Norwegian School of Economics

Bergen, Spring 2022

Usage Trend Analysis and Forecasting for Ride Sharing: A case of Bildeleringen

An empirical approach using the car-specific data

King Solomon Evbogbai

Supervisor: Julio Cesar Góez

Master thesis, MSc in Economics and Business Administration,

Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Contents

Abstract

Car-sharing is gaining a lot of popularity amongst users, as more people are finding various instances and benefits to use this service. With this development, there is increasing number of companies setting up car-sharing platforms to satisfy this growing demand. As is characteristic of highly competitive industries, the players win market share by effective planning and efficient operations. One aspect of effective planning is ensuring that the car-sharing fleet of cars is suitable to the needs of the target customers. The goal of this paper is to use past data to analyse the car features that are affecting the demand of cars and propose a model to predict the future demand of cars using these features.

To achieve this, we obtained the ride data from Bildeleringen, the leading car sharing operator in Bergen (Norway). We analysed all of the data tables and picked the variables that were essential to our study. After cleaning up the data, we created a new dataset that gave car level information on the car type, car features, the availability period, and the usage variable.

We obtained two measures of usage from the data – time driven and kilometres driven. Based on the business model of Bildeleringen where more of the cost of usage is attributed to the driven time, we chose the time driven as the more appropriate usage measure. Also, we noticed that some cars were available on the platform for way longer than others, hence we went a step further to define the measure of usage as the kilometre driven as a ratio of the time available on the platform.

Using charts, histograms, and box plots, we investigated the possible relation in the car features and the usage of these cars on first glance. We then proceeded to run a multiple linear regression on our data set. We then used 10 data prediction methods to model the car usage and tested the predictive performance of the models using cross validation. The models used belonged to the Linear regression, Ensembles, Decision tree, Bagging and Boosting.

The results of the show that are the car level features that affect the demand are transmission type, wheel drive system, baby pillow availability, child seat installed, and roof box installed. Based on the Mean Squared Error comparison, we also found that the Decision tree is the best model to use for the prediction.

1. Introduction

1.1 Car Sharing

What is car sharing?

Car sharing (also known as the name "car clubs") is growing popular as a new form of car rental and is assuming a promising solution for sustainable transportation (Hartl & Hofmann, 2021). Car sharing can be non-profits, for-profits, or cooperatives, and is defined as providing social and environmental benefits (CarSharing Association, 2011). Car sharing is a reliable, convenient, and flexible alternative to car ownership (CarSharing Association, 2011). Given that many cars are not used to their full capacity, car-sharing communities allow individuals to share access to a car, supporting community transit for the local population at lower transportation costs, including those less able to afford car ownership (Hartl & Hofmann, 2021). On the other hand, the aims of car-sharing to reduce individual car ownership and the number of vehicles driven on the roads ease the burden on the public road infrastructure and improve urban land use and development (CarSharing Association, 2011). The reduction of vehicle miles travelled also reduces greenhouse gas emissions by decreasing dependence on fossil fuels (CarSharing Association, 2011).

How does car sharing work?

Compared to traditional car rental, car sharing is intended for short-time and short-distance journeys as an extension of the transportation network, offering public services to improve mobility alternatives (CarSharing Association, 2011). Some of the largest car-sharing companies in the world include ZipCar in the US, TappCar in Canada, and DiDi Rider in China. Individuals first look for the car-sharing operators in their community and figure out what conditions each operator put on membership. Car-sharing companies provide membership-based services to eligible drivers in the community (CarSharing Association, 2011). Once individuals meet the requirements, members can pick a rate plan and choose and reserve the car from numerous vehicles online or by phone (Millard-Ball, Murray, Ter Schure, & Fox, 2005). Each car-sharing company usually has multiple car parks in the city or community, company's website usually shows individuals a map of their area with the locations of all the reserved parking. Members visit the nearest parking spacing, unlock the cars with their own membership or electronic key card, individuals can then drive off and begin their journey (Millard-Ball, Murray, Ter Schure, & Fox, 2005). Depending on car-sharing companies, members can return vehicles to any car park or within the free-floating operation zone in a one-way car-sharing system, while a two-way system requires members to return vehicles to the location of origin (Di Febbraro, Sacco, & Saeednia, 2012). Once members return the vehicles, that should be all. Car sharing provides members access to vehicles on an hourly basis. Car usage is unrestricted and available at reasonable hourly and/or per mile or kilometre fees that include gasoline, insurance, and maintenance (CarSharing Association, 2011). Car sharing companies provide members with access to a dispersed network of shared automobiles at unattended self-service locations 24 hours a day, 7 days a week (CarSharing Association, 2011).

