Norwegian School of Economics
Bergen, Spring 2020

Analysing the demand for car-sharing in Bergen

An empirical approach using car-specific and demographic data

Jinhua Fang & Eivind Opedal

Supervisor: Julio Cesar Góez & Mario Guajardo

Master thesis, MSc in Economics and Business Administration,

Business Analytics/Economic Analysis

NORWEGIAN SCHOOL OF ECONOMICS

Contents

Abstract

Background: Car-sharing is gaining popularity throughout the world. Currently, there is limited research on car-sharing in Norway. This paper aims to identify empirically, based on past data, what demographic and car-specific variables determine demand for a car-sharing service in Bergen. Additionally, it aims to predict future car performance for Bildeleringen's cars.

Research question: What are the drivers of demand for the utilization of a car-sharing service?

Method: Using data from Bildeleringen and Statistics Norway, we use multiple linear regression to determine drivers of demand. Furthermore, linear regression is also used for predicting future car performance of Bildeleringen's cars. Linear regression is chosen out of five possible models based on cross-validation error. **Results:** Several variables significantly increase or decrease car performance of Bildeleringen's cars. The variables "car type", "higher average amount of cars in the parking spot", "electric cars", "cars with automatic gear shift", "higher median income", "spring", and "households without car" increase performance, while "higher population density", "age 20-24 years old", "retired", "male", and "child cushion" decrease performance. **Analysis:** Out of the significant demographic variables, only "households without car", "age 20-24 years old", "retired" and "male" affect car performance to a noteworthy degree. Out of the car-specific variables "car type", "electric cars", "cars with automatic gear shift", "spring", and "child cushion" affect car performance to a noteworthy degree. **Conclusion:** The relationship between car performance and demographic variables is not strong. Car-specific variables seem to show a higher degree of correspondence with car performance.

Key words: Car-sharing, car-sharing locations, car performance, linear regression, demographic variables, car-specific variables, Norway.

1. Introduction

Considering the growth of transport demand due to increasing populations and the pressure on time efficiency in the modern world, transport networks have been expanded all over the world. The increasing use of private transport in industrialized countries provides greater accessibility. However, in the long-term increased transport has had many negative consequences, such as traffic congestion, lack of parking spaces, increased noise levels, emission of environmental pollutants, as well as consumption of energy. This has occurred mainly in urban areas where demand is concentrated in peak hours. Moreover, the costs of owning a car are increasing. These costs include fuel prices, parking, the cost of purchasing, and insurance fees. Additionally, some of these costs are sunk costs even before a mile is driven, which means that they are unrecoverable even if the vehicle is not being used. In addition, use of each private car is very low. In America, for example, vehicles spend around 90% of their time parked (Jorge & Correia, 2013). Public transport could be a good alternative, but it has several disadvantages. For instance, public transportation does not provide door-to-door service even in cities with sophisticated public transport systems. More importantly, schedules are not flexible, and services lack personalization. Utilizing public transport during the peak hour demand also means that vehicles are idle for the rest of the day, decreasing the vehicles efficiency (Jorge & Correia, 2013). Efforts have been made in the last few decades to provide new urban transport alternatives. One of these is car-sharing, which involves a fleet of vehicles scattered around a city for use by a group of members. It is a system that is somewhere between private and public transport.

One of the most important problems for a car-sharing company is to find the best locations to place their cars. The best locations should be chosen based on the demographic features that can influence future demand. In this paper, data from all of Bildeleringen's cars and demographic data from Statistics Norway (SSB) are analyzed to determine what factors are critical to car performance. This paper identifies drivers of demand using a multiple linear regression model, and then chooses the best predictive model to predict car performance.

The method used in this paper is a two-step approach. The first step is to fit a multiple linear regression model to data of all of Bildeleringen's cars and uncover the drivers of demand.

The dependent variable is the performance of the car-sharing service, and the measure of performance is the average amount of minutes each vehicle is being used per day. The independent variables are car-specific features, for example car type, fuel type, child seats, and demographic characteristics in the region around the parking spot, for example population density, share of households without car, age distribution and income level. In the second step, five predictive models with different statistical methods are created and the best model is chosen based on model performance.

1.1 Goals

The goals of this paper are to identify the demand drivers of car performance, create a model to predict future demand for Bildeleringen, test this model with new data from Bildeleringen and use this model to make practical recommendations for Bildeleringen on where they should place new possible parking locations to have high utilization of their cars.

1.2 Research question

What are the drivers of demand for the utilization of a car-sharing service?

1.3 Theory

1.3.1 Car-sharing

The origins of car-sharing services date back to 1948, when a housing cooperative known as Sefage provided its service to their clients in Zürich, Switzerland (Jorge & Correia, 2013). Later, in the 1970s, further experiments with car-sharing services were implemented in other European countries. However, they were short lived due to the small number of cars available. At the end of the 1980s, the number of car-sharing projects increased and some of them were a success. Car-sharing services continued to see a rapid increase during the 1990s, and major development in car-sharing started from 2000 on onward. It became increasingly popular in Europe, Asia and North America. In 2014, Europe accounted for 46% of the

global car-sharing business, and North America came second with 34% (Mindur, Sierpiński, & Turoń, 2018). It is worth noting that during the period from 2006 to 2014, Asia recorded the most rapid growth in the number of users registered with car-sharing systems. In 2017 in Shanghai alone the number of users was 1.2 million, while it was around 2 million in Germany, which is the leader in Europe (Mindur et al., 2018).

Worldwide, the number of car-sharing users is forecasted to grow from 2.3 million in 2013 to 12 million in 2020. The largest car-sharing firms, Car2go and ZipCar, initiated by large car producers BMW and Daimler, operate in multiple cities with total fleet sizes of over 10,000 cars each (Ferrero, Perboli, Rosano, & Vesco, 2018). Car-sharing services work as a new and more sustainable way of transportation, which is shifting the private mobility from ownership to service use. The economic benefits for the users are clear, which means that it can increase the low utilization rate of private cars, decrease the high fixed cost to own such as maintenance, parking and insurance fees (Jorge & Correia, 2013). Shared cars have much higher utilization rates than private vehicles because each vehicle spends more time on the road and less time parked, thereby reducing the sunk costs. When cars are being used and not occupying parking places, higher utilization rates mean that less land is needed for parking. Martin, Shaheen and Lidicker (2010) conducted a stated-preference survey in North America and concluded that car sharing members reduced their vehicle holdings significantly, from an average of 0.47 vehicles per household to 0.24 vehicles per household (Martin et al., 2010). From the point of view of building a sustainable city, the vehicles used in car-sharing are typically fuel efficient and lead to positive effects in reduction of urban emissions and city congestion.

There are three main types of car-sharing services regarding the way vehicles are hired and returned. According to Ferrero et al. (2018), who analyzed papers on car-sharing services, almost 47% of the papers they analyzed studied one-way mode, 19% studied two-way mode, 19% studied free floating mode, and 15% of the papers studied other modes (Ferrero et al., 2018).

One-way (station based): The car is taken from one station and returned to another station. This business model needs to consider the vehicle reallocation problem and the imbalance issue in the different parking stations.

Two-way (station based): The car is taken from and returned to the same station.

Free-floating: The car is taken from and returned to any accessible location in the city. It is the last mode to arrive in the market (Ferrero et al., 2018).

Car-sharing services also classify their services using engine type. There has been a growing interest in electric car-sharing service in recent years, especially an increasing investment in electric vehicles by car-sharing operators in China (Mindur et al., 2018).

Fully thermic: These fleets are composed of vehicles powered by traditional fuels such as gasoline or diesel.

Green: Green cars are adopted by car-sharing companies who are environmentally aware. Specifically, the vehicles have less-polluting engines, such as electrical, hybrid, plug-in, natural gas and liquefied petroleum gas (Ferrero et al., 2018).

1.3.2 Bildeleringen

Bildeleringen is a car-sharing company located in Bergen, Norway. The company was established in 1996 and now has more than 2100 private and corporate memberships. There are more than 200 cars placed in more than 80 parking spots in Bergen. Bildeleringen is organized as a cooperative and is owned by all the members. The company is operated on a non-commercial basis and the eventual profit goes back to operation (Bildeleringen, 2020). Bildeleringen operates in a classical two-way mode, which means that customers hire and return the vehicles in the same place. There are 7 types of vehicles: minicar, small car, wagon, large wagon, van, 9 seats and SUV. They have 3 types of fuel engines: gasoline, diesel and electric. Some cars have automatic gear shift, while others are manual cars.

To reserve a car in Bildeleringen, you need to become a member on their website. You need to pay a deposit, per kilometer and per hour for each trip. All other expenses are included in the price. You can then login with your membership, view all available cars sorted by geographical proximity, and choose when you wish to reserve a car. You can also choose car type and extra items. It is not necessary to reserve a car in advance, which means that you

can make a reservation spontaneously and drive a car immediately if it is available. When your trip is finished, you return the car to the place where you got it (Bildeleringen, 2020).

1.4 Literature review

With the increase in car-sharing services around the world the field has also gained attention from academia. Many papers have written about this field and in this section some of the papers are presented to try to pinpoint areas of interest that have yet to receive attention. The papers have been divided into three groups to logically correspond with the structure of this paper. The papers were chosen based on certain informal inclusion and exclusion criteria. They had to be in English, had to be not too old, and also needed to be highly relevant to this study's topic. The following does not claim to be an exhaustive list of papers.

1.4.1 Drivers of demand

The following section includes papers that try to determine which factors are important drivers of demand. These studies have been conducted with data from the US (Khan & Machemehl, 2017; Millard-Ball, 2005; Stillwater, Mokhtarian, & Shaheen, 2009) and Canada (De Lorimier & El-Geneidy, 2013), and are concerned with slightly different goals.

Goals

One paper looked at what types of markets car-sharing appeals to, and in what types of neighborhoods it succeeds in (Millard-Ball, 2005). Another paper studied the use of car-sharing vehicles of a major car-sharing operator (Stillwater et al., 2009). Yet another paper studied the effect of several variables on the use of free-floating car-sharing vehicles (Khan & Machemehl, 2017). Lastly, one paper investigated the relationship between the use and the availability of car-sharing vehicles at a car-sharing company (De Lorimier & El-Geneidy, 2013).

Findings

One paper found that car-sharing appeals to households with higher education, but not necessarily with higher income households. They also found that the neighborhoods that have car-sharing are characterized by other characteristics than demographic ones. They found that establishing car-sharing locations in neighborhoods with low car ownership was the key to success (Millard-Ball, 2005). Another paper found positive relationships with the occurrence of light rail stations and with households with less cars. They found no relationships with population density or other demographic variables (Stillwater et al., 2009). Yet another paper also found positive relationships with households with less cars, the number of transit stops nearby, the number of adults over 18 years of age nearby and the income of the parking spot neighborhood. They also noted that income may serve as a proxy variable for education level (Khan & Machemehl, 2017). Lastly, one paper found a positive relationship on availability on the number of cars at the parking location, the occurrence of a child seat in the car and the vehicle's age (De Lorimier & El-Geneidy, 2013).

Method

One paper did an analysis of neighborhood characteristics around car-sharing locations (Millard-Ball, 2005). Another paper used a logistic regression model (Khan & Machemehl, 2017). Yet another paper used a multilevel regression model (De Lorimier & El-Geneidy, 2013). Lastly, one paper investigated the relationship between the performance of the car-sharing vehicles and several variables including transportation, demographic and location specific variables. For this they used multivariate regression (Stillwater et al., 2009).

1.4.2 Demand estimation and prediction

This section includes papers on predictive models that estimate demand for the future. These studies have been conducted with data from Palermo, Italy (Catalano, Lo Casto, & Migliore, 2008), Zürich, Switzerland (Ciari, Schüssler, & Axhausen, 2010), the University of Wisconsin-Madison (Zheng et al., 2009), and 13 US regions which had car-sharing (Celsor & Millard-Ball, 2007).

Goals

One paper created a travel demand model (Catalano et al., 2008). Another paper modeled car-sharing and estimated travel demand (Ciari et al., 2010). A third paper studied the potential car-sharing demand at a university (Zheng et al., 2009). One paper investigated site selection based on user preferences. They suggested a method to help decision makers plan for new car-sharing sites (Ion, Cucu, Boussier, Teng, & Breuil, 2009). Another paper created a tool to assess which neighborhoods are good for car-sharing (Celsor & Millard-Ball, 2007). Lastly, one paper predicted future bookings for a free-floating car-sharing system (Seign, Schüßler, & Bogenberger, 2015).

Findings

One paper found the attributes that proved to be the most important were travel time and cost, specific attributes for the car and the number of cars a household had (Catalano et al., 2008). Another paper found that the variables that could best explain the variation in bookings were population density, closeness to the city center, house rent prices in the area and hotel and restaurant density. Their result supports the assumption that urban locations result in greater amounts of bookings (Seign et al., 2015). Yet another paper found that transportation characteristics have a stronger relationship to car-sharing performance than demographic characteristics. They found that low vehicle ownership, especially, had a strong relationship with car-sharing performance (Celsor & Millard-Ball, 2007). Lastly, one paper found that the respondent's status at the university, more so than their socio-economic status, affected their willingness to join car-sharing (Zheng et al., 2009).

Method

One paper carried out a survey asking about the preference between transport alternatives when car-sharing was one of the alternatives. They used a multinomial logit model (Catalano et al., 2008). Another paper used an open source software, called MATSim (Ciari et al., 2010). Yet another paper carried out a preference survey and developed probabilistic models based on this. They then were able to predict car-sharing market shares under different scenarios (Zheng et al., 2009). Lastly, one paper created a regression model to predict future demand, with several independent variables. (Seign et al., 2015).

1.4.3 Optimizing locations

This section includes papers that create optimization models where parking locations are one of the variables. These studies have been conducted with data from Nice, France (Kumar & Bierlaire, 2012), Hanover, Germany (Rickenberg, Gebhardt, & Breitner, 2013), Chengdu, China (Cheng, Chen, Ding, & Zeng, 2019) and San Diego, USA (He, Mak, Rong, & Shen, 2017).

Goals

One paper optimized new parking locations for a car-sharing operator (Kumar & Bierlaire, 2012). Another paper created an optimization model to help decision makers decide the location and size of the parking locations for a car-sharing operator. The goal was to maximize profit (Rickenberg et al., 2013). Another paper used machine learning to help car-sharing operators choose locations for their parking locations (Cheng et al., 2019). Lastly, one paper aimed to help a car-sharing operator choose locations in which to operate (He et al., 2017). All in all, these goals are very similar.

Findings

One paper found population density, higher income and higher education to be important factors explaining the success or failure of parking locations (Kumar & Bierlaire, 2012). Another paper found that high population density had a positive effect on utilization of the cars. They also found that higher population density leads to shorter average distances driven (Rickenberg et al., 2013). Another paper did not find population characteristics to be an important factor (Cheng et al., 2019). Lastly, one paper found that to ensure a high probability that a car will be available for the customer, the area of operations needs to be reduced (He et al., 2017)

Method

One paper used a two-step model. In the first step they created a regression model to determine which factors were important for a successful car-sharing location. They used the average amount of bookings at the location as the dependent variable and used several different independent variables (Kumar & Bierlaire, 2012). Another paper divided the city

into many small districts and assigned a binary value to each, signifying whether or not there was a demand for car-sharing in the district. They then ran several regression models to determine important factors for choosing parking locations (Cheng et al., 2019). Lastly, one paper used a mixed integer program (He et al., 2017)

To summarize, several studies which concern car-sharing have been conducted. None of the papers have utilized data from Norway. A study utilizing Norwegian data would enrich the current state of knowledge. It is relevant to see if the same patterns observed in the studies mentioned above would emerge with data from Bildeleringen in Norway, or if the results tell a different story.

These studies share many of the same goals, which in short is to answer the question: “What makes a successful location for car-sharing?”. This is also the goal of this study.

The findings in the literature are inconsistent. Some have found a relationship between demographic characteristics and car performance, while others have not. This will be discussed further in the methods section of this paper.

These studies utilize several different methods, all specifically adapted to their context, while in most cases also utilizing a regression model. This study has a similar approach.

2. Method

2.1 The independent variables

Several of the independent variables have been studied before. Here is an overview of some of the studies that have used identical or similar variables as the ones used in this study.

Car-specific features: Several car-specific features can affect the rate of use of different cars. An important variable is the age of the car (De Lorimier & El-Geneidy, 2013). Since all of Bildeleringen's cars are relatively new, this has not been chosen as a relevant variable for the study. Another important variable are special attributes of a car (Catalano et al., 2008). One study shows that the occurrence of a child seat in a car increases the car's availability (De Lorimier & El-Geneidy, 2013), which means it is less used. Also, the car type and fuel used are important factors. Finally, many car-sharing users find environmental regards important, and electrical cars are therefore more popular (Firnkorn & Müller, 2015).

Number of cars at the parking location: This factor is undoubtedly relevant. The direction of the relationship is not clear. More cars at a parking location will cannibalize upon each other's demand, which will drive down the average performance (Kumar & Bierlaire, 2012).

More cars at a parking location might drive in new customers, though, and create a hot-spot for car-sharing (Khan & Machemehl, 2017), which would increase the average performance.

Population density: It is intuitive that population density increases demand for car-sharing. More people equal more customers, which leads to more use of the cars. Several studies have shown this positive relationship (Cohen, Shaheen, & McKenzie, 2008; De Lorimier & El-Geneidy, 2013; Dias et al., 2017; Kumar & Bierlaire, 2012; Seign et al., 2015). On the other hand, others have not found any relationship (Stillwater et al., 2009).

Share of households without a car: Less cars in the household leads to an increase in car-sharing use. Cars that are parked near areas with a large number of households with few or no cars would therefore perform better. Several studies have shown this relationship

(Catalano et al., 2008; Celsor & Millard-Ball, 2007; Clewlow, 2016; Khan & Machemehl, 2017; Millard-Ball, 2005; Stillwater et al., 2009).

Seasonal variability: One paper shows that car performance varies a lot between seasons (De Lorimier & El-Geneidy, 2013). For instance, the summer months have an impact on the type of demand for the cars. In summer there would be less demand for using the cars for commuting and more demand for longer recreational trips.

Age and gender: Studies have shown that users of car-sharing services tend to be young (Dias et al., 2017). As previously mentioned, Khan and Machemehl (2017) show that there is a positive relationship between car performance and the number of adults over 18 years old in the neighborhood. Some studies have shown that there are more male car-sharing users than female (Kumar & Bierlaire, 2012).

Share of higher income and higher education: Many studies have shown that car-sharing users tend to have higher education (Coll, Vandersmissen, & Thériault, 2014; Dias et al., 2017; Kumar & Bierlaire, 2012; Millard-Ball, 2005). Some studies have shown that they also tend to be in the higher income group (Dias et al., 2017; Khan & Machemehl, 2017; Kumar & Bierlaire, 2012) while others have shown that they tend to be in the lower to medium income group (Efthymiou & Antoniou, 2016). These variables might be too closely linked though, as pointed out by Khan and Machemehl (2017).

Share of different working status: Logic dictates that working status should affect one's travel habits. For instance, a retired person and an employed person have different travel habits. One paper shows that students and employed university workers have different travel habits (Zheng et al., 2009). It might be natural to think that the share of employed people and car performance are positively correlated, since these people need to get to work, but Khan and Machemehl (2017) argue that car-sharing is not well suited for commuting. They found a negative relationship between the share of employed and car performance (Khan & Machemehl, 2017).

Walking distance: Walking distance is not a variable in the model, but it is used to calculate the demographic variables. It is intuitive that the walking distance between a household and the parking location influences the probability that the household would be a car-sharing

user. Zoepf and Keith (2016) show that an increase in walking distance of one mile is equivalent to an increase in price of US2$ per hour in vehicle renting cost for the user. Another paper found convenience, including walking distance, to be an important driver of demand (Lindloff, Pieper, Bandelow, & Woisetschläger, 2014).

2.2 Data collection

To examine drivers of demand for Bildeleringen, we need dependent and independent variables, and we need to decide how this should be measured. Bildeleringen has more than 80 parking spots in Bergen, and more than 200 cars spread out over these parking spots. The cars differ in type and other variables that affect its demand. These car-specific variables are important factors for the study and need to be included. To do that, the cars need to be the sample unit of the study. The data that is used for the amount and length of the trips of the cars are from 1st of January 2019 to 12th of January 2020. Some of the cars were moved in this time period from one parking spot to another. This is a challenge that needs to be addressed, since the demographic data around the parking spots are highly important in this study. Though this might seem like a challenge, it rather becomes a strength. Bildeleringen has access to where each car had been parked at any given time, as well as the time period it had been parked there. Therefore, the sample unit of the study is the cars that Bildeleringen has, but with one observation for each parking spot the cars had been placed at. This increases the sample size in the study, which then increases the power in the regression.

