LUT
Lappeenranta
University of Technology

Department of Business Administration
Master's in International Marketing Management

# USING ONLINE SEARCH QUERY DATA TO FORECAST BARGAIN SALE CAMPAIGN OUTCOMES

The examiner of the thesis was Professor Asta Salmi

Minna Virtaneva

2018

**TIIVISTELMÄ**


| | |
|---|---|
| Tekijä: | Minna Virtaneva |
| Otsikko: | Alennusmyyntikampanjan myynnin ennustaminen hakukonedatan avulla |
| Tiedekunta: | Kauppatieteellinen tiedekunta |
| Vuosi: | 2018 |
| Pro Gradu tutkielma: | Lappeenrannan Teknillinen Yliopisto |
| | 72 sivua, 17 kuviota, 2 taulukkoa ja 6 liitettä |
| Tarkastaja: | Asta Salmi |
| Hakusanat: | hakusana, hakukone, Google Trends, ennustusmalli, verkkokäyttäytyminen, lineaarinen regressioanalyysi, WoM, eWom, alennusmyynti, kvantitatiivinen tutkimus |



Tämän Pro Gradu tutkielman tavoite on tutkia Google hakusanojen volyymien yhteyttä asiakaskiinnostukseen sekä sitä voiko hakusanadatalla ennustaa alennusmyynnin myyntituloksia. Pro Gradu on tehty case-yrityksen avulla. Yritys on mahdollistanut todellisen alennusmyynnin tuloksien analysoimisen tarjoamalla alennusmyyntidatan tähän tutkimukseen. Hakusanadata on ladattu ilmaiseksi Google Trends- sivuilta. Hakusanadatan on todettu olevan harhaton otos asiakkaiden hiljaisesta kiinnostuksesta. Hakusanojen ollessa kaupallisia, niitä voidaan tulkita ostohalukkuuden mittarina haettua aihetta kohtaan. Tämä mahdollistaa hakusanavolyymien käyttämisen prokuurana asiakaskiinnostukselle ja siten hakusanadataa voidaan hyödyntää myyntituloksien ennustamiseen. Käyttämällä Google Trends hakusanadataa voidaan mahdollisesti säästää yrityksen käyttämää rahaa ja aikaa vähentämällä perinteisten asiakaskyselyiden määrää. Tutkimuksen päätutkimuskysymys on ”*Voiko Google-hakukoneen hakusanavolyymeillä ennustaa alennusmyynnin myyntituloksia, kun hakusanavolyymit toimivat prokuurana asiakaskiinnostukselle?”* Tätä tutkimuskysymystä lähdetään ratkaisemaan kolmella teoriaosuudella, jotka käsittelevät verkkokäyttäytymistä, hakukonedataa sekä hakusanadatalla ennustamista. Tutkimus on kvantitatiivinen ja toteutetaan klassisella lineaarisella regressiomallilla. Tutkimuksen tulokset ovat lupaavia. Hakukonedata onnistui ennustamaan merkityksellisesti kivijalkaliikkeen myyntiä sekä kokonaismyyntiä, muttei kuitenkaan onlinekauppamyyntiä. Tulokset viittaavat siihen, että hakukonedataan perustuva selittävä muuttuja parantaa nykyisiä ennustusmalleja, koska muuttuja pystyy ennustamaan sekä alennusmyynnin yleisiä trendejä että potentiaalisia käännekohtia.

## ABSTRACT


| | |
|---|---|
| Author: | Minna Virtaneva |
| Title: | Using Online Search Query Data to Forecast Bargain Sale Campaign Outcomes |
| Department: | School of Business and Management |
| Year: | 2018 |
| Master's Thesis: | Lappeenranta University of Technology |
| | 72 pages, 17 figures, 2 tables and 6 appendices |
| Examiners: | Asta Salmi |
| Keywords: | Search query, search engine, Google Trends, online behavior, forecasting models, WoM, eWoM, bargain sale, linear regression model, quantitative research |



The objective of this master's thesis is to examine search query volume- based predictor's predictive power towards the case company operated bargain sale. The phenomena of using search queries for forecasting has gained interest during the last decades due to the increasing usage of search engines. Search query data is an unbiased sample of the population and it represents genuine and unspoken interest towards the query topic. The searches include commercial queries, which can be considered as proxies for customer interest and thus can be used for analyzing and predicting sale sizes. Google Trends provides free to download files that include search query volumes from chosen timelines. This recently discovered data opens a new opportunity for companies to study their consumer interests without having to conduct costly and time-consuming consumer surveys. The main research question is "*Can Google search engine query volumes help estimating bargain sale outcomes from the perspective of representing a proxy for customer interest?"* There are three theory parts to find answers to this research question. These three theoretical parts are customer online behavior, search engine data and forecasting with search queries. The thesis is a quantitative research and the forecasting will be done with a classic linear regression model. The results are prominent. The search query data was able to forecast significantly the brick store sale and the total sales. However, not the online store sale. The results suggest that the Google predictor improves the current forecasting models by being able to detect public bargain sale trends as well as the potential turning points of the sale.

**ACKNOWLEDGEMENTS**

*The Case Company deserves a big thank you for taking interest on this topic and providing the necessary datasets. Also, my thesis examiner has been a great help at making me achieve my ambitious schedule. I would want to thank my friends as well for helping me through this writing process. Your support and especially patience during my frustrations have been appreciated. Lastly, I want to thank my parents. Thank you for facilitating me with food and good advice when most needed.*

# Table of content


*1. Introduction* ..... *1*

**1.1 Research topic** ..... **1**

**1.2 Main Keywords** ..... **2**

**1.3 Delimitation** ..... **3**

**1.4 Preliminary Literature Review** ..... **4**

**1.5 Theoretical framework** ..... **6**

**1.6 Research questions** ..... **7**

**1.7 Study outline** ..... **8**

*2. Literature Review* ..... *9*

**2.1 Customer online behavior** ..... **9**

**2.2 Use of online data for forecasting** ..... **10**

**2.3 Data handling methodology** ..... **12**

*3. Customer online behavior* ..... *14*

**3.1. Chapter outline** ..... **14**

**3.2 Search engine user motives** ..... **16**

**3.3 WoM creates searching motivations** ..... **18**

**3.4 Triggers creating word-of-mouth** ..... **21**

**3.5 Discussion** ..... **24**

*4. Search engine data* ..... *27*

**4.1 Search engine as a personalized tool** ..... **27**

**4.2 Google Trends** ..... **29**

**4.3 Search query data limitations** ..... **31**

**4.4 Discussion about Google Trends data benefits** ..... **33**

*5. Forecasting with search queries* ..... *35*

**5.1 Data analytics and data modifying** ..... **35**

**5.2 Search query forecasting methods** ..... **36**

5.2.1 Data visualization ..... 38

5.2.2 Classical linear regression model ..... 39

5.2.3 Autoregressive models ..... 40

5.2.4 Search query data in Nowcasting ..... 41

*6. Methodology* ..... *44*

**6.1 Research background** ..... **44**

**6.2 Bargain sale data** ..... **46**

6.2.1 Bargain sale data structure ..... 46

6.2.2 Bargain sale data limitations ..... 49

**6.3 Google Trends Search Query Data** ..... **49**

6.3.1 Search query data structure ..... 50
6.3.2 Search query data limitations ..... 52
**6.4 Data analysis ..... 54**
6.4.1 Visualizing the datasets ..... 54
6.4.2 Linear regression analysis ..... 56
6.4.3 Residual analysis ..... 57
**6.5 Forecasts and their interpretations ..... 62**
***7. Conclusions ..... 65***
**7.1 Research results and findings ..... 65**
**7.2 Answers to research questions ..... 67**
**7.3 Suggestions for further research ..... 69**
***8. Executive summary ..... 71***
***References ..... 73***
***Appendices ..... 77***

**List of figures:**

Figure 1: Theoretical Framework
Figure 2: Causal Map of Search Engine Use Motives and Triggers
Figure 3: Causality map of observed datasets
Figure 4: Sale outcomes subsets together (indexed)
Figure 5: Marketing budget versus Actual Sale Outcomes
Figure 6: Search volume index vs. Actual sales outcome indexes
Figure 7: Search volume index vs. Budgeted sales outcome indexes
Figure 8: SVI versus Marketing Budget
Figure 9: Actual brick store sales plotted against SVI
Figure 10: Actual total sales plotted against SVI
Figure 11: Residual plot from ACT Brick dependent
Figure 12: Residual plot from ACT total dependent
Figure 13: Residual plot against time (ACT Brick sales)
Figure 14: Residual plot against time (ACT total sales)
Figure 15: Normal Probability plot for Actual brick store sales residuals
Figure 16: Normal Probability plot for Actual total sales residuals
Figure 17: Forecasts for actual total and brick store sales

**List of tables:**

Table 1: SVI Forecasting Methods collected
Table 2: Search volume index division per bargain sale week

# 1. Introduction

This thesis starts off by introducing the research topic. The most relevant and often used term and keywords are explained after covering the research topic. In the third sub header the research limitations are covered. Under the fourth sub header there is going to be a preliminary literature review to introduce the main authors and research done towards the topic in hand. After this there will be the theoretical framework is presented, and research questions are introduced after this. The introduction chapter will end with an outline of the whole thesis.

## 1.1 Research topic

This thesis' research is conducted to see if customer interest can be evaluated through online search engines' search query volumes. More precisely can search query volumes help predicting case company's bargain sales' financial outcomes. The use of search engine search volume data has been proven to be at least helpful as a forecasting tool and thus, it should be examined further (Choi & Varian 2011). Forecasting methods have been previously used for predicting housing markets, employment and product prices just to name a few (Varian 2014). The topic is current and relevant because of the continuously increasing use of the internet and search engines (Statista 2018b). This user increase reinforces to study search query forecasting further by trying it on different economic measures (McLaren & Shanbhoge 2011).

This master's thesis is done with the help of a case company. This means that the search query data predicting power will be experimented on actual bargain sales data. The case company will provide the data about the studied bargain sale campaign. This sales data will work as the dependent variable for this thesis. The search query data is going to work as the explanatory variable. The used search query data is emitted from Google Trends. Google Trends provides search volume weekly statistics for different search queries. Just by looking at the histograms for the studied search query from Finland there is a notable repeating shape and distinctive peaks on search volumes. These peaks make it interesting to conduct this study and see if there is interconnection between the bargain sale sizes and search query volumes.

## 1.2 Main Keywords

The topic of this thesis has some specific keywords that will be repeated throughout the whole thesis. In this sub-chapter these keywords will be defined and explained. Some words are the product of other researchers. Due to the similarity of the research these keywords are also used in this thesis and author credits are given to these specific words.

**Bargain sale** and **Clearance** work as synonyms for each other throughout the whole thesis. They refer to the sale that is held twice a year by to the case company and which will work as the dependent variable and explained phenomena of this thesis.

**Search word** or **term** is the word or words that a person enters into the search engine to get wanted results. Refers specifically to one individual word.

**Search word query** is a more extensive when compared to just 'search word'. The search word query represents everything that a person might enter into the search engine at once. It can be a question, a line of words, a title etc. This study focuses on studying a search word query.

**Google Trends** is a search query data explorer of Google. It allows the user to see the unbiased data sample of a certain search word and query that users have entered into Googles search engine. The user can download the data as a CSV-file and explore it further. (Rogers 2016) In this thesis the data used to examine the forecasting power of the search word query is obtained from Google Trends.

**Google predictor** was used by the authors Askitas and Zimmerman (2009). By this term the authors describe the relationship between the examined phenomenon and the Google search query data. When talking about the search query data and its relationship on the bargain sale the variable will be called the Google predictor.

**SVI** is an abbreviation for *Search Volume Index* and it refers to the Google Trends datasets index value per week (Da, Engelberg & Gao 2011). In other words, SVI refers to the search word query volume that is in index form. The query index is used to describe the Google Trends data. The data is in index form instead of raw levels of queries for a certain search word. The index numbers are normalized values from 0 to 100 determining the total query volume for search term divided by the total number of queries at a point in time (Choi & Varian 200). Simplified the index is: search interest index = [(number of queries for search term in time t) / (Maximum

search query amount)] * 100. If the index is 100 it is the so far maximum of searches done during the examined period. If there is next value of 50 it means that the search volume was half of the maximum.

**Search query volumes** in this thesis represent a proxy for customer interest. This is an assumption made about the search query volumes that if the volumes are for example high the customer interest is also high.

**WOM** is the abbreviation of Word-of-Mouth. WOM is the activity of person-to-person information sharing (Buttle 1998). In the thesis WOM is held as face-to-face information sharing.

**eWOM** and **Online WOM** are word-of-mouth interaction that happen on online platform such as in social media, chatrooms or blogs (Davis and Khazanchi 2008). eWOM is an abbreviation from electronic word-of-mouth and as the name represents it is the same as WoM only the exception that the communication does to happen face-to-face.

## 1.3 Delimitation

All of the used data is secondary. The studied search query data is obtained from Google Trends (https://trends.google.com/). Sales data is obtained from the case company with the help of their employees. This thesis is going to be solely focusing on forecasting the bargain sale. The methods used in this thesis are going to precisely optimized for forecasting the studied data. Due to the confidentiality agreement of the company in question the information provided by the company such as the name of the company or the sale numbers will not be shown in this thesis. Nevertheless, this will not affect the study results or the interpretations of the results in any way.

The delimitation for this paper is mainly going to be the used datasets structures. For both datasets the main limitation is time. There exists valid data that includes the online store from the clearance back from 2013 and onwards. It is necessary to include the online store outcomes into the dataset since they are so current and might possess features that are neglected with the search query volumes. Thus, the Google Trends data is obtained from between years 2013-2017. This search query data will be the independent variable to study the sale outcomes. The dependent variables for the study are separately going to be the outcomes of the bargains sale campaigns. The bargain sale is held twice a year, spring and autumn. Google Trends provides weekly data from a 5-year-scope (2013-2017). In these five years there is going to be ten different sales which can be studied. From these ten

campaigns there will be data collected from the company. This data is going to be the bargain sale dates, the profit of the sale, the budgeted profit of the sale and the marketing budgets per sale campaign. Another limiting aspect for sales data and search query data is going to be location. All of the studied data is going to be solely collected from Finland. This is because the sales are only held in Finland and in Google Trends a user can limit the data to any country location or 'worldwide'. To keep the study coherent the search query data is going to be obtained only from Finland.

To understand the relationship between search queries and customer interest this thesis will focus on the reasons and drivers for why people use search engines. These drivers are limited to motivations and trigger because the goal is to find the connection between search engine usage and customer interest so that the customer interest can for certain be called the proxy of search query volumes. For search query data the information is limited to what Google Trends provides and thus the methodology will follow the limitations that this data form creates. Forecasting methodology will follow previous literature and the actual study is conducted with methods that are proven to be useful. Delimitations for methodology is the aspect that this thesis is not supposed to be exhaustive. More precise goal is to see if there exist an interconnection between variables and if the data in hand has potential for future forecasting. This keeps the methods on the basic econometric level which can be easily conducted with the help of Excel analytic tools.

## 1.4 Preliminary Literature Review

The Concept of search word query forecasting was first examined by authors Ettredge, Gerdes & Karuga (2005). After that there has been a significant peak in the amount of literature on the topic. Choi and Varian (2009 & 2011) have followed the research of Ettredge et. al. and made some groundwork for future research on this topic. Based on the mentioned literature, this forecasting method has been proven to be helpful in predicting the future of different economic phenomena. Depending on the nature of the phenomenon the result accuracy differs. This result variation was studied by Choi and Varian (2009) who studied search query forecasting on home sales, automotive sales and tourism. In general, this topic seems to be a trending topic since it is still generally new at the scientific field and it has shown promising results.

A doctoral dissertation from Gauri M. Kulkarni (2010) argues that search terms indicate purchasing interest and studies this on the predicting power of Google search queries on new product sales. This literature is the closest to this thesis and thus has helped structure

and choose the relevant theoretical frames. Chosen theories such as word-of-mouth have evolved from the research done by Buttle (1998) which had been continued by Godes and Mayzlin (2004). Another continuum from Buttles (1998) research has been the theory of electronic word-of-mouth which was studied further by Henning-Thurau, Gwinner, Walsh and Gremler (2004). These authors studied the motives behind word-of-mouth activity. Online behavior theory has gained texture from the work of Rose and Levinson (2004) who continued Broder's (2002) research on search engine user goals. These goals can be considered significant when studying the search engine user behavior. These behavioral factors are found interesting when evaluating the search query data's accuracy towards customer interest.

Purcell, Brenner & Rainie (2012) and Fallows (2005) from Pew Research Center have stressed out the increase of search engine usage and especially the position that Google as the most used search engine. Brin & Page (1998) and Evans (2007) have literature on the search engine behavior that due to the increase of search engine usage, is also evolving. Rogers (2016) and Choi & Varian (2009, 2011) discuss about Google Trends and Google Trend data possibilities for forecasting. With the comparative study between survey-based and Google Trends indicators Schmidt and Vosen (2009) support the future of using Google Trends for forecasting private consumption.

From the literature used for this thesis Google Trends data has been studied by authors Choi and Varian (2009, 2011); Kristoufek (2013); Kulkarni (2010); McLaren and Shanbhoge (2011); Vaughan and Romero-Frías (2013); Askitas and Zimmermann (2009); Da, Engelberg & Gao 2011; Schmidt & Vosen 2009. Especially Hal Varian (2009, 2011 & 2014) and Hyunyoung Choi (2009 & 2011) have been studying Google Trends data and observing the possible methods to get the most out of the data. Classical linear regressions have been the most used modelling for predicting econometrics phenomena but paper by Varian (2014) goes deeper into this subject and reflects that there are many other regression models that will perform better when handling big data. One of these possibilities is using the ARMA-family models for forecasting (Choi & Varian 2011). Mainly the results from research have been to the direction that Google predictors helps to predict the present and help the models' performance.

## 1.5 Theoretical framework

In figure 1 there is a theoretical framework showing how the thesis shall proceed. As mentioned the research starts by finding the reasons that create search query volumes in the first place. First step to doing this is by finding the main drivers of online search behavior. By doing this the credibility and preciseness of the search word query data can be evaluated.

The main drivers for customer searching behavior will be retrieved with the help of similar scientific articles and studies done by other researchers. The search query volumes are also observed from the direction of Google as a search engine. Since both attributes have their own features that affect the search query volumes they are both under observation to create reliability to the thesis' results. Furthermore reassuring the search query data it is necessary to aknowledge the limitation that come up through the theories.

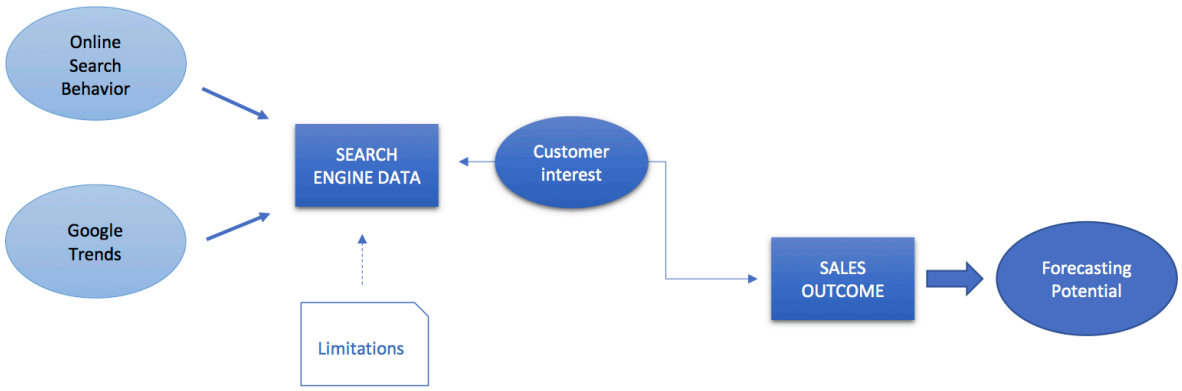



Figure 1: Theoretical Framework

After establishing the background and included features of the search engine data the theoretical framework moves on to customer interest. Customer interest is associated to be the proxy of search query volumes. This meaning the assumption that if the search volumes are high the consumer interest is high and vice versa. This is studied through previous literature by finding the causality between these two. To see if this assumption is true on hands-in level there is the bargain sale data to study. The possible relationship assumptions are tested by seeing how well the search engine data and the sales data correlate. Depending on the correlation level the forecasts are created. These forecasts will then be plotted against the actualized values to see how well the forecasts perform.

In order to find the optimal forecasting model, the previously used forecasting tools and methods from similar literature are observed. After evaluating the suitable forecasting methods there is a theoretical part about the forecasting methodology to avoid unnecessary testing and mistakes. Based on the methodology theory the thesis aims to find the optimal forecasting tool for data in-hand. After the model testing and forecast creation there should be some solid results that will help to give out answers to the research questions and show if the Google predictor possesses predicting power towards the bargain sale.

## 1.6 Research questions

To have clear goals for this thesis there exist the main research question and sub-questions to divide the main research question into smaller parts to solve. The main research question for this thesis is the following:

***Can Google search engine query volumes help estimating bargain sale outcomes from the perspective of representing a proxy for customer interest?***

Sub-questions are formed from this research question considering separately customer online behavior, search engines and forecasting methodology. The sub-questions for customer online behavior are following:

- Can search query volumes represent customer interest?

If so…

- What customer needs affects search engine activity?
- Is search engine activity a good measurement for customer interest?

Sub-questions for search query data are the following:

- Is Google Trends data a valid measurement for customer interest?
- Is Google Trends data valid for forecasting?

Sub-questions for forecasting methodology:

- What forecasting methods are efficient for handling search query data?
- Is search query-based forecasting beneficial?

## 1.7 Study outline

The thesis will start with a literature review. The literature review gives a representation about what there has been written about on the topic in hand and also, to see who have done the first and most significant studies. After this, the paper will move on to the theoretical parts. There are three main theoretical parts. The theory starts by going through customer online behavior as a phenomenon. The goal is to define what are the drivers that make customers use search engines. After defining the drivers there will be a more concrete causality relationship between the search word data and customer interest. The main theoretical frames under this chapter are going to be the goals behind search engine usage and triggers behind these goals.

After this the thesis moves on to theory about search engine data. The goal for this theory parts is to examine search engines and search query data. More precisely the target is on Google as a search engine. Hence, this theory part will also cover the theory about Google Trends. Important for the reliability for this study is to evaluate Google search engine and Google Trends data. Also, the structure of the search query data is under the observation and this is why there is a collection of search query data limitation from devoted authors.

The last theoretical part is going to handle the forecasting methodology. Aim is to find the methods that are used to analyze the relationship between the search query volume and sales profits. By evaluating previously used forecasting and data handling methods the goals is to find the best one to conduct this thesis' study. There will be also discussion about these methods and reasoning about their benefits and faults.

