fits well to the overall logic of any project. The principle of reverse loading by Harry guides to select the issue to be investigated and to collect data based on the question to be answered. Best of the Best and Worst of the Worst, as well as the Progressive Search from the Shainin System can act as triggers for Exploratory Data Analysis. The set of statistical tools are the ones that Watson presents for EDA process. Tukey's guidelines, such as looking at a level deeper, fits well to the loop idea of the mental model. Most of the ideas of the mental model have been stated by different authors, but now they are put together into a more holistic approach.

4.2. Step 1 – Possible Issue Identification and Selection

The first step of the Mental Model is "Possible Issue Identification and Selection." The "Plan" step in the PDCA cycle contains activities to decide what to do and what are the desired results, as well as methods. From a DMAIC perspective, the first step should be the "Define" phase activity – defining the issue to investigate and the approach for the subsequent study. Figure 29 illustrates graphically the position of the first step of the model.



Mental Model for Exploratory Data Analysis – Step 1

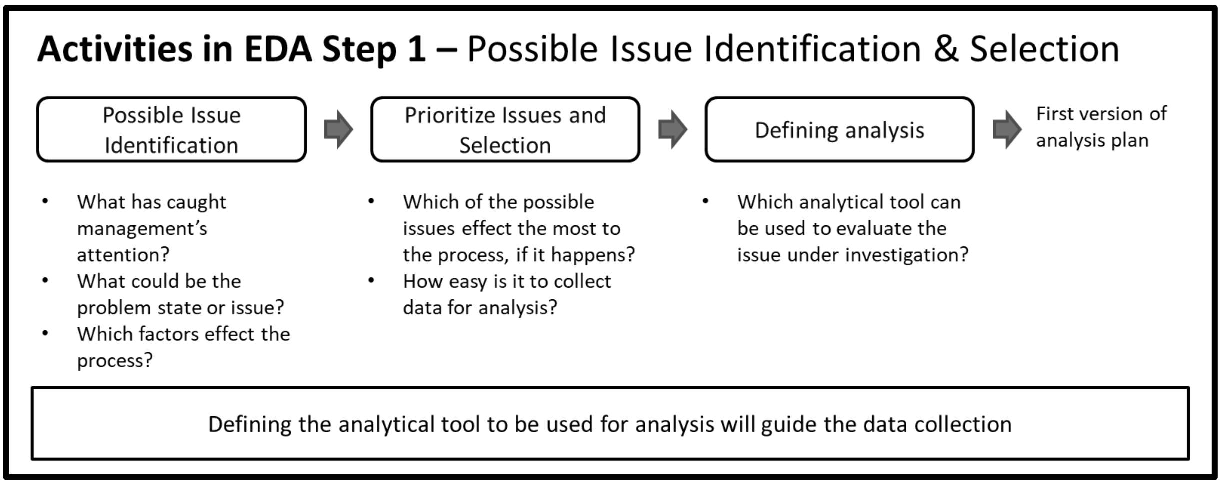
Possible Issue Identification & Selection

Input:
- Issue statement

Output:
- Analysis plan
  - Selected issue
  - Selected analysis method

Figure 29. Step 1 of Mental Model for Exploratory Data Analysis – Possible Issue Identification and Selection

The model uses an issue statement as its input. This is a narrative of the situation that occurs when the performance expectations of a process or product have not been met. Figure 30 presents the activities in the first step of the model. The reason for the issue identification is to generate ideas about what is or can be wrong in the process. An example of such an issue is management's concern that final assembly time is too long in a production operation. Customers could have complained that products were delivered late or that their order lead time was too long, but these presenting situations are generic and do not reveal an understanding of the causal system that created this dissatisfaction. The initial activity for an investigation of this case would be to identify what products have problems with long assembly times, which process steps are taking too long or if there is any specific event which occurs and is correlated with these longer assembly times. Frequently, different factors affect the process and it is important to breakdown the process into its component parts to determine which factors drive the unacceptable performance. Decomposing the data content into rational sub-groups helps to localize the issue into a specific logical category which makes it easier to investigate. Developing a graph which illustrates the rational subgroup factors can be accomplished using a Mind Map and this aids greatly in identifying where to focus the step-by-step EDA inquiry.



Activities in EDA Step 1 – Possible Issue Identification & Selection

Possible Issue Identification

- What has caught management's attention?
- What could be the problem state or issue?
- Which factors effect the process?

Prioritize Issues and Selection

- Which of the possible issues effect the most to the process, if it happens?
- How easy is it to collect data for analysis?

Defining analysis

- Which analytical tool can be used to evaluate the issue under investigation?

First version of analysis plan

Defining the analytical tool to be used for analysis will guide the data collection

Figure 30. Activities in Exploratory Data Analysis Step 1

The second activity is "Prioritize Issues and Selection." There are two principal factors to consider: The first is to evaluate the identified issues based on their severity of impact upon the process. In other words, if that issue occurs, how severe will be its effect on the process outcome? The second influencer of priority should be: How readily is data available? For some analyses, there is data exists; however, in other cases an extensive data collection program could be needed. This factor will delay the start of an investigation, but it should not determine whether or not an issue is selected.

Based on the issue selected for investigation, the specific methods and tools for the analysis are chosen. Figure 31 presents a template of the first version of an analysis plan. The initial activity is to develop an issue statement, which describes the situation that will be investigated. When considering a situation where the product assembly time is too long, the issue statement could be written as: “the assembly time of a product x is longer than that planned”. The next step is to state which factors could affect the issue and to determine the magnitude of the “miss” in expected performance. For the assembly time of a specific product x, a factor could be for example different accessories or options that are included in the design of customized versions of that product which were not considered in the original planning process.

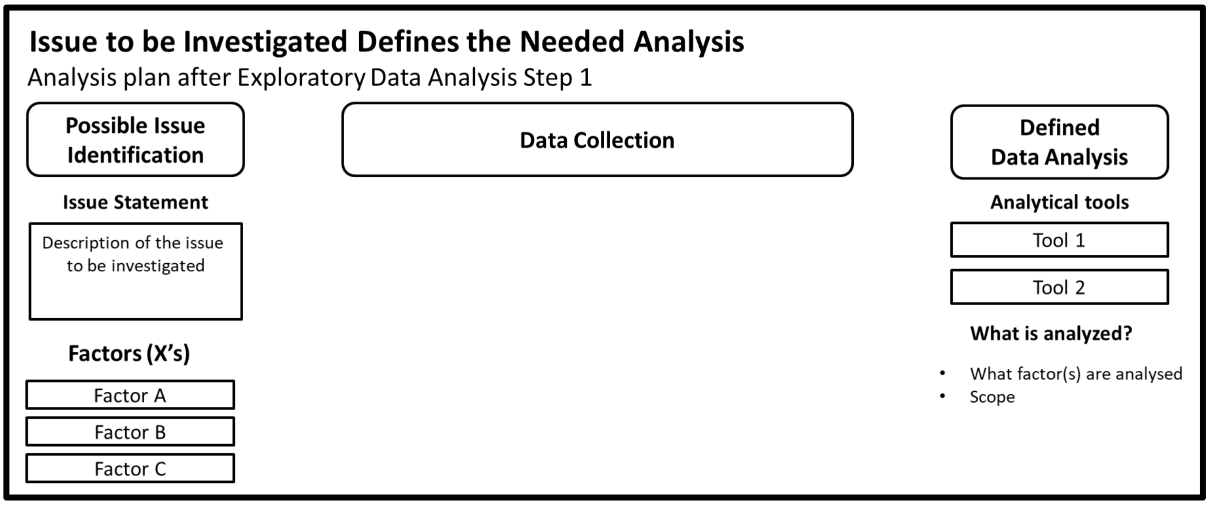
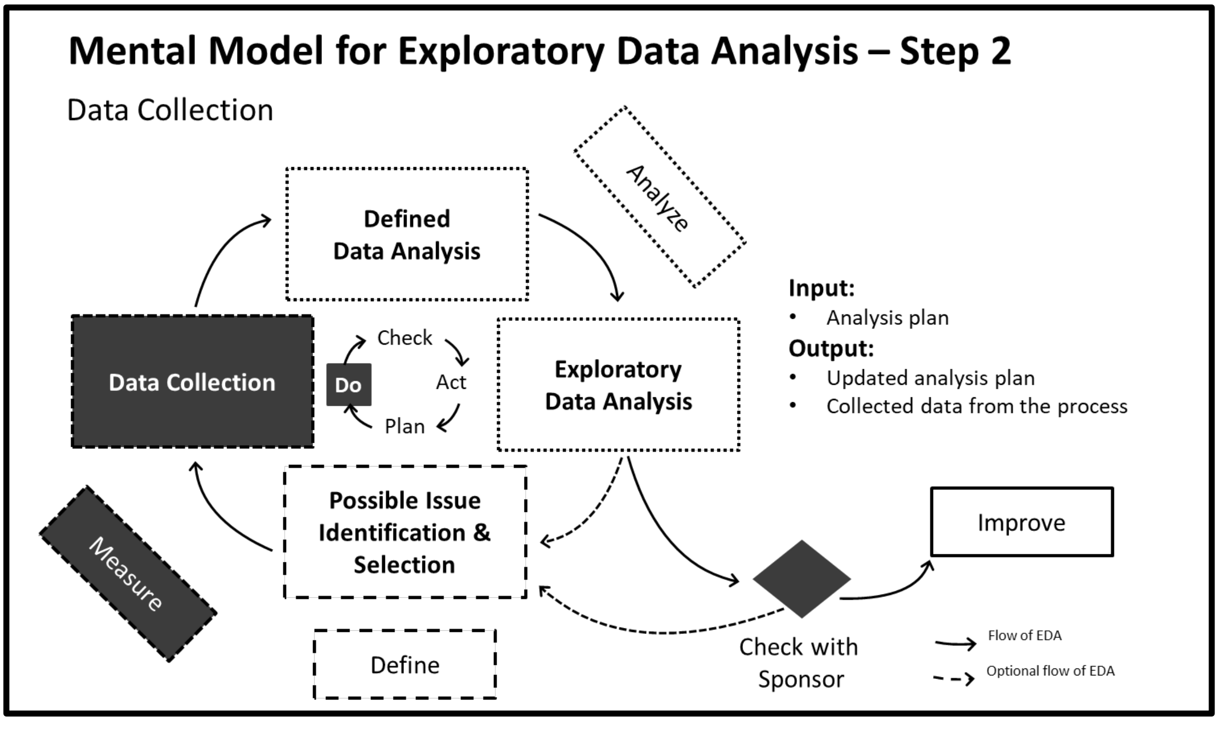

Issue to be Investigated Defines the Needed Analysis
Analysis plan after Exploratory Data Analysis Step 1
Possible Issue Identification
Data Collection
Defined Data Analysis
Issue Statement
Description of the issue to be investigated
Factors (X’s)
Factor A
Factor B
Factor C
Analytical tools
Tool 1
Tool 2
What is analyzed?
• What factor(s) are analysed
• Scope

Figure 31. First Version of Analysis plan

Before competing Data Collection, the analysis needs to be planned. In this activity, analytical methods and tools are selected, and the exact analysis to be performed is characterized and prototyped to assure that it will yield the desired information. This activity specifies the factor or factors to be analyzed and the scope and procedures for the analysis. Scope in this case refers to the specification of factors such as types of products, events or time frame to be included in the analysis. By specifying the analysis in this manner, clarification of the data collection requirements is achieved. In the assembly time issue, for example, the Defined Data Analysis part can be stated simply as: “use Boxplots to identify the ‘difference in assembly times between combinations of accessories and options chosen”. The scope of the analysis could be restricted to consider only those specific products that have identified assembly times longer than the upper limit of expected assembly time.

4.3. Step 2 – Data Collection

Data collection is the second step in this EDA model. It occurs in the Do step of PDCA in which the developed plan is executed. In this case, doing means collecting the data. In DMAIC, the Measure phase describes the activities that occur during the process of data collection. Figure 32 identifies the positioning of data collection in this EDA mental model. The preliminary analysis plan is the input and the outputs of this activity are an updated analysis plan and the data that has been collected.



Mental Model for Exploratory Data Analysis – Step 2

Data Collection

Input:
- Analysis plan

Output:
- Updated analysis plan
- Collected data from the process

Figure 32. Step 2 of Mental Model for Exploratory Data Analysis – Data Collection

The first activity in the data collection step is to plan the data collection process by addressing the questions formulated in the analysis plan. The questions and template for the updated analysis plan are illustrated in Figure 33. First the data to be collected needs to be identified and specified based upon the preliminary issue statement and the initial analysis plan. The question to be addressed in this inquiry is: what kind of data needs to be collected for which sub-groups of the possible population of data observations? For the assembly time example, the following statement could apply to this specification: "Assembly cycle time of products x, y and x from commencement of the production to the acceptance of the product at final testing."

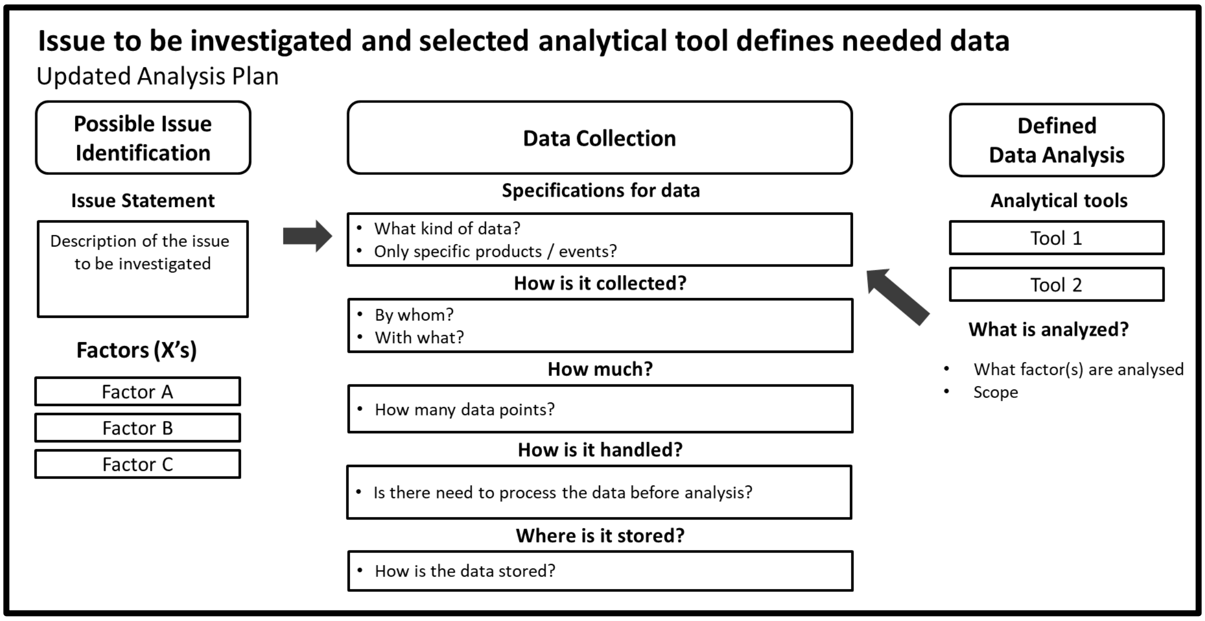


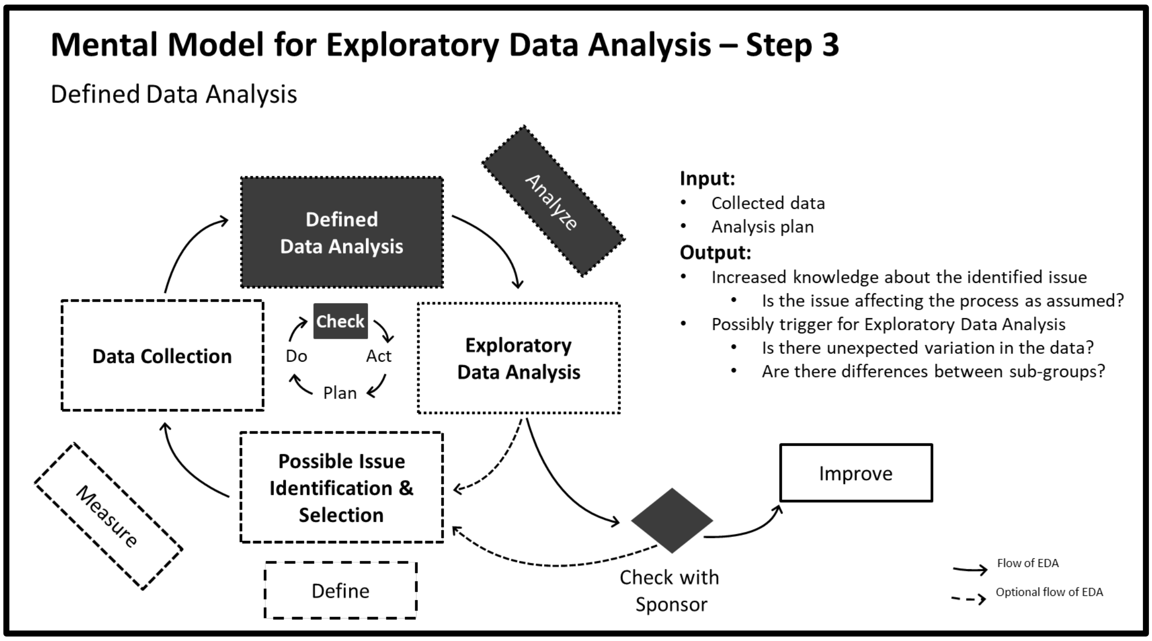
Figure 33. Updated Analysis Plan

There are four aspects in this Data Collection Plan. The first is how the data is to be collected. The person responsible is identified and the means by which the data is to be collected is described. The collection method can vary from an automatic device to a manual data collection process. Secondly, how much data is required needs to be specified to assure sufficient data for sound analysis. This includes both the amount of data observations and the frequency with which the data collection occurs. These decisions need to be made to assure that data will be collected in the right amount and often enough to capture expected cycles in the process throughput, so the data reflects the expected operational process. Thirdly, data may need to be processed for subsequent analysis so it is in an acceptable format. An example of this situation occurs when timestamps need to be changed from time observations to indicate the elapsed cycle time or lead time. Finally, how the data is stored is another item to consider, because access to the data for all interested parties, not just the analyst, needs to be assured. For instance, there might be a need in the future to have access to this data by others for subsequent analysis or for use as a base case comparison to future states of process performance. The main activity in the Data Collection step is the actual collection of the data, which follows the analysis plan.

4.4. Step 3 – Defined Data Analysis

The third step in the Mental Model for Exploratory Data Analysis is "Defined Data Analysis." This step executes the analysis plan previously developed. In the analysis plan, an issue has been identified to be investigated in detail. The characterization of the issue is accomplished by the analysis in this step. From a PDCA point of view, the goal of the "Check" step is to ensure that the desired results have been achieved and to make observations about what actions should be taken in the "Act" step.

Step 3 is identified in this EDA model in Figure 34. Step 3 uses the collected data and analysis plan as inputs and formulates an output of increased knowledge about the issue under investigation. If the issue affects the process to some extent, it may trigger additional analyses, which is the in-depth process of Exploratory Data Analysis. A trigger for executing this in depth exploring can, for example, be observed performance differences between the sub-groups analyzed or unexpected variation patterns within the data. DMAIC applies many statistical tools identified in Figure 35.



Mental Model for Exploratory Data Analysis – Step 3

Defined Data Analysis

Input:

- Collected data
- Analysis plan

Output:

- Increased knowledge about the identified issue
  - Is the issue affecting the process as assumed?
- Possibly trigger for Exploratory Data Analysis
  - Is there unexpected variation in the data?
  - Are there differences between sub-groups?

Figure 34. Step 3 of Mental Model for Exploratory Data Analysis – Defined Data Analysis

A variety of analytical tools can be used in the Defined Data Analysis. The set of tools will be identical to the Exploratory Data Analysis methods proposed as a sequence for analysis in the process defined by Watson. Depending on the issue selected for investigation, all tools presented in Figure 35 can be used for analysis or a selected subset of the tools may be applied, depending on the requirements for investigation of special causes of variation or identification of waste and losses that are observed. However, when a process is analyzed using all the suggested analytical tools, a "deep dive" into the performance characteristics of the issue can be obtained. These tools create an "information intensity" that aids in characterization of the problem by linking it to the process steps that are causal. The most value from applying this Mental Model for Exploratory Data Analysis is gained in the Defined Analysis, where observations uncover a need for deeper analysis as triggered by observed patterns in variation that requires investigation of the issue at a deeper level by repeating the analysis within a problematic subgroup that has been identified. This is a process of sequential analysis through "drill down" by seeking the level of the process where the true cause of the problem can be observed. This process has historically been called "Root Cause Analysis (RCA)".

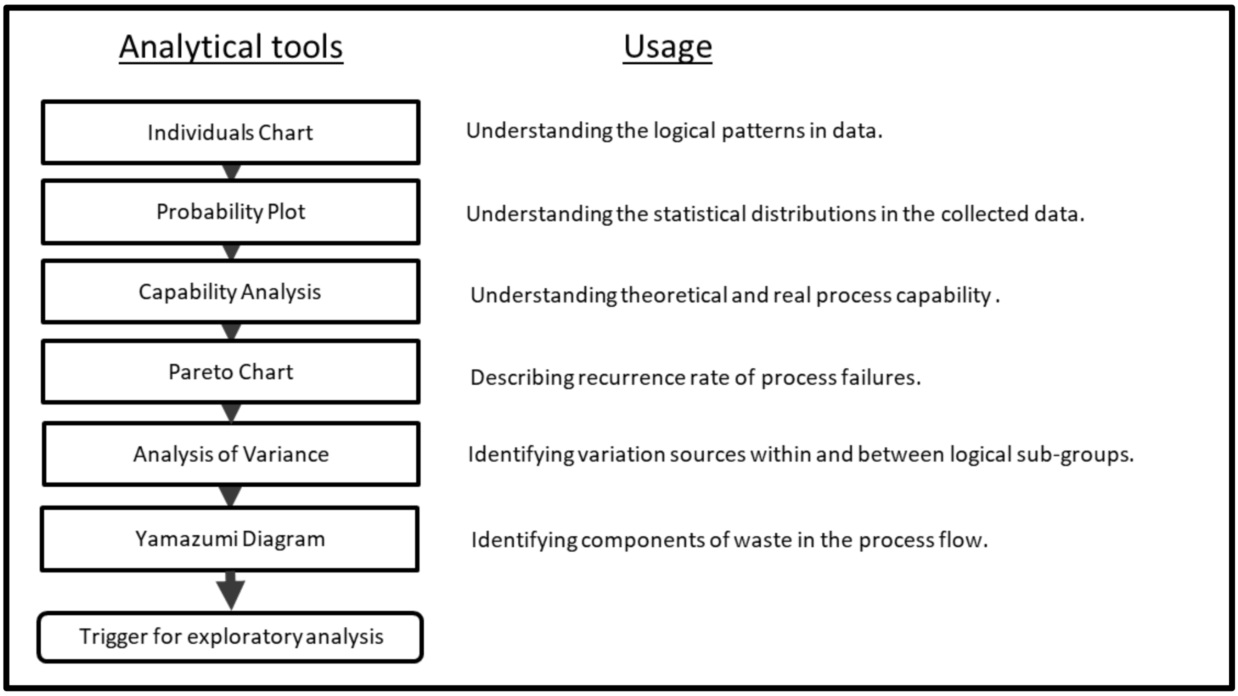

Analytical tools
Usage
Individuals Chart
Understanding the logical patterns in data.
Probability Plot
Understanding the statistical distributions in the collected data.
Capability Analysis
Understanding theoretical and real process capability .
Pareto Chart
Describing recurrence rate of process failures.
Analysis of Variance
Identifying variation sources within and between logical sub-groups.
Yamazumi Diagram
Identifying components of waste in the process flow.
Trigger for exploratory analysis

Figure 35. Set of Analytical Tools Used in Defined Data Analysis

4.5. Step 4 – Exploratory Data Analysis

The fourth step of the Mental Model for Exploratory Data Analysis is conducting the Exploratory Data Analysis. Figure 36 presents the logic for step 4. From a PDCA point of view, Exploratory Data Analysis in this mental model is related to the Act step, where actions are taken or adjustments to the process based on the observations from the Check step. From a DMAIC point of view this step relates to the Analyze phase activity, where the focus is placed on identifying the sources of variation in the data by applying a variety of analytical tools.

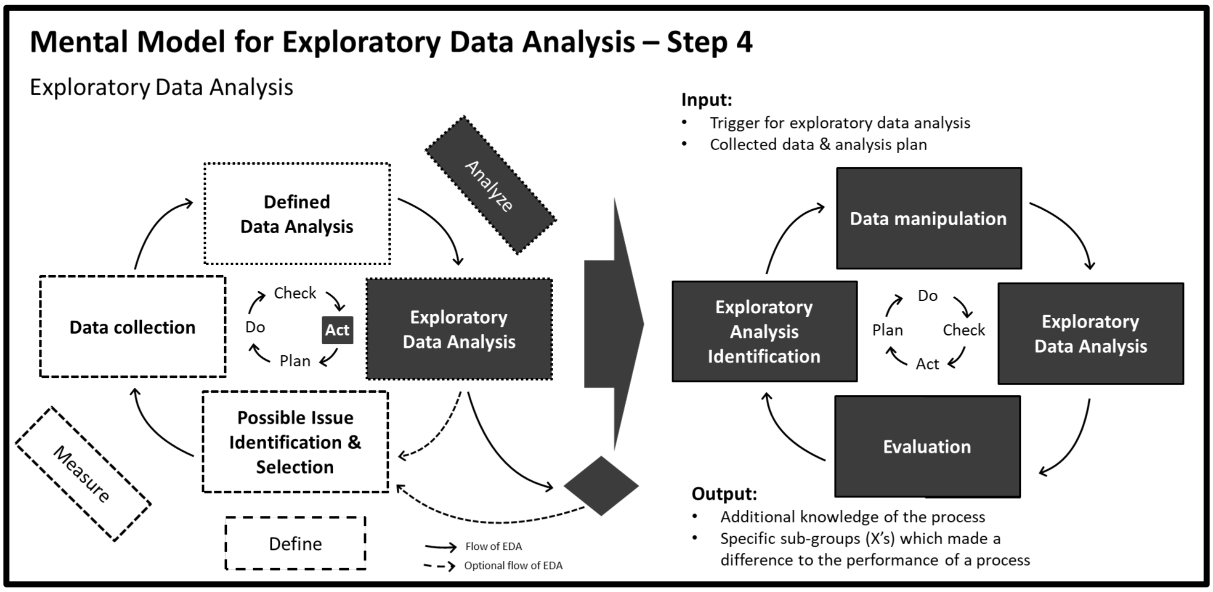

Mental Model for Exploratory Data Analysis – Step 4
Exploratory Data Analysis
Input:
• Trigger for exploratory data analysis
• Collected data & analysis plan
Analyze
Defined Data Analysis
Data collection
Check
Do
Act
Plan
Exploratory Data Analysis
Possible Issue Identification & Selection
Measure
Define
Flow of EDA
Optional flow of EDA
Data manipulation
Exploratory Analysis Identification
Do
Plan
Check
Act
Exploratory Data Analysis
Evaluation
Output:
• Additional knowledge of the process
• Specific sub-groups (X's) which made a difference to the performance of a process

Figure 36. Step 4 of Exploratory Data Analysis Mental Model – Exploratory Data Analysis

The most critical input to Step 4 is the data observations that trigger the Exploratory Data Analysis. This means that unexpected variation or differences between the performance of sub-groups has been discovered in the Defined Data Analysis step. Other inputs for this step are the data collected and the analysis plan. Figure 37 represents the investigative journey in this detailed Exploratory Data Analysis “drill down” of analysis factors. Factor X1 has been defined as a factor which might be causing issues to the performance of the process. In the defined data analysis step, the difference between X1.1 and X1.2 has been analyzed. The variation in performance of the process with X1.1. has been more than X1.2, which has triggered the need for Exploratory Data Analysis.

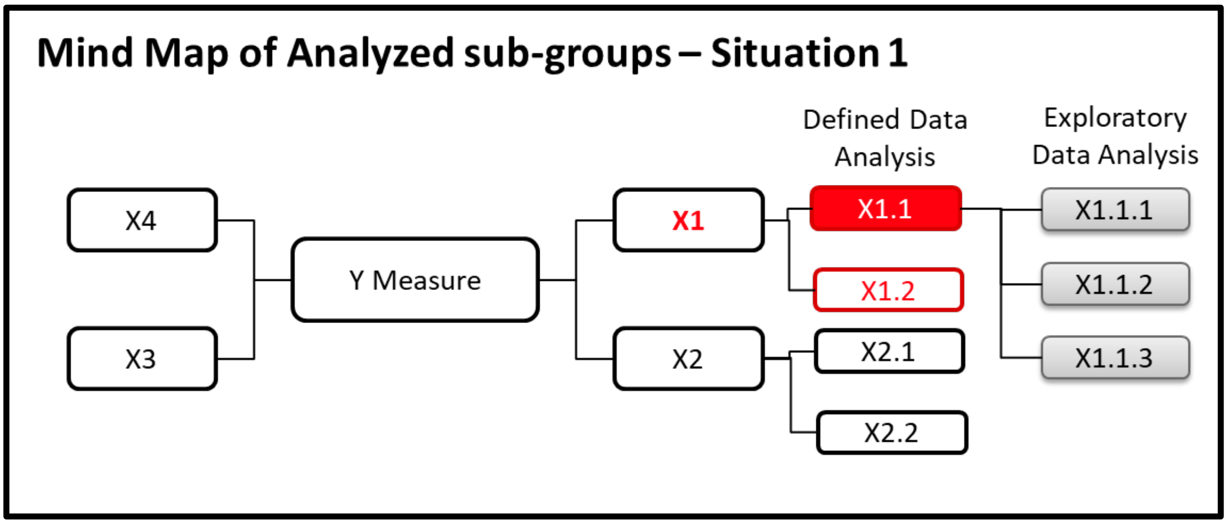


Figure 37. Mind Map of Analyzed sub-groups during Exploratory Data Analysis – Situation 1

The logic of the inside EDA loop for Exploratory Data Analysis is the same as applied for the external loop in the comprehensive Mental Model for Exploratory Data Analysis. This more detailed "drill down" using Exploratory Data Analysis logic is based on the identification of "interesting variation" that is initiated based as a trigger for this sub-step. In this sub-step suitable analytical methods and tools are selected and the need for data manipulation so the data conforms to the analysis methodology is determined. From a PDCA point of view, the first step is Plan or Planning the Exploratory Analysis. The "Do" step first addresses the need for data manipulations, which, in this context, means filtering or sorting the collected data so it is in the format and structure that is required for the analysis. Some calculations of the collected data can be done, for example to change the form of the data from observed levels of performance to reflect the rate of change between observations. The third step is to conduct the Exploratory Data Analysis using the analytical tools.

When viewed from the perspective of a Y=f (x) function, Exploratory Data Analysis dives deeper into more detailed level of x's than the top tier in a measurement hierarchy. An illustration of this situation is represented in Figure 37. As X1.1 has been a trigger for a deeper Exploratory Data Analysis, differences between X1.1.1, X1.1.2. and X1.1.3. are evaluated at a deeper level of the measurement hierarchy to determine where any detected variation in X1.1 has originated. That deep dive can be accomplished by combining ANOVA to identify the sub-set with the most variation and then drilling down into that sub-set by applying the sequence of

EDA methods starting with the Individuals control chart as illustrated previously in Figure 35. In the Evaluation step a decision is made to continue with this additional analysis. The results of the analysis can be presented to the improvement project sponsor or a different issue may be selected for analysis using the entire Mental Model.

This inner loop of Exploratory Data Analysis can rotate as many times as required to obtain an understanding of the causal situation. Figure 38 represents the circumstance where the Evaluation step continues into a further cycle of Exploratory Data Analysis. The reason for this continuation of the analysis could have been that the performance of the process was different for factor X1.1.2 from X1.1.1 and that the variation explained by X11.1.1 was not sufficient to gain confidence that it acted alone to drive the performance shortfall, and thus, additional factors affecting the performance. The next Exploratory Data Analysis would be performed for data representing only X1.1.2. This means in the Data Manipulation step; the observed data set has been filtered so that only observations of X1.1.2 are used in the subsequent analysis. In this case, the Exploratory Data Analysis compares differences between X2.1 and X2.2. A similar kind of analysis can be done to investigate other relationships.