The dependent variable is the performance of these cars, while demographic data is used for many of the independent variables, as well as for control variables for each car. As it is not obvious how the demographic data should be presented, neither is the meaning of demographic data “near” a parking spot, a lot of data processing has to be done for the data to be usable for the analysis. This section covers some of the critical decisions that have to be made, as well as some of the assumptions that are made. First, the data sources and data content that are used in the analysis are covered, as well as an explanation for some key terms needed to understand the data. Second, the data processing, from individual demographic data and trip data from Bildeleringen to demographic data and performance

data for each car are covered. Finally, some of the weaknesses in the data, as well as the assumptions that could potentially be a weakness for the analysis are covered.

2.2.1 The demographic data

All of the demographic data is from Bergen municipality. Some terms need to be explained to clarify where the data comes from. The data is collected at three different geographical levels. These are "Bydel", "Grunnkrets" and individual addresses. "Bydel" is the biggest geographic area and could be translated to a city district. Throughout the paper the English translation City District is used. There are eight such City Districts in Bergen municipality. These are: Arna, Bergenhus, Fana, Fyllingsdalen, Laksevåg, Ytrebygda, Årstad and Åsane. As Bergen municipality is larger than 400 km^2, the City Districts are still quite big. "Grunnkrets" is a geographic area used by SSB to display statistics for analysis on a regional or municipality level. "Grunnkrets" is much smaller than City Districts and there are 363 "Grunnkrets" in Bergen municipality. Throughout the paper "Grunnkrets" is abbreviated to "GK". There are about 150 000 addresses in Bergen. An address consists of a street name, a number, and sometimes a letter and an apartment number if there are several households connected to one address. It is important to be aware of these terms because different data is given at different levels.

Data given on the City District level

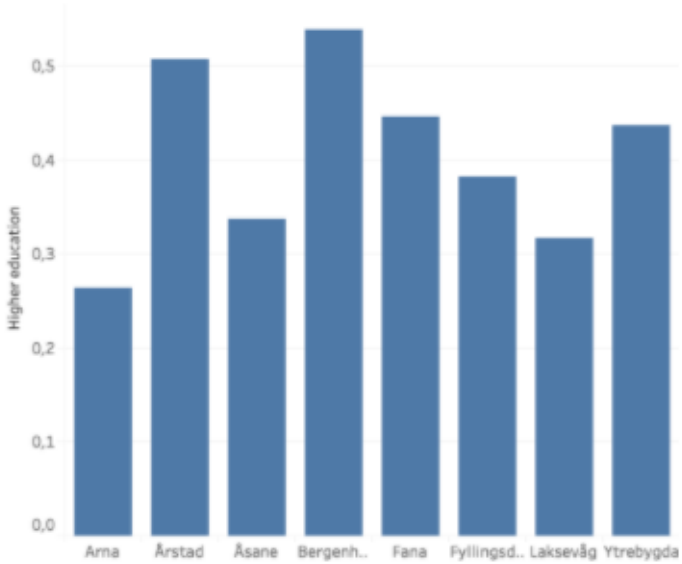
Bydel	
Arna	540 000
Årstad	418 000
Åsane	570 000
Bergenhus	403 000
Fana	642 000
Fyllingsdalen	563 000
Laksevåg	501 000
Ytrebygda	699 000

Median income: As previously mentioned, there are eight City Districts in Bergen, as shown in Table 1. The median income varies between the different City Districts. The two areas with the highest median income are Ytrebygda and Fana, both located south of the city center. The two areas with the lowest median income are Bergenhus and Årstad. Bergenhus consists of the city center, while Årstad borders Bergenhus to the south.

Table 1: Median income in City Districts in Bergen, Norway (SSB, 2020e).

Level of Education: There are five levels of education: “Grunnskole”, “Videregående skole”, “Universitets- og høgskoleutdanning, 1-4 år”, “Universitets- og høgskoleutdanning, over 4 år” and “Uoppgitt eller ingen fullført utdanning”. Each level denotes the number of people with this level of education as their highest achieved education. For instance, a person currently attending “Videregående skole”, or high school, has “Grunnskole”, primary and middle school as their highest achieved level of education. Bar chart 1 shows level of higher education in percentage of the population.

Only the two highest levels of education are considered to be higher education, which are “Universitets- og høgskoleutdanning, 1-4 år” (University and college degree, 1-4 years), “Universitets- og høgskoleutdanning, over 4 år” (University and college degree, more than 4 years). Therefore, a theory is that only these two would have an effect on car use. The two levels are combined in Bar chart 1 to show distribution of higher education, but are kept as separate variables in the study. The chart therefore shows the percentage of the population that have achieved higher education. The highest percentage is found in “Bergenhus”, the city center, while the lowest is found in “Arna”.



Bar chart 1: Higher education by city district in Bergen, Norway (SSB, 2020b).

Data given on the Grunnkrets level

As previously mentioned, the GK is a geographic area used by SSB to display statistics. SSB tries to make the regions as homogenous as possible in respect to building structure and economic base (SSB, 2018). Some of the GK were deleted because of missing data. This should not affect the study though, because these GK are the mountainous regions around Bergen or GK with a very small population.

Population density: SSB has data on the total population of each GK. These data are later used to calculate the population density around the parking spots (SSB, 2020a).

Age and gender distribution: Age distribution, taken from SSB (SSB, 2020a) is split into 11 categories. The numbers are given as total numbers, which have been recalculated into the percentage of the total population of that GK. The categories are: “0-5”, “6-15” , “16-19”, “20-24”, “25-29”, “30-49”, “50-59”, “60-66”, “67-69”, “70-79” and “80-”. Data on gender distribution is also given in total number and is recalculated into percentages (SSB, 2020a).

Working status: This data consists of five categories: Employed, unemployed, retired, under education and other. Persons 15 years old or younger are not included in the data. The percentages are therefore calculated using the total number of persons over 16 years old from the age distribution. The percentage of employed in a GK ranges from 30% at Hatleberg student village, to 82% at Haukeland, near the city hospital (SSB, 2020d).

Households without a car: This data consists of two categories and is given at the household level, not on the level of individual people. The number of cars a household has is not distinguished in the data source. Either a household has one or more cars, or it does not have any car. The percentage of households without a car ranges from 6% in rural areas, to 95% near urban student villages (SSB, 2020c).

Data at the address level

Data for all the addresses in Bergen municipality comes from Kartverket (2018). This data consists of 149 430 addresses in Bergen. Each address is given with the corresponding GK, as well as the postal code and its geographic coordinates. With the postal code it is possible

to know which City District it is in. Each address is therefore connected to all the previously mentioned above demographic data. All the demographic data are then stored with the individual addresses. If for instance a GK has 25% men, all the addresses in that GK also have 25% men. The population density is handled differently. The total population in a GK is divided between all the addresses in that GK. The address therefore has an approximation on the demographic data in the household, as well as an estimate on how many people live at this address.

2.2.2 Car-specific data

All the data relating to cars, trips and parking locations comes from several different tables in Bildeleringen's database. Importantly, each car has a car ID, and each location has a location ID. These are used to assign the correct data to the correct car or location. The study uses several different car-specific variables. An explanation on how these variables are formed are given here.

Car type: Each car is just one of seven different car types. Each of the car types are represented as a binary variable in the study. In Diagram 1, the proportions of Bildeleringen's car types are shown.

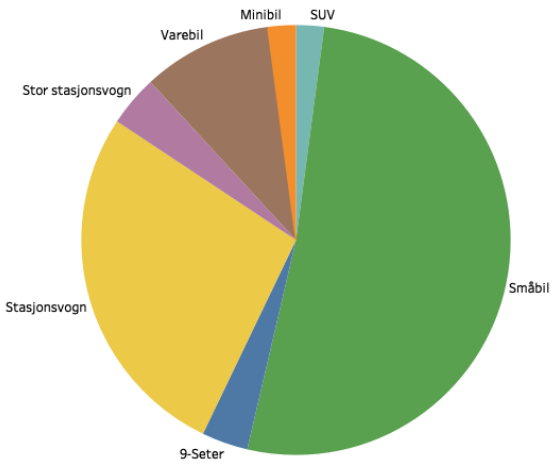
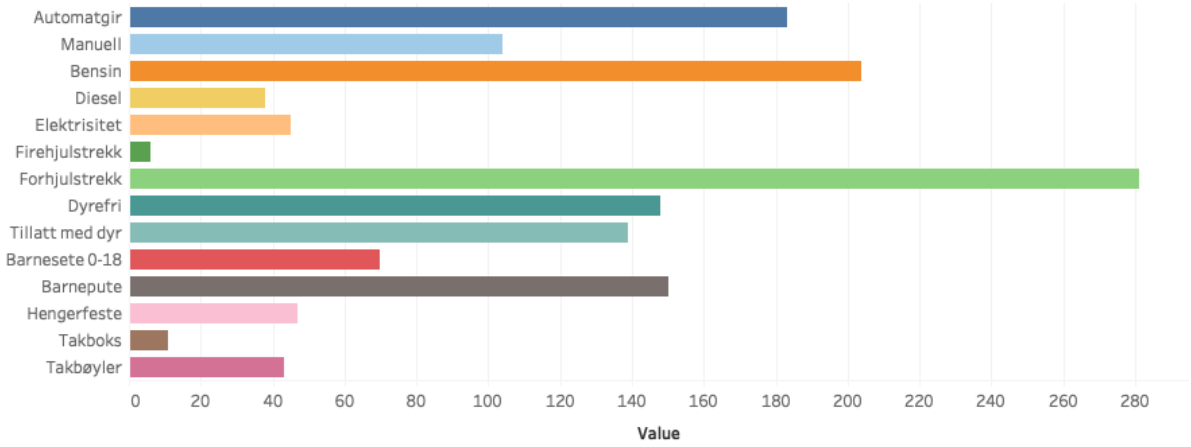


Diagram 1: Share of car type for Bildeleringen.

Car-specific features: Each car can have several different binary features, and an overview of how many cars have each feature is given in Bar chart 2. Five of the features can be simply understood as either a car has the features, or it does not. These five features are "barnesete 0-18", "barnepute", "hengerfeste", "takboks" and "takbøyler". Bar chart 2 shows that about 50% of the cars have "barnepute". Binary variables are created for these features, 1 represents the availability of the facilities, 0 means the opposite.

The remaining 9 features are not so simply understood. These features are grouped by category and distinguished by color in Bar chart 2 to denote that a car has to have one, and only one of these features. For instance, for the blue category, concerning type of gear shift, a car can have either "Automatgir" or "Manuell". For the orange/yellow category, concerning type of fuel, a car can have either "Diesel", "Bensin" or "Elektrisitet". For the green category, concerning type of wheel driving mode, a car can have either "Firehjulstrekk" or "Forhjulstrekk". For the turquoise category, concerning animals allowed or not, a car can either be "Dyrefri" or "Tillatt med dyr". In each category, the number of cars sum up to the total number of observations in this study.



Bar chart 2: Share of cars with car-specific features for Bildeleringen.

Seasonal variables: In Bildeleringen's database the start and end dates of operation for a car at a location is given. Since the time period of the data spans approximately one year (1st of January 2019 – 12th of January 2020) it is possible to create binary variables to control for the four seasons a car has been available for. The seasons are Winter (December-February),

Spring (March-May), Summer (June-August) and Fall (September-November). The binary variable of the different seasons equals 1 if the car was available for more than 50 % of the season.

Number of cars at the parking spot: As previously mentioned the start and end date of operation for a car at a location is given in Bildeleringen's database. This makes it possible to know how many cars have been at a parking spot for a certain period of time, which is calculated for each car at each location. An example of how this is calculated is given below:

1. The period of time is decided, for example 1st of May 2019 to 30th of May 2019. This is the time the car was in use at a parking spot before it was moved or put out of service. This is car A.

2. All the other cars that have been parked at this parking spot are taken into account. Two cars are given as an example here. Car number 1 has been parked at this parking spot from 1st of April 2019 to 1st of June 2019. The overlap period in this example is 30 days. Car number 2 has been parked at the parking spot from 10th of May 2019 to 25th of May 2019. The overlap period is 15 days. We sum up the two overlap periods and get 45 days.

3. The sum of the overlap periods is then divided by the length of the time period being investigated. In this example the time period is 30 days. We therefore have 45/30 = 1.5. This number represents the average amount of other cars that have been parked together with car A in the period of time the car has been available at a location. This number is used in the study to control for the number of cars at a parking spot.

The number calculated to control for the number of cars at a parking spot varies quite a bit for the different cars. Certain parking spots have many cars, while many parking spots only have one car. Therefore, the number of cars at a parking spot varies between 0 (no other cars at the parking spot during the time period) and 16.

2.2.3 Connecting the data

Up until this point, the car-specific variables are connected to the cars while the demographic variables are connected to addresses. Now then to how we connect these two types of data. The answer lies in the fact that there are coordinates for the addresses and the parking spots. It is therefore possible to connect the demographic data near a parking spot to the cars parked at that parking spot. The idea is illustrated in Figure 1 below:

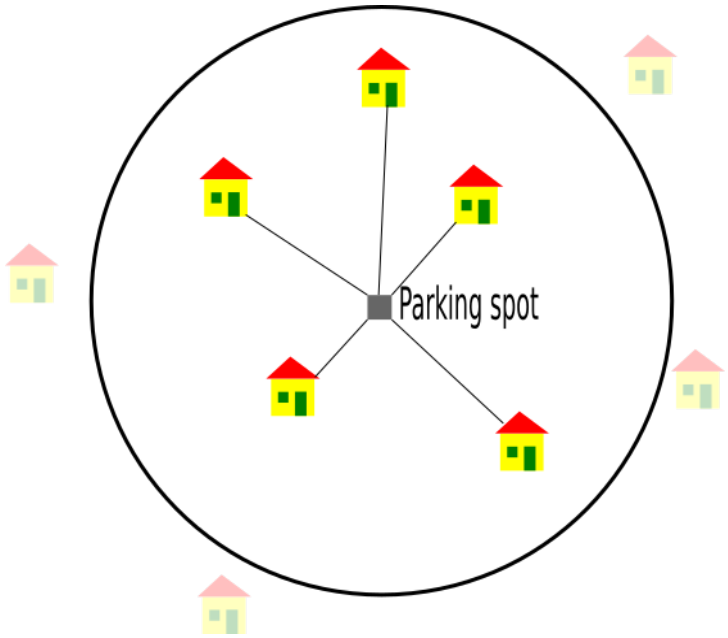


Figure 1: An idea on walking distance.

The idea is to find some limit to how far out from the parking spot the demographic data would be used. The limit is represented with the black line in the figure.

Calculating walking distance

It is assumed that the distance between the parking spots and households have no effect on demographic data. Meaning, on average, people are the same no matter how close or far away they live to a parking spot. The distance households have to a parking spot should affect their use of the cars at the parking spot. The assumption is that households closer to

the parking spots have a higher chance of using the cars. Therefore, there is a need to calculate a maximum walking distance for the households. This could be a fixed amount, like in Figure 1, that is to say that all households within walking distance to the parking spot are included in the demographic data. This seems unrealistic though, for several reasons. Households differ in how far they are willing to walk. The previous illustration is therefore not accurate enough to capture the demographic data around the parking spot. An alternative method is using a scaled-out approach. The idea here is that households are segmented based on the walking distance to the parking spot. Groups closer to the parking spot have a higher chance of using the cars, so they are given a higher weight when calculating the demographic data. A simple example is given using population density and gender:

In the example we have two groups of households. The first group are the households within 200 meters. The second group are the households between 200 and 500 meters. The percentage of households that find the parking spot to be within walking distance of 200 meters is 100 %, for the other group the percentage is 50 %. There are 20 men and 10 women within 200 meters, 40 men and 100 women between 200 meters and 500 meters. The data is summarized below in Table 2:

Distance	Percentage of households within walking distance	Men	Women
Within 200m	100%	20	10
200m-500m	50%	40	100

Table 2: Example of households within different walking distances.

The first group contributes 20 men and 10 women to the demographic pool of the parking spot. The second group contributes 20 men and 50 women to the pool, since only 50% of the population find the parking spot to be within walking distance. We end up with a pool of 40 men and 60 women. The demographic data for all the cars at this parking spot would therefore be as in Table 3:

Population in the vicinity of the parking spot	100 people
Men	40%
Women	60%

Table 3: Example of demographic data at a parking spot.

The walking distance used in this study could have been calculated if the addresses of the users were available. Unfortunately, this data is restricted. Kabra, Belavina and Girotra (2019) find in their study that 80% of bike-sharing users traveled less than 300 meters to their pick up location. This implies that the walking distance is relatively short for a majority of people. Using data from Kumar & Bierlaire (2012), the walking distance could be calculated, though this data is not from Bergen. Kumar & Bierlaire (2012) divide households into 9 different groups based on the distance from where the car is picked up. The five groups that are furthest away from the pick-up location contribute less than 20 % combined. To make the scripts more manageable to run, these groups are dropped. The remaining groups are named A, B, C and D. The paper by Kumar & Bierlaire (2012) contains data on the percentage of trips made from the different groups. This data, as well as other data needed to calculate maximum walking distance, is given in Table 4 below.

Group	Walking distance	Percentage of trips	Trips compared to A	Total area	Total area compared to A	Total trips in relation to total area, compared with A
A	Less than 100m	15%	1x	31415 m^2	1x	100%
B	100m-200m	14%	0.933x	94247 m^2	3x	31.1%
C	200m-500m	23%	1.533x	659734 m^2	21x	7.3%
D	500m-1000m	30%	2x	2356194 m^2	75x	2.67%

Table 4: Calculating maximum walking distance.

The column “Percentage of trips” is taken from Kumar & Bierlaire (2012) and represents the percentage of trips made within this group of households. Note that the percentages amount to 82%. This is because some of the groups are dropped, which amount to 18 % of the total trips. “Trips compared to A” is the number of trips made within this group compared to group A. So, for instance group D has twice as many trips as group A. “Total area” is the area these groups cover. For group A this is a circle with a radius of 100 meters. For the other groups the center of the circle has been left out to get the actual area the group covers. “Total area compared to A” is very similar to the previous column. It compares the total area with group A.

Three assumptions need to be made to calculate the last column:

1. Population is on average uniformly distributed around the pick-up spots. That is to say that on average there is no difference in population density closer or further away from the pick-up location, up to 1000 meters away.

2. On average, the only reason the households closer to the pick-up spots use the cars more than households further away is because of walking distance. There is no difference on average demographic data closer or further away from the pick-up spots, up to 1000 meters.