After the three theoretical parts have been covered the thesis moves on to the empirical part of the thesis. In the empirical analysis the goal is to see if there truly are any forecasting attributes between the search query data and the customer interest to creates sales. First in this chapter are the reasons behind the conduction of the study and connections between the theory and the conducted study. After this the thesis moves on to data structures. First thing to be explained is the company dependent bargain sales data and then Google Trends search query data. After this there is the constructing of the used forecasting method and then the used models. After the testing there will be the results and discussion about the results. The discussion will focus on the analysis of the results and on the reasons of the outcome. After the discussion there is going to be the study result and findings. The thesis ends with an executive summary. List of references and appendices are located at the very end of the thesis.

# 2. Literature Review

In this chapter there will be a narration about the main literature that has had an influence on the thesis topic and that has been directional to the conduction of the research. The first part is going to review about the literature used to find the main drivers for online search engine usage. This is done by examining customer online behavior. The second part is a see-through about the used literature to evaluate the search engine- based forecasting methods. The last part focuses on the data analyzing methods and on the forecasting models that are used to evaluate the actual correlation between company's sale data and Google Trends' search query data.

## 2.1 Customer online behavior

The amount of research on customer behavior is vast. For this study the focus has been mainly on online behavior, eWoM and triggers for these mentioned actions. The increase of internet usage has also increased the amount of research done towards understanding better the online behavior of customers and search engine functionality. After the millennium change there seems to be quite many and very specific articles about user online search behavior (**Rose & Levinson 2004; Broder 2002; White & Drucker 2007**), articles about who are the search engine users (**Deborah 2005**) and what is expected of search engines by the users (**Teevan, Alvarado, Ackerman & Karger 2004**). The most common motivation behind the conductions of these research articles has been the lack of previous research on the topic of online user behavior.

**Buttle (1998)** stresses out in his article "Word of mouth: understanding and managing referral marketing" the significance that WoM has on customer behavior. He discusses how WoM influences conditions such as awareness, expectations, perceptions, attitudes, behavioral intentions and behavior of the consumer. Buttle (1998) says that WoM is stronger than any marketer-controlled sources. Thus, WOM works as unconscious advertising for the marketer and it is free for the company. **Godes and Mayzlin (2004)** support this idea by stressing out the credibility that WoM has in sharing information. WoM is a possible driver for online searching because of the studied efficiency and hopefully also therefore a driver for consumer interest shifting.

Rose and Levinson (2004) as well as Broder (2002) study the motives search engine user have for their searching behavior. They try to find what are the users' needs and goals when entering a query into the search engine. They also aim to find the answer to the question:

'Why do people use search engines?'. Broder (2002) conducted his study by making a pop-up-window survey when using the search engine. This survey was presented to random users. The survey simply asked what the user was looking for in very broad question forms and all the answers had a different goal behind them. From these answers Broder collected a trichotomy of goals (navigational, informational and transactional) that searchers most probably have behind their search action. Rose and Levinson (2004) continue the research that Broder (2002) has done by adding their own search word query research into it. By doing so a new motivated goal was found: The recourse searching goal. Both of the papers help out to define the undefined goals that lie behind the customers' motives. The four search goals can all be connected to the reasons to why a potential customer would use search engines for information. This extended trichotomy of goals has worked as groundwork for further research about online behavior.

In article "Electronic Word-of-Mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?" by **Henning-Thurau et. al. (2004)** the authors described how word-of-mouth is evolving to be a vital part of the internet since the internet is becoming a big platform for discussions about different subjects. This change has created a new form of WoM which is called *electronic word-of-mouth (eWoM).* This is a change that has gained interest since eWOM creates obtainable data into the web that can actually be studied easily. The eWoM data is in more permanent form and thus, it is easy to use for research, which has not been possible from traditional WoM. Henning-Thurau et. al. (2004) discuss about the importance of the consumer motives created by WoM. These motives seem to be similar with eWoM based on previous literature. This is important since there is more literature based on traditional WoM than on eWom. If the core human motives stay the same, it is easier to understand the eWOM actions that can be obtained from the web and put under research. This list of motives created by WoM has worked as a base for the search engine user motives for this thesis.

## 2.2 Use of online data for forecasting

**Gauri M. Kulkarni (2010)** has done a dissertation about the topic *'Using online search data to forecast new product sales'*. This has been an important literature for creating the structure of this thesis and also, an important literature source for the studied topics. Kulkarni (2010) studies the online behavior of customers and how it might have a helping impact on forecasting future product sales. As for this thesis, Kulkarni also uses data that is provided by Google Trends to test the possible interdependence. Technical report written by **Choi and Varian (2011)** shows more technically how well Google Trends data can be

used to help predicting the future and the present for economic indicators. Their paper "Predicting the Present with Google Trends" from years 2009 and 2011 has worked as a groundwork for other researchers as well. Author **Askitas & Zimmermann (2009); McLaren & Shanbhoge (2011); Schmidt & Vosen (2009)** to name a few have been influenced from the research done by Choi and Varian and have had their own influencing input into the field of search query forecasting. Also, other researches done by Choi and Varian have worked as referential study to many other research papers.

Using search word data to forecast the future of an economic phenomenon is turning more common with the rise of the internet usage. Research done under the **Pew Research Center (2005 & 2012)** there are clear numbers on how the search engines usage is turning to be the most popular internet activity for people. Askitas & Zimmermann (2009) and **Ettredge, Gerdes & Karuga (2005)** have studied the effect of search word volumes on determining the employment rates. The forecasting of unemployment rate has been a rising topic and vastly studied after the Ettredge at. al. (2005) research. Both, Askitas & Zimmermann (2009) and Ettredge et. al (2005) studies found that there is significant or notable correlation between the economic phenomena and search query volumes. Choi and Varian (2011) followed the paper made by Ettredge et. al. (2005) and said it would be the first paper to study the predictive abilities of search query volumes (Choi & Varian 2011).

The big boom of bitcoins also made it quite common to use search word data to predict the currency floating because of the lack of an appropriate forecasting tools (Kristoufek 2013; Young, Lee, Park, Choo, Jong-Hyun & Kim 2017). Connected papers have come to the conclusion that the search query volumes have predictive attributes toward bitcoin value. After the year 2005 the number of articles about search engine-based forecasting is increasing. Search query forecasting has also created a boom in forecasting flu and influenza epidemics and such studies have been conducted by **Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009)** as well as authors **Goel, Hofman, Lahaie, Pennock and Watts (2010).**

Research by **McLaren and Shanbhogue (2011)** focuses on examining the labor and housing markets in the UK and can it be predicted by search query data. They discuss the problems that search word-based forecasting can entail when used on economic activity such as the newness of the data and stiffness of search terms. Another problem they debate about is the inevitable index form that Google Trends data possesses. The index property of the data narrows down the possibilities that the data can give. It also gives biased result when two different queries are compared to each other since the index is done only inside of the individual queries data. The indexes from different queries are then not compatible

with each other. **Kristoufek (2013)** mentioned another search query data problem in his research article. The problem is that the search data cannot be divided into positive and negative interest. Thus, it is not possible to study singularly the positive search interest. **Castle, Fawcett & Hendry (2009)** also adds some problematic features to the Google predictor. They characterize the forecasting methods as a trade-off between timeliness and data quality. Last salient data limitations were discussed by **Vaughan & Romero-Frías (2013**) and they were barriers that the language and location create to the data. Discussing problems around used language and word for the entered query are authors **Teevan, Alvarado, Ackerman and Karger (2004)** (different keywords), **Brin and Page (1998)** (misspelling queries) and **Vaughan and Romero-Frías (2013)** (language). So, even if the search query-based forecasting seems to have prominent results there are still issues that limit the usage and create debate between researcher. The debate is mainly about the reliability of the forecasting results when Google predictors are involved.

## 2.3 Data handling methodology

There are four categories into which data analysis in statistics and econometrics can be divided into. These categories were put together by **Varian (2014)** and they are: 1) predicting, 2) summarizing, 3) estimating, and 4) hypothesis testing. Even if not in the list, visualization is also an important part of econometric research and it is a good starting point if the data size and shape allow it. (Varian 2014)

The most vitalizing starting point for studying time-series search query data is indeed to visualize the dataset and to see how the variables compare to each other when plotted against each other on a common timeline. This then helps to validate if there is a possible relationship to study and post-evaluate the result by returning to the graph shapes. Visualization has been used by various researches such as Kulkarni 2010; Askitas & Zimmermann 2009; Choi & Varian 2011; Kristoufek (2013); Shimshoni et. al. (2009); Goel, Hofman, Lahaie, Pennock & Watts (2010). Using graphical methods in the beginning of the data analysis can also help to detect possible unwanted features and thus help to improve the dataset before actual model fitting (Brooks 2008, 133; 140-141; 165-167).

Varian (2014) discusses that regression models are the optimal tools for summarization. **Koop and Onorante (2013)** discuss in their paper how linear regression models are very useful for search query data and it is a straightforward method to create predictions. **Hal Varian (2014)** agrees that linear regression models are efficient and especially large dataset allow flexibility into the simple linear models. Varian also discusses about the

possibilities that other regression models than just linear regression models can offer for handling big data in forecasting. One possible method is using regression trees. The research of regression tools was continued by **Varian and Choi (2011)** by adding ARMA (Autoregressive Moving Average)- family forecasting tools as an efficient way of using search query data for forecasting economic metrics.

There is debate between researcher weather Google Trends data is relevant in short-term and long-term forecasting. **Varian (2014)** conclude from various authors that Google search queries have significant short-term predictive power when it comes to economic metrics. **Dimpfl and Jank (2012)** shows that search queries can improve long-run predictions when they study stock market volatility. The authors speak about autoregressive time series models being relevant for handling search queries and they help to created good predictions for long-term studies. **Choi and Varian** (2009, 2011) support this claim with a result that search query data can significantly help creating a predictor that helps predicting long-term values. **Koop and Onorante (2013**) debate against Choi and Varian's finding by saying that Google Trends data is rarely useful in broad macroeconomic variables. Though, Koop and Onorante (2013) do agree that for more specific variables (private consumption, housing or labor marker) the Google Trends data is proven to be helpful and necessary. Papers from **Choi and Varian** (2009, 2011) stress the predictive feature of Google Trends data for short-term predicting and predicting the present. In other word search query data is good for *nowcasting*.

**Schmidt and Vosen (2009)** conducted a relevant study about comparing survey-based indicators with Google Trends. They conclude their findings to the fact that Google Trends outperform survey-based indicators when forecasting private consumption. **McLaren and Shanbhoge (2011)** support this ideology and add the possibility that surveys are consciously collected based on pre-determined questions whereas the search query data is not and might help to answer unexpected and unsupervised issues. Search query data is collected based on behavior and is correlated with the current time thus it can provide more information about that time period.

In various articled used to conduct this thesis, the usage of search query data created Google predictor as one of predictors increases the model performance significantly. If not straightforwardly at least through decreasing the mean squared or mean average errors.

# 3. Customer online behavior

This chapter will handle the topic of customer online behavior. It is a relevant subject because it is expected to have causality to the action of online searching. The goal of this part is to find the drivers that lead the customers to use internet search engines. These established main drivers can be used to evaluate the search data more carefully and thus evaluate how well search engine usage behavior actually represent customer interest. Also, by defining the motivations and triggers to the drivers it is more possible to see if the search action can be controlled by the company or if the company has generated them possibly in the first place.

This chapter starts off with explaining the structure of the chapter and how the handled theories connect with each other. Then there will be theory about the main motives for search engine usage. After this there will be more theory about the main driver which is word-of-mouth. Before ending the chapter, the main triggers for WoM shall be covered. This chapter will end with a discussion about the connection relevance of the covered theories with the research subject of the thesis.

## 3.1. Chapter outline

This third chapter starts by examining the motives behind users for using a search engine. After this there is going to be a closer examination of those motives and how they have been triggered originally. In Figure 2 there is a causal map about the reasons which create search query volumes. From the figure 2 there is a link between ‘Customer interest’ and different goals. This goal- approach is created to describe the motivations that people have to use search engines and it was originally created by Broder (2002) but was shaped into the form that it is used in this thesis by Rose and Levinson (2004). These goals are created through different needs and lead to different kinds of search query forms. Through finding the user needs it is possible to track where these goals were originated from.

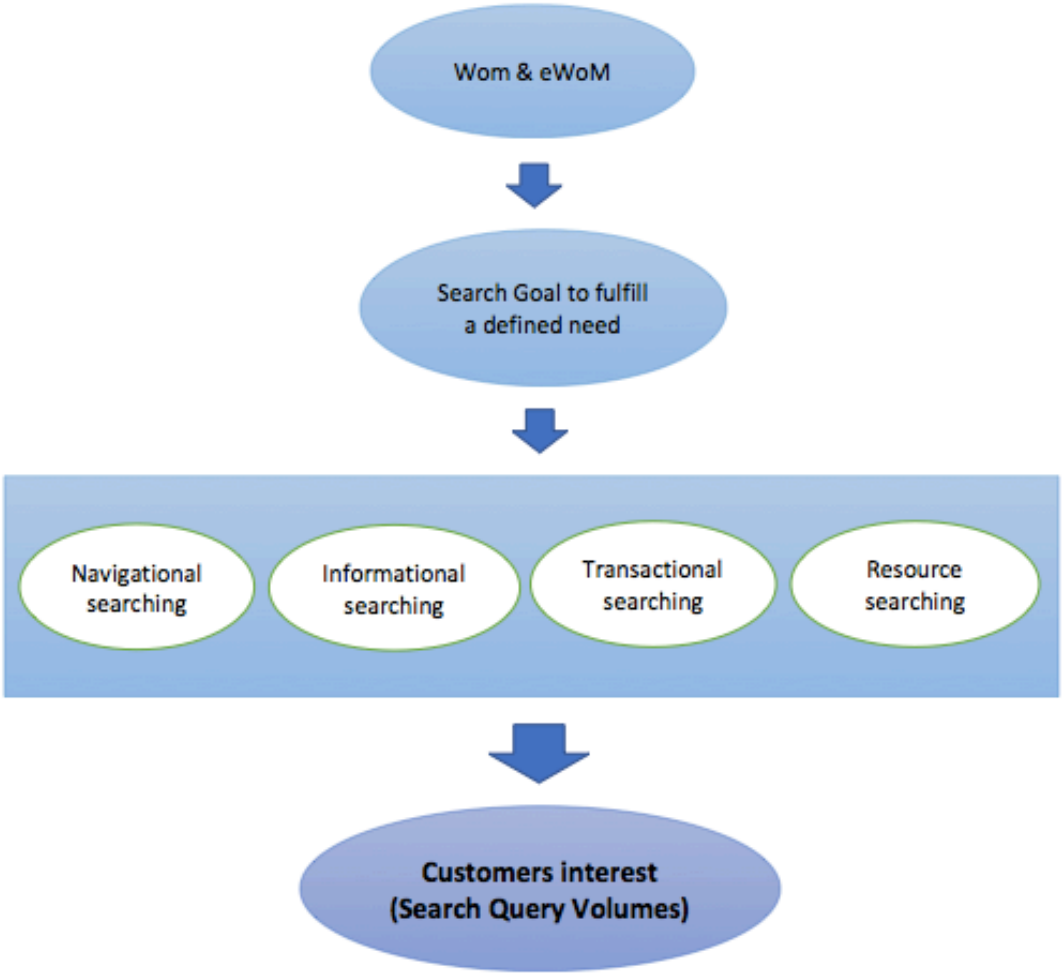



Figure 2: Causal Map of Search Engine Use Motives and Triggers

Figure 2 shows the main drivers that have been chosen to be the most accurate ones for this thesis. The main drivers are: Traditional word-of-mouth (WOM) and internet based electronic word-of-mouth (eWOM). There are certain needs that trigger people to share information to one another. In part 3.4 there is going to be more detailed theory about the reasons for why people take part in electronic as well as traditional word-of-mouth. Simplified the theory parts show that WoM/eWoM creates needs which turn in to goals. These goals are satisfied through search engines. Depending on the need the search query is formed. Together all these search goals create search volumes that can be examined.

## 3.2 Search engine user motives

The amount of search engine users in the United States in 2015 was 219.7 million users and the predicted user amount for the year 2020 is 239.1 million users. This will be approximately a 20 million user increase. After this the coverage is going to be over 70% of the whole population in the United States. (Statista 2018b) This 70% is only search engine users. When including all of the internet users the percentage will be even higher. These search engine users search for personally relevant information as well as answers to trivial questions. People use search engines for knowledge and also just for fun. Traditionally information-searching has been used for learning objectives. Nowadays, when search engines have made the information-search less challenging, it is more common to search information aka trivia just out of curiosity. (Fallows 2005) Based on studies it seems apparent that people search for information to fill out some need for example out of curiosity or due to an information gap.

Before the internet the information search has been through books and libraries. Now with the digitalization the traditional searching activity is moving more to the internet. But the surmise is that the motivations are still similar to as why people search for information, especially customers. From a commercial point of view the internet has opened up a whole new scale of opportunities. Now when a potential customer hears something interesting for example about a bargain sale they can simply just type in what they know about the topic into the search engine and expect relevant results. In order to understand the search engine usage goals, it is necessary to sort out first the motivations behind the search queries themselves. (Teevan et. al. 2004) This digitalization of information search seems also to be changing the form of consumer behavior since in traditional information search a customer should turn into a fellow product user for advice (in other word participate in word-of-mouth communication). Nowadays it is possible to evaluate products through the search engine results and find the optimal one just by browsing the web. Even though there are shifts towards more efficient information search possibilities the triggers for finding information should still be the same. These triggers, more commonly said these needs are the ones that should be established in order to understand the customer mindset before even starting the information search in the first place.

People use search engines to fill out an underlying need (Rose & Levinson 2004). Broder (2002) has created a trichotomy of web search types to understand these needs. These search types describe the need and the goal of the search that the user wants to fulfil. The types are: *navigational* searches, *informational* searches and *transactional* searches. *Navigational* searches have the goal to find a specific URL of a website. For example, a

customer could enter the search query command “example store” to the search engine and expect to be navigated to the official company website. *Informational* searches are considered to be the most traditional search types because it is similar to the traditional library information searching method (Rose & Levinson 2004). These informational searches have the goal to find further information about a topic in interest. For example, a potential customer could enter search command for ‘new car’ just to get more information about it. The expected result could be for example a catalogue of new car models, what are the cars like and where are they sold at. In other words, the informational search result is expected to be broad information stream about anything that is related to the topic of interest. *Transactional* searches aim to find a website in which an activity can be performed. For example, a user could search for a specific clothing brand and expect to get a website that offers the possibility to shop online for the specific brand products.

Rose and Levinson (2004) continue this Broder’s study (2002) by adding one more type into the web search types. This type is *resource* search. The goal of this type is to find something specific to have as a resource. This means that the searcher does not just want the transitory information. For example, a customer could enter the search command for a clearance in the hope to find the sale catalogue. Not just to view the catalogue at that moment but also to have it during the whole week when that sale is held. This search has not been done to learn about the topic but more about having the information at hand when it is needed. The authors described the resource searching from the searchers point of view as “My goal is to obtain a resource (not information) available on web pages”. Rose and Levinson (2004) categorized the need for the resources as download, entertainment, interact and obtain. Downloading is an indication for owning a resource for further need and retention. Entertainment indicates to videos and pictures. Items that can be watched that are available to find from the result page. Interact resources are results that help the current situation for example the weather or measurement converter tool. Last resource form was to obtained resources which the authors indicates as something that is possible to print out or just look from the screen. These are thing as lecture plans and scientific articles.

Originated from Broder’s (2002) research authors Rose and Levinson (2004) specify informational searches into two categories: Direct and Undirected. The separation is based on what kind of further information is the searcher aiming to get and how it effects the search term entered to the search engine. Direct goal is targeted to find a particular further information on the topic for example, using the search word query ‘When is the clearance held’. This kind of a search query has only one unambiguous answer. The second one, unidirectional goal, describes the searches desire to learn anything or everything about the

topic. In this case the search query is very open for example only entering a company name and the expected search results are any information where the search word query is mentioned.

When understanding the user goals and what expectation lie behind the entered search query, it is possible to increase the performance of the search engine on result accuracy. Also, there is the possibility to gain relevant data about what kind of information search engine users are looking for and what triggers them to search the topic. (Teevan et. al. 2004) By understanding these goals that the potential customer might have it is also easier to evaluate how well the search word query volume can predict the actual interest towards the query-subject. When thinking about the search engine user goals, they are limited to personal levels and are quite hedonistic, which is understandable since the main goal is to fulfill a personal need. (Rose & Levinson 2004) But even as a personal need the need can be one that evolves around social interaction. With further research there seem to be more community-based motives that might trigger information needs for search engine user. These motivations are created by person-to-person interactions or in other words through word-of-mouth. These community-based approaches study more precisely the motives that happen before entering a query to a search engine. These community-based approached are also more pervasive. (Godes & Mayzlin 2004)

## 3.3 WoM creates searching motivations

Vakratsas and Ambler (1999) discusses in their article that the ignition for the actual customer behavior comes from the mental effect that is generated by an outside trigger. The authors have done this study based on more than 250 articles to observe how advertising works on consumer behavior. Their paper focuses on how advertising can work as the ignitor for the mental effect. Important for this thesis is to stress out from the article the fact that adverting does not only mean the conscious advertising from the marketer. Advertising can also be implicit and interactive, and it may happen unconsciously from the marketer, but it has the same advertising effect on the consumer. This unconscious advertising can be for example conversations with fellow consumer about a product excellence. Studied by Lindgreen and Vanhamme (2005) this is also called viral marketing or as the thesis focuses on calling it, the word-of-mouth marketing. Word-of-mouth marketing creates marketer unconscious advertising which might then triggers the customer to buy the product or at ‘google’ it and see what it is all about.

Word-of-Mouth is one of the oldest information-sharing methods and has existed as long as the humankind has been able to communicate to one another. WOM is person-to-person communicating where there is an information communicator and the receiver. The information that is shared is non-commercial, but the message has a positive or negative approach toward a company, brand, product, service and so on. WOM-activity is said to have a strong influencing ability on people's knowledge, feelings and thus their behavior. This activity is said to be stronger than any marketer-controlled sources. Thus, word of mouth could work as a company promoter and it is free of charge if the company gets its consumer to share their product/service experience. This is only one of the reasons why WOM has been studied lately more and more. The goal would be to find a way to control the information that customers share and thus create good company image and increase sales. (Buttle 1998; Lindgreen & Vanhamme 2005)

Godes and Mayzlin (2004) are supporting this idea by stressing out the credibility that WOM has compared to anything else. Peer-to-peer communication has the authenticity that is hard to mimic by a company. This is because the authentic feel comes from the assumption that a company cannot be involved in the interaction. The honest agenda of simply sharing information about positive or negative experience reinforces the power of WOM. Lindgreen and Vanhamme (2005) continues this mentality by saying that WOM marketing has the power it has because of the emotion of surprise. This emotion of surprise refers to the experience the customer has had on the company or product. When consumer experiences a surprising emotion of satisfaction or dissatisfaction it is more prominent that the person will share this experience to others (Henning et. al. 2004; Lindgreen & Vanhamme 2005).

