An exploratory time series data analysis on leasing vehicles' maintenance costs

Lappeenranta–Lahti University of Technology LUT

Bachelor's Programme in Business Administration, Bachelor's thesis

2023

Jussi Joutsjärvi

Examiner: Jyrki Savolainen

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Jussi Joutsjärvi

Tutkiva aikasarja-analyysi leasing-ajoneuvojen huoltokustannuksista

Kauppatieteiden kandidaatintutkielma

2023

33 sivua, 20 kuvaa ja 4 taulukkoa

Tarkastaja: Jyrki Savolainen

Avainsanat: Tutkiva data-analyysi, Aikasarja-analyysi, Huoltokustannukset

Työ tutkii nimeämättömän firma X:n leasing-autojen huoltokustannuksia Suomessa vuosina 2018–2022. Analyysissä sovelletaan tutkivan analyysin keinoja aikasarjadataan firman leasingautokannan huoltokuluista, sekä kuluttajahintaindeksin alakategoriaan autojen huolloista. Analyysivaihe on toteutettu Python-ohjelmointikieltä hyödyntäen.

Vuoden 2022 aikana kiihtynyt inflaatio on lisännyt tarvetta tutkia huoltokustannusten kehitystä. Tutkimus alkaa tutkivan data-analyysin ja aikasarjojen keskeisten käsitteiden ja teorian katsauksella, josta edetään tutkimaan autojen määräaikaishuoltojen hintakehitystä. Huoltokustannusanalyysi perustuu aikasarjadatasta piirrettyihin kuvaajiin ja tunnuslukuihin. Tutkivan analyysin keinoja hyödynnetään myös ARIMA-mallin estimoinnissa.

Työ vahvistaa määräaikaishuoltokustannusten riippuvan muun muassa auton kilometrilukemasta, merkistä, kuin myös polttoainetyypistä. Analyysissä havaintoja aggregoidaankin mainittujen autojen ominaisuuksien mukaan mediaaniarvoiksi eri mittaisille ajanjaksoille. Eri aggregaateista tehdyistä kuvaajista on useimmiten havaittavissa kuluttajahintaindeksin kehityssuunnan kaltainen nouseva trendi. Yleisen inflaation kehityksestä havaittujen poikkeamien arvioidaan johtuvan autokannan uniikista koostumuksesta, sekä asiakkaiden muuttuneista ajotottumuksista.

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

School of Business and Management

Business Administration

Jussi Joutsjärvi

An exploratory time series data analysis on leasing vehicles' maintenance costs

Bachelor's Programme in Business Administration, Bachelor's thesis

2023

33 pages, 20 figures and 4 tables

Examiner: Jyrki Savolainen

Keywords: Exploratory data analysis, Time series analysis, Maintenance costs

The thesis examines the maintenance costs of an unnamed company X's leasing vehicles in Finland for the years 2018 to 2022. The analysis applies exploratory analysis methods on time series data on the company's leasing fleet maintenance costs and the consumer price index subcategory on vehicle maintenance. The analysis is conducted using the Python programming language.

Rising inflation in 2022 has increased the need to examine the development of maintenance costs. The study begins by reviewing fundamental concepts in exploratory data analytics and time series, which leads to the analysis of cost development for the vehicles' scheduled maintenance. Maintenance cost analysis is based on visual representations of the data and summary statistics. Exploratory data analysis methods are also utilized for estimating an ARIMA model.

The study verifies that scheduled maintenance costs depend on factors such as the vehicle's mileage, make, and fuel type. The analysis aggregates observations based on the aforementioned vehicle attributes as median values for periods of varying length. Most figures on the aggregated data show ascending trends similar to the consumer price index. Deviations from the rate of general inflation are hypothesized to be caused by the unique composition of the fleet, as well as the customers' changing driving habits.

SYMBOLS AND ABBREVIATIONS

μ	Mean
ε	Error
β, θ	Coefficients
y	Dependent variable
x	Explanatory variable
ACF	Autocorrelation function
ADF	Augmented Dickey-Fuller
AIC	Akaike information criterion
AR	Autoregressive model
DGP	Data-generating process
EDA	Exploratory data analysis
ICE	Internal combustion engine
IQR	Interquartile range
KDE	Kernel density estimation
MA	Moving average model
PACF	Partial autocorrelation function

Table of contents

1 Introduction

A survey by the Finnish automotive industry's information center has shown (Autoalan tiedotuskeskus 2021) that in addition to worries about devaluation, 55% of consumers also worry about the cost of maintaining their vehicles. This is understandable since maintenance forms a sizeable portion of a vehicle's yearly expenses. In a Yle news article, The Finnish automobile union approximated the average cost of a privately owned vehicle's maintenance to be around 500 to 600 euros in 2018, with the total yearly average cost of owning a car with devaluation not considered being about 3000 euros (Hanhinen 2018). This Bachelor's thesis examines the development of maintenance costs for an unnamed vehicle leasing provider operating in Finland, referred to as company X.

In company X's leasing contracts, vehicle maintenance is included in the monthly leasing fee. The company operates a large fleet of several thousand vehicles in Finland, and since this fleet of cars needs to be maintained like all other vehicles in private use, maintenance fees from repairs and usual scheduled services form a significant expenditure for the company. It goes without saying that a company maintaining a large fleet of vehicles should consider the effect of sudden changes in the cost of maintenance due to factors like inflation.

Leasing vehicle maintenance costs and inflation's effect on the cost of leasing have been addressed in earlier studies. For example, Baklouti et al. (2022) studied the profitability of leasing through mathematical modeling, and Feng et al. (2018) examined the cost of leasing for the lessee when fluctuation caused by inflation is considered. However, no studies can directly address the importance of factors such as general inflation with respect to how company X could apply its unique data in the current economic climate.

Due to increased uncertainty over the development of vehicle maintenance costs, it is essential to examine how factors such as inflation have affected costs in the past and what can be expected in the future. The effect of inflation on consumer prices is well-kept track of, but it is not sure how well it is reflected in the prices for a company servicing large masses of vehicles regularly.

1.1 Background

In 2022, rising inflation due to recent global crises has become a problem for both consumers and businesses. According to Lindholm et al. (2016), one of the most important objectives of economic policy in western countries is to maintain a slow rate of inflation, as rapid inflation causes uncertainty that complicates the operation of businesses and slows down economic growth. The European central bank aims for an inflation rate of 2% to "maintain price stability" (ECB 2022). In Finland, the inflation rate was well below the target since 2013 until it started rising steadily in 2021. In October of 2021, the rate was 2,5%, and since then inflation has already reached over 7%, with the latest measurement from August 2022 showing a rate of 7,6%. (Statistics Finland 2022a)

No matter the cause, the effects of rising inflation can be seen in the prices of all commodities, affecting the automobile industry most notably through a rise in fuel prices. According to a survey by LähiTapiola (2022), rising energy and fuel prices are causing uncertainty among consumers, delaying or affecting the vehicle purchase decision of 23% of the respondents with 18% feeling unsure about what fuel type they should choose. Inflation also causes additional costs for vehicle leasing providers, that need to pay more for services to maintain their vehicles. This affects the operation of large leasing fleets, where prices for new leasing contracts are affected by maintenance costs, as are profits from active contracts.

The maintenance costs of leasing vehicles have been addressed in earlier studies. In a recent study, Baklouti et al. (2022) created an analytical model to determine whether it is more profitable to sell, or lease used vehicles considering many factors, including expected maintenance costs. However, the study doesn't consider changes in maintenance costs, and rather than exploring existing data, is building a mathematical model partly based on assumptions. Another study by Feng et al. (2018) examined the cost of leasing for the lessee when changes in the consumer price index are considered. In this study, however, the rental price of the leasing contract also fluctuated with inflation, which is not the case with company X's vehicle leasing where monthly installments are predetermined.

No theory or pre-made model can directly be applied to company X's maintenance data before its characteristics are well known. As different leasing companies have different kinds of vehicle fleets, operating in different climates with different customs, it is unlikely a study based on entirely different vehicle data could be directly applied to another company's fleet.

By conducting exploratory data analysis on the company's maintenance data, underlying trends are revealed to understand the formation of maintenance costs.

1.2 Data and Methodology

The study is conducted with exploratory data analysis methods. The author has been provided with data on a leasing provider's fleet of vehicles. The goal is to find out what kind of analysis methods can be used on the data provided and to determine how inflation has affected the cost of the vehicles' maintenance. The provided dataset contains data concerning company X's vehicle maintenance over the last five years, from the beginning of 2018 till August 2022. This data will also be compared with Statistics Finland's consumer price index on repair and maintenance of privately owned vehicles (Statistics Finland 2022a).

The provided data includes all vehicle maintenance on all fleet vehicles, unique vehicle numbers, mileage, make, model, event cost, dates, vehicle types, and fuel types. The analysis will be conducted using programmatic methods with Python, utilizing several of its libraries made for data analysis. The data will first be cleaned and processed, then econometric analysis methods will be applied to gain insight and accomplish the goals set for the study. An exploratory data analysis approach has been chosen, because it enables an open-minded exploration of the data where conclusions can be drawn from both graphical visualizations and numerical summaries (Hartwig & Dearing 1979).

1.3 Research Questions

The study aims to make meaningful observations from the provided dataset using suitable exploratory data analysis methods. Points of interest are especially to find out how the company's maintenance costs have developed compared to consumer prices, and how factors like vehicles' fuel type or brand affect price development. The research questions set for the study are the following:

RQ1: ***What are the most commonly used exploratory data analysis methods, and how to apply them to time series data?***

RQ2: ***What kind of development can be seen in the vehicle maintenance cost dataset, and what are the key insights with respect to general inflation?***

1.4 Structure of the Thesis

In the following chapters, the basics of exploratory data analysis and time series data will be introduced, which leads to using programmatic analysis methods on the data. The findings will be presented with the help of clever visualizations that hopefully lead to the formation of valuable information that helps to better understand the development of the company's vehicle maintenance costs. The last chapter concludes the findings and discusses their meaning.

2 Methods

This chapter describes the analytical methods later utilized on the data used in this study. First, the basics of exploratory data analysis and time series data are introduced, followed by a summary of essential concepts for time series data and how they fit into the exploratory data analysis methods. All of the theory for exploratory data analysis and time series analysis described in this chapter is based on the following works: Exploratory data analysis (Hartwig & Dearing 1979), Practical time-series analysis : master time series data processing, visualization, and modeling using python (Pal & Prakash 2017), Introduction to Time Series Analysis (Pickup 2015), Applied Time Series Analysis and Forecasting with Python (Huang & Petukhina 2022), and LUT University Advanced Course in Statistical Methods materials (Jabłońska-Sabuka 2018).