3. No households have a maximum walking distance less than 100 meters.

If these assumptions are true, then we can calculate the “Total trips in relation to total area, compared with A”. If households had no maximum walking distance, then “Trips compared to A” should be the same as “Total area compared to A”. However, they are not the same, which implies that maximum walking distance plays a role. For example, for group B it is only 31.1% (0.933/3) of what it should be. This means that 68.9% are excluded because the walking distance is too far. This indicates that 31.1% should be used as weight for group B when calculating population density and demographic data. The same calculation is done for the other groups. A final illustration is given below in Figure 2:

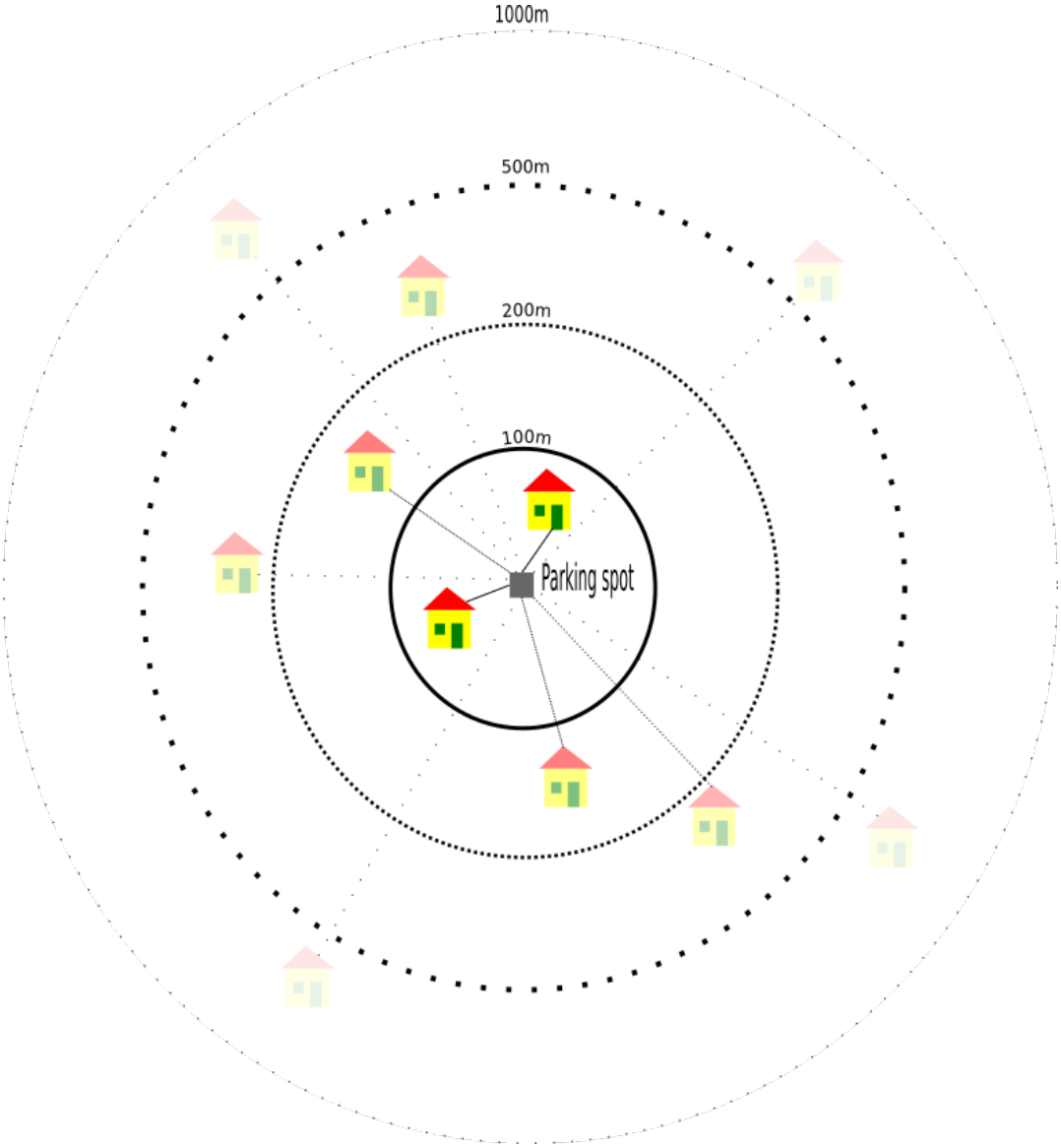


Figure 2: Illustration of scaled-out approach.

2.2.4 The dependent variable

The dependent variable is trying to capture the performance of a parking spot. This can be measured in many different ways. The first thing to consider is if it is the car performance or the parking spot performance that is being studied. As previously mentioned, there are

several car-specific features that need to be controlled for. Because of this, car performance is chosen as a sample unit for the study.

The measure of performance also needs to be considered. Three potential candidates are considered in our study: Number of trips, kilometers driven and minutes of car use. Looking at the data source, minutes of car use is the one that seems the most reliable of the three.

The last thing to consider is if the dependent variable should be the total minutes of car use or an average. Since the number of days the cars have been available at a parking spot varies quite a bit, it is decided that the average minutes of car use per day is the best candidate for the dependent variable in the study. This is calculated by adding together all the periods of usage time a car had, and then dividing it by the number of days the car had been available at that parking spot.

2.2.5 Weaknesses in the data

There are several potential weaknesses in the data. The weaknesses that have been observed in the data are listed here:

Missing or inaccurate data: This applies to several different areas. Some of the GK have missing demographic data. These GK should not affect the study though, since all of them have few or no population and they are located far away from any of the parking spots of Bildeleringen. All of the GK with missing data were deleted, as well as the addresses connected to this GK. Around 100 addresses were deleted because of this. Two of the cars were deleted because they were listed as test cars in the data source. Some of the trip data was deleted for the same reason.

Long trips: Some of the cars have been rented for longer periods of time. The effect this has on the dependent variable would not be accurately captured by the independent variables. The amount of these kinds of trips are relatively small compared to the total amounts of trips, and are therefore kept in the study.

City District data: Data from larger geographical areas lead to more inaccuracy in our analysis. This applies to the data on the CD-level. Data from larger geographical areas lead to less different data inputs, which means less variation. This means that many households over a large area would have the same demographic data, which is inaccurate. The necessity of these variables should be considered.

Freedom in parking: Many parking spots have some freedom in where you can park. This can be a street or a neighborhood. If the areas are large, this may distort the effect of the demographic variables. Samples have shown that these areas are not very big, but it is important to be aware of this.

Parameters are from Nice, France: To calculate the maximum walking distance we use data from Kumar & Bierlaire (2012). This data is from Nice, France. It is safe to assume that maximum walking distance will vary to some extent between cities and countries. So, the parameter "Percentage of households within maximum walking distance" may be inaccurate to some degree, depending on the differences between Bergen and Nice regarding city layout and topography.

Flexibility in pick-up location: One study found that several customers were likely to be flexible with the time or space of their pick-up location for relatively little compensation (Ströhle, Flath, & Gärttner, 2019). This implies that demographic data could be less important.

2.3 Preliminary analysis

The following section shows how analysis of the data and modeling is implemented, using R, version 3.6.1.

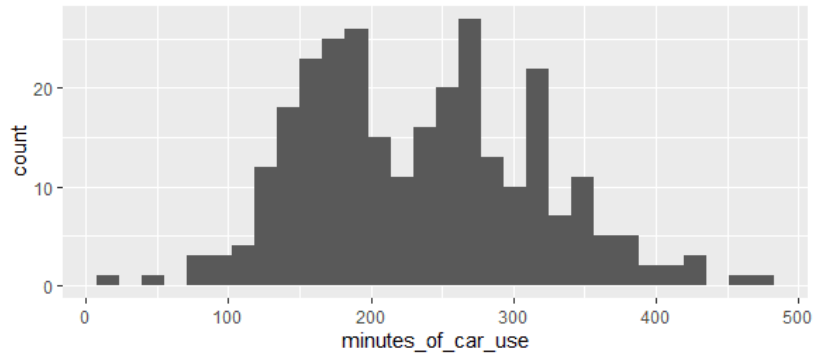
Before statistical models can be fitted to the data, some visualization of the current dataset has to be done. The dataset contains data from a total of 287 cars. After removing some irrelevant variables such as car ID, location ID, start and end time, there are one dependent variable and 41 independent variables. The visualization shows that some variables might be promising predictors of car performance. From Table 5, Histogram 1 and Density plot 1

which show the distribution of car performance, we see that car performance is approximately normally distributed, with minimum value at 22.91 minutes, mean value at 233.56 minutes and maximum value at 482.12 minutes. Most values fall into the range between 150 and 350 minutes. From Boxplot 1 of car performance, we see that there is one outlier, which is the maximum value of 482.12 minutes. The outlier from "small car performance" can be seen in Boxplot 2, which shows car performance for each car type. The outlier observation is then removed, so that it does not give any noise to the analysis. The distribution of car performance without outliers can be seen in Table 6, Boxplot 3 and Boxplot 4.

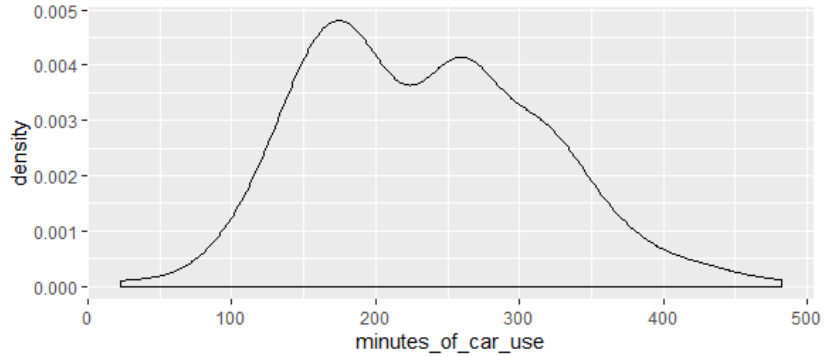
After removing the outlier, the dataset consists of 286 car observations. An overview of the first few lines of the dataset can be seen in Plot 1 in Appendix. There are 147 small car observations, 78 wagon observations and 28 van observations, in addition to a few other types of car observations, which is shown in Bar chart 3. Regarding average car performance per day, we see from Bar chart 4 that minicar, small car and van have the best performance, which means these car types are more popular. In terms of fuel type, 204 cars use gasoline, compared to only 38 cars using diesel and 44 cars using electricity, which can be seen in Bar chart 5. Average car performance is higher with electric cars, as seen in Bar chart 6. This shows the popularity of electric cars. We can also see that cars with automatic gear shift, whether the company has defined a car as animal free or not, wheel driving mode, i.e. if the car is 4WD, back-wheel driven or front-wheel driven, children's cushion, population density, share of households without cars, and average number of cars in the parking spot seems to have a promising positive relationship with car performance. The detailed average car performance for cars with different features can be seen from Plot 2 to Plot 7 in Appendix.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.91	170.77	235.09	233.56	288.44	482.12

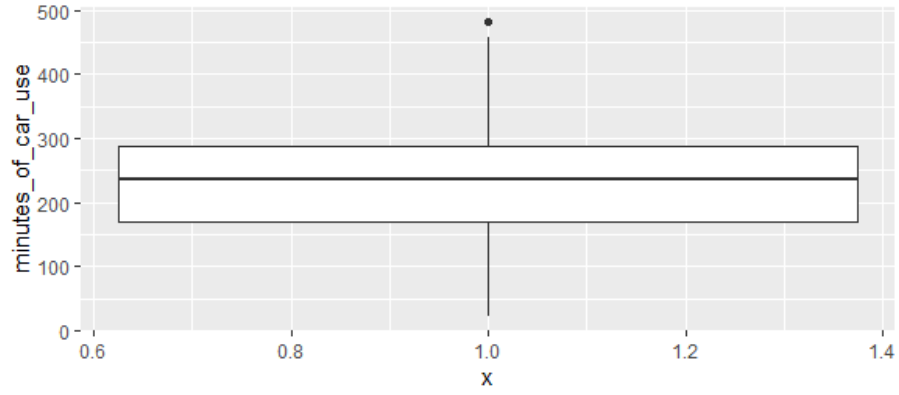
Table 5: Distribution of car performance.



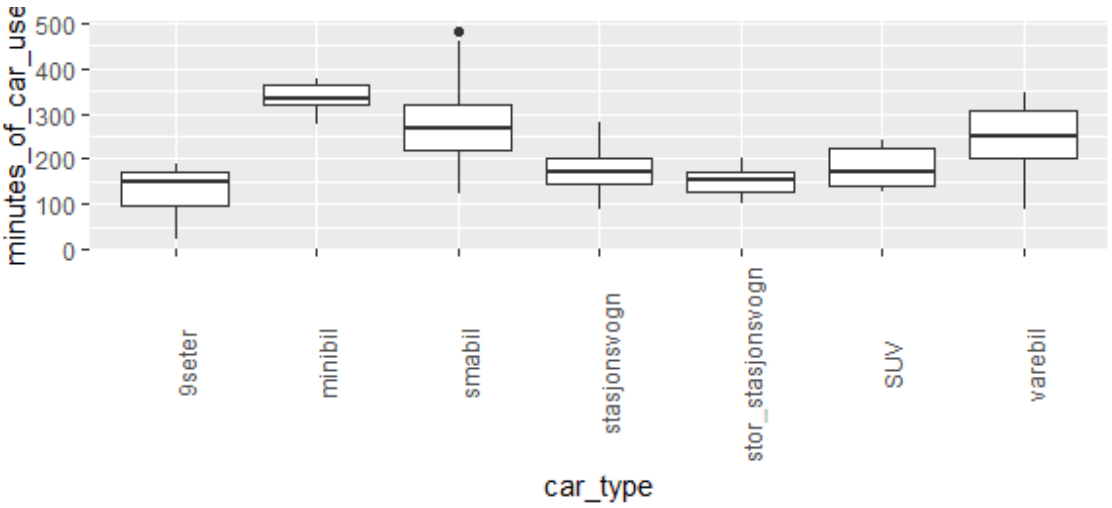
Histogram 1: Distribution of car performance.



Density plot 1: Distribution of car performance.



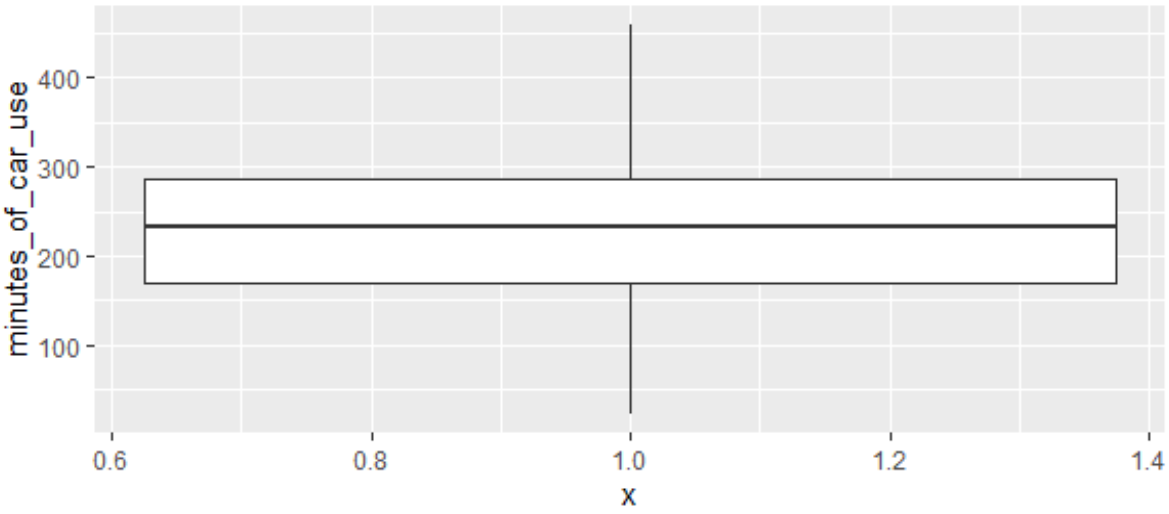
Boxplot 1: Distribution of car performance.



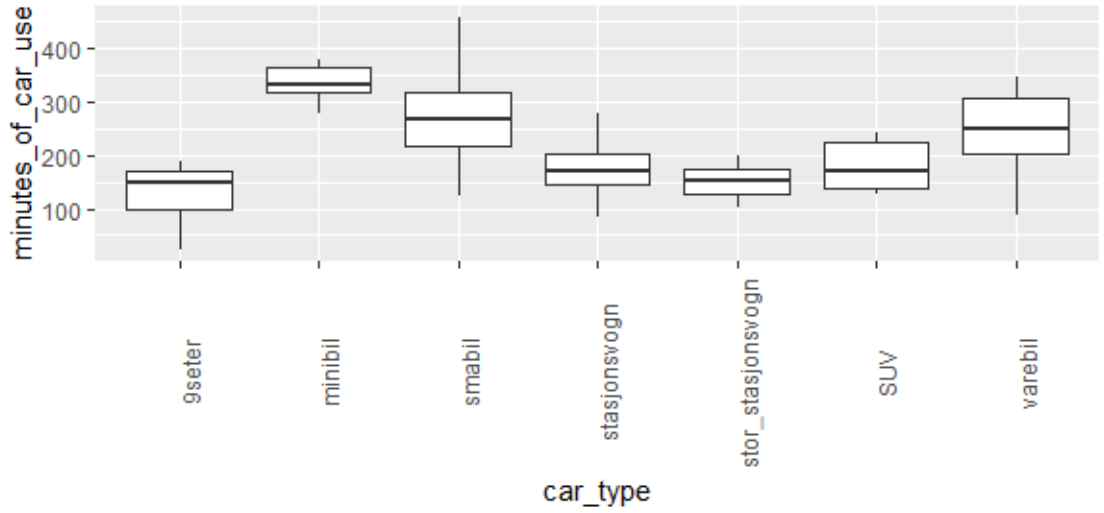
Boxplot 2: Distribution of car performance for each car type.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.91	170.49	232.65	232.69	286.28	459.04

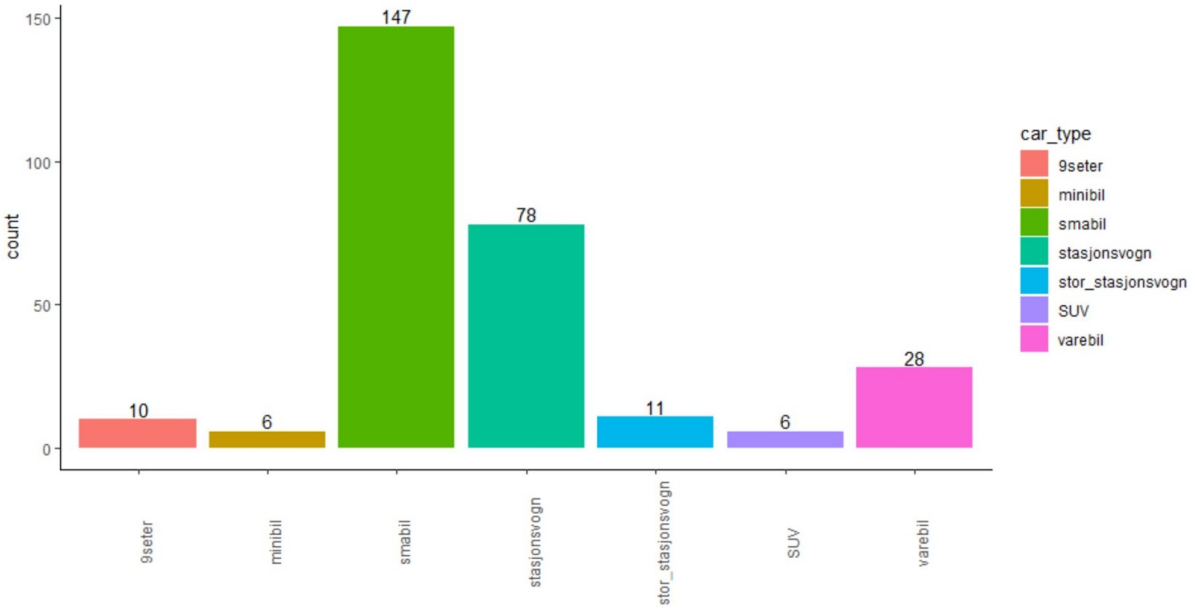
Table 6: Distribution of car performance without outliers.



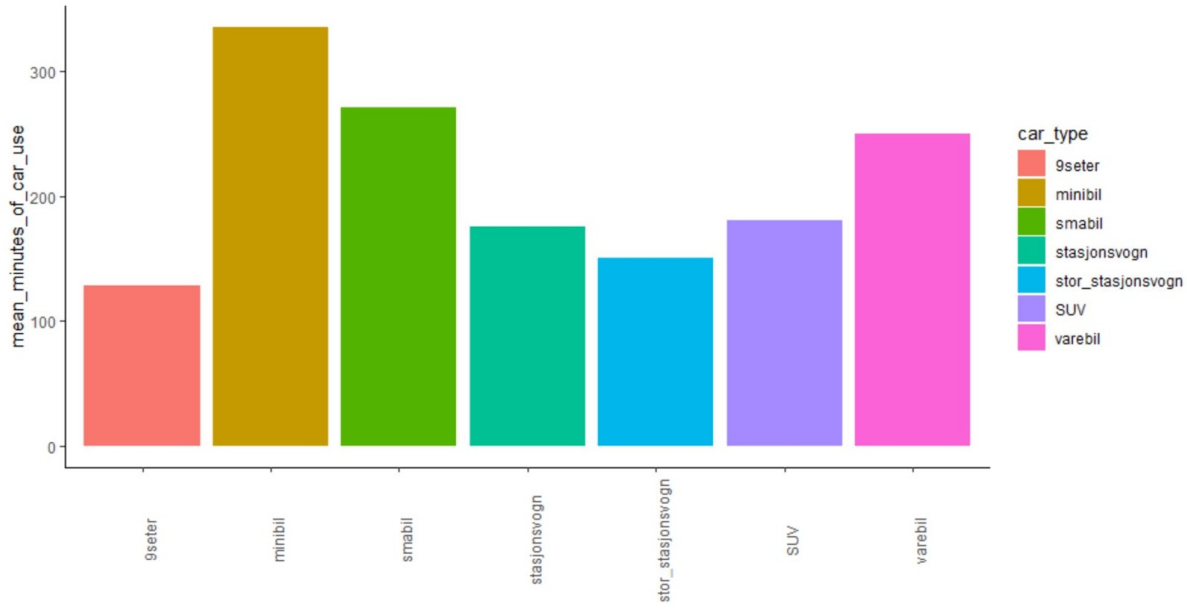
Boxplot 3: Distribution of car performance without outliers.