Impacts of car sharing

Environmental impacts

When evaluating the environmental impacts of car sharing, researchers usually investigate the effects of car sharing on car ownership, car use, and CO2 emissions. CO2 emissions are positively related to car ownership and car use. First, the availability of car-sharing services reduces the number of car ownership. People indicated that they are willing to dispose of their cars because of car-sharing or would have purchased an additional vehicle if car-sharing services are not available (Nijland& van Meerkerk, 2017). On average, a shared car is 39% smaller than a privately-owned vehicle in terms of carbon footprint (Nijland & van Meerkerk, 2017). In terms of car usage, it is usually measured in terms of vehicle kilometers travelled (VKT). Although many studies have shown that car sharing has reduced VKT, some studies have also indicated that car sharing has increased VKT, especially among car-sharing members who do not own a vehicle, who would have previously been travelled using more environmentally friendly forms of transportation, or not at all (Cyriac & Erik Julsrud, 2018; Nijland &van Meerkerk, 2017). By offering access to a car to people who did not previously have such access, car-sharing may increase driving demand and hence increase carbon emissions (Cyriac & Erik Julsrud, 2018). Hence, whether people did own a vehicle before becoming car-sharing members do matter. The net impact on the environment is positive if the car-sharing can offset a large amount of VKT of former vehicle owners and or the increase in VKT by the car-sharing users who got access to a car is lower than what their usage would have been if s/he had owned a vehicle (Cyriac & Erik Julsrud, 2018).

Social impacts

Car sharing is a form of shared mobility, and it provides an additional option to the transportation system in the location in which it operates. It provides individuals with greater mobility and convenience (Millard-Ball, Murray, Ter Schure, & Fox, 2005). First, car-sharing service is usually available 24/7, it provides temporal accessibility to traditional public transportation systems (Marsden, 2022). Second, car-sharing provides people with the potential for the first and last-mile trips that connect to mainline public transportation services or other less accessible locations, car sharing provides greater spatial accessibility and expands the geographical options that can be reached in an acceptable time (Marsden, 2022). Third, car-sharing provides individuals with cost savings by reducing the unit costs of each journey by making better use of assets (Millard-Ball, Murray, Ter Schure, & Fox, 2005). Aside from route time (spatial accessibility) and cost savings, there may be other aspects of journey quality that car-sharing improves, such as increased social advantages from arranging journeys with co-workers or friends and family, as well as emotions of pleasure and well-being (Marsden, 2022). Lastly, car-sharing improves the availability of parking and reduces traffic congestion by reducing vehicle ownership.

Challenges of car-sharing

Car-sharing is a niche product that will likely thrive only in a limited number of most metropolitan areas. For example, in Norway, car-sharing remains geographically focused on the Oslo metropolitan region, with some established cooperatives in Bergen and Trondheim (Cyriac & Erik Julsrud, 2018). If car-sharing would like to contribute to Norway's transition to a more sustainable transportation system, the government, policymakers, and service providers must provide the framework for the practice to expand beyond Oslo, at the very least in other urban regions (Cyriac & Erik Julsrud, 2018). Car sharing, on the other hand, will remain a niche practice for the foreseeable future. In many countries, until recently, car sharing has flourished with little to no support from public authorities, either financially or in terms of regulations, or infrastructure support (Cyriac & Erik Julsrud, 2018). Governments and policymakers should think about methods to give such incentives to car-sharing users and service providers if they want to reduce the usage of private cars and promote the use of car sharing.