2.1 Exploratory data analysis

Exploratory data analysis (EDA) is a more flexible and open way of analyzing data. It focuses more on the graphical visualization of data, putting less weight on the statistical side of data analysis. Whereas traditional confirmatory data analysis tests a hypothesis by

calculating values from the data and testing the likelihood of them occurring by chance, EDA aims to better understand the relationships of variables and give insight into them. (Hartwig & Dearing 1979)

According to Hartwig & Dearing, the most fundamental concept of EDA is the smoothness and roughness of data. The smooth is the estimation, it describes the patterns in the data, such as a line fitted between two variables. It describes the general shape of the data, whereas the rough is the form the data takes after the smooth has been extracted. They are the residuals, the points not explained by the general patterns, and should not contain any patterns themselves. If the rough data contains patterns, it can still be "smoothened" to achieve the roughness where no pattern can be found.

As mentioned, EDA relies heavily on the use of visual representations of data, both on individual variables, and the relationships of variables. In the case of statistical summaries, emphasis is placed on resistant statistics. They are statistics less affected by unusual outlier observations, making them better at identifying the roughness, which could be blurred by using nonresistant statistics. For example, the usual measurement of spread, standard deviation, is highly nonresistant due to its formula containing the sum of squares, which is the deviation from the mean squared. The deviation from the mean being squared means that unusual outlier observations increase the standard deviation at an increasing rate. Although less sensitive, the mean is also a nonresistant value. Order statistics such as the median and the interquartile range are good, more resistant alternatives for describing the distribution of values. (Hartwig & Dearing 1979)

The motive for conducting exploratory analysis is that by simply fitting some smooth model to a dataset and testing its summary statistic for statistical significance, the relationship in the data might not actually be smooth even if the statistic would so indicate, or the statistic might not be significant because the data should be tested for some other smooth relationship. Exploratory data analysis aims to uncover the true relationships of variables in the data, thus also enabling the testing of the correct models. (Hartwig & Dearing 1979)

2.2 Time series

Quantitative data collected at sequential points in time, usually at even intervals is a time series. Time series data of varying frequencies can be utilized to understand what affects the state of a process over time and to forecast the future state of the process by its characteristics. (Pal & Prakash 2017)

In time series analysis, the dependent time series variable Y_t is a sequence of time-ordered observations:

$$(1)\ \mathrm{Y_t} = \mathrm{y_1}, \mathrm{y_2}, \mathrm{y_3}, \dots, \mathrm{y_T}, \mathrm{T} = number\ of\ timepoints$$

These observations are the outcome of a data-generating process (DGP). The DGP of a time series dataset is considered a stochastic or random process with infinite possible outcomes, where only one of these potential outcomes has been drawn for each time point. An example of a time series DGP:

$$(2)\ y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, for\ t = 1,2, \dots, T$$

where only one value of explanatory variable x_t and error term ε_t is drawn for each time point t, forming the observed values y_t. The data-generating process is an unknown reality, which the analysis aims to uncover. However, when estimating a model from data, the DGP is usually thought to be more complicated than the model. (Pickup 2015)

2.3 Time series attributes

Time series data differs from cross-sectional data by its characteristics, and so do the analysis methods applicable to it. A time series can be expressed as a sum or multiplication of the following characteristics: trend, seasonality, cyclical movement, and unexpected variations. In an additive model, seasonal variation does not depend on time, whereas in a multiplicative model it changes in amplitude. The additive model is expressed as the following:

$$(3)\ y_t = f_t + s_t + c_t + e_t, t = 1,2,3, \dots, N$$

Managing to decompose the time series to its elements and create models for them helps to understand observed development and enables forecasts for future values. (Pal & Prakash 2017)

The first element, the general trend, is the upward or downward movement of measured values over a long period of time. It is often modeled as a regression where time and other known factors are the explanatory variables. The trend line produced can then be used as a predictor for the long run. EDA is a good way of approximating, which kind of regression should be used for modeling. Residuals of the trend line are analyzed for the rest of the time series characteristics.

The second element, seasonality, means repeating periodical variations in observed values. Seasonality can be seen as systemic and predictable fluctuations from a de-trended time series. Cyclical changes are similar to seasonality but occur less frequently and might not have a fixed period. Seasonal variations usually happen within a year, whereas cyclical changes can take several years.

Transforming the data to not include any trend or seasonality should leave a stationary time series. Its values are not dependent on time, meaning it has a constant mean, a constant variance, and a correlation of two points depending only on the time lag between the points, measured as autocovariance or autocorrelation. Autocorrelation is the normalized version of autocovariance and gets its values in the range of -1 to 1. They both measure the linear correlation of two points y_t and y_{t+h} or y_{t-h}, where $h = time\ lag$. Positive autocorrelation means that values h steps ahead move in the same direction as present values, and the opposite direction with negative autocorrelation. If the time series has statistically significant autocorrelation, it can be used to predict values h steps ahead.

Related to autocorrelation, autoregression means that one or more of a variable's previous (lagged) values affect the following value. EDA on autocorrelation helps identify the order of an autoregressive (AR) model, meaning the number of lags with a statistically significant correlation. Autoregressive models can be utilized to make predictions from time series data. An autoregressive model includes one or more lags of the dependent variable as explanatory variables. If a time series process Y_t is a function of its lagged value Y_{t-1} and a stochastic error ε_t, it is an autoregressive process of order 1. The equation for an autoregressive process of order p, AR(p) is:

$$(4)\ y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots + \theta_p y_{t-p} + \varepsilon_t$$

where $\theta_1, \theta_2, \ldots, \theta_p$ are the coefficients and errors ε_t are independent and normally distributed with mean $E(\varepsilon_t) = 0$ and variance $var(\varepsilon_t) = \sigma_\varepsilon^2$.

Another model for time series data is the moving average (MA) model, which has lagged error terms as explanatory variables to predict future values. A moving average model of order q, MA(q) is:

$$(5)\ y_t = \mu + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q}$$

Where, μ is mean intercept, $\theta_1, \theta_2, \ldots, \theta_q$ are error coefficients, and errors ε_t are independent and normally distributed with mean $E(\varepsilon_t) = 0$ and variance $var(\varepsilon_t) = \sigma_\varepsilon^2$. Depending on the time series characteristics, AR and MA models can be used separately or together for value predictions.

Having modeled a time series, only unexpected variations are left. It is the error component that is random and cannot be explained by the other characteristics. It is caused by a lack of information, or by random noise in the data. Having fitted a model to a time series, the left-over error term should be white noise, meaning it should have a mean of 0, constant variance over time, and no autocorrelation (Jabłońska-Sabuka 2018).

2.4 Exploratory analysis methods for time series data

Time series analysis aims to create a mathematical model for explaining observed behavior and forecasting future values. Conducting exploratory data analysis on a time series gives suggestions and insight on how to construct the mathematical model through graphical and statistical summaries of the data. A time series model is either an addition, multiplication or a mix of addition and multiplication of the time series characteristics discussed earlier. To create a model for a time series, each of the components needs to be considered. Conducting EDA helps to determine the composition of the model.

When conducting EDA on a time series, the logical starting point is to plot the time series data as a time plot, where the horizontal axis represents the time index, and observed values are on the vertical axis. From this simple plot, the evolution of the time series can be clearly seen, including two important characteristics of time series data, trend and seasonality. (Huang & Petukhina 2022) Seasonality with a constant variance indicates an additive model, whereas seasonality with a changing variance indicates a multiplicative model. Seasonality is best recognized after de-trending a time series, which is possible with methods such as regression line fitting.

Fitting and plotting a regression line is a good way of revealing long-run trends in the data. The trend can either be a simple linear regression or a non-linear polynomial regression. Having found the best-fitting regression, its residuals can be examined for seasonality. If no seasonality is present, the residuals are stationary. In case there is seasonality, ways of removing it include taking the differences of consecutive values or using a method of moving averages, different from the moving averages MA model. Differencing can also be used to remove the trend in the data. The purpose of removing trends and seasonality is to make the data stationary. Stationarity can be tested with an Augmented Dickey-Fuller (ADF) test, which tests for a unit root or non-stationarity in the data. Being able to reject the null hypothesis means the data is stationary. Python can be used to easily run an ADF test on a dataset. (Pal & Prakash 2017)

The autocorrelation function (ACF) plot shows whether there is a presence of autocorrelation. A partial autocorrelation (PACF) plot leaves out indirect effects, only showing the direct effect of lagged variables, which helps determine the order of an autoregressive model. The order of a moving average model can be approximated from the autocorrelation ACF plot. If both autocorrelation plots show statistically significant autocorrelation, an ARMA model can be considered. Depending on the time series attributes other models such as the ARIMA model can be considered.

3 Exploratory analysis of the case data

Python was selected as the tool for data analysis. It has numerous libraries that help with manipulating, analyzing, and visualizing data, such as NumPy, Pandas, and Matplotlib. The analysis is conducted on data provided by company X, as well as Statistics Finland's (2022a) consumer price index on the category "Car service", a subcategory to "Privately owned vehicles' service and repairs". In this chapter, the steps taken to prepare the data for analysis are explained, followed by an exploratory analysis of maintenance events.

3.1 Preparing the data

The author was provided with several hundred thousand rows of data in 13 columns on vehicles' maintenance and repairs over the last 5 years. The data includes event dates, amounts, descriptions, mileage, vehicle type, make, model, and fuel type, as well as unique invoice and vehicle numbers. The data is time series data, with daily values from January 1st, 2018, to August 31st, 2022. In the dataset, there are multiple values for each point in time, and some gaps most likely caused by weekends and holidays. Because there are multiple measurements for each date, and to make the examination of long-run changes easier, the data is transformed into aggregates such as median values for time periods of varying length.

There are over 200 types of repair and maintenance events, which were divided into 9 categories by their repair codes. In the dataset, each description of a maintenance event ends with a four-digit code, which specifies the type of event. The events were categorized with a Python program made to iterate through each row and assign them a category number based on the repair codes. This study focuses on events categorized as usual scheduled maintenance, as these are performed on all vehicles at regular intervals. The other categories include irregular and unpredictable repairs on different components of the car and will not be examined.

After preparations, the scheduled maintenance category includes around 70 thousand rows of data. The data was cleaned by excluding unnecessary rows, such as events concerning vehicles with the type "van". These are commercially used utility vehicles, which are dropped from the dataset because of their sometimes unpredictably high maintenance costs, incomparable with consumer prices. Also, rows concerning maintenance events such as different filter replacements and windshield wiper changes were dropped due to their low costs, distorting the measurements of location and spread. In the analysis, different car brands will be referred to as A, B, C, et cetera.

3.2 Describing the data

Utilizing the kernel density estimation (KDE) plot to visualize the distribution of maintenance costs shows why resistant statistics are preferred in exploratory data analysis.