With the importance of WOM for a company there is a problem. Godes and Mayzlin (2004) address there is a difficulty of actually measuring WOM and control its usage on promotion. This is because information change happens in conversation forms. Word of mouth usually happens face-to-face and is strongly linked to the attitudes of the communicator and receiver (Lindgreen & Vanhamme 2005). Godes and Mayzlin (2004) aims to solve this problem by focusing on the current trend which is the electronic word of mouth (eWOM). The eWOM is easier to examine since it leaves a permanent trace into the internet which then can be obtained and observed. Electronic word-of-mouth message is also more straightforward since the sender needs to articulate his or her message into writing thus it is easier to understand if the message is positive or negative approach. (Lindgreen & Vanhamme 2005) Another noteworthy feature is that when studying eWOM it is possible to examine the whole conversation from beginning to end and see how the topic evolves during conversation (Hung & Yiyan Li 2007).

Cheung and Thadani (2010) discusses that the advent of the internet has accelerated the power of the (electronic) word-of-mouth activity. The internet provides platforms that are easy to access, participate in conversation and share information of their own. These platforms also offer on-date information about current topics and consumer products. Online platforms allow people whom don't know each other to spread information between one another which then broadens the scale of information receivers. This concluding to the fact that the internet provides bigger networks in which consumer can have word-of-mouth activity between each other. (Henning-Thurau et. al. 2004) Davis and Khazanchi (2008) summarizes online based WOM as: "The ability to exchange opinions and experiences online is known as online word of mouth (WOM)".

Henning-Thurau et. al. (2004) stress out the power of WOM and how the movement towards eWOM also shift the power from the company to the consumers. This is through the network that the online platform provides and thus gives the consumers the freedom to share any kind of criticism to others. This power of consumer to consumer influence turns big when the topic gets a buzz on it. This buzz is interesting since it is created by word-of-mouth, but it also creates word-of-mouth. The term 'buzz' refers to the phenomenon when a topic gets an unexpected online interest around it and this then creates web searching and discussion around the topic (Kulkarni 2010). If a company manages to get a buzz about its products, it will be free worldwide advertisement. An optimal goal would be to gain WOM on a positive product experience which then would turn into an online buzz and catch the attention of people that would never otherwise be in touch with the company.

In Buttle's (1998) article there was an important notation that WOM does not have to be only consumer-to-consumer based information sharing. It can also be a bit more commercial without being straightforward company advertisement. This kind of "half-commercial" WOM happens when the other person is a company employee and the other is a normal consumer but the conversation between the two does not evolve around the goal to gain sales. The WoM that happens between for example, the sales assistant and the customer may trigger the customer to search more information about the topic he or she heard about due to the knowledge of the sale assistant. When the company employee is part of the WoM there might be skepticism about the genuineness of the conversation. When including a friendship aspect to the conversation it makes the expectation for the conversation to be genuine. Thus, this kind of "half-commercial" WoM can either work as peer-to-peer information sharing or it can feel like a face-to-face advertisement.

As stated WOM has an important place when evaluating consumer communication and information sharing and thus, customer behavior. To understand customer behavior even

deeper the theory will look into the reasons behind word-of-mouth. By finding the reasons why people share some things and some they don't it is possible to understand if these triggers could be possibly controlled by the company. Through controlling these triggers there is a possibility to control the creation of brand awareness and customer interest (Chaudhuri & Holbrook 2001).

## 3.4 Triggers creating word-of-mouth

There exist reasons behind why people participate in word-of-mouth activity and these reasons are similar to electronic WoM. Authors Henning-Thurau et. al. (2004) collected together the primitive reasons why people share information between each other. In their article "Electronic Word-of-Mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet?" the authors discuss about the electronic word-of-mouth motives. They state that the consumer motivations created from WoM are similar with eWoM, based on previous literature. The main WoM drivers were collected together from papers by authors: Dichter (1966); Engel, Blackwell & Miniard (1993) and Sundaram, Mitra & Webster (1998). Since the papers are continuations of each other there were same motives listed between authors. For this list these duplications were removed. These itemized motivations collected by authors Henning-Thurau et. al. (2004) are:

> *From Dichter (1966):* Product-involvement, self-involvement, message-involvement and other-involvement. The word involvement indicates to the strong connection that the communicator has towards the discussed topic. *Product-involvement* refers to a strong feeling about the product and thus the need to share it further on to gain personal reassurance. *Self-involvement* represents the persons need to share information about a product to gratify personal emotional needs. *Message-involvement* is the outcome of a message trigger of an advertisement, commercial etc. that was strong enough to simulate the need to discuss about the topic further with a peer consumer. *Other-involvement* refers to the activity where the communicator has a need to share the topic to the receiver for more specific personal or social reasons.
>
> *From Engel et. al (1993):* Involvement, self-enhancement, concern for others and dissonance reduction. *Involvement* is the involvement the communicator has on the topic and this simulates general need for further discussions on the topic. *Self-enhancement* refers to the need of the communicator to gain attention with the

product information and expressing him- or herself as an intelligent shopper. *Concern for others* is the genuine need to help the other such as friend or relative to make good purchasing decision. *Dissonance reduction* is the feeling after a bad purchase and my conversation the goal is to reduce this doubt.

*From Sundaram et. al. (1998):* Altruism (positive and negative WOM), helping the company, anxiety reduction, vengeance and advice seeking. *Altruism* from positive or negative WOM refers to the act of sharing own experiences and through that the target of helping the other to make a good purchasing decision or avoiding it. *Helping the company* as the name says is simply the desire to support the company due to personal involvement or through good experience. *Anxiety reduction* refers to the ease that a discussion gives to the communicator on his or her negative feeling of anger, anxiety and frustration that the product purchase or advertisement has created. *Vengeance* is to share personal negative experience to make the company look bad as a revenge of the personal bad experience. *Advice seeking* is the act of discussion when the goal for the communicator is to resolve a problem by obtaining advice from others.

Henning-Thurau et. al. (2004) compile from several authors that consumers share WOM mainly when their consumption-related expectations are not reinforced. Depending if the WOM is negative or positive the sharing motivations differ. Also depending on how strong/involving the need or the feeling has been for the communicator affect the message effectiveness (Lindgreen & Vanhamme 2005). A strong emotional feel to the subject (Lindgreen & Vanhamme 2005) and negative experience (Anderson 1998) are the most probable conditions for a consumer to share their experiences to others. This is reinforced by the idea that when one consumer finds something strong emotionally (negative or positive) it would be expected to be the same for peer-consumer and thus sharing the experience would be considered to be highly useful for the receiver as the experience was for the communicator (Lindgreen & Vanhamme 2005). The assumption behind the communicator is that there is some utility for the information which could be moved forwards to the receiver. This utility approach was explored by Balasubramanian and Mahajan (2001), Lindgreen and Vanhamme (2005) and also used by Henning-Thurau et. al. (2004) to understand the need for sharing a message to one another from a different direction than just personal emotion.

When combining all of the mentioned WoM motivations together and observing them together the result is that they all have some kind of utility behind them (Henning-Thurau et. al. 2004). Balasubramanian and Mahajan (2001) describes this utility framework to

integrate social and economic activity together. These utilities are *focus-related utility*, *consumption utility* and *approval utility*. When examining the collected motivations, it can be seen that all of them fit into one of these utilities and hence gives a bigger picture about consumer behavior (Henning-Thurau et. al. 2004). The *focus-related utility* describes a motive that support and strengthens the community. *Consumption utility* reflect the gain from the contribution of others in other words the consumption of the conversation creates further knowledge to the communicator and the receiver. *Approval utility* focuses on the idea that the communication generates social approval. (Balasubramanian & Mahajan 2001) The collected motivations can be divided in to these utility- groups based on their character (Henning-Thurau et. al. 2004). The original group division for motivations was roughly done by Henning-Thuraus et. al. in their article (2004). With the adjustments of the list of motivations this division was adjusted further in this thesis. Here is a list of these motivations divided into utility groups based on their utility purpose:

> *In focus-related utility group there are:* Other-involvement, Concern for others, Altruism (positive and negative WOM) and Helping the company. Vengeance also is included in this group since it has a community impact and the utility groups are narrowly assumed to be including only positive motivations. Message-involvement is included into this group if the messages main goal is to support the receiving end in some way.
>
> *In consumption utility- group there are:* Product-involvement, Self-involvement and Advice seeking. Message-involvement is included in this group if the message is shared to the receiver personally in order to gain further information about the topic and the receiver is the one who offers the information. Dissonance reduction is included in this group since the goal of the communicator is to share the word in order to reduce personal dissonance and this same goal applies also for anxiety reduction- motive
>
> *In approval utility- group there are:* Self-enhancement and Involvement.

People use search engines to fulfill an underlying need. This need is created by interactions that happen through word-of-mouth or through electronic/online word-of-mouth. People share consumer related word-of-mouth when they have had a strong feeling of involvement about the consumer experience and then the shared information will work as a utility for the communicator, receiver or community. This user need then forms into a goal that the search engine is expected to fulfill. Thus, hits on specific search queries will increase.

## 3.5 Discussion

The observed bargain sale is held twice a year and by so has evolved to be a part of some consumers yearly traditions. The “rareness” of the sale creates word-of-mouth around the sale when the sales dates are approaching, or this could be assumed based on the notable search volume increases around the sale dates. The WoM will be spread by dedicated brand loyalists (Chaudhuri & Holbrook 2001) and by consumers who find the sale convenient. Due to the increase of internet usage the bargain sale has gained to be a part of also the electronic platforms and creation of eWoM. But in any scenario, there are motivations that create buzz around the sale and these motivations can be associated with the ones mentioned in the theory. Especially current with the bargain sale search query volumes would be mostly the positive motivations. This with the assumption that a consumer would not look up bargain sales which they have not heard good feedback about.

The data about the search query volumes of the bargain sale are only numbers, but they can be assumed to be created through navigational, informational, transactional and resource search goals. The navigational target is to find the website of the company hosting the bargain sale or the bargain sales own website. Informational search would give the goal of finding where and when the bargain sale is held. Transactional goal could target to find the online shop of the bargain sale and avoid going to the physical store. Lastly the resource goal would be in the bargain sale case to have the catalogue and timetable of the sale. These are the objective query hits that show as high or low index values. To see how the bargain sale queries are created the motivations behind WoM might give explanations.

Main motivations behind WoM about the bargain sale would be product-involvement, self-involvement, message-involvement, self-enhancement, concern for others, dissonance reduction, helping the company, advice seeking and other-involvement. The bargain sale is held only around one brand, and this increases the product-involvement aspect for the customer to share their involvement in the brand (Chaudhuri & Holbrook 2001). Self-involvement is strong also because of the trust towards the brand and for many the mindset that they return to the bargain sale every time to reinforces the gratification of personal needs. Message-involvement was the trigger created via advertisements and around the bargain sale there is marketing around the sale and hopefully these ads also create WoM amongst potential customers. Self-enhancement can be considered as a part of the fact that the focus in on a bargain sale which immediately might refer to an intelligent buying habit when compared to buy the brand-product normal priced. When observing more the question why a person would share information about the bargain sale the concern for

others- motivation is definitely one. The dissonance reduction- motive is not that positive from first glance but since bargain sale might have the tendency to push the consumer to buy something out of low-price rather than need the consumer might need to discuss the purchase in order to realize it was a good purchase. Helping the company- motivation might mainly come from old employees or some real brand loyalists. Advice seeking- motivation might most definitely be a leading factor to the bargain sale since there might exists peer-consumers who need advice where to get the specific brand with affordable price. Lastly the other-involvement should be mentioned since there can be many different personal, historical and social reasons why a customer would share the word about the bargain sale.

These motivations towards the bargain sale can also be categorized to the utility- groups. The bargain sale connects people and brings similar consumers together. This need for connection might create the focus-related utilities for the consumers in the need of strengthening the community. Consumption utility mainly would refer to the need to discuss the sale advantages and why the consumer should buy a product that is offered only during the sale. Since the bargain sale is a brand product sale as mentioned it does evolve around the approval utility. Consumers want to feel a part of a social group and product brands are modern ways to show in what lifestyle the individual is involved in or wants to be involved in (Chaudhuri & Holbrook 2001). The bargain could be considered most strongly as the approval utility since the bargain sale lures customer with rare items and reduces prices. This would reflect mostly to the need to make a good buy and gain specifically product from that specific brand.

Google predictor (search query data) aims to provide a measure for "*consumers' preparatory steps to spend by employing the volume of consumption related search queries*" (Schmidt & Vosen 2009). These preparatory steps can be thought as a part of the consumer purchasing process. The process steps are 1. Problem recognition, 2. Information search, 3. Evaluating alternatives, 4. Purchase decision and 5. Post-purchase evaluation. When considering the triggers and motives for search engine usage they go hand-in-hand with process steps 1, 2 and 3.

The first step of problem/interest recognition can be considered as the trigger for the action which in this case can be seen as the utility-based approach. The second step of information search is quite straight-forward. This happened between WOM conversation and using the search engine or solely in the current internet era the consumer goes straight to the search engine for information digging. The third step 'Evaluating the alternatives' is easily done through the search engine as well since it is quite simple to just enter queries, find results and then compare these results to each other to find the best choice (step 4). Step five can

work as the ignitor for new triggers. The post-purchase opinion can be spreader as WOM or eWOM and thus creating need to other consumers and/or helping other to evaluate their product choices. The use of search engines has evolved to be such a strong part of our lives that it is starting to be essential to include it into many traditional consumer behaviors analyzing tools (Koufaris 2002).

# 4. Search engine data

The fourth chapter focuses on examining the theory around search engines and the creation of search query data. The first part here is going to be about online search engines in general. The second part is going to handle more precisely the features about Google as a search engine and Google Trends as a data provider. The third part aims to criticize the usage of search query data and discover the main limitations that search query data has and how the problems can be avoided and possibly solved. The last sub-header there is going to be discussion about the Google Trends data benefits. Main focus of this sub-header is how search engine data can be useful and specifically on a company-level how it can accelerate measuring metrics performance.

## 4.1 Search engine as a personalized tool

Search engines have been around almost as long as the internet. Search engines help internet users to find what they are looking for from the vast amount of provided information. The first search generators were made and used between 1995-1997. From there on the search engines have evolved remarkably. The first namely official search engines were made and used in 1998-1999 and it could support navigational and informational queries. In other words, the search engine could lead to the pages were the information was held but not yet provide pictures, maps of downloadable material. This is the period when Google search engine was born. Google is a revolutionary search engine since it was the first one to use link structure of the web for quality ranking and first to utilize links to improve search query result relevance (Brin & Page 1889). After the year 1999 search engines have evolved further on to target the user need behind the search queries. (Broder 2002) Ever since the search engines have been advancing in a rapid pace and searching is currently one of the most popular activities to do in the internet (Purcell, Brenner & Rainie 2012).

All search engine users are individuals and there exists user divergence among the 219.7 million search engine users (Statista 2018b) that can be distributed into more perceivable groups. The probability that all search engine users with different geographic and demographic features would be homologous search engine users is doubtful. Research done by Pew Research Center and written by Deborah Fallows (2005) concludes that the most probable search engines user is going to be a young, high earning and educated man. The article showed that 88% of men and 79% of women are habitual search engine users.

In study also made by Pew Research Center in year 2012 showed that 83% of the whole focus group preferred Google as their number one search engine (Purcell et. al. 2012).

There seems to be more difference between age groups than between genders. Fallow (2005) continues in her research that 89 percent of under 30-year-olds use search engines. This makes the age group the most active group of search engine users. After this there is a pattern. The higher the age of the group is the less is the search engine usage is in percentage. In the age group 30 to 49 the search engine users are 85% of the age group. In the age group 50 to 64 the usage percentage drops to 79%. But even with the differences it is remarkable that every group mentioned has over 50% user coverage. This study by Pew Research Center was conducted to the users who live in the US during years 2004 and 2005. From the geographic point of view and the time span there might be some differences compared to the Finnish population search engine usage in 2018. But again, the amounts probably won't differ that much between western countries. It is safe to say that at least half of the western population uses search engines.

On daily base user activity Pew Research Center made a study in 2012. The research result was that the daily search engine usage is 60% of the users in group 18 years to 49 years. In group of 50-years and more the daily usage is only 41%. But in 50+ group the weekly search engine usage is higher (39%) when compared to group 18-29 and 30-46 who had a 26-27percentage for search engine daily usage. The search engine usage is significantly higher in groups that have the highest education and income levels. (Purcell et. al. 2012) Concluding the researches it is more probable that the search engine user is a young male with high education and high-income level.

Due to the huge user base and individuality enhancement in the 2010- century the search engines' functions have been evolving towards personalization (Brin & Page 1998; Hannák, Sapiezyski, Kakhki, Lazer, Mislove & Wilson 2017). This personalization is the activity where the search query results are optimized according to the users' previous search activity, browsing activity, location and other collected broeser cookies (Hannák et. al. 2017). This targeted goal is interesting since 65% of search engine users agreed in a research done by Pew Research Center that: "*It's a BAD thing if a search engine collected information about your searches and then used it to rank your future search results, because it may limit the information you get online and what search result you see*" (Purcell et. al. 2012). This act of personalization is problematic because it is often used only for commercial purposes rather than just helping the user to navigate the internet (Brin & Page 1998; Purcell et. al. 2012). However, it does optimize the search query results and helps company marketing success when their ads reach the correct target group.

Search engines have existed only approximately 20 years and have evolved extremely fast from nothing to something that millions of users use on daily bases (Broder 2002). By the evolution of search engines, it is turning more and more common that internet users 'google' everything they feel of interest and most commonly the internet path starts with the use of a search engine (Evans 2007). It is also turning more common that users see only the first ten results of the entered search query and then resume with another search query (Brin & Page 1998; Evans 2007). This means that the search engines need to be more precise on what kind of result come out first and user learn to the fact that even with the simplest search query the right result is expected to pop-up (Teevan et. al. 2004). Also, the learned ease of 'googling' necessary things at the needed time has made it more obvious that users search for certain things when they are current for example the query volume for "diet" is connected with new year's resolutions and query volume peaks in the beginning of every year and is the lowest on holiday seasons (Da, Engelberg & Gao 2011).

Hannák et. al. (2017) discusses in their paper that the personalization was originally used to give the users better search results but due to commercialization and the lack of personal privacy more negative sides are appearing. The referred authors debate that there should be more research done towards these negative side effects since it has been studied and proved that personalization is in general better for the search engines functionality. The main negative side is that in order to personalize search results personal data needs to be collected but then it is important to secure this data. However, the whole purpose behind personalization is to get more people to use search engines due to their invincibleness. Since search engines are not information themselves but a tool to find what the user is looking for, it is just natural for the result to be personalized for the query in question. A search engine is a tool for the user to browse the web and find what they are looking for. Thus, the mechanism of search engines should evolve around the user and how to fulfill the user search goal most efficiently (Teevan et. al. 2004). However, the question here is to fulfill the search query need, not to optimize the entire internet to the user.

## 4.2 Google Trends

There exist many search engines in the world such as Bing, Yahoo, Altavista, Ecosia just to mention a few but Google is the most used search engine in the world (Statista 2018a; Purcell et. al. 2012). Google as the most used search engine has obviously the vastest user base. This big user base creates better credibility to the search data. The bigger the sample the more there is data to support the dissertations and thus making the study reliable.

Googles datasets are also unbiased, anonymous, classified and aggregated (Rogers 2016). These are the reasons why Google search data was chosen to for thesis.

Google has made it possible for users to observe the search query data themselves by offering Google Trends platform where one can download search data for a precise search query from a certain time period and location. Google search data is free to view and download. It is available at: https://trends.google.com/trends/. The site allows the user to type in the search query of interest and see graphically how many hits the query has had in for example, the past 3 years. The user can adjust the search timeline to the furthest as to 2004 and change the region from worldwide to only including one region. Google Trends also shows the most used related search words to the one enquired. The search query includes all hits that has had the search word in it. (McLaren & Shanbhoge 2011) For example, if the search query is for the word *'data'* the query includes everything that has the word *'data'* in it such as *'data analysis'*, '*data forecasting'* etc. The user can also compare search words and see how they compare to each other in a graph form.

The user can download the data as a CSV file and can easily reformed it into an excel-file. The data shows the date and the weekly index of search queries. Because of the increase of search engine usage Google needed to create and index to count the used search queries. This is because the number of actual searches would be too big to monitor. This weekly index is calculated by diving the number of searches that include the query term by total number of online search queries submitted during the week and the highest index number is normalized to be maximum of 100 (Choi & Varian 2009).

Google Trends data is time series data and depending on the observed time period the data can be divided into *real time* and *non-real time* data. The difference between these datasets is the time period from which the random sample of search data has been obtained. Since Google Trends provides weekly indexes the real time datasets include everything within the last 7 days and the non-real time datasets include everything from 2004 till the current hours. (Rogers 2016) For this thesis the observation is done with non-real time datasets.

Because of the informative and reliable data that Google Trends provide journalist have been using Google trends information to stay on track on the most recent topics (Barrett 2015). It is easy to see the interest spikes and the topic trend line against time. An election study is a good example of how the Google Trends data is informative. Since the search queries can be studied by location the user/journalist/academic can compare how different states in America are interested or curious about opposing candidates. (Rogers 2016)

## 4.3 Search query data limitations

So far, this paper has focused on the benefits of using search word data for forecasting purposes. However, it is important to go through the problems and limitations that the forecasting method might pose. As mentioned, there are differences in the search engine users and in their behavior. There are also limitations in the search query data itself and by acknowledging these limitations the data can be analyzed reliably and comprehensively. The most common limitations are collected together from various authors that have come upon these method limitations and these limitations are handled next.