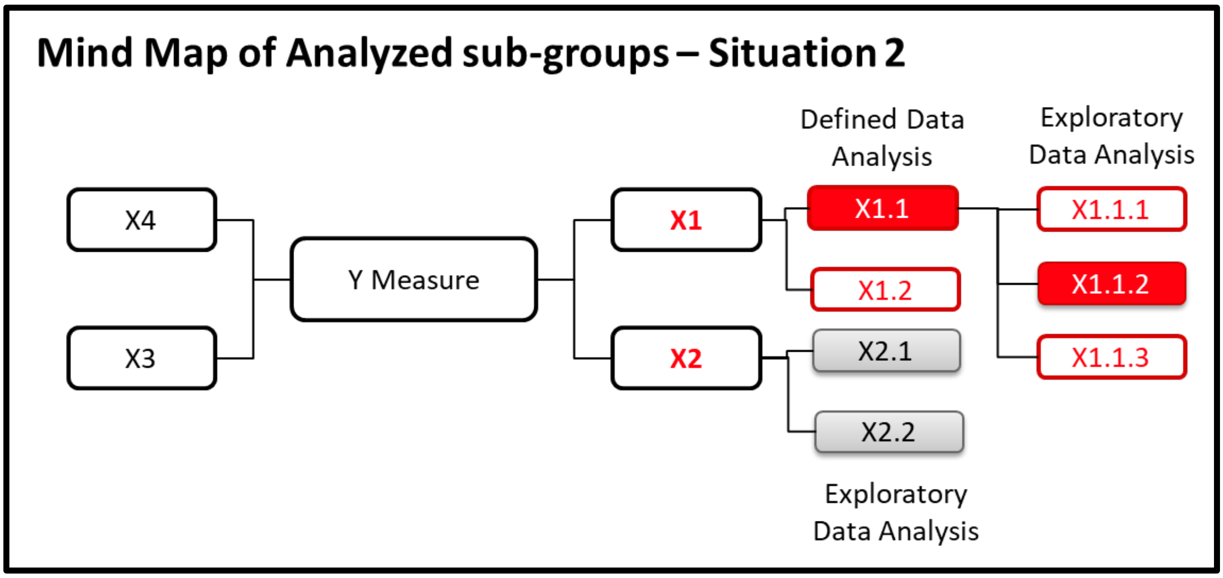

Mind Map of Analyzed sub-groups – Situation 2
Defined Data Analysis
Exploratory Data Analysis
X4
X3
Y Measure
X1
X2
X1.1
X1.2
X2.1
X2.2
X1.1.1
X1.1.2
X1.1.3
Exploratory Data Analysis

Figure 38. Mind Map of Analyzed sub-groups during Exploratory Data Analysis – Situation 2

5. ANALYSIS OF THE PREMILINARY MENTAL MODEL FOR EXPLORATORY DATA ANALYSIS

The first version of the developed Thinking Model for Exploratory Data Analysis was presented to five experienced Lean Six Sigma Master Black Belts. Four of them worked for the same company. After the presentation, an interview was conducted by seven questions about the model. The presentation and questions asked are included in Appendices A.

5.1. Relevance of the Mental Model to the Real-World Projects

5.1.1. Progress of Analysis

Sometimes, the first step of a problem-solving process can be skipped, which is illustrated by the comment from Master Black Belt interviewee number two (MBB 2): *"Problem should be defined based on some measure and analysis, but sometimes problems and projects are defined based on opinions."* The model can be seen helpful at the beginning of the project when defining the problem statement. MBB 4 comments the usefulness of the model in the following way: "*The model guides to select the right direction for a project starting at a high-level. Among the models I know, this one presents best the analytical journey of DMAIC projects. It can also lower the threshold to use structured problem-solving approach, because some people can see DMAIC projects as resource intensive*." The later comment highlights the possibility of the model to be used for structured problem solving just by itself.

According to Tukey (1977) a desired approach to data analysis is to conduct exploratory and confirmatory data analysis at the same time. The model integrates to some extent both approaches. MBB 1 comments the following on the need for both approaches:" The *model describes well how data analysis progress in the real world. We need to keep the following facts in mind when analyzing problems: We are conducting both types of analysis: Investigating hypothesis as well as exploring what the data has to say and we need to be ready to adjust our views if needed."*

The loop type of approach of the mental model was seen as an advantage of the model. According to MBB 2:*"In real life the analysis process is a kind of loop. At first, we have some* *kind of data, which develops our understanding of the situation and gives room for new speculations. Based on the speculations, additional data is collected, which guides us deeper into details and closer to solution.*" This comment illustrates that in reality, data needs to be collected many times and the answers given by the analysis of the current data set guides the project forwards.

5.1.2. Right Type of Data to the Question in Hand

The comment by MBB 3 illustrates well the situation engineers encounter many times: *"For myself, the selection of the tool takes place many times after data has been collected."* This comment is similar to Michael Harry's (1997) principle of reverse loading to avoid situations where the result of the analysis is limited by pre-defined measurement system and type of data available. The developed mental model can do the same based on the comment of MBB 1: *"The model forces to discuss collecting the right data for the projects, which is ideal because in some cases project teams are satisfied with insufficient data. It is impossible to do a sufficient analysis based on insufficient data."*

The problems caused by the fixed measurement systems to problems solving is well illustrated by the comment of MBB 2:" *Collecting data is guided by the model from the real process instead of just analyzing the data already collected by SAP. Many times, people have preconceptions that if there is no available data already, it cannot be used as a part of development project. The model discloses that collecting and analyzing data is a loop, with every rotation bringing additional knowledge on the problem."* In other words, the model forces to look for and possibly collect the right type of data as part of the problem-solving project.

5.1.3. Systematic Approach to Analysis

General problems in decision making are described by this comment from MBB 1: *"Many people develop hypotheses without data or just based on historical data. Many people also draw conclusions without any data analysis."* Using the model can help to overcome the pitfalls. That view is supported by the following comments. First by MBB 1: "*The model offers a systematic approach to data-analysis. It guides to conduct the analysis in a repeatable and predictable way. That decreases randomness and increases the confidence in the final analysis.*

MBB 4 had a similar type of view on the benefits of the model: "*It gives structure to data analysis, which follows DMAIC logic, and it can decrease error estimates and how much opinions effect on the decision making.*"

As EDA has been described more as an art than a science, learning it can take a long time. Providing some structure to the analysis can help learning it, which was noted by MBB 3: "*The Thinking Model can give more structure to make an analysis, especially when the analyst is inexperienced. If I did not have experience on the exploratory data analysis, the model would bring more structure and systematics to it. A great model in its own context!*"

Integrating a qualitative tool to represent sub-groups is supported by the comment from MBB 3: "*The usage of fishbone diagram and integrating it into process of doing an analysis has helped myself in my projects*"

The mental model is flexible because it allows the analyst to select the appropriate tools to be used for the analysis. The view of Tukey that there are many ways to analyze the same data set and come to a good conclusion is supported by the comment from MBB 3: " *In the end, the tool used is not the most important point of the analysis as long as the end result is good.*"

5.2. Development Points for the Model

5.2.1. Measurement System Analysis and Content Knowledge

An important point for the development points for the model is how the model is used for data analysis. If it is purely for exploring and generating new questions, data quality is not a big issue as comment from MBB 1 illustrates: "*Data contains a lot of variation and faulty points, which may lead to a situation where the quality of data is questioned. Exploratory Data Analysis should not concentrate very much on the quality of data. My personal view on exploratory data analysis is that new hypotheses are formed based on the graphical outputs, without taking much stance on the format of the data.*"

In case the model is applied for hypothesis testing type of analysis, the importance of Measurement System Analysis increases. MBB 1 commented the following on the importance of MSA: *"The model does not take a stance on the usability or the possible problems of the data. Measurement system analysis of some kind should be done in order to predict errors in hypothesis testing."* As a summary, if the model would be applied also to hypothesis testing type of analysis, a measurement system analysis should be added to it as an extra step. The most natural position for measurement systems analysis would be within the data collection step.

5.2.2. Data Manipulation

Data manipulation, which is in the fourth step, Exploratory Data Analysis, received comments from two MBBs. The first comment was that the term data manipulation can have many meanings, which may lead to different actions than intended. MBB 1 commented about it this way: *"The word data manipulation has a bad connotation, because some people might delete datapoints in order to be able to analyze the data better. Data editing could be a better word."* What the author intends that data manipulation would be, is organizing and possibly connecting some data set together so that they can be analyzed with statistical tools rather than sorting out data for exclusion from the analysis.

The other MBB pointed out, that data manipulation is not discussed in many cases when data-analysis is taught. Comment by MBB 3: *"Data manipulation is often somewhat neglected, which is, however an important part of data analysis."* On the other hand, the model does not describe in detail, what kind of data manipulation is needed. That can become an extension of the model for a future consideration, because different analytical software used might require different kinds of data manipulation.

5.2.3. Other Possible Additions

Interviewed MBBs pointed out a few additions to the model. The easiest to add is suggestion from MBB 5 to add 5 Why's and 1 H method to support the issue identification. De Mast also mentions that in his writing.

MBB 1 suggested the following about the tools: *"Grouping the tools on the basis which tools are used and in which part of the model. Description of what kind of data can be analyzed with different tools."* The comment's both points are very valid for the usage of the model. The grouping of tools based on the part of the model could make sense, but basically the statistical tools are used on analysis steps and the qualitative tools can be used throughout the model. Potentially Yamazumi diagram can raise some questions, as it can also be used in the fourth EDA step, but the logic remains the same – to look one layer deeper, which means to inquire about activities inside the process step. Commenting what type of data can be analyzed with which tool is very relevant when people are learning about the statistical tools. This approach could also be added later on, as it is out of the research scope for this thesis.

A third development point was raised regarding adding the point of view of management and coaching of the analyst within the EDA process. MBB 3 commented: *"I would emphasize how coaching is part of doing an analysis, Management reviews can be added to different parts of the model. Another point to consider is how results are communicated inside the company."* The check point after the loop has been intended to be the management review, but the way the model is currently presented in Figure 28 (overview of the model) the review would be only before Improve phase of DMAIC project. This is also something to consider for future developments of EDA.

5.3. Possible Limiting Factors of the Model

The model can be used in various ways, which creates a possibility to consider how to limit factors based on a variety of points of view. This was emphasized in the comment of MBB 2: *"Weaknesses of the model are hard to point out, because it depends on how the model is used. It is hard to conduct a good analysis if the analyst does not understand the quality of data, what is being analyzed or how the things analyzed are linked together."* MBB 4 also emphasized the need for content knowledge of the analyst: *"To produce meaningful analyses and successful projects, the analyst needs to have a certain amount of content knowledge to understand the meaningfulness of factors in real life."* On the other hand, different authors have pointed out, that included within any analysis process there should be content matter experts to supplement the analyst. That is a consideration which the current model does not explicitly require. Content matter knowledge is needed in every step of the model from its initiation in defining possible issues, to planning the data collection, and through the interpretation of results.

Two MBBs pointed out that the mental model might limit the thinking of the analyst in certain situations. MBB 4 commented about focusing too much on a specific process step: *"The model might guide to focus on a specific process step, where the problems seen might be caused by a previous process step. If the previous process steps are not considered, there is a possibility that the conclusion and improvement actions based on them can be completely incorrect. In order to proceed to improve phase, we should understand the biggest causes of the problem:"* MBB 1 concentrated more on the way the model guides to choose a specific tool or tools: *"One possible pitfall is that the Thinking Model might possibly limit the dynamic of the analysis, if the end result of the analysis is not as expected. It might lock thinking just to specific tools, whose use is developed by usage and coaching".* How the model is coached is an important aspect in developing familiarity as analysts first learn about the model, but subject is also outside the scope of this research.

The selected tools for the model are able to show only one-way interaction effects, which was pointed out in the comment of MBB 3: *"The presented analytical tools for the model do not analyze if there are interaction effects between multiple factors, but the presented way to analyze might give hints about the interaction effects."* Possibly additional tools could be added to the model analyze interaction effects, but this is also outside the scope for this research as these methods would typically be engaged in the "deep dive" following EDA formulation of the problem for further investigation of causality.

5.4. Views on Different types of Exploratory Data Analysis

The interviews also questioned how the nature of EDA changes in different phases of DMAIC projects. The trigger for this question was the following statement by Dr. Watson (2018a, p. 168): "EDA is done in Define phase, but the tools and logic can be used in the later phases of the project." The biggest benefit of EDA comes from its initial application in the Define phase for shaping the problem to be attached in more detail. This was pointed out by MBB 3*: "First we recognize the problem and don't start to develop anything that is not a big problem. If the problem is defined on the basis of opinions, data may not support the description of the problem,* *which at worst can lead to excessive use of resources. We can prioritize the development resources to use them for more difficult problems. A well conducted EDA at the define stage helps to confine the problem, speeding the project, because it makes it possible to focus on the essential process step."*

Depending on the organization, the Define phase of a project can be done by different people than the Measure and Analyze phases. This was pointed out in a comment from MBB 3: *"In the context of Green Belt projects, the Define phase can be done by MBB or BB. The role of the Green Belt would be to determine the cause or causes the problem."* One reason for this is that a goal of define phase is to create a problem statement. MBB 4 commented about it in the following way: *"Define phase should define the problem statement, which is unambiguous, and by means of* [performing an] *EDA you get a clearer problem statement. It defines the target from big Y to big X's."*

Difference between Y and X's were pointed out by many MBBs, which can be a good way to explain how the EDA differs in different phases of DMAIC projects. MBB 2 commented the following about the journey from Y to x's: *"The high-level analyses of the Define phase creates a rough understanding of the process output and of the problem, which later iterates as a more detailed resolution as for x's at the Analysis phase. In structured problem-solving, the variation of Y must be understood. To start focusing on possible effects of x's."* MBB 1 had a similar view, but instead of x's a term "deeper" was used in this context: *"In the beginning, the problem is defined, and in the Measure and the Analyze phases the problem is analyzed deeper, and the focus of improvements is defined based on the analysis."*

MBB 4 pointed out three different levels of EDA: *"To define when analysis is detailed enough in order to improve the detected problem. The example in the presentation identifies the factors, but not yet what causes the problem. Linkage to another type of analytical tool could be created, for example, 5 why's. Usage of the model could be described with a function of Y -> X -> x, where analysis drills down to deeper levels."* The logic presented is similar to the Y = f(x) in the beginning of this thesis. This link would be a good addition to the model.

MBB 5 described the journey of EDA for Y to small x's in detail*:"*

1. *The first iteration of EDA on the Y identifies the process step for further investigation, this is done only on output data, which normally exists for every process. Output of it is an issue statement.*
2. *The second iteration of EDA drills into the process steps, that the ANOVA shows that creates the most variation to achieve this we must do a Yamazumi diagram at this lower level and then the full EDA. In this step the data is typically automatically captured by SAP but it may also require manual data capture by process workers*
3. *The third iteration of EDA concentrates on process detail data that maybe observed by automated data collection systems or testing data (voltages, etc.)*

As a summary the flow of EDA goes from Y to big X's and then to small x's. They can be called as high-level sub-groups for Y measure (performance results), middle-level detail which is high-level X (process flow details) and operational level analysis using process and product x's (process operations details)."

All these comments from these MBB interviews regarding the journey of EDA are considered in the next stage development of the EDA mental model.

5.5. Summary of Comments and Development Points

In general, the feedback on the model was very positive. Many interviewees commented that the proposed model brings structure to data analysis as well as describes quite well how it progresses in the real world. The model also forces thinking about data collection from the data point of view: does the project have proper data to answer its most important questions; the most critical questions that reveal the best insights into the problem situation.

There were several improvement suggestions. Two comments were regarding data quality and measurement system analysis. The model itself does not take a stand with respect to how good is the collected data. Some kind of measurement system analysis could be done, to increase the confidence in the analysis. Natural point for including this capability would be to include it as part of data collection phase.

It was suggested to rename the second EDA step due to negative connotations of the term manipulation. Additionally, the name of the overall mental model was suggested to be changed from exploratory data analysis, because EDA is positioned as a part of the model. It was also suggested that tools to be grouped according to the data which they can be used to analyze.

The deepest discussions during the interviews were regarding positioning the model with respect to DMAIC phases. The biggest discovery was made regarding the fact that EDA can operate at many different levels of detail. It starts with Y measure and progress to detailed x's. This helps to shape the issue statement into a focused problem statement, which can be then made even more precise by adjusting scope of the project.

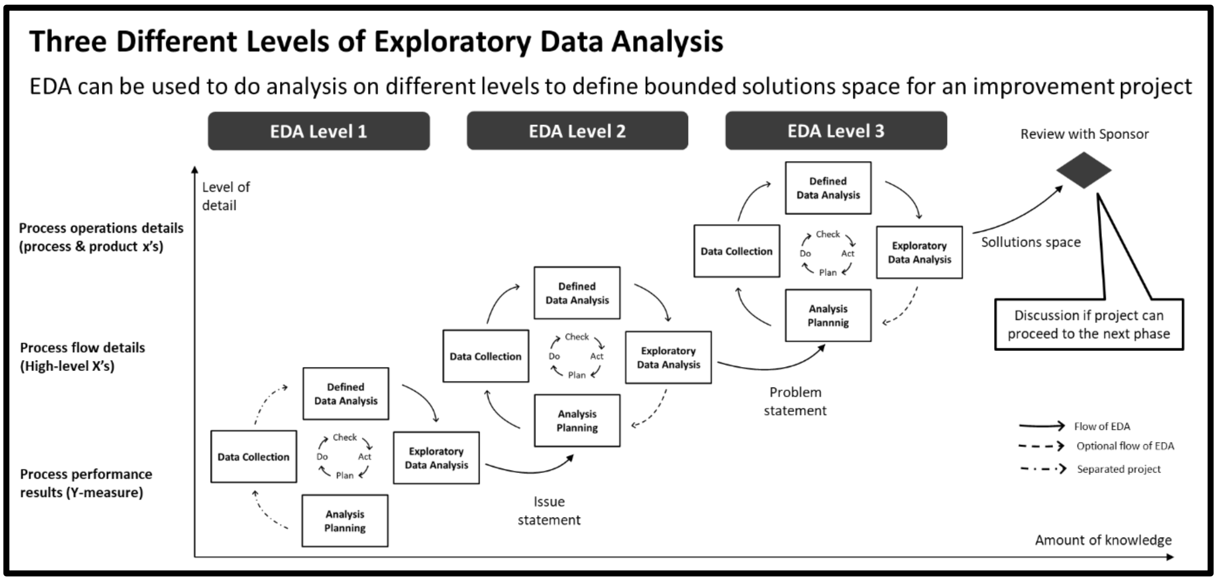
6. UPDATED MENTAL MODEL OF EXPLORATORY DATA ANALYSIS

This chapter presents the revised version of the Thinking Model for Exploratory Data Analysis based on the comments of the interviews described in the previous chapter. This developed model was presented to four Lean Six Sigma Master Black Belts for critique. After the presentation, an interview was conducted using five questions. The presentation and questions asked are included in Appendices A and B.

6.1. Updated Mental Model for Exploratory Data Analysis

The biggest update to the model has been to add different levels of the analysis. Figure 39 presents the updated model. The model operates at three levels starting with the Y-measure and ending with the process operations details. The following sub-chapters present each of these three analysis levels in more detail. As a guideline, the deeper the analysis dives, the more content knowledge the analyst has about the problem and its related processes.

Idea of the mental model for exploratory analysis is that at first a performance issue is noticed, which is analyzed deeper and based on this analysis an issue statement is formed. Based on the issue statement process flows are analyzed, which may engage output measures from several different process steps. This EDA level 2 analysis should yield a problem statement. The problem statement should identify the location, measure and desired direction of change regarding the problem. Based on the defined problem statement an operational analysis is conducted to define boundary of the potential solutions space.



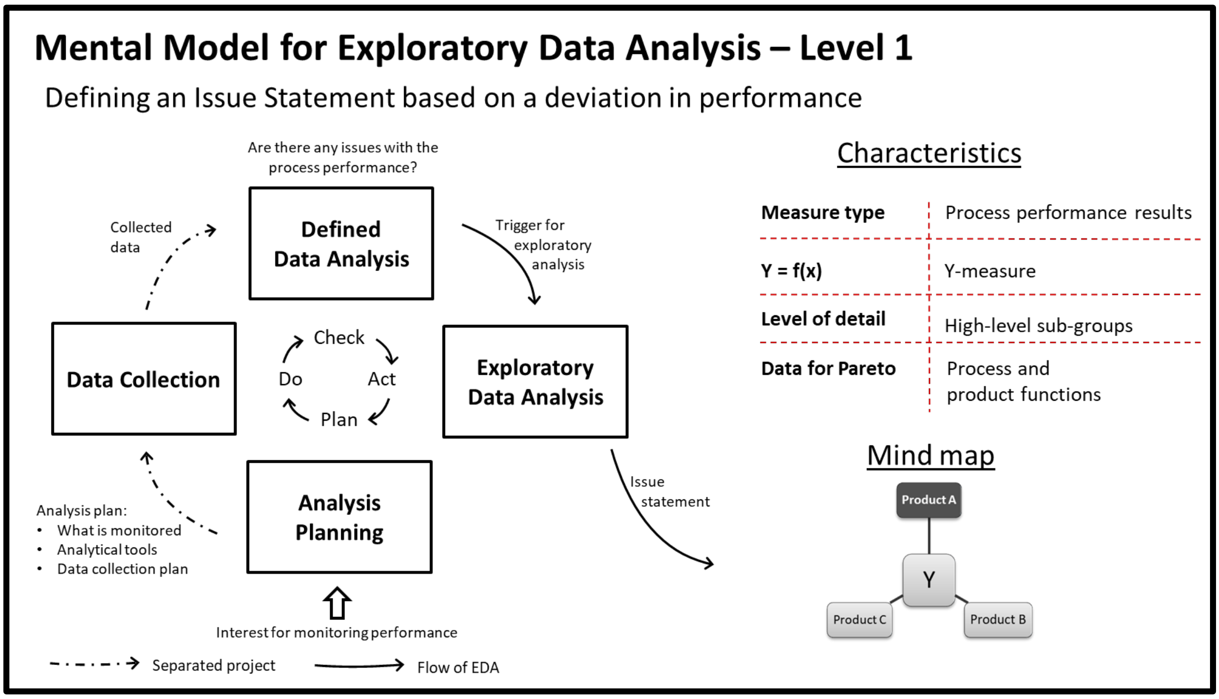
Three Different Levels of Exploratory Data Analysis

EDA can be used to do analysis on different levels to define bounded solutions space for an improvement project

Figure 39. Three Different Levels of the Mental Model for Exploratory Data Analysis

6.1.1. Exploratory Data Analysis – Level 1

The first level of Exploratory Data Analysis is different from the others, because it does not start with analysis planning but instead, it starts when Defined Analysis triggers a need for Exploratory Data Analysis. Figure 40 shows the details of EDA level 1. The first EDA is done for the Y-measure which indicates the productive output of the system under investigation. This is the measure the management uses to evaluate goodness of the process. Details that are included this initial analysis are drawn from high-level sub-groups in the process and are aimed at defining where to focus on the detailed drill-down that follows. The mind map shows an example of possible sub-groups which can be different products (or principal functions of products) or production lines (or primary processes in production operations). Output from this first level EDA is an issue statement, which indicates which performance indicator is out of control, by how much, and what is the desired state to be achieved (although not yet a target or goal as the ability to achieve this result has not yet been determined). An example of such an issue statement is: *“Production throughput time of product A is five hours longer than planned.”*



Mental Model for Exploratory Data Analysis – Level 1

Defining an Issue Statement based on a deviation in performance

Are there any issues with the process performance?

Defined Data Analysis

Trigger for exploratory analysis

Exploratory Data Analysis

Issue statement

Analysis Planning

Interest for monitoring performance

Analysis plan:
- What is monitored
- Analytical tools
- Data collection plan

Data Collection

Collected data

Check
Do
Act
Plan

Separated project

Flow of EDA

Characteristics

Measure type	Process performance results
Y = f(x)	Y-measure
Level of detail	High-level sub-groups
Data for Pareto	Process and product functions

Mind map

Figure 40. Mental Model for Exploratory Data Analysis – Level 1

Defined Data Analysis is very different compared to the other levels. In the initiating case this approach is supported by the normal performance reporting system used by the management. At the subsequent EDA levels, it is defined based on the need of the analysis. Level 1 defined analysis planning and execution is conducted separately from any Exploratory Data Analysis, because the organization gathers and analyzes this information on a regular basis and uses the outcome of those analysis as a continuous input for their process of daily management of routine work. Detailed planning and design of that system of analysis should be accomplished as part of the development of the production system architecture and its data collection and analysis methods should be automated as much as possible as it is used for regular reporting. EDA should focus on the detectable exceptions to the standard reporting system or those observations that are noticed which do not make sense according to the assumptions made regarding the operating processes.

6.1.2. Exploratory Data Analysis – Level 2

Second level EDA takes the issue statement as an input and seeks to structure a problem in a way that it can be further investigated. Figure 41 represents the details of EDA level 2. Arrows indicate which steps in this level are different from the activities that occurred in the level 1

EDA. The output of analysis planning is an analysis plan which describes how the specific issue will be investigated. The defined data analysis step will then analyze the occurrence of the issue and collect more information about its circumstances. Based on this collected data, the EDA can be conducted to inquire about the data and discover what is interesting about its patterns and relationships.

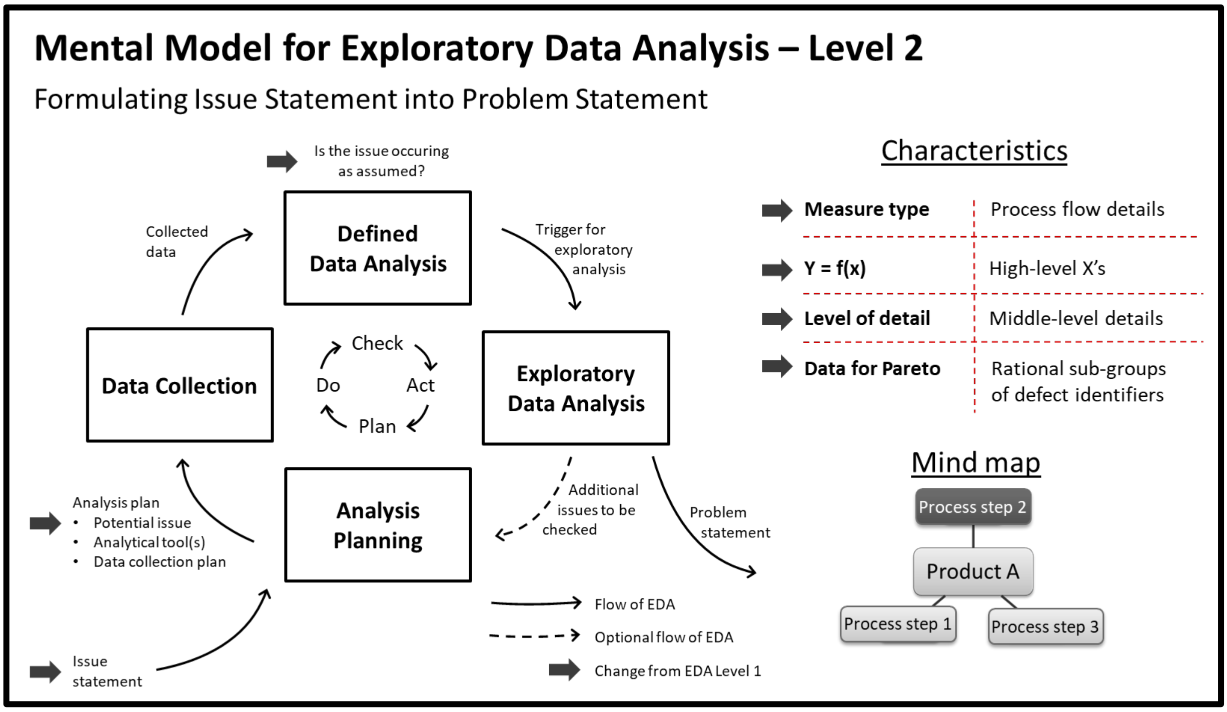
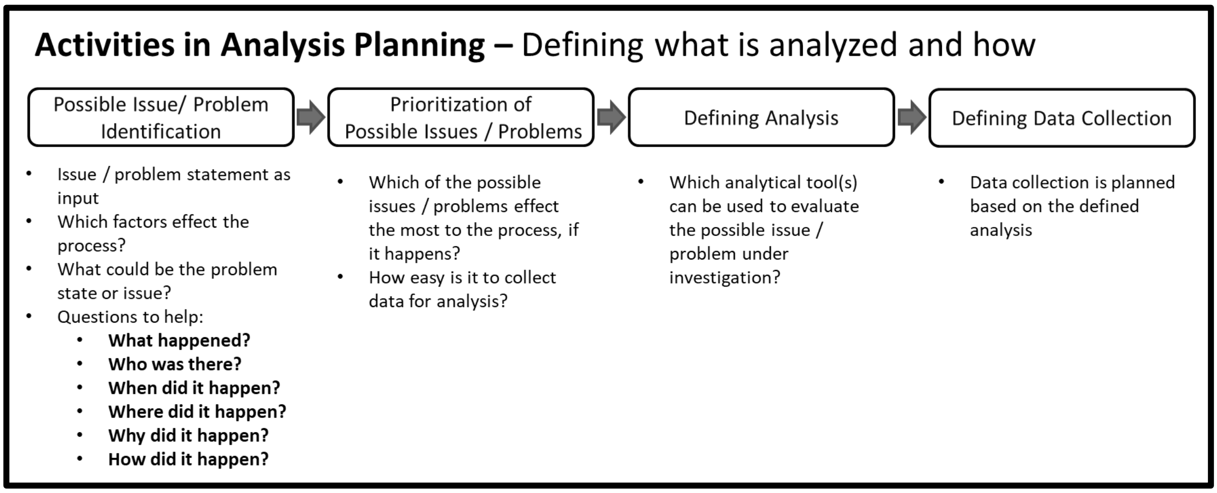


Figure 41. Mental Model for Exploratory Data Analysis – Level 2

Data that is measured or collected to support this second level of EDA is different compared to the first EDA, which uses standard process performance results from the management reporting system. In this second level the focus is on the process flow details, where each process step and the flow between these steps becomes the focus of analysis. Problems that occur in different process steps can be identified by comparing the defects from the Pareto analysis with their origins in the process flow. The level of information detail that comes from this data analysis is middle-level detail which means that data is not the summary data presented to management for evaluating process performance, but, on the other hand, it is specific enough to trace performance contributions of each individual process step. Output of second level EDA is a problem statement which specifies the undesirable circumstance in more detail than the issue statement which just described the broad situation composed of symptomatic observations. The goal of the second level EDA is to identify the problem through data observations of the gap in performance and discover where the problem occurrence can be observed. This adds to the information in the issue statement and initiates a more detailed diagnosis as the problem statement is formulated. If the first rotation of the process does not result in development of a good problem statement, then the issue statement may be reformulated and pursued from a different angle by starting again with the analysis planning step.

The Analysis Planning step has been updated based on feedback from the first round of interviews. Originally this step was called Issue Identification and Selection. Figure 42 presents the activities in the Analysis Planning step for both levels 2 and 3 of the EDA Mental Model. The reason that both possible issues and possible problems are mentioned is so the same model can be applied in both levels. The objective to pursue during this step depends on the input for the step. Additionally, the “5W and 1H" questions have been added to help specify the issue as a problem. This figure also shows that Data Collection is included in this step.



Activities in Analysis Planning – Defining what is analyzed and how

Possible Issue/ Problem Identification

- Issue / problem statement as input
- Which factors effect the process?
- What could be the problem state or issue?
- Questions to help:
  - **What happened?**
  - **Who was there?**
  - **When did it happen?**
  - **Where did it happen?**
  - **Why did it happen?**
  - **How did it happen?**

Prioritization of Possible Issues / Problems

- Which of the possible issues / problems effect the most to the process, if it happens?
- How easy is it to collect data for analysis?

Defining Analysis

- Which analytical tool(s) can be used to evaluate the possible issue / problem under investigation?