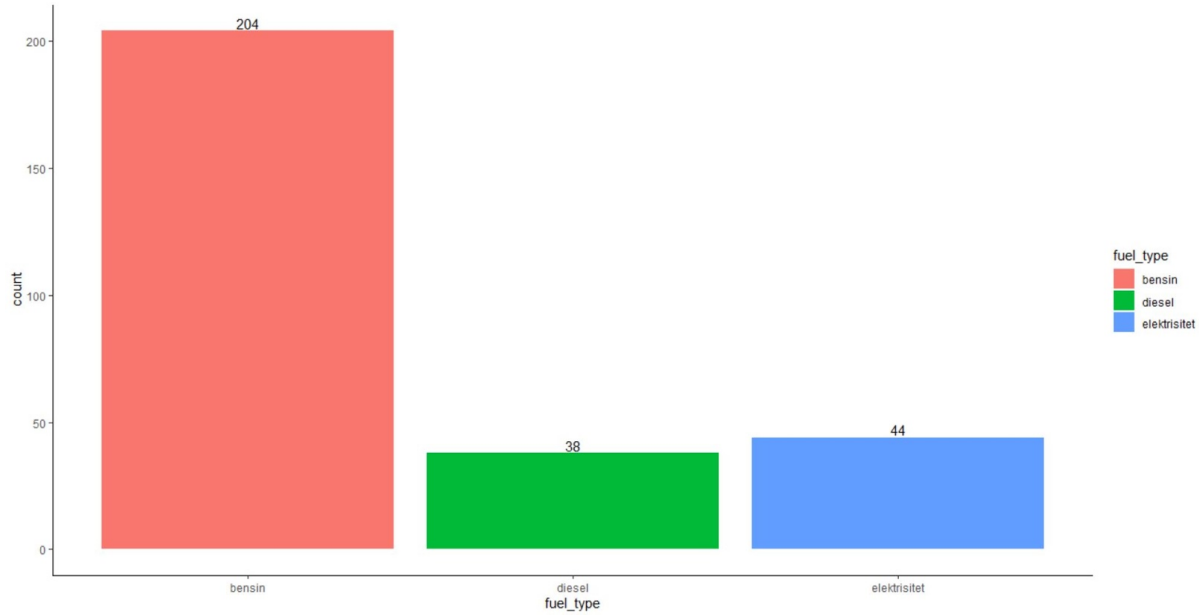
Boxplot 4: Distribution of car performance for each car type without outliers.



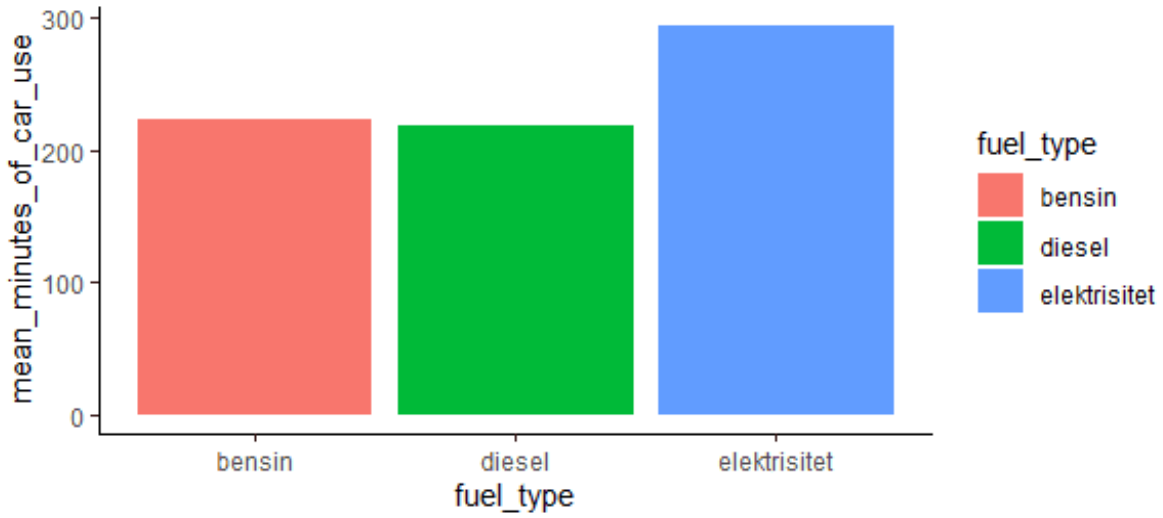
Bar chart 3: The number of cars for each car type without outliers.



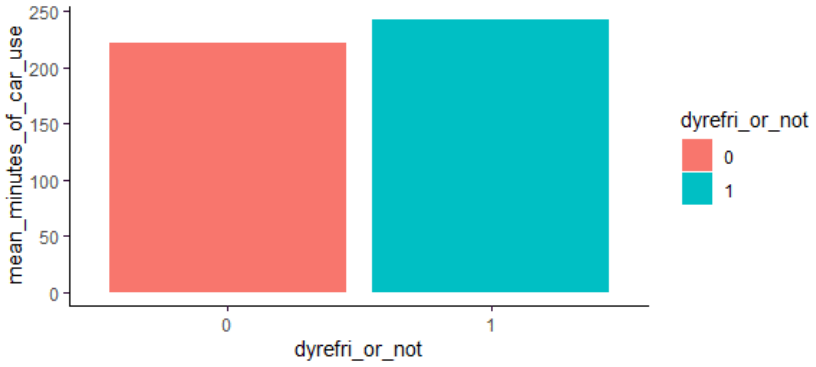
Bar chart 4: Car performance for each car type without outliers.



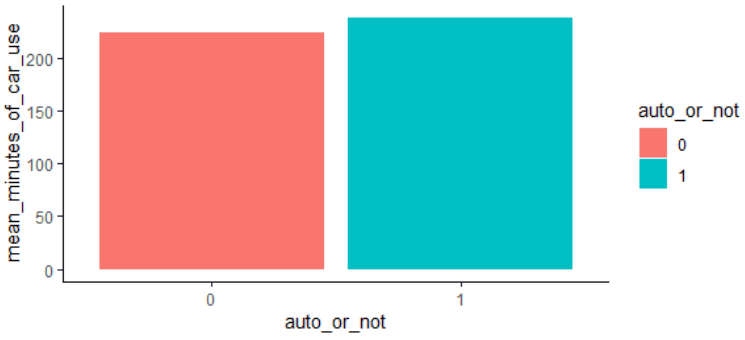
Bar chart 5: Number of cars with different fuel types.



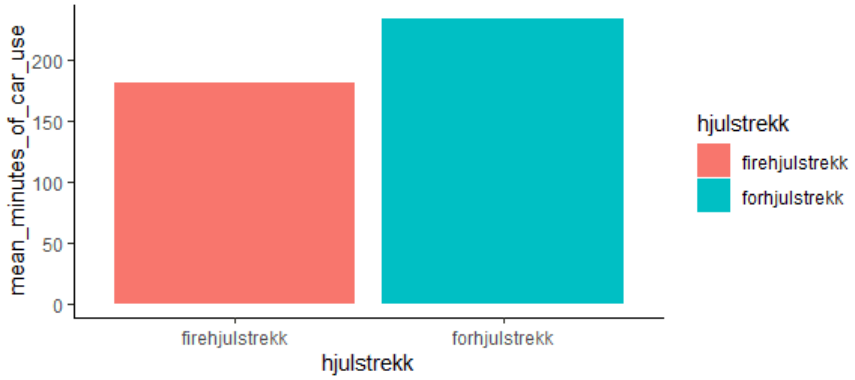
Bar chart 6: Car performance for cars with different fuel types.



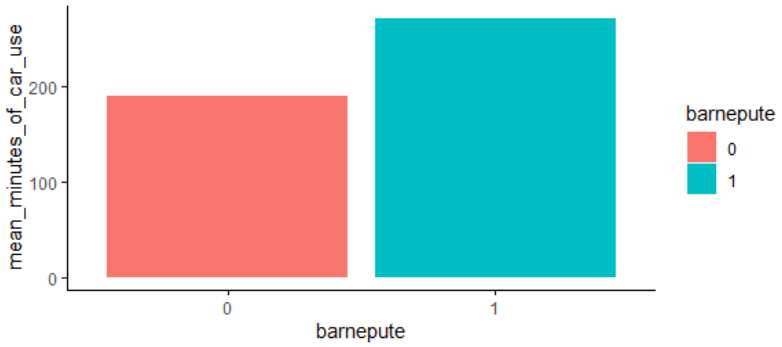
Bar chart 7: Car performance for cars with or without animal.



Bar chart 8: Car performance for cars with automatic gear shift or not.



Bar chart 9: Car performance for cars with or without 4WD.



Bar chart 10: Car performance for cars with or without child cushion.

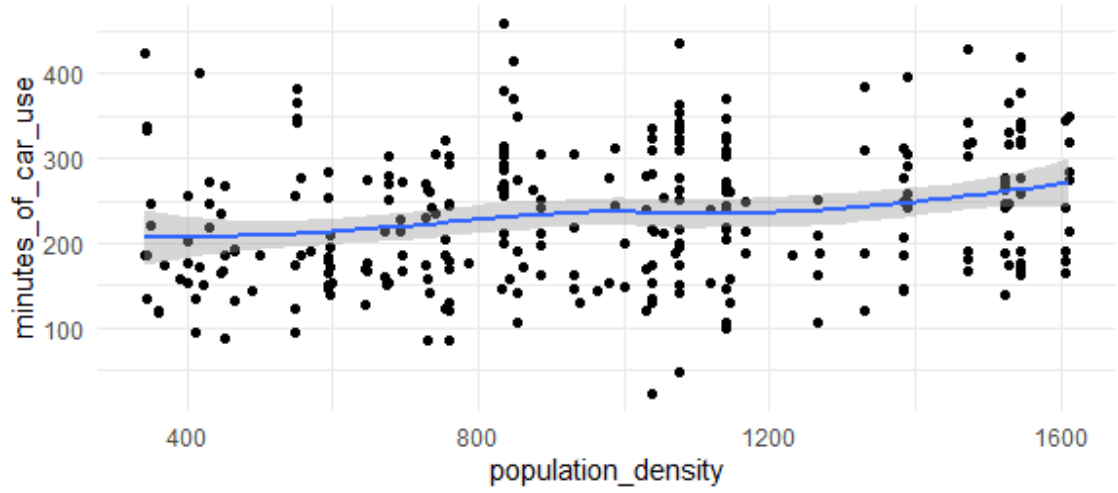


Figure 3: The relationship between population density and car performance.

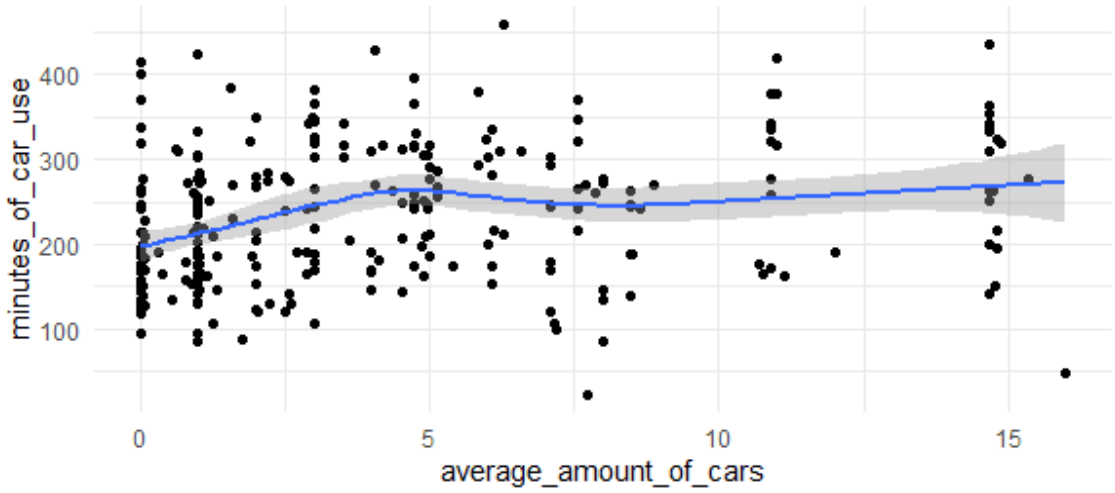


Figure 4: The relationship between average amount of cars and car performance.

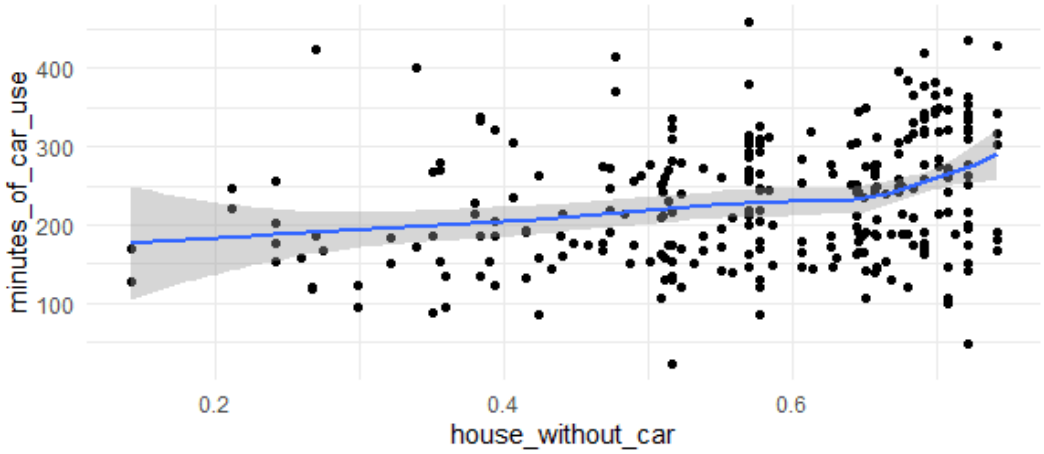


Figure 5: The relationship between households without cars and car performance.

2.4 Step one: Method used to determine drivers of demand

Multiple linear regression is used to determine drivers of demand. The aim of step one is to show which demographic and car-specific variables most affect car performance, as well as how strong that effect is. Multiple linear regression is chosen because inference is important for quantifying the importance of demand drivers. Multiple linear regression is a very straightforward approach for predicting a quantitative response Y on the basis of multiple

predictor variables X1, X2,...Xp (James, Witten, Hastie, & Tibshirani, 2013). It assumes that there is approximately a linear relationship between X and Y. The function can be written $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$. The coefficient estimates β0, β1, ...,βp can be found by minimizing the residual sum of squares (RSS) with least squares approach. RSS is the sum of difference between the ith observed response value and predicted ith response value by our multiple linear regression model, and the formula for RSS is given in Equation 1 below. Although there might be some bias regarding the true relationship between Y and X, multiple linear regression methods have a clear advantage in terms of interpretation. This means that one of the main assumptions related to the regression model is that the relationship between car performance and the independent variables is linear in nature. While assumption of linearity is fairly strong and restrictive, it is still pursued for its ease and simplicity in measuring performance (James et al., 2013).

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Equation 1: Formula for RSS residual sum of squares.

In the multiple regression setting with p predictors, the question of whether all the regression coefficients are zero presents itself, i.e. whether $\beta1=\beta2=\ldots\beta p=0$. We use a hypothesis test to answer this question, and we test the null hypothesis, H_0: $\beta1=\beta2=\ldots\beta p=0$ versus the alternative, Hα: at least one βj is non-zero. The null hypothesis is that there is no relationship between the predictors and the response, while the alternative hypothesis is that there is some relationship between at least one predictor and the response. This hypothesis is performed by computing the F-statistic. When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if Hα is true, then F is expected to be greater than one. Based on the p-value associated with the F-statistic, one can determine whether or not to reject H_0. More importantly, in order to determine which predictors are related to the response, and which are the noise variables, the individual p-value associated with the t-statistic should be examined. In addition, the coefficients estimate from the multiple linear regression indicates a positive or negative relationship between the predictors and the response, and how strong that relationship is (James et al., 2013).

2.4.1 Method for drivers of demand

Multiple linear regression

In the multiple linear regression model, variables which can be defined as individual attributes are included. These variables include demographic features such as gender, age, income level, education level, share of households without cars, as well as car features, such as car type, fuel type, and child seats. Employing the multiple linear regression model, the aim is to find the statistically significant variables, which could be the drivers of demand for Bildeleringen. The drivers of demand are the factors affecting car performance, which can reveal user preferences and local characteristics. These drivers of demand are important building blocks for the prediction model for car performance.

The dataset has a total of 286 observations and 42 variables including the dependent variable 'minutes_of_car_use'. Since the number of predictors is 41, which is small compared to the overall observations of 286, it is determined that the null hypothesis test and F-statistic are appropriate for the multiple linear regression model, which uses the least square fitting. Dummy variables are used for qualitative variables. They are car type, fuel type, automatic gear shift or not, wheel driving mode, animal free or not, child seat, child cushion, luggage rack or roof box, tow hitch and the four seasons. Other predictors are all continuous numeric variables.

- Car type: small car, wagon, van, minicar, big wagon, SUV or 9 seats
- Average amount of cars in the parking locations
- Fuel type: diesel, electricity or gasoline
- Automatic gear shift or manual gear shift
- Wheel driving mode, i.e. if the car is 4WD, BWD or FWD
- Animal free or not, i.e. if animals are allowed in the car or not
- Child seat
- Child cushion
- Luggage rack

- Roof box
- Tow hitch
- Population density
- Share of households without cars
- Share of people with different levels of education-five categories
- Share of people within different age groups-eleven categories
- Share of men/women
- Share of people with different working status-five categories
- Median income
- Seasons car has been available

Before the data set is fitted with multiple linear regression, the correlation between all the independent variables is checked, and those with high collinearity are removed. After some highly correlated variables are removed 26 independent variables remain. Then all the 26 variables are put into the multiple linear regression model to predict car performance, and the insignificant variables are removed one by one based on the largest p-value which is greater than 0.05. This procedure continues until all the remaining variables have a p-value below 0.05. Finally, there are 15 statistically significant variables obtained from this linear regression model. The details of the results can be seen in Table 7 in the next section. In terms of model assessment, the study uses p-value, F statistic and adjusted R^2.

In the multiple linear regression model, the response is the average minutes of car use for each car per day. The potential predictors are the 41 independent variables, which correspond to each vehicle of Bildeleringen. These variables are selected for the linear regression model based on the literature review and analysis of the dataset from Bildeleringen.

2.5 Step two: Methods used to create prediction models

In the second part, our study aims to build a predictive model for Bildeleringen in order to predict the car usage performance. Our paper utilizes five statistical methods on the same database and assesses the model performances with prediction accuracy, which is the cross-validation MSE and RMSE.

Five prediction models are evaluated: Linear Regression, Linear Regression-Forward stepwise selection, Lasso, Random forest, and Boosting. Since there are 41 independent variables and some of them are highly correlated, variable selection is required for the linear models. The first two models are multiple linear regression models. The first model, Linear Regression, uses the 15 significant variables found to be the drivers of demand in step one. The second model, Linear Regression-Forward stepwise selection, is an approach for automatically performing feature selection in R, i.e. for excluding irrelevant variables from a multiple linear regression model. A subset selection approach includes best-subset selection, Forward stepwise selection, and Backward stepwise selection methods, and it involves identifying a subset of the p predictors that are believed to be related to the response. The model with the subset of p predictors then uses least squares on the reduced set of variables. Because the demographic and the car specific variables amount to a total of 41 predictors, which is a high-dimension data set, it is not possible to use the best-subset selection method. In general, 2^p models involve all the combination of p predictors, therefore best-subset selection becomes computationally infeasible for values of p greater than around 40, even with extremely fast modern computers (James et al., 2013). Forward stepwise selection is a good alternative to best-subset selection. The third model, Lasso, is a shrinkage method which forces some of the coefficients to be exactly zero, after which variable selection can be performed. Lastly, Random Forest and Boosting are aggregated tree-based models which can also capture the non-linear relationship between response and predictors (James et al., 2013).

2.5.1 Cross-validation

Regarding the performance assessment of the five different models, the model with the lowest test error is chosen. In order to select the best model with respect to test error, the test error needs to be estimated. Specifically, 10-folds cross-validation RMSE as an estimate of test error is used in our study. The reason why our study chooses 10-fold cross-validation is that there is a bias-variance trade-off associated with the choice of K. When K=10, it has been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance (James et al., 2013).