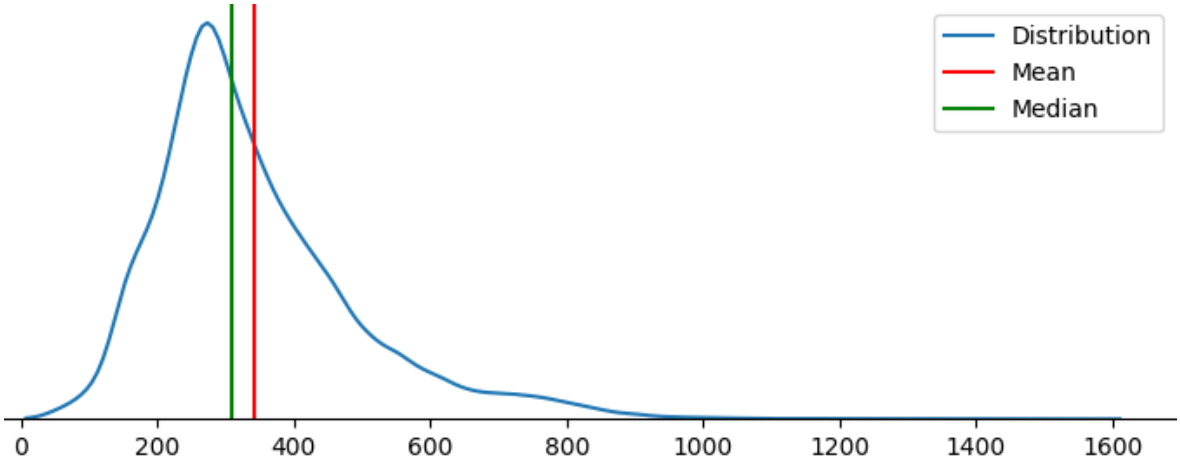

Figure 1 Distribution of maintenance costs

As can be seen from Figure 1, the distribution is not normal but slightly skewed. The skewness can be caused by many different factors, such as more uncommon premium brands having higher maintenance costs, and the fact that cars need more extensive servicing at different points in their lifespan. The absence of a second peak could be caused by many vehicles not being owned by the company long enough for the more extensive servicing to take place. Also, as will be discovered, the number of extensive service events has likely decreased in the analysis period. Furthermore, the period and cost for the more extensive service events vary by make and model. Filtering by make showed two-peaked distributions for some brands, describing the different price points for different events. When filtering the data to only include certain brands at certain mileages the distributions were still not quite normal, and the median was always the better estimate for the center of the distribution. Therefore, the median will be preferred over the mean for the analysis going forward. Also, outliers are removed using the interquartile range (IQR) method, as it does not rely on non-resistant statistics. The IQR method removes outliers 1.5 times the interquartile range below the 25th or above the 75th percentiles. The below Table 1 describes the distribution of the data numerically.

Table 1 Distribution of values (Outliers included)

Count	mean	std	25 %	50 %	75 %
70311	342.99	151.66	243.32	309.16	412.71

In the following analysis of maintenance costs, different aggregates for factors such as brand and mileage will be used. Outliers will always be removed after aggregating because doing so before could cut out meaningful values. For example, using the IQR method for the

unfiltered data as it is described in Table 1, would leave out over 3000 observations above 667 euros, which is still a completely realistic cost for some of the more extensive services done at a franchised dealership. Overall, the distribution of the costs seems completely plausible, meaning the cleaning and categorizing of maintenance events was successful. Yearly price development is quantified in Table 2 below.

Table 2 Yearly median service cost

Year	2018	2019	2020	2021	2022
Median	303.33	311.28	309.76	307.69	316.84

The maintenance costs of the vehicles in the dataset are not completely comparable to measures of consumer prices such as the consumer price index on vehicle maintenance, due to the vehicles being considerably newer than the national average. All of company X's vehicles start off as new, being leased off for a few years depending on the contract length. During this time only the first few maintenance events are performed following factory-recommended servicing schedules. In fact, 75% of the scheduled maintenance events were performed on vehicles with less than 83 thousand kilometers of mileage. In comparison, the average age of passenger vehicles in Finland was 12.6 years in 2021 (Statistics Finland 2022b). Also, the company's vehicle fleet's composition may not be comparable to the general public due to just a few brands representing a large portion of the dataset. The following Table 3 has the number of occurrences of the most popular brands representing 92% of the whole dataset.

Table 3 Occurrences of most popular makes

Make	A	B	C	D	E	F	G	H	I	J
Occurrences	16758	10502	8974	6610	5735	5136	4135	3865	1563	1388

From Table 3, it can be seen that there are large differences in the popularity of different makes, with only the three most popular brands A, B, and C accounting for about half of all scheduled maintenance events with 52% of all 70 311 occurrences. Sudden changes in the service costs of popular brands could unproportionally affect overall median values, as could sudden changes in the composition of brands in the fleet. Below, Figure 2 shows the most popular brands' yearly percentage out of all services. The composition of the fleet shows some changes, most notably brand C seems to have been gaining popularity compared to the others.

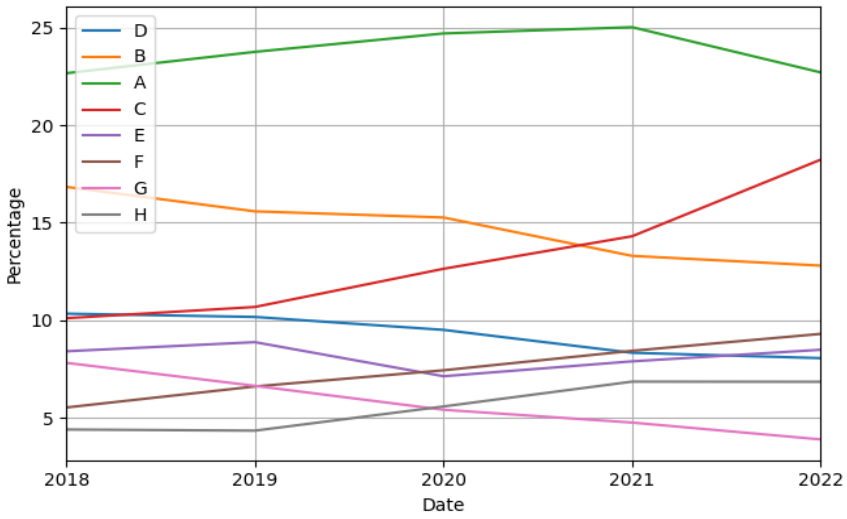

Figure 2 Percentages of most popular brands

As mentioned, company X operates a fleet of comparably new vehicles, which is also seen in Table 4, showing the top 10 most frequent maintenance events representing 84% of all observations.

Table 4 Occurrences of different service events

Event name	30 000km	Yearly service	Longlife service	60 000km	90 000km	120 000km	15 000km	20 000km	25 000km	45 000km
Occurrences	11030	10774	10684	10105	5989	3499	2287	1546	1478	1465

In addition to mileage-specific events, both “Yearly service” and “Longlife service” are in the top three events, and unlike the rest, these are not bound to a specific point of a vehicle’s lifespan. “Longlife service” is a maintenance event specific to Volkswagen group vehicles. It does not have a specific service interval, and instead, the service interval adapts to driving style and conditions of use with a maximum interval of 30 000km or 2 years (Volkswagen 2022). The “Yearly service” event has vehicles from several manufacturers with different mileages and varying costs. When plotting the data, mileage will be accounted for by only selecting portions of the dataset by giving limits to the vehicle mileage variable. Much of the analysis will concentrate on service events done on cars with 20 to 40 thousand kilometers driven, as this is often the time for the first servicing of most cars and is most likely similar in content for all brands.

The following heatmap in Figure 3 shows how median prices of different mileage-specific maintenance events are distributed for different brands. In the plot, each column describes an individual brand’s service costs at various points of a car’s lifespan. From the plot, mileage for the more extensive services can be recognized as the rows with the highest values. It can be seen that the service interval for the more extensive service events varies by brand as

different columns have their highest values at different mileages. Also, it would seem that the average cost of maintenance varies greatly by brand.

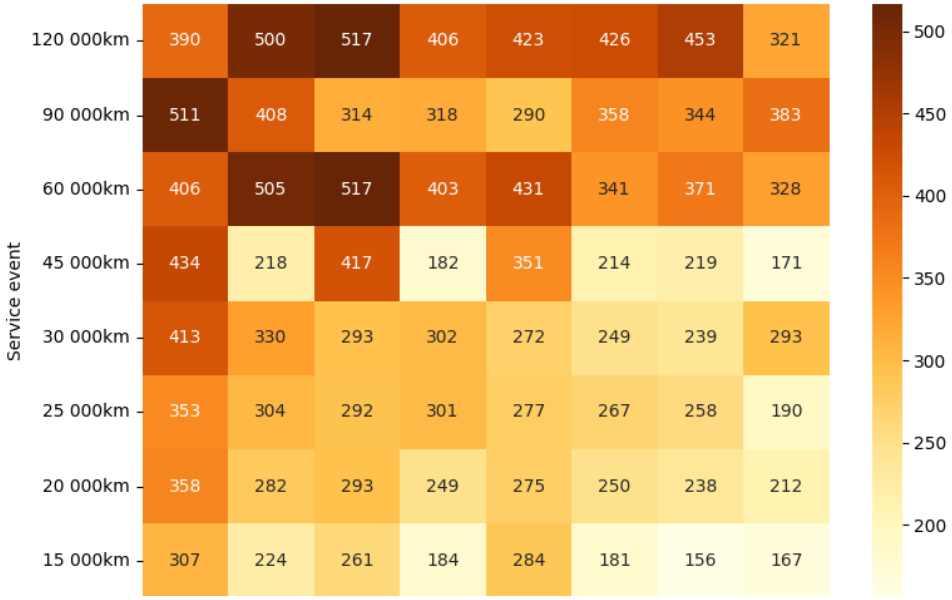

Figure 3 Heatmap of median cost by service event and brand

3.3 Analyzing overall vehicle maintenance cost development

Figure 4 has a time plot of company X's monthly median service costs with a 4^{th}-degree polynomial trend line fitted with least squares estimation, as well as the consumer price index trend line. The consumer price index line y-axis position is not relevant, as it only includes data on the percentual monthly change and is there only for comparison of the line slope. We can see the cost of maintenance fluctuate and go down at times, even though consumer prices of vehicle maintenance have steadily increased. This leads to the conclusion that other factors besides time have affected the average monthly cost. One such factor is the median mileage of the vehicles, which is plotted below the cost development and inherits a trend very similar to the one in the cost plot. Percentual changes for the trends of both the cost and kilometers were calculated, and a strong Pearson correlation of 0.81 was measured between them, indicating a connection. It was mentioned earlier that services at different points of a vehicle's lifespan can have different costs, and the below plots suggest that as the customers have driven fewer kilometers, the costlier services have been delayed. Also, an examination of the monthly count of maintenance events showed no clear declining trend, indicating that the vehicles were serviced according to the time between services as opposed to mileage.