Defining Data Collection

- Data collection is planned based on the defined analysis

Figure 42. Activities in Analysis Planning

Example issue statement from EDA level 1 was: “Production throughput time of product A is five hours longer than planned.” The level 2 EDA focuses on where the cause or causes of the issue are located. To continue the same example, the problem statement developed as result of EDA level 2 could be stated as: “*Decrease cycle time of process step 2 for product A.*”

6.1.3. Exploratory Data Analysis – Level 3

The third level of EDA takes a problem statement as its input and defines the solutions space for the problem. Figure 43 illustrates the detailed sequence of activities in the Mental Model for EDA level 3. The problem statement is the input for this level, and it concentrates the investigation for this level. In concrete terms the difference between EDA level 2 and 3 is that on level 2 focus was placed upon the source of the issue, while on level 3 the focus is placed upon under the set of conditions that trigger the problem occurrence. To gain this understanding the analysis needs to be conducted at an operational level. This means that the sub-activities within a process step must be analyzed.

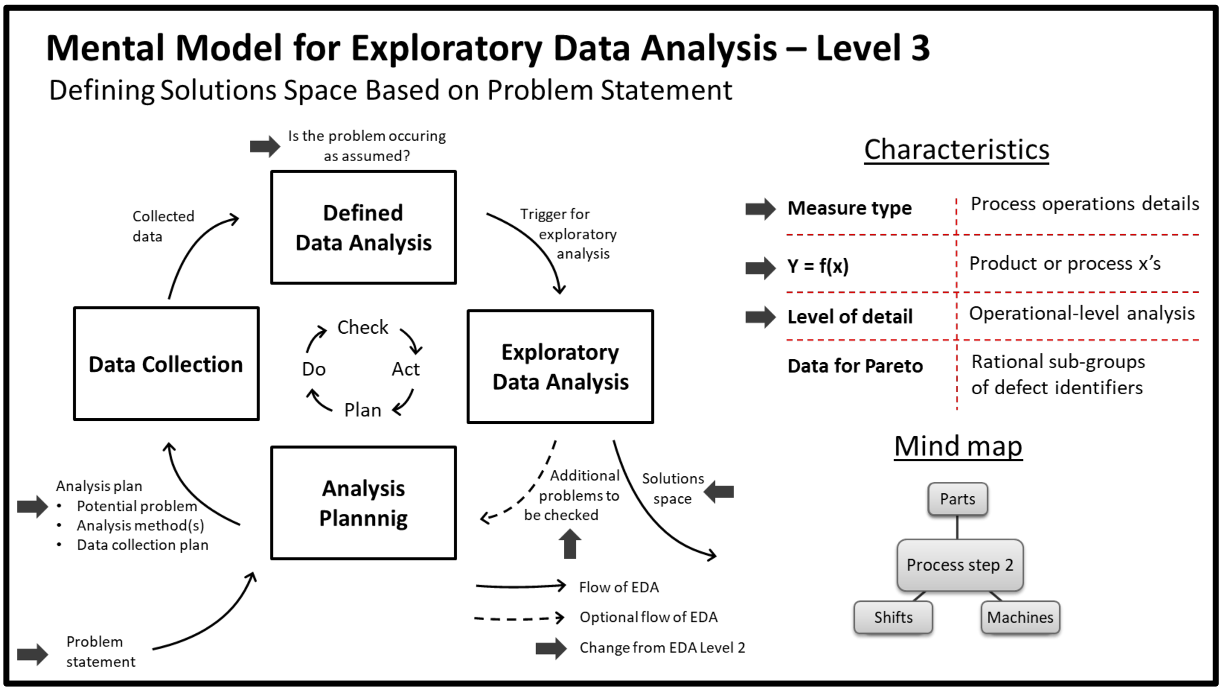


Figure 43. Mental Model for Exploratory Data Analysis – Level 3

The Solutions space can be described using a Mind Map. Basically, the solutions space defines alternatives or opportunities for improvement which have been uncovered through the EDA process. In the example described previously, the problem statement was to: "*Decrease the cycle time for process step 2 for product A.*" In this situation, the solutions space may consist of the specification of the shift which generated the problem; the particular machine which contributed the most defective parts; the individual operator who had the worst performance and needs additional training; and the specific defects found in the parts which may indicate a physical issue within the equipment or material which produced the part. This solutions space becomes the focus for the detailed experimentation and analysis that follows by applying the DMAIC method.

6.1.4. Logic of Exploratory Data Analysis

The name assigned to the finalized model is "Mental Model for Exploratory Data Analysis" and just one step will be referred to as "Exploratory Data Analysis." Figure 44 presents the logic of this Exploratory Data Analysis step. It starts with Exploratory Analysis Planning, where the analysis is planned based on the collected data which triggers the Exploratory Analysis. In this planning step the analytical methods and tools are selected, and interest is generated in the topic to be analyzed. If the data observations require processing by filtering, sorting, converting, or sub-grouping, then this activity is done in the data organizing step. The data analysis is performed in the exploratory analysis step and results are analyzed in the evaluation step.

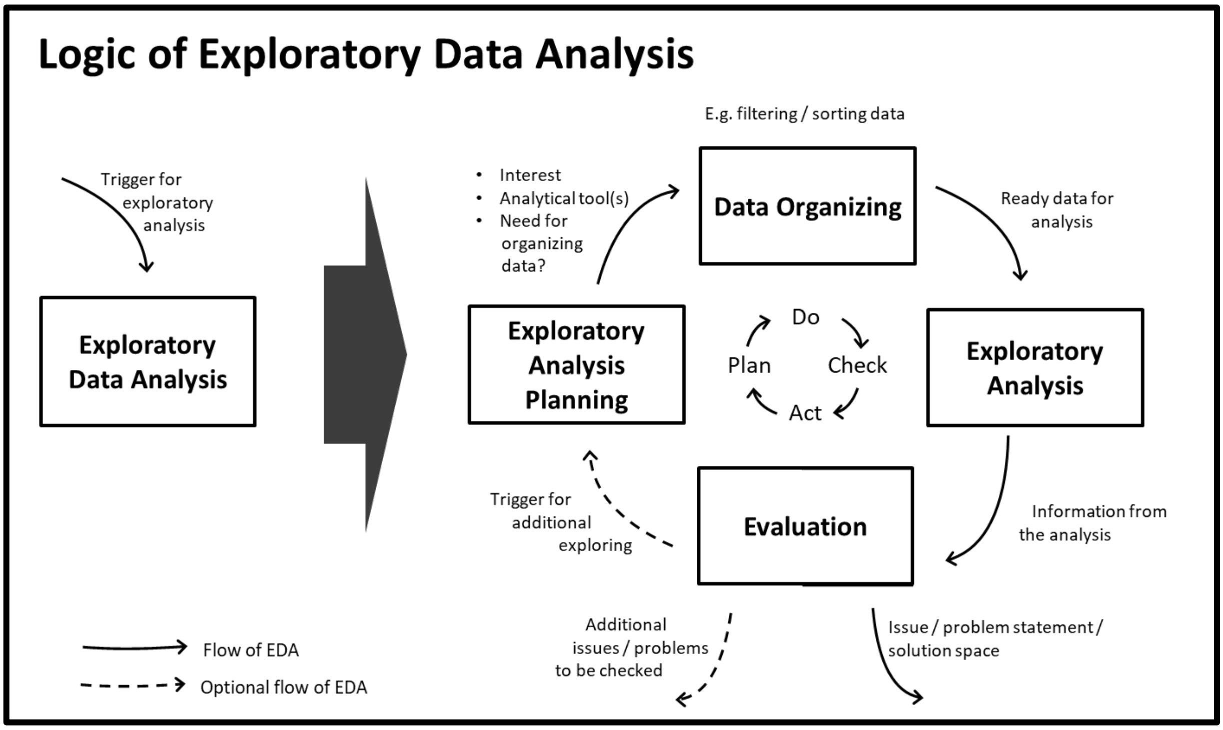

Logic of Exploratory Data Analysis
Trigger for exploratory analysis
Exploratory Data Analysis
E.g. filtering / sorting data
• Interest
• Analytical tool(s)
• Need for organizing data?
Data Organizing
Ready data for analysis
Exploratory Analysis Planning
Do
Plan
Check
Act
Exploratory Analysis
Trigger for additional exploring
Evaluation
Information from the analysis
Flow of EDA
Optional flow of EDA
Additional issues / problems to be checked
Issue / problem statement / solution space

Figure 44. Logic of Exploratory Data Analysis

The Evaluation step can have three different types of outcomes. The first one provides output for the next phase of the Mental Model. This may consist of an issue statement, problem statement or solutions space. Second type of outcome forms the trigger to proceed to the next round of the EDA on the same analysis level by suggesting an additional issue or problem to investigate. Third type of outcome from the Evaluation step is initiation of the need to continue with additional data exploring. This third outcome means that the Exploratory step identified something interesting to analyze in more detail or that there is something remains to be analyzed in the data set. This type of iteration with the Exploratory step occurs frequently and the iteration acts like a progressive revelation of the process knowledge. On some occasions, the detailed evaluation of the physics, or chemistry of the parts needs to be accomplished in collaboration with a content matter expert, where the analysis is performed jointly in order to develop an appropriate problem statement.

An example of a trigger for Exploratory Data Analysis can be illustrated in the example shown in Figure 45. Here there is observed high variation in the cycle times of process step 2 for product containing part B. This figure represents a situation where the factor “shift” is selected as the interesting sub-group for further analysis.. Because situations where products containing part B were the focus of the analysis, observations which only contain them are analyzed with respect to different shifts. In practice the analysis of this example would proceed such that Data organizing would filter the observations so only products with part B would be included. Then in the Exploratory Analysis step, analysis of the cycle times for the different shifts would be made using ANOVA or an I Chart with stages representing the shifts. After this, the Evaluation step would analyze the results of this sub-set of the entire problem or further decompose the situation to look at operators of the equipment within the shift if one shift dominates the variation production.

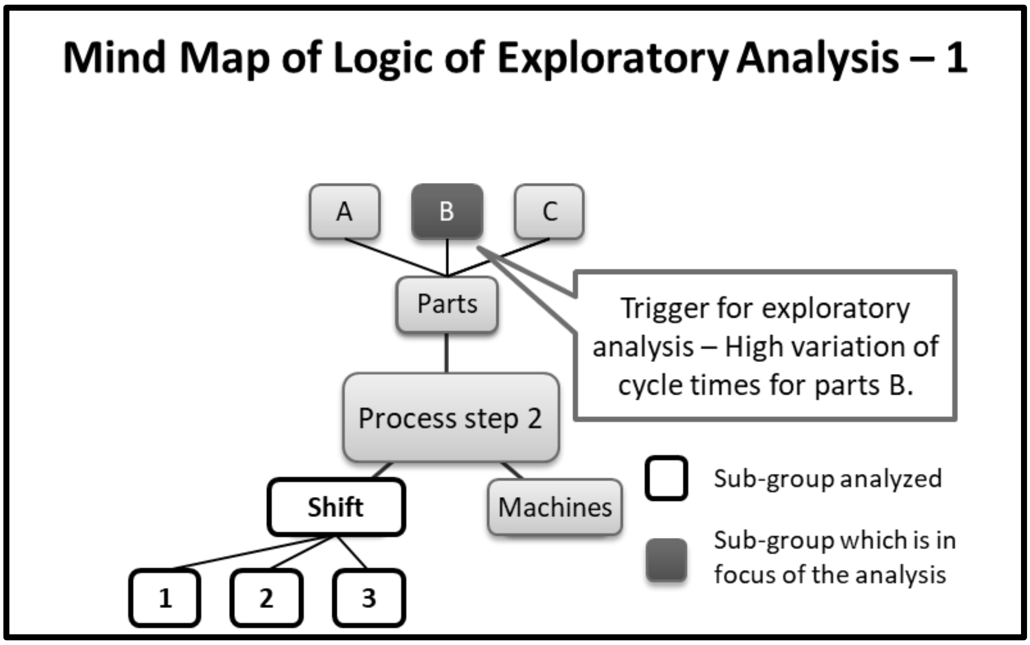

Mind Map of Logic of Exploratory Analysis – 1
A
B
C
Parts
Process step 2
Shift
Machines
1
2
3
Trigger for exploratory analysis – High variation of cycle times for parts B.
Sub-group analyzed
Sub-group which is in focus of the analysis

Figure 45. Mind Map of logic of Exploratory Analysis – situation 1

This example can be expanded to the next level as presented in Figure 46. The first round of Exploratory Analysis showed that shift 2 had high variation in cycle times for process step 2. That finding yielded a trigger for exploring further to analyze differences in cycle time between machines for products containing part B as well as restricting the inquiry to focus only on the second shift. Now, a similar procedure repeats the analysis work that was done in the first situation. Now, the Data organizing step filters data to contain only values for the second shift and just products with part B. Analysis by ANOVA or I chart with stages for different machines can then be used in the Exploratory analysis step. Results of that analysis can be evaluated to determine the possible solutions space. In this example, the solution spaces could be identified as: "*Products with part B produced on machine M during the second shift.*"

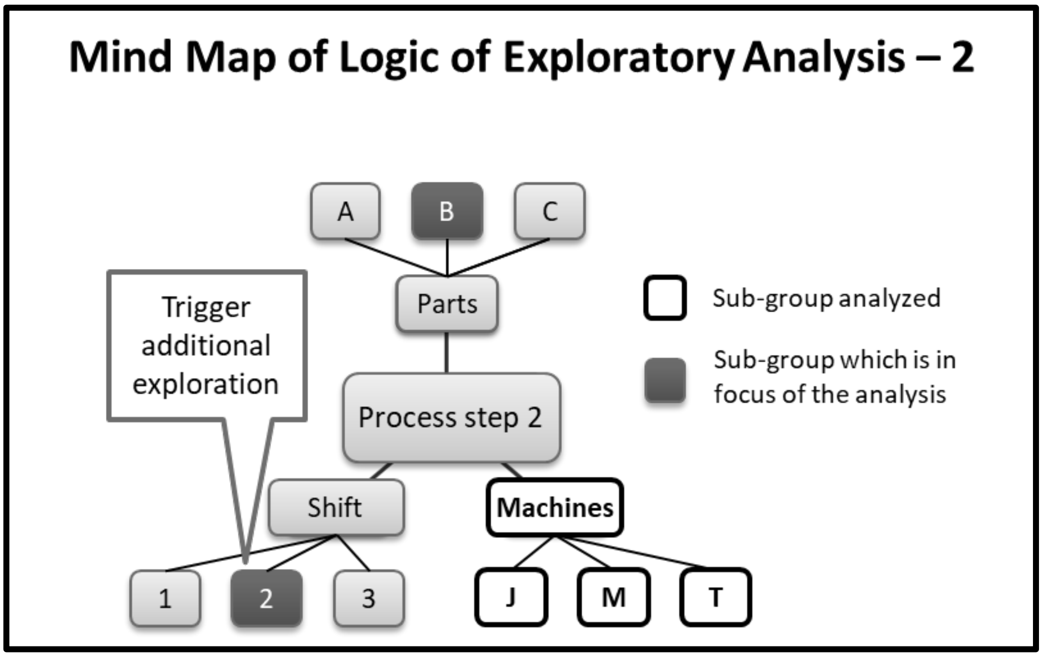

Mind Map of Logic of Exploratory Analysis – 2
A
B
C
Parts
Trigger additional exploration
Process step 2
Shift
Machines
1
2
3
J
M
T
Sub-group analyzed
Sub-group which is in focus of the analysis

Figure 46. Mind Map of Logic of Exploratory Analysis – Situation 2

6.1.5. From Undesired Performance to Solutions Space

The journey from initially observing a performance issue to establishing a bounded solutions space may be divergent rather than direct. Figure 47 illustrates the journey from issue statement to solutions space. This example is the same as used throughout this chapter. It starts with the issue statement: “Assembly time of product A is too high” This is quite vague. To understand the problem more precisely the location of the observation is needed. This is uncovered during the second level EDA. The potential issue can be identified that process step 2 is the bottleneck. Once analysis shows that this is true, then the following problem statement was formulated: “Decrease cycle time of process step 2 for product A”.

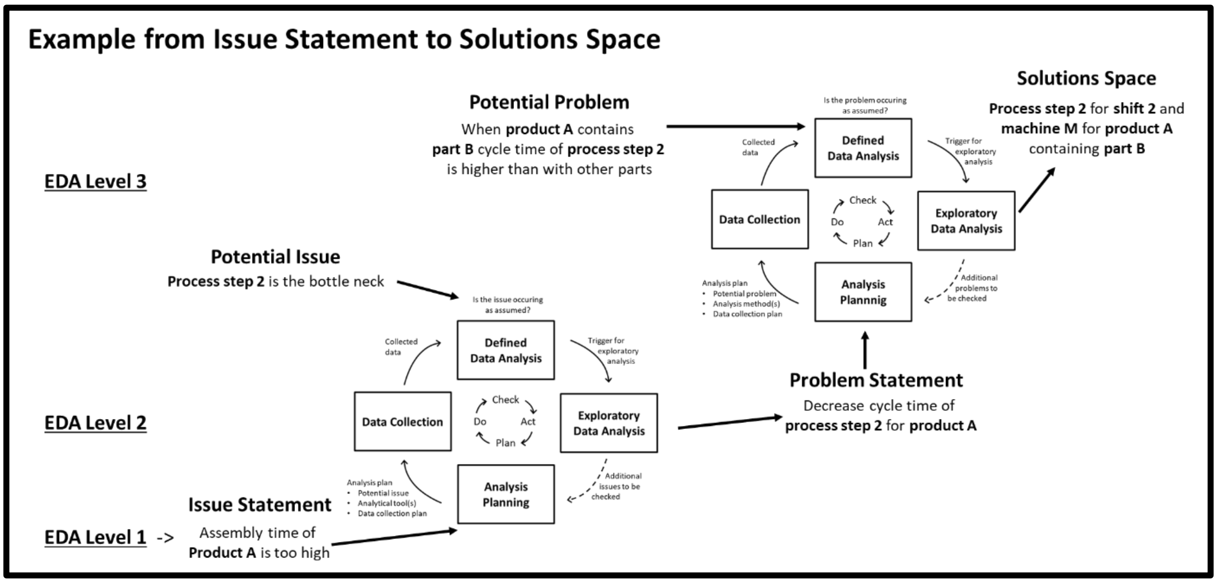


Figure 47. Example from Issue Statement to Solutions Space

The third level of EDA investigates possible reasons for the problem. For the example the potential problem investigated was that the cycle time is higher when product A contains part B. After deeper analysis, the solutions space was formulated as the follows: “Process step 2 for shift 2 and machine M for product A containing part B.” The potential problem could have also been formulated to look at the machine M or second shift, and the exploratory data analysis would most likely have yielded the same solutions space if the collected data would have included all the necessary problem factors.

7. DISCUSSION OF THE MENTAL MODEL FOR EXPLORATORY DATA ANALYSIS

The revised version of the Mental Model for Exploratory Data Analysis was presented to four Lean Six Sigma Master Black Belts. These content matter experts are different ones than those who were interviewed in the first development round of the model. All of these experts worked for different companies. After a presentation of the revised model, five questions were asked.. The presentation and questions asked are included in Appendices B.

7.1. Usage of Exploratory Data Analysis Differs from Company to Company

EDA focuses on data analysis, therefore, the more relevant data available, the better the ability to conduct the analysis. MBB 8 commented: *"In order to use the presented model effectively, the organization needs be quite mature in process management where needed data is easily available."* When interpreting the comments from the content matter experts interviewed in this second round, it is good remember that they all are working for different companies so merging their ideas into a consensus is not appropriate.

Some content matter experts perform EDA before the Define Phase. Some call these activities before Define Phase the Recognize Phase for DMAIC. MBB 7 commented that he uses this way: *"I do quite a lot of EDA before Define phase. Sometimes the whole GB and BB project can be exploratory in nature and the whole project goes around DMA and based on that we define a new project. This model describes what we are doing there."* The next step after DMA is to Improve and then Control (completing the DMAIC sequence). Sometimes these steps can be done as a separate project. This is the approach used by MBB 7: *"Many times we get an engineering project after DMA and specifications for engineering team are given based on those findings."*

In the beginning of structured problem-solving projects EDA can be performed on the Y-measure which means that data of interest to manager is analyzed. Interestingly, this implies that organizations can get more value out of EDA, if the managers can use the EDA methods themselves. MBB 7 commented the following for using EDA more widely in organizations:

"*Anther way EDA is done is by process owners on monthly basis, where they are doing EDA 1, There Quality department updates the standard issues and data collection once a year to update analysis system."* Interestingly, the role of quality department which acts in a data collection and screening function. With respect to the mental model, this means that the first planned analysis is done by the quality department and that the organization has developed its data management system to create this data for reporting purposes. Once something unusual is detected from this screening, Exploratory Data Analysis is immediately initiated and it can, at least partly, be done by the process owner.

Based on MBB comments, EDA is not only used in as a Define activity. MBB 6 commented about the nature of problem definition: *"Describing problem more clearly is an iterative process, which takes place in Define and Measure phases"*. MBB 8 linked the reality of EDA and mental model: *"Repetitive EDA cycles are good, and they reflect well the reality of exploratory data analysis when new data sets are collected to drill down to the problem."*

For some the mental model describes what happens later in DMAIC projects. MBB 6 pointed out that: *"Based on my experience this describes more measure phase. This might be due to the fact the problem description is not that clear after define phase."* View of MBB 9 was supporting the view of MBB 6: *"EDA is used quite rarely in Define phase; however, it is very good to use EDA in the Define phase. If there is any data on the process, EDA helps to quantify the magnitude of the problem as well as to scope it."* Most likely the analysis activities that are mentioned by MMB 9 need to be done at some point of the project, and the earlier they are done, the better.

7.2. Benefits of Using the Presented Mental Model

In general, some benefits of applying EDA methodology as part of DMAIC projects were pointed out by MBB 9: *"Doing EDA in Define phase on high level guides the project to focus on sub-groups on more detailed level in later phases of the project. Doing EDA in Define and Measure phases usually helps to get to the root causes faster."* Similar points of view were presented by content matter experts in the first round of interviews.

MBB 7 would do EDA before starting DMAIC projects: *"Exploratory Data Analysis is most useful before DMAIC and this should be part of Recognize phase. Based on my experience many GBs and BBs get confused on EDA because they are looking for solutions and strong signals and they think that they are in analyze phase."* Discussion about additional steps for DMAIC are out of scope for this thesis, but the applying this mental model might decrease the probability of confusion of GBs and BBs about EDA as there has not been clear mental model for EDA included in most training of GBs or BBs. That argument was supported by MBB 8: "*Previously EDA has been presented as set of tools which are used to do graphical analysis. This model presents a structure and process for using the tools when doing EDA with three different type of cycles which work together."* MBB 6 comment was more neutral about the model: *"This might be a good model to describe the thinking logic in EDA."*

7.3. Target Group of the Mental Model

The model was perceived useful especially for application by MBBs and BBs. MBB 8 commented about the use case for MBBs: *"This model is useful for a person who manages LSS project portfolio of an organization. This is for upper competence persons, especially MBB's who is in role of LSS deployment champion, because the analysis of one issue might yield to many projects."* MBB 9 had similar a view: *"If I would be an MBB in a very data rich environment, I would use the model by myself to give very detailed projects."*

Mental models are very helpful when basic analysis skills are initially taught to people. MBB 7 had the following view on the usefulness of the model: *"The model helps on teaching new Black Belts on project definition, because the questions are very different there than when looking at the solutions to the problems. On the other hand, many times when EDA is done well, the solutions can be a "just-do-it" type of project, because the problem wasn't perceived as a problem before."* Extending this idea, it is important to consider that teaching EDA more effectively has the potential to significantly increase the effectiveness of Black Belts for solving problems which may not have been discovered otherwise. MBB 9's view was more related to executing of DMAIC projects: *"For scoping and quantifying the problem the model would be helpful for black belts."*

Champions represent management's view during DMAIC projects. Based on of the interviewed MBBs, Champions do not necessarily need to know the mental model for EDA. MBB 9 suggested: *"Sponsors and champions should know what EDA is in general. Tollgate reviews are in nature more making a decision to continue the project than to understand the whole loops of EDA."* MBB 8 had a similar point of view: "*Champion don't need to know the model itself, but the results of the analysis need to be understood by the sponsor as the projects proposed by the deployment champion."* MBB added more to the role of Champions: *"Champions should challenge MBBs and BBs to give them better problem statements. This model helps to describe the journey to define problem statement."* In order to satisfy this challenge for MBBs and BBs to develop better problem statements, the EDA journey can help. Perhaps champions don't need to know the details of the journey but understanding it in a high-level of abstraction could help them to be better in their role as an assessor of the project and steerer to assure that the business benefits are obtained.

The complete mental model became quite complex. This was noted by a question which MBB 9 asked after observing the model during the interview: "*What all have you developed*?" My answer to that was: "*A mental model to represent what happens in EDA for a single data set and how similar logic can be applied in different levels of detail with different type of data.*" In addition to this answer to the question, it can also be noted that the model integrates the journey from a vague, abstract, ambiguous issue through different levels of details to achieve a clear problem statement that leads to a solutions space where the search for improvement solutions can commence. People using the full mental model need to have good understanding about the meaning of Y = f(x) for their particular case. That is probably why MBB 6 commented: "*The Model might be a bit too complex and it might be easier to talk about to others if the model is simpler. The presented model might be confusing for Green Belts if that is presented in addition to DMAIC"*

7.4. Limitations of the Mental Model

The biggest limitations seen in the mental model was the concern about validating the data quality. MBB 7 stated: *"The model doesn't take into account what to do with new data sets and what to do with financial data. Model does not state how quality of raw data is validated."*

Similar comments were stated by the first pool of content matter experts. That is something to consider in the future research. MBB 6 had a similar kind of point of view: *"The model doesn't into account the reliability of data, especially when going into flow data."* Dr. Watson (2019d) felt that this is not an important objection as: *"Management is already making decisions on this data. If it is good enough for their decision-making, then it is good enough for starting the exploration of the data. Cleaning data or other preliminary steps are not necessary."*

As pointed out previously, the maturity of an organization in conducting data analysis, effects its ease in applying the EDA mental model. MBB 8 commented: *"This model can be applied more easily in mature organizations."* It seems that the different versions of EDA that have been developed in the past might have shifted the focus of the interviewee because MBB 9 commented about the usability of the model: *"How the model is now positioned limits it's usage in case there is no data available straight away."*

This comment contradicts comments made by the first pool of content matter experts. They liked the idea that the first version of the EDA mental model forced the collection of the right type of data. That part of the model was not changed in the revised model, because the second step of the revised EDA model still conducts data collection. However, the way that the revised model was presented to the content matter experts as part of the second-round interview was different from the first round, and this might have caused the perception by the second pool of MBBs. Despite this observation, data collection can take a long time, which makes it harder to apply the EDA mental model as management typically wants fast answers. Thus, it is easier to apply the EDA mental model if an organization is more mature in their approach to process management and they have already developed a sound measurement system for management information so the d is readily data available for analysis.

In some cases, using the EDA approach might change the way projects are accomplished, which may negatively affect the way that the model is perceived. MBB 9 observed: *"Theoretically it can slow down the start of the project, because some companies do a project kick-off in the end of Define phase where the decision to start the project is made."* To avoid that kind of situation the benefits of the preliminary data analysis need to be communicated to the management.

7.5. Suggested Development Points for the Mental Model

The only qualitative method presented in this thesis was Ishikawa analysis. MBB 7 suggested: "*Exploratory mindset is larger than what the model presents, e.g., Gemba walks could be added to it. There should be understanding if the perceived results can even happen in real life."* Several authors and some of the interviewed content matter experts have emphasized the need of process knowledge when conducting analysis. The analyst does not need to personally possess deep knowledge about the process, but that knowledge should be gained from some individuals who work with the team to do the EDA inquiry. Gemba walks and utilization of people with process knowledge are good great additions to the model in the future. Other non-quantitative methods and tools can also be added to the model. This was proposed by MBB 6*: "The X-Y matrix could be a useful tool for defining the factors chosen for investigation."*

Adding operational definitions of the terms and description of goals was also suggested. MBB 6 commented: *"Adding Operational definitions of characteristics as well as describing the goals of each step and relations ships to tools would be good additions."* MBB 9 had a similar view: *"Clear operational definitions of the terms would be useful."* Concrete illustration of the value of operational definitions was pointed out by an additional comment made by MBB 9: "*Personally as a non-native English speaker, the difference between issue and problem statement is a bit hard to distinguish, because in my own language they are the same word. How you have now presented the problem statement, can be used as project goal in some cases."*

Because data quality was commented upon as a weakness in the model, it was suggested that this capability should be developed in a future model. MBB 8 commented: *"If the model is used in Define, there should be a focus on some of the Measure phase activities, especially MSA and data quality."* MBB 7 was pointed that*: "The role of the quality team in gathering and integrating EDA level 2 data could be added to the model."* Using the quality team for data management and preliminary analysis and reporting would help to increase the organization's process maturity and the EDA model could serve as a framework for guiding the organization to a new level of sophistication in data analytics.

Linking advanced methods such as the statistical Design of Experiments (DoE) was suggested by MBB 7: *"DOE methods can be used to analyze failures which customers have reported. That can be called as exploratory testing to replicate the failure mode and understand better what has caused it."* MBB 2 had a similar type of view. These methods are typically used in the Improve phase of DMAIC and should be considered in the future when EDA becomes more closely tied to DMAIC.

Clearer presentation of the mental model was suggested by MBB 6: "*Showing characteristics of different EDAs on same slide would makes it easier to understand them."* MBB 9 suggested: *"You could map the model better by showing the example on the overview slide."*
Improving the graphical clarity of the model will also become a future consideration when the model is extended.

7.6. Summary of the Critique of the Model

The interviewed content matter experts did not disagree on the fundamentals of the EDA mental model. They pointed out how ease of its application depends on the process maturity of the organization where it is applied. The model was seen particularly helpful for MBBs to define multiple projects in a system-wide improvement program as well as for BBs to develop better problem statements and focus their analysis.

The biggest concerns were raised about data validation. Previously EDA has been seen as something to do for an existing historical data, but since the model links its use to include data collection, there are some grounds to be more concerned about data quality. That is an excellent topic for further research and could be coupled with adding additional concepts, methods or tools to the EDA mental model.

8. CASE EXAMPLE OF APPLICATION OF THE MENTAL MODEL FOR EDA

This chapter presents the application of the model in a real-world context. The mental model was applied in a case company to support a structured problem-solving project which was conducted in collaboration with a supplier.

8.1. Case and Company Introduction

The case company was founded more than one hundred years ago. It has grown into a global company with global operations. The EDA case project was conducted at one of its manufacturing units in Finland.

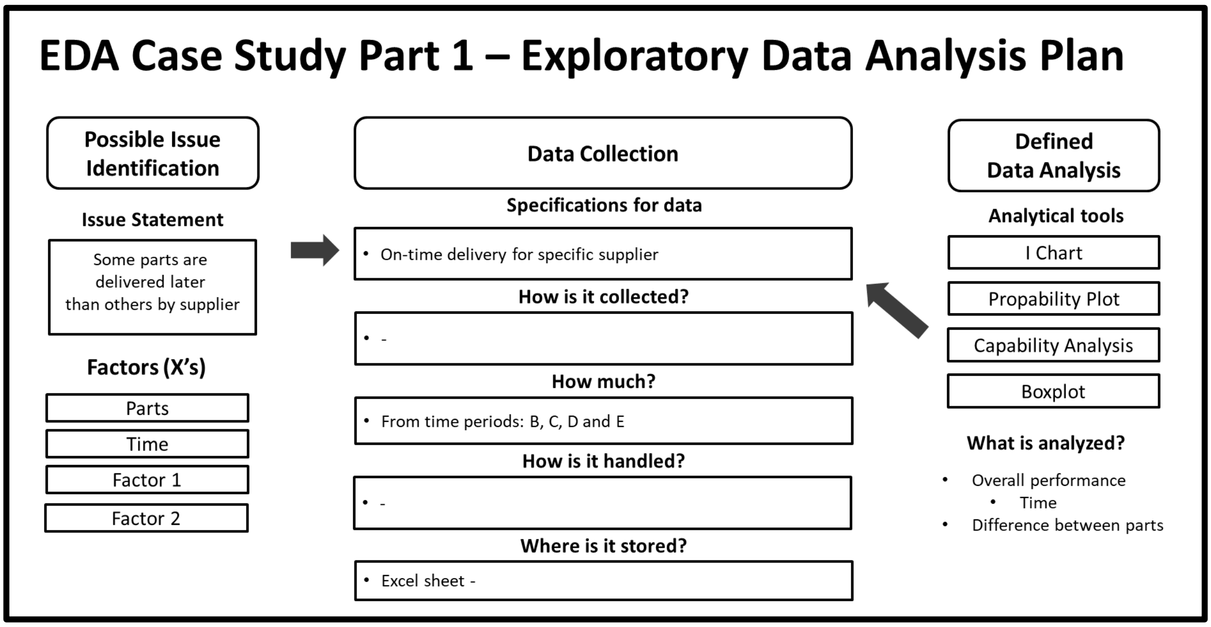
Since 2007 the company has been training its employees as Lean Six Sigma Yellow Belts, Green Belts, Black Belts, Master Black Belts, and everyone is required to attend awareness training as a White Belt. This is an engineering company and employees are accustomed to making calculations and data analysis as part of its daily activities.