McLaren and Shanbhoge (2011) discussed in their paper that customers vary in their way of using search engines, which might affect the size and reliability of the observed sample. The difference appears through the searching habits. Customers might search for the exact same thing, but they might use completely different search queries and keywords. Authors Teevan, Alvarado, Ackerman and Karger (2004) discussed about this problem of search engines users using different keywords. When studying only a certain search query this means that the relevant data that is emitted from Google Trends might be missing an important sample of interested consumers. Instead of using the specific search query that is studied the users might enter search queries that lead to finding the official website of the bargains sale by using completely different keywords. Brin and Page (1998) support this by suggesting a user habit that has effect on the search query which is misspelling entered keywords and Vaughan and Romero-Frías (2013) add the language difference for keywords. A misspelled word and using a different language works the same as another keyword for a search engine. Google has the tendency to correct obvious misspellings and with the help of Google Translate include more language options thus give the right search result anyways. This means that these search engine users should be included in the sample of interest but due to the restriction of the observed search query they are not. It is not certain though that all languages and misspelled words are corrected by the search engine. These different keyword queries are not prone to be included into the datasets unless they are all individually collected together by the researcher (Ginsberg et. al. 2009). It would be a lot of survey work to find all the misspelling, different language versions and all the different query variation that it would demand a lot of time.

Since the obtained data from Google Trend is in weekly index form instead of in amounts, it creates limitations to the data usage. It is expected to do the analysis so that it includes the index formalities since there is no fixed numbers of search amounts. McLaren and Shanbhoge (2011) criticize the index form that Google provides. The index form is valid for research only when one search query is examined singularly. Different queries cannot be

compared because the index numbers are related to each other in that specific dataset. For example, if a search term query data has an index value of 100 and another that is 50. The one with the value 50 had half the search interest as the one that scored 100. So, when compared to another dataset the index value 50 is not necessary half of that datasets index of 100. The first dataset might indicate 100 on over a million search terms as the second dataset could indicate 100 on the only on a third of a million search queries. The value 100 is only given to the maximum about of searches that the specific search term has had during that time period. Value of 100 could easily be the indicator of only 10 search queries if that is the maximum about of searches that the search term ever has had.

Kristoufek (2013) discussed in his scientific report about Bitcoin value forecasting that the search data does not divide the positive and negative search interest from each other. This matter was also brought up by McLaren and Shanbhoge (2011) because searching just out of curiosity without any actual interest creates significant background noise to the dataset. This can also fortify a negative agenda thus giving a false interpretation of positive interest (Henning-Thurau et. al. 2004). This is a troublesome disadvantage in a situation where the goal is to see if the search volumes correlate with positive customer interest in the assumption that positive interest creates sales.

McLaren and Shanbhoge (2011) debate that the correlation between demographic features and search engine users make the used sample uncomprehensive. This meaning that not all people are search engine users or in this case Google search engine users. This creates a misinterpretation of the customer interest since the non-search engine users are not included into the data. This problem makes the studied customer interest data (search query data) sample homogenous. Also, for the purposes of this thesis it is good to acknowledge the fact that the search engine user is most probably a highly educated and mid-earning young man (Deborah Fallows 2005). This means that the least search engine using sample is not a part of the customer interest sample which is studied.

Last limiting problem that was collected from McLaren and Shanbhoge (2011) is that the search query data is relevantly new and has a short backrun. This is at least when compared to other economic indicators that have had more time to evolved during time. This short backrun undermines the reliability of the data. The Google Trends data backs only up to 2004 (Rogers 2016) and all correlations studies beyond that are impossible to conduct. This limits the possibilities of the search query data. Also, the lack of data makes it harder to draw solid conclusions. This short backrun also makes it harder to remove possible seasonality from the data since Google Trends does not do that yet (Schmidt & Vosen 2009). It is not a problem at every forecasting situation but for example when evaluating

employment rates is should be considered and then the Google Trends data need adjustment to meet the expected variable criteria (Schmidt & Vosen 2009; Shimshoni, Efron & Matias 2009).

## 4.4 Discussion about Google Trends data benefits

The most obvious benefit of using search query data from Google Trends for forecasting is that the data is free to view, download and study (Rogers 2016). This is significant when compared to a traditional survey that might take a lot of time and resources from the company to conduct. Google Trends provides unbiased and a big sample of data possibly from the exact same thing that the conducted survey would be about but in current time and with minimum resources. Search query volumes also represent numeric data about general interest toward a topic. This makes it also easy for the company to evaluate results with other numeric metrics as well. (Schmidt & Vosen 2009)

Choi and Varian (2011) found in their research that search queries are useful and can be used as indicators for the consumer purchase interest. Especially in situations where the consumer evaluates the purchase before the actual decision. This way the query data is most effective for the company to use when the studied cause is something that has information about interest before the actual purchase happens. For example, information search is relevant especially before a car purchase but in this internet era information search can be included to almost anything from kitchen shopping to service providers.

Studies also show that search query forecasting is especially rewarding for examining private consumption. Search query volume- based indicator outperformed survey-based indicators in almost all in-sample and out-of-sample forecasting experiments. This is because survey-based indicators do not accurately capture the link between actual spending decision and what have been expected to be. Search data indicator avoids this because it gives out an unsupervised response from customer, potential customers and non-customers. (Schmidt & Vosen 2009) Furthermore, it has been studied that the search query volumes help to predict general volatility better in time periods where volatility is high and thus give more precise predictions. This enhancement increases the model reliability since high volatility has been problematic in forecasting models. With precise volatility predicting, it is possible to remove it safely from the data and/or study the volatility further. (Dimpfl & Jank 2012)

Google Trends can also help a company to see trends and features that they have not noticed before. Possibly a clear seasonality can be discovered and then used to accelerate production. Google Trend also provides locations in which that search query has been used the most which is also beneficial information for example for production goals as well. There is also a possibility to find product ideas through queries that are most likely linked with the company's products or the query in question. When it comes to competition it is easy to see if the neighboring companies' products get bigger volumes or if there exist content others use that gather attention around it. These are good starting point when increasing business or staying in the game. (Collins 2016)

# 5. Forecasting with search queries

In this chapter the goal is to find the relevant methods that previous literature has proven to be useful when handling and predicting with search query data. The aim is to scout out the appropriate methods and models that will be used to conduct this thesis' research. The chapter starts with data analytics and data modifying then moves on to the forecasting methods. Noteworthy information about the upcoming theory is that all in all there exists numerous different models to use for forecasting, however, due to the thesis' limitation the main focus is going to be on the most frequently used simple models that are suitable for search query data. The goal is also to handle models that would be the most suitable for solving the research question of this thesis.

## 5.1 Data analytics and data modifying

Before starting to construct any econometric model, it is important to make sure that the data is valid itself. This is done by cleaning the data from outliers, non-numeric values or from other possible distracting features that create distortion into the dataset and thus, to the model and results (Varian 2014, Da et. al. 2011). Specific feature of Google Trends data is that the values are given only in weekly indexes and this needs to be acknowledged and the other datasets modified according to it. In this thesis it is done by converting the dependent variables data into adequate form by also indexing the values, which is generally a preferred method. (McLaren & Shanbhoge 2011)

Data modifying is a common way to reassure compatible results. By dividing time-series data into separate periods helps to create in-sample period data and out-of-sample period data which can be used for parameter estimation and to test out the model performance (Schmidt & Vosen 2009). This was used by and also by Schmidt and Vosen (2009) in their research. Shimshoni et. al. (2009) modified their data by dividing the search query dataset into yearly periods. After the division they used historical data to forecast the yearly trends. Then they compared the forecasted data with the actual datasets to see if the forecast is worthy. For creating the forecasting model, they created an in-sample period data for creating the model and evaluated its performance based on the out-of-sample (test data) output.

If the search query data is not comprehensive enough by itself, the search query data can be used to create more exogenous variables. This method was used by Schmidt and Vosen (2009). They created exogenous variables to use in the regression by extracting known and

unobserved factors from the search query data. For extracting variables, the authors used unweighted least squares method. These artificially created exogenous variables have the potentiality to provide useful information to the researched through how the variables react at each given time of the dataset and with other variables. (Schmidt & Vosen 2009; Koop & Onorante 2013) Another good policyfor modifying the search query data in order to get desirable results is to create the Google predictor in the form of probabilities rather than including them as they are into the regression. This is a good method if the goal is to predict long-term results since the further the predicted period is the fuzzier the results will be and thus probabilities will be more helpful than actual numbers. (Koop & Onorante 2013).

In the case of forecasting with a classical linear regression (CLR) model there exist some unwanted data features that should be acknowledged and adjusted on the dataset before beginning. Especially time related features might be an issue since the search query data is time series. When using the CLR model the creation of the estimators demands data inspection since the estimator is created through the use of ordinary least squares (OLS) method. The OLS estimator needs to be the best, unbiased, linear estimator (BLUE) that is possible to extracted from the data. To make sure that the estimator is BLUE the CLR model needs to follow five assumptions, which are the following:

(1) The errors have zero mean

(2) The variance of the errors is constant and finite over all values of Xt

(3) The errors are linearly independent of one another

(4) There is no relationship between the error and corresponding x variate

(5) The error terms are normally distributed.

(Brooks 2008, 43-44)

These assumptions are now presented in the form of using CLR model, however, these assumptions are also present in many other forecasting models and hence are important to acknowledged before starting any forecasting.

## 5.2 Search query forecasting methods

The search engine forecasting is based on the idea that the search query data has predictive power that represents unspoken interest form the search engine users. (Ettredge et. al. 2005; Shimshoni et. al. 2009; Choi & Varian 2011). Goel, Hofman, Lahaie, Pennock

and Watts (2010) describes in their article this interest causality as: "…*it is a short step to conclude that what people are searching for today is predictive of what they will do in the near future*" and data journalist at Google describes this connection "*Examining what people search for provides a unique perspective in what they are currently interested in and curious about*" (Rogers 2016). It is conducive to see how this customer interest has been handled in econometric forms by researchers. Before actually performing the correlation study it is important to choose the right methods to do it. For the goals of this study the main data handling methods are collected together from similar studies. These methods and the authors who have used the method are collected together into Table 1.

| Data visualizing | Classic Linear Regression Model | Autoregressive models | Nowcasting |
|---|---|---|---|
| • Kulkarni (2010)<br>• Askitas & Zimmermann (2009)<br>• Choi & Varian (2011)<br>• Kristoufek (2013)<br>• Shimshoni, Efron & Matias (2009)<br>• Goel, Hofman, Lahaie, Pennock & Watts (2010)<br>• Varian (2014) | • Ettredge, Gerdes & Karuga (2005)<br>• Goel, Hofman, Lahaie, Pennock & Watts (2010)<br>• Ginsberg, Mohebbi, Patel, Brammer, Smolinski & Brilliant (2009)<br>• Varian (2014) | • McLaren & Shanbhoge (2011)<br>• Choi & Varian (2011)<br>• Goel, Hofman, Lahaie, Pennock and Watts (2010)<br>• Choi & Varian (2009)<br>• Kristoufek (2013)<br>• Da, Engelberg & Gao (2011)<br>• Schmidt & Vosen (2009)<br>• Dimpfl & Jank (2012) | • McLaren & Shanbhoge (2011)<br>• Choi & Varian (2011 & 2009) Castle, Fawcett & Hendry (2009) |

Table 1: SVI Forecasting Methods collected

The best starting point for searching variable causalities is to understand the bigger picture of the data through data visualization. Data plotting for example is a method that is used to examine the data behavior before adding any models into it. Also, different kind of graphs help to evaluate the data reliability for example by spotting out data distortion. The data plotting method was used by all of the authors named in Table 1, to pre-study their data and see how the data works after inserting it into the model. For example, Varian (2014) uses plotting to demonstrate the difference between the actual data, predicted data and model fit when observing website visit amounts.

From relevant literature it can be summarized that the most simple and useful model to use for forecasting is the classical linear regression model which is based on variable correlation and thus can be helpful for creating predictions and forecasting the near future (Goel, Hofman, Lahaie, Pennock and Watts 2010). Another quite commonly used and simple method is the autoregressive model. Choi and Varian 2011 support this argument by concluding their findings to: "*That simple seasonal AR model that include relevant Google Trends variables tend to outperform models that exclude these predictors by 5% to 20%".* The seasonal-AR model gains more predictive power when the Google Trends data is added as Google predictor for forecasting (Choi & Varian 2009)

Choi and Varian (2011) discussed in their technical paper that the more information (data) a research has for forecasting the more accurate the result will be. This is an obvious assumption, but the authors did add, though, that based on research it is possible to create a reliable google predictor even without having centuries worth of time-series data. This google predictor is noticed to be best and most helpful for present forecasting and the short-term forecasting which in other words is called *nowcasting*.

### 5.2.1 Data visualization

Visualization is an important part of econometric research and it is a good starting point because it helps to understand the whole data and how it is shaped (Varian 2014). Using graphical methods in the beginning of the data analysis can also help to detect possible unwanted feature such as heteroscedasticity, autocorrelation or outliers from the datasets. (Brooks 2008, 133; 140-141; 165-167; 230). It is also a common policy to visualize the outcome after fitting the model into the data. This is to see for example how the predicted data look against the actualized data and if it is trustworthy.

Data visualization is an easy way to avoid unnecessary pitfalls. When the data is visualized in appropriate way such at plotting the data or creating statistic bars it is possible to see what the data is capable of or if the data need some observation cleaning or in the worst-case scenario if the data is not suitable at all. These are situations that are good to acknowledge before starting the analysis because it saves time and it will reassure the results reliability. (Stedman 2012)

Especially now in the era of big data it is more essential to pre-evaluate the data before starting the analyzing and modelling. This is to sort out the unnecessary information and only focus on the wanted data. (Varian 2014) With the usage of big data there is more need for more advanced analytics where machine learning is becoming an important part. When

dealing with machine learning it is important to visualize also the output to see if that the models are performing as planned. (Rouse 2017)

### 5.2.2 Classical linear regression model

Classical linear regressions have been the most used models for predicting econometric phenomena (Koop & Onorante 2013). The CLR model can be used to create prediction but also in many cases the regression values work as the reference values for model variable evaluation (Ginsberg et. al. 2009). The basic regression is the most important tool of econometrics and also the easiest to comprehend, so, no wonder it is the most frequently used predicting tool (Brooks 2008, 27).

When using a regression model for forecasting the predictive assets of the Google search data are used by creating an explanatory variable from the search query data and adding it into the model in order to explain the dependent variable. The regression will then give out the significance and correlation between the dependent variable and the Google predictor. The idea behind the regression tool is to find correlation between the variables and by so help to understand the change that an explanatory variable has on the dependent variable. After finding the possible intercept and correlations it is possible to make linear predictions to future values. (Brooks 2008, 27)

For the regression analysis it is not unfortunately always this straightforward. The explanatory variable can be for example a collection of different features where the search query data is only a part whole data, or the dataset is too large to simplify it into one regression. (Koop & Onorante 2013) Since the most important task of the regression tool is to detect and summarize relationships. It is necessary that the analyzing tool can handle the data in hand. In a situation where the variables are not straightforward the future of machine learning becomes helpful. Mostly these machine learning tools help to summarize various sort of nonlinear relationships in the data. There are three features that work as indicators for the need of having more powerful regression tools for the data. First there is the size of the data that might demand more power from the regression tool to conduct results. The second feature is that there might exist more appropriate predictors than the estimator might allow and thus the help from a variable selector is welcome. Third feature links to the third one that the bigger the dataset is the more reliable the estimations are and thus it is good to aim for big data and not leave the power of the predicting tool get in the way of this. The machine learning tools can thus provide more flexible relationships than

simple linear models. Just an example of these machine learning tools are decision trees, support vector machines, neural nets and deep learning. (Varian 2014)

There is debate on whether or not Google variables are good for linear predicting (Choi & Varian 2011). Simple linear or logistic regression tool is the safe and secure way to handle a predicting problem but when datasets are big and complex more trained data analyze could be in order (Varian 2014). However, it is necessary to say that all analyzes are dependent on the data and the relationships of the variables and reshapes the Google predictors predicting power. For some macroeconomic phenomena the Google predictor is a perfect match and for some it works only as a directional estimation. One thing can be proven accurate is that Google variables are good for signaling turning points and model changes (Koop & Onorante 2013, Goel et. al. 2010). Some research also discusses about the collecting features that the Google search queries have. This 'collective wisdom' is integrated into the search query and can create, through regression, information about the studied econometric variables in different points of time. (Koop & Onorante 2013)

### 5.2.3 Autoregressive models

Since search queries create time series data it is expected that researchers would use the most popular stochastic time series models, which are the autoregressive integrated moving average (ARIMA) models or the solely the ARMA models, which are more suitable for univariate time series modeling (Adhikari & Agrawal 2013, 9; 18-19). The ARIMA model and especially the simple seasonal AR model are proven to be useful when handling search query data. This is because all of the models that have the Google predictor in them outperforms models that excluded the predictor. (Choi & Varian 2009; 2011, Schmidt & Vosen 2009) An example of this outperformance is that when Google predictors are included into the AR-model the forecasting performance increases the baseline model performance as well as the out-of-sample models performance (Schmdt & Vosen 2009). This outperformance also makes the autoregressive time series model better at succeeding in long-run predictions and also when the system is iterated forwards. (Dimpfl & Jank 2012).

When predicting volatility, the simple autoregressive models is not enough anymore and to gain better results it is better to use an AR model that understand model heteroscedasticity (Adhikari & Agrawal 2013, 18; Dimpfl & Jank 2012). The basic AR-model, even with search query data, managed to create models with a good fit and predict volatility accurately in calm times, however, the models does not provide information about future volatility. It is also interesting to see that when handling volatility, the model that includes the search query

data will outperform a univariate realized volatility model, but unfortunately at the same time the predictive power of the search query predictor decreases.

When using a model that understands heteroskedasticity for example HAR- model (heterogeneous autoregressive model) the model with search query data also manages to keep the search query predictive power across time. This model efficiency is because the HAR-models has been proven to be good at capturing long-memory properties of realized volatility. Need to refine that the HAR-models were chosen to be the outperforming models when using search queries to predict individual's interest towards aggregate stock market. Interesting outcome was that when the individuals interest rose to invest the volatility also increased. (Dimpfl & Jank 2012)

Another useful model of the autoregressive models' family is the vector autoregressive model (VAR), which is most suitable for situations when there is more than on dependent variable (Brook 2008, 209). The model is not equally popular when forecasting interest through search queries but is has been used by researchers at some level and it has given noteworthy and sufficient models to use for predicting (Dimpfl & Jank 2012, Kristoufek 2013). For example, Da et. al. (2011) used the VAR model to study the interest volatility of investors and found out that the effect of news and extreme returns increase the interest level.

Autoregressive models should be used when the dataset is large enough since the model creates prediction from its own lagged values, which means that the larger the dataset is the more useful the predictive model will be (Goel et. al. 2010). If there is considerable volatility in the search query dataset then an autoregressive model with heteroskedastic features should be chosen. Since the model selection is not all black and white there is a good possibility that the predictive power increases when combining models in orderly moderation (Goel et. al. 2013)

### 5.2.4 Search query data in Nowcasting

McLaren and Shanbhoge (2011) raised the topic of nowcasting in their paper. The definition of nowcasting is to predict the present hence the combined name 'Now' and 'Forecasting'. Previous research has been focusing mainly between search volumes and brick store sales, but the focus has been shifting more on to 'predicting the present' thus seeing how well the search volumes correlate with contemporary phenomena (Goel et. al. 2010). The use of search query volumes for nowcasting or short-term forecasting has been an investigated

topic in the current literature and the topic has gained interest due to the positive results (Koop & Onorante 2013; Castle et. al. 2009; Schmidt & Vosen 2009).

Nowcasting is not a forecasting model itself but it is essential to bring up since almost all of the used models are trustworthy and suitable in short-term predicting. This is not a negative aspect since currently the ability to forecast the present is becoming more important. Castle et. al. (2009) addressed the reasons for the need of nowcasting in their article: “Nowcasting is not just Contemporaneous Forecasting”. The main four reasons are the following:

1. The first and most important reason for the need of acknowledging and using nowcasting is that the obtained data is almost never time accurate. This means that the collected data is usually published with a time lag and thus does not represent the current day when the forecasting is made (Koop & Onorante 2013).

2. The second reason is that economic time series are only a glimpse or a flash estimate of the phenomena. This means that the data is in interconnection with the current time and when the data is used with a lag then it represents potentially the past environment. This is problematic when predicting long-term forecast, the data is correlated with the context of issues that happened at the time when the data was collected. These features have the possibility to decrease the data reliability the further in time the forecast goes.

3. The third reason is the inconsistency of the subsets, which are problematic. This means that since the time series data is computed from different components to form a systematic model these components should be available through different periods, but this does not always happen thus leading to ‘changing dataset problem’. Thus, leading to the situation that the closer the forecast is to the actual time period of the data the better the forecast is and in many cases the data lag bound the forecast to turn into nowcasting.

4. The fourth reason for creating a nowcast is to find early warning signs even if the timely current datasets are available. If the nowcasts are already significantly different from the measured series, the nowcast values work as an evaluating sign that the measures wont most probably work on the long run. So, when predicting a nowcast it might help to check the used datasets timeliness and see how lagged they are according to the current time.

Forecasting is most often used to predict private consumption which is 70 % of the US-GPD. When measuring GPD and private consumption the time is always in the essence of the process, but GDP (in the US) is published only once a month which inevitably creates a measurement lag of possibly one month into the data. With Google Trends data there is no significant lag in the data and thus it will help to create an accurate nowcast with the lagged data. Nowcast can be very helpful when great uncertainty, volatility and unique shocks are at hand because all of these three are features that make it very difficult to do long(er) time prediction. It is desirable to have accurate forecast for the current moment since the nowcasts can then be used to support another forecast that can predict a period further. In these situations, having a search query data assisted model can be particularly useful since past value indicators lose their predictive power and the Google predictor assist them to be used accurately once more. (Schmidt & Vosen 2009)

Nowcast can be done without including Google predictors into the forecast but the Google predictor does give additional forecasting power when nowcasting conventional monthly macroeconomic variables (Koop & Onorante 2013). When creating a nowcasts with the help of the Google predictor it has been noticed to create valuable information to understand future volatility. Valuable pieces of information about volatility can be in many cases help to understand the current time and possibly know how to react against volatility. (Dimpfl & Jank 2012)

Google variables have been used as regressors to indicate possible correlation but Google variable regressors have been pointed out to be useful for choosing the optimal nowcast model in each point of time. This is quite smart since a search query represents a certain time period. It is possible to see from the search query correlation which nowcast represent which time period and the best correlated values will increase the models' accuracy when preforming the forecast. This method has been proven to be useful when using dynamic model selection method (DMS) for finding the best nowcast for specific macroeconomic variables. The usage of Google predictor in the dynamic model selection outperformed the conventional DMS most of the times. The DMS models are very useful since the Google predictor has been proven to be useful at most of the time but at sometimes it is still found to be unnecessary (Choi & Varian 2011) and by deploying the DMS the periods can be found when the Google predictor increases the predictability. (Koop & Onorante 2013)

# 6. Methodology

The aim of this sixth part is to see how well the bargain sale variable and the Google predictor correlate together. Also, it is necessary to see if there exists predictive power in the Google search history toward sales forecasting. From this possible correlation simple forecasting models are used to see short-term prediction accuracy. The thesis does not aim to create an exhaustive predictor or model. More important is to validate the predictive power of the Google predictor towards the bargain sale outcomes

The methodology part will start by explaining the research background. The dataset structures and characters will be introduced right after this. Both the bargain sale data and Google Trends data limitations will be acknowledged in the structure parts. After this there is the data analysis where the actual forecasting will be conducted. The analysis part will start with an evaluation of the dependent variables. This will be done through a regression analysis. After this the forecasting model will evaluated. Naturally after this there will be a forecasting model for all of the sufficient dependent variables and a forecast can be created. Very last of this thesis is going to be the forecast results.