In addition, more car-sharing companies are entering the European and North American markets. The application of current and new technology is important to provide great growth prospects for this business (Wagner & Shaheen, 1998). To succeed in the car-sharing market, access should be easy, low cost, simple and straightforward payment, and have many vehicle options (Wagner & Shaheen, 1998). Electronic and wireless technology advancements are assisting in the provision of security, reliable vehicle access, and data tracking (Wagner & Shaheen, 1998). This automation makes sense for a large number of shared-use cars. The rapidly growing market for navigational equipment, such as global positioning systems (GPS), will almost certainly have a substantial influence on the vehicle sharing business and mobility management (Wagner & Shaheen, 1998).

1.2 Car Sharing in Norway

Car sharing has existed in Norway for about two decades. As of late 2018, Norway has 11 car-sharing service providers or platforms that provide over 7,000 vehicles to more than 200,000 registered members. The development of car sharing in Norway can be broken down into 4 periods describing different business models, operational models, and user profiles. Four stages can be characterized as: (1) the development and expansion of cooperative providers; (2) the entry of corporate and multinational players; (3) the arrival and rapid growth of peer-to-peer platforms; and (4) the blurring of boundaries between platforms and types of service.

The first period happened between 1995 and about 2004. Bilkollektivet, the first car-sharing company in Norway was established in 1995 in Oslo, which was largely inspired by the successful establishment of Swiss and German cooperatives that existed for a decade ago. In 1996, similar car-sharing companies were established in Bergen and Trondheim. These car-sharing companies were member-owned cooperatives, demonstrating an example of user-driven innovation in urban mobility. Households seeking access to a vehicle without owning one were the primary customer base for the early Norwegian car-sharing market. Another segment of the car-sharing industry was businesses and government organizations, which utilized car-sharing services as needed rather than owning and maintaining a fleet of designated corporate vehicles. Car sharing stations were mostly located in central regions with a dense enough residential or commercial population to support a successful client base. Each of the three original car-sharing cooperatives had a specific regional focus. This lack of rivalry allowed the cooperatives to work together; for instance, members of any of the organizations had access to vehicles on all three platforms.

The entry of private corporations into the car-sharing industry marked the second period of Norwegian car-sharing history, which lasted from about 2004 to 2014. Besides providing car-sharing services exclusively in highly populated metropolitan regions, car-sharing stations were also frequently found in the parking lots of large shopping malls, particularly those with an Ikea. In 2007, a second for-profit car-sharing company, Move About, was established. Move About was unique in terms of the cars it provided and the consumers it served. Move About was the first Norwegian car-sharing company to provide 100% battery electric vehicles and to focus on serving corporate clients rather than individual households.

The third period, which began in 2015 and is still currently ongoing, was the emergence of peer-to-peer platforms and the extension of geographic coverage beyond major urban regions. The entry of Nabobil and GoMore brought the first large-scale official P2P car-sharing platforms to the Norwegian car-sharing market. With P2P, car-sharing companies do not own and maintain any cars, they act as facilitators rather than vehicle providers. Instead, individual owners are the car providers and they rent out their personal vehicles to others for a short period of time. P2P platforms attract people who want to rent a vehicle for personal use and members who want to earn additional money from the surplus capacity of a capital asset. For the P2P platforms to function, the companies must recruit an adequate and balanced number of member users and providers.

Recently, it has become difficult and unclear to distinguish the boundaries that define previous periods and the numerous types of car-sharing services available in Norway. Platforms have started embracing hybrid business models that include elements of the peer-to-peer, business-to-business, business-to-consumer, and cooperative models. There are also indications that car-sharing is becoming more closely associated with residential organizations e.g., OBOS, the largest housing cooperative in Norway. In addition, in late 2018, Norway's first free-floating car-sharing system was introduced.