The 10-fold cross validation approach involves that one dataset is divided randomly into 10 groups or folds of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining 9 folds for model training. The mean squared error, MSE_1 is then computed on the observations in the held-out fold with Equation 2 below. This procedure is repeated 10 times; each time, a different fold of observations is treated as a validation set. This process results in 10 estimates of the test error, MSE_1, MSE_2, ..., MSE_{10}. The 10-fold cross-validation error $CV_{(10)}$ is computed by averaging the 10 test MSEs with Equation 3 below, and the cross-validation RMSE is the square root of MSE (James et al., 2013). MSE measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. RMSE root-mean-square error is the square root of MSE.

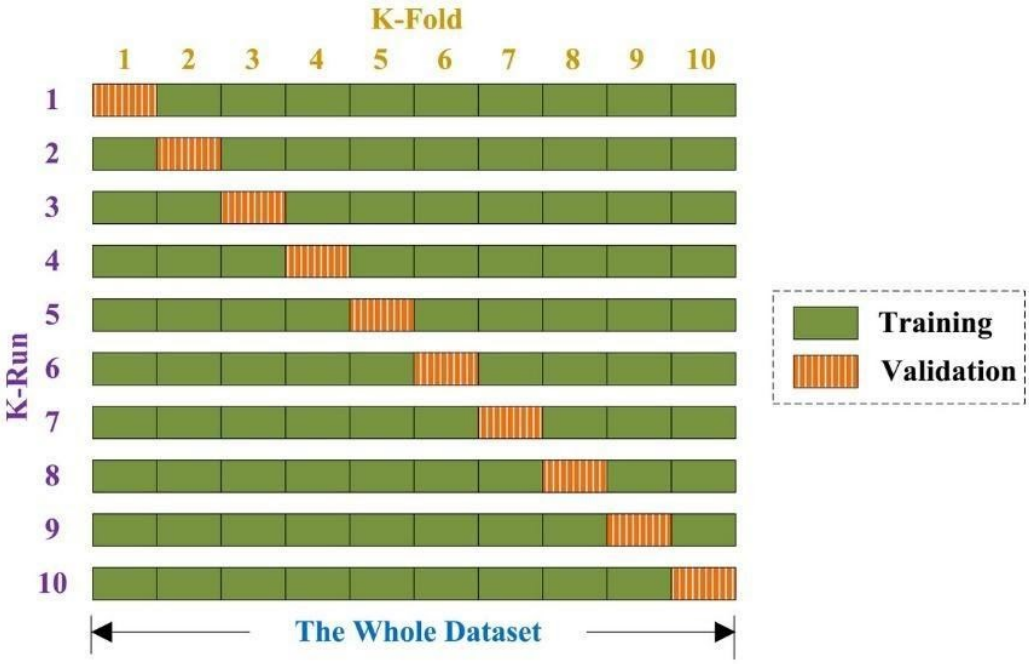


Figure 6: Illustration of 10-fold cross validation.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Equation 2: Equation for mean squared error.

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$

Equation 3: Equation for k-fold CV error (James et al., 2013).

Linear Regression

The 15 variables found to be significant in step one are used in this model. The variables in the model are presented below in Table 7. The model is applied on the full dataset with 10-fold cross validation in order to obtain the cross-validation MSE and RMSE as an estimate of the test error.

```
Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.924e+02  1.006e+02   1.912 0.056971 .
car_typeminibil           2.663e+02  4.159e+01   6.403 6.79e-10 ***
car_typesmabil            1.943e+02  3.505e+01   5.544 7.08e-08 ***
car_typestasjonsvogn      5.686e+01  2.163e+01   2.629 0.009053 **
car_typestor_stasjonsvogn 2.700e+00  2.665e+01   0.101 0.919361
car_typeSUV               3.377e+01  3.131e+01   1.078 0.281870
car_typevarebil           1.292e+02  2.129e+01   6.070 4.35e-09 ***
average_amount_of_cars    4.105e+00  1.103e+00   3.722 0.000241 ***
fuel_typediesel                  NA         NA      NA       NA
fuel_typeelektrisitet     2.421e+01  1.143e+01   2.119 0.034996 *
auto_or_not1              2.666e+01  8.500e+00   3.137 0.001896 **
barnepute1               -5.354e+01  2.920e+01  -1.834 0.067793 .
population_density       -4.352e-02  1.867e-02  -2.331 0.020495 *
house_without_car         3.574e+02  1.158e+02   3.087 0.002231 **
year20_24                -3.974e+02  1.807e+02  -2.199 0.028730 *
man                      -6.173e+02  2.382e+02  -2.591 0.010083 *
retired                  -2.798e+02  1.227e+02  -2.280 0.023374 *
median_income             2.818e-04  1.126e-04   2.503 0.012906 *
spring1                   2.506e+01  7.996e+00   3.134 0.001918 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.5 on 268 degrees of freedom
Multiple R-squared:  0.5432,    Adjusted R-squared:  0.5142
F-statistic: 18.75 on 17 and 268 DF,  p-value: < 2.2e-16
```

Table 7: The results of linear regression.

The 10-fold cross-validation error is 3272.167, and the RMSE is 57.20. This means that the estimate of test error from the model is an average of +/- 57.20 minutes from the actual value.

Linear Regression - Forward stepwise selection

For computational reasons, best subset selection cannot be applied with very large p predictors above 40 variables. Moreover, an enormous search space in the best subset selection method can lead to overfitting and high variance of the coefficient estimates. For both of these reasons, a forward stepwise method which explores a far more restricted set of models is an attractive alternative to best subset selection. Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. In more detail, at each step the variable that gives

the greatest additional improvement to the fit is added to the model. This is done by simply choosing the model with the smallest RSS or highest R^2 (James et al., 2013).

It is noteworthy that in order to yield an accurate estimate of the test error, only the training observations can be used to perform all aspects of model-fitting, including variable selection. Therefore, the determination of which model of a given size is best must be made using only the training observations. In this study the 10-fold cross -validation approach is used for both variable selection and optimal model choosing. More specifically, R automatically performs forward stepwise selection within each 9 folds training set and decides upon the 41 best models with different numbers of variables ranging from 1 to 41. They are then fitted on the remaining fold test dataset respectively to obtain the test error. This has resulted in a 10×41 matrix, of which the (i, j)th element corresponds to the test MSE for the ith cross-validation fold for the best j-variable model. The 10×41 matrix can be seen in Plot 8 in Appendix. The columns of this matrix are averaged in order to obtain a vector for which the jth element is the cross-validation error for the j-variable model (James et al., 2013). The vector can be found in Plot 9 in Appendix. The model with 17 variables has the lowest cross-validation error, which can be seen in Figure 7. Finally, the forward stepwise selection is fitted on the full dataset and the 17-variable model is obtained which is shown in Table 8. Again, 10-fold cross-validation is used on the multiple linear regression to get the cross-validation MSE and RMSE.

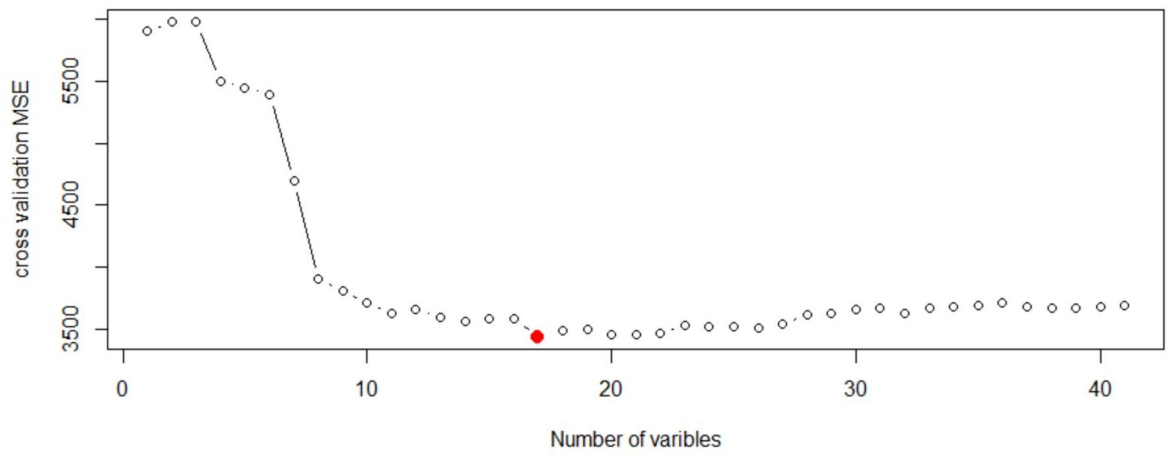


Figure 7: Models with different number of variables and cross validation MSE.

```
> coef(reg.best,min)  #17 variable on full data-set
                (Intercept)              car_typeminibil
                  157.480648                   149.510461
             car_typesmabil car_typestor_stasjonsvogn
                   81.645379                   -49.334851
           car_typevarebil    average_amount_of_cars
                  134.454433                    4.807429
      fuel_typeelektrisitet               auto_or_not1
                   25.620731                   20.263729
           dyrefri_or_not1          house_without_car
                    1.361737                   13.296553
               high_school              university1_4
                 -121.396408                   16.082311
                   year0_5                    year6_15
                  -47.281462                   60.440859
                  year30_49             fuel_typediesel
                  -14.476190                  -65.515341
    hjulstrekkforhjulstrekk                      year80
                   12.821105                  142.705564
```

Table 8: Results from the forward stepwise selection method.

The cross-validation MSE is 3490.926, which means that RMSE is 59.08. RMSE 59.08 implies that the estimated deviation from the model is an average of +/-59.08 minutes from the actual car performance.

Lasso

Lasso is one of the famous shrinkage methods. To fit this model some of the coefficient estimates are forced to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, Lasso performs variable selection, and it turns out the shrinkage can significantly reduce the variance. Therefore, compared to least squares, Lasso might increase both the prediction accuracy and interpretability due to the bias-variance trade-off. As λ increases, the flexibility of the Lasso regression fit decreases due to more coefficient estimates are shrunk towards zero, and the shrinkage of the coefficient estimates leads to a substantial reduction in the variance of the predictors, at the expense of a slight increase in bias; thereby leading to a decrease in test MSE. Similarly, to least squares fitting which minimizes RSS, Lasso regression coefficient estimates β, which is the values that minimize the quantity in Equation 5. Lasso uses a L1 penalty which is given by $\|\boldsymbol{\beta}\|_1=\Sigma|\boldsymbol{\beta}_j\|$, and the

tuning parameter λ serves to control the relative impact of the shrinkage penalty which is given in Equation 4 on the regression coefficient estimates (James et al., 2013).

$$\lambda \sum_{j=1}^{p} |\beta j|$$

Equation 4: Shrinkage penalty for Lasso.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Equation 5: The minimizing quantity for RSS plus shrinkage penalty in Lasso.

The λ in the model has a grid of values ranging from 10^{-2} to 10^{10}, and the best λ determined by the lowest 10-fold cross-validation is 2.04. The plot of different λ and the corresponding cross-validation error can be illustrated by Figure 8 below. The details of different λ and the corresponding cross-validation MSE can be found in Plot 10 in Appendix. The Lasso model with the best λ is then fit on the full dataset, and 17 variables are found to have non-zero coefficient estimates, as seen in Table 9. Finally, the 10-fold cross-validation is used in order to get the estimate of the test MSE. The MSE for each fold can be seen in Plot 11 in Appendix.

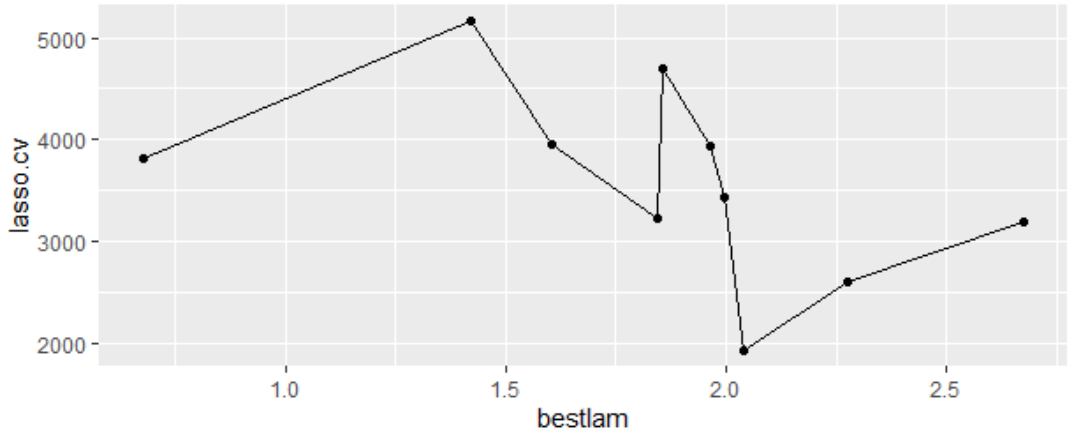


Figure 8: λ and cross validation MSE for Lasso.

```
> lasso.coef[lasso.coef!=0] #17 variables
             (Intercept)           car_typeminibil
              122.657990                129.640984
          car_typesmabil car_typestor_stasjonsvogn
               79.514572                -37.170704
        car_typevarebil    average_amount_of_cars
               72.969306                  2.903150
        fuel_typediesel     fuel_typeelektrisitet
               -2.150090                 21.080350
           auto_or_not1              hengerfeste1
               19.411799                -11.041167
            takboyler1               high_school
               -4.102650                 -3.640447
         university1_4          university_over4
               31.371524                 35.033848
            year16_19                 year30_49
             -138.497699                -67.184130
            year50_59                  employed
             -134.312858                91.532118
```

Table 9: Results from Lasso method.

Averaging the MSEs, the cross-validation MSE from the Lasso model is 3522.377 and the corresponding RMSE is 59.35 minutes. The result from Lasso is quite similar to linear regression in this study.

Random forest

A regression tree involves dividing the predictor space into J distinct and non-overlapping regions R_1, R_2, …, R_j. It consists of a series of splitting rules, starting at the top of the tree. The goal is to find boxes R_1, R_2, …, R_j that minimize the RSS, which can be seen in Equation 6. For every observation that falls into the region Rj, we make the same prediction, which is simply the mean of the response values for the training observations in Rj. A simple example of a regression tree and the three-region partition is given in Figure 9 and Figure 10. It is an example of a regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year. These three regions can be written as R_1={X|Years<4.5}, R_2={X|Years>=4.5, Hits<117.5}, and R_3={X|Years>=4.5, Hits>=117.5}.The predicted salaries for these three groups are $\$1{,}000\times e^{5.11}=\$165{,}670$, $\$1{,}000\times e^{6}=\$403{,}428$, and $\$1{,}000\times e^{6.74}=\$845{,}560$ respectively.

Regression trees are advantageous in that they are very easy to explain and can be displayed graphically. Some people believe they resemble human decision making more closely than other regression approaches. However, a single regression tree can be very non-robust, which means that it has high variance because of the random split of training data sets and test data sets.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Equation 6: RSS for regression tree.

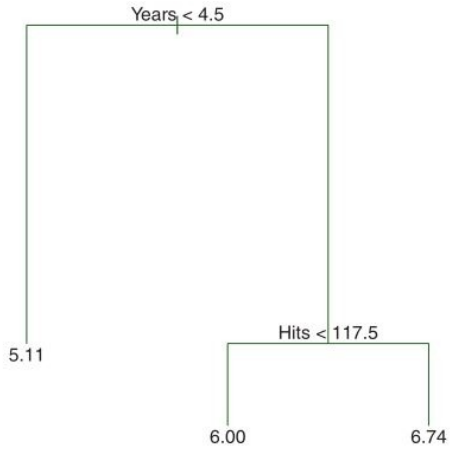


Figure 9: A simple regression tree.

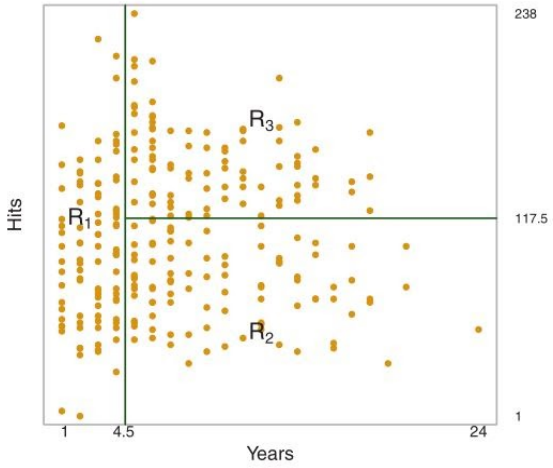


Figure 10: A three-region partition.

Random forest is a method which can reduce the variance from a single regression tree by aggregating many decision trees, which in turn can increase the prediction accuracy. The algorithm involves taking repeated samples from our dataset and generating B different training data sets. We then train our regression tree model on the bth training set in order to get $f^b(x)$, and finally average all the predictions with Equation 7, to obtain the final prediction results. When building these regression trees, Random forest forces each split to consider only a random sample of p/3 predictors which is 14 independent variables in our

study. The goal is to decorrelate the trees so that they do not look similar, thereby making the average of the resulting trees less variable and hence more reliable (James et al., 2013).

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x).$$

Equation 7: The average of all B regression trees as the final prediction.

Considering that there may be non-linear and complex relationships between the features and the response, a regression tree might outperform the linear regression model. This study makes use of the 10-fold cross-validation approach for the Random Forest model to get the cross validation MSE and RMSE. The MSE for each fold can be seen in Plot 12 in Appendix. The importance of each variable is shown in Figure 11 and Figure 12. Figure 11 is based on the mean decrease of accuracy in predictions on the test dataset when a given variable is excluded from the model. Figure 12 shows a measure of the total decrease in node impurity which in this study is training RSS that results from splits over that variables, averaged over all trees. The results show that the most important variables are car type, children cushions, children seats, average amount of cars in the parking spot, fuel type, household without car and share of employed people.

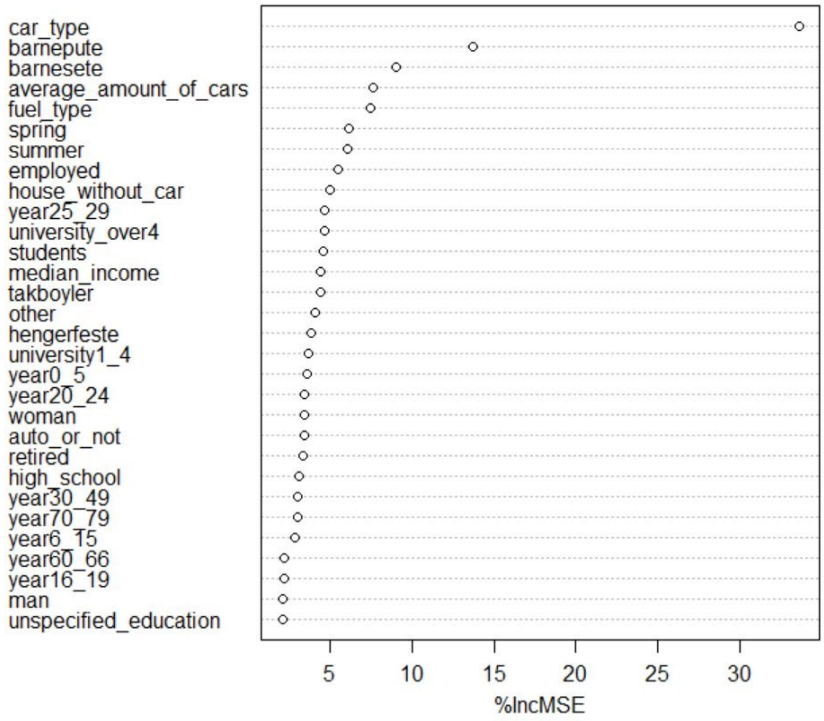


Figure 11: The importance of variables for decrease of MSE.