The case project was performed within a Lean Six Sigma Black Belt project, which addressed the improvement in delivery performance of a supplier. The EDA mental model was applied to support the data analysis for the project. The analysis was limited to investigate only high-level data (the output or Y-measure) for on-time deliveries.

8.2. **Applying the Mental Model for Exploratory Data Analysis**

The initial condition at the start of the project was that the performance measurement of the supplier on-time delivery showed undesired performance, which caught management’s attention. An improvement project was initiated to get performance back on track. This case study demonstrates how the mental model could have been applied in this case project using the measurements that were collected by the company. To start this improvement project an issue statement needed to be developed. The trigger for this was the observed poor performance by this supplier. Management had a preliminary “hunch,” based on their past experience with this company, that some parts delivered by the supplier were later than others. This intuitive insight acted as a trigger to initiate the Exploratory Data Analysis. The data analysis methods regularly used by the company did not permit directly conducting an Exploratory Data Analysis, so the data needed to be extracted from its SAP system and analyzed with statistical software called Minitab.

As the first step, the analysis was planned. Figure 48 presents the analysis plan. The potential issue to be checked was defined: ”*Some parts are delivered later than others by supplier JDK*.” That type of issue definition is very broad and could not be presented for discussions with the supplier. The following factors were identified for the simplified analysis presented in the case: parts, time, factor 1 and factor 2. Details of the data collection are not shared as part of the case. The outcome measure selected for analysis is on-time delivery. Analytical tools selected were I-Chart, probability plot, capability analysis, and boxplot.



EDA Case Study Part 1 – Exploratory Data Analysis Plan

Possible Issue Identification

Issue Statement

Some parts are delivered later than others by supplier

Factors (X’s)

Parts

Time

Factor 1

Factor 2

Data Collection

Specifications for data

- On-time delivery for specific supplier

How is it collected?

- -

How much?

- From time periods: B, C, D and E

How is it handled?

- -

Where is it stored?

- Excel sheet -

Defined Data Analysis

Analytical tools

I Chart

Propability Plot

Capability Analysis

Boxplot

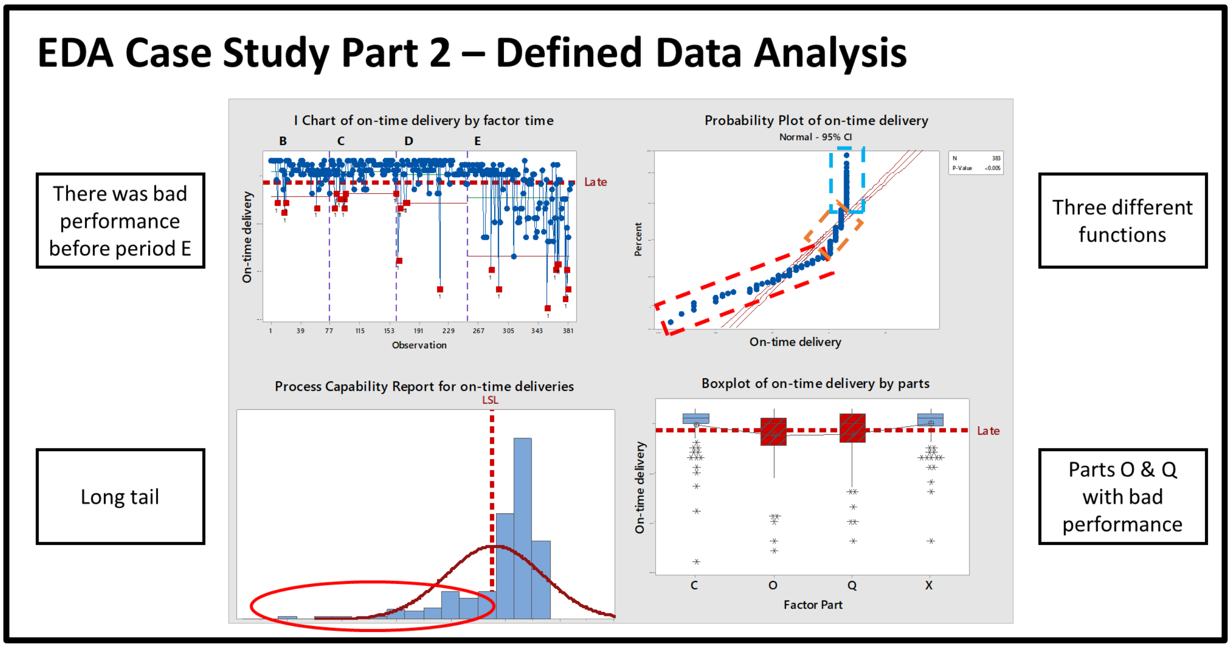
What is analyzed?

- Overall performance
  - Time
- Difference between parts

Figure 48. Exploratory Data Analysis Plan

Output for the defined data analysis is shown in Figure 49. For the performance indicator selected the rule for interpretation is “bigger performance is better,” because on-time deliveries are calculated by subtracting target delivery date from the actual delivery date. On the I-Chart, Process capability analysis, and boxplot a red line is added to indicate the time after which the deliveries are considered to be late.

I-Chart with stages provides an overview of the performance across the historical time periods. In period E the performance was the worst, but it cannot be said to be bad only in that period, because there have been late deliveries during every time period. The Probability plot shows that there have been three different functions of performance, which are marked in the picture. Process capability shows a long tail on left, which indicates that some of the deliveries have been late a lot more than others. The Box plot indicates that parts O and Q have been the parts with the worst performance, because the majority of the data in these boxes are partly below the red line. Some deliveries of parts C and have also been late, but not as consistently as these others.



EDA Case Study Part 2 – Defined Data Analysis

There was bad performance before period E

Three different functions

Long tail

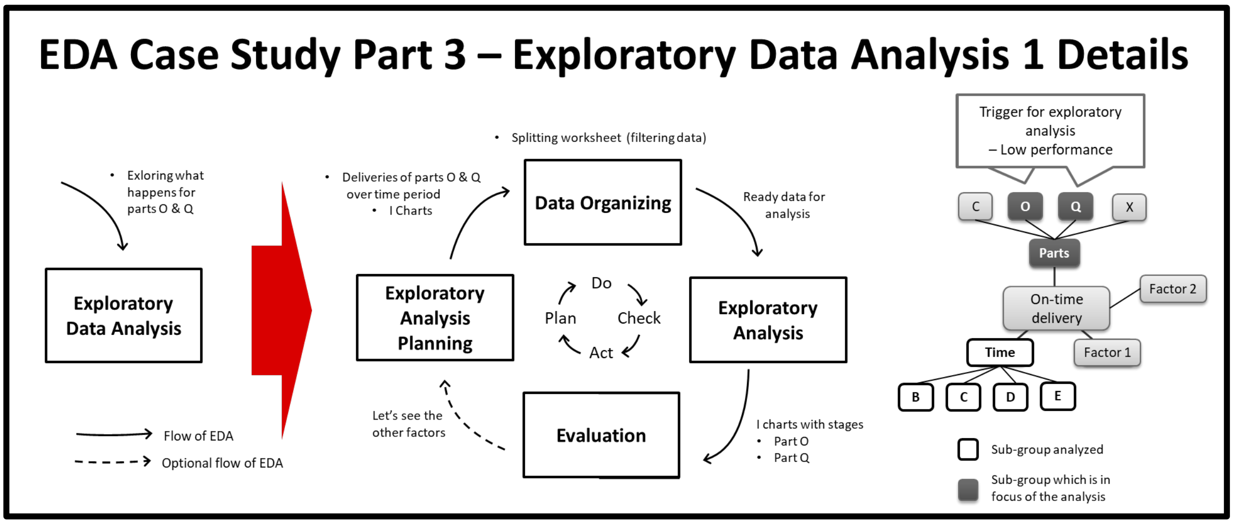
Parts O & Q with bad performance

Figure 49. Defined Data Analysis

The planned analysis indicates that there clearly was a performance problem with parts O and Q. Based on that information the issue statement could be updated to: "*Some of the parts O ja Q have been delivered late by supplier.*" Still there is additional information available in the data set, which can be used to focus the analysis even more precisely. One of the goals of EDA is to look at least one level deeper, which is accomplished by engaging in the Exploratory Data Analysis step of the mental model.

Observed bad performance for parts O and Q trigger the Exploratory Data Analysis, which is the fourth step in this loop of the mental model. Figure 50 illustrates the details of Exploratory Data Analysis 1. Ishikawa Analysis is presented on the right side of the figure in mind map format to provide an overview of the situation. The dark boxes are the parts, which are in focus of the Exploratory Data Analysis. Boxes with white background illustrates the time periods which are interest in this first iteration of the Exploratory Data Analysis. Boxes with light grey are the factors which have been identified in planning the analysis as possibly important ones and the data set also contains information about these factors.

In the Exploratory Analysis Planning step an I-Chart was selected to present the deliveries of parts O and Q over the historical time period. This highlighted a need to split the analysis into two views by organizing the data independently for parts O and Q in the data organizing step. After this step the data was ready to be used for the Exploratory Analysis step where the two graphs in Figure 51 were made with Minitab.



EDA Case Study Part 3 – Exploratory Data Analysis 1 Details

- Exloring what happens for parts O & Q

Exploratory Data Analysis

Flow of EDA

Optional flow of EDA

- Deliveries of parts O & Q over time period
  - I Charts

Exploratory Analysis Planning

- Splitting worksheet (filtering data)

Data Organizing

Ready data for analysis

Do

Plan

Check

Act

Exploratory Analysis

I charts with stages
- Part O
- Part Q

Evaluation

Let's see the other factors

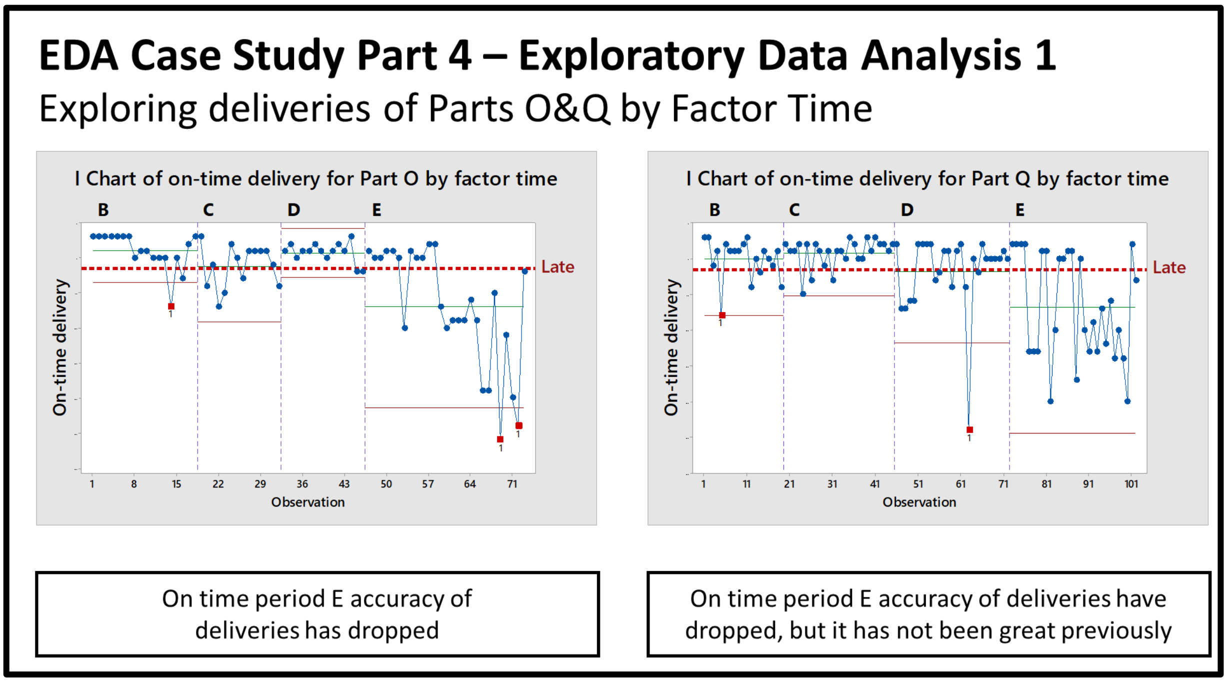
Trigger for exploratory analysis – Low performance

□ Sub-group analyzed

■ Sub-group which is in focus of the analysis

Figure 50. Exploratory Data Analysis 1 Details

These two I-Charts (with stages) are presented in Figure 51 as analyzed in the evaluation step of the inner EDA loop. These I-Charts reveal interesting information about the performance of the supplier. In general, there have always been some Q parts, which have not been delivered on time during each time period. Deliveries of O parts have been more constant prior to time period E, but some parts have always been late. In time period E, there has been a significant increase in the amount of both parts O and Q that were delivered late.



EDA Case Study Part 4 – Exploratory Data Analysis 1

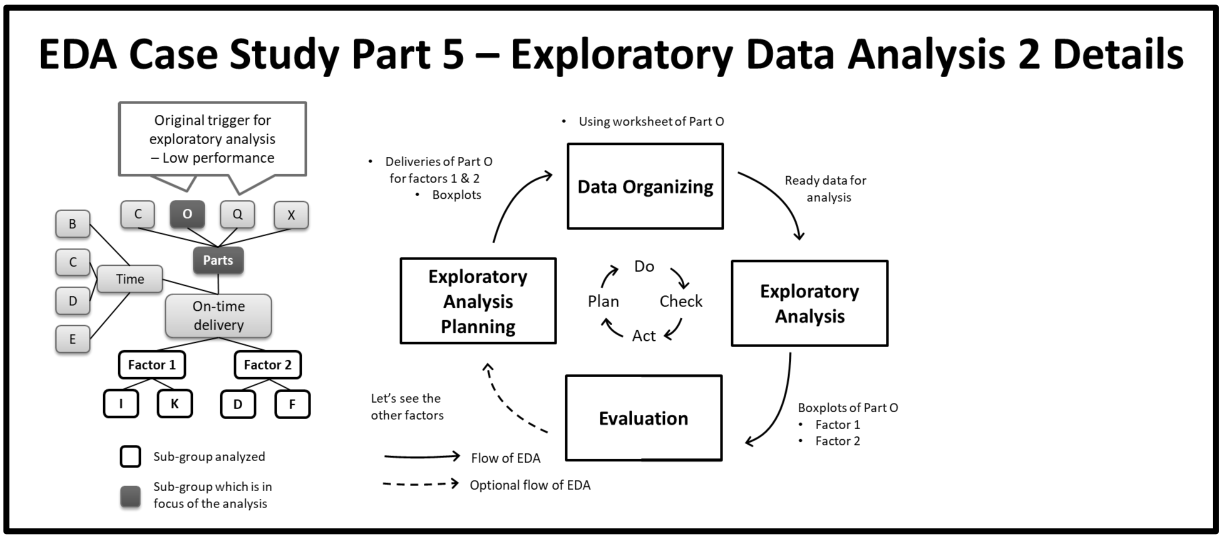
Exploring deliveries of Parts O&Q by Factor Time

On time period E accuracy of deliveries has dropped

On time period E accuracy of deliveries have dropped, but it has not been great previously

Figure 51. Exploratory Data Analysis 1 Results

The first iteration of Exploratory Data Analysis revealed that especially part Q has had on-time delivery issues even before time period E. There are additional factors to investigate in the data set. Figure 52 represents the details of the second iteration of Exploratory Analysis. The interest is in performance of part O with regards factors 1 and 2. Boxplot were selected as statistical tool for analysis of this situation; however, I-Charts with stages could also have been used. In the Data organizing step the only concern is to use the right work sheet in Minitab, as it was split in the first iteration of the EDA loop. The output of this Exploratory Analysis is the two boxplots shown in Figure 53.



EDA Case Study Part 5 – Exploratory Data Analysis 2 Details

Original trigger for exploratory analysis – Low performance

B
C
D
E
Time
C
O
Q
X
Parts
On-time delivery
Factor 1
Factor 2
I
K
D
F

□ Sub-group analyzed

■ Sub-group which is in focus of the analysis

- Using worksheet of Part O

Data Organizing

Ready data for analysis

Exploratory Analysis

- Deliveries of Part O for factors 1 & 2
  - Boxplots

Exploratory Analysis Planning

Do
Plan
Check
Act

Boxplots of Part O
- Factor 1
- Factor 2

Evaluation

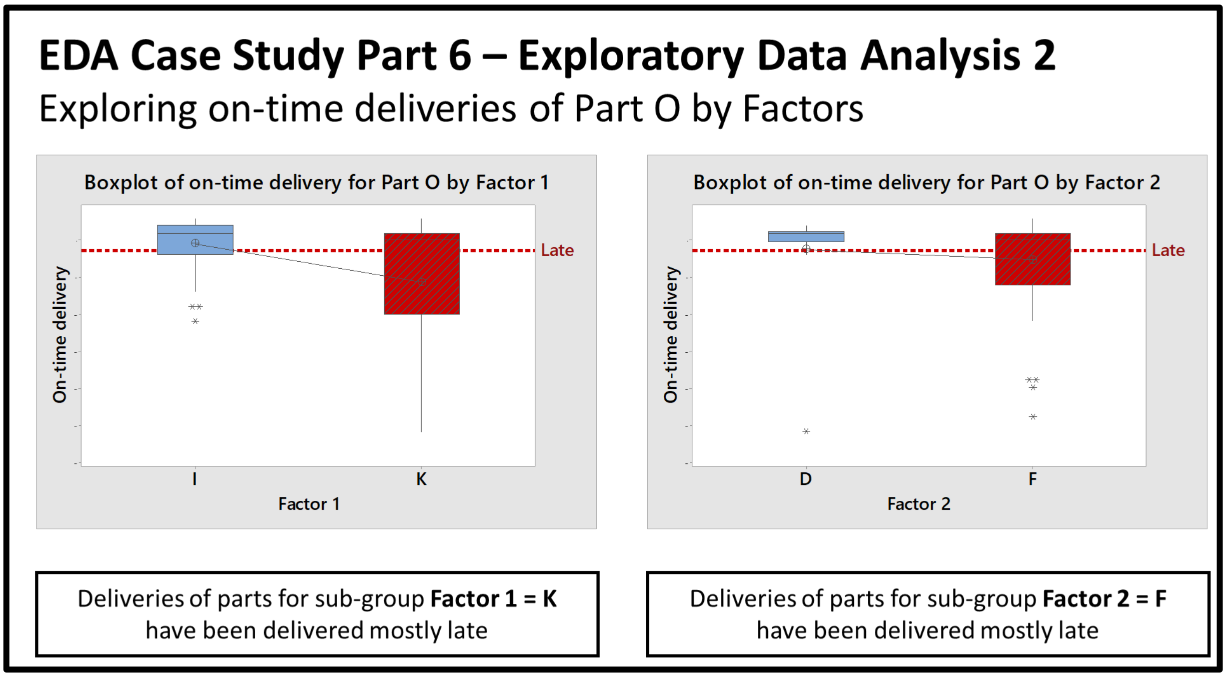
Let's see the other factors

Flow of EDA

Optional flow of EDA

Figure 52. Exploratory Data Analysis 2 Details

Analysis of the different factors increased understanding of the situation. Exploratory Analysis 2 presented by figure 53 shows that for factor 1 value K has been a character, which has been identifying late deliveries. Same logic can be applied for F value for factor 2. That observation leads to a desire to analyze the performance of part Q using these two same factors.



EDA Case Study Part 6 – Exploratory Data Analysis 2

Exploring on-time deliveries of Part O by Factors

Boxplot of on-time delivery for Part O by Factor 1

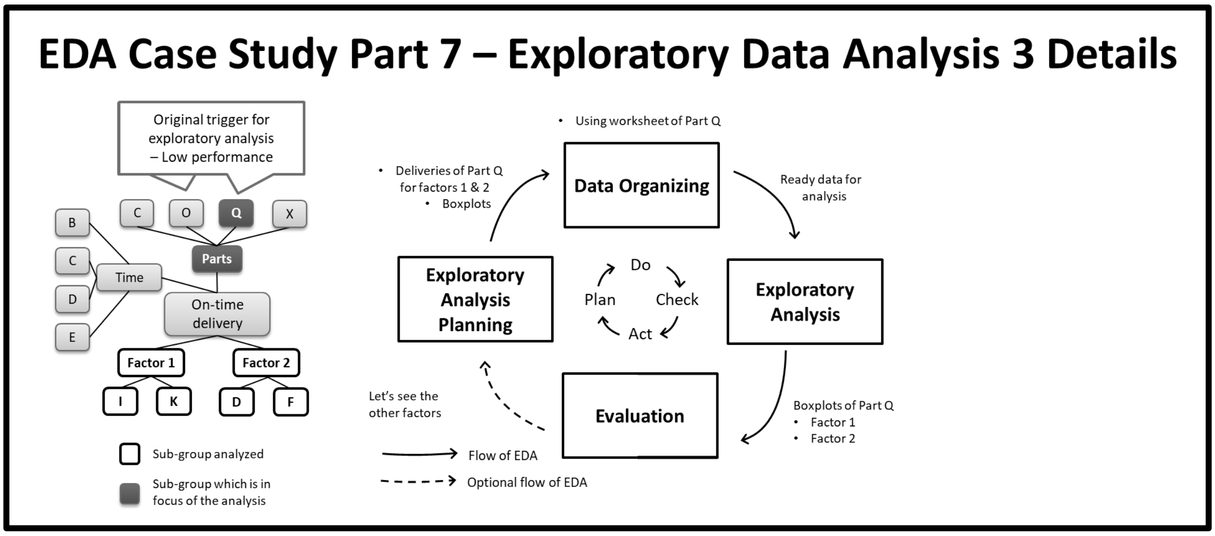
Boxplot of on-time delivery for Part O by Factor 2

Deliveries of parts for sub-group **Factor 1 = K** have been delivered mostly late

Deliveries of parts for sub-group **Factor 2 = F** have been delivered mostly late

Figure 53. Exploratory Data Analysis 2 Results

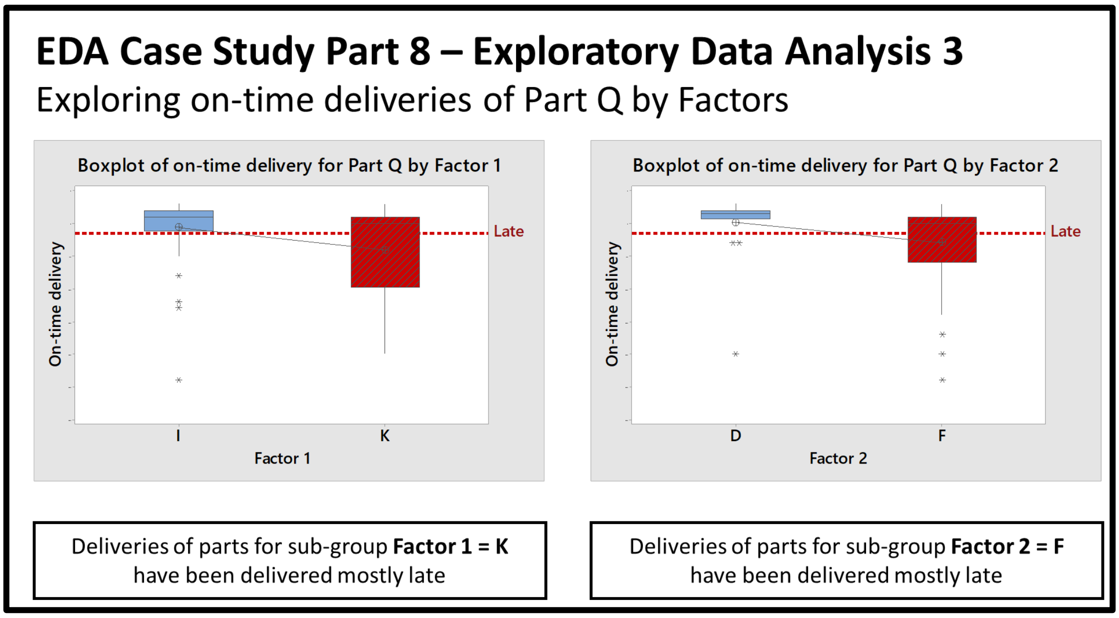
Details of the Exploratory Data Analysis are shown in Figure 54. The details are the same as for the previous analysis shown in Figure 51 with one exception, that the part analyzed there was O, instead of Q.



EDA Case Study Part 7 – Exploratory Data Analysis 3 Details

Figure 54. Exploratory Data Analysis 3 Details

After conducting the same analysis on part Q the results are illustrated in Figure 55. This shows similar results as did the analysis for part O. For factor 1 parts with value K has been delivered late as well as factor 2 = F.



EDA Case Study Part 8 – Exploratory Data Analysis 3

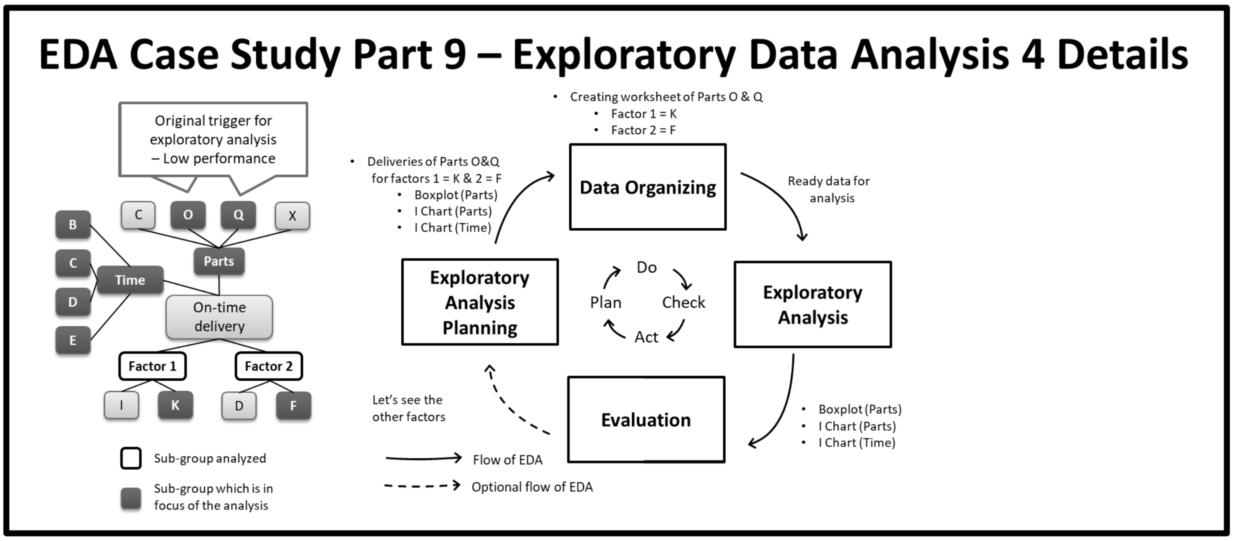
Exploring on-time deliveries of Part Q by Factors

Deliveries of parts for sub-group **Factor 1 = K** have been delivered mostly late

Deliveries of parts for sub-group **Factor 2 = F** have been delivered mostly late

Figure 55. Exploratory Data Analysis 3 Results

The next natural step goes a level deeper in the analysis by inquiring about the performance of both parts when factor 1 = K and factor 2 = F, because both parts were performing badly with those factor values. Overview of the details of the next iteration of Exploratory Data Analysis loop is presented by Figure 56. In this case boxplots were selected to show differences between parts O and Q. Additionally, I-Charts with stages were used to see how the performance of both parts O and Q with the selected values for factors performed during different time periods. In this case some organizing of data was needed to create a worksheet in Minitab with values for the parts O and Q with these selected values.


EDA Case Study Part 9 – Exploratory Data Analysis 4 Details

Original trigger for exploratory analysis – Low performance

- Creating worksheet of Parts O & Q
  - Factor 1 = K
  - Factor 2 = F

- Deliveries of Parts O&Q for factors 1 = K & 2 = F
  - Boxplot (Parts)
  - I Chart (Parts)
  - I Chart (Time)

Ready data for analysis

- Boxplot (Parts)
- I Chart (Parts)
- I Chart (Time)

Let's see the other factors

Sub-group analyzed

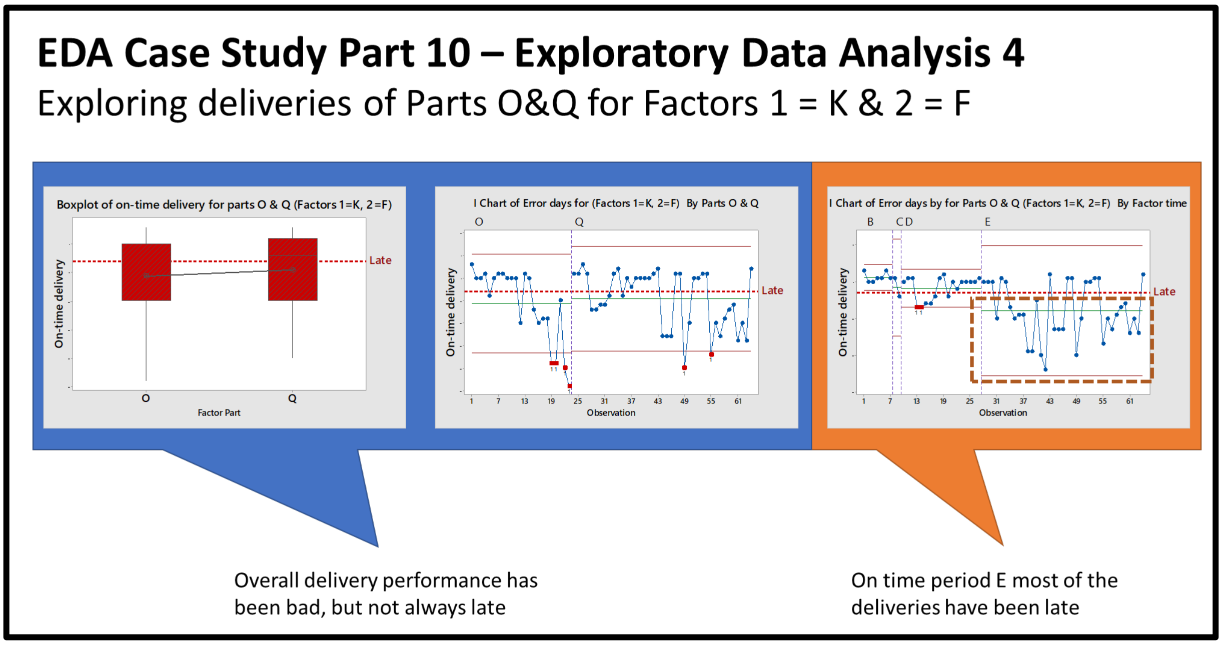
Sub-group which is in focus of the analysis

Flow of EDA

Optional flow of EDA

Figure 56. Exploratory Data Analysis 4 Details

Analysis results of fourth iteration of the EDA loop are presented by Figure 57. Boxplots for both parts look quite similar, which means that both parts have had delivery issues. Additionally, the I-Chart in the middle of the figure provides the information that deliveries of O parts have been a bit later than deliveries of Q parts. The assumption here is that the data is in time order inside each specific stage. Additionally, the latest delivery of part Q has been on-time. The I-Chart on the right shows that most of the deliveries have been late during time period E.