## 6.1 Research background

Based on the previous research it seems very plausible that the search query volumes might be able to predict the bargain sale outcomes. Putting aside previous literature for a while the main supporting reason to believe that Google Trends search query data could forecast the bargain sale outcomes came from the search data graph (Appendix 1). From the graph it is easy to see how the weekly index for search query volume peaks uniformly. When looking at the timeline on X-axis the peaks happen twice a year. One peak around spring time and one peak around autumn time. This is quite of a coincidence since the bargain sales are held twice a year, one in spring and one in autumn. Some of the peaks differ significantly in size from each other. This has some strong indications that there might be some actions that have affected the size of the index value. From the graph it seems plausible that changes of the peak sizes could be the result of changing customer interest towards the bargain sale hence plausible to be predicted through the Google predictor. The second supporting reason is the shape of the sales outcomes (appendix 2). The bargain sales outcomes follow somewhat similar increasing shape as the search query volumes graph. This means that there is an increasing trend in both, the search query data as well as in the bargain sales data.

Due to the increasing interest towards big data analysis it makes it important for the company to know how to take advantage of free, online information sources. The search query data is only a tip of the iceberg when it comes to analyzing data from online, however, it is a very insightful and easy beginning point. The more there is research about the online data the more there will be efficient analyzing tools to help the company. This is another reason that gave meaning to conduct this research.

In figure 3 there is a simplified presentation of the causality between the bargain sale data and the search query data. The actual sales forecast represents solely itself and is the main target of the study. If you can forecast the actual sales, you can optimize all other company actions. The budgeted sales data proxy as the company actions. Since the budgets are created through company decision and resources the dataset is a comprehensive representation of the actions that the company will carry out in order to obtain the budgeted sales. The budgeted sales are presumed to represent all company actions since it would not be fruitful to study all company actions singularly. This way it is possible to get a clear view of the company actions and if they are indeed a part of the process of creating the customer interest and thus search query volumes. The search query volumes, as mentioned, stand for a measurement for the customer interest since customer interest creates search query volumes. The company actions are expected to influence the customer interest and through the customer interest it is expected to create search query volumes and therefore sales. When examining the company actions and the customer interest it is possible to forecast the actual sales.

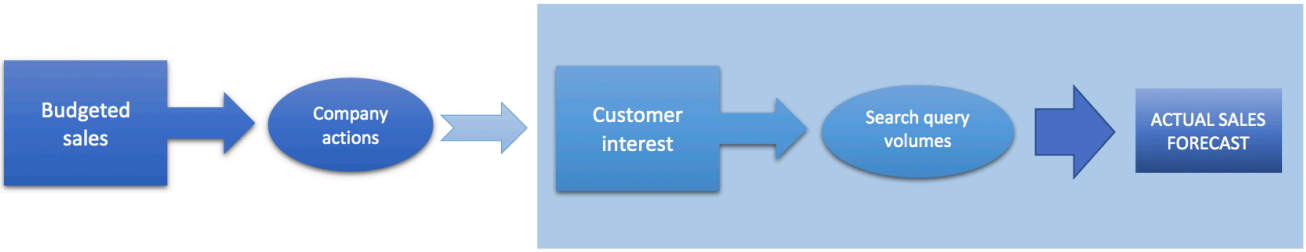



Figure 3: Causality map of observed datasets

## 6.2 Bargain sale data

In this part the main bargain sale data characteristics are explained. The part starts with a description of the data structure and then moves on to data limitations. In the data limitations there will be limitations that impacts the usage of the data and also the interpretations of the forecasting results. The data is obtained through the bargain sale arranging case company. Six different datasets were obtained. These obtained datasets are theactualized outcomes and the budgeted outcomes of the bargain sales from each spring and autumn period. Since the search query data is time-series data the time periods of the bargain sales are also included into the data. This is to understand the trend lines and the evolvement of the data as time passes. The company performed marketing budgets are also collected since they might possess vital information for further understanding about the research results.

### 6.2.1 Bargain sale data structure

The bargain sale duration is one week, and the observations of the sales are weekly sums. This induces that there will be one observation per bargain sale. This means that there are ten different sale outcomes to study. To create reliability to the research the sale outcomes are also studied together but mainly as separate subsets. These subsets are online store outcome, brick store outcome and total outcomes. From all these subsets there are the actual and the budgeted sale outcomes. The subsets are studied individually to see if some of the them correlates better with the Google predictor than the others. To make the bargain sale easier to evaluate with the search query data the bargain sale data is also changed to index form. This indexing is done the same way as for the search query data.

The bargain sale, indexed subsets can be observed from Figure 4. All of the subsets differ from each other, however, there seems to be similar shapes in all of them. There is an increasing trend all the way to ‘Autumn 2016’ and ‘Spring 2017’ when there is a small drop in all except the budgeted total sales. It is interesting to observe when the variables hit the index value 100. The outcomes hits index 100 the first time in ‘Spring 2016’. The budgeted total (BUD total) stays after that close to the value 100. This could indicate a shift in the company expectations. In ‘Autumn 2017’ the actualized outcomes pass their budgeted values and actual total and actual online sales hit the index value 100. This increase can be considered as a lagged outcome from the company actions.

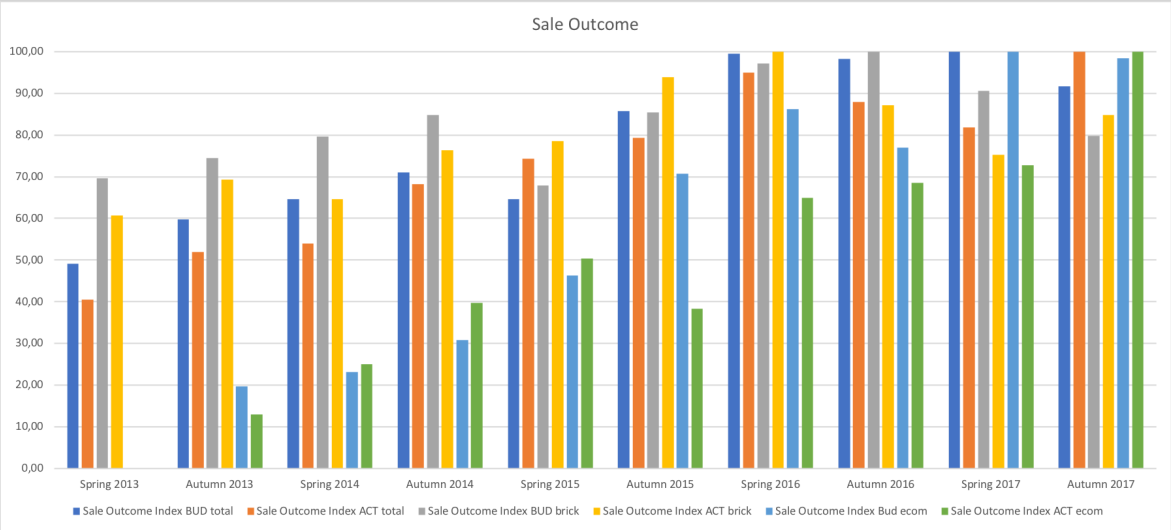



Figure 4: Sale outcomes subsets together (indexed)

The budgeted sales proxy as company actions, which means that the budgeted sales are also observed from the causality direction. In appendix 6 all the budgeted sales and actualized sales are plotted against each other. The shapes are very coherent. This would suggest that the case company is very close at predicting the actual sales. Especially for total sales and brick store sales the budgeted sales are very coherent with the actualized sales. The total budgeted sales are in size most accurate when the brick store sales are better at following similar shapes. This would mean that the company actions are working in order to obtain the budgeted sale amounts. The only curiosity are the values after 'Spring 2017'. Budgeted sales are expected to drop but in fact the actual sales grow over the budgeted line. Hopefully the SVI will give some extra predicting power to this unexpected shift.

A part of the bargain sale data are also the marketing budgets that were made for each bargain sale. The marketing budget are examined in order to further understand possible changes and differences in the actualized and budgeted sales. It is obviously expected that the marketing budgets have impact towards the actual sale outcomes and also, they are a significant part of supporting the budgeted sales realization.

If after the data analysis it seems that the search query volumes do not give significant results or there exist turning point that are hard to understand, the research can turn to see if the marketing budgets might have an explaining impact on the actualized sales. The marketing budget data values are also indexed and are summarized in figure 5 together with the actualized sale outcomes. The marketing budgets includes offline-media prints, radio and outdoor advertising.

As a presumption is that it is obvious that the budgeted outcomes would correlate with the marketing budgets since the marketing budget is to support the budgeted sales. This is why the budgeted outcomes were not plotted against the marketing budgets since both variables are company created numbers, so the gained information from the graph would not bring any essential knowledge into this research.

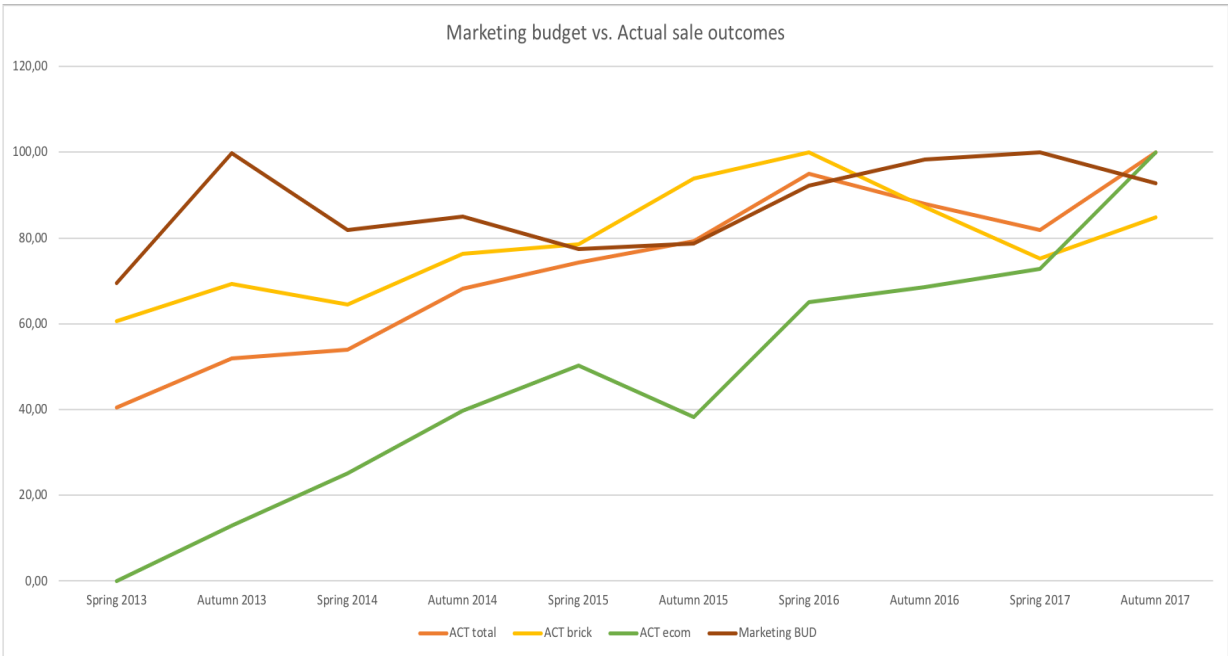



Figure 5: Marketing budget versus Actual Sale Outcomes

When observing the marketing budget against the actual sale outcomes (Figure 5) there is a slightly increasing shape. Until 'Autumn 2016' the marketing budget and the sales outcomes peak and drop quite similarly, however, the marketing budgets shift more radically up and down. The sale outcome and the marketing budgets have some similarity which would indicate that the marketing budget might be explanatory about the actual sales. It would be only reasonable that the increase in marketing budgets would help to gain more customer attentions and thus lead to larger sales. It is only positive and expected that the marketing budget and the actual sale values move consistently with each other. Based on figure 5 it would indicate the obvious that the marketing budget have causality towards the sale sizes

### 6.2.2 Bargain sale data limitations

The bargain sale time scope was chosen mainly to be able to follow the brick store success and the online store success simultaneously. The online store shopping possibility was added to be a part of the bargain sale in 2013. Because the search engines correlate well with the digitalization it was thought important to have the online store (ecom) included into the research. Another factor affecting this decision was that the data from Google trends was more valuable to study between years 2013 and 2017. This time period showed the most significant peaks that could be examined. The bargain sale is held twice a year which means that there are two summed values of the sale to study per year. These decisions resulted to the outcome that there are in total ten bargain sales under observation.

The data size is the most concrete limitation since the results can be only considered directional since there is not a lot of previous data to create forecast values. However, there would be no amount of data that could give exhaustive results since forecasting is always, in reality, estimating the future values. But in general, the bigger the data size is the better and reliable the results are (Choi & Varian 2011). Nevertheless, the goal of this thesis is only to find out the forecasting possibilities behind the Google predictors and the bargain sale data, not the exhaustive prediction model. However, as mentioned in the structure part to solve this data scarcity the bargain sale data was divided to actual and budgeted total sales, brick store sales and online store sales. This way there is more ground to evaluate the data and give more precise results about which sale outcome is the most correlated with the search query data and what these correlations might tell about the forecasting possibilities.

## 6.3 Google Trends Search Query Data

This part will describe the search query data characteristics. The part will start by explaining the dataset structures and then move on to the limitations. The limitations will go through features that limited the structure of the search data but also limitations that might affect the research conduction and results.

The observed query was chosen based on intuition and testing. Same way as McLaren and Shanbhoge (2011) chose theirs. Different kind of relative queries where examined but the only query that gave significantly more observations and volatility was the chosen one (the company name + bargain sale name). Google Trends has the ability to provide related queries when searching for query volumes (Rogers 2016). However, all of the related queries where the same as the observed one only with an added year in the end. This

reassured the decision to go with the current query since there is no related queries that should be included in the research. Thus, the bargain sale name was chosen to be the observed search query.

Assumption that is made about the data is that all consumers searching the chosen query are interested in the bargain sale one way or another and thus supporting the assumption that the search queries represent consumer interest towards the bargain sale. The search query volumes are only seen as numbers, but these numbers have been created by actions. They have been studied to be the creation of navigational, informational, transactional and resource searching goals. All these queries have the name of the bargain sale in them and thus they are included into the search query data. These are search queries such as 'bargain sale website' and 'where can I buy this brand on sale', 'when is *bargain sale name* held' etcetera. These represent the need and goals of the potential customers.

The concept of the bargain sale and the sales name is very limited and thus can be assumed to be representing only itself in the search queries. The query is expected not to indicate anything else than what it is when entered to the search machine. Also, since the query is very specific it constrains the individual searchers to people who are actually interested in the topic in hand. This is mentioned in the previously handled assumption that the search queries would present more positive interest rather than negative (see Kristoufek (2013) theory about negative and positive interest). This is another assumption made about the search query data that it would represents positive customer interest towards the bargain sale. If it would not represent positive customer interest, it would be very curious to see how negative interest can create sales. However, in order to keep the thesis straightforward the sales are expected to increase through positive searching interest.

### 6.3.1 Search query data structure

The emitted search query data is time-series data and the search volumes are presented in index form (see keyword SVI). This index form forces the data to be studied solely together since the indexes are created in a way that they are dependent on each other's values and the indexes indicates distances from one value to another. This means that the dataset cannot be studied if the time-series from 2013 to 2017 would have been obtained separately from Google Trends. For this study it is important to state that the focus is only on the entered search queries to the Google search engine. The study does not count the search results that the users have received as an outcome from their search query request. This is because the search result of the search query is irrelevant because the core focus

is on the means of expressing interest toward the search query by doing the action of entering the query.

Search word queries are unbiased in a sense that the results are not affected by any experimental atmospheres. The data is created by single users without any surveillance and thus the data can be considered as true to the reality as possible hence the SVI will provide a good insight on the amount of interest toward the search query phenomena at a given time period. The original form of the SVI data can be seen in appendix 1. In the appendix-figure it clearly shows peaks two times a year and probably not so accidentally the peaks are all around the same time as the bargain sale dates. To see how well the peaks sizes correlate with the bargain sale data some modifications to the search query data were necessary.

The actual search query data was downloaded 29th of January 2018 from Google Trends. The data is from weeks between 3.2.2013 - 24.12.2017 and is consisted from 256 SVI observations. Since the bargain sale data is consisted of ten sales and thus ten weeks, the search query data is divided into ten periods as well. In table 2 there is a precise division of which search query periods are included into which bargain sale. This was done since the weekly data was not compatible to be compared with the bargain sale data in its original form. To study the interest toward a precise bargain sale the highest search volume index was emitted from that search index period and acknowledged as the search query volume of that sale. For example, sale of 'Spring 2013' has the maximum search index value between periods 3.2.2013 to 30.6.2013.

| Bargain Sale | Weeks included |
|---|---|
| Spring 2013 | 3.2.- 30.6.2013 |
| Autumn 2013 | 7.7.-29.12.2013 |
| Spring 2014 | 5.1.-22.6.2014 |
| Autumn 2014 | 6.7.-28.12.2014 |
| Spring 2015 | 4.1.-28.6.2015 |
| Autumn 2015 | 5.7.-27.12.2015 |
| Spring 2016 | 3.1.-26.6.2016 |
| Autumn 2016 | 3.7.-27.12.2016 |
| Spring 2017 | 1.1.-25.6.2017 |
| Autumn 2017 | 2.7.-24.12.2017 |

Table 2: Search volume index division per bargain sale week

In appendix 3 (Weekly SVI per bargain sale week) the search query data is graphed separately depending on which bargain sale the search query data is included. It is interesting to see that the peaks are almost separate from all of the other variables. The difference between the peak and the values before is almost from 0 to 100. This is a strong indication that the company is doing something right to gain so much more attention around the bargain sale. This might be the reason of a good marketing allocation or integrated company actions or just due to customer loyalty. However, based on the graphs the structure of the search query data is pretty clear and potential to represent customer interest toward the sale.

### 6.3.2 Search query data limitations

From the theory about search query data it is clear that the data has certain limitation and some of these limitations are also present in the studied search query- data. Main limitations for the search query data in hand are the following:

- Geographic differences
- Only weekly indexes available
- There is no separation between negative and positive interest
- Synonymic keywords and misspellings are not included
- Dataset only includes interest from Google search engine users

First and foremost, limitation that was made to the data at the point of obtaining was limiting the entered search queries only to the region of Finland. This was because the bargain sale is held only held in Finland. Also, when comparing Finland-region search query volumes with worldwide volumes the difference was insignificant but to be sure the limitation was made.

The Google Trends data is only available on a weekly level and as mentioned before the data is in index-form. This means that we have only a certain mean value for a weekly search amount. This amount is in index form, which also limits the research since there is no possibility to see the actual day to day search amount. Hence the result will also be in approximate amounts. The result will then be more directional rather than factual.

Another problem which will be strongly present throughout the research is the problems that the studied search query data does not show negative and positive interest a part of each other. So, if the volumes are high because for example a scandal around the company it

will be accidentally assumed to be positive. This problem has been tried to avoid by choosing the bargain sale name since it has not had any public negativity around it yet. Limitation that is related to this same problem is the power of marketing. It is possible that a peak in search volumes are solely based on effective marketing. However, this will be handled just as positive interest towards the sale and thus may just gain more customers to the sale. It will be a separate study to see how well search query volumes can predict marketing campaign success.

Possible attribute to lower the data reliability is that query synonyms and misspellings are not included into the dataset. This is difficult to pass-by since the sale is still in quite a small scale in the scale of general worldwide query volumes. Since the sale is only in Finland and the data is studied in Finland there are no relevant keywords associated with the sale to include into the datasets. Also, language is difficult since the sale does not have an official English version name thus a non-Finnish speaking person can use any kind of synonym to find the sale. In theory part about 'Online Search engine' there was an important mention about the possibility of having more search engine users than shown in the search query volumes. These users are the ones who do not use the specific query that this thesis is looking into. The evolvement of the Google search engine it might be that people searching just for a topic of 'Finnish bargain sale' or just misspelling the bargain sale name might get the right answer on the first try and then the user won't rewrite his or her query to the correct form. This means that the potential customer never has to use the search query of the bargain sale name thus never creating a hit to the observer search query.

The dataset only has interest from Google search engine users. This means that there is going to be some kind of biasness in the user sample. Since it is not possible to assume that all customers use solely Google search engine or if at all search engines, there will be certain undefined limitation in the users who creates the search query data. For example, it can be that the main bargain sale buyers are the opposite from the main search engine users. This information should be acknowledged when evaluating the sample that represent the customer interest. The sample will not be exhaustive to include all the actual bargain sale customers.

## 6.4 Data analysis

In this part there is going to be an evaluation of forecasting models and then finding the model to predict bargain sale outcomes. The part starts with an introduction about the phases to find the optimal forecasting tool. Then the part moves on to validation of the model predictor. After this there will be a brief conclusion of the study findings. More profound discussion about the result will be in chapter 7.

### 6.4.1 Visualizing the datasets

There are six different datasets in the bargain sale data that can be considered as different variables and can be studies as the dependents. When observing the data visually there appears to be an increasing connection between the search volume index values and the actual sales outcomes as well as the budgeted sales. This can be seen from figures 6 and 7. Based on the figures it would seem that the SVI correlates better with the actual sales-values rather than the budgeted ones. This is because looking at the figure 6 of actual sales there is a more coherent increasing pattern that follows the same path as the SVI. The actual values are a bit higher than the SVI values. The highest peak of the actual sales comes with a lag when compared to the SVI. When looking at the budgeted values from figure 7 the budgeted sales are more stable around index value 80. Only exception is the budgeted online store (BUD ecom) sales which is by glance the most accurate with the SVI values.