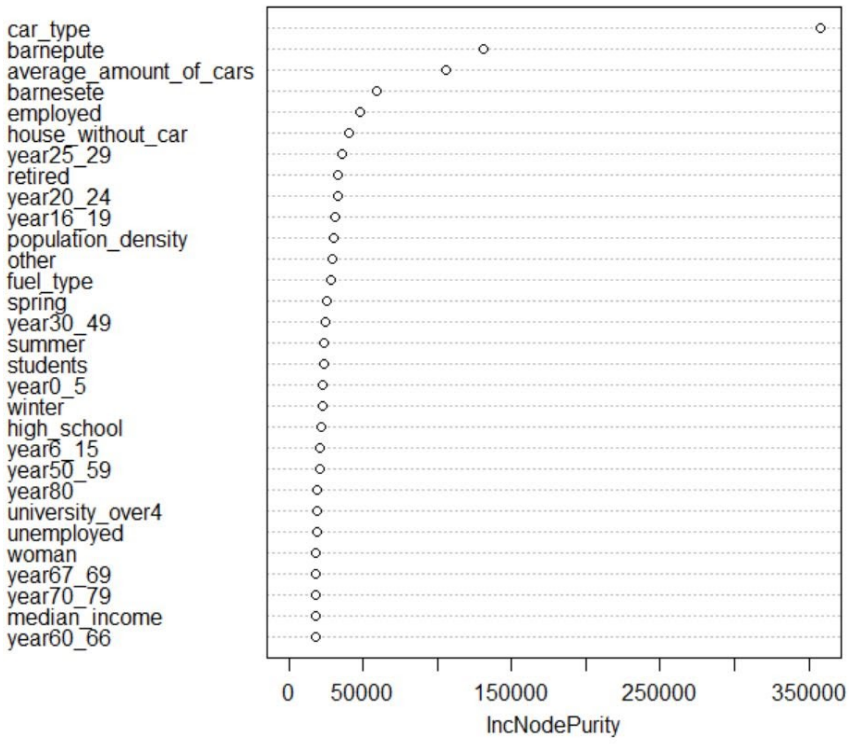


Figure 12: The importance of variables for decrease in node impurity.

The cross-validation MSE is 3935.296, and the RMSE is 62.73 minutes. The results are higher than for Linear Regression, Linear Regression - Forward stepwise selection, and Lasso.

Boosting

Boosting is another approach for improving the predictions from a decision tree. Unlike fitting a single large decision tree to the data, which amounts to fitting the data hard and also potentially overfitting, the boosting approach instead learns slowly. Similar to Random Forest, Boosting also involves averaging predictions over many trees, except that all the predictors can be used in each split and the trees are grown sequentially: Each tree is grown using information from previously grown trees. Boosting does not involve resampling from the original dataset, instead each tree is fit on a modified version of the original data set. That is, a regression tree is fitted using the current residuals, rather than the outcome Y as the response. A new decision tree is then added into the fitted function in order to update the residuals. Each of these trees can be rather small, and this model training process slowly improves the prediction. The algorithm of Boosting for regression trees can be seen in Algorithm 1 below, where r_i means residual (James et al., 2013). Similar to other models, the 10-fold cross validation is used for Boosting model. The MSE for each fold can be seen in Plot 13 in Appendix.

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x). \tag{8.10}$$
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \tag{8.11}$$
3. Output the boosted model,
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \tag{8.12}$$

Algorithm 1: The algorithm of Boosting for regression trees (James et al., 2013).

Compared to Random Forest, the growth of a particular tree in boosting takes into account the other trees that have already been grown, and smaller trees are typically sufficient. In the boosting model of this study, the number of splits in each tree d=1 is used, in which case each tree is a stump, consisting of a single split, and there are a total of 5000 trees grown. The number in Table 10 illustrates the marginal effect of the selected variables on the response after integrating out the other variables. It can then be seen, in Table 10, that the most important variables are car type, average amount of cars, share of employed and the age group 30-49. Figure 13 to Figure 16 plot the relationship between these important variables and car performance.

```
> summary(boost.fit)
                                           var   rel.inf
car_type                             car_type 8.7045387
average_amount_of_cars average_amount_of_cars 6.5666073
employed                             employed 5.3523361
year30_49                           year30_49 5.3045274
year80                                 year80 5.1925696
population_density         population_density 5.0869682
year16_19                           year16_19 4.2414120
other                                   other 4.1097318
year6_15                             year6_15 3.4028848
fuel_type                           fuel_type 3.1334093
unspecified_education   unspecified_education 2.9396297
unemployed                         unemployed 2.9072412
spring                                 spring 2.8501664
year0_5                               year0_5 2.8038144
house_without_car           house_without_car 2.7279228
retired                               retired 2.5489830
students                             students 2.4369213
summer                                 summer 2.3431444
year25_29                           year25_29 2.3377489
year70_79                           year70_79 2.3248420
year67_69                           year67_69 2.2297332
university_over4             university_over4 2.0385319
year60_66                           year60_66 1.8754037
year20_24                           year20_24 1.8454980
```

primary_school	primary_school 1.7282360
man	man 1.7205412
university1_4	university1_4 1.6460435
high_school	high_school 1.5299619
year50_59	year50_59 1.3501011
woman	woman 1.2846994
median_income	median_income 1.2778477
fall	fall 0.9256654
auto_or_not	auto_or_not 0.8353368
winter	winter 0.6545057
dyrefri_or_not	dyrefri_or_not 0.6271740
barnepute	barnepute 0.5649546
takboyler	takboyler 0.2875139
hengerfeste	hengerfeste 0.1605402
barnesete	barnesete 0.1023125
hjulstrekk	hjulstrekk 0.0000000
takboks	takboks 0.0000000

Table 10: The relevant importance of variables from Boosting.

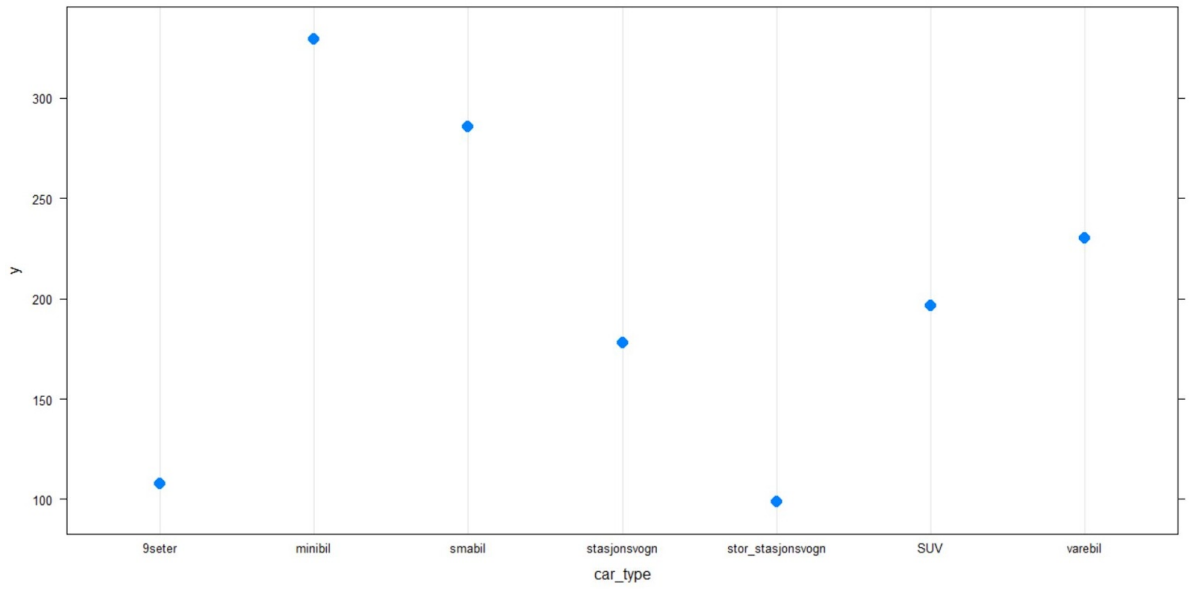


Figure 13: The relationship between car type and car performance.

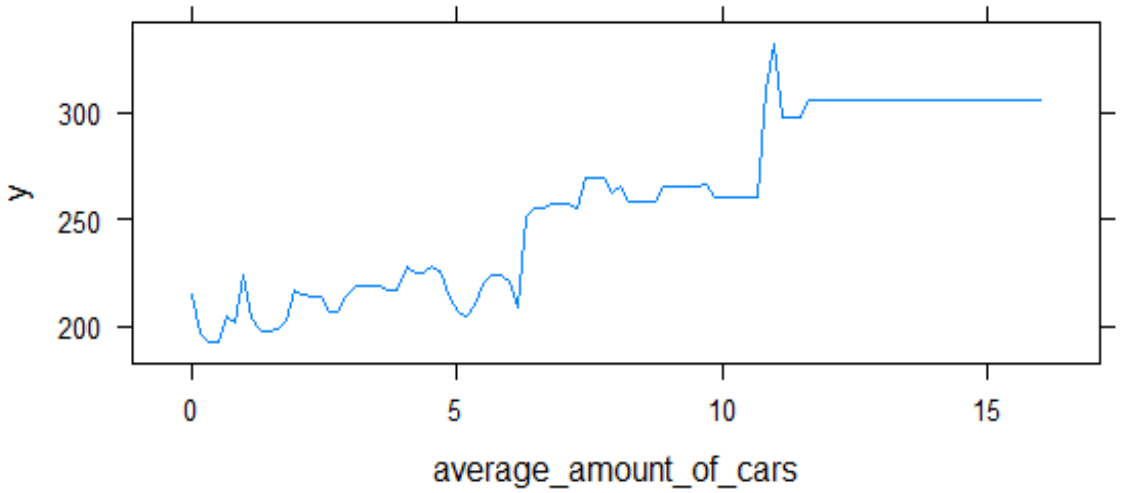


Figure 14: The relationship between average amount of cars and car performance.

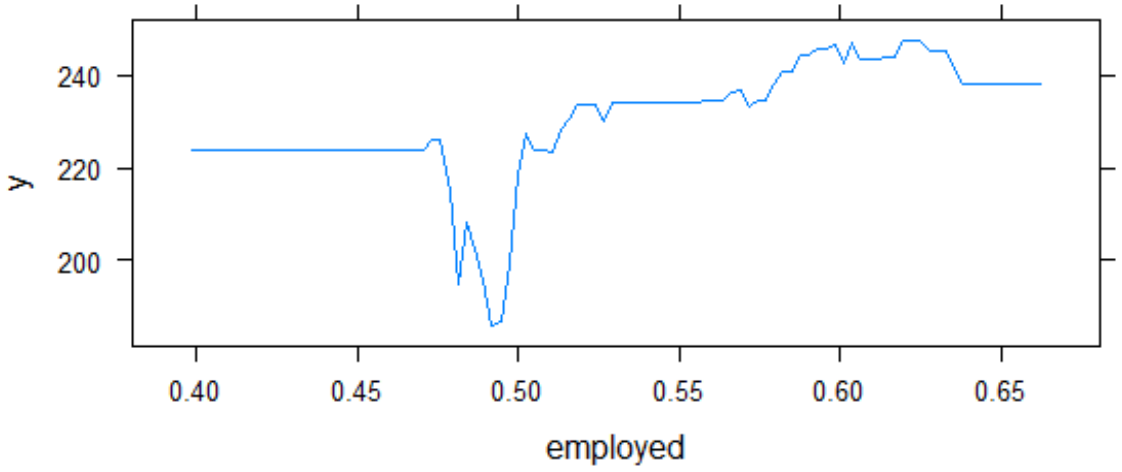


Figure 15: The relationship between the share of employed and car performance.

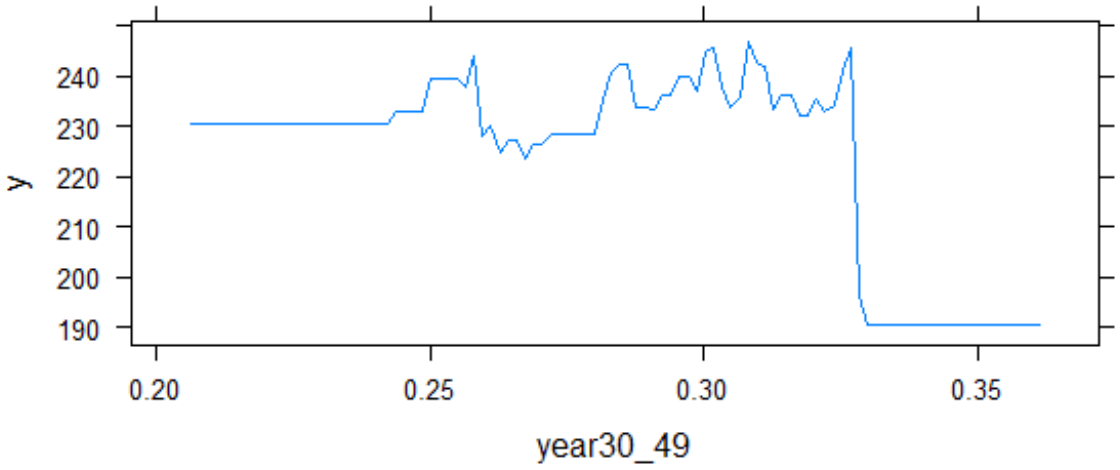


Figure 16: The relationship between the age group 30_39 and car performance.

The cross-validation MSE is 4877.104, and RMSE is 69.83 minutes from the Boosting model.

2.5.2 Model assessment summary

The number of variables, cross-validation MSE and RMSE are summarized in Table 11, and Figure 17 illustrates the cross-validation RMSE from each model. It clearly shows that the Linear regression model has the smallest cross-validation RMSE which indicates the best predictive performance.

	Number of variables	Cross-validation MSE	Cross-validation RMSE
Linear regression	15	3272.167	57.20
Linear regression - Forward stepwise selection	17	3490.926	59.08
Lasso	17	3522.377	59.35

Random forest	N/A	3935.296	62.73
Boosting	N/A	4877.104	69.83

Table 11: Summary of model assessment.

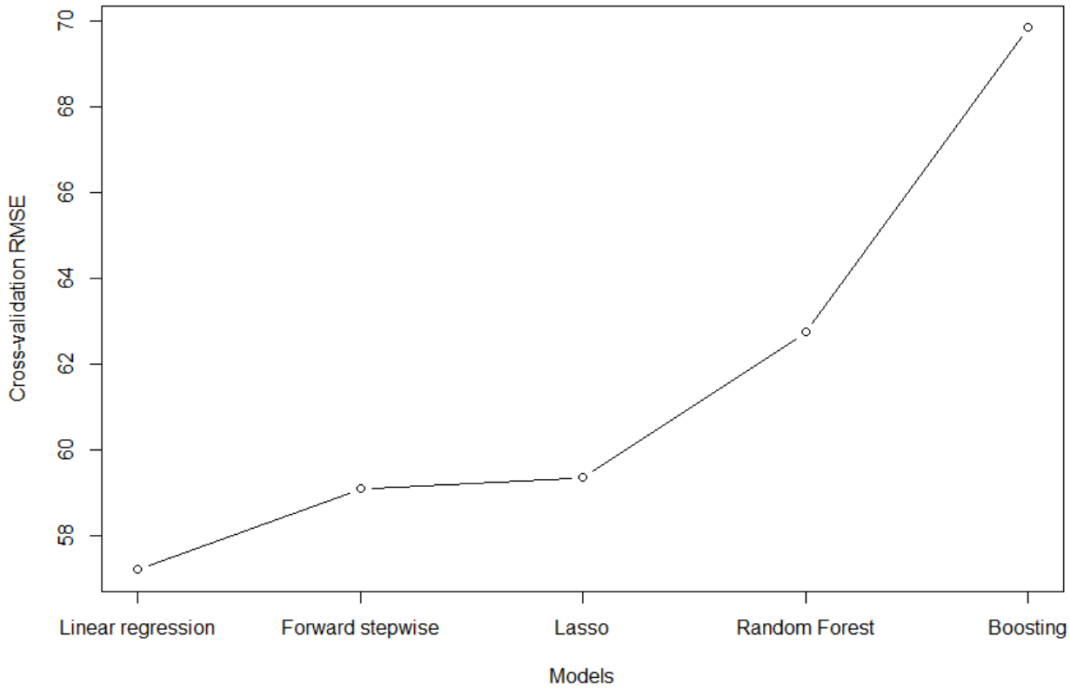


Figure 17: The cross-validation RMSE of different models.

Based on the cross-validation MSE and RMSE results from the five prediction models, the linear models perform better compared to non-linear models in this paper's dataset. Both the results from Random forest and Boosting are worse than the linear models, probably because the true relationship between the predictors and the response in nature is linear. Despite the advantageous graphical display of variable importance possible with the tree-based methods, the models have higher cross-validation MSE. This indicates worse predictive accuracy compared to linear methods. Therefore, Random forest and Boosting are not considered as the optimal predictive model for this study.

In addition to the smallest cross-validation error, the reason for our study to choose the linear regression to be the optimal predictive model also attributes to the reasonable variables it chooses, compared to other models. The first linear regression model, with the 15 significant

drivers of demand, performs slightly better than the forward stepwise selection and Lasso methods. More importantly, although some common variables in these three models have similar coefficient estimates, for instance the influence of car type, fuel type, automatic gear shift and the average amount of cars in parking locations, the variables selected in the first linear model are more preferable than the 17 variables selected in the other two models because of better interpretation. For example, the share of households without cars is important, but it is not included in the Lasso method. On the other hand, the variables tow hitch and luggage rack in Lasso method are highly correlated with car type, as usually big cars have these facilities; therefore, these two variables are not very good options.

Regarding the model forward stepwise selection, seasons are not included in the model although it is an important factor. The front wheel drive variable has a positive relationship with car performance, possibly because it is highly correlated with car type, as usually small cars which have good car performance have front wheel drive. The age group 6-15 years old, as well as over 80 years old are included in the model, but the impacts are positive which is unreasonable as people in these age groups usually do not use car-sharing service based on our intuition and literature review. Based on the analysis of the results, the variables in the first model have least correlation and include all important variables. Therefore, the first multilinear model with 15 variables, and the adjusted R^2 51.4% is chosen as the optimal predictive model because of both high prediction accuracy and interpretability.

2.5.3 Using the prediction model on the new validation dataset

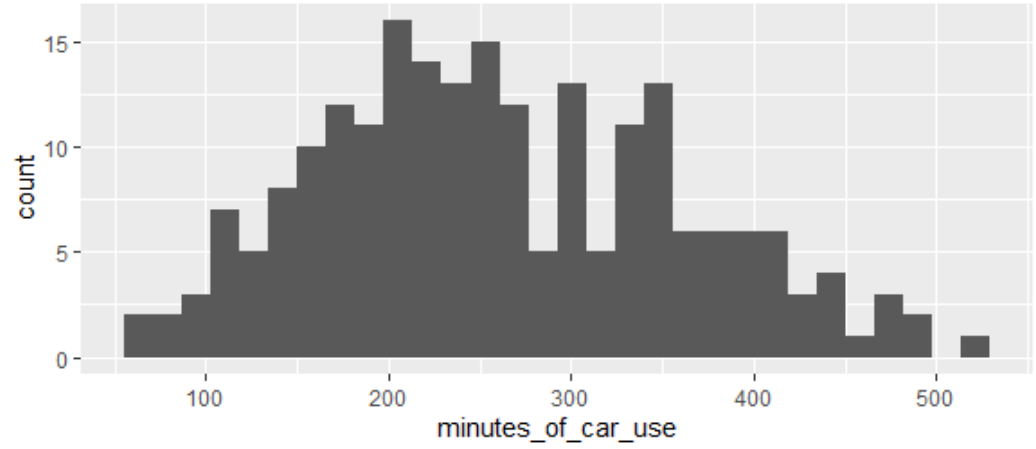
In this section the prediction model tests data from 12th of January 2020 to 12th of March 2020. The new data consists of trip data, and all demographics remain unchanged while car specific variables are updated for the new time period. 12th of March has been chosen as the end date, because movement restriction measures were implemented in Norway soon thereafter due to the COVID-19 pandemic.