1.3 Bildeleringen

Bildeleringen is a car-sharing company operating in Bergen, Norway. The company was established in 1996 and the goal of Bildeleringen is not to make a profit but to make car-sharing accessible and to promote a smarter, more solidarity, and environmentally friendly way of using a car ("Bildeleringen: Norges Smarteste Bilkollektiv," 2020). It is operated as a non-commercial collaboration and is 100% member financed. Members are the owners; all profits go directly back to the members. Bildeleringen currently provides car-sharing services to 4,000 private and business members in Bergen ("Bildeleringen: Norges Smarteste Bilkollektiv," 2020). It has over 350 vehicles located in over 100 places in Bergen ("Bildeleringen: Norges Smarteste Bilkollektiv," 2020).

To rent a car with Bildeleringen, a person must be over 18 or above with a valid driver's license, and you have to first register as a member on Bildeleringen's website. When you sign up as a member, you can choose to either sign up as a private or a company. The membership fee for private is 600 kr per half-year, while the membership for a company is 1800 kr per half-year. Both types of membership require 7000 kr of deposit, which is refundable upon withdrawal. A private membership provides the opportunity to register two users and order two cars at the same time; a corporate membership has no restrictions on the number of users or concurrent reservations. Once you have registered for membership, you can find where the cars are located and reserve a car through the app, called Dele. Bildeleringen is operated in a two-way system, which means customers must rent and return the vehicle in the same location.

When renting, you pay an hourly rate of up to 12 hours, price per kilometre, and a start-up fee of 49 kr. The price for use of the vehicles depends on the type of car mileage. There are 16 types of vehicles on the price list, all charged 2.30 kr per km up to 150 km and 1.30 kr per km after 150 km. The hourly rates of renting vary between types of vehicles. Expenses for car passes and parking fees are invoiced in addition. Vehicles can be categorized into 7 types: small car, station wagon, SUV, mini car, large station wagon, van, and 8/9 seats. There are 3 fuel types: gasoline, diesel, and electric. Some cars have automatic gear shifts, while some are manual cars.

1.4 Problem Description

Supply chain planning is very important for car-sharing companies in ensuring that they always have the right cars in the right places at the right amount. In this study, we focus on the first part of the supply chain problem, which is how the car sharing companies can identify how car-specific features affect the car-usage.

1.5 Research Goals

In this study, we aim to identify the car specific properties that affect car usage in car sharing, formulate models that uses these features as the regressors (prediction variables) to predict the micro-level demand on a car level and test these models using data from Bildeleringen. We will then choose which of the models based on model indicators is the best to predict the micro-level demand for Bildeleringen.

2. Literature Review

2.1 Drivers of Demand

A review of related literature suggests that there are several factors affecting the demand for car sharing.

Geographic factors

Car-sharing is a supplement to other public transportation options. It is possible only as part of a wide transportation package in areas where public transportation, walking, and cycling are all feasible possibilities (Millard-Ball, Murray, Ter Schure, & Fox, 2005). Car sharing services are mainly concentrated in metropolitan areas (Millard-Ball, Murray, Ter Schure, & Fox, 2005). A highly dense, good pedestrian environment, and a mix of uses and parking pressures all contribute to the success of car-sharing (Millard-Ball, Murray, Ter Schure, & Fox, 2005). The demand for car-sharing increases because of overcrowded public transportation and increased traffic congestion in urban areas (Ankita Bhutani & Pallavi Bhardwaj, 2018). The ability to live without a vehicle, or with just one vehicle, is also important. Low vehicle ownership rates are a strong indicator of having a strong car-sharing market (Millard-Ball, Murray, Ter Schure, & Fox, 2005). Many car-sharing companies first focused on the residential market. Some have discovered, however, that business users are the primary source of growth (Millard-Ball, Murray, Ter Schure, & Fox, 2005). University campuses are also a valuable market niche (Millard-Ball, Murray, Ter Schure, & Fox, 2005).