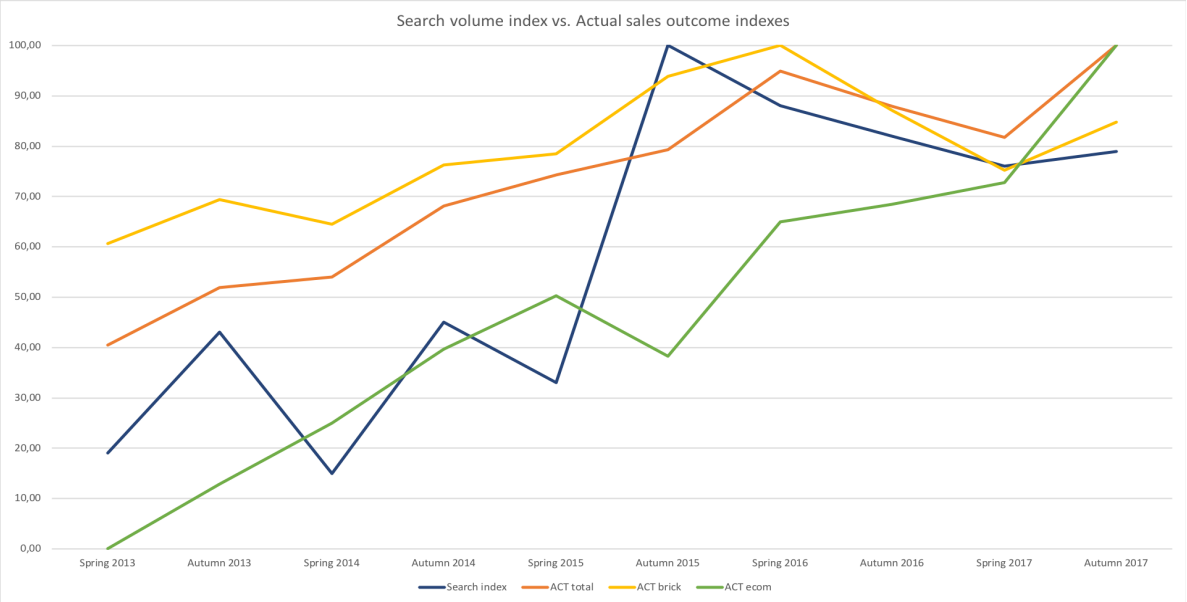



Figure 6: Search volume index vs. Actual sales outcome indexes

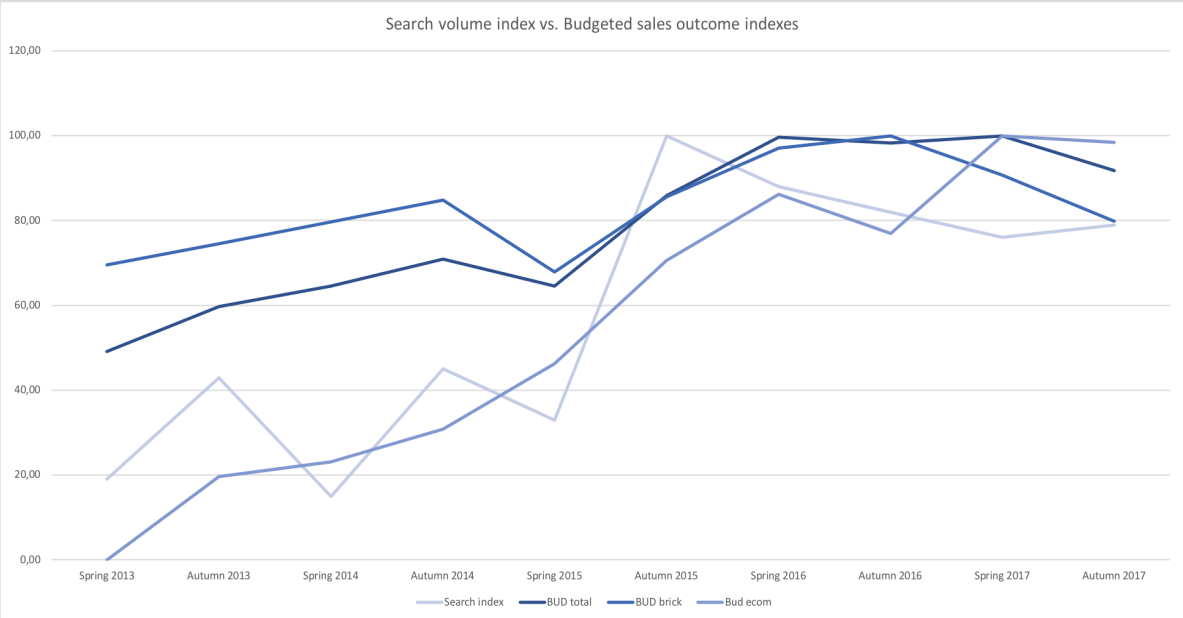



Figure 7: Search volume index vs. Budgeted sales outcome indexes

From the information obtained from the data structure investigation the search query volumes were also plotted against the marketing budgets in Figure 8. This was done to see how well the marketing budget would correlate with the customer interest rates. The main ups and downs are quite similar until 'Autumn 2015' when the peaks start to be the opposite form each other. However, there does not seem to be any indications that the marketing budget would affect the search query volumes.

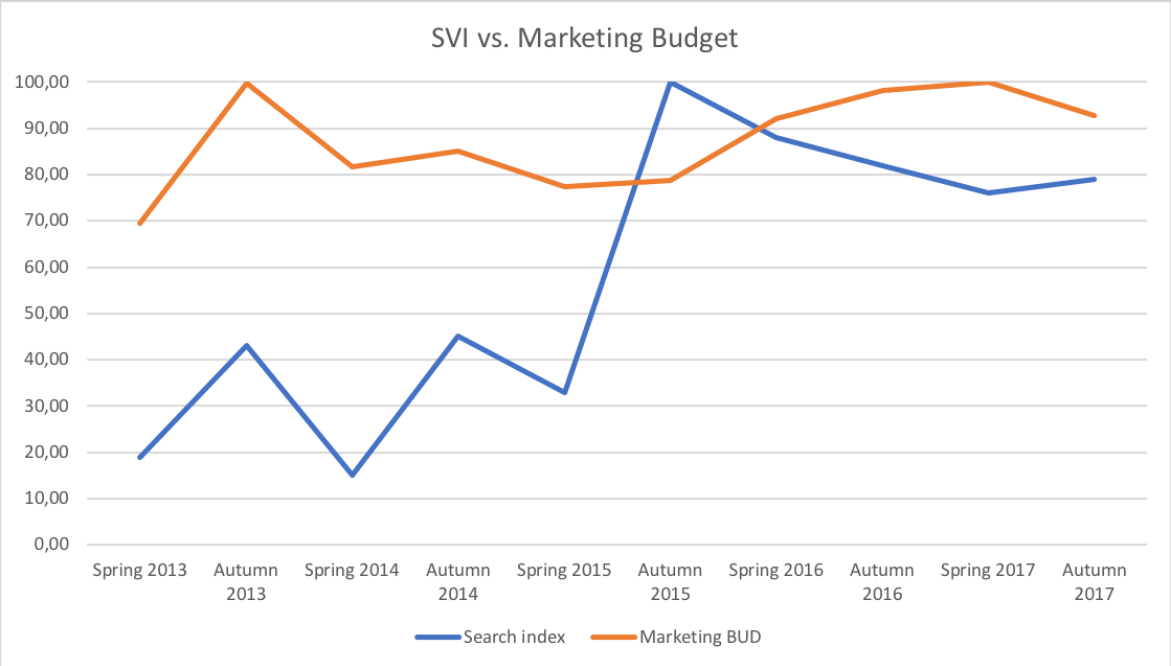



Figure 8: SVI versus Marketing Budget

These figures (figure 6, 7 and 8) give good insight of the upcoming possibilities for predicting and therefore also reinforces the research to be continue with the current data and use it for forecast modelling. The data structure suggests that the forecasting will be done through classic linear regression model. Through the modelling it is possible to see which one of the dependent variables are actually correlating the best with the search query volumes and thus, are efficient for forecasting.

### 6.4.2 Linear regression analysis

The bargain sale dataset is small and only one subset (dependent variable) will be under observation at a time. The search query volumes will be the explanatory variable and it is called the Google predictor. Time is important feature in both of the datasets. Since the datasets are simple and small the most appropriate model would be the classical linear model. The research will start by seeing through linear regression results and by so see which of the bargain sale subsets correlate the best with the Google predictor.

Using Microsoft Excel, the regression data analysis was performed individually to all of the actual and budgeted sales outcomes to see which of them holds the strongest connection with the search query data. The performed regressions were in form *sale outcome ~ SVI + error term* and the results were coherent. All gained coefficients were significant since the significance level of the F statistics were always under 0.05 (Appendix 4 and 5). R-squared values had more difference between subsets. Nevertheless, all of the values except actual online-store (ACT ecom, $R^2 \sim 0{,}45$) and budgeted brick store (BUD Brick, $R^2 \sim 0{,}52$) showed correlation around 70%. The highest explanatory power was for budgeted total (*BUD total*) which was 77,4%, however, actual brick store (*ACT Brick*) was also very good with a fit of 77%. Close to second was the actualized total sales (*ACT total*) with $R^2 \sim 0{,}68$.

The marketing budget was also expected to be correlating with the Google predictor. To be sure a regression analysis was performed where the SVI was explained with the marketing budget. The regression form was *SVI ~ marketing budget + error term*. As already notices the results were equally insignificant as the graph gave to expect since the explanation level was only 23% ($R^2 \sim 0{,}23$). It is positive that the marketing budgets are more coherent with the actual sales, however, it is unexpected that the marketing budget size does not affect the search engine query amounts. This would have been an obvious assumption to be made in the beginning of the research.

From conduction of a simple regression there seems to be a connection between two dependent variables: BUD total and ACT Brick (Appendix 4 and 5). The budgeted values

represented the company actions toward sale generating. This makes it interesting to see that the Google predictor correlates the best with the total budgeted sales. This could signal that the company actions and goals are consistent at creating customer interest and hence search query volumes. This would give hope that the company expectations and the level of customer interest are on the same level. The variable ACT Brick (brick store) is the most interesting to study since it is the best correlated actual sales value. Since the actual sales are the forecasting target the study will also include the variable ACT total (actual total sales) into further research. The variable was chosen because it had $R^2 \sim 0{,}68$ (Appendix 4) which is still in the level of being significant. The actual sales values are the most important to study further since they are the unknown values for the company and forecasting them will help out on budgeting and strategy creation.

### 6.4.3 Residual analysis

The classical linear regression model is not always the most appropriate model for the data which means that the appropriateness of the data needs to be tested and this can be done through a residual analysis. When performing a linear regression model there are some assumptions made towards the variables and their relationships. To reassure that these assumptions are not neglected it is good to analyze the error terms of the model. Since residuals work as the estimations for the error terms and they are easily obtained it is valid to analyze the residuals that are created during the linear regression model.

The OLS assumption that need to be verified are the following: (1) the errors have zero mean, (2) the variance of the errors is constant and finite over all values of Xt, (3) the errors are linearly independent of one another, (4) there is no relationship between the error and corresponding x variate, and lastly (5) the error terms are normally distributed. (Brooks 2008, 43-44) In other word the goal is to have weak exogeneity, linearity, standard variance (homoskedasticity), independence (autocorrelation) and the non-existence of multicollinearity. The assumption (3) will be discarded since there is only one independent variable in the model which means that there cannot be multi-collinearity in the model. Since dependent variables for actual brick store sales and total sales had the highest correlation with the Google predictor and they are the values that need to be forecasted, their residual will be studied next. In real life none of these assumptions cannot be fulfilled perfectly but the goal is to have values close to perfect.

Before going to the first official assumption it is good to see the nature of the connection between the dependent variable and the independent variable. This referring to the situation weather there is a linear or non-linear connection between the two variables because this

will determine the how the variables are included in the model. In figure 9 and 10 there are plots between the dependent variables (ACT Brick and ACT Total) and the Google predictor. In figure 9 there is the actual brick store sales (ACT Brick). With the help of the trend line there can be seen a clear linear growth. In figure 10 where the actual total sale (ACT total) is plotted against the SVI the connection also seems to be linear. This means that there is no need to modify variables in the model to include non-linearity. The regression model can be held the same.

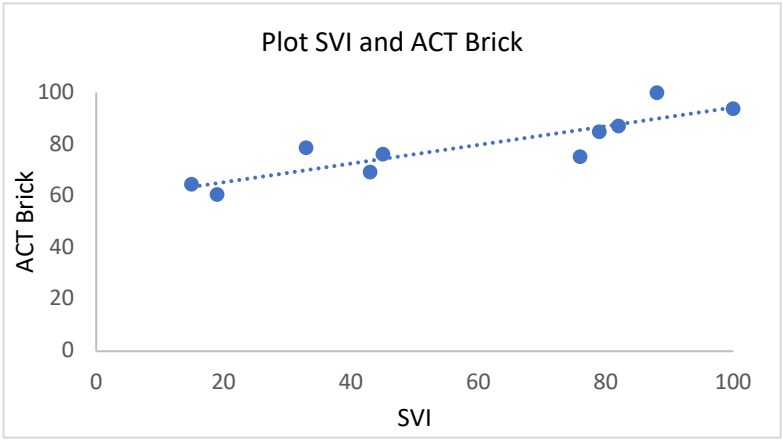



Figure 9: Actual brick store sales plotted against SVI

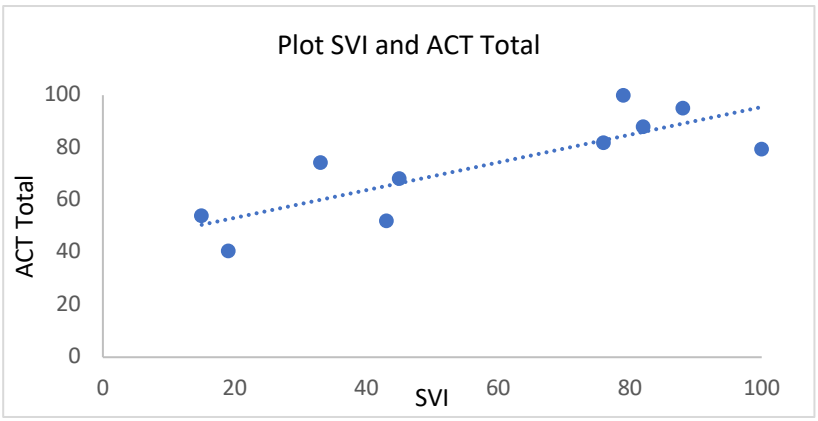



Figure 10: Actual total sales plotted against SVI

The first assumption was that the errors to have a zero mean. For actual brick store sales, the residuals do have a zero mean or at least very close to it with a negative value og 0,0000000000000178. For the actual sales the residuals also had a mean value very close to zero (0,0000000000000057). This means that both of these dependent variables pass the first OLS assumption. (Appendix 4)

The second assumption reassures that the variance for the error term is constant. In other words that the error terms are homoscedastic. This can be examined through the residual plots (figure 11 and 12). If heteroskedasticity is present the plotted residuals would perform a shape that systematically changes as the values increase. When looking at figures 11 and 12 there is not systematic shapes to be noticed. Of course, the reliability weakens because of the small dataset but based on this there is no significant heteroskedasticity to be found.

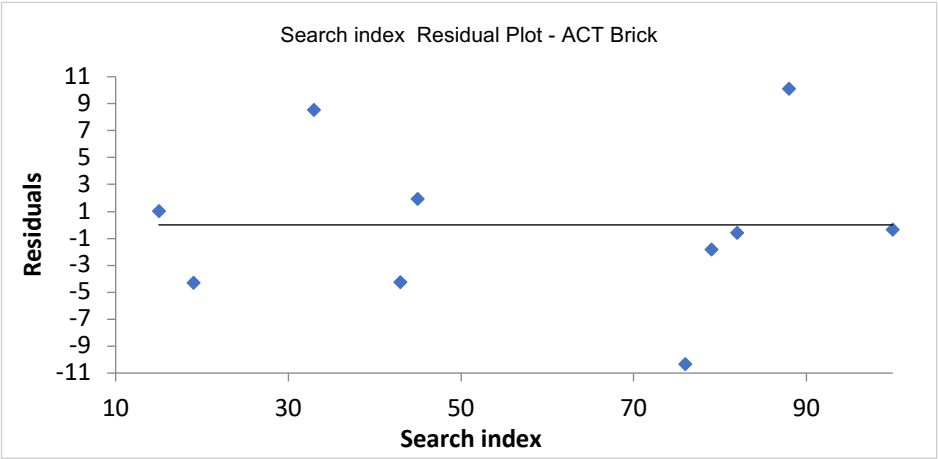



Figure 11: Residual plot from ACT Brick dependent

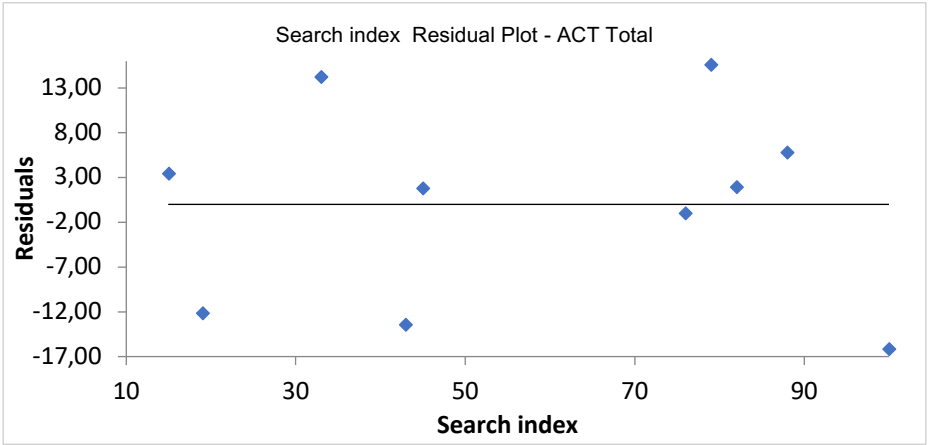



Figure 12: Residual plot from ACT total dependent

The fourth assumption reassures that there is no relationship between the error and corresponding x variate. This can also be checked from figures 11 and 12 where the residuals (error terms) are plotted against the corresponding variate (search index). When looking at the plots and the trendline there is no linear connection between these two variables. This would verify that the fourth assumption is assured.

The OLS assumption of 'no autocorrelation' says that the error terms of different observations should not be correlated with each other. With time series models it is very common that the following value is somehow dependent on the previous values and thus the rule of autocorrelation will most likely be violated. However, to be sure it is good to check the residuals plotted against time. This is done in figures 13 and 14. If autocorrelation would be present there would be repeating patterns, however, from looking at the figures there seems to be no autocorrelation in neither of the residual plots.

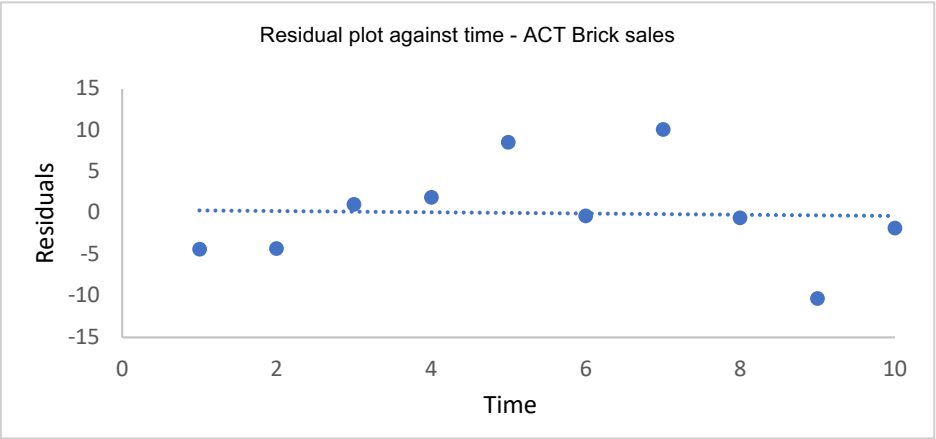



Figure 13: Residual plot against time (ACT Brick sales)

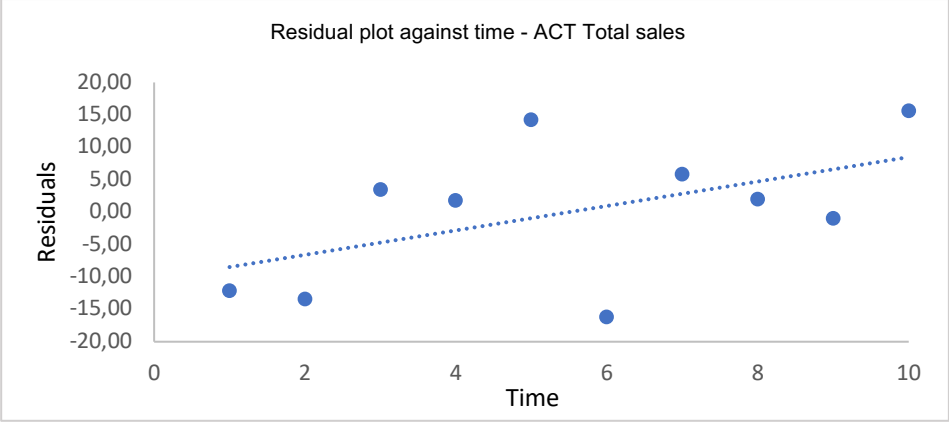



Figure 14: Residual plot against time (ACT total sales)

The last OLS assumption is that the error terms are normally distributed. This can be easily check through the normal probability plot. These plots can be seen in figures 15 and 16. The residuals are plotted against the theoretical normal distribution and in order to be normally distributed the residuals should form approximately a straight line. Both of the plots seem to form approximately a straight line and this affirms the last assumption.