The test dataset has 215 observations. The number of observations for this new time period is lower than the number of observations from the original time period, because less cars were relocated by the company. The 42 variables remain the same. An overview of the first

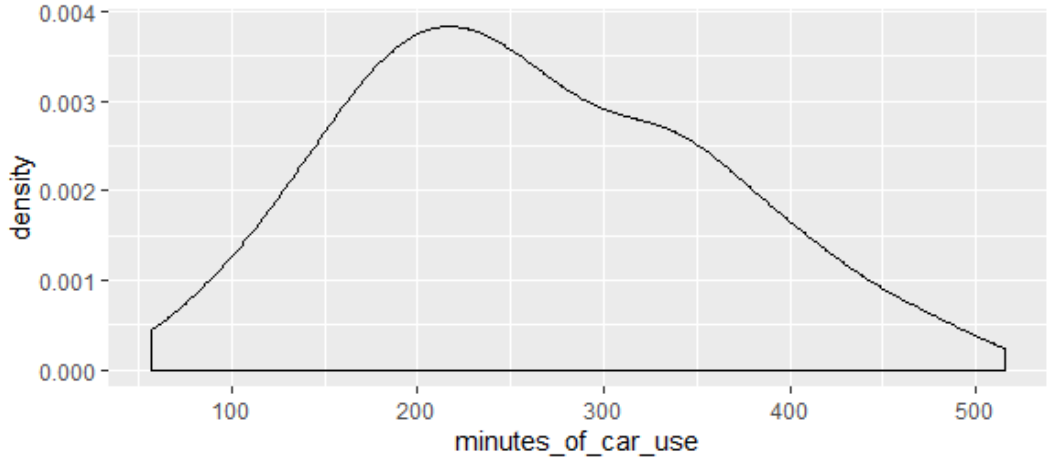
few lines of the validation data set can be seen in Plot 14 in Appendix. From Table 12, we can see that car performance per day is at an approximately normal level, with a mean value of 260.97 minutes, minimum value of 57.11 minutes, maximum value of 515.47 minutes, and standard deviation at 98.51. From the Histogram 2 and Density plot 2 below, we can see that the car performance is approximately normally distributed, and the Boxplot 5 shows that there is no outlier in the dataset.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
57.11	187.58	250.30	260.97	334.24	515.47

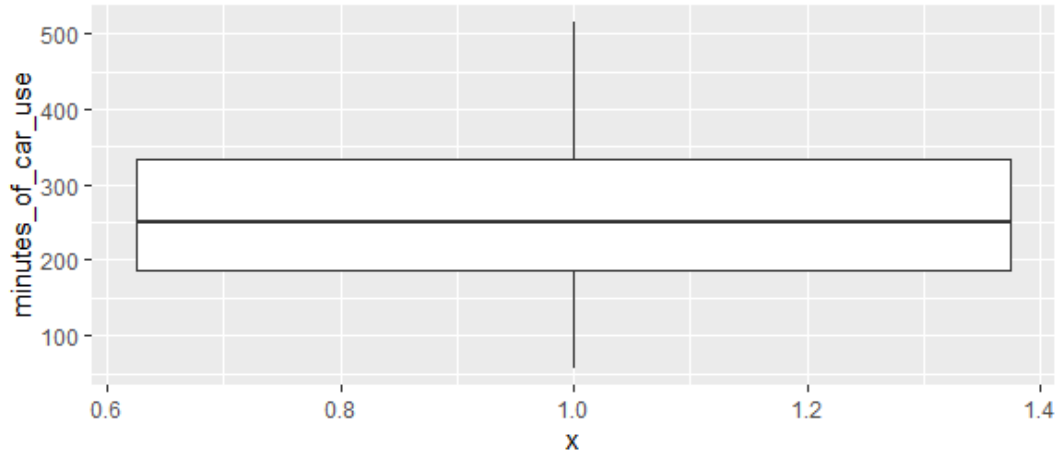
Table 12: Distribution of car performance in the validation dataset.



Histogram 2: Distribution of car performance in the validation dataset.



Density plot 2: Distribution of car performance in the validation dataset.



Boxplot 5: Distribution of car performance in the validation dataset.

3. Result

3.1 Drivers of demand

There are 15 statistically significant variables, which for car type include minicar, small car, wagon and van, in addition to average amount of cars in the parking spots, electric car, cars with automatic gear shift, child cushion, population density, share of households without a car, the age group 20-24 years old, share of males, share of retirees, median income and spring which can be seen in Table 7. Looking at the F-statistic, it was apparent that the large number 18.75 and the corresponding p-value < 2.2e-16 indicated that there is some relationship between the response and the predictors. More importantly, the adjusted R square is 51.42%, which implies that the model can explain 51.42% of the variability of the response. The performance of the linear regression model is a good fit to the dataset and can capture the drivers of demand effectively. The significant variables are stated below:

- Car type-minicar, small car, wagon, van (+)
- Average amount of cars in the parking spots (+)
- Fuel type-electric car (+)
- Car with automatic gear shift (+)
- Child cushion (-)
- Population density (-)
- The share of households without a car (+)
- The age group 20-24 years old (-)
- The share of men (-)
- The share of retired (-)
- Median income (+)
- Spring (+)

3.2 The prediction model - Linear Regression

The optimal predictive model is determined to be the Linear Regression method. The 10-fold cross-validation error is 3272.167, and the RMSE is 57.20, which means that the estimate of test error from the model is an average of +/-57.20 minutes from the actual value.

3.3 Using the prediction model on data from January to March 2020

The test dataset from 12 January to 12 March 2020 is fitted by the Linear Regression predictive model. The testMSE is 6862.709, testRMSE is 82.84.

4. Analysis and discussion

4.1 Drivers of demand

From the multiple linear regression results, we can interpret the relationship between the car performance and each of the predictors as well as how much impact the predictors have on the car performance. Most of the variables are positively related to the car performance, for example some car types, the average amount of cars in the parking spots, electrical car, auto car, the share of households without car, median income and spring. While some variables are negatively related to car performance, such as child cushion, population density, age group 20_24, the share of men and retired people. Each of them are discussed here.

According to the results car type is the most important variable. The estimated coefficients indicate that if the car type is a mini car holding other variables fixed, the car performance increases by an average of 266 minutes each day relative to a 9-seat car which is the baseline. Similarly, the car performance increases by 194.3 minutes for small cars, 56.8 minutes for wagon, and 129.2 minutes for van respectively. It seems that small cars are quite popular, probably because people usually use them for short distance travel with few people together, and smaller cars are cheaper to rent. In addition, we might be able to imagine that a van is preferable for long distance travel or moving to a new house.

If the average amount of cars in the particular parking spot increases by 1 unit, the car performance increases by 4.1 minutes. It seems that a larger supply of car capacity leads to more demand for the cars located at that location. Logic dictates that the cars would cannibalize on each other's demand, but this does not seem to be the case. There are, however, other explanations for these results. It might be that a car is only used up to a certain amount of time, and by introducing a new car to the same parking location, more demand can be fulfilled. There would of course be a cut off point for demand at some point, but it does not seem as if that point has been reached by Bildeleringen at these locations yet. Another explanation may be that placing many cars together creates a hot-spot for car-sharing services. Khan and Machemehl (2017) found a similar effect with a free-floating car-sharing service.

The effect of automatic cars, as opposed to manual stick shift cars, is extra car use of 26.6 minutes. These findings comply with our intuition that people prefer auto cars. People might prefer automatic cars because they are not able to drive cars with a manual stick shift.

The effect of electric cars, as opposed to fossil fuel cars, is extra car use of 24.2 minutes. These findings seem reasonable, and are reflected by Efthymiou & Antoniou (2016), who found that people who are concerned with being environmentally friendly are more willing to join car-sharing services.

The results show that the availability of child cushion in cars leads to less minutes of car use. If a car has a child cushion, the car usage decreases by 53.54 minutes. In the paper by De Lorimier and El-Geneidy (2013) a child seat in the car had a positive relationship with availability of the car, which means that car performance decreased. Child cushion and child seat are treated as two different variables in this study. While child seats are insignificant in this study, it can be seen that the findings in this paper is similar to the findings of De Lorimier and El-Geneidy (2013).

Though the effect is not strong, the negative relationship between car performance and population density needs to be discussed. The results show that if the population around a parking spot increases by 100 people, the car usage decreases by 4.35 minutes. Logic dictates that this should not be. There are several arguments for the fact that the relationship might not be strongly positive in this study.

As previously mentioned, not all studies have found population density to be a relevant factor. Stillwater et al. (2009) did not find population density to be a relevant factor in the model they built. Seeing that the study described in this paper investigates new data, it might be that population density is not a relevant factor for this study as well.

If we assume that the parking spots of Bildeleringen are randomly distributed around Bergen, it would be logical that we would find a stronger positive relationship. Placing parking spots at random would include locations not suited for car sharing, due to very low population density. This is not the case though. Parking spots are not placed at random. On the other hand, they are carefully considered by Bildeleringen's staff. All of the parking spots are therefore already placed in dense neighborhoods, compared to many other locations

in Bergen. One possible solution might be that population density and car performance only have a positive relationship up to a certain point, i.e. that the curve is concave and flattens when population increases. Possibly, all of Bildeleringen's parking spots are then near the "top of the curve", and a possible positive relationship between population density and car performance would not show.

Moreover, urban locations have specific features that may negatively affect the use of cars. These are higher traffic congestion, which would make people less willing to drive at all, more toll booths, and better public transport alternatives than other less populated areas.

Another finding comes from investigating some of the more popular parking spots. Some of them are close to parks, and one could argue that this creates "false ruralism" for these areas. In this study the population density is measured by counting the number of people in a radius of up to 1000 meters from the parking spot, and a park located within this radius would lower the measured population density. This is not accurate in that it does not capture real circumstances, but instead "false ruralism".

Some of the parking spots are located right next to large university campuses, which is a variable this study has not controlled for. It might be that closeness to a university campus is an important driver of demand. This was found by Kumar and Bierlaire (2012). Also, this paper has not controlled for large public facilities or leisure facilities, for example hospitals, shopping centers, large public transport hubs, sports stadiums or other places where large groups regularly gather.

Lastly, some of the parking spots are large in size, which would slightly affect measurements of population density.

The results show that when the share of households without a car increases by 1 percentage point, the car performance increases by 3.57 minutes. This is in line with previous literature (Catalano et al., 2008; Celsor & Millard-Ball, 2007; Clewlow, 2016; Khan & Machemehl, 2017; Millard-Ball, 2005; Stillwater et al., 2009), and might be a factor for Bildeleringen to consider when deciding where to put new parking locations.

People in the age group 20-24 years old are negatively related to car performance. If the share of the age group 20-24 increases by 1 percentage point, the car performance decreases

by 3.97 minutes. This seems reasonable, because people in this age group have lower disposable income and many of them are students. In this study the effect is seen clearly close to student villages, which naturally have high occurrences of people in the age group 20-24 years old. In the data source it can be seen that some student village locations perform average (“Alrek Studenthjem”, “Fantoft Studentby”, “Nattland Studentby”), while others perform poorly (“Gyldenprisveien Studentboliger”, “Løbergsveien Studentboliger”).

The share of men is also negatively related to car performance. If the share of men increases by 1 percentage point, the car performance decreases by 6.17 minutes. Previous literature, from France, has shown that car-sharing users are predominantly male (Kumar & Bierlaire, 2012). It is unclear if this is the case in Norway. It is also possible that more men than women have a driving license in France compared to Norway.

The share of retired people is negatively related to car performance too. If the share of retirees increases by 1 percentage point, the car performance decreases by 2.79 minutes. This is in line with previous literature, which has shown that car-sharing users tend to be young (Dias et al., 2017).

Although the median income is a positive significant variable, the impact is very small. More specifically, if the median income increases by 10,000 NOK, the car use increases only 2.81minutes. Previous literature has not been clear on what the effect would be when it comes to median income. Also, median income is only provided at the level of city districts, which might make the data insufficient for analysis.

During spring car performance increases by 25.06 minutes. However, data on seasons are only provided for one year. Therefore, these findings are vulnerable to any number of effects, which are not controlled for in this paper.

4.2 Prediction on the validation dataset

The average testRMSE is 82.84 minutes for the validation data, which is gathered between 12th of January 2020 and 12th of March 2020. Since the mean value of car usage is 260.97 minutes with standard deviation of 98.51, the testRMSE is within one standard deviation

from the mean value. In addition, the test error is 31.74% of the actual mean car performance, which is 260.97. This demonstrates that the linear prediction model has good predictive power. The results of prediction and the actual car performance is shown in Figure 18, in which the blue line is the Loess regression line with 95% confidence interval, and the red line is the 45-degree line on which the prediction value is equal to the actual value. In order to analyze the results in more detail, actual car performances are colored with 25%, 50%, 75% quantiles, which are 188, 250 and 334 minutes respectively in Figure 19. It can be seen that the blue line is close to the red line when the actual value is between 150 and 350 minutes. However, there are some values with large deviations when the actual value is smaller than 150 or greater than 350 minutes. The model tends to overestimate car performance when the actual value is under 150 minutes but tends to underestimate it when the actual value is over 350 minutes.

When our study fits the model on a subset of the dataset with actual car performance between 150 and 350 minutes, the testRMSE reduces to 61.27 minutes, which is very close to the cross-validation error 57.2 minutes. The comparison of prediction and the actual results can be seen in Figure 20, which shows a zoomed in subset of the data, specifically between 150 and 350 minutes. In Figure 20 the blue line is close to the red line. This implies that the predictive model has good performance on the dataset when actual car performance is between 150 and 350 minutes, which consists of the most middle part, 67%, of the validation dataset.

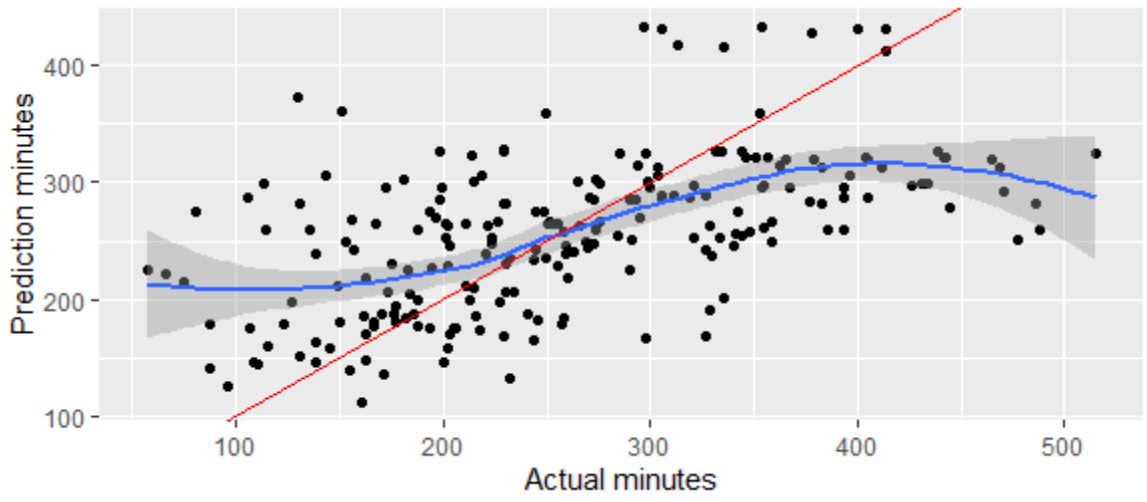


Figure 18: The prediction and actual value of car performance on the validation dataset.

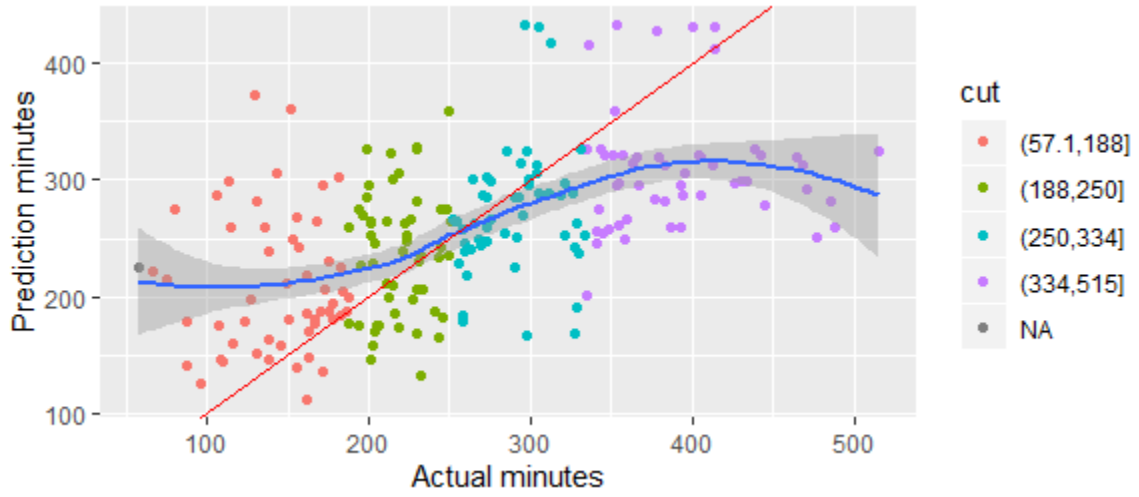


Figure 19: The prediction and actual value of car performance on the validation dataset colored by quantiles.

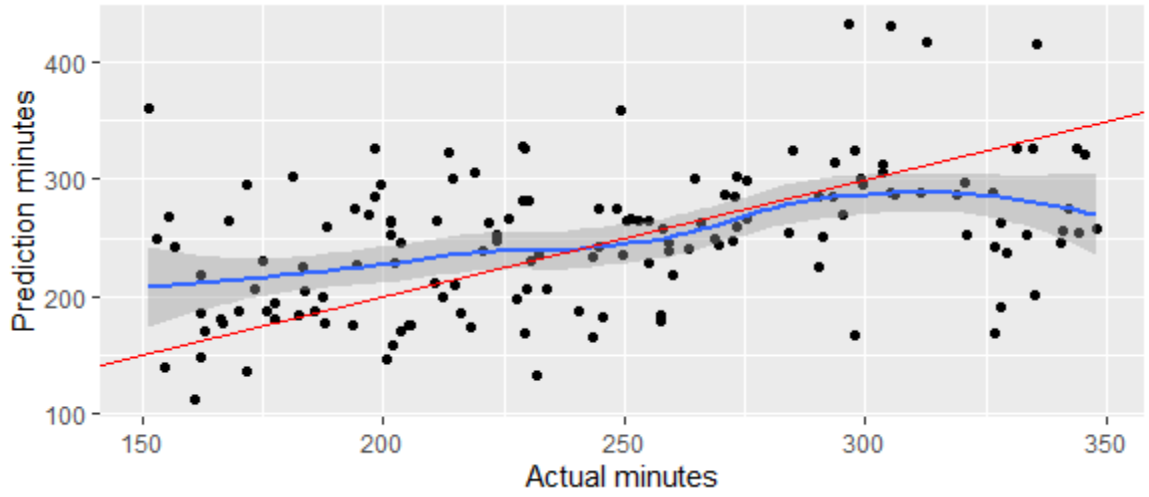


Figure 20: The prediction and actual value of car performance on the subset of validation dataset between 150 and 350 minutes.

4.3 Practical recommendations for Bildeleringen

As previously mentioned, the demographic variables and car performance are not found to have a strong relationship in this study. We expect Bildeleringen to have some ideas on

which car features are popular, and the results from this study might confirm or give more insight into this. Though the relationship between the demographic variables and car performance is not strong, it is still useful to discuss the practical recommendations the results imply.

1. Small cars, minicars, wagons and vans are important drivers of demand for car performance. In addition, electric cars and cars with automatic gear shift are significantly popular among customers, while the occurrence of child cushions decreases car performance. This is useful information for Bildeleringen when it comes to new product development.

2. The percentage of male population has a negative relationship with car performance. No areas of the city have a large imbalance in male/female population, and the share of male/female stays at around 50% for the whole city. It is therefore difficult to give any recommendations on this finding.

3. The share of 20-24 year olds has negative relationships with car performance. As previously discussed, the negative relationship of the age group 20-24 years old might be because of parking spots near student villages that perform badly to mediocre. According to our results, expansion should not be concentrated near student villages in the future.

   Though it is unclear why 20-24 year olds do not use car-sharing as much, there are several possible ways of increasing use in this age group. One possibility would be promotions, especially at times when people in this age group are likely to need a car. For instance, students need to move their possessions and buy items both early and late in the semester.

4. The share of retirees has a negative relationship with car performance. Also, the share of retirees is lower in the city center and higher further away from the city center. According to our results, parking spots should therefore be located in urban locations close to the city center.

5. The share of households without a car has a positive relationship with car performance. Areas in Bergen that have large shares of households without a car are the city center, and to some degree the areas immediately north and south of the city center.

   Bildeleringen can target specific areas in the city where the share of households without a car is high. One possibility is to use data from Statistics Norway on number of households with or without a car at the GK-level, as explained in this paper. Another possibility is to target specific residences that do not offer the possibility of owning a car, for example large apartment complexes with limited parking space.

5. Conclusion

The drivers of demand for the utilization of a car-sharing service are 1) car-specific variables, including being a smaller car or a moving van, being an electric car and having automatic gear shift, average amount of cars at the parking spot and having a child cushion, 2) high percentage of households without a car, 3) age-specific variables, especially not early 20s or retired, 4) season, i.e. spring brings higher demand, and 5) gender, i.e. large percentage of males lessens demand.

The relationship between demographic variables and car performance is not strong. It is therefore difficult to determine where Bildeleringen should place their new parking spots to have high utilization of their cars. However, it is possible to say that urban locations closer to the city center perform better.

Some variables are not tested in this study. These include public transport hubs, large public facilities, such as shopping malls, hospitals and universities, as well as different housing categorizes, such as student villages. All of these might affect car performance and can be included in further study on car-sharing. Further research can also be done in other cities in Norway using many of the same data sources. We believe that our study can serve as an important input regarding further research on car-sharing location problems in Norway.