Demographics factors

Demographic factors measure a population's age, sex, race, income level, education level, etc. One study suggests that young individuals, primarily between the ages of 25 and 45, appear to be the driving force behind car sharing (Chun, Matsumoto, Tahara, Chinen, & Endo, 2019). Car sharing is more likely to appeal to young individuals because of their attitude and preference for less auto-oriented transportation (Chun, Matsumoto, Tahara, Chinen, & Endo, 2019).

Car-sharing users are characterized by their high level of education (Millard-Ball, Murray, Ter Schure, & Fox, 2005). On the other hand, one study suggests that education level has little bearing on the demand for car sharing. The relationships that many car-sharing schemes have created with schools and universities are likely to have an influence on the trend that car share members are more educated in North America and other developed countries (Chun, Matsumoto, Tahara, Chinen, & Endo, 2019). Individuals with post-graduation education levels or higher in developing countries like India are shown to be more ready to own a vehicle, thus, it is difficult to conclude that people in developing countries with higher education levels will prefer to share a car rather than purchase one (Chun, Matsumoto, Tahara, Chinen, & Endo, 2019).

It is unclear if car sharing is more appealing to certain income level groups. To some extent, car sharing is more common among individuals with lower and middle incomes (Efthymiou, Antoniou, & Waddell, 2013). However, some studies suggest that car-sharing members tend to have middle- to higher incomes (Clewlow, 2016). It is likely that different income levels have different motives regarding car-sharing (Burkhardt & Millard-Ball, 2006).

Psychological factors

The non-observable variables such as psychological factors give further insights into factors affecting the demand for car sharing. Recent studies have stressed the importance of psychological factors such as perceptions, attitudes, social norms, culture, lifestyle, or habits on the decision of car ownership and choice of transport mode (Fujii & Kitamura, 2003). Almost all members are concerned about environmental and social issues and are more interested in the functionality of a vehicle than its appearance or brand (Millard-Ball, Murray, Ter Schure, & Fox, 2005). One's pursuit of convenience, value-seeking, and symbolic lifestyle can impact one's demand for car-sharing (Schaefers, 2013). The demand for car sharing is also positively related to one's risk perceptions on car ownership, leading to a choice to minimize car ownership (Schaefers, Lawson, & Kukar-Kinney, 2015).

Technology

Technological advancements in car sharing such as the growing popularity of digital car keys enable renters to have access to their vehicles directly through the car-sharing app. When picking up the car, renters can simply use their smartphones to unlock the doors (Ankita Bhutani & Pallavi Bhardwaj, 2018).

Politics

Stringent government regulations to control vehicular emissions and government incentives for the adoption of car sharing drive the demand for car sharing (Ankita Bhutani & Pallavi Bhardwaj, 2018).

Trip purposes/intentions

Car-sharing is used for a variety of trips; however, it is rarely used for the everyday commute to work (Millard-Ball, Murray, Ter Schure, & Fox, 2005). Members can take public transportation, bike, or walk for most of their daily journeys, but they still have access to a car when needed. Members often use the car-sharing services when they have items to transport, require a car to get to their destination, or have multiple stops to make (Millard-Ball, Murray, Ter Schure, & Fox, 2005).

2.2 Demand Prediction

Demand forecasting is a very useful tool for companies in designing their operations. This method is used extensively in all industries and is particularly useful in transportation. It has been used to address problems in car sharing, bike sharing, ride hailing and automobile sales. Forecasting the demand for vehicles based on different micro and macro factors help companies plan their supply chain, station location and fleet distribution. This consequently leads to operational efficiencies and profit maximization. Hence, there has been a lot of work done by researchers on this subject.

We have reviewed existing literature across Car Sharing based research (which has the same base as this study) as well as literature in related applications. These related applications include Ride hailing, Bike Sharing and Automobile sales.

GOAL

The goals of the papers have generally inclined towards predicting service level demand of car sharing operators and services, with a goal of efficiently implementing vehicle relocation and station placement. These papers have set out to predict to analyse the different factors that affect the demand for Car Sharing services on a service or station level. The research reviewed has been primarily based on Station Based Car Sharing (SBS) and Free-Floating Car Sharing systems.