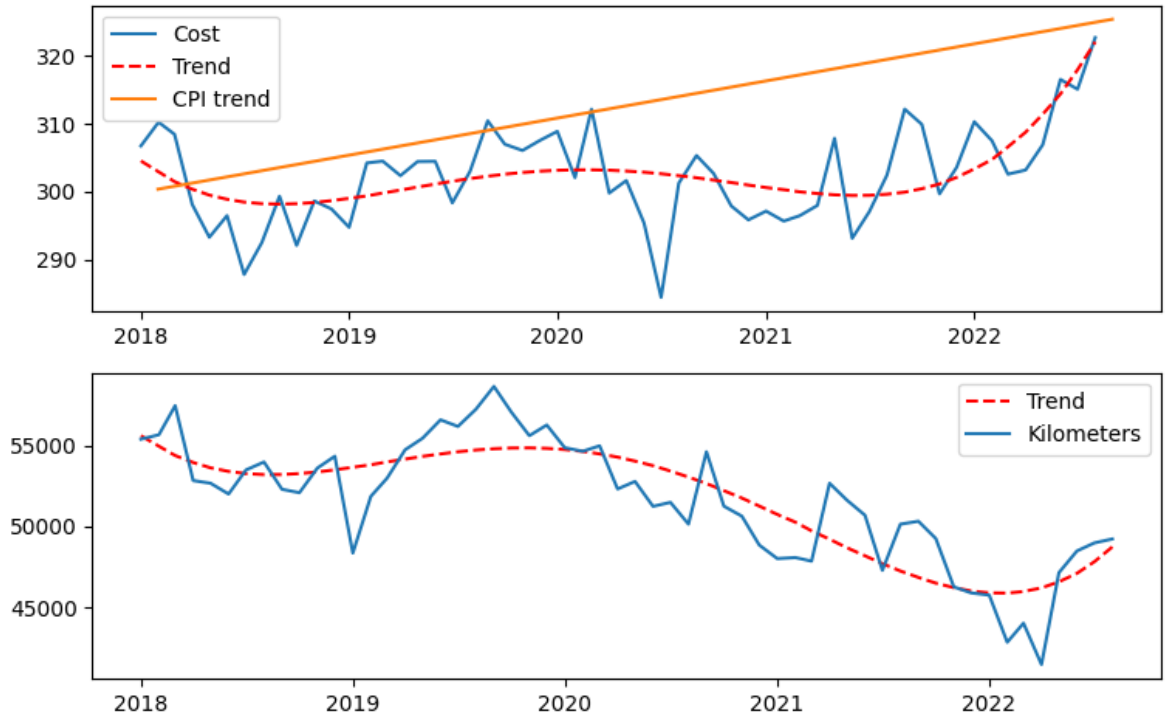

Figure 4Median monthly service cost & mileage

According to The Finnish Transport and Communications Agency Traficom (2022a & 2022b), total car kilometers in Finland decreased by 4% in 2020 due to the Covid-19 pandemic and continued to decline in 2021 by 0.8%, suggesting the pandemic could be one reason for the fluctuations in median service costs.

3.3.1 Decomposition of overall maintenance cost development

Figure 4 cost plot displays some months clearly deviate from the fitted trend, giving reason to also examine the data for seasonality and autocorrelation. Decomposing the time series to trend, seasonality, and residuals can be done with Python's statsmodels module seasonal decompose function. This time weekly medians were taken as opposed to monthly medians as in Figure 4. No clear change in variance amplitude over time was visible, so an additive model was estimated with a seasonal frequency of one year. In Figure 5, the plots produced by the function show the polynomial trend, which was already known, as well as a yearly seasonality with a decrease in costs around July every year. The residuals were tested for stationarity, which was confirmed by an augmented Dickey-Fuller test p-value of 0.001, rejecting the null hypothesis of non-stationarity. As the function estimates a systematic seasonality, the drop every July might be a bit exaggerated due to the larger decline in July

2020. Nevertheless, looking closely at the weekly medians does show some decline in costs almost every summer, which could be caused by summer holidays for example.

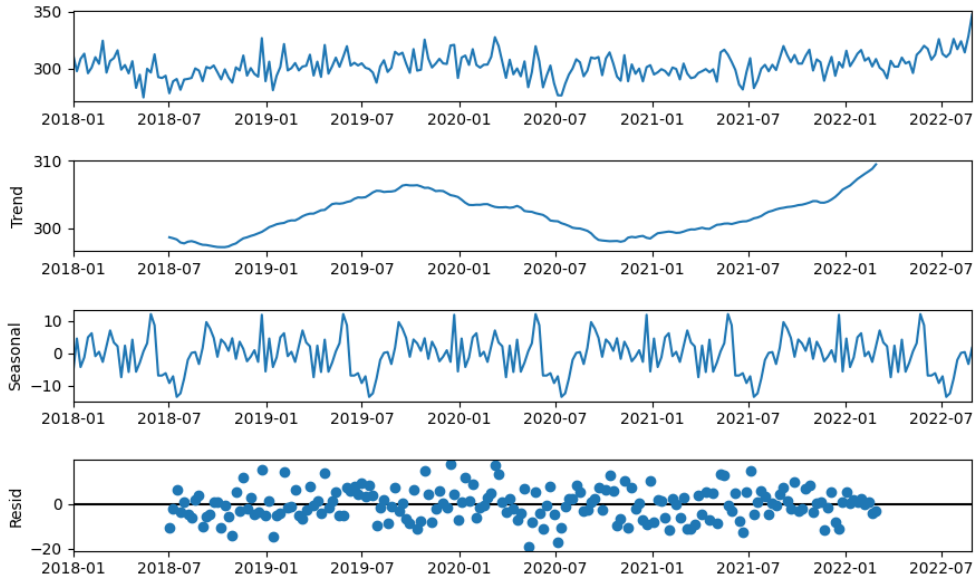

Figure 5 Weekly median cost decomposition

In Figure 6, plotting the autocorrelation and partial autocorrelation plots for the residuals show barely statistically significant autocorrelation at lag 7, indicating that seasonality and trend do not completely describe the time series. This means more complex models with AR and/or MA elements could be considered for forecasting.

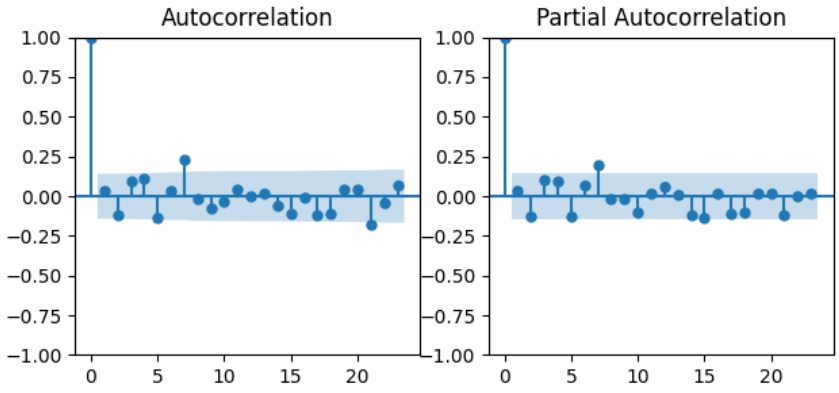

Figure 6 ACF and PACF plots for decomposition residuals

3.3.2 Comparison of the inflation rate and median cost development

In Figure 4, a linear trend was shown for the consumer price index, even though general inflation rate is known to have increased significantly from the end of 2021. Figure 7 shows the development of the monthly change percentages for the consumer price index by Statistics Finland and the observed values in the dataset. It can be seen that the rising overall

inflation rate (CPI) has not been reflected in the consumer prices for car services (Car Service CPI) as of August 2022. Looking at the monthly change rates for the fleet's observed median values clearly shows inflation is not the only factor affecting their development. The monthly change rates for the fleet's observed values do not fall anywhere near the rate of inflation for car services due to factors inherent in the dataset. It is already assumed that the changes in overall median mileage are in part causing the decline in overall costs but limiting the mileage to only show the rate for first services done at 20 to 40 thousand kilometers also shows a differing rate far higher than the measured consumer prices. The following analysis aims to uncover more factors determining the cost development of the fleet's services.

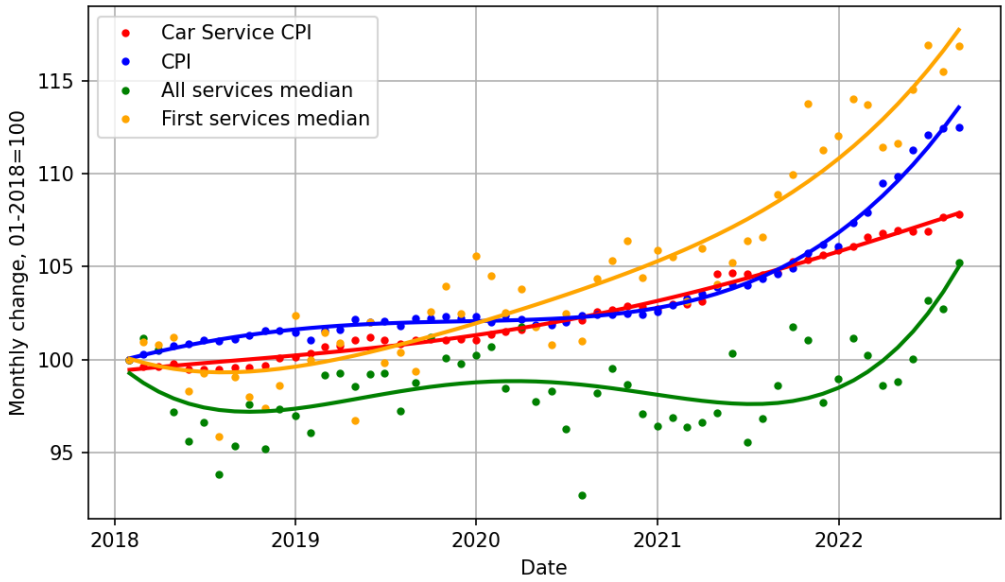

Figure 7 Monthly change in consumer prices and median fleet service costs

3.4 Analyzing vehicle maintenance by mileage and fuel type

It should now be confirmed that different service events do in fact have different costs, and the following plot of the eight most popular mileage-specific services shows just that.