EDA Case Study Part 10 – Exploratory Data Analysis 4

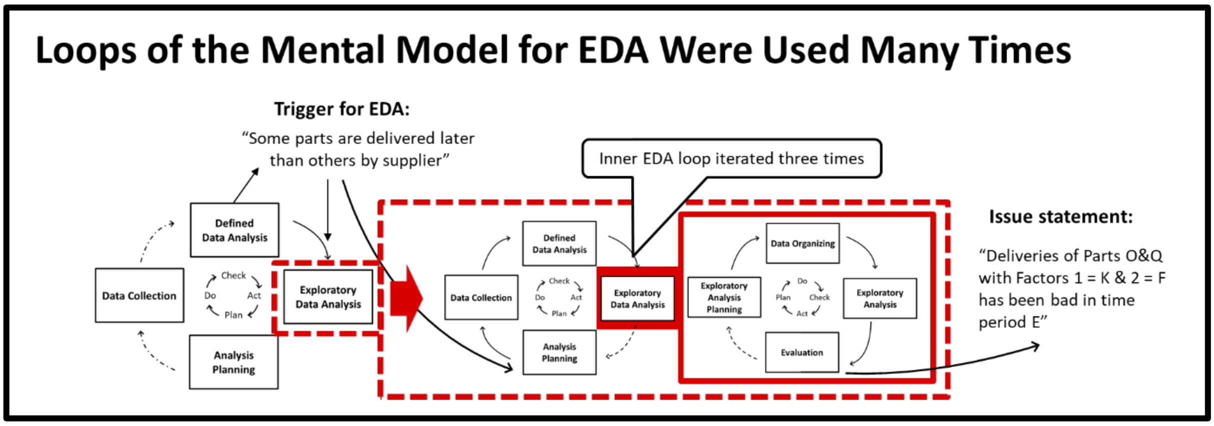
Exploring deliveries of Parts O&Q for Factors 1 = K & 2 = F

Overall delivery performance has been bad, but not always late

On time period E most of the deliveries have been late

Figure 57. Exploratory Data Analysis 4 Results

The outcome of the Exploratory Data Analysis resulted in the construction of the following issue statement: "*Deliveries of Parts O and Q with Factors 1 = K and 2 = F have been bad in time period E.*" A critical point for the analysis is when there is no need to analyze the data set further. In this case the point where it was possible to construct the issue statement acted as a point to end the analysis. Figure 58 present the journey of the Exploratory Data Analysis and its application of the mental model. Defined analysis of Y-measure used by the management triggered need for EDA. Utilizing the EDA loop with three iterations of the inner EDA loop enabled the analyst to construct the issue statement.



Loops of the Mental Model for EDA Were Used Many Times

Trigger for EDA:

"Some parts are delivered later than others by supplier"

Inner EDA loop iterated three times

Issue statement:

"Deliveries of Parts O&Q with Factors 1 = K & 2 = F has been bad in time period E"

Figure 58. Utilization of the Mental Model in the Case study

9. CONCLUSIONS

9.1. Summary

The goal of this thesis was to develop a mental model for exploratory data analysis in the context of structured problem-solving. The first research question: *What kind of mental models are there of Exploratory Data Analysis for structured problem-solving?* was addressed in chapters two and three, where different mental models in the historical development of structured problem-solving and Exploratory Data Analysis were described. Seven different approaches to Exploratory Data Analysis were introduced in chapter four.

A mental model for EDA was developed in an iterative manner in chapters four through seven, which answers the second research question: "*What type of mental model could be developed to support Exploratory Data Analysis applications for structured problem-solving?"* In the model development process nine Lean Six Sigma Master Black Belts were interviewed in their role as content matter experts on the topics of structured problem-solving and EDA. In general, the feedback from these content matter experts was very positive with respect to the proposed EDA mental model.

A real-life case study that applied the mental model was presented in chapter eight. It answered the third research question: *"How can the mental model developed for Exploratory Data Analysis be used in a real-world case?*" The case study demonstrates how the proposed EDA mental model can be applied in the context of a real-world situation to focus improvement activities. In the case study the proposed EDA mental model is applied only to the output performance measure, in other words, only high-level or summary data is used in the case study.

9.2. **Theoretical Contribution**

This master's thesis developed a new mental model for Exploratory Data Analysis for use in the context of structured problem-solving. Feedback from a pool of highly qualified content matter experts was positive about the model construction and they have noted benefits of applying EDA in their organizations. The most significant benefits of the proposed mental model for exploratory data-analysis include its structured approach to analysis, iterative sequence of analytical steps, and the way that it forces the analyst to think about the questions to be addressed before analyzing at the data.

EDA has been described: “more as an art than a science.” This model has potential to change that perception by linking the tools and activities into a comprehensive mental model for interrogation of performance issues. Additionally, the thesis presents a concrete case study demonstrating how Exploratory Data Analysis can be conducted in a structured way using the proposed mental model. The model might be used in the future to train Lean Six Sigma Black and Master Black Belts and accelerate their learning and ability to conduct their preliminary analyses in a structured way.

9.3. **Future Research**

Interviews of the content matter experts disclosed several topics for future research. Most frequently commented upon was the issue of data quality. The assurance of measurement system integrity and data reliability for performance measures should be integrated into the mental model and this would extend its usability for drilling deeper into the causal structure behind problems. Another way to extend the usage of the proposed EDA mental model is to develop a simpler version for use by people at the operating level of the organization who prefer to have simpler methods for conducting analysis.

Since the case study only used high-level process data, it would be a natural next step to demonstrate how the proposed EDA mental model operates on more detailed process data. Additionally, the ease of application of the proposed EDA mental model should be tested in different business and commercial environments and in organizations with different levels of process maturity. Another point of view regarding model applicability is to apply it for investigating a wider variety of processes and problems in order to demonstrate the breadth of its applicability or the boundaries where it may need to be modified to work in special circumstances.

Some methods and analysis tools were suggested to be added to the model: X-Y matrices and Gemba walks. Increasing the scope of methods and tools may add richness to the proposed

EDA mental model; however, making it overly complex could also limits the value of its application as a “quick-and-dirty” analysis method. Combining the process knowledge of other people to support the analyst in conducting the analysis could help increase the strength of interpretation. This approach to “collaborative analysis” would increase the strength of the analysis by increasing the number of perspectives from which the problem would be investigated. Finally, adding the ability to analyze interaction effects between multiple factors would permit the use of the proposed EDA mental model for more complex engineering or scientific problems

In his book Exploratory Data Analysis (1977) Tukey emphasized that the role of computers would change how data analysis will be conducted. Possibly some parts of the journey described by the mental model for EDA could be accomplished more efficiency by incorporating the latest digital technology. It would be especially helpful to extend the use of statistical tools as management is always interested in answers to the question: ”By how much?”

10. REFERENCES

Abrahamian, J. 2009. Red X Problem Solving. ASQ Joint Meeting – ASQ Automotive Division, Hartford, United States, September 10.

Anderson-Cook, C. M, Lu L., Clark, G., Stephanie, P., DeHart, C., Hoerl, R., , Jones, B., MacKay, J., Montgomery, D., Parker, P. A., Simpson, J., Snee, R., Steiner S., H., Van Mullekom, J., Vining G. G., Wilson, A. G., 2012. Statistical Engineering—Forming the Foundations. *Quality Engineering*. Vol. 24(2), pp. 110-132.

Anand, G. & Kodali, R. 2008. Benchmarking the benchmarking models. *Benchmarking: An International Journal*. Vol. 15, nro. 3, s. 257–291.

American Society for Quality (2019). Tree Diagram. Online. Available: https://asq.org/quality-resources/tree-diagram [Accessed: 06.12.2019]

Cogollo-Florez, J. M., Florez, M. C. & Florez, A. L., 2017. Estimating Process Capability Indices for Inaccurate and Non-Normal Data: A Systematic Literature Review. *Calitatea,* 18(158), pp. 50-59.

Compaq 1991: A Structured Approach to the Design of Statistical Training Courses: Pedagogical Body of Knowledge for Statistical Methods Volume 1. United States: Corporate Quality Office

Craik, K. 1967. The Nature of Explanation. 1 updated ed. Cambridge: Cambridge University Press.

Deming, W. E., 1975. On Probability as a Basis for Action. *The American Statistician*, Vol. 29(4), pp. 146-152.

de Mast, J. & Kemper, B. P. H. 2009. Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case?' *Quality Engineering*. Vol. 21(4), pp. 366-375.

Harry, M. J. 1997. The Vision of Six Sigma: A Roadmap for Breakthrough Volume 2. 5 ed. Phoenix, Arizona: Tri Star Publishing.

Brook, Q. 2014. Lean Six Sigma and Minitab: The Complete Toolbox Guide for Business Improvement. 4 ed. Winchester, United Kingdom: OPEX Resources Ltd.

Hoerl, R. W. & Snee, R. D. 2002. STATISTICAL THINKING Improving Business Performance. Hoboken, New Jersey: John Wiley & Sons.

Kipling, R. 1902. Just So Stories. New York: The Country Life Press.

Minitab (2019a). Minitab 18 Support - Example of Individuals Chart [online] [Cited 2019-08-13] Available: https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/control-charts/how-to/variables-charts-for-individuals/individuals-chart/before-you-start/example/

Minitab (2019b). Minitab 18 Support - Add stages to show how a process changed [online] [Cited 2019-08-13] Available:https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/control-charts/supporting-topics/options/add-stages-to-show-how-a-process-changed/

Minitab (2019c). Minitab 18 Support - Example of Normal Capability Analysis [online] [Cited 2019-08-13] Available: https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/capability-analysis/how-to/capability-analysis/normal-capability-analysis/before-you-start/example/

Minitab (2019d). Minitab 18 Support -Overview for Probability Plot [online] [Cited 2019-08-13] Available: (https://support.minitab.com/en-us/minitab/18/help-and-how-to/graphs/how-to/probability-plot/before-you-start/overview/)

Minitab (2019e). Minitab 18 Support - Pareto chart basics [online] [Cited 2019-08-13] Available: https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/quality-tools/supporting-topics/pareto-chart-basics/

Minitab (2019f). Minitab Express Support – Boxplot overview [online] [Cited 2019-12-04] Available: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/graphs/boxplot/before-you-start/overview/

Sabadka, D., Molnár, V., Fedorko, G. & Jachowicz, T. 2017. OPTIMIZATION OF PRODUCTION PROCESSES USING THE YAMAZUMI METHOD. *Advances in Science and Technology Research Journal.* Vol 11(4), pp. 175-182.

Sheehy, P., Navarro, P., Silvers, R. Keyes, V., Dixon D., Picard D. 2002. The Black Belt Memory Jogger: A Pocket Guide for Six Sigma Success Spiral-bound. United States: GOAL/QPC

Shewhart, W. A. 1939. Statistical Method from the Viewpoint of Quality Control. New York: Dover Publications inc.

Spiridonova, E. & Watson, G. H. 2018. Using Capability Studies to Improve Supply Chain Customer Service. Lean and Six Sigma Conference, Phoenix, United States, March 8.

Steiner, H. S., Mackay R. J.& Ramberg J. S. 2008. An Overview of the Shainin System™ for Quality Improvement. *Quality Engineering*. Vol. 20(1), pp. 6-19.

Notz, W. I. 2012. Statistical Engineering, a Missing Ingredient in the Introductory Statistics Course. *Quality Engineering*. Vol. 24(2), pp. 193-200.

Tukey, J. W. 1977. Exploratory Data Analysis. 1 ed. United States: Addison-Wesley Publishing Company Inc.

Vining, G. 2009. Geoff Vining's Discussion of “Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn From a Well-Known Case?”" *Quality Engineering*. Vol. 21(4), pp. 380-381.

Watson, G. and DeYong, C. 2010. "Design for Six Sigma: caveat emptor". *International Journal of Lean Six Sigma*. Vol. 1(1), pp. 66-84.

Watson, G. H. 2015a. Managing for Quality using LSS Methods. Lean Six Sigma Black Belt Course, Helsinki, Finland, August 19-21.

Watson, G. H. 2015b. Performing Statistical Analyses. Lean Six Sigma Black Belt Course, Helsinki, Finland, August 19-21.

Watson, G. H. 2015c. Structured Problem-Solving Process. Lean Six Sigma Black Belt Course, Helsinki, Finland, August 19-21.

Watson, G. H. 2015d. Business Process Analysis. Lean Six Sigma Black Belt Course, Helsinki, Finland, November 7-11.

Watson, G. H. 2015e. Performance Measurement. Lean Six Sigma Black Belt Course, Helsinki, Finland, November 7-11.

Watson, G. H. 2015f. Exploratory Data Analysis. Lean Six Sigma Black Belt Course, Helsinki, Finland, November 7-11.

Watson, G. W. 2018a. Theory and Practice of Profound Knowledge: An Inquiry Into Quality and Strategy Management. Dissertation. Oklahoma, United States: Oklahoma State University.

Watson, G. W. 2018b. Lean Six Sigma Workshop: Exploratory Data Analysis Method, Master Class, Helsinki, Finland, October 11.

Watson, G. W. 2019a. Basic Problem Solving for Quality Improvement, Advanced Quality Professional Certification Program, Banjul, The Gambia, January 16.

Watson, G. H., 2019b. Shewhart process control and History of Lean Six Sigma [Interview]. Date October 5.

Watson, G. H., 2019c. Kipling Method and nature of X's [Interview]. Date November 2

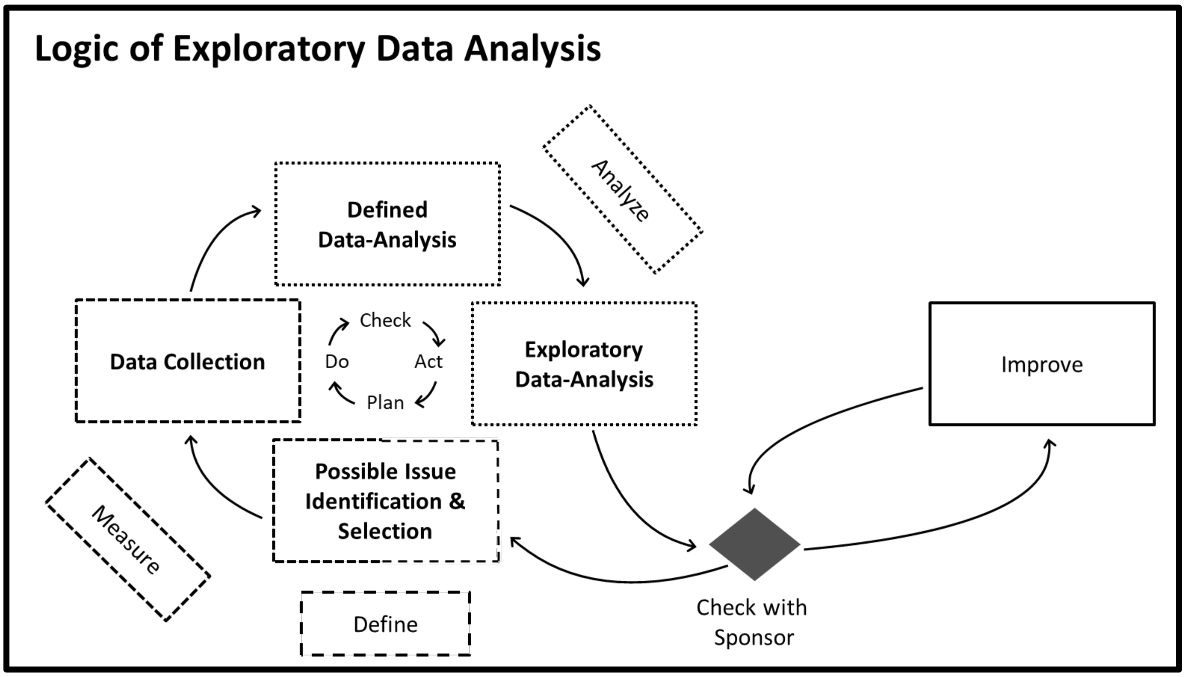
Watson, G. H., 2019d. Need of Measurement System Analysis in the Digital Age [Interview]. Date December 1

Watson, G. H. & Spiridonova, E. 2019b. FISH(BONE) STORIES. *Quality Progress*. Vol. 52(8), pp. 14-23.

World Economic Forum (2018). The Future of Jobs Report. Online. Available: https://www.weforum.org/reports/the-future-of-jobs-report-2018 [Accessed: 31.10.2019]

11. APPENDIX

APPENDIX A: PRESENTATION OF THE PRELIMINARY VERSION OF THE MENTAL MODEL



Logic of Exploratory Data Analysis

Defined Data-Analysis

Analyze

Data Collection

Check

Do

Act

Plan

Exploratory Data-Analysis

Improve

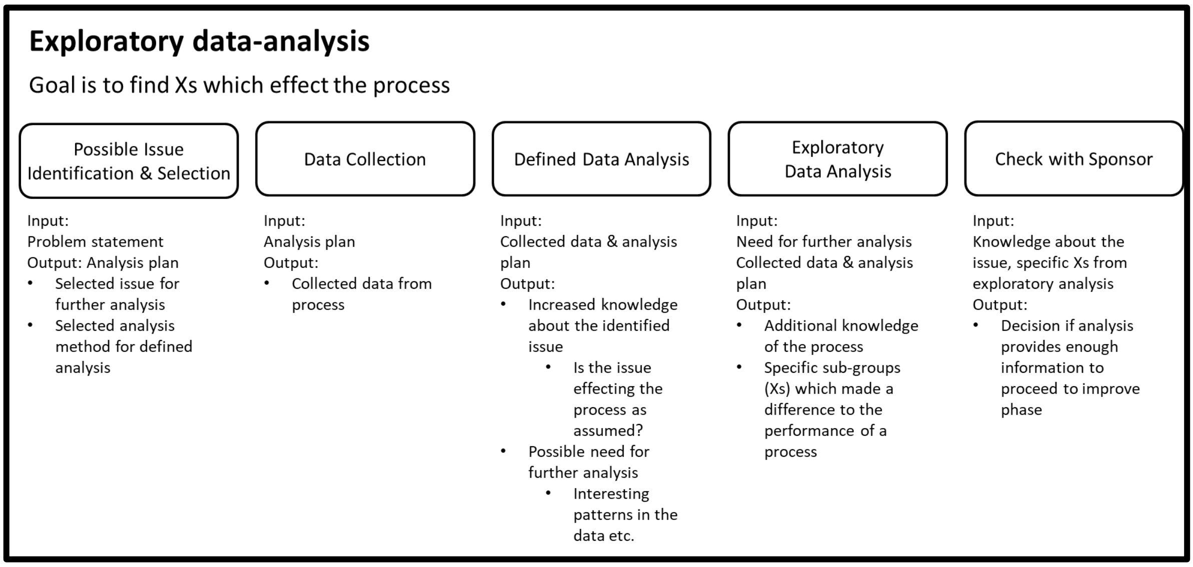
Possible Issue Identification & Selection

Measure

Define

Check with Sponsor

Figure 1. Slide 1



Exploratory data-analysis

Goal is to find Xs which effect the process

Possible Issue Identification & Selection

Input:
Problem statement
Output: Analysis plan

- Selected issue for further analysis
- Selected analysis method for defined analysis

Data Collection

Input:
Analysis plan
Output:

- Collected data from process

Defined Data Analysis

Input:
Collected data & analysis plan
Output:

- Increased knowledge about the identified issue
    - Is the issue effecting the process as assumed?
- Possible need for further analysis
    - Interesting patterns in the data etc.

Exploratory Data Analysis

Input:
Need for further analysis
Collected data & analysis plan
Output:

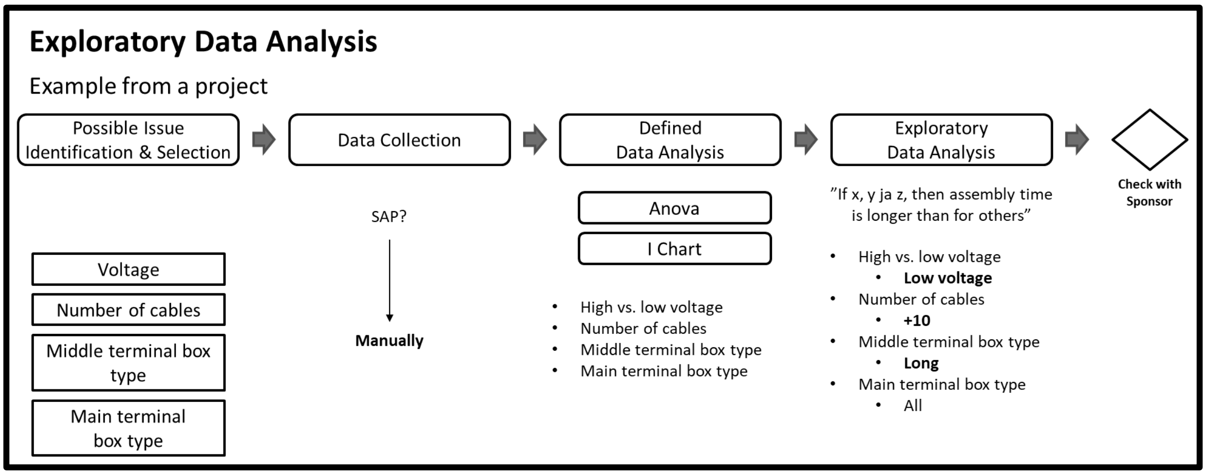
- Additional knowledge of the process
- Specific sub-groups (Xs) which made a difference to the performance of a process

Check with Sponsor

Input:
Knowledge about the issue, specific Xs from exploratory analysis
Output:

- Decision if analysis provides enough information to proceed to improve phase

Figure 2. Slide 2



Exploratory Data Analysis

Example from a project

Possible Issue Identification & Selection → Data Collection → Defined Data Analysis → Exploratory Data Analysis → Check with Sponsor

Votage

Voltage

Number of cables

Middle terminal box type

Main terminal box type

SAP?

Manually

Anova

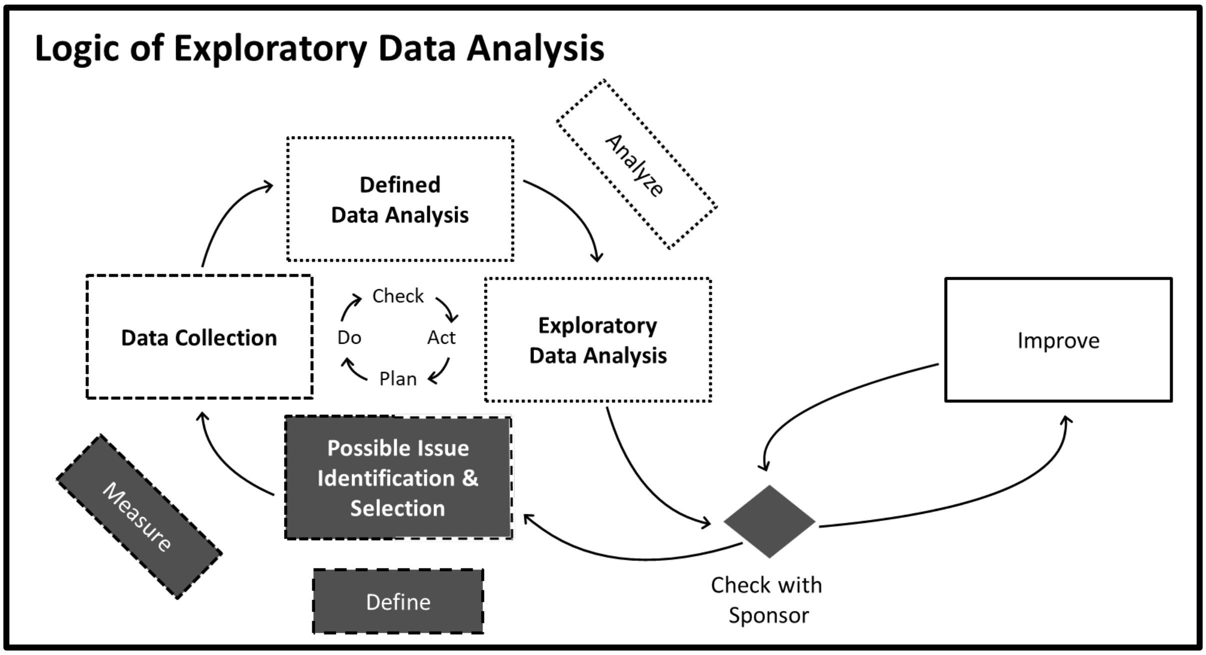
I Chart

- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

”If x, y ja z, then assembly time is longer than for others”

- High vs. low voltage
  - **Low voltage**
- Number of cables
  - **+10**
- Middle terminal box type
  - **Long**
- Main terminal box type
  - All

Figure 3. Slide3



Logic of Exploratory Data Analysis

Defined Data Analysis

Analyze

Data Collection

Check

Do

Act

Plan

Exploratory Data Analysis

Improve

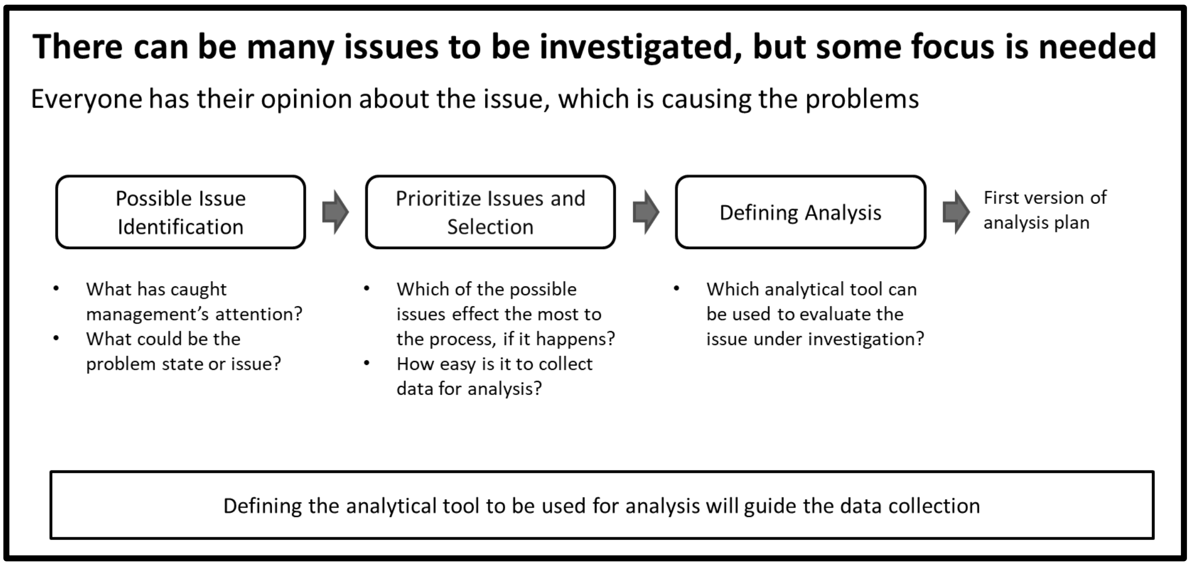
Measure

Possible Issue Identification & Selection

Define

Check with Sponsor

Figure 4. Slide 4



There can be many issues to be investigated, but some focus is needed

Everyone has their opinion about the issue, which is causing the problems

Possible Issue Identification → Prioritize Issues and Selection → Defining Analysis → First version of analysis plan

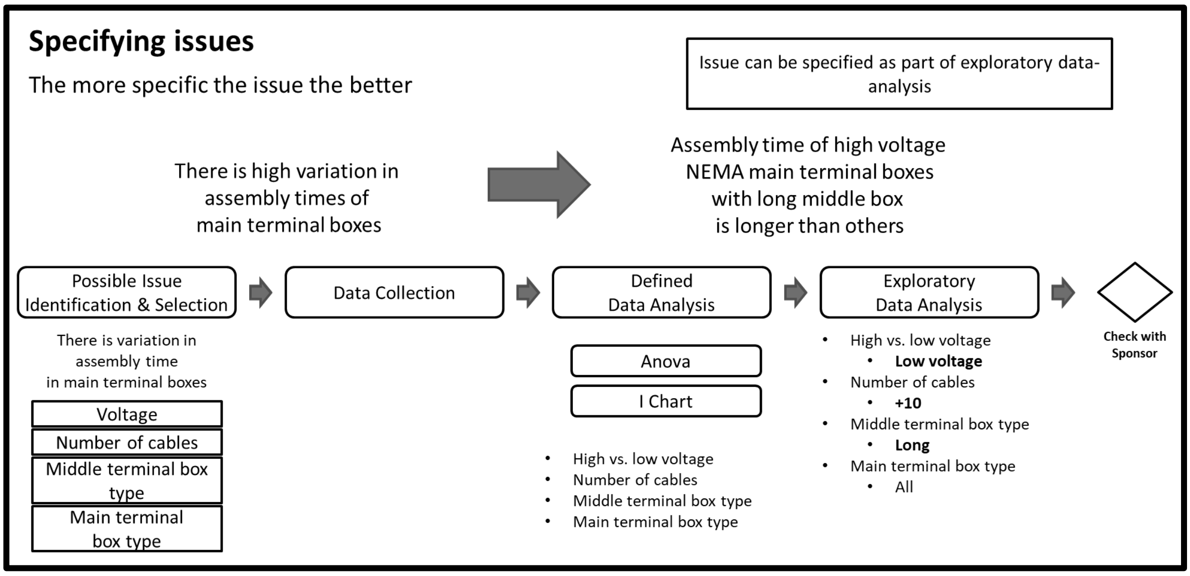
Possible Issue Identification
- What has caught management's attention?
- What could be the problem state or issue?

Prioritize Issues and Selection
- Which of the possible issues effect the most to the process, if it happens?
- How easy is it to collect data for analysis?

Defining Analysis
- Which analytical tool can be used to evaluate the issue under investigation?

Defining the analytical tool to be used for analysis will guide the data collection

Figure 5. Slide 5



Specifying issues

The more specific the issue the better

Issue can be specified as part of exploratory data-analysis

There is high variation in assembly times of main terminal boxes → Assembly time of high voltage NEMA main terminal boxes with long middle box is longer than others

Possible Issue Identification & Selection → Data Collection → Defined Data Analysis → Exploratory Data Analysis → Check with Sponsor

Possible Issue Identification & Selection: There is variation in assembly time in main terminal boxes

- Voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Defined Data Analysis:

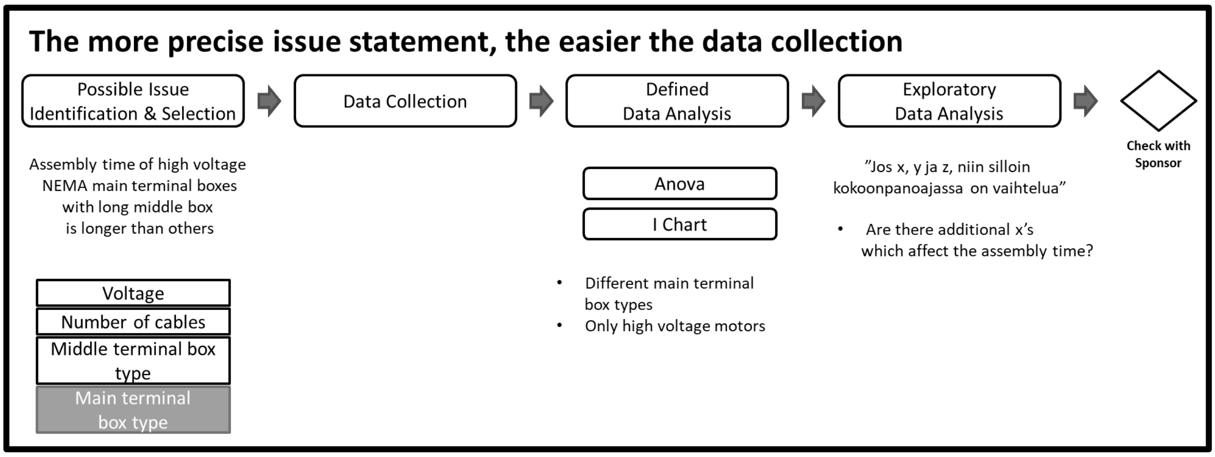
- Anova
- I Chart

- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Exploratory Data Analysis:

- High vs. low voltage
  - **Low voltage**
- Number of cables
  - **+10**
- Middle terminal box type
  - **Long**
- Main terminal box type
  - All

Figure 6. Slide 6



The more precise issue statement, the easier the data collection

Possible Issue Identification & Selection → Data Collection → Defined Data Analysis → Exploratory Data Analysis → Check with Sponsor

Assembly time of high voltage NEMA main terminal boxes with long middle box is longer than others

Voltage

Number of cables

Middle terminal box type

Main terminal box type

Anova

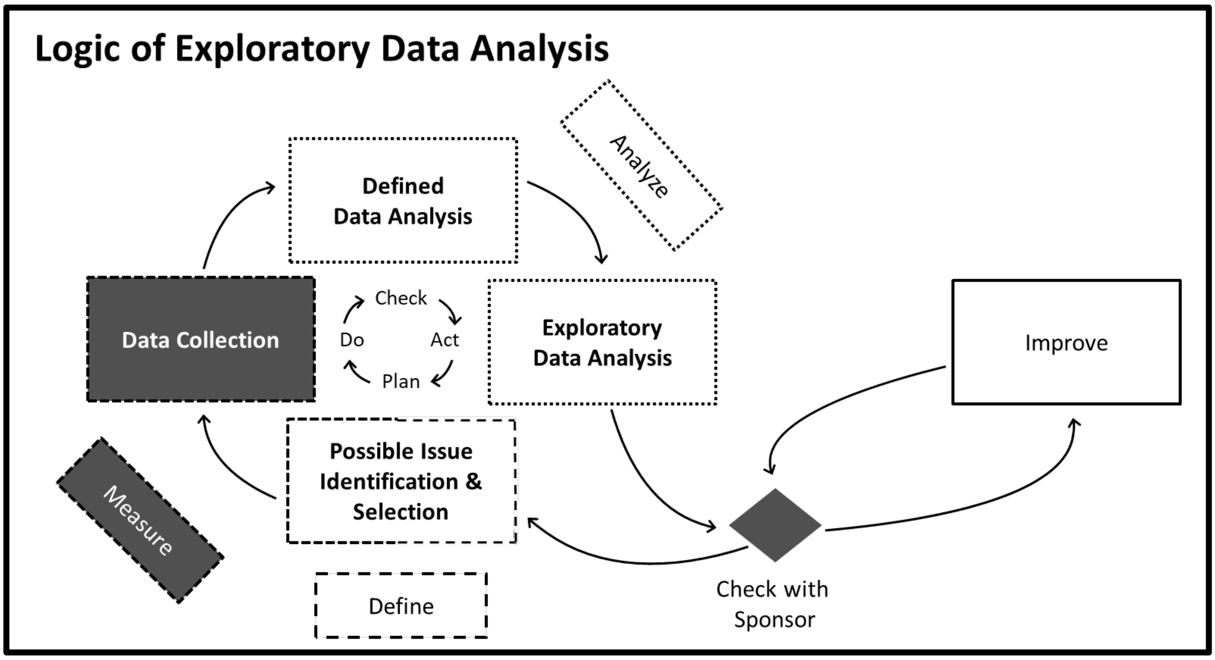
I Chart

- Different main terminal box types
- Only high voltage motors

"Jos x, y ja z, niin silloin kokoonpanoajassa on vaihtelua"

- Are there additional x's which affect the assembly time?