Chiara et al. (2016) aims to model the station-based demand for cars for an electric car sharing operator operating a station-based model. This model developed using system data is then used to show usage patterns and furthermore propose a classifier for determining the potentials for profitability of the various stations. Yu et al. (2020) set out to detect the operation status of ride sharing systems and create a short term forecast on the usage trend of vehicles in a station-based model. A Long Short-Term Memory approach was employed, with the effectiveness tested with real world data.

Wang et al. (2019) aims to identify causal variables for Car Sharing demand at an operational level using the Granger Causal Test. Furthermore, this analysis is then used to create a micro demand forecasting model using the Long Short-Term Approach for the one-way (Free Floating Car Sharing) model. Schmöller et al. (2014) aims to detect the influence of various factors (spatial and temporal data) on the demand for a Free-Floating Car Sharing model. The study also aims to identify interdependencies that could exist through analysing the parameters and identify demand cold spots and hot spots.

METHOD

Chiara et al. (2016) defines car sharing systems on the lines of station capacity and station utilization. Using 30 days of Data from a car sharing operator in France, this paper studies the spatial and temporal patterns of the car sharing service; modelling the pickup and drop-off rates of cars in its 960 stations. The stations are also clustered according to usage(demand) trends. The paper also proposed a classifier to represent the potential profitability of the stations using a heuristic model. This model uses the pick-up rate and availability of cars as the factors for this classifier. The scope of this paper is very relevant to our study as it is based on a Station Based Car Sharing Model. Although this study focuses on station level demand (as opposed to the car level demand which our study explores), the choice of demand factors is very relevant to our study.

Yu et al. (2020) uses a deep learning approach to predict the demand for car sharing in a Station Based Car Sharing Model. The Long Short-Term Model is used by the authors here to forecast the demand for car sharing on the hypothesis that this kind of problem could not be solved by a feedforward network using fixed-size time windows. The Discrete event model is used to highlight vehicle usage behaviours and consequently analyse the results of the predictions from the Long-Short Term Memory model. The study used temporal features such as Pickup and drop off time, arbitrary time, and weather as the time series variables. These are used to predict the future vehicle pickup and drop-offs. The prediction accuracy is measured by RMSE (Root-mean-square-error).

Wang et al. (2019) identifies 11 raw indicators that can be used to predict the station level demand for various stations. The study also employs the use of the Granger causality test and the Long Short-Term Model. The Granger causality test is used to test the indicators such as ‘Driving distance’, ‘Weekday’, ‘Daily highest temperature’, ‘Daily lowest temperature’,

and ‘Fee per kilometre’. The effect mechanism for the indicators is then analysed using the impulse function while variance decomposition is used to measure the contribution of each indicator to the model error. The prediction model used in this study is the Long Short-Term Memory model.

Schmöller et al. (2014) focuses on analysing the difference in demand trends for HCS (Hybrid Car Sharing) and FFCS (Free Floating Car Sharing). It attempts to link the spatial and temporal instances of maximum and minimum demand data to the usage trends and external factors such as Climate and Socio-demographic factors. To achieve this, a temporal analysis of the two types of car sharing is carried out first. The aim here is to identify the peaks of demand across the two types of car sharing and compare them for insights. Same thing is done using spatial analysis using Kernel Density, to identify the geographical trend of the demand for the car sharing types.

FINDINGS

Chiara et al. (2016) detects a patterns of station utilisation where there are stations that attract cars more in the morning and others that attract cars in the evening. This trend is then successfully linked to the type of area where the station is located (residential vs business). A classifier is also used to identify stations that are profitable or not.

Yu et al. (2020) found out that the Long Short-Term Memory prediction model worked best in prediction the short-term car sharing demand as compared to Artificial Neural Network and Convolutional Neural Network. LSTM had the lowest values for Root Mean Square Error, R-Square and Mean Absolute Error. The report also shows that there are high peaks of demand in the Afternoon and very low peaks in the early mornings and evenings. It is highlighted that these trends are like the general transportation trends in the city of Chengdu.