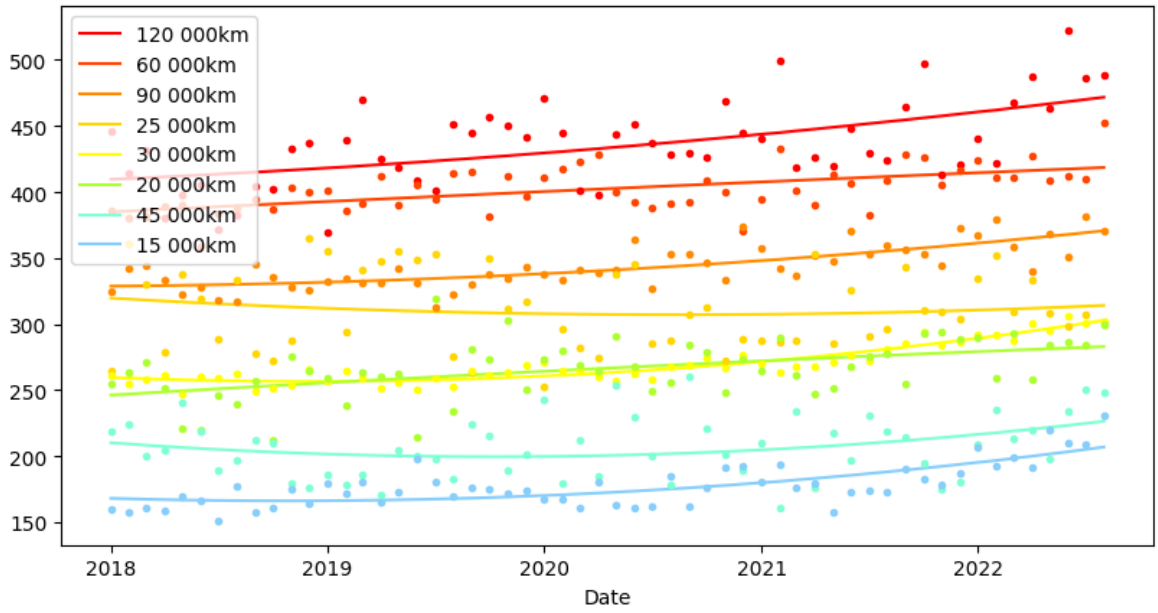

Figure 8 Monthly median cost by mileage with trendlines

Figure 8, a plot of the median costs of the eight most common mileage-specific events shows the effect of mileage on costs, with each event clearly in its own price range. From the graph it could be said that 60 thousand kilometers is probably the interval for most of the fleet's more extensive maintenance events, as 60- and 120-thousand-kilometer services seem to be the most expensive ones. This is supported in a Finnish vehicle workshop Oilpoint's service rates (2022), where depending on the manufacturer, extensive servicing is recommended every 40 to 90 thousand, and intermediate servicing every 15 to 30 thousand kilometers.

A similar comparison was plotted for weekly medians by fuel type in Figure 9. Electric and liquid petroleum gas (LPG) vehicles were left out due to them being too uncommon for meaningful trend fitting.

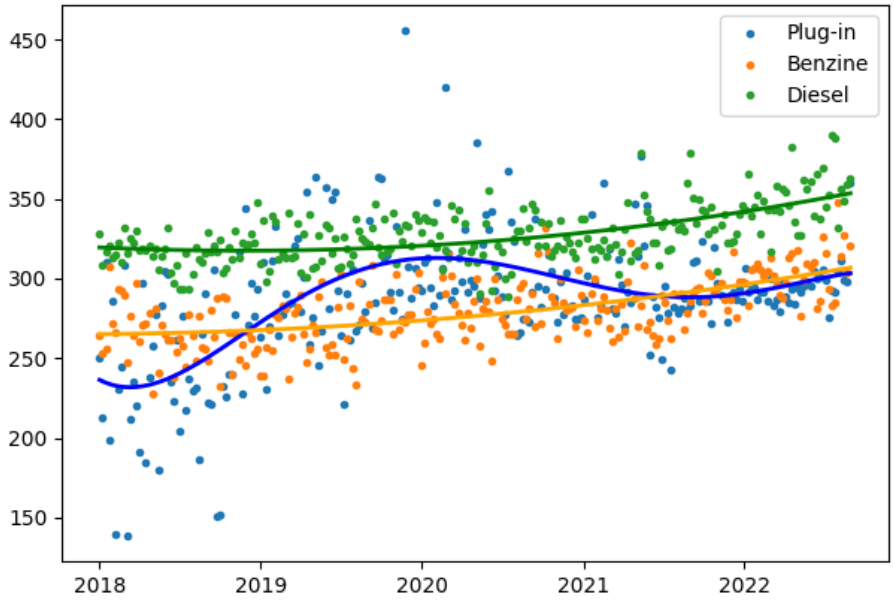

Figure 9 Weekly median cost by fuel type

From the plot, it can be seen that diesel and gasoline cars have a very similar trend upward, while Plug-in hybrids have a wildly fluctuating trend. One reason for gasoline cars' services being continuously less expensive than those of diesel cars could be that larger and more expensive premium cars often use diesel, whereas small economy cars are usually gasoline powered. Figure 10 of the weekly median kilometers by fuel type also gives some insight into the differences in the observed cost trends.

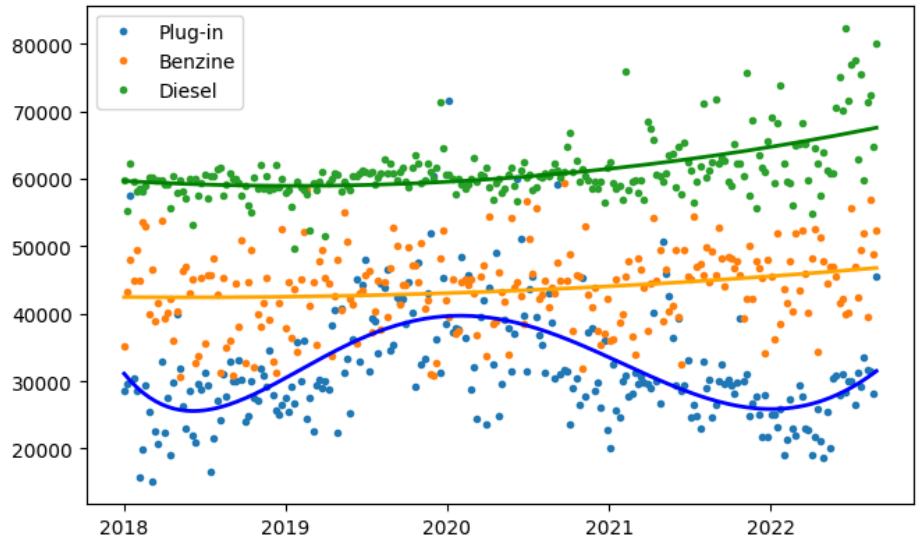

Figure 10 Weekly median kilometers by fuel type

A vehicle's fuel type is strongly connected to its annual mileage, likely due to differences in fuel consumption, fuel price, and range. Figure 10 suggests the average mileage at service

explains some of the difference in costs for different fuel types. Diesel cars seem to be driven a lot more compared to gasoline and plug-in hybrid cars, which is reflected in the cost of maintenance. Diesel cars being driven more makes sense because in Finland they are often considered the better choice for high mileage use, as the lower fuel consumption and often a lower fuel price can outweigh the higher taxation. Like the polynomial trends in Figure 4 overall cost and mileage, the polynomial trends for plug-in hybrids' cost and mileage have similar fluctuations, indicating a connection. As plug-ins are the least common type in the plot, they are more sensitive to fluctuations caused by specific models being popular at different times.

From Figure 10 it can be seen that both diesel and gasoline cars have a fairly linear trend upward, somewhat contradicting what was seen in figure 4 median monthly mileage plot where the values fluctuated and seemed to have a declining overall trend. The following Figure 11 shows the percentage for the most popular fuel types each month, displaying a drastic decline in the popularity of diesel vehicles and an increase in the popularity of plug-in hybrid vehicles. This would indicate another reason for the decline in figure 4 median mileage, besides the pandemic affecting people's driving habits, which could be the declining number of diesel vehicles in the fleet as they usually have the highest mileage.

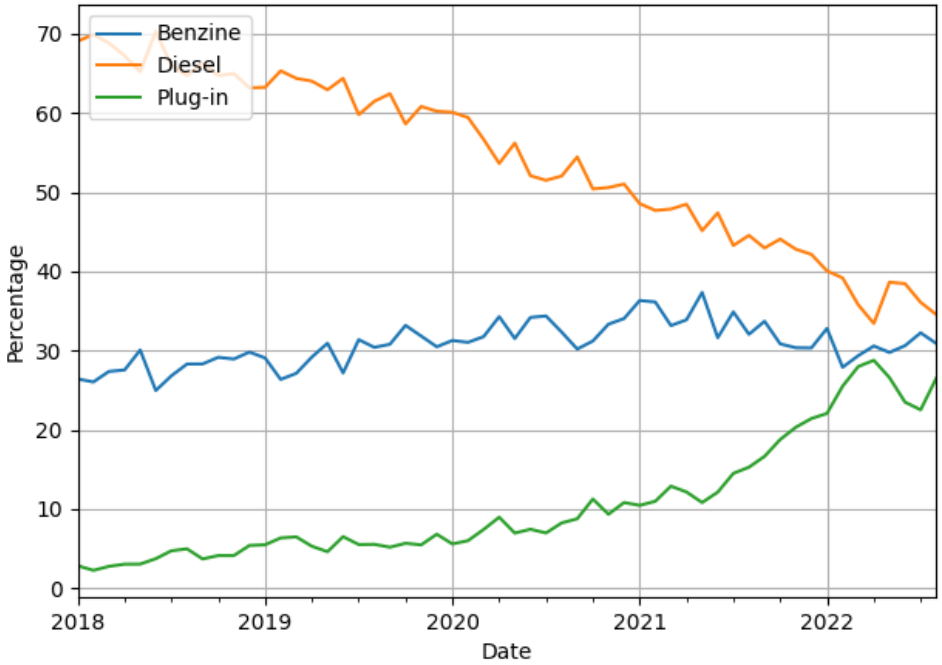

Figure 11 Monthly percentage of fuel types in fleet vehicles

3.5 Analyzing vehicle maintenance by brand

Figure 12 is a plot of the median cost development for the most popular makes with a 2nd-degree regression line fitted for each. The plot shows how brands are in different price categories, which can cause fluctuations in overall median cost development as the concentrations of different brands in the fleet change. Similarly to what was discussed with different fuel types, the size of the vehicle could have some effect on the brand comparison as well. As some makes can be more popular for gasoline-powered and often smaller economy cars while others might concentrate on offering more "premium" vehicles often also larger in size.