Figure 7. Slide 7



Logic of Exploratory Data Analysis

Defined Data Analysis

Analyze

Data Collection

Check

Do

Act

Plan

Exploratory Data Analysis

Improve

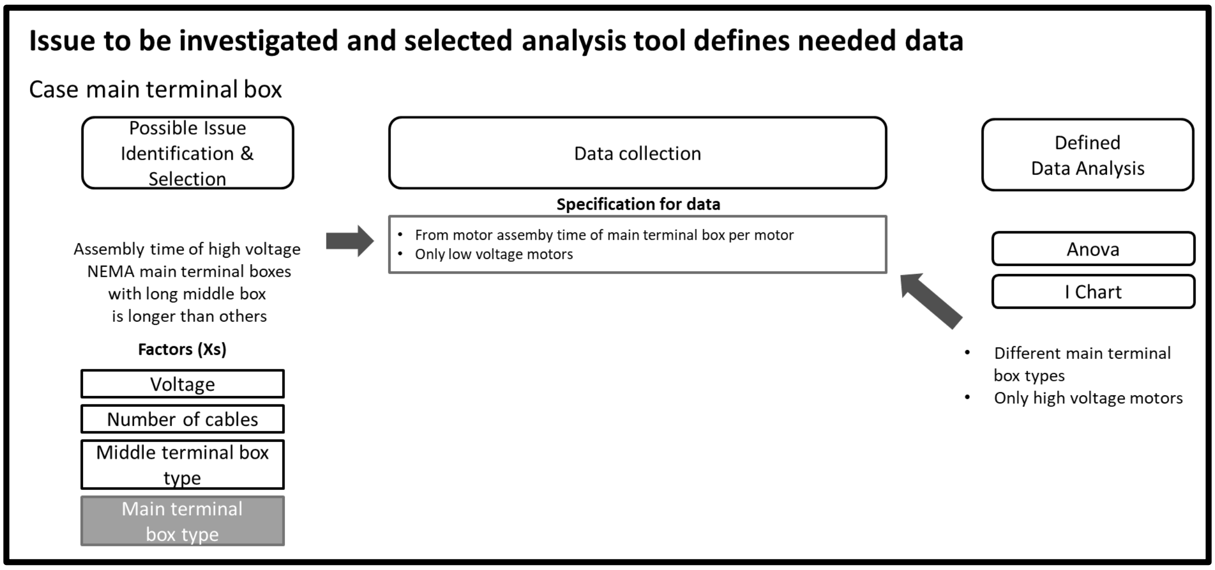
Measure

Possible Issue Identification & Selection

Define

Check with Sponsor

Figure 8. Slide 8



Issue to be investigated and selected analysis tool defines needed data

Case main terminal box

Possible Issue Identification & Selection

Assembly time of high voltage NEMA main terminal boxes with long middle box is longer than others

Factors (Xs)

Voltage

Number of cables

Middle terminal box type

Main terminal box type

Data collection

Specification for data

- From motor assemby time of main terminal box per motor
- Only low voltage motors

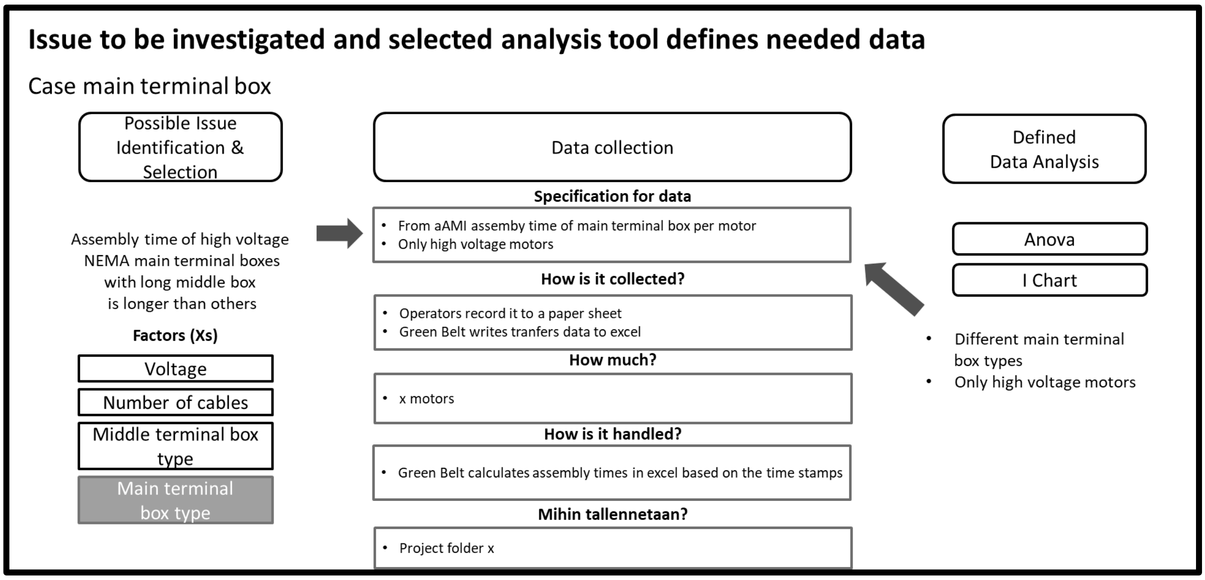
Defined Data Analysis

Anova

I Chart

- Different main terminal box types
- Only high voltage motors

Figure 9. Slide 9



Issue to be investigated and selected analysis tool defines needed data

Case main terminal box

Possible Issue Identification & Selection

Assembly time of high voltage NEMA main terminal boxes with long middle box is longer than others

Factors (Xs)

Voltage

Number of cables

Middle terminal box type

Main terminal box type

Data collection

Specification for data

- From aAMI assemby time of main terminal box per motor
- Only high voltage motors

How is it collected?

- Operators record it to a paper sheet
- Green Belt writes tranfers data to excel

How much?

- x motors

How is it handled?

- Green Belt calculates assembly times in excel based on the time stamps

Mihin tallennetaan?

- Project folder x

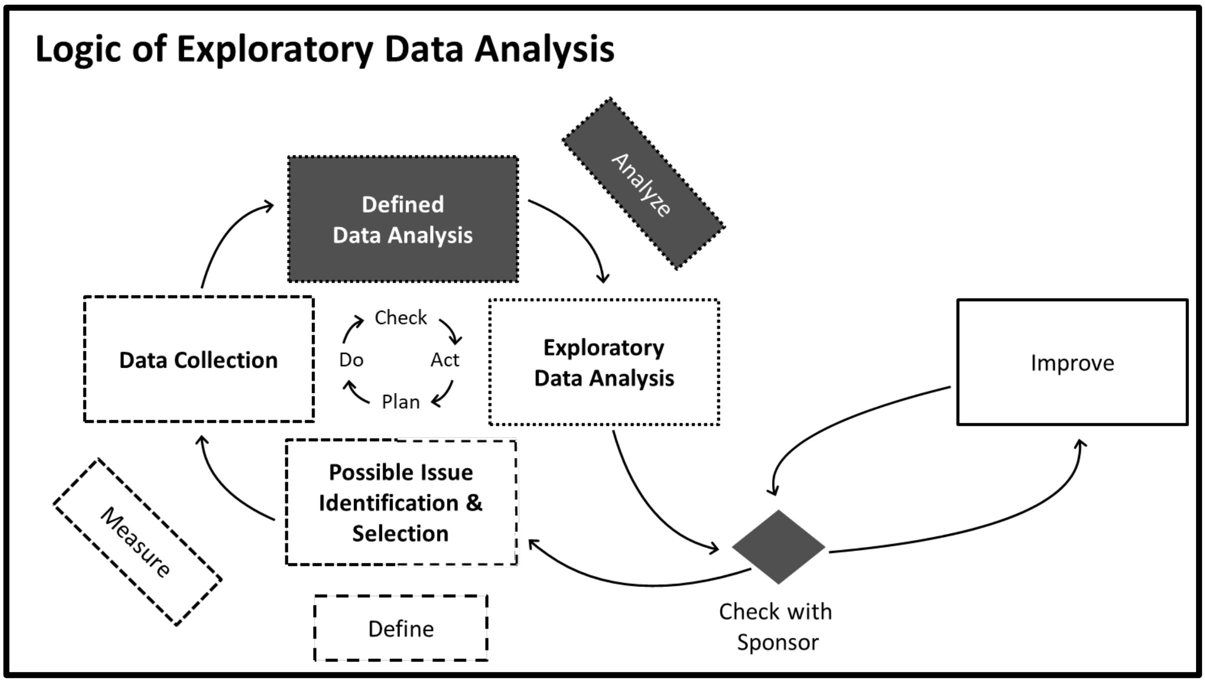
Defined Data Analysis

Anova

I Chart

- Different main terminal box types
- Only high voltage motors

Figure 10. Slide 10



Logic of Exploratory Data Analysis

Defined Data Analysis

Analyze

Data Collection

Check

Do

Act

Plan

Exploratory Data Analysis

Improve

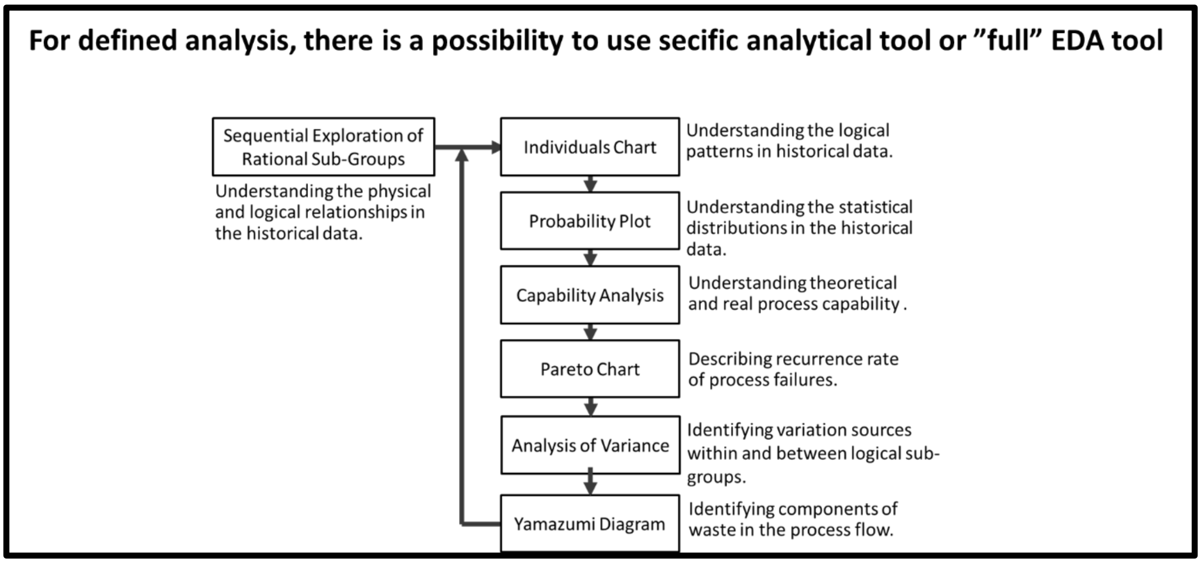
Possible Issue Identification & Selection

Measure

Check with Sponsor

Define

Figure 11. Slide 11



For defined analysis, there is a possibility to use secific analytical tool or ”full” EDA tool

Sequential Exploration of Rational Sub-Groups

Understanding the physical and logical relationships in the historical data.

Individuals Chart – Understanding the logical patterns in historical data.

Probability Plot – Understanding the statistical distributions in the historical data.

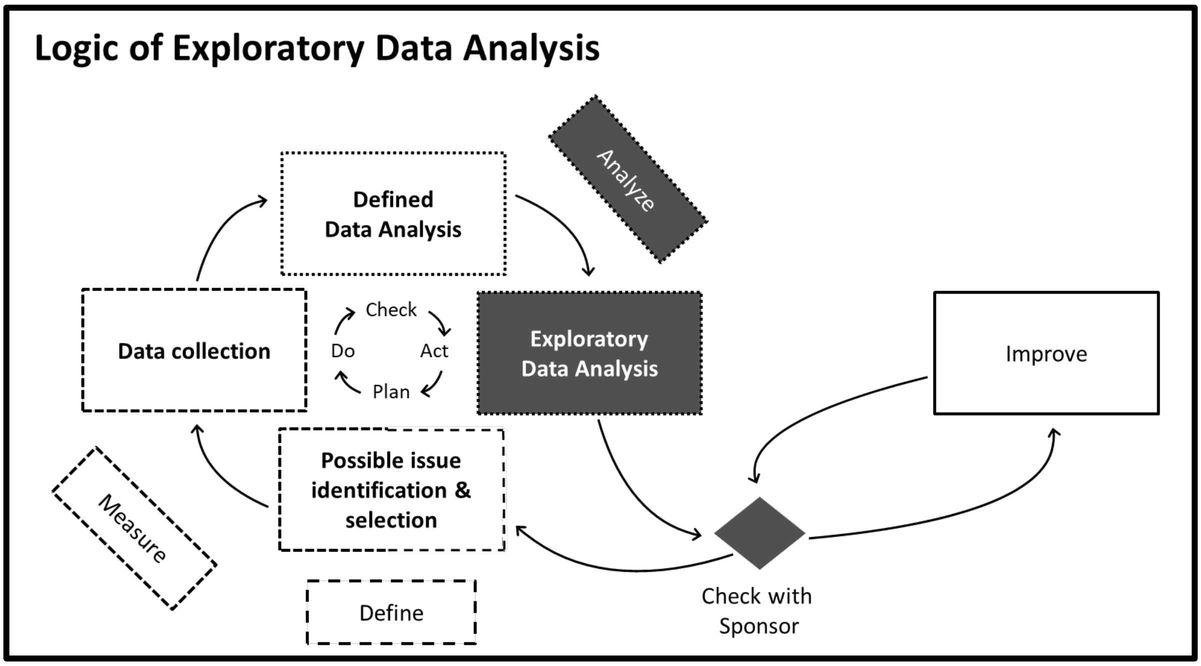
Capability Analysis – Understanding theoretical and real process capability .

Pareto Chart – Describing recurrence rate of process failures.

Analysis of Variance – Identifying variation sources within and between logical sub-groups.

Yamazumi Diagram – Identifying components of waste in the process flow.

Figure 12. Slide 12



Logic of Exploratory Data Analysis

Analyze

Defined Data Analysis

Data collection

Check

Do

Act

Plan

Exploratory Data Analysis

Improve

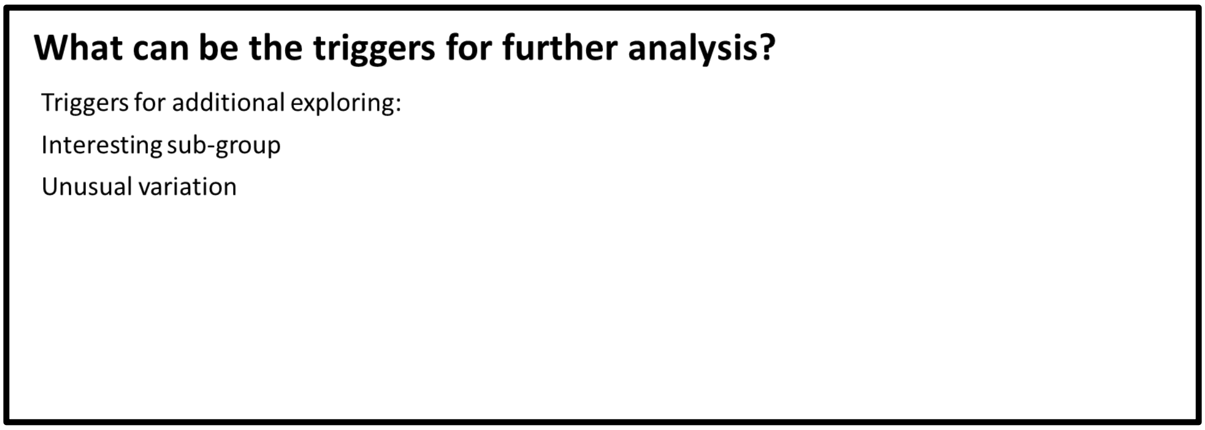
Possible issue identification & selection

Measure

Define

Check with Sponsor

Figure 13. Slide 13



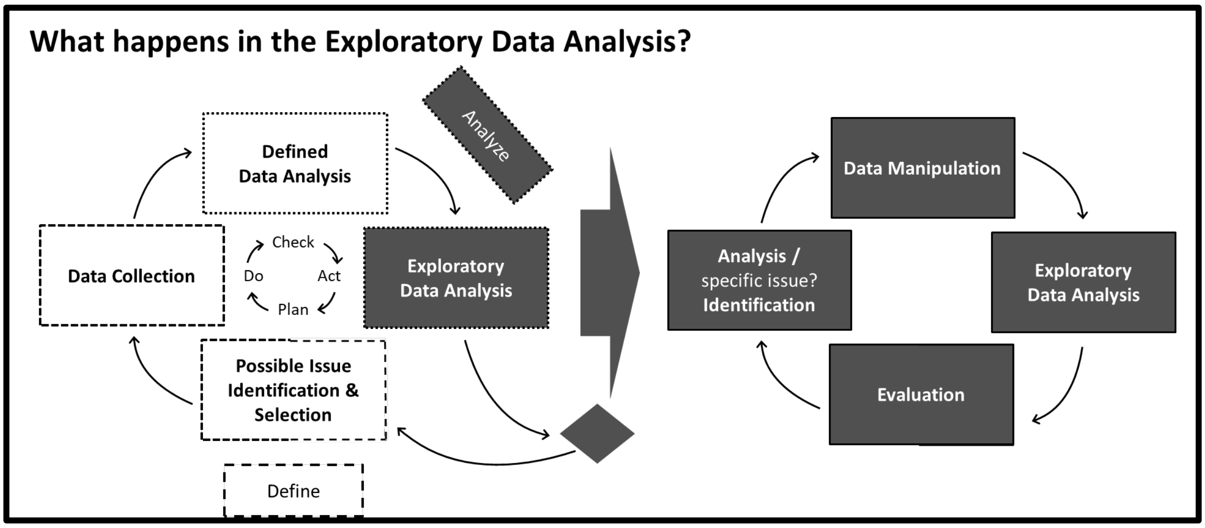
What can be the triggers for further analysis?

Triggers for additional exploring:

Interesting sub-group

Unusual variation

Figure 14. Slide 14



What happens in the Exploratory Data Analysis?

Analyze
Defined Data Analysis
Data Collection
Check
Do
Act
Plan
Exploratory Data Analysis
Possible Issue Identification & Selection
Define
Data Manipulation
Analysis / specific issue? Identification
Exploratory Data Analysis
Evaluation

Figure 15. Slide 15

What happens in the Exploratory Data Analysis?

Analysis Identification

Input:
Need for further analysis
Analysis plan
Output:

- Analysis method for exploratory analysis

Data Manipulation

Input:
Collected data & selected exploratory analysis method
Output:

- Data in the right format for analysis

Exploratory Data Analysis

Input:
Selected analysis method
Manipulated data
Output:

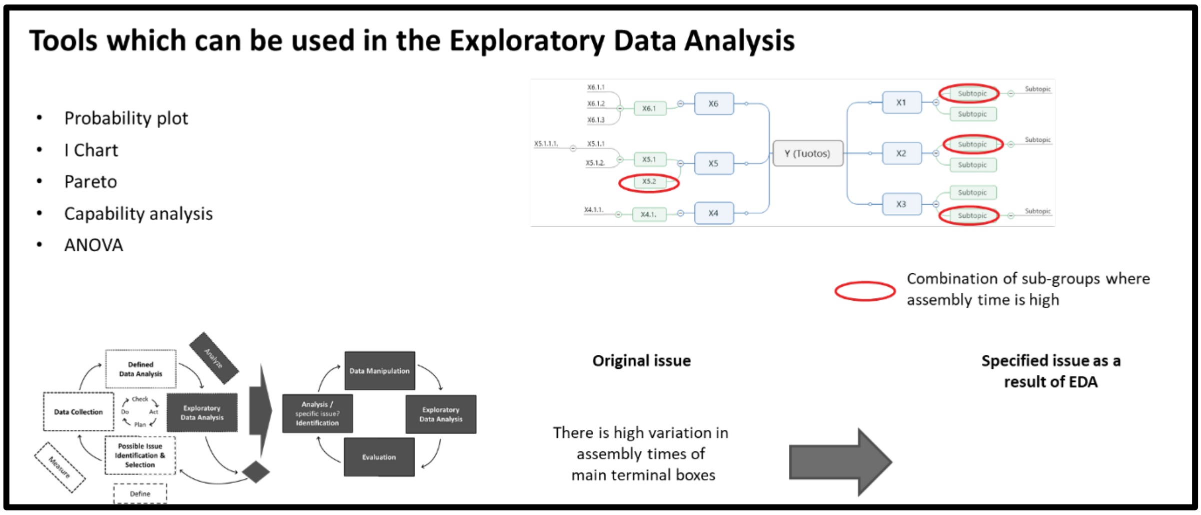
- Results of the exploratory analysis by selected method

Evaluation

Input:
Results of the exploratory analysis by selected method
Output:

- Decision if analysis provides enough information to proceed to improve phase
- Decision if there is still something additional to analyze from the dataset

Figure 16. Slide 16



Tools which can be used in the Exploratory Data Analysis

- Probability plot
- I Chart
- Pareto
- Capability analysis
- ANOVA

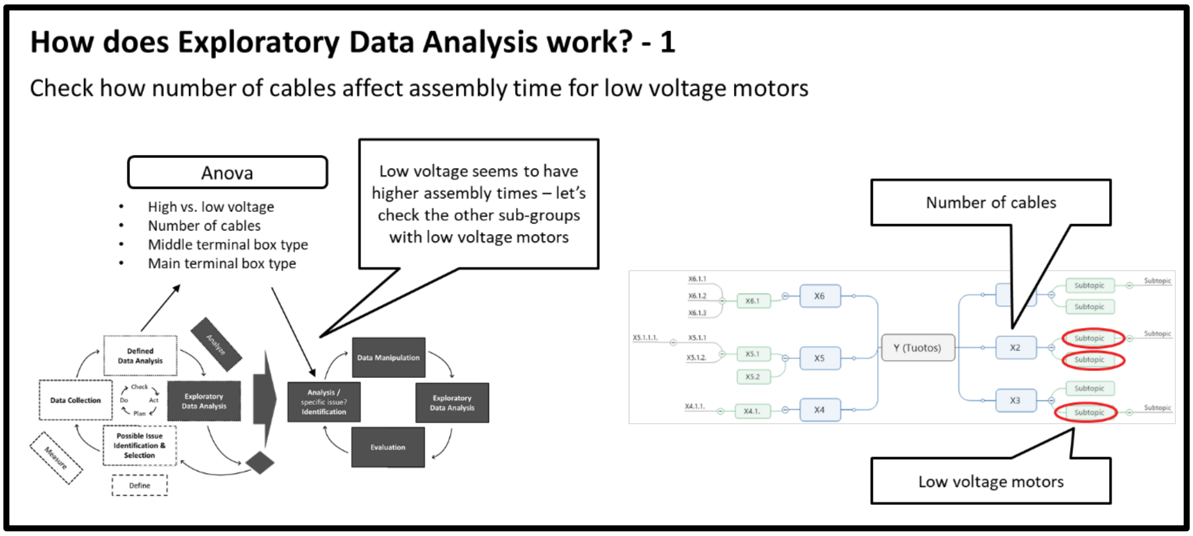
Combination of sub-groups where assembly time is high

Original issue

Specified issue as a result of EDA

There is high variation in assembly times of main terminal boxes

Figure 17. Slide 17



How does Exploratory Data Analysis work? - 1

Check how number of cables affect assembly time for low voltage motors

Anova

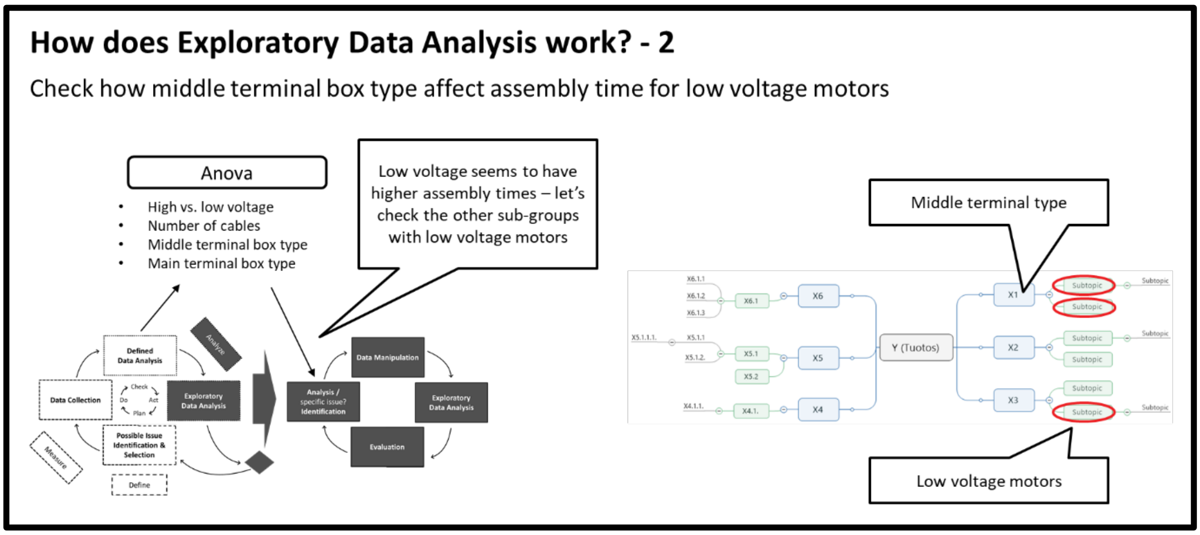
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Low voltage seems to have higher assembly times – let's check the other sub-groups with low voltage motors

Number of cables

Low voltage motors

Figure 18. Slide 18



How does Exploratory Data Analysis work? - 2

Check how middle terminal box type affect assembly time for low voltage motors

Anova

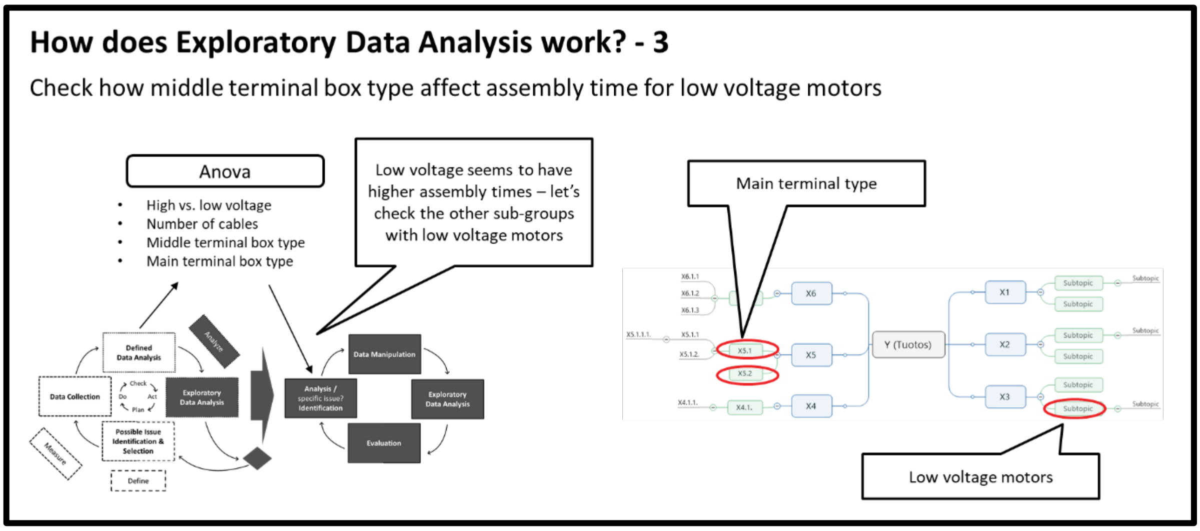
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Low voltage seems to have higher assembly times – let's check the other sub-groups with low voltage motors

Middle terminal type

Low voltage motors

Figure 19. Slide 19



How does Exploratory Data Analysis work? - 3

Check how middle terminal box type affect assembly time for low voltage motors

Anova

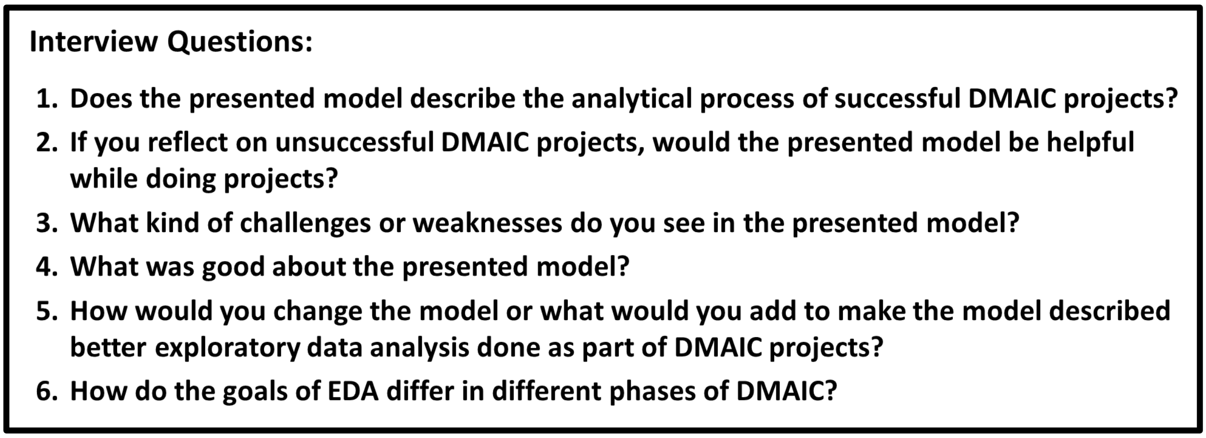
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Low voltage seems to have higher assembly times – let's check the other sub-groups with low voltage motors

Main terminal type

Low voltage motors

Figure 20. Slide 20

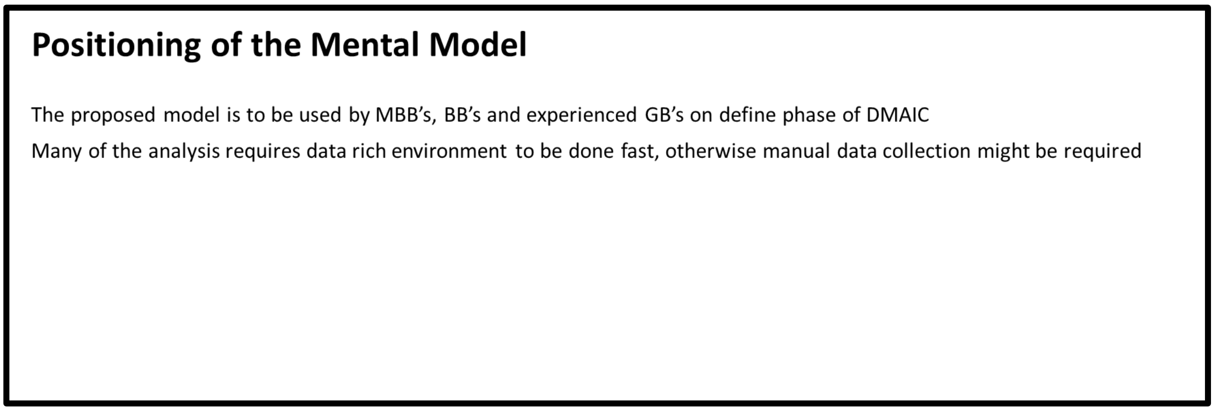


Interview Questions:

1. Does the presented model describe the analytical process of successful DMAIC projects?
2. If you reflect on unsuccessful DMAIC projects, would the presented model be helpful while doing projects?
3. What kind of challenges or weaknesses do you see in the presented model?
4. What was good about the presented model?
5. How would you change the model or what would you add to make the model described better exploratory data analysis done as part of DMAIC projects?
6. How do the goals of EDA differ in different phases of DMAIC?