Wang et al. (2019) claims that the long short-term network model produces the most accurate prediction as compared to other methods with deviation rate of 57.5%. It also reports that ‘Car pick-up interval’, ‘Trip time’, ‘Workday or not’, ‘Daily weather condition’, ‘Fee per minute’, and ‘Order volume’ are significant predictors of car usage while ‘Driving distance’, ‘Weekday’, ‘Daily highest temperature’, ‘Daily lowest temperature’ and ‘Fee per kilometre’ are not relevant predictors in the demand forecast model.

Schmöller et al. (2014) finds that there is a big difference in the usage trends of HCS and FFCS and low correlation between usage and external factors. The study purports that Hybrid Car Sharing is used for longer trips compared to the Free-Floating Car Sharing. Also, FFCS is used more frequently for night trips and the paper hypothesizes that this may be due to users not wanting to return cars to stations at night. Spatial analysis reveals that FFCS is used more for shorter trips to the city centres and business or shopping districts while HCS is used for longer round trips. External Factors such as weather and presence of students in the area have a low correlation coefficient (less than 0.1).

3. Data Description

3.1 Description of Available Data

The data used in this analysis was sourced from the Bildeleringen AS for their operations in Norway. This operation data is valid from 12th January 2019 till 21st April 2020. This data spans over a year and gives specific insight into the operations of Bildeleringen across trips, cars, and other business components. The data shows the relationship between Bildeleringen 256 cars across 86 parking spots used in 86,000+ trips.

The database contains the tables listed below:

i. Assignment: This shows the assignment task of cars to parking locations. This shows the history of cars locations across the XX parking spots in Bergen. This gives some insight into the relocation efforts of Bildeleringen in response to geographical demand trends which were not covered in this study but also essential to demand analysis.
ii. bildeleringen_price_model: This shows the different pricing models used by Bildeleringen for different cars. Bildeleringen as expected, charges differently across different cars depending on their features. This will be further examined in this study as this also impacts the focus of this study.
iii. car_category: This shows the categorization of cars on the Bildeleringen car network. This wide-level categorization is done with the size as the differentiating factor.
iv. car_property_group: This groups the car properties based on the common factor. This helps to organize the features into a smaller set of easy to classify groups. As seen in the table below, the FUEL_TYPE group refers to the variables that define the fuelling mode of the cars.
v. car_property: This shows the full list of car property variables and the assignment of these properties to the respective property group. This study is hinged on the car properties contained in this table.
vi. car_to_property_mapping: This shows all the cars under the Bildeleringen network and the properties that they possess. This table will be essential for defining the features composition of the cars and highlighting the relationship between these features and the demand.
vii. car: This shows car level information with the most granular detail level. It shows the operational details of all the cars as it relates to availability, servicing, location, and

damages. The availability aspects of this data are used in this study, and the effects of the servicing status, odometer reading, and damages are not considered in the scope of this study. However, these aspects should be investigated in future studies of this topic.

viii. driver: This shows the assignment of drivers on each reservation (all trips started and not started) to the persons (customers). This data could have been useful in analysing the demand trends as well, but the *persons* data does not contain any demographic or geographic data. This can form the basis for future studies to show the relationship between the demographic features of the users and the car usage trends. Also, in a FFCS system, we can use this data to analyse the urban-rural transportation trends of car sharing systems.

ix. invoice_line: This shows a list of the invoices and the calculation components of these invoices. It also gives information about the related cars to each invoice.

x. invoice: This shows the invoice status for the generated invoices. It shows the amount of the invoice, the status (paid or unpaid), the dates (invoicing, due and payment dates) as well as the recipient data. In studying payment trends in car sharing systems, this table will be relevant, however this is outside the focus of the study. Hence it will not be used.

xi. location: This gives detailed geographical description of the location of the cars. This helps to determine the geographical effects of the demand for cars. This is a heavily researched area and not considered in this study.