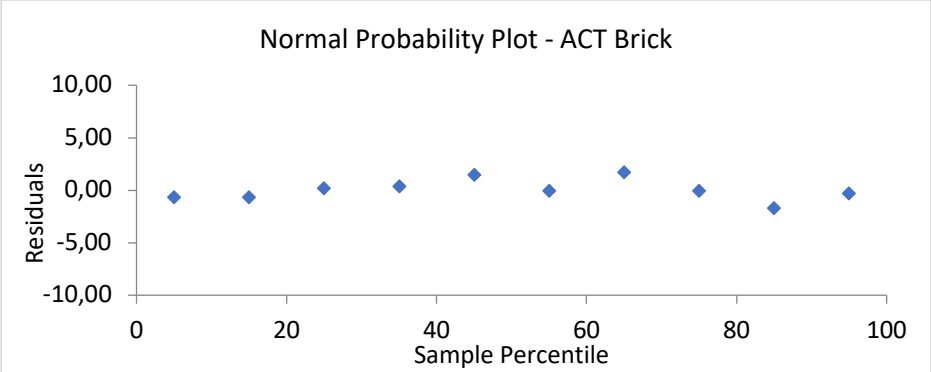



Figure 15: Normal Probability plot for Actual brick store sales residuals

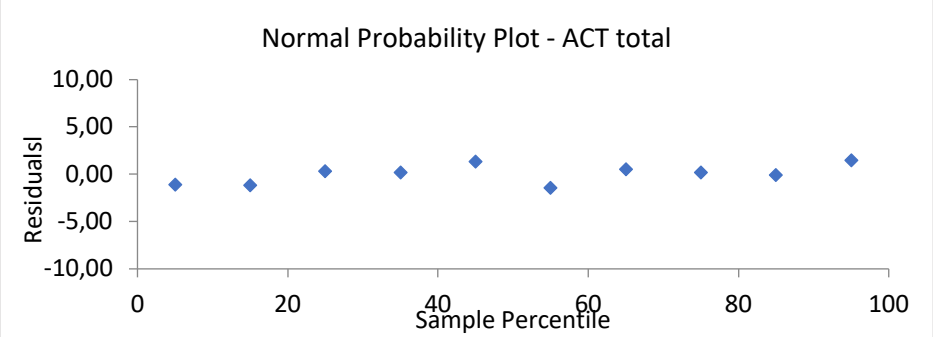



Figure 16: Normal Probability plot for Actual total sales residuals

All OLS assumptions were verified. This means that the estimator is BLUE and the regression analysis is valid. This then concludes to the fact that the research can be continued with the Google predictor and use it for forecasting purposes.

## 6.5 Forecasts and their interpretations

Forecasting will be done with the help of a linear regression model and the forecasts will be conducted through Excel. The forecasts are done with the original euro (€) values for all dependent variables. The forecasts have been done using the search query volumes and the previous actualized sale values. After the forecasts have been obtained the values are indexed against each other in the 0 to 100 value scale. To visualize the forecasting accuracy the forecasting values are plotted against the actualized values.

The budgeted values are not forecasted because they are dependent on the predictions of the actualized values and this would violate the OLS assumptions. This means that the forecasts are only done to the actual sale values. Since the actualized online store (ACT ecom) did not get a significant correlation with the Google predictor and thus the model has not been verified with the OLS assumptions, the online store values are not forecasted. The forecast for actualized total sales (ACT total) and brick store sales (ACT Brick) are presented in figure 17. The blue line represents the forecast and the orange line shows the actualized sale values.

From the figure 17 it is possible to see that the actual total sale forecasts are not that far apart from the actualized sale outcomes. The overall average seems to be the same even though the forecasts have more volatility in their values. The actualized brick store (ACT Brick) got the best correlation with the Google predictor and it can be seen in these graphs. The lines are seemingly more alike when comparing to the total sale values. In average the brick store forecasts are not a lot better from the actual total sale forecasts, however, the forecasted value distances from the actualized values per observations are generally smaller than for the brick store forecasts. Especially between ‘Autumn 2014’ and ‘Autumn 2016’ the predictions are close to perfect for the actual brick store values.

After ‘Autumn 2016’ the forecasts show higher sales than the actualized sales, which is interesting since budgeted values were also expected to drop from 100 to around 80. This could be explained if there has been some undetected factor that has affected the sales between ‘Autumn 2016’ and ‘Spring 2017’, however, not the search volumes. This could be for example an incident that has affected the brand image. An unfortunate event on the brand image does not necessarily drop the search query volumes since people can ‘google’ the bargain sale also with a negative mindset which then does not lead to sales. The marketing budget was plotted against the actual sales in the ‘Bargain sale data’- part. Curious to see is that in ‘Autumn 2016’ the marketing budget drops a little bit. This might have caused a one period lagged effect on the sales. This change could have cumulated

to the actual sales a bit stronger than expected. Probably the marketing cuts have been accidentally made somewhere where the marketing has been effective before but where it has not impacted the 'googling' of the bargain sale. This would explain why the SVI does not drop.


Figure 17: Forecasts for actual total and brick store sales

The forecasts were also plotted against the budgeted values to see if these company actions are in possible causality towards the actualized sale changes. These plots can be found from appendix 6. For budgeted brick store sales there was a significant budget drop after ‘Autumn 2016’. This is apparent in the actualized sales, however, not in the forecast. This indicates an information lack in the budgeted sales forecasting model. The search query volumes do not include this expected drop in ‘Autumn 2016’. The budgeted values were planned to keep dropping after ‘Autumn 2016’, however, in reality the actualized values kept growing past the budgeted values. The forecasts done with the help of the Google predictor are more accurate with predicting the sales growth when compared to the budgeted sales expectation. This result would follow the previous literature and reassure that the search query predictor can help predicting unexpected interventions when the models excluding the Google predictor cannot.

# 7. Conclusions

The thesis ends by concluding the main research results and findings and answering the research questions. Also, recommendations for further research are presented at the end.

## 7.1 Research results and findings

The research results follow the same line as previous literature. This means that there is potential to use Google predictors in order to predict bargain sale outcomes. It also confirms that the research done for the bargain sale data is done correctly and the results can be considered valid and potential for further research. The goal of this thesis was to study the search query volumes predicting power toward the bargain sale separately from other factors. This was done to see how well the predictor works on its own and thus, understanding what predicting power comes solely through the Google predictor. The results showed that there are some unexpected features in the forecasted values. The forecasting model was not targeted to be exhaustive, which means that it is not advisable to use the Google predictor singularly when predicting the bargain sale in the further research but rather find supporting predictors to add into the model. Hence it would be advisable to try out this thesis' Google predictor for the bargain sale by adding it into a multivariate forecasting model. Possible variables to add into the model could be variables representing the level of the case company's brand image and a variable representing the chosen marketing allocation in social media.

Varian (2014) said: "*As economists know well, there is a big difference between correlation and causation*" in one of his many researches and it is very true. In this thesis the correlation has been in the core of analyzing results. Correlation between the Google predictor and the bargain sales total sales and brick store values are significant. However, the online store values were weakly correlated with the Google predictor, but this does not exclude the possibility of causality. Varian (2014) continued his statement "*So even though a predictive model will not necessarily allow one to conclude anything about causality by itself, such models may help in estimating the causal impact of an intervention when it occurs".* Since there seems to be an interesting connection between the Google predictor and the bargain sale and if the correlation does not significate causality there is definitely proof that the Google predictor will be useful at signaling possible turning points and it will help the research result accuracy.

This Varian's (2014) expression was actualized when the budgeted values were adjusted so that they start dropping after 'Spring 2017' when in fact the actualized sales kept on growing. The Google predictor was able to forecast this growth trend and thus prove its significance on predicting unexpected incidents. From literature is was mentioned that the search query data can help prediction possible trends. For a company this could be very beneficial since the Google predictor was able to predict the sale trend. Also, from the theoretical part it is important to stress out the possibility that Google Trends can help indicating what product would be the next big thing. The company just needs to go explore product search queries and their volumes.

Even though the goal was to forecast the actual sales it was good to acknowledge that there was no significant difference between the Google predictor and budgeted sale values. It would have been worrying if the Google predictor would have not correlated at all with the budgeted values. The reason for this statement is that if the budgeted sales which represented a proxy for company actions, would have parted a lot from the search query volumes it would have indicated biased expectations from the company towards the upcoming sales and hence customer interest. It is also good to see that there is some interconnection since this might show the power of the company action towards the customer interest. By acknowledging this there might be a path to understanding and measuring customer interest on a new level and by so creating better sales.

The fact that the Google predictor forecasted the brick store (ACT Brick) values the best was surprising. The actual total sales (ACT total) would have been the most ideal to forecast since it represents the main goal. By having the ability to predict the incoming sales it is easier to manage investments, future ideas and strategical moves when there is a certain financial limit that the company need to follow. However, presumably it is quite obvious that the actual sale outcomes have more features to consider and thus more volatility, so it would complicate the predicting accuracy. It would have been more expected that the online store sales would correlate the best with the search query amounts. Especially since the online store sales are in size going past the brick store sales. The actual online store outcomes (ACT ecom) had the highest volatility from all of the dependent variables. The variable had the lowest $R^2$ value, which confirms the Google predictors weakness towards predicting variables with high volatility. This is coherent with previous research.

Based on the result that the Google predictor did not correlate with the online store sales, it would be quite wise to assume that the people who 'google' the bargain sale are mainly looking at the information that will be helpful when shopping at the brick store. These searching goals would be mainly navigational and informational searching goals. Another

reason for this insignificant correlation would be that the online store provides shopping possibilities to people outside of Finland. This would suggest the expected search query limitation that the data is consisted from only one search query and for this thesis the query was in Finnish. This will automatically exclude the different language versions. The bargain sale does not have an official name in for example English thus a non-Finnish speaking person can use any kind of synonym to find the sale. With the current Finnish language dependent query, the Google predictor might be able to only predict the online stores sales that are created by Finnish speaking customer.

Another curious finding was the fact that the marketing budgets did not correlate with the search query volumes. It would have been very well expected that when marketing budgets increase it would increase the visibility of the bargain sale and then create higher customer interests. It would be interesting to study this further and find out why there is no correlation.

## 7.2 Answers to research questions

Before answering the main research question the sub-question will be answered first. First the online customer behavior question, then the search query data related questions and lastly the forecasting methodology sub-question are answered.

*Can search query volumes represent customer interest? If so what customer needs affects search engine activity and is search engine activity a good measurement for customer interest?*

It is possible to explain the search query volumes with customer purchasing interest. Anybody using a search engine has a conscious or unconscious goal behind their search activity. When considering the search engine user as a potential customer searching information about the bargain sale it is obvious that there exists some level interest towards the sale. This interest can then move from the computer keyboard into the bargain sale cashier and create sales.

The user need that create the search activities are the needs for personal empowerment or feel of belonging into bigger community. These needs are most probably created through WoM or eWoM. These needs create goals and the goals represent themselves through the form of the query. The query forms can be either *navigational* searches, *informational* searches, *transactional* searches or resource obtaining searches depending on the search goal.

*Is Google Trends data a valid measurement for customer interest and is the data valid for forecasting?*

Search engine data is valid for explaining a sample of the whole population of interest towards the bargain sale. Google is the most popular search engine and thus it should be the best at representing the biggest sample search of the customer interest. When narrowing the research to include only the Google search engine some of parts of the whole population are lost. Also, by studying only the search engine users the non-users are automatically excluded from the forecasts. However, when acknowledging these limitations, the Google Trends data can be used for forecasting and it is easy to model. The data represents an unbiased sample of Google search engines users' interests.

*What forecasting methods are efficient for handling search query data and is search query-based forecasting beneficial?*

The efficient forecasting method depends strongly on the datasets and variables that are predicted with the search query data. The efficiency comes from the data itself and determines the model which can be used for forecasting. There does not exist only one way to forecast with search queries that would be better than the other. For this thesis' data and variables, the classic linear regression model forecasting was the best one because the observed data verified necessary assumptions.

Based on this research and also on the relatively current literature the search query data shows significant helpful features towards forecasting. From the research results gained from this thesis' result it would confirm the helpfulness of the Google predictor when testing it towards the bargain sale. The Google predictor towards the bargain sale was indeed found beneficial.

With the help of the answers to the sub-question the main research will be answered next. The main research question is the following:

*Can Google search engine query volumes help estimating bargain sale outcomes from the perspective of representing a proxy for customer interest?*

Straightforward answer is that the Google predictor is helpful when predicting bargain sale outcomes. It does indicate customer interest on theoretical level as well as on practical level. The secondary data gives out promising results for the Google predictor predicting power. Especially towards predicting two out of three actualized sale sizes. The results recommend adding the Google predictor as a part of the future forecasting. It will represent the proportion of interest that the search engine users have, and it will help out on creating sales estimations. The search query volumes also help out following the upcoming trends towards the bargain sale popularity.

## 7.3 Suggestions for further research

From literature and theory, it seems that it is possible that the search query forecasting research reveals something completely new from the datasets. In this thesis this was how the Google predictor managed to predict sales growth better than the company itself with its budgeted sale sizes. By adding the Google predictor into the sales estimations, it is possible to estimate causal impacts when something unexpected occurs. One future research suggestion would then be to test out the Google predictor into bigger datasets to find more of the causalities that the predictor is able to estimate or even disclose.

Another suggested future research is to create a more comprehensive forecasting model for the bargain sale. This means that new explanatory variables are added into the model to fill out the gaps that are not acknowledged when forecasting solely with the Google predictor. These explanatory variables could be measures for the popularity of the company brand, measures about the popularity of the products sold at the bargain sale or measures expressing the general knowledge of the existence of the sale. One possible research could be around finding these needed variables in order to improve the forecasting model.

It would also be interesting to test out the Google predictor on forecasting marketing campaign success. In this thesis' research there was only a glimpse of the marketing budget and there was no correlation with the search query volumes. Thus, it would be interesting to study why doesn't the marketing budgets show in the search volumes. Another search about marketing would be to examine the marketing campaign success with the help of the google predictor. For future research it could be profiting to see the customer interest level from Google search amount and reflect these with marketing campaign success. Possible research questions for this could be for example the following:

- Is it possible to predict the marketing campaign success with the help of the Google predictor?
- Will the Google search volumes change based on the marketing campaign?

# 8. Executive summary

This research provides analyzes and evaluations of the predictive power of search query volumes (Google predictor) when forecasting the case company's bargain sale. The used methods for testing out the Google predictor's forecasting possibilities towards the different bargain sale dependent variables were classic linear regression models. The CLR models were also used to create the forecasts. All dependent variables were checked to verify the linear regression model demands. All the regression outcomes and results can be found from the appendices and data analytics -part.

The research results are promising. The obtained forecasts follow similar patterns as the studied previous literature. In practice the Google predictor was able to predict two out of three bargain sale outcomes and therefore would signal that there is prominent predicting power. The Google predictor forecasted the total sales and brick store sales the most accurately. The Google predictor was able to predict the actual sales growth trends when the budgeted sales were not able to do it. However, it was surprising to see that the Google predictor was not able to predict the online store sales. It would have been expected that these two variables would have had the highest correlations. Also, it was unexpected to find that the marketing budgets did not have any significant correlation with the search query volumes. This is a significant finding since the marketing should be the causality for creating customer attention and thus make customers 'google' for the bargain sale and create search query volumes.

The results cannot be considered exhaustive for the whole customer base since the research is done only to study the search engine users. This means that only search engine users are included into the forecast models. The small dataset restrains the amount of previous observations that can be used to create next period forecasts. However, based on the results it is confident to say that the Google predictor for the bargain sale forecasting is helpful and has the advantage of explaining previously unknown phenomena and also upcoming trends. Thus, the search query data should be considered as a part of the decision-making process firstly through Google Trends observation and secondly by adding the Google predictor into the company's future forecasting models.

From the theoretical part the main findings for the bargain sale were the following:

- Google Trends helps at detecting upcoming product trends
- Search query volumes help to study singular product demands
- Analyzing what people search for can help choosing what products should be added to the collections

From the empirical part the main findings for the bargain sale were the following:

- The Google predictor was significant at predicting brick store sales
- The online store sales cannot be predicted when solely using the Google predictor
- The marketing budgets were not helpful at explaining the search query volumes
- It is possible to create a forecasting model for the bargain sale that contains the Google predictor

Future recommendations for the company:

- Improve the bargain sale forecasts by adding the Google predictor as one of the explanatory variables
- Test the Google predictor for bigger datasets such as the continuing company sales
- Improve the forecasting results by adding complementary explanatory variables into the forecasting model
- Try the Google predictor on marketing campaign forecasting

# References

**Adhikari R. & Agrawal R.K. 2013.** An Introductory Study on Time Series Modeling and Forecasting. LAP Lambert Academic Publishing, Germany

**Anderson E.W. 1998.** Customer Satisfaction and Word of Mouth. Journal of Service Research, Volume 1, No. 1. Pp. 5-17. Sage Publications, Inc.

**Askitas N. & Zimmermann K. L. 2009**. Google Econometrics and Unemployment Forecasting. Discussion Paper No. 4201. Applied Economics Quarterly 55 (2), 107-120.

**Barrett B. 2015**. Google Trends Now Shows the Web's Obsessions in Real Time. Wired Magazine, CNMN Collection [online document]. [Accessed 30 January 2018]. Available: https://www.wired.com/2015/06/google-trends-real-time/

**Brin S. & Page L. 1998.** The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, Vol. 30 No. 1-7, pp. 107-117

**Broder A. 2002.** A taxonomy of web search. SIGIR Forum. Fall 2002, Vol. 36, No. 2.

**Brooks C. 2008.** Introductory Econometrics for finance, 2nd edition. Cambridge University Press. Cambridge

**Buttle F.A. 1998**. Word of mouth: Understanding and managing referral marketing. Journal of Strategic Marketing (6), 241-254. Routledge

**Castle J L., Fawcett N.W.P. & Hendry D.F. 2009.** Nowcasting is not just Contemporaneous Forecasting. Department of Economics. Special Issue of the National Institute Economic Review.

**Chaudhuri A. & Holbrook M.B. 2001.** The Chain of Effect from Brand Trust and Brand Affect to Brand Performance: The Role of Brand Loyalty. Journal of Marketing, 65:2. ABI/INFROM Global

**Cheung C.M.K. & Thadani D.R. 2010**. The effectiveness of Electronic Word-o-Mouth Communications: A Literature Analysis. 23rd Bled eConference, eTrusts: Implications for the Individual, Enterprises and Society. Bled, Slovenia

**Choi H. & Varian H. 2009**. Predicting the Present with Google Trends. Technical Report, Google Inc.

**Choi H. & Varian H. 2011**. Predicting the Present with Google Trends. Technical Report, Google Inc.

**Collins M. 2016.** How to use Google Trends to Gain a Competitive edge. Hallam Internet Ltd. [online document]. [Accessed 15 March 2018]. Available: https://www.hallaminternet.com/google-trends-introduction-business/

**Da, Engelberg & Gao 2011.** In Search of Attention. The Journal of Finance, Vol LXVI (66) No. 5

**Davis A. & Khazanchi D. 2008**. An Empirical Study of Online Word of Mouth as a Predictor for Multi-Product Category E-Commerce Sales. Information Systems and Quantitative Analysis Faculty Publications. 17.

**Dimpfl T. & Jank S. 2012.** Can internet search queries help to predict stock market volatility? Eur Financial Management, 22: 171–192.

**Ettredge M., Gerdes J. & Karuga G. 2005**. Using Web-based Search Data to Predict Macroeconomic Statistics. Communications of the ACM, 48 (11): 87-92

**Fallows D. 2005**. Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. Pew Internet & American Life Project 1615, Washington

**Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S. & Brilliant L. 2009.** Detecting influenza epidemics using search engine query data. Nature, Vol 457, pp. 1012–1014

**Godes D. and Mayzlin D. 2004**. Using Online Conversations to Study Word-of-Mouth Communication. Marketing Science, Vol. 23, No. 4, 545–560. Informs

**Hannák A., Sapiezyski P., Kakhki A.M., Lazer D., Mislove A. & Wilson C. 2017.** Measuring Personalization of Web Search [extention paper]. WWW 2013 [25]

**Henning-Thurau T., Gwinner K.P., Walsh G. & Gremler D.D. 2004**. Electronic Word-of-Mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? Journal of Interactive Marketing, Vol. 18. No. 1, Wiley Periodicals, Inc. and Direct Marketing Educational Foundation, Inc.

**Hung K. & Yiyan Li S. 2007.** The Influence of eWOM on Virtual Consumer Communities: Social Capital, Consumer Learning and Behavioral Outcomes. Journal of Advertising Research pp. 485-495

**Koop G. & Onorante L. 2013.** Macroeconomic Nowcasting Using Google Probabilities. Working paper, ESRC

**Koufaris M. 2002.** Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior. Information Systems Research, Vol. 13. No. 2, pp. 205-223

**Kristoufek L. 2013**. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific Reports 3:3415

**Kulkarni G.M. 2010**. Using online search data to forecast new product sales. Doctoral Dissertation. Faculty of the Graduate School of the University of Maryland.

**Lindgreen, A. & Vanhamme, J. 2005**. Viral marketing: the use of surprise. In Advances in Electronic Marketing, Clarke, I. and Flaherty, T.B. (Eds.) Hershey, PA: Idea Group Publishing, 122-138.

**McLaren N. & Shanbhoge R. 2011**. Using internet search data as economic indicators. Bank of England Quarterly Bulletin 2011/Q2

**Omar N.N. & Yee T.M. 2017.** Effectiveness of hidden messages in advertisements towards viewers' buying intention. International Journal of Social Sciences. Vol. 3, Issue 1, pp. 542-553. GRDS Publishing

**Purcell K., Brenner J. & Rainie L. 2012**. Search Engine Use 2012. Pew Research Center's Internet & American Life Project, Washington

**Rogers S. 2016.** What is Google Trends data – and what does it mean? Google News Lab [online document]. [Accessed 13 February 2018]. Available: https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8

**Rose D. E. & Levinson D. 2004.** Understanding User Goals in Web Search. WWW 2004, May 17-22. New York, USA

**Rouse M. 2017.** Data visualization. Essential guide: Enterprise data analytics strategy: A guide for CIOs. TechTarget Network

**Schmidt T. & Vosen S. 2009.** Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. Ruhr economic papers. No. 155

**Shimshoni Y., Efron N. & Matias Y. 2009.** On the Predictability of Search Trends. Unpublished paper. Google, Israel Labs

**Statista 2018a**. Worldwide desktop market share of leading search engine from January 2010 to October 2017 [online document]. [Accessed 29 January 2018]. Available: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

**Statista 2018b**. Number of search engine user in the United States from 2014 to 2020 (in millions) [online document]. [Accessed 29 January 2018]. Available: https://www.statista.com/statistics/253795/number-of-search-engine-users-in-the-united-states/

**Stedman C. 2012.** Data visualization get more advanced – and more complicated. Essential guide: dashboard development and data visualization tools for effective BI. TechTarget Network

**Sundaram, D.S., Mitra, K., & Webster, C. 1998**. Word- of-Mouth Communications: A Motivational Analysis. Advances in Consumer Research, 25, 527–531

**Teevan J., Alvarado C., Ackerman M. S. & Karger D. R. 2004.** The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Direct Search. CHI Vol 6, No. 1. April 24-29, Vienna, Austria

**Vakratsas D. and Ambler T. 1999**. How Advertising Works: What Do We Really Know? Journal of Marketing. Vol. 63 (January), 26-43

**Varian H. R 2014.** Big Data: New Tricks for Econometrics. Journal of Economic Perspectives. Vo. 28, N0. 2, pp. 3-28

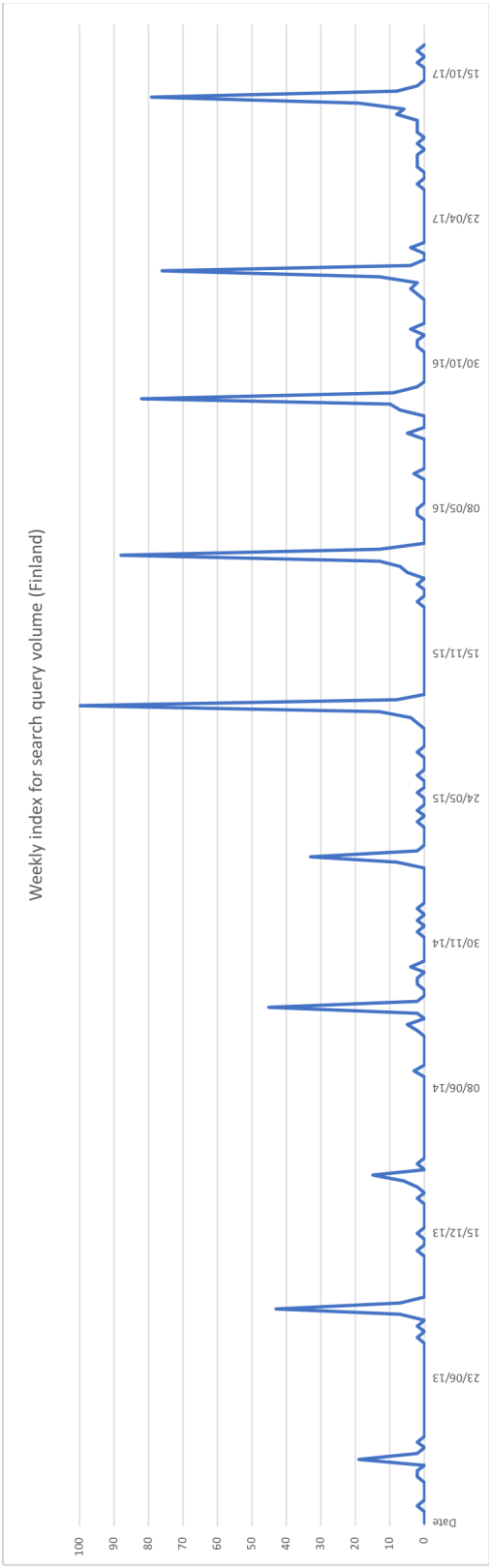
**Vaughan, L. & Romero-Frías, E. 2013.** Web Search Volume as a Predictor of Academic Fame: An Exploration of Google Trends. Journal of the American Society for Information Science and Technology, 75(4): 707-720.