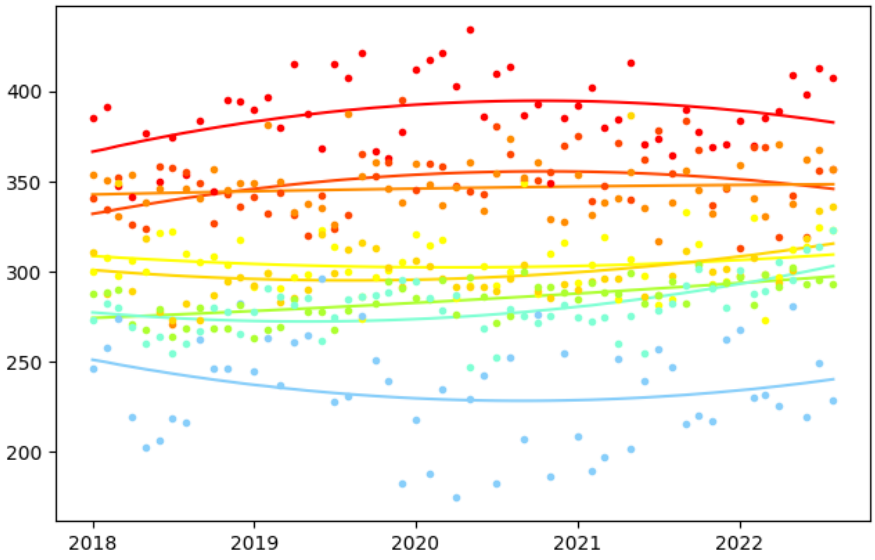

Figure 12 Monthly median cost by brand

The data is now filtered for only the first maintenance events, usually done at 20 to 40 thousand kilometers, and the effect of different concentrations of different makes is removed by taking the weekly medians of all car brands and taking their mean. Now the following time plot in Figure 13 can be drawn, with a clear upwards trend, visualized with a least square fitted trend line. Also, the car service consumer price index monthly change values were utilized to calculate the orange CPI line. Again, it does not measure average consumer price as the consumer price index only measures the change in prices. This means its position on the y-axis is meaningless and it was only plotted to show the similarity of the slope with the trend in the data. The plot places equal weight on all vehicle makes, meaning it does not describe the weekly average cost of the fleet well. However, it suggests that the prices charged on the first services of the fleet vehicles have increased at a similar rate to consumer

prices, which also consist of a wider range of makes and models. Compared to Figure 7 where the median for the first services exhibited a trend far steeper than that of the consumer price index, it seems that either individual makes or their concentrations have affected the cost development.

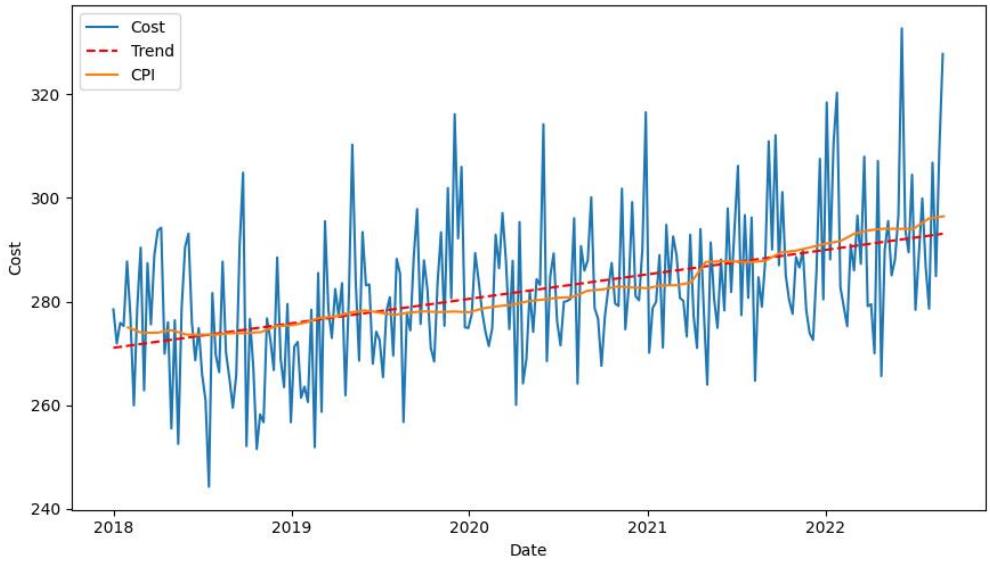

Figure 13 20-40 thousand km weekly service cost mean of brand medians

3.5.1 Brand A's maintenance cost development

Not visible in Figure 12 brand cost comparison where different brands displayed fairly linear trend lines, limiting the observations by mileage shows interesting brand-specific development affecting the average prices of the whole fleet. By far the most popular car brand in the dataset is brand A with over 16 thousand service observations. The following Figure 14 describes the price development of internal combustion engine (ICE) powered brand A cars' first services. The analysis is narrowed down to include only ICE vehicles to get more accurate estimates.

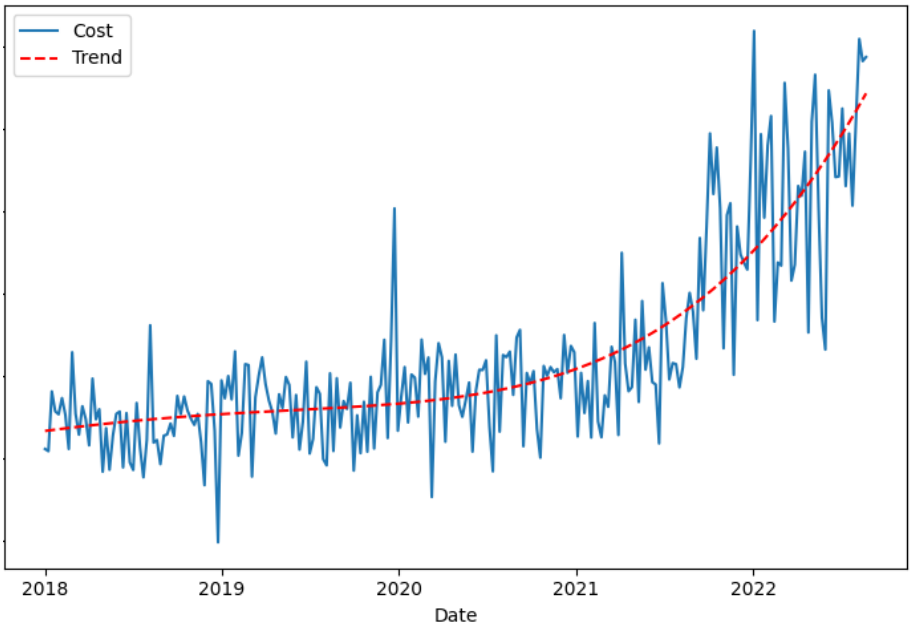

Figure 14 20–40-thousand-kilometer ICE brand A weekly median

There is a clear accelerating trend in the data, best described by a polynomial trend line. Service rates for brand A vehicles explained some of the development was due to a notable increase in the service cost for their most popular model, caused by a launch of a new generation of the model. However, similar development was also observed in the dataset for brand A models that had not been updated. The content of at least the most popular model's service was somehow fundamentally changed when the new generation was introduced, but this does not explain the change in costs for other models. The observed increase in cost for models surpassing their quoted cost would indicate the rising cost is not only caused by the new model's changed service rates but might also have something to do with the service provider.

Brands B and D are also popular and known to have comparable service programs to brand A, so their development will also be examined.

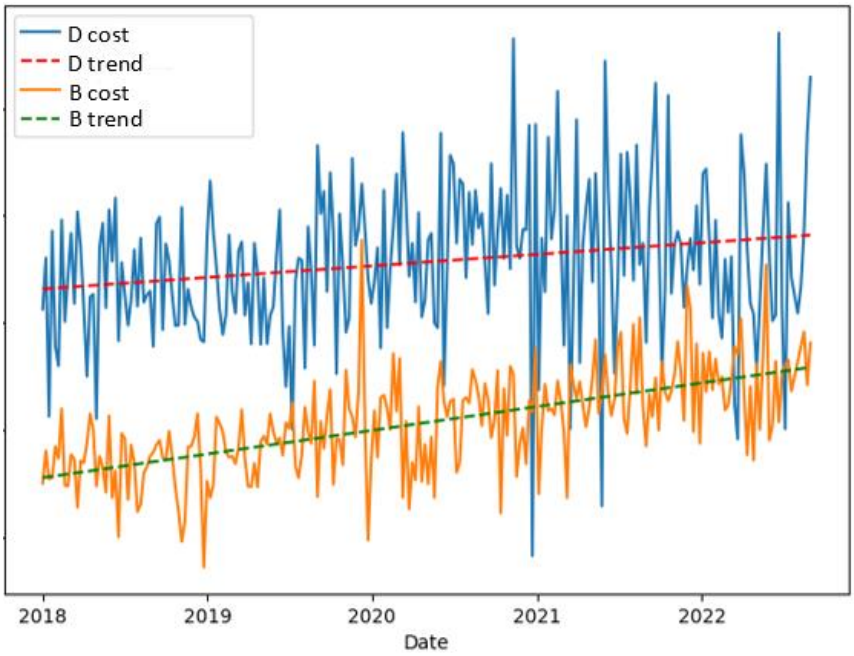

Figure 15 20-40 thousand km service cost brands B & D weekly median

In Figure 15, both brands B and D show an upward linear trend with B having a slightly better fit thanks to less spread. Brand D's line fit is not as good due to a higher spread of values, especially near the end of the period when the number of observations had significantly decreased. Neither brand B nor D shows similar development to brand A, giving reason to believe the increase in brand A cost is due to factors not concerning other brands. It also seems that the observed change in the monthly change rate for first services in figure 7 was at least in part due to the increase in brand A's service costs.

3.6 ARIMA model for internal combustion engine vehicles

So far, the analysis has shown how various factors influence the development of average maintenance costs, causing deviations from the development observed in consumer prices. Changes in factors such as mileage, fleet composition, and costs for individual makes or models affect the measured average cost. As these unpredictable factors are unique to the fleet in question, they make the estimation of future costs based on inflation alone inaccurate. To forecast the fleets' maintenance costs, models based on the fleet data are needed.

Estimating mathematical forecasting models is out of scope for exploratory data analysis, but as available Python libraries make it fairly simple, a model will be estimated focusing on graphical visualizations when determining the composition. The model estimates the median cost of maintenance considering only ICE-powered vehicles' first services. Mileage and fuel types were limited for more accurate estimates. An ARIMA model was selected as it has an integrated component for removing non-stationarity. AR models are good for forecasting values of a function where a trend is expected to retain, such as a stock price of a company (Pal & Prakash 2017). In this case, the costs are expected to rise in the long run due to factors such as inflation.