Figure 21. Slide 21

APPENDIX B: PRESENTATION OF THE UPDATED VERSION OF THE MENTAL MODEL

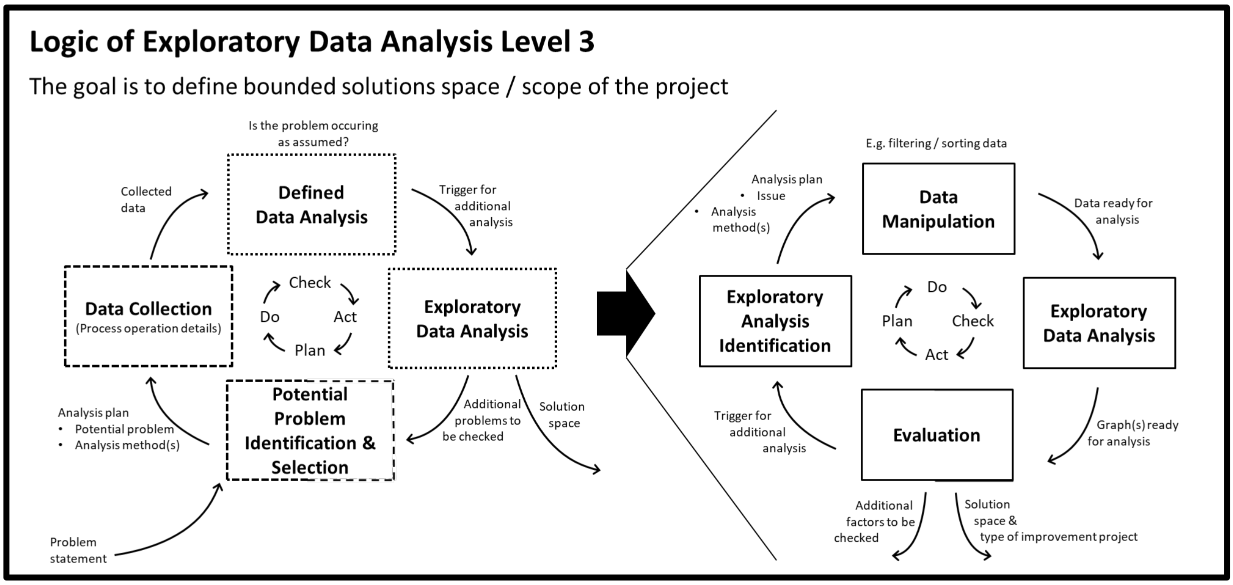


Positioning of the Mental Model

The proposed model is to be used by MBB's, BB's and experienced GB's on define phase of DMAIC

Many of the analysis requires data rich environment to be done fast, otherwise manual data collection might be required

Figure 1. Slide 1



Logic of Exploratory Data Analysis Level 3

The goal is to define bounded solutions space / scope of the project

Is the problem occuring as assumed?

Defined Data Analysis

Collected data

Trigger for additional analysis

Data Collection (Process operation details)

Check

Do

Act

Plan

Exploratory Data Analysis

Analysis plan
- Potential problem
- Analysis method(s)

Potential Problem Identification & Selection

Additional problems to be checked

Solution space

Problem statement

E.g. filtering / sorting data

Analysis plan
- Issue
- Analysis method(s)

Data Manipulation

Data ready for analysis

Exploratory Analysis Identification

Do

Plan

Check

Act

Exploratory Data Analysis

Trigger for additional analysis

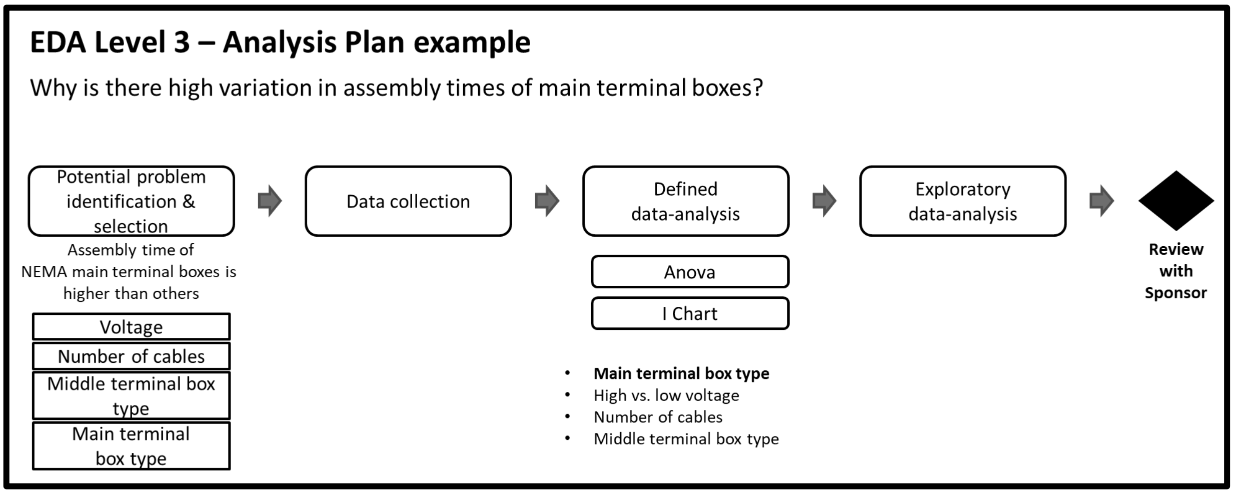
Evaluation

Graph(s) ready for analysis

Additional factors to be checked

Solution space & type of improvement project

Figure 2. Slide 2



EDA Level 3 – Analysis Plan example

Why is there high variation in assembly times of main terminal boxes?

Potential problem identification & selection → Data collection → Defined data-analysis → Exploratory data-analysis → Review with Sponsor

Assembly time of NEMA main terminal boxes is higher than others

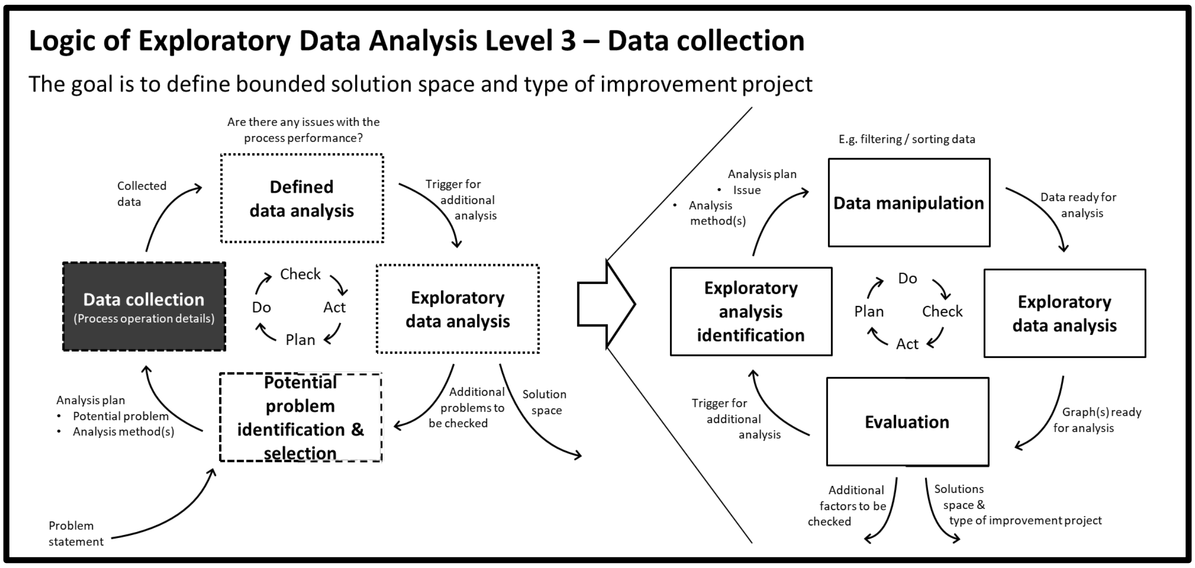
- Voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Anova

I Chart

- **Main terminal box type**
- High vs. low voltage
- Number of cables
- Middle terminal box type

Figure 3. Slide 3

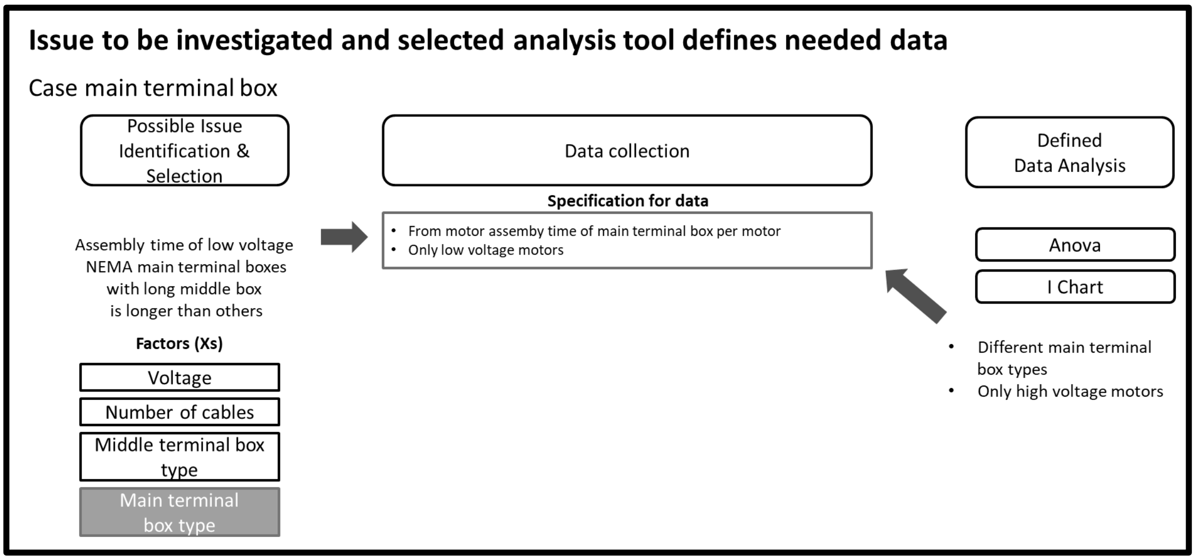


Logic of Exploratory Data Analysis Level 3 – Data collection

The goal is to define bounded solution space and type of improvement project

Are there any issues with the process performance?
Defined data analysis
Collected data
Trigger for additional analysis
Data collection (Process operation details)
Check
Do
Act
Plan
Exploratory data analysis
Analysis plan
• Potential problem
• Analysis method(s)
Potential problem identification & selection
Additional problems to be checked
Solution space
Problem statement
E.g. filtering / sorting data
Analysis plan
• Issue
• Analysis method(s)
Data manipulation
Data ready for analysis
Exploratory analysis identification
Do
Plan
Check
Act
Exploratory data analysis
Trigger for additional analysis
Evaluation
Graph(s) ready for analysis
Additional factors to be checked
Solutions space & type of improvement project

Figure 4. Slide 4



Issue to be investigated and selected analysis tool defines needed data

Case main terminal box

Possible Issue Identification & Selection

Assembly time of low voltage NEMA main terminal boxes with long middle box is longer than others

Factors (Xs)

- Voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Data collection

Specification for data

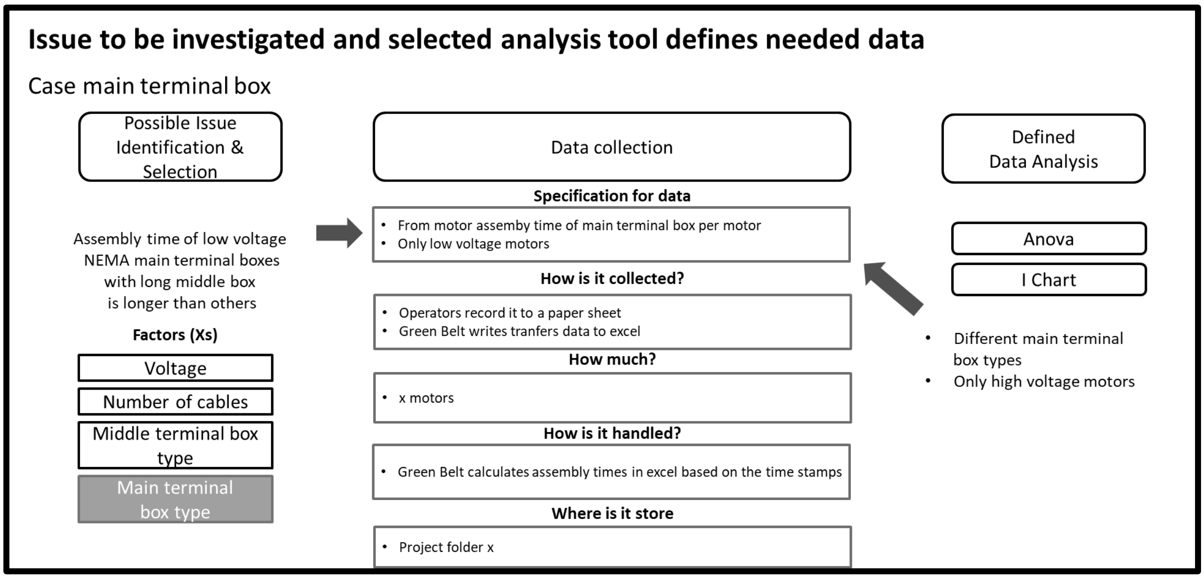
- From motor assemby time of main terminal box per motor
- Only low voltage motors

Defined Data Analysis

- Anova
- I Chart

- Different main terminal box types
- Only high voltage motors

Figure 5. Slide 5



Issue to be investigated and selected analysis tool defines needed data

Case main terminal box

Possible Issue Identification & Selection

Assembly time of low voltage NEMA main terminal boxes with long middle box is longer than others

Factors (Xs)

- Voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Data collection

Specification for data

- From motor assembly time of main terminal box per motor
- Only low voltage motors

How is it collected?

- Operators record it to a paper sheet
- Green Belt writes tranfers data to excel

How much?

- x motors

How is it handled?

- Green Belt calculates assembly times in excel based on the time stamps

Where is it store

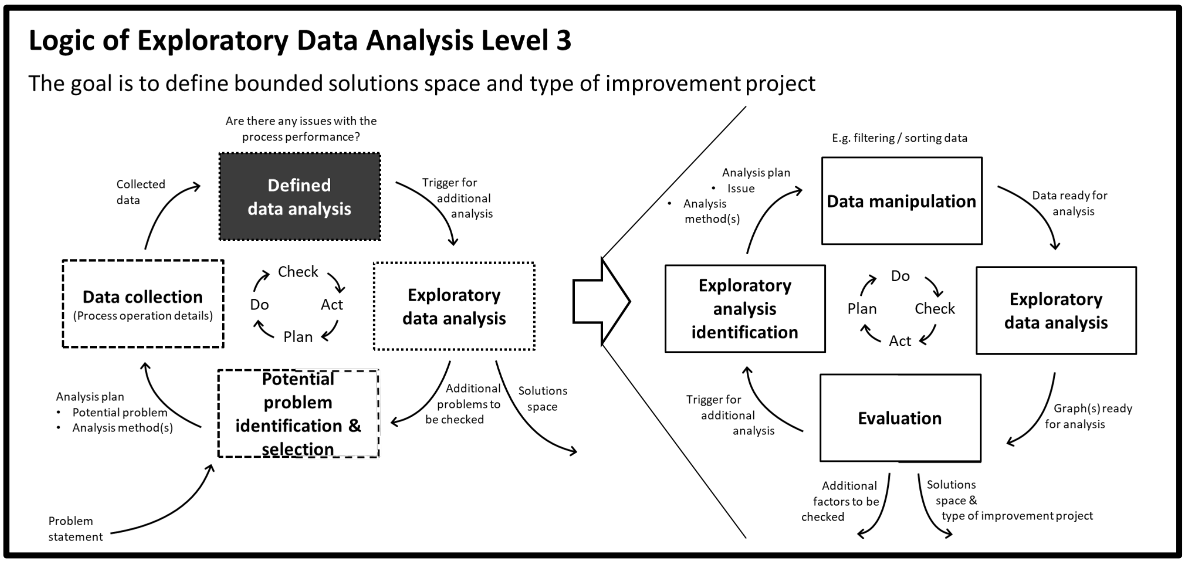
- Project folder x

Defined Data Analysis

- Anova
- I Chart

- Different main terminal box types
- Only high voltage motors

Figure 6. Slide 6



Logic of Exploratory Data Analysis Level 3

The goal is to define bounded solutions space and type of improvement project

Are there any issues with the process performance?

Defined data analysis

Collected data

Trigger for additional analysis

Data collection (Process operation details)

Check

Do

Act

Plan

Exploratory data analysis

Analysis plan
- Potential problem
- Analysis method(s)

Potential problem identification & selection

Additional problems to be checked

Solutions space

Problem statement

E.g. filtering / sorting data

Analysis plan
- Issue
- Analysis method(s)

Data manipulation

Data ready for analysis

Exploratory analysis identification

Do

Plan

Check

Act

Exploratory data analysis

Trigger for additional analysis

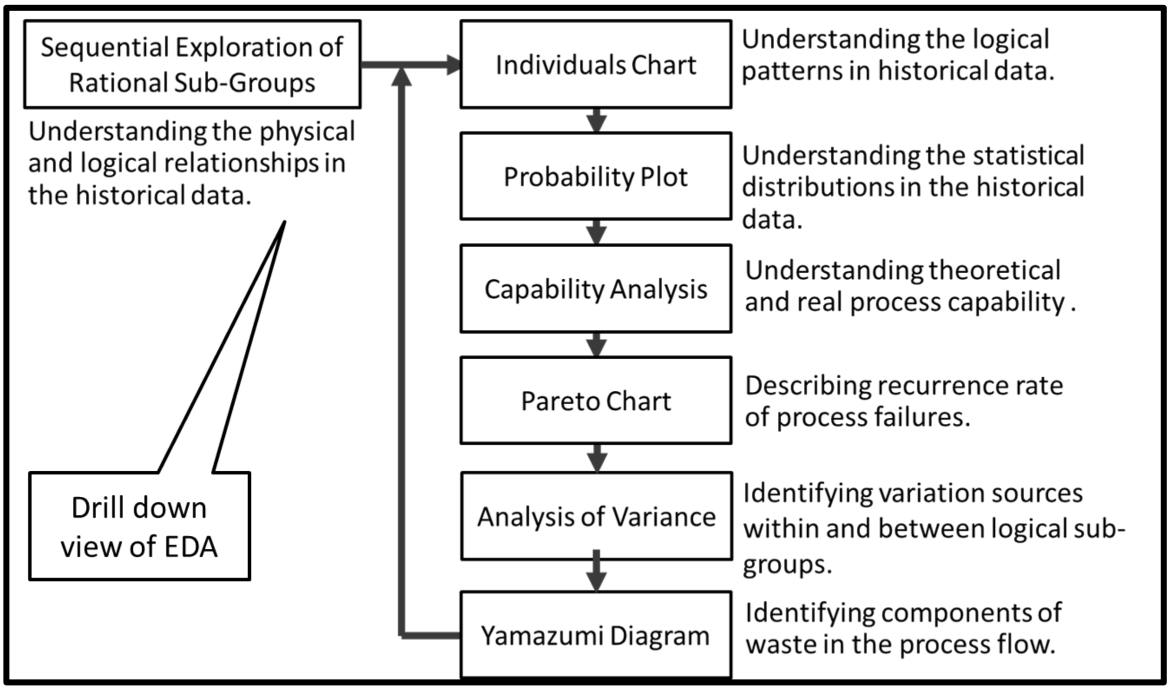
Evaluation

Graph(s) ready for analysis

Additional factors to be checked

Solutions space & type of improvement project

Figure 7. Slide 7



Sequential Exploration of Rational Sub-Groups

Understanding the physical and logical relationships in the historical data.

Drill down view of EDA

Individuals Chart — Understanding the logical patterns in historical data.

Probability Plot — Understanding the statistical distributions in the historical data.

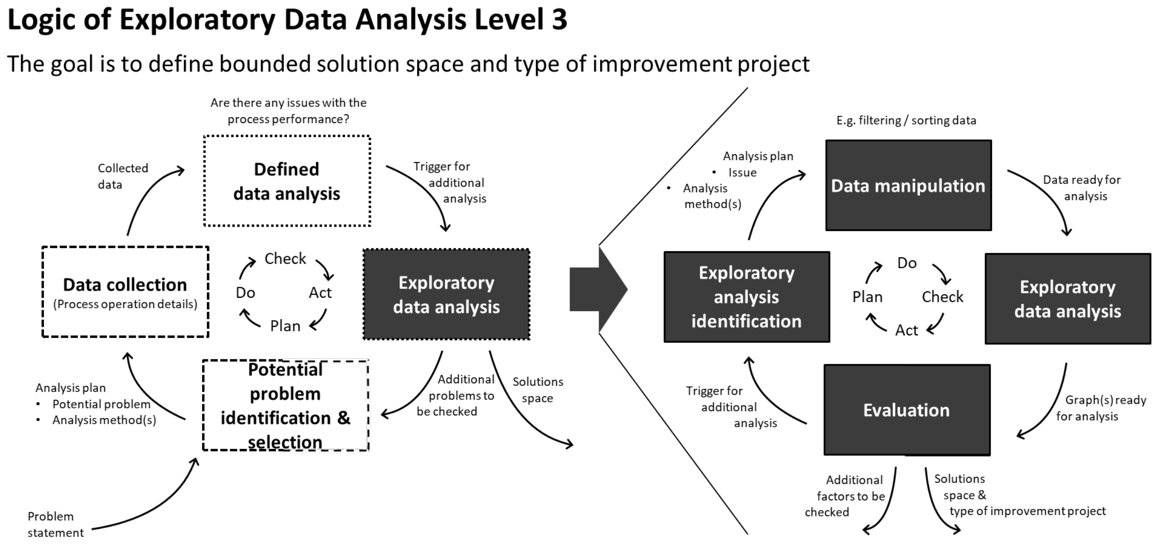
Capability Analysis — Understanding theoretical and real process capability .

Pareto Chart — Describing recurrence rate of process failures.

Analysis of Variance — Identifying variation sources within and between logical sub-groups.

Yamazumi Diagram — Identifying components of waste in the process flow.

Figure 8. Slide 8



Logic of Exploratory Data Analysis Level 3

The goal is to define bounded solution space and type of improvement project

Are there any issues with the process performance?

Defined data analysis

Collected data

Trigger for additional analysis

Data collection (Process operation details)

Check

Do

Act

Plan

Exploratory data analysis

Analysis plan
- Potential problem
- Analysis method(s)

Potential problem identification & selection

Additional problems to be checked

Solutions space

Problem statement

E.g. filtering / sorting data

Analysis plan
- Issue
- Analysis method(s)

Data manipulation

Data ready for analysis

Exploratory analysis identification

Do

Plan

Check

Act

Exploratory data analysis

Trigger for additional analysis

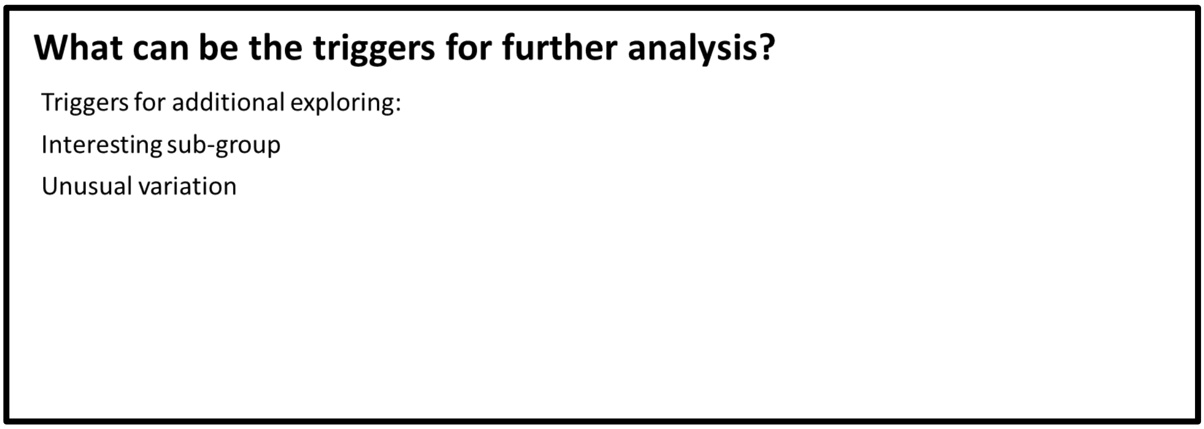
Evaluation

Graph(s) ready for analysis

Additional factors to be checked

Solutions space & type of improvement project

Figure 9. Slide 9



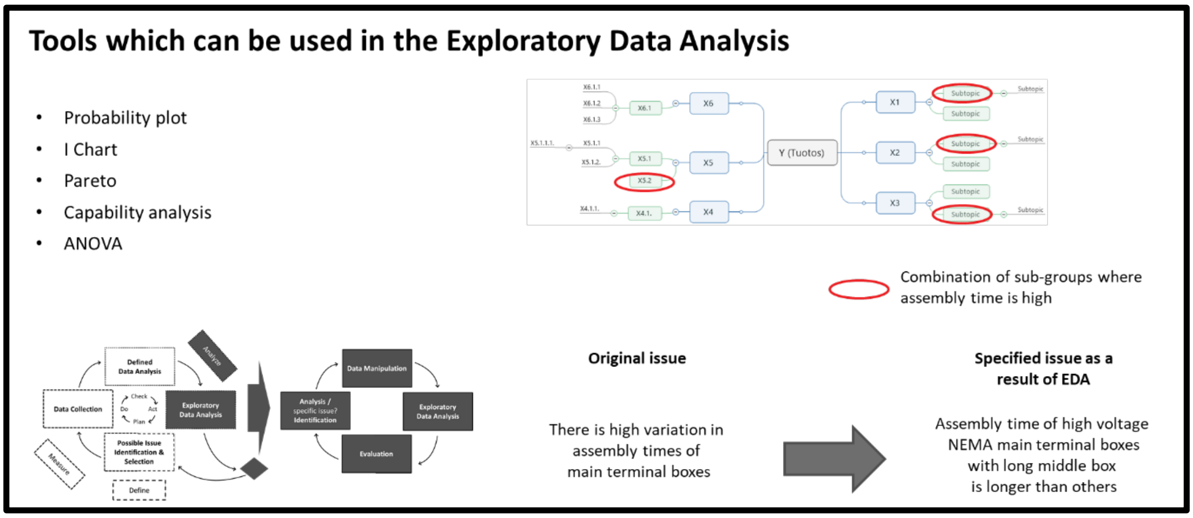
What can be the triggers for further analysis?