**White R. W. & Drucker S. M. 2007.** Investigating Behavioral Variability in Web Search. WWW 2007 / Track: Browsers and User Interfaces. May 8-12. Canada

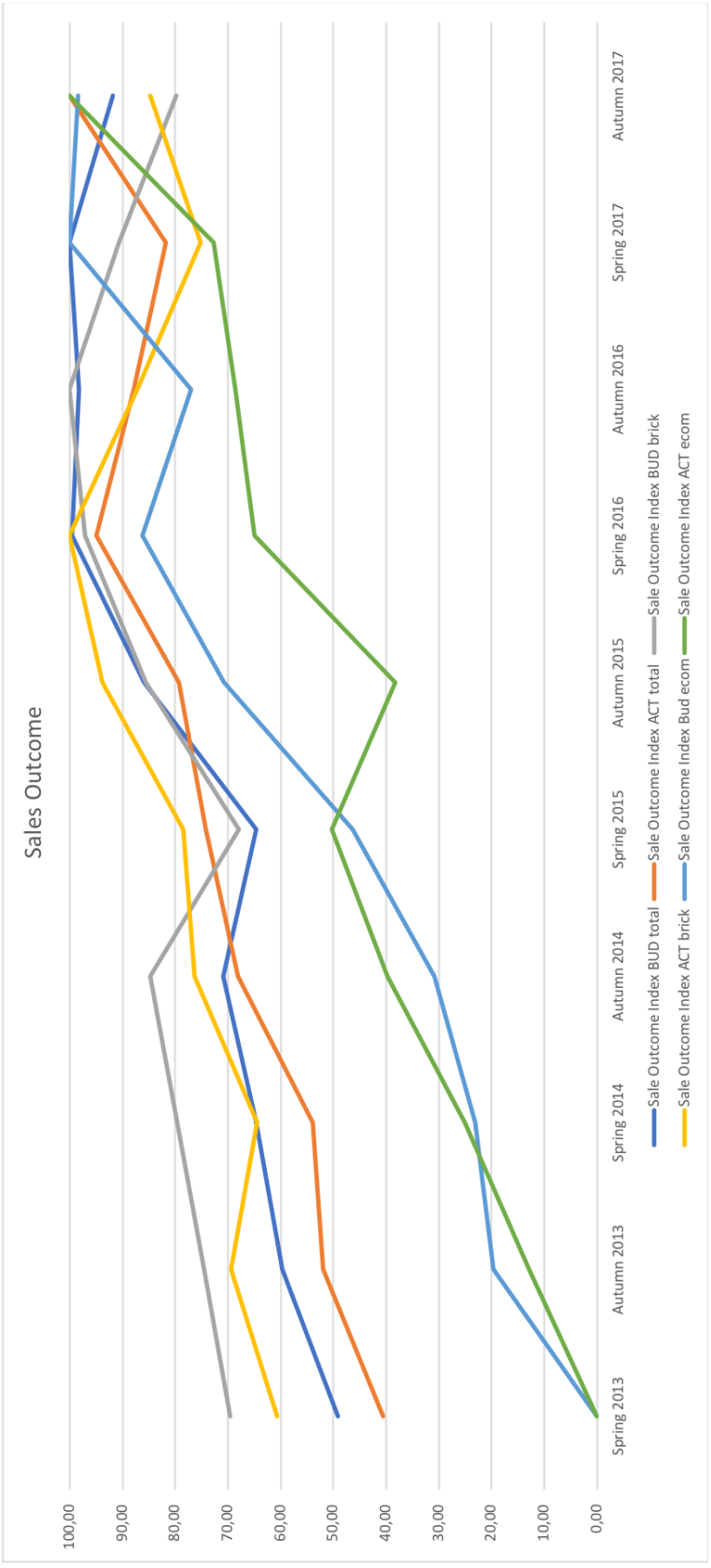
**Young B.K., Lee J., Park N., Choo J., Jong-Hyun K. & Kim C.H. 2017**. When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. PLoS One, vol. 12, no. 5.

# Appendices

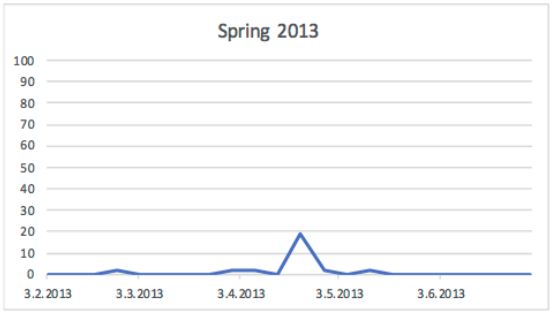
**Appendix 1:** Weekly index for search query volume in Finland (Google Trends, data accessed 29 January 2018)

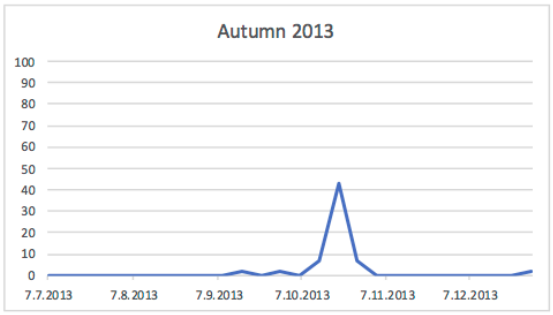


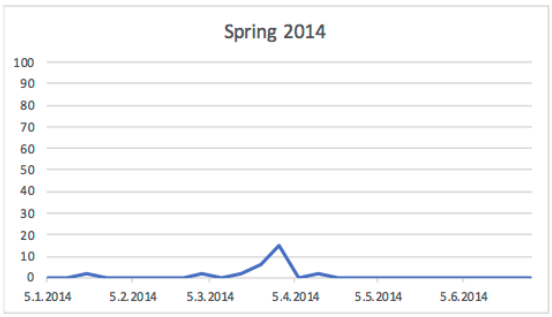
**Appendix 2:** Sales outcomes

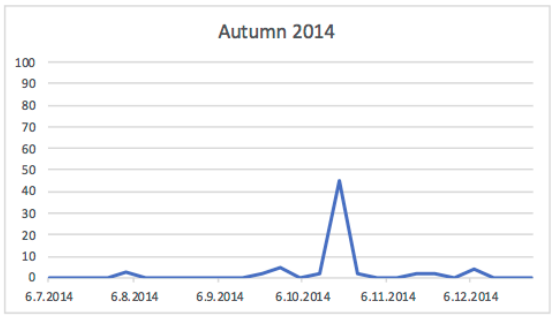


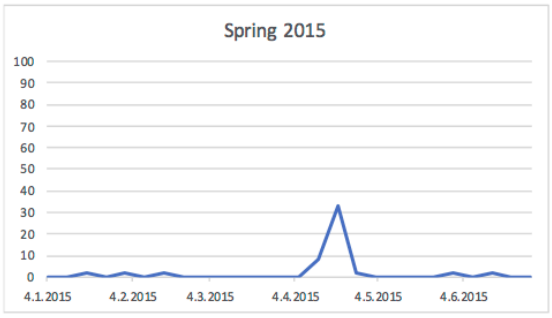
**Appendix 3**: Weekly SVI per Bargain sale week

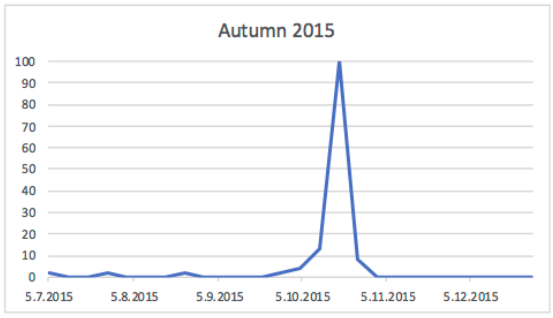


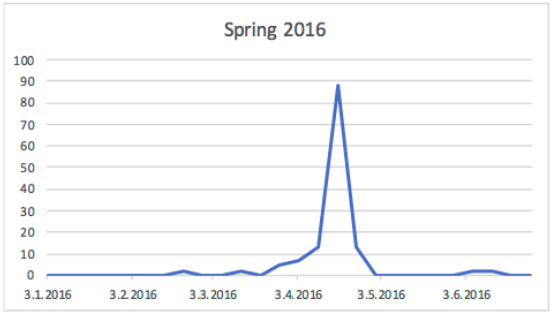



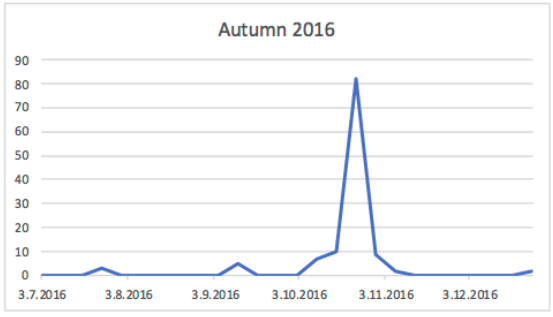



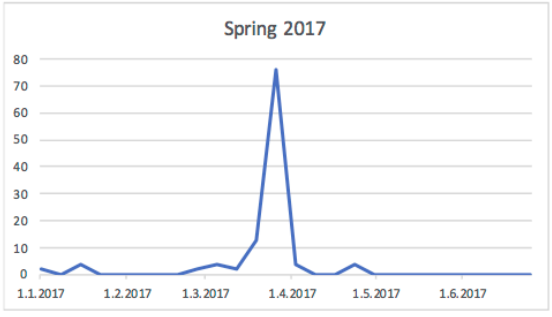



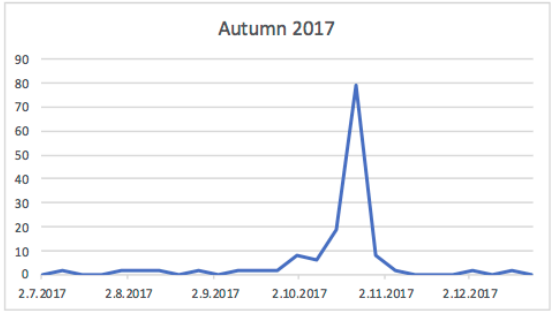



# **Appendix 4**: Regressions analysis for Actual Sales Outcome

SUMMARY OUTPUT ACT Brick

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,877673359 |
| R Square | 0,770310525 |
| Adjusted R Square | 0,74159934 |
| Standard Error | 6,392669941 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 1096,425903 | 1096,425903 | 26,8296325 | 0,00084303 |
| Residual | 8 | 326,9298318 | 40,86622897 | | |
| Total | 9 | 1423,355735 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 58,08179588 | 4,523745452 | 12,83931567 | 1,2787E-06 | 47,65002016 | 68,5135716 | 47,6500202 | 68,5135716 |
| Search index | 0,361414158 | 0,069774671 | 5,179732856 | 0,00084303 | 0,200513477 | 0,522314838 | 0,20051348 | 0,52231484 |

RESIDUAL OUTPUT

| *Observation* | *Predicted Index ACT brick* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 64,95 | -4,31 | -0,72 |
| 2 | 73,62 | -4,27 | -0,71 |
| 3 | 63,50 | 1,05 | 0,17 |
| 4 | 74,35 | 1,94 | 0,32 |
| 5 | 70,01 | 8,53 | 1,42 |
| 6 | 94,22 | -0,35 | -0,06 |
| 7 | 89,89 | 10,11 | 1,68 |
| 8 | 87,72 | -0,58 | -0,10 |
| 9 | 85,55 | -10,32 | -1,71 |
| 10 | 86,63 | -1,80 | -0,30 |
| Mean | | 0,00 | |
| Max | | 10,11 | |
| Min | | -10,32 | |

PROBABILITY OUTPUT

| *Percentile* | *Index ACT brick* |
|---|---|
| 5 | 60,63739754 |
| 15 | 64,55092417 |
| 25 | 69,35237253 |
| 35 | 75,23043598 |
| 45 | 76,28203497 |
| 55 | 78,53970286 |
| 65 | 84,82862982 |
| 75 | 87,14144538 |
| 85 | 93,87522701 |
| 95 | 100 |

SUMMARY OUTPUT Act total

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,8264782 |
| R Square | 0,683066216 |
| Adjusted R Square | 0,643449493 |
| Standard Error | 11,66825547 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 2347,448726 | 2347,448726 | 17,2418656 | 0,003198709 |
| Residual | 8 | 1089,185486 | 136,1481857 | | |
| Total | 9 | 3436,634212 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 42,6059363 | 8,256990915 | 5,15998343 | 0,00086368 | 23,56528111 | 61,6465915 | 23,5652811 | 61,6465915 |
| Search index | 0,528826945 | 0,127356597 | 4,152332554 | 0,00319871 | 0,235142105 | 0,822511784 | 0,2351421 | 0,82251178 |

RESIDUAL OUTPUT

| *Observation* | *Predicted Index ACT total* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 52,65 | -12,14 | -1,10 |
| 2 | 65,35 | -13,42 | -1,22 |
| 3 | 50,54 | 3,44 | 0,31 |
| 4 | 66,40 | 1,75 | 0,16 |
| 5 | 60,06 | 14,21 | 1,29 |
| 6 | 95,49 | -16,19 | -1,47 |
| 7 | 89,14 | 5,83 | 0,53 |
| 8 | 85,97 | 1,94 | 0,18 |
| 9 | 82,80 | -1,02 | -0,09 |
| 10 | 84,38 | 15,62 | 1,42 |
| Mean | | 0,00 | |
| Max | | 15,62 | |
| Min | | -16,19 | |

PROBABILITY OUTPUT

| *Percentile* | *Index ACT total* |
|---|---|
| 5 | 40,51103093 |
| 15 | 51,92126306 |
| 25 | 53,97634921 |
| 35 | 68,1580995 |
| 45 | 74,26256534 |
| 55 | 79,29676801 |
| 65 | 81,77352823 |
| 75 | 87,90739917 |
| 85 | 94,97198746 |
| 95 | 100 |

SUMMARY OUTPUT ACT ecom

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,668916242 |
| R Square | 0,447448939 |
| Adjusted R Square | 0,378380057 |
| Standard Error | 23,87464682 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 3692,622948 | 3692,622948 | 6,47829996 | 0,034428236 |
| Residual | 8 | 4559,990086 | 569,9987608 | | |
| Total | 9 | 8252,613034 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 8,775721757 | 16,89479137 | 0,519433568 | 0,6175197 | -30,18373702 | 47,73518053 | -30,183737 | 47,7351805 |
| Search index | 0,663258768 | 0,260586836 | 2,545250471 | 0,03442824 | 0,062344446 | 1,26417309 | 0,06234445 | 1,26417309 |

RESIDUAL OUTPUT

| *Observation* | *Predicted Index ACT ecom* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 21,38 | -21,38 | -0,95 |
| 2 | 37,30 | -24,40 | -1,08 |
| 3 | 18,72 | 6,32 | 0,28 |
| 4 | 38,62 | 1,06 | 0,05 |
| 5 | 30,66 | 19,63 | 0,87 |
| 6 | 75,10 | -36,83 | -1,64 |
| 7 | 67,14 | -2,14 | -0,10 |
| 8 | 63,16 | 5,36 | 0,24 |
| 9 | 59,18 | 13,55 | 0,60 |
| 10 | 61,17 | 38,83 | 1,72 |
| Mean | | 0,00 | |
| Max | | 38,83 | |
| Min | | -36,83 | |

PROBABILITY OUTPUT

| *Percentile* | *Index ACT ecom* |
|---|---|
| 5 | 0 |
| 15 | 12,89694048 |
| 25 | 25,0437619 |
| 35 | 38,26699104 |
| 45 | 39,68667147 |
| 55 | 50,2946856 |
| 65 | 65,00160379 |
| 75 | 68,52373716 |
| 85 | 72,73291142 |
| 95 | 100 |

## **Appendix 5**: Regressions analysis for Budgeted Sales Outcome

**SUMMARY OUTPUT SVI - BUD total**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,879794993 |
| R Square | 0,77403923 |
| Adjusted R Square | 0,745794134 |
| Standard Error | 9,484683862 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 2465,276486 | 2465,276486 | 27,4043758 | 0,000788145 |
| Residual | 8 | 719,6738238 | 89,95922797 | | |
| Total | 9 | 3184,95031 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 47,02149142 | 6,711795836 | 7,005798831 | 0,00011199 | 31,54406247 | 62,4989204 | 31,5440625 | 62,4989204 |
| Search index | 0,541936415 | 0,103523364 | 5,234918887 | 0,00078814 | 0,303211109 | 0,78066172 | 0,30321111 | 0,78066172 |

RESIDUAL OUTPUT

| *Observation* | *Predicted BUD total* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 57,31828331 | -8,158952119 | -0,912405272 |
| 2 | 70,32475728 | -10,6383602 | -1,189674334 |
| 3 | 55,15053765 | 9,461953359 | 1,058118248 |
| 4 | 71,40863011 | -0,434464042 | -0,048585563 |
| 5 | 64,90539313 | -0,273540496 | -0,030589687 |
| 6 | 101,2151329 | -15,39132369 | -1,721192215 |
| 7 | 94,71189596 | 4,862148401 | 0,543727892 |
| 8 | 91,46027747 | 6,841588434 | 0,765086161 |
| 9 | 88,20865898 | 11,79134102 | 1,318610717 |
| 10 | 89,83446823 | 1,93960933 | 0,216904052 |

**SUMMARY OUTPUT SVI - BUD Brick**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,722699257 |
| R Square | 0,522294217 |
| Adjusted R Square | 0,462580994 |
| Standard Error | 7,943436399 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 551,9014687 | 551,9014687 | 8,74670954 | 0,018216122 |
| Residual | 8 | 504,7854545 | 63,09818182 | | |
| Total | 9 | 1056,686923 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 68,08846016 | 5,621138682 | 12,11293014 | 1,9961E-06 | 55,12609111 | 81,0508292 | 55,1260911 | 81,0508292 |
| Search index | 0,256416692 | 0,086700967 | 2,95748365 | 0,01821612 | 0,056483904 | 0,45634948 | 0,0564839 | 0,45634948 |

RESIDUAL OUTPUT

| *Observation* | *Predicted BUD brick* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 72,9603773 | -3,371244172 | -0,450150822 |
| 2 | 79,1143779 | -4,607643361 | -0,61524302 |
| 3 | 71,93471053 | 7,783378742 | 1,039288215 |
| 4 | 79,62721128 | 5,180932809 | 0,691792419 |
| 5 | 76,55021098 | -8,6357239 | -1,153098978 |
| 6 | 93,73012932 | -8,217208409 | -1,097216022 |
| 7 | 90,65312902 | 6,449455164 | 0,861173915 |
| 8 | 89,11462887 | 10,88537113 | 1,453486758 |
| 9 | 87,57612872 | 3,081662981 | 0,411484026 |
| 10 | 88,34537879 | -8,548980986 | -1,141516491 |

**SUMMARY OUTPUT SVI - BUD ecom**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,85468967 |
| R Square | 0,730494431 |
| Adjusted R Square | 0,696806235 |
| Standard Error | 19,74203394 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 8451,288776 | 8451,288776 | 21,6839877 | 0,001630619 |
| Residual | 8 | 3117,983234 | 389,7479043 | | |
| Total | 9 | 11569,27201 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95,0%* | *Upper 95,0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -2,99725165 | 13,97036561 | -0,214543537 | 0,835491 | -35,21297251 | 29,2184692 | -35,212973 | 29,2184692 |
| Search index | 1,003406681 | 0,215480221 | 4,656606891 | 0,00163062 | 0,506508401 | 1,50030496 | 0,5065084 | 1,50030496 |

RESIDUAL OUTPUT

| *Observation* | *Predicted Bud ecom* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 16,06747529 | -16,06747529 | -0,863240897 |
| 2 | 40,14923563 | -20,53385102 | -1,103201322 |
| 3 | 12,05384856 | 11,02307451 | 0,592225509 |
| 4 | 42,15604899 | -11,38681822 | -0,611767998 |
| 5 | 30,11516882 | 16,20513887 | 0,870637009 |
| 6 | 97,34341645 | -26,66080106 | -1,432377733 |
| 7 | 85,30253628 | 0,851309879 | 0,04573746 |
| 8 | 79,28209619 | -2,358583882 | -0,126717237 |
| 9 | 73,2616561 | 26,7383439 | 1,436543798 |
| 10 | 76,27187615 | 22,18966232 | 1,192161411 |

**Appendix 6**: Budgeted total sales plotted against actual sales