Figure 16 depicts the development of all ICE-powered vehicles' monthly median service cost. There is clearly a non-linear increasing trend in the data, that needs to be removed to make the data stationary. The trend will be removed using first-order differencing, which means taking the difference between each value and the value preceding it. The differenced time series is plotted in figure 17.

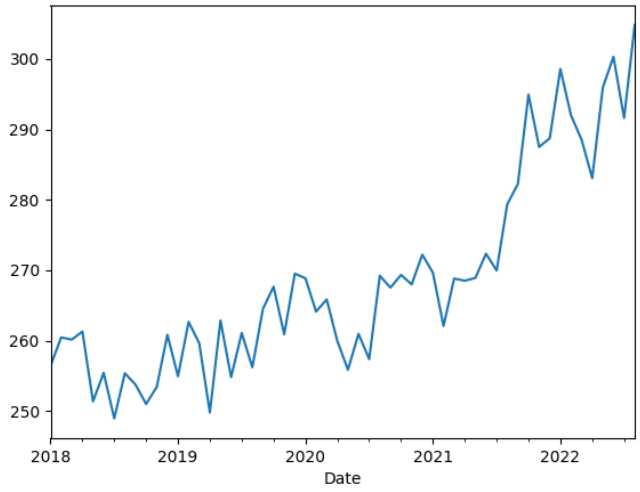

Figure 16 ICE vehicles 20-40 thousand km monthly median cost

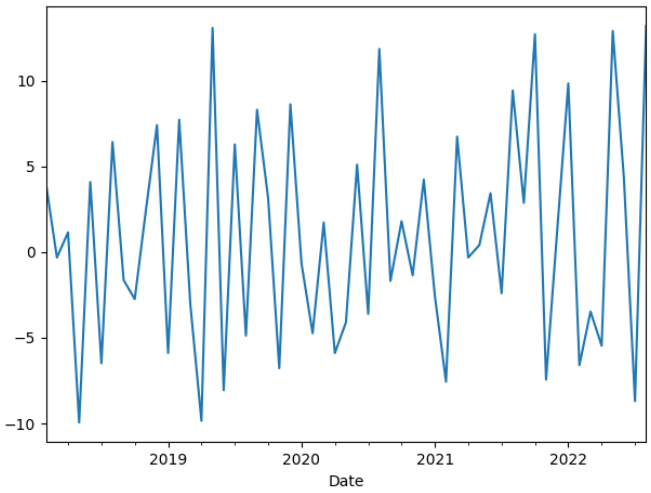

Figure 17 ICE vehicles 20-40 thousand km monthly median first difference

In Figure 17, the residuals from the differencing seem stationary, as the mean and variance appear to be constant. Python module statsmodels has an augmented Dickey-Fuller test, which was used to test the stationarity of the residuals. The test gave a p-value of less than 0.01, meaning the null hypothesis is rejected at a 99% confidence level and the data is considered stationary.

Next, the autocorrelation and partial autocorrelation were plotted, to see if AR or MA models can be utilized. From the plots in Figure 18, it can be seen that there is statistically significant autocorrelation in both plots. As the data had to be differenced to remove the trend, and there is reason to include both AR and MA components, an ARIMA model can be utilized to create estimates. From the partial autocorrelation function we could estimate the order of the autoregressive component to be 1 and from the autocorrelation plot the order of the moving average component to be 1 as well. The I or integrated component is 1, from the first order difference.

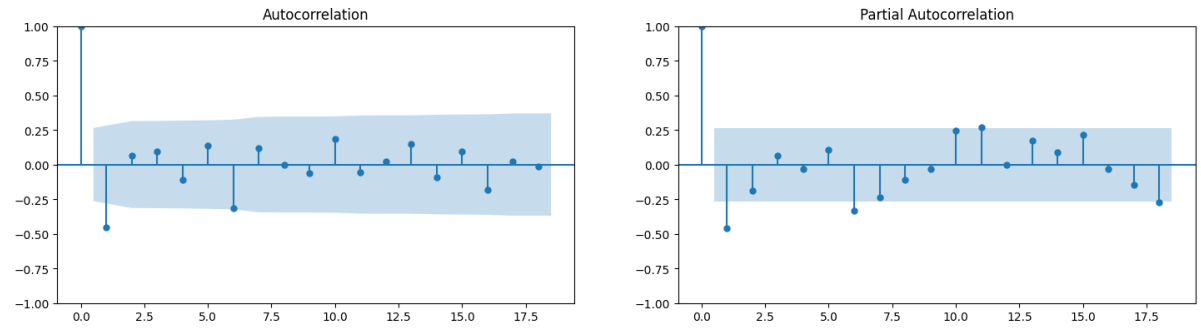

Figure 18 ACF and PACF plots for the ARIMA model

After a few iterations, an ARIMA(1,2,1) model with a second order of differencing was found to give the most accurate estimates. This means that the model estimates are based on 1st-order autoregression, 2nd-order differencing, and 1st-order moving averages. From a stats-models summary table, the model coefficients were all checked to be statistically significant, as well as the Akaike information criterion (AIC) which measures the overall goodness of fit was checked. The AR, I, and MA values for the model were selected by iterating over all reasonable combinations and selecting the one with the best AIC value. The following Figure 19 forecast plot was made using the selected model. It seems to capture the overall trend well, and as the observed values end, the fluctuations settle down and an increasing linear trend is shown. The 95% confidence interval widens as forecasting the future becomes more difficult.

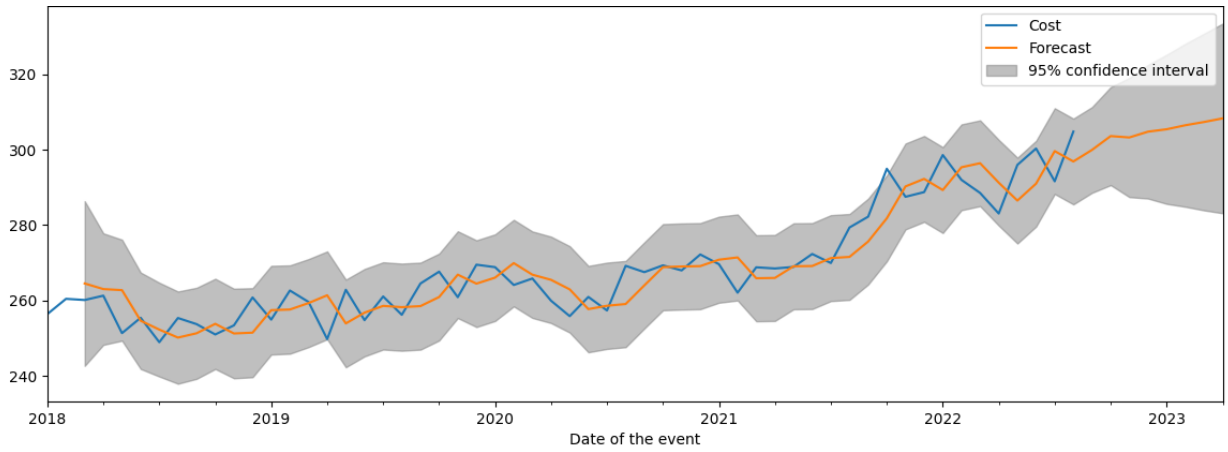

Figure 19 ARIMA model forecast for ICE 20-40 thousand km service cost

Finally, the model residuals are plotted to check if they are white noise. Python statsmodels has a diagnostics function that can be used to plot the distribution of values. In Figure 20, the standardized residual plot shows mean 0 and constant variance. The normal Q-Q plot shows the sample and estimated data values roughly come from the same normal distribution. The kernel density estimate (KDE) over the histogram shows a distribution similar to the zero mean normal distribution. And lastly, the correlogram shows a barely significant autocorrelation at lag 6, indicating that something is left unexplained by the model, however further modeling goes beyond the scope of this study, and the estimated model is considered good enough.

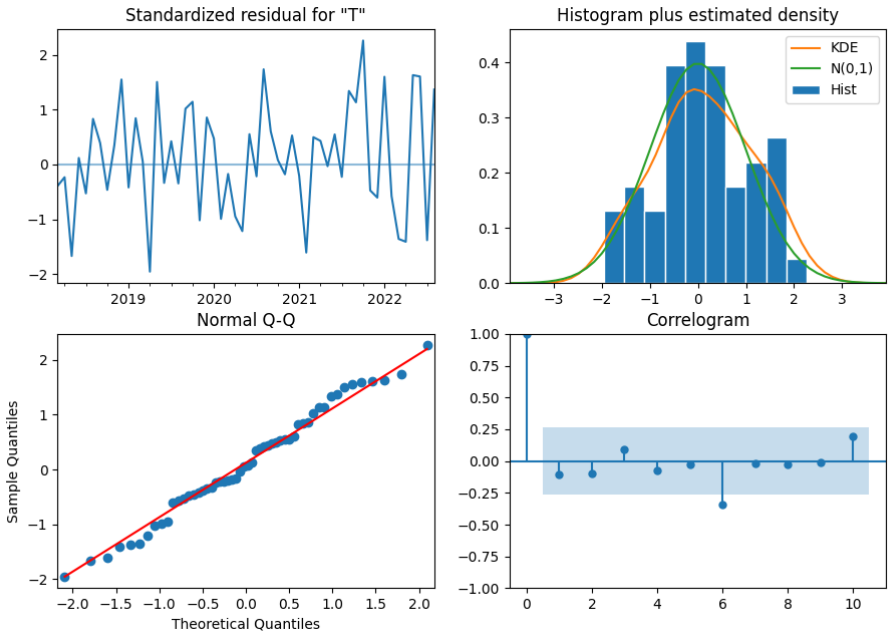

Figure 20 ARIMA model residuals

3.7 Results analysis

Starting with the overall median maintenance cost development in Figure 4, it was shown that the median cost is influenced by the fluctuating median mileage, causing the cost development to look completely different in comparison to the development in the consumer price index. One possible reason for the observed decline in median mileage could be the covid-19 pandemic which caused a decline in driven kilometers on a national level. The decline in the dataset was however larger than the national decline reported by Traficom, which could in part be explained by the radical decrease of diesel-powered vehicles in the fleet as was shown in Figure 11. In Figure 10 aggregating the data by fuel type showed each fuel type exhibiting its own range for median mileage. Diesel cars get the highest mileage, whereas plug-in hybrids get the lowest. The median mileage for diesel cars seemed to only have increased, suggesting that the declining overall median mileage is partly caused by the decline in diesel car popularity and an increase in plug-in hybrids' popularity. It would seem that either the customers are switching to electrified vehicles and driving less, or the customer base is changing with customers who drive a lot leaving and new customers favoring electrified powertrains.