References

Bildeleringen. (2020). Bildeleringen. Retrieved from https://bildeleringen.no

Catalano, M., Lo Casto, B., & Migliore, M. (2008). Car sharing demand estimation and urban transport demand modelling using stated preference techniques.

Celsor, C., & Millard-Ball, A. (2007). Where does carsharing work? Using geographic information systems to assess market potential. *Transportation Research Record, 1992*(1), 61-69.

Cheng, Y., Chen, X., Ding, X., & Zeng, L. (2019). Optimizing location of car-sharing stations based on potential travel demand and present operation characteristics: The case of chengdu. *Journal of Advanced Transportation, 2019*.

Ciari, F., Schüssler, N., & Axhausen, K. W. (2010). Estimation of car-sharing demand using an activity-based microsimulation approach: Model discussion and preliminary results. *Arbeitsberichte Verkehrs-und Raumplanung, 632*.

Clewlow, R. R. (2016). Carsharing and sustainable travel behavior: Results from the San Francisco Bay Area. *Transport Policy, 51*, 158-164.

Cohen, A. P., Shaheen, S., & McKenzie, R. (2008). Carsharing: A guide for local planners.

Coll, M.-H., Vandersmissen, M.-H., & Thériault, M. (2014). Modeling spatio-temporal diffusion of carsharing membership in Québec City. *Journal of Transport Geography, 38*, 22-37.

De Lorimier, A., & El-Geneidy, A. M. (2013). Understanding the factors affecting vehicle usage and availability in carsharing networks: A case study of Communauto carsharing system from Montréal, Canada. *International Journal of Sustainable Transportation, 7*(1), 35-51.

Dias, F. F., Lavieri, P. S., Garikapati, V. M., Astroza, S., Pendyala, R. M., & Bhat, C. R. (2017). A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation, 44*(6), 1307-1323.

Efthymiou, D., & Antoniou, C. (2016). Modeling the propensity to join carsharing using hybrid choice models and mixed survey data. *Transport Policy, 51*, 143-149.

Ferrero, F., Perboli, G., Rosano, M., & Vesco, A. (2018). Car-sharing services: An annotated review. *Sustainable Cities and Society, 37*, 501-518.

Firnkorn, J., & Müller, M. (2015). Free-floating electric carsharing-fleets in smart cities: The dawning of a post-private car era in urban environments? *Environmental Science & Policy, 45*, 30-40.

He, L., Mak, H.-Y., Rong, Y., & Shen, Z.-J. M. (2017). Service region design for urban electric vehicle sharing systems. *Manufacturing & Service Operations Management, 19*(2), 309-327.

Ion, L., Cucu, T., Boussier, J.-M., Teng, F., & Breuil, D. (2009). Site selection for electric cars of a car-sharing service. *World Electric Vehicle Journal, 3*(3), 531-540.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Jorge, D., & Correia, G. (2013). Carsharing systems demand estimation and defined operations: a literature review. *European Journal of Transport and Infrastructure Research, 13*(3).

Kabra, A., Belavina, E., & Girotra, K. (2019). Bike-share systems: Accessibility and availability. *Management Science*.

Kartverket. (2018). *Matrikkelen - Adresse*. Retrieved from: https://kartkatalog.geonorge.no/metadata/matrikkelen-adresse/f7df7a18-b30f-4745-bd64-d0863812350c

Khan, M., & Machemehl, R. (2017). The impact of land-use variables on free-floating carsharing vehicle rental choice and parking duration. In *Seeing Cities Through Big Data* (pp. 331-347): Springer.

Kumar, P., & Bierlaire, M. (2012). *Optimizing locations for a vehicle sharing system.* Paper presented at the Swiss Transport Research Conference.

Lindloff, K., Pieper, N., Bandelow, N. C., & Woisetschläger, D. M. (2014). Drivers of carsharing diffusion in Germany: an actor-centred approach. *International Journal of Automotive Technology and Management, 14*(3/4), 217.

Martin, E., Shaheen, S. A., & Lidicker, J. (2010). Impact of carsharing on household vehicle holdings: Results from North American shared-use vehicle survey. *Transportation Research Record, 2143*(1), 150-158.

Millard-Ball, A. (2005). *Car-sharing: Where and how it succeeds* (Vol. 108): Transportation Research Board.

Mindur, L., Sierpiński, G., & Turoń, K. (2018). Car-Sharing Development–Current State and Perspective. *Logistics and Transport, 39*.

Rickenberg, T. A., Gebhardt, A., & Breitner, M. H. (2013). A decision support system for the optimization of car sharing stations.

Seign, R., Schüßler, M., & Bogenberger, K. (2015). Enabling sustainable transportation: The model-based determination of business/operating areas of free-floating carsharing systems. *Research in Transportation Economics, 51*, 104-114.

SSB. (2018). Concept variable: Basic statistical unit. Retrieved from https://www.ssb.no/a/metadata/conceptvariable/vardok/135/en

SSB. (2020a). Befolkning. https://www.ssb.no/statbank/table/04362/

SSB. (2020b). Befolkningens utdanningsnivå. https://www.ssb.no/statbank/table/09434/

SSB. (2020c). Folke- og boligtellingen, husholdninger (opphørt). https://www.ssb.no/statbank/table/10181

SSB. (2020d). Folke- og boligtellingen, sysselsetting og utdanning (opphørt). https://www.ssb.no/statbank/table/10182

SSB. (2020e). Inntekts- og formuestatistikk for husholdninger. https://www.ssb.no/statbank/table/06944/

Stillwater, T., Mokhtarian, P. L., & Shaheen, S. A. (2009). Carsharing and the built environment: Geographic information system–based study of one US operator. *Transportation Research Record, 2110*(1), 27-34.

Ströhle, P., Flath, C. M., & Gärttner, J. (2019). Leveraging customer flexibility for car-sharing fleet optimization. *Transportation Science, 53*(1), 42-61.

Zheng, J., Scott, M., Rodriguez, M., Sierzchula, W., Platz, D., Guo, J. Y., & Adams, T. M. (2009). Carsharing in a university community: Assessing potential demand and distinct market characteristics. *Transportation Research Record, 2110*(1), 18-26.

Zoepf, S. M., & Keith, D. R. (2016). User decision-making and technology choices in the US carsharing market. *Transport Policy, 51*, 150-157.

Appendix

minutes_of_car_use	car_type	average_amount_of_cars	fuel_type	auto_or_not	hjulstrekk	dyrefri_or_not	hengerfeste
347.0268	smabil	7.548495	bensin	0	forhjulstrekk	1	0
121.4595	stasjonsvogn	0	bensin	1	forhjulstrekk	0	0
269.3156	stasjonsvogn	7.697778	bensin	1	forhjulstrekk	0	0
323.6178	varebil	5.973333	diesel	0	forhjulstrekk	0	0
146.2407	stasjonsvogn	1.012346	bensin	0	forhjulstrekk	0	1
85.44253	stasjonsvogn	1	bensin	0	forhjulstrekk	0	1

barnesete	takboyler	takboks	barnepute	population_density	house_without_car	primary_school	high_school
0	0	0	1	1141.682	0.70739	0.1522454	0.2935168
1	0	0	0	548.8394	0.2989952	0.243602	0.398432
1	0	0	0	1141.682	0.70739	0.1522454	0.2935168
0	0	0	0	1038.447	0.5171172	0.1858577	0.2959504
1	0	0	0	593.3089	0.6066958	0.1559856	0.2937876
1	1	0	0	731.7404	0.4246031	0.2810496	0.4137541

university1_4	university_over4	unspecified_education	year0_5	year6_15	year16_19	year20_24	year25_29
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.243082	0.091888	0.022996	0.05726481	0.08936884	0.03412444	0.0509835	0.0703048
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.2902	0.154318	0.07367394	0.08866255	0.1056745	0.03760133	0.07512262	0.1203797
0.3099835	0.1811728	0.05907065	0.04451222	0.05308485	0.0279506	0.2074264	0.1597385
0.208815	0.06379695	0.03258446	0.07243227	0.1023386	0.03870304	0.07516281	0.1042121

year30_49	year50_59	year60_66	year67_69	year70_79	year80	man	woman
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2452461	0.1329986	0.08409433	0.03338685	0.1220932	0.08013463	0.4897309	0.5102691
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2994689	0.09695314	0.05413124	0.024026	0.05209396	0.04588604	0.4744531	0.5255469
0.2545951	0.07743917	0.05231679	0.02006179	0.057234	0.04564068	0.5012488	0.4987512
0.2803326	0.1243237	0.06920721	0.02879319	0.07856201	0.02593254	0.4862234	0.5137766

employed	unemployed	retired	students	other	median_income	winter	spring	summer	fall
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	0	1	0	0
0.4767079	0.008705152	0.267872	0.0519571	0.0481242	563000	1	0	0	1
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	0	1	1	1
0.4692755	0.009066237	0.1819784	0.05804458	0.0872983	418000	0	1	1	1
0.5079472	0.007121444	0.1344084	0.1526385	0.1002874	404669.1	0	1	1	0
0.5166047	0.0101405	0.1799668	0.04073125	0.07778588	501000	0	1	1	0

Plot 1: Overview of first few lines of data set.

car_type <fctr>	mean_minutes_of_car_use <dbl>
9seter	128.1600
minibil	335.3594
smabil	270.6299
stasjonsvogn	176.0646
stor_stasjonsvogn	150.5946
SUV	180.9336
varebil	249.8802

Plot 2: Average car performance for different car types

fuel_type <fctr>	mean_minutes_of_car_use <dbl>
bensin	222.3021
diesel	217.8486
elektrisitet	293.6433

Plot 3: Average car performance for cars with different fuel types

dyrefri_or_not <fctr>	mean_minutes_of_car_use <dbl>
0	222.3361
1	242.4725

Plot 4: Average car performance for cars with or without animals

auto_or_not <fctr>	mean_minutes_of_car_use <dbl>
0	223.9133
1	237.6989

Plot 5: Average car performance for cars with automatic gear shift or manual stick shift

hjulstrekk <fctr>	mean_minutes_of_car_use <dbl>
firehjulstrekk	180.9336
forhjulstrekk	233.7949

Plot 6: Average car performance for cars with 4DW or FDW

barnepute <fctr>	mean_minutes_of_car_use <dbl>
0	190.0900
1	271.8514

Plot 7: Average car performance for cars with or without child cushion

```
            1        2        3        4        5        6        7
 [1,] 8044.968 8004.957 8004.091 6321.976 6338.579 6167.291 6177.132
 [2,] 7217.891 7078.926 7034.184 6688.941 6265.368 5788.296 5461.015
 [3,] 6491.000 6487.638 6403.839 5091.458 5101.071 4791.276 4637.701
 [4,] 4671.623 4596.891 4761.914 4466.240 4436.705 4314.325 4326.761
 [5,] 4879.370 4960.400 4926.472 4099.916 4097.213 3922.877 2004.168
 [6,] 3544.037 3942.733 3669.689 3969.671 4638.714 5138.955 3029.729
 [7,] 5657.006 5886.882 6246.696 5546.661 5576.617 5568.248 5260.455
 [8,] 4608.012 4628.761 4620.859 4621.839 4639.631 4704.778 2434.505
 [9,] 8001.018 7732.587 7701.605 8163.268 7967.937 8058.412 7969.940
[10,] 5898.803 6381.140 6368.433 6038.142 5395.614 5441.133 5658.000
            8        9       10       11       12       13       14
 [1,] 4356.033 4354.208 4326.490 4309.488 4358.212 4608.417 4655.013
 [2,] 3187.195 3364.744 3340.941 3331.127 3435.480 2972.733 2959.712
 [3,] 3950.104 3901.216 3904.574 3763.320 4014.364 3995.250 4007.777
 [4,] 3756.353 3444.181 3336.066 3276.672 3252.961 3182.315 3180.322
 [5,] 2031.849 2050.063 1980.697 2000.542 1930.550 1938.533 1963.310
 [6,] 3347.906 3377.215 3351.488 3096.545 3146.844 3417.874 3425.432
 [7,] 5162.565 4174.094 3467.159 3459.068 3469.646 3351.483 3321.600
 [8,] 2455.496 2450.716 2394.557 2377.953 2373.781 2395.777 2270.402
 [9,] 5566.548 5828.620 5821.986 5826.672 5741.829 5262.346 5360.666
[10,] 5233.134 5197.143 5264.034 4829.960 4916.481 4916.481 4545.917

           15       16       17       18       19       20       21
 [1,] 4780.966 4566.194 4628.716 4796.794 4806.176 4821.789 4856.911
 [2,] 2967.618 3006.195 3038.375 3078.990 3070.219 3052.215 3009.124
 [3,] 3954.498 3845.762 3876.753 3890.751 3877.238 3953.401 4010.969
 [4,] 3286.715 3300.715 3277.246 3371.626 3529.416 3230.636 3248.696
 [5,] 1956.949 1979.092 2006.912 1988.953 1990.222 1898.481 1748.899
 [6,] 3564.509 3573.915 3553.485 3610.984 3563.687 3489.536 3470.125
 [7,] 3308.519 3306.693 3285.775 3334.262 3365.943 3369.674 3681.037
 [8,] 2279.370 2318.891 2397.695 2459.010 2462.313 2493.550 2490.738
 [9,] 5374.235 5318.179 5348.670 5385.236 5408.727 5302.052 5117.806
[10,] 4450.469 4648.858 3014.824 3019.109 2993.201 3024.109 2996.473
           22       23       24       25       26       27       28
 [1,] 5258.702 5287.999 5305.636 5076.515 5095.250 5105.997 5307.568
 [2,] 3033.513 3093.028 3169.817 3169.817 3078.696 3077.116 3207.903
 [3,] 3933.821 3886.442 3835.955 3839.299 3839.299 3807.615 3852.939
 [4,] 3089.614 3198.218 3198.218 3281.716 3305.430 3307.891 3307.891
 [5,] 1748.325 1750.913 1752.296 1772.176 1772.176 1751.555 1754.166
 [6,] 3053.247 3183.547 3191.434 3191.434 3167.153 3284.445 3542.161
 [7,] 3726.201 4020.983 4020.983 4027.841 3984.454 4090.693 4176.873
 [8,] 2841.628 2848.676 2846.883 2846.883 2851.115 2824.738 2845.584
 [9,] 4980.576 5011.901 4972.862 5086.229 5073.122 5073.122 5073.122
[10,] 2996.933 3041.668 2957.865 2954.051 3021.735 3177.920 3096.733
```

```
            29       30       31       32       33       34       35
 [1,] 5307.568 5277.050 5293.752 5234.444 5806.625 5936.374 5988.909
 [2,] 3320.515 3320.515 3331.572 3153.634 3149.802 3237.862 3237.862
 [3,] 3852.939 4018.400 3997.311 3988.904 3981.917 3984.711 3971.856
 [4,] 3421.782 3399.696 3260.306 3148.035 3137.382 3137.382 3154.291
 [5,] 1750.883 1722.594 1701.430 1671.192 1684.077 1719.297 1707.989
 [6,] 3409.189 3418.119 3360.359 3266.397 3266.397 3267.169 3267.169
 [7,] 4181.387 4352.580 4311.437 4352.039 4222.704 4217.808 4238.322
 [8,] 2842.222 2842.222 2851.509 2870.354 2901.274 2901.274 2939.174
 [9,] 5130.841 5140.503 5260.350 5194.214 5125.473 5147.119 5145.333
[10,] 3131.616 3110.096 3418.863 3418.863 3439.660 3252.237 3265.306
            36       37       38       39       40       41
 [1,] 5988.909 5920.770 5743.971 5743.971 5745.631 5756.601
 [2,] 3105.333 3128.017 2918.012 2811.308 2818.863 2720.507
 [3,] 4150.943 4106.341 4109.488 4251.233 4386.161 4400.156
 [4,] 3135.986 3064.750 3359.445 3359.445 3361.184 3348.058
 [5,] 1781.369 1781.673 1806.272 1754.408 1754.408 1856.547
 [6,] 3193.563 3255.570 3400.716 3460.679 3568.739 3567.884
 [7,] 4242.176 4242.176 4237.636 4246.882 4239.990 4461.195
 [8,] 2973.098 2752.610 2750.866 2757.795 2563.665 2482.019
 [9,] 5269.489 5269.114 5257.443 5257.443 5257.443 5141.563
[10,] 3283.824 3318.019 3129.603 3129.603 3121.751 3264.368
```

Plot 8: The 10×41 matrix with MSE for 41 models in each fold in forward stepwise selection

```
       1        2        3        4        5        6        7
5901.373 5970.092 5973.778 5500.811 5445.745 5389.559 4695.941
       8        9       10       11       12       13       14
3904.718 3814.220 3718.799 3627.135 3664.015 3604.121 3569.015
      15       16       17       18       19       20       21
3592.385 3586.449 3442.845 3493.571 3506.714 3463.544 3463.078
      22       23       24       25       26       27       28
3466.256 3532.338 3525.195 3524.596 3518.843 3550.109 3616.494
      29       30       31       32       33       34       35
3634.894 3660.178 3678.689 3629.808 3671.531 3680.123 3691.621
      36       37       38       39       40       41
3712.469 3683.904 3671.345 3677.277 3681.784 3699.890
```

Plot 9: 10-fold cross validation error for each of the 41 models in forward stepwise selection.

bestlam <dbl>	lasso.cv <dbl>
1.8551554	4691.814
1.8441419	3218.991
1.9633345	3934.308
1.9944497	3434.281
2.0400247	1931.102
0.6775981	3819.068
1.6043768	3949.692
2.2768623	2594.864
1.4221953	5164.169
2.6759002	3193.647

Plot 10: Different λ and the corresponding cross-validation MSE in Lasso

```
          [,1]
 [1,] 4694.871
 [2,] 3239.247
 [3,] 3938.722
 [4,] 3436.130
 [5,] 1931.102
 [6,] 3168.706
 [7,] 4050.566
 [8,] 2612.923
 [9,] 5061.858
[10,] 3089.640
```

Plot 11: MSE for each fold in Lasso

```
          [,1]
 [1,] 5590.064
 [2,] 2646.922
 [3,] 4612.401
 [4,] 4231.818
 [5,] 2211.843
 [6,] 3177.821
 [7,] 5176.725
 [8,] 2583.710
 [9,] 6159.379
[10,] 2962.279
```

Plot 12: MSE for each fold in Random forest

```
       [,1]
 [1,] 7393.235
 [2,] 2731.978
 [3,] 4687.180
 [4,] 3544.483
 [5,] 2874.218
 [6,] 6989.104
 [7,] 5472.519
 [8,] 3513.799
 [9,] 6953.280
[10,] 4611.247
```

Plot 13: MSE for each fold in Boosting

minutes_of_car_use	car_type	average_amount_of_cars	fuel_type	auto_or_not	hjulstrekk	dyrefri_or_not	hengerfeste
393.1017	smabil	6.271186	bensin	0	forhjulstrekk	0	0
487.7458	smabil	6.271186	bensin	0	forhjulstrekk	1	0
87.66102	SUV	6.271186	bensin	1	firehjulstrekk	1	1
273.4407	smabil	6.271186	bensin	0	forhjulstrekk	1	0
335.0678	stasjonsvogn	6.271186	bensin	1	forhjulstrekk	0	0
162.1864	stor_stasjonsvogn	6.271186	bensin	1	forhjulstrekk	1	0

barnesete	takboyler	takboks	barnepute	population_density	house_without_car	primary_school	high_school
0	0	0	1	1141.682	0.70739	0.1522454	0.2935168
0	0	0	1	1141.682	0.70739	0.1522454	0.2935168
0	1	0	0	1141.682	0.70739	0.1522454	0.2935168
0	0	0	1	1141.682	0.70739	0.1522454	0.2935168
1	1	0	0	1141.682	0.70739	0.1522454	0.2935168
0	1	0	0	1141.682	0.70739	0.1522454	0.2935168

university1_4	university_over4	unspecified_education	year0_5	year6_15	year16_19	year20_24	year25_29
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098
0.3124605	0.1845351	0.05724225	0.03170461	0.0396502	0.01571556	0.1478827	0.1499098

year30_49	year50_59	year60_66	year67_69	year70_79	year80	man	woman
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605
0.2832296	0.09107267	0.0572665	0.0217288	0.06751965	0.09431993	0.5226395	0.4773605

employed	unemployed	retired	students	other	median_income	winter	spring	summer	fall
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	1	1	0	0
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	1	1	0	0
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	1	1	0	0
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	1	1	0	0
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	1	1	0	0
0.5648262	0.01338487	0.2152139	0.06524088	0.06997941	403000	1	1	0	0

Plot 14: Overview of the first few lines of validation data set