Triggers for additional exploring:

Interesting sub-group

Unusual variation

Figure 10. Slide 10



Tools which can be used in the Exploratory Data Analysis

- Probability plot
- I Chart
- Pareto
- Capability analysis
- ANOVA

Combination of sub-groups where assembly time is high

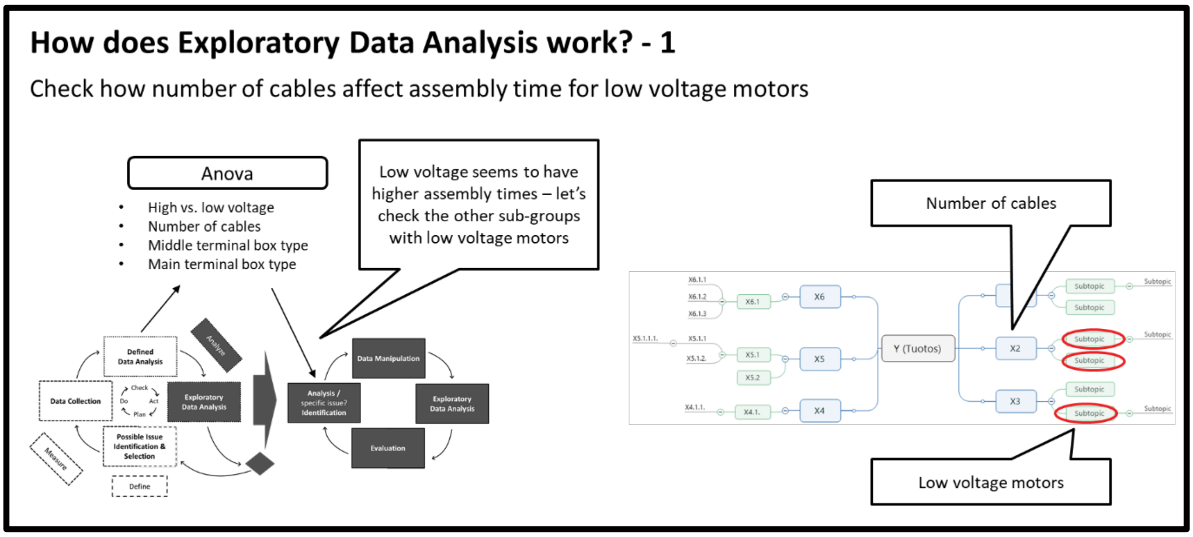
Original issue

There is high variation in assembly times of main terminal boxes

Specified issue as a result of EDA

Assembly time of high voltage NEMA main terminal boxes with long middle box is longer than others

Figure 11. Slide 11



How does Exploratory Data Analysis work? - 1

Check how number of cables affect assembly time for low voltage motors

Anova

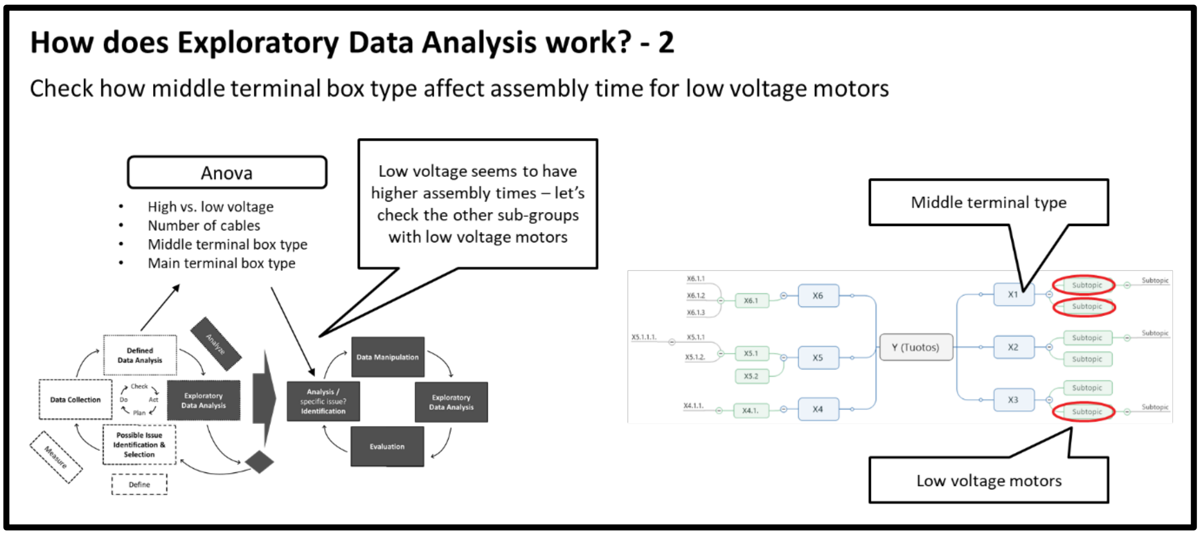
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Low voltage seems to have higher assembly times – let's check the other sub-groups with low voltage motors

Number of cables

Low voltage motors

Figure 12. Slide 12



How does Exploratory Data Analysis work? - 2

Check how middle terminal box type affect assembly time for low voltage motors

Anova

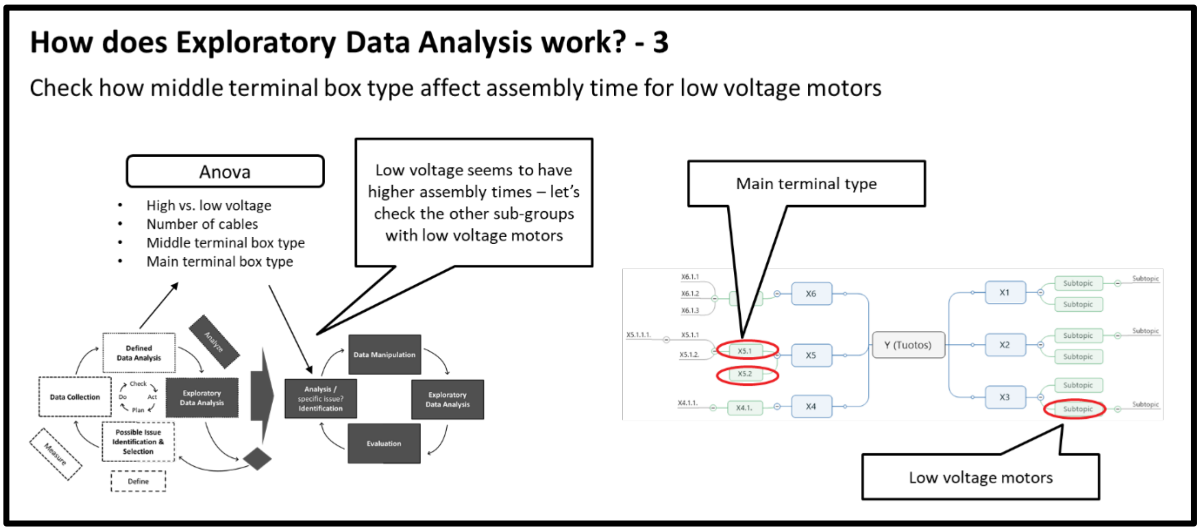
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Low voltage seems to have higher assembly times – let's check the other sub-groups with low voltage motors

Middle terminal type

Low voltage motors

Figure 13. Slide 13



How does Exploratory Data Analysis work? - 3

Check how middle terminal box type affect assembly time for low voltage motors

Anova

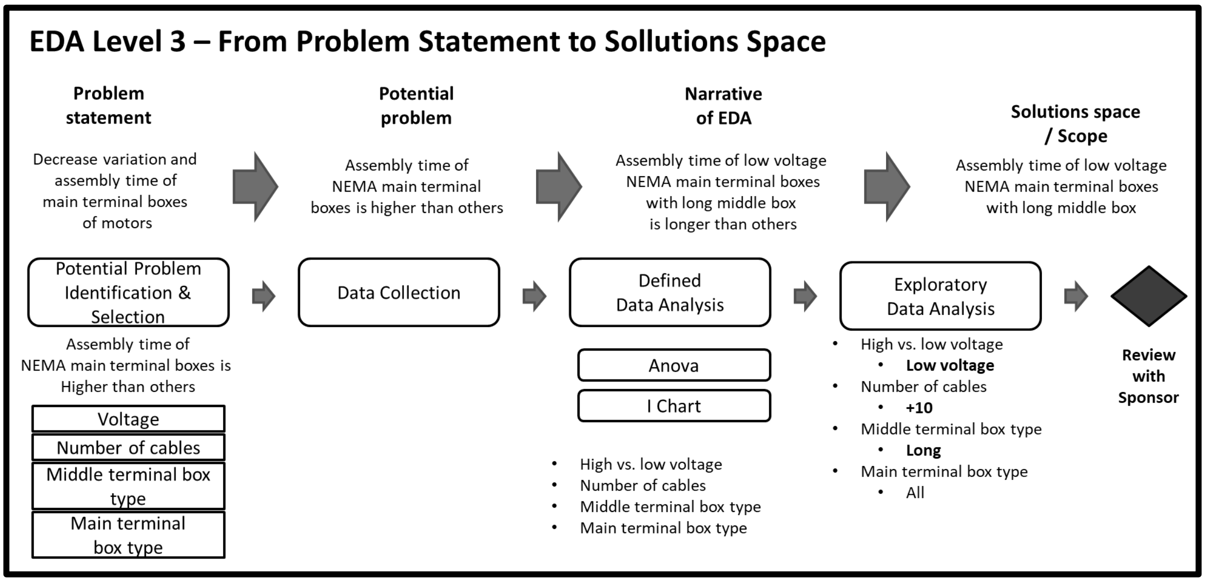
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Low voltage seems to have higher assembly times – let's check the other sub-groups with low voltage motors

Main terminal type

Low voltage motors

Figure 14. Slide 14



EDA Level 3 – From Problem Statement to Sollutions Space

Problem statement

Decrease variation and assembly time of main terminal boxes of motors

Potential problem

Assembly time of NEMA main terminal boxes is higher than others

Narrative of EDA

Assembly time of low voltage NEMA main terminal boxes with long middle box is longer than others

Solutions space / Scope

Assembly time of low voltage NEMA main terminal boxes with long middle box

Potential Problem Identification & Selection

Assembly time of NEMA main terminal boxes is Higher than others

Voltage

Number of cables

Middle terminal box type

Main terminal box type

Data Collection

Defined Data Analysis

Anova

I Chart

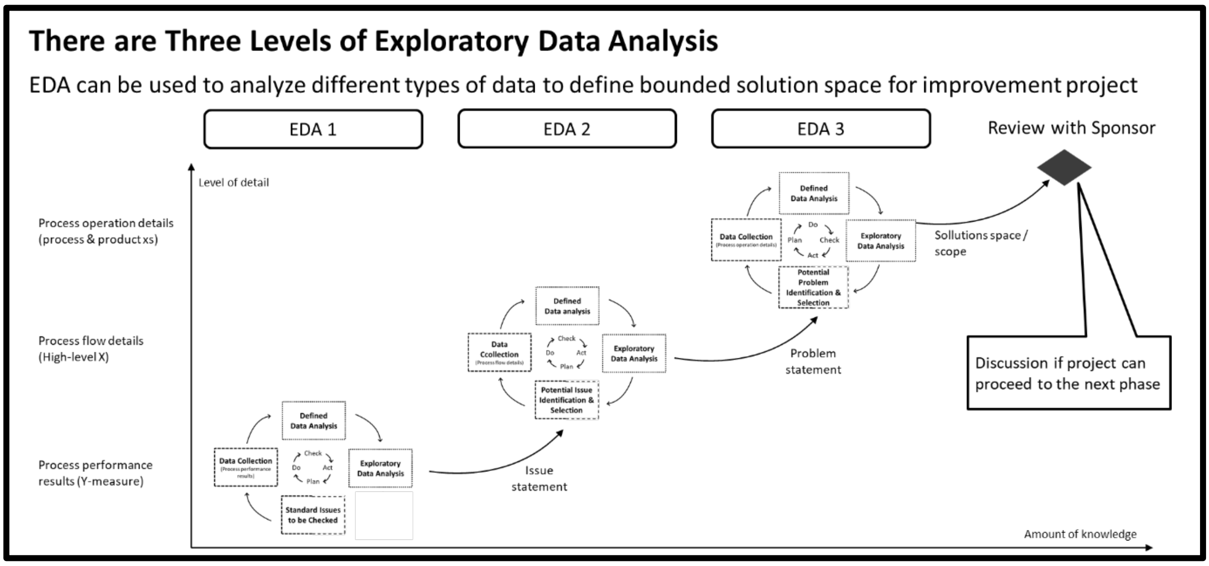
- High vs. low voltage
- Number of cables
- Middle terminal box type
- Main terminal box type

Exploratory Data Analysis

- High vs. low voltage
  - **Low voltage**
- Number of cables
  - **+10**
- Middle terminal box type
  - **Long**
- Main terminal box type
  - All

Review with Sponsor

Figure 15. Slide 15



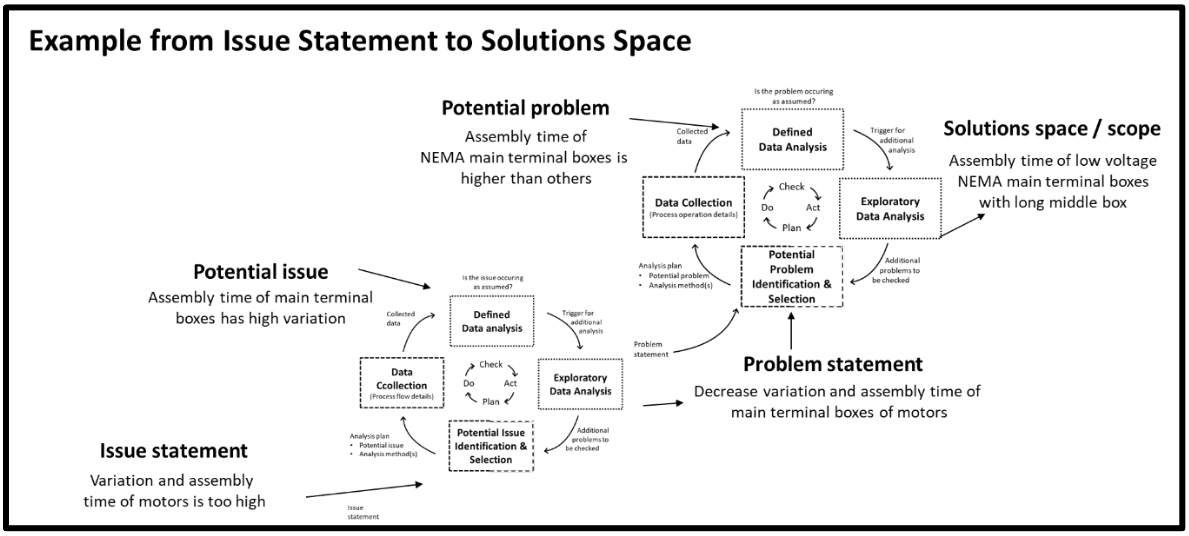
There are Three Levels of Exploratory Data Analysis

EDA can be used to analyze different types of data to define bounded solution space for improvement project

EDA 1
EDA 2
EDA 3
Review with Sponsor
Level of detail
Process operation details (process & product xs)
Process flow details (High-level X)
Process performance results (Y-measure)
Defined Data Analysis
Data Collection (Process performance results)
Check
Do
Act
Plan
Exploratory Data Analysis
Standard Issues to be Checked
Issue statement
Defined Data analysis
Data Ccollection (Process flow details)
Potential Issue Identification & Selection
Problem statement
Defined Data Analysis
Data Collection (Process operation details)
Potential Problem Identification & Selection
Sollutions space / scope
Amount of knowledge

Discussion if project can proceed to the next phase

Figure 16. Slide 16



Example from Issue Statement to Solutions Space

Potential problem

Assembly time of NEMA main terminal boxes is higher than others

Solutions space / scope

Assembly time of low voltage NEMA main terminal boxes with long middle box

Potential issue

Assembly time of main terminal boxes has high variation

Problem statement

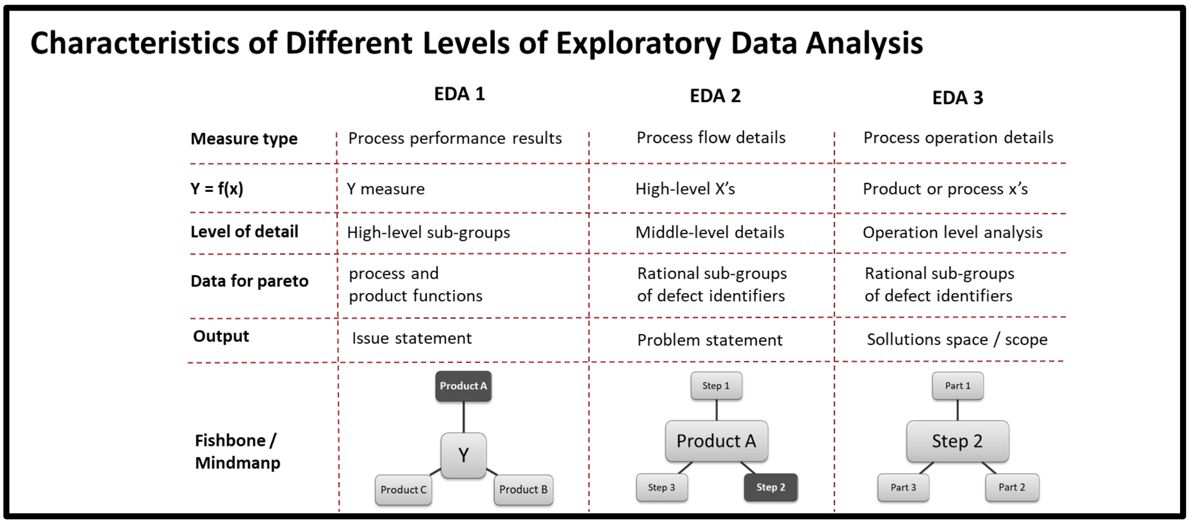
Decrease variation and assembly time of main terminal boxes of motors

Issue statement

Variation and assembly time of motors is too high

Is the problem occuring as assumed?
Collected data
Defined Data Analysis
Trigger for additional analysis
Data Collection (Process operation details)
Check
Do
Act
Plan
Exploratory Data Analysis
Analysis plan
• Potential problem
• Analysis method(s)
Potential Problem Identification & Selection
Additional problems to be checked
Problem statement
Is the issue occuring as assumed?
Collected data
Defined Data analysis
Trigger for additional analysis
Data Ccollection (Process flow details)
Check
Do
Act
Plan
Exploratory Data Analysis
Analysis plan
• Potential issue
• Analysis method(s)
Potential Issue Identification & Selection
Additional problems to be checked
Issue statement

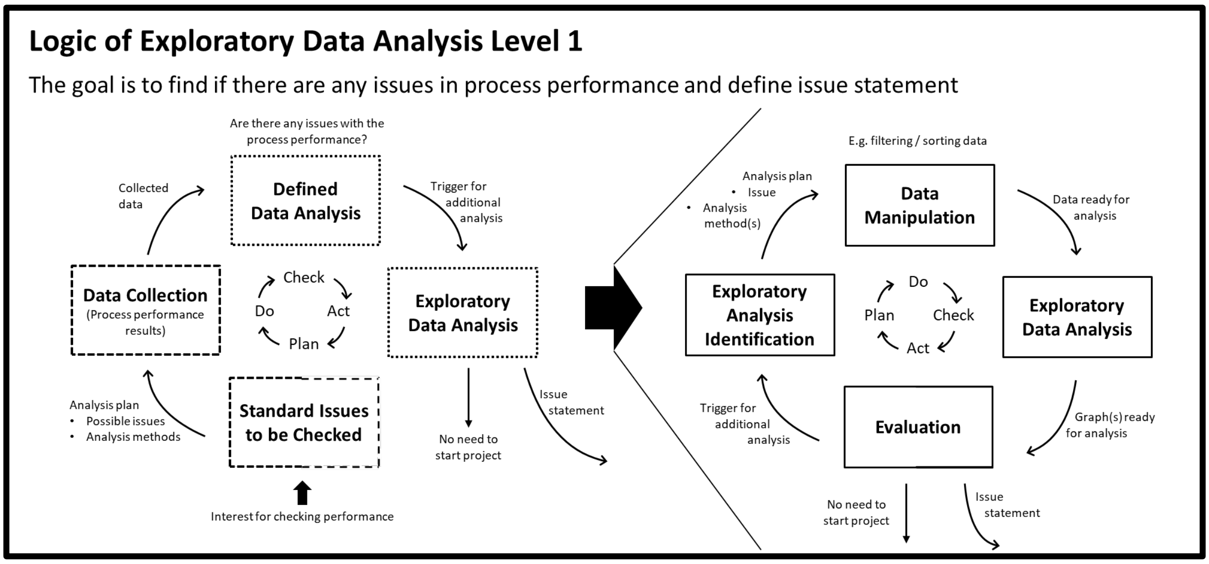
Figure 17. Slide 17



Characteristics of Different Levels of Exploratory Data Analysis

	EDA 1	EDA 2	EDA 3
Measure type	Process performance results	Process flow details	Process operation details
Y = f(x)	Y measure	High-level X's	Product or process x's
Level of detail	High-level sub-groups	Middle-level details	Operation level analysis
Data for pareto	process and product functions	Rational sub-groups of defect identifiers	Rational sub-groups of defect identifiers
Output	Issue statement	Problem statement	Sollutions space / scope
Fishbone / Mindmanp	Product A Y Product C Product B	Step 1 Product A Step 3 Step 2	Part 1 Step 2 Part 3 Part 2

Figure 18. Slide 18



Logic of Exploratory Data Analysis Level 1

The goal is to find if there are any issues in process performance and define issue statement

Are there any issues with the process performance?

Defined Data Analysis

Trigger for additional analysis

Collected data

Data Collection (Process performance results)

Check Do Act Plan

Exploratory Data Analysis

Analysis plan
- Possible issues
- Analysis methods

Standard Issues to be Checked

No need to start project

Issue statement

Interest for checking performance

E.g. filtering / sorting data

Analysis plan
- Issue
- Analysis method(s)

Data Manipulation

Data ready for analysis

Exploratory Analysis Identification

Do Plan Check Act

Exploratory Data Analysis

Trigger for additional analysis

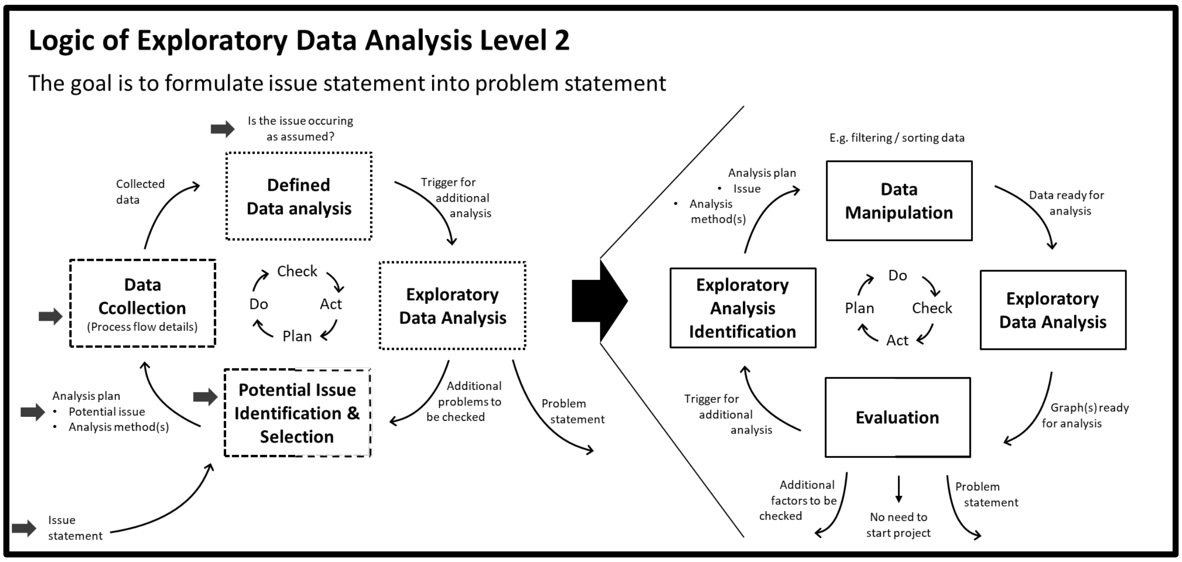
Evaluation

Graph(s) ready for analysis

No need to start project

Issue statement

Figure 19. Slide 19



Logic of Exploratory Data Analysis Level 2

The goal is to formulate issue statement into problem statement

Is the issue occuring as assumed?

Defined Data analysis

Trigger for additional analysis

Collected data

Data Ccollection (Process flow details)

Check Do Act Plan

Exploratory Data Analysis

Analysis plan
- Potential issue
- Analysis method(s)

Potential Issue Identification & Selection

Additional problems to be checked

Problem statement

Issue statement

E.g. filtering / sorting data

Analysis plan
- Issue
- Analysis method(s)

Data Manipulation

Data ready for analysis

Exploratory Analysis Identification

Do Plan Check Act

Exploratory Data Analysis

Trigger for additional analysis

Evaluation

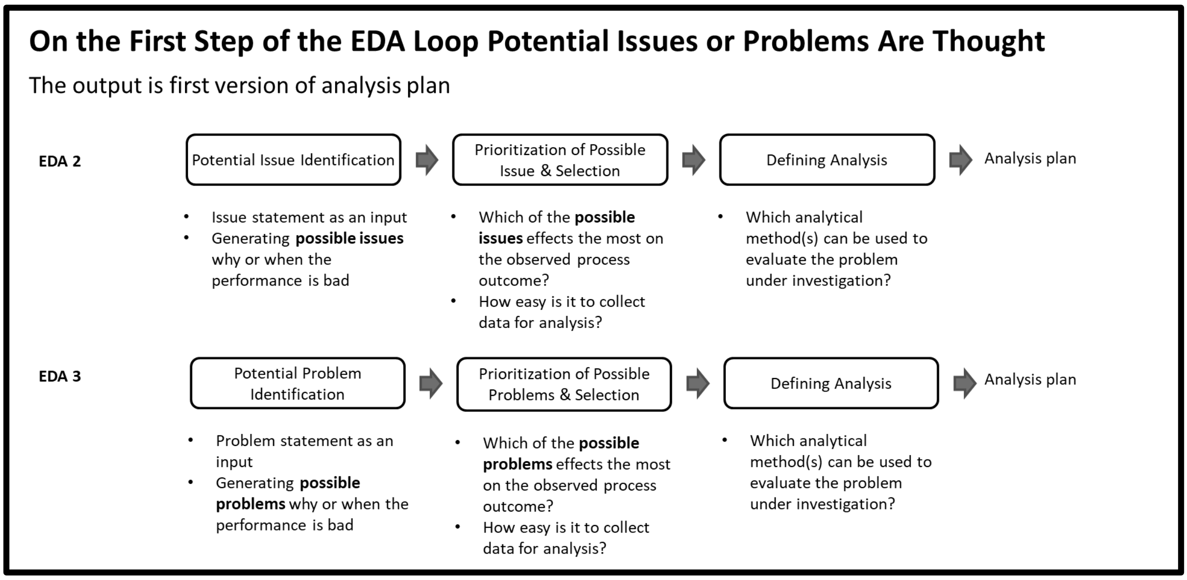
Graph(s) ready for analysis

Additional factors to be checked

No need to start project

Problem statement

Figure 20. Slide 20



On the First Step of the EDA Loop Potential Issues or Problems Are Thought

The output is first version of analysis plan

EDA 2

Potential Issue Identification → Prioritization of Possible Issue & Selection → Defining Analysis → Analysis plan

- Issue statement as an input
- Generating **possible issues** why or when the performance is bad

- Which of the **possible issues** effects the most on the observed process outcome?
- How easy is it to collect data for analysis?

- Which analytical method(s) can be used to evaluate the problem under investigation?

EDA 3

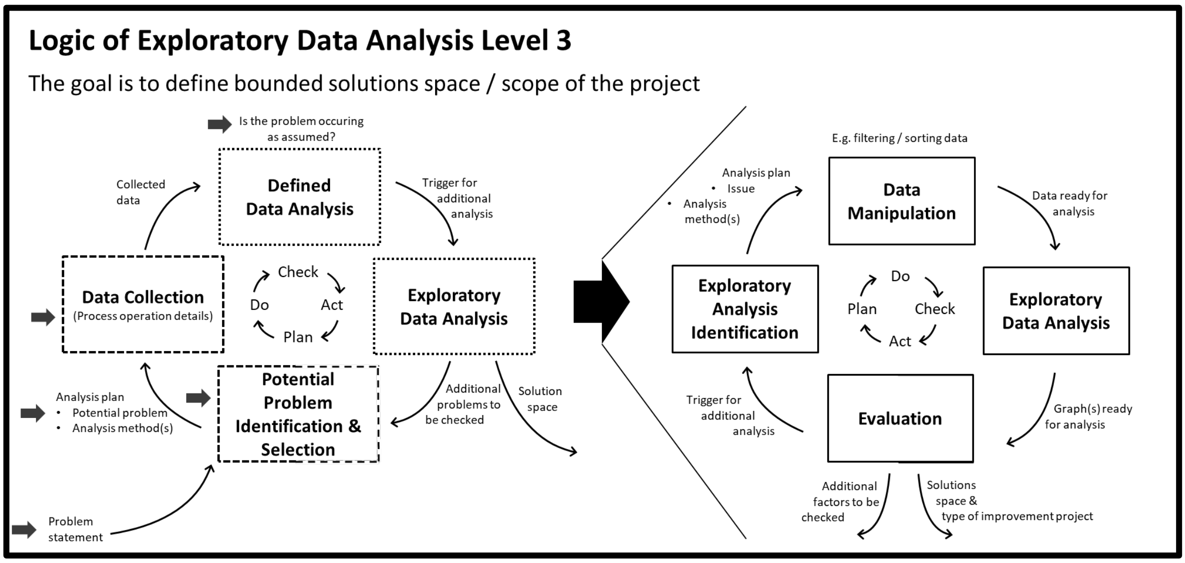
Potential Problem Identification → Prioritization of Possible Problems & Selection → Defining Analysis → Analysis plan

- Problem statement as an input
- Generating **possible problems** why or when the performance is bad

- Which of the **possible problems** effects the most on the observed process outcome?
- How easy is it to collect data for analysis?

- Which analytical method(s) can be used to evaluate the problem under investigation?

Figure 21. Slide 21



Logic of Exploratory Data Analysis Level 3

The goal is to define bounded solutions space / scope of the project

Is the problem occuring as assumed?

Defined Data Analysis

Collected data

Trigger for additional analysis

Data Collection (Process operation details)

Check

Do

Act

Plan

Exploratory Data Analysis

Analysis plan
- Potential problem
- Analysis method(s)

Potential Problem Identification & Selection

Additional problems to be checked

Solution space

Problem statement

E.g. filtering / sorting data

Analysis plan
- Issue
- Analysis method(s)

Data Manipulation

Data ready for analysis

Exploratory Analysis Identification

Do

Plan

Check

Act

Exploratory Data Analysis

Trigger for additional analysis

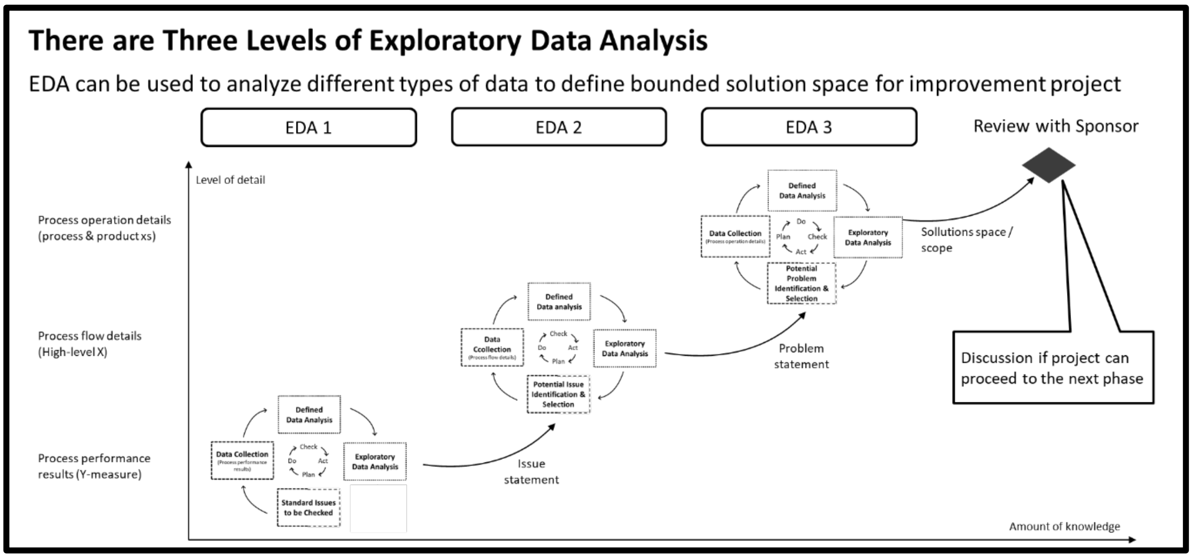
Evaluation

Graph(s) ready for analysis

Additional factors to be checked

Solutions space & type of improvement project

Figure 22. Slide 22



There are Three Levels of Exploratory Data Analysis

EDA can be used to analyze different types of data to define bounded solution space for improvement project

EDA 1

EDA 2

EDA 3

Review with Sponsor

Level of detail

Process operation details (process & product xs)

Process flow details (High-level X)

Process performance results (Y-measure)

Defined Data Analysis

Data Collection (Process performance results)

Check

Do

Act

Plan

Exploratory Data Analysis

Standard Issues to be Checked

Issue statement

Defined Data analysis

Data Ccollection (Process flow details)

Check

Do

Act

Plan

Exploratory Data Analysis

Potential Issue Identification & Selection

Problem statement

Defined Data Analysis

Data Collection (Process operation details)

Do

Plan

Check

Act

Exploratory Data Analysis

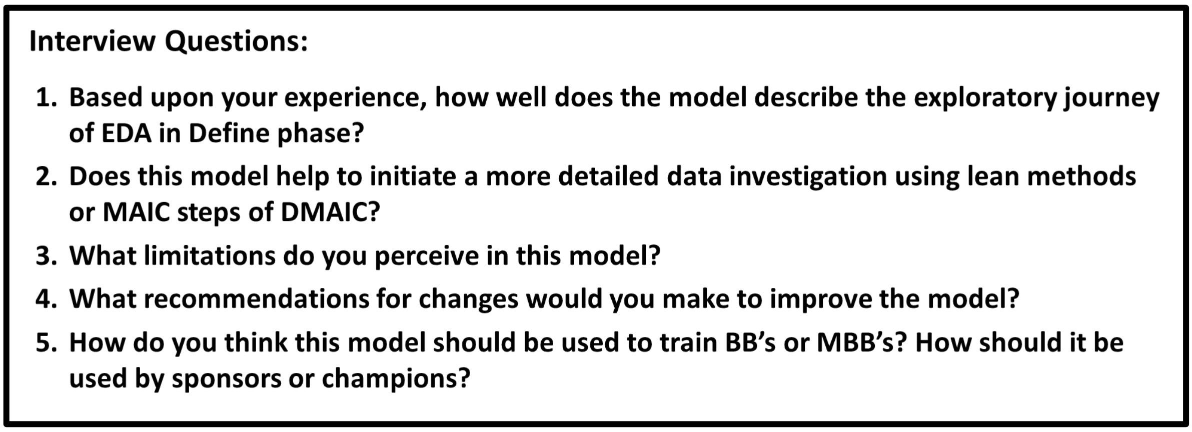
Potential Problem Identification & Selection

Sollutions space / scope

Discussion if project can proceed to the next phase

Amount of knowledge

Figure 23. Slide 23



Interview Questions:

1. Based upon your experience, how well does the model describe the exploratory journey of EDA in Define phase?
2. Does this model help to initiate a more detailed data investigation using lean methods or MAIC steps of DMAIC?
3. What limitations do you perceive in this model?
4. What recommendations for changes would you make to improve the model?
5. How do you think this model should be used to train BB's or MBB's? How should it be used by sponsors or champions?

Figure 24. Slide 24