It was apparent from Statistics Finland's consumer price index monthly change percentages, that the increasing rate of general inflation was not reflected in the inflation rate for car services, which had a linear trend for the observed period. It would be fair to assume that the cost development of specific service events for individual brands would roughly follow the rate of inflation for car services, however, as was shown in Figure 14 of brand A's first services, this can be a misleading assumption to make. A seemingly unexplainable increase in the service costs for internal combustion engine brand A services done at 20 to 40 thousand kilometers was observed, with no similar development for two other brands B and D. The service provider for the by far most popular brand in terms of the number of service events, brand A, seemed to have significantly increased their service rates. The service cost for the new generation of brand A's most popular model could be higher because of more advanced technology for example, but it does not explain the similar jump in older models' costs. The increase in brand A service costs showed how sudden changes in service costs can happen for individual brands, affecting the overall costs.

Some kind of vehicle class variable based on models could be a good addition, as it could possibly capture a combination of the effects of multiple factors. Assuming vehicles in the same segment by different manufacturers are competing for the same customers, they would have similar attributes. Calculating the median cost for different vehicle segments could capture the unique distributions of different fuel types, brands, and average mileages inherent to the different classes. Class variables could be used to create forecasts for similar vehicles, as filtering the data to specific cars might not leave enough observations for meaningful model estimates.

As the fleet vehicles are serviced in the same workshops as consumer vehicles, the costs of individual service events most likely develop at the same rate as is measured in the consumer price index. However, the unique composition of the fleet and customers' changing driving habits make predictions of average costs based on inflationary trends unreliable. An ARIMA model forecast for 20-to-40-thousand-kilometer services of internal combustion engine vehicles attempted to capture all factors affecting cost development by predicting costs based on previous values. Using it to forecast a few months ahead showed a similar upwards trend as was captured in the observation period not considering the sudden increase in 2021, possibly caused by the change in brand A's service rates.

4 Conclusion

In the introduction section, two research questions were presented. They were: “What are the most commonly used exploratory data analysis methods, and how to apply them on time series data?” and “What kind of development can be seen in the vehicle maintenance cost data, and what are the key insights with respect to general inflation?” These questions were answered by first reviewing central EDA and time series analysis methods, followed by an analysis of the case dataset with the aforementioned methods. This chapter concludes the study by bringing together and summarizing the most important findings.

4.1 RQ1: The most common EDA methods and how they were applied

Exploratory data analysis relies heavily on the use of graphical visualizations and summary statistics. The data was aggregated and filtered to represent different aspects of the dataset, and then plotted using the most suitable graphical methods. Categorizing the data allowed the examination of different variables separately both with numerical summaries and different plots.

The basic composition of a time series was deconstructed with the help of both manual trend fitting, as well as Python statsmodels decomposition function. Plotting both linear and polynomial trend lines proved to be the most effective way of understanding the long-term evolution of the data. The data was made stationary with the decomposition function, as well as differencing, after which the residuals were plotted for autocorrelation using both ACF and PACF plots. Stationarity of the residuals was tested with an augmented Dickey-Fuller test, which was also readily available as a function of Python statsmodels.

Time series autoregressive integrated moving average (ARIMA) modeling was also touched on with a focus on recognizing attributes from different graphical visualizations, such as the usual time plot, residual plot, and autocorrelation plots. The residuals were also examined for white noise by observing their distribution from a series of different plots.

4.2 RQ2: Development in the data and key insights

The analysis was narrowed down to only include the costs from usual scheduled maintenance, as it is done on all vehicles regularly, making it predictable and easier to interpret. Also, the price for scheduled maintenance is usually predetermined, which is not the case with repairs caused by sudden breakdowns or accidents.

The development of maintenance costs was analyzed by looking at both the overall trend, as well as breaking the dataset down to show the effect of different variables including makes, mileages, and fuel types separately. Plotting by categorical variables was always done for values with most observations to get meaningful trends with reasonable variances. When plotting by mileage-specific events, 60- and 120-thousand-kilometer services were found out to be the most expensive. Plotting by make showed considerable differences in costs for different brands and plotting by fuel type showed diesel cars' services to be more expensive than those of gasoline cars.

The plots for different makes and fuel types could be taken to suggest that the size of the vehicle affects service prices. Unfortunately, there is no feasible way for filtering the data by vehicle size. Having some kind of a vehicle class variable based on vehicle weight or model for example would enable testing of cost predictions for similar vehicles from different brands. This could be valuable, as filtering the data by model narrows the number of observations down too much to create meaningful estimates. Also, as car models change over time, long-term examination does not produce meaningful results. Vehicle classes however stay the same, with large saloons and estates likely always having a higher cost of maintenance compared to smaller economy cars.

One major finding was the effect of reduced median mileage possibly caused by the Covid-19 pandemic and the declining number of diesel vehicles. A clear reduction in the number of kilometers driven correlated strongly with the abnormal decrease in median maintenance costs. Also, filtering the data to only include service events of 20- to 40-thousand-kilometers and weighing all makes equally showed a trend that almost perfectly followed the consumer price index monthly change. This could be interpreted to mean that company X's maintenance costs do in fact follow the CPI when the effect of mileage and make is minimized.

Finally, an ARIMA model was estimated for forecasting values of internal combustion engine cars' first services at 20- to 40-thousand kilometers. The estimated model residuals did give reason to rethink the composition of the model, but the overall model fit was good and gave a completely reasonable forecast for the end of the year 2022, with costs continuing to rise but at a slower rate than during the last measured year where the prices of brand A's services suddenly increased.

4.3 Future research

Different hypotheses for the development of service costs have been formed with exploratory data analysis methods. Further research can be done to confirm the validity of the made claims and to further study the development of recognized factors.

Cost development of individual models should be researched, especially for brand A vehicles, for which it would seem new generations of current models will have noticeably higher maintenance costs. Also, as brand A's newer generation models start to get their extensive servicing done, similar development to that of their first services might be seen for these services as well which could have a noticeable effect on overall costs. Similar changes can of course be made for any brand as new models are released.

The residual value analysis for the ARIMA model showed something was not accounted for as there was presence of autocorrelation. The model could be improved and similar forecasts could be estimated for different aggregates, such as the monthly median for a specific service event for different brands. Further models could be estimated taking into account that the forecasts' accuracy declines fast the further it is estimated, and sudden fundamental changes such as the one observed in brand A's service costs are difficult to predict.

One factor that should also be investigated further is the declining median mileage. Is the company's customer base changing, or are the driving habits of the customers changing at a rate faster than the national average? The decline seems to be in some part due to the portion of diesel cars shrinking. If the amount of diesel cars is going down, how come other fuel types' median mileages do not go up to fulfill the customers' need for personal transport?

One more interesting topic for future research is the electrification of the fleet, which was apparent from the growing proportion of plug-in hybrid vehicles in the data. Fully electric

vehicles are also becoming more popular, but their proportion of the dataset was still very small as cars are usually serviced a year after registration at the earliest. The maintenance costs of electric vehicles would be a good future research topic, as fully electric drivetrains are completely different from internal combustion engines and have different service procedures.

References

Autoalan tiedotuskeskus. (2021) Tutkimus auton valinnasta ja hankintatavoista. [online]. [Accessed 7 October, 2022]. Available https://www.aut.fi/files/2498/autojen_hankintatapatutkimus_2021.pdf

Baklouti, A., Schutz, J., Dellagi, S. & Chelbi, A. (2022) Selling or leasing used vehicles considering their energetic type, the potential demand for leasing, and the expected maintenance costs. Energy Reports 8, 9, 1125-1135.

European Central Bank. (2022) Measuring inflation – the Harmonised Index of Consumer Prices (HICP). [online]. [Accessed 8 October, 2022]. Available https://www.ecb.europa.eu/stats/macroeconomic_and_sectoral/hicp/html/index.en.html

Feng, X., Xu, Y., Ni, G. & Dai, Y. (2018) Online leasing problem with price fluctuations under the consumer price index. Journal of combinatorial optimization 36, 2, 493-507.

Hanhinen, H. (2018) Kuinka paljon autoilu maksaa? Yle pyysi ja Autoliitto laski esimerkit – ”Itsensä huijaamista, jos ajattelee vain käyttökustannuksia”. Yle. [online]. [Accessed 8 October, 2022]. Available https://yle.fi/uutiset/3-10042081

Hartwig, F. & Dearing, B. E. (1979) Exploratory data analysis. Newbury Park, SAGE.

Huang, C. & Petukhina, A. (2022) Applied Time Series Analysis and Forecasting with Python. Cham: Springer International Publishing AG.

Jabłońska-Sabuka, M. (2018) Advanced Course in Statistical Methods. Week 5. LUT University, Computational Engineering.

Lindholm, T., Kettunen, J. & Turunen, J. (2016) Globaali kansantalous. 1st ed. Helsinki, Edita.

Lähitapiola. (2022) Poikkeukselliset ajat autokaupassa – kokosimme vinkkejä auton hankintaa harkitsevalle. [online]. [Accessed 8 October, 2022]. Available https://www.lahitapiola.fi/tietoa-lahitapiolasta/uutishuone/uutiset-ja-tiedotteet/uutiset/uutinen/1509577903102

Oilpoint. (2022) Autohuollon ja määräaikaishuollon hinnasto. [online]. [Accessed 29 November, 2022]. Available https://oilpoint.fi/hinnastot/huoltohinnasto

Pal, A. & Prakash, P. (2017) Practical time-series analysis : master time series data processing, visualization, and modeling using python. 1st edition. Birmingham, England, Packt.

Pickup, M. (2015) Introduction to Time Series Analysis. Los Angeles, SAGE.

Statistics Finland. (2022a) Consumer Price Index. [online]. [Accessed 8 October, 2022]. Available https://www.stat.fi/julkaisut

Statistics Finland. (2022b) Motor vehicle stock. [online]. [Accessed 29 November, 2022]. Available https://www.stat.fi/til/mkan/2021/mkan_2021_2022-03-01_tie_001_en.html

Traficom. (2022a) Liikennesuoritteen kehitys. [online]. [Accessed 30 November, 2022]. Available https://liikennefakta.fi/fi/turvallisuus/tieliikenne/liikennesuoritteen-kehitys

Traficom. (2022b) Maanteiden liikennesuorite valtakunnallisesti ja suurimmilla kaupunkiseuduilla. [online]. [Accessed 30 November, 2022]. Available https://tieto.traficom.fi/fi/tilastot/maanteiden-liikennesuorite-valtakunnallisesti-ja-suurimmilla-kaupunkiseuduilla

Volkswagen. (2022) Huolto ja palvelut, moottoriöljy. [online]. [Accessed 9 December, 2022]. Available https://www.volkswagen.fi/fi/huolto-ja-palvelut/moottorioljy-ja-nesteet/moottorioljy.html