xii. membership_fee: This shows the payment information for membership plans over time. This is done by linking the membership ID and invoiceline ID.

xiii. membership_type: This shows the different types of memberships offered by Bildeleringen to their customers.

xiv. membership: This shows the details on the membership registrations on the Bildeleringen platform. It shows all the membership transactions that are active or cancelled. It shows details on the start and end date as well as the related person_id of the owner. Along with the membership_fee and membership_type information, this data can be used to analyse the effect of subscription on the usage of cars and demand trend of users. This analysis can be used to influence strategy in terms of membership offerings, pricing, and marketing.

xv. Model: This shows the model level details for cars. Model level data such as the car category and number of seats are highlighted here.

xvi. person_membership: This shows a list of all the users(person) on the Bildeleringen platform and an analysis of whether or not they are subscribed to a membership.

xvii. person: This shows a list of all the users on the Bildeleringen platform and their basic information. In other to align with data security restrictions, personal information such email, phone number, social security number and address have either been masked or deleted. This table also notably lacks the sex and age demographic data. These variables could have been useful in analyse the personal demand trends.

xviii. product: This shows the extra billing scenarios on Bildeleringen. These are triggered on special scenarios such as parking, kilometres driven and tolls. This is irrelevant to the scope of this study but can be used to analyse the revenue generation components and their potentials for Bildeleringen.

xix. reservation_message: This shows the reservation messages sent to customers upon reservation. This is also irrelevant o this study but can be used to analyse reservation cancelation requests in subsequent studies.

xx. Reservation: This shows all the reservation on the Bildeleringen system over the analysed period. It shows the timestamps, the users who booked the reservation and their membership status. Furthermore, this also shows the invoice status and the car damage status at the end of the reservation. This table contains the reservations that resulted in trips and those that did not. Hence, there is a lot of noise here.

xxi. Trip: This table forms the basis for our analysis as it shows all of the car usage in terms of kilometres driven and time usage (these two variables are our primary measure of measuring usage). This information is presented on a per trip level and will be aggregated on a car level to repurpose this data for our analysis.

To summarize, the car_category, car_property_group, car_property, car_to_property_mapping, car, and trip data will be used in this study to analyse the effect of car-specific data on the usage trend of these cars. Assignment, bildeleringen_price_model, driver, invoice_line, invoice, location, membership_fee, membership, Model, person_membership, person, product, reservation_message, and reservation tables are relevant outside the scope of this study and can be used to understand the demand based on a different cluster of factors. Some tables were also not reviewable as they contained no data or incomplete data. Hence, they were not reviewed in this study.

3.2 Derived Data

The goal of this study is to firstly understand the usage (also referred to as the demand) of cars based on the car specific characteristics. The analysis is then used as a basis to predict the car level usage for the cars in the Bildeleringen network. It is hence important to define what the measure of usage will be.

In previous literature, the demand (usage) has been represented either by the time dimension measured in minutes or hours or by the distance driven dimension measured in kilometres or miles. These two approaches highlight different components of the demand and can be applied differently based on the business models of the car sharing systems. More companies charge more for time expended as compared to the distance driven by the car. For these companies the usage will be more focused on the time dimension and vice versa.

$$Usage\_minutes = \sum_{t \in T} endts(t) - startts(t) \; \forall c \in C$$

Equation 1 : Usage Minutes

$$Usage\_km = \sum_{t \in T} km\_driven(t) \; \forall c \in C$$

Equation 2 : Usage Kilometres

where T is a set all Trips and C is a set of all the cars

In this study, we will go one step forward to contextualize the usage as a factor of car availability. Hence, we take into the account the time a car was available for when calculating the usage measure. We will introduce two new variables.

a. *Available_days:* This is the amount of time for which the car was available to be used during the analysed period. It is measured in days and derived by summing up the availability periods as retrieved from the ***assignment*** table.

$$Available\_days = \sum_{a \in A} endts(a) - startts(a)\ \forall c \in C$$

where the A is a set of car assignments and C is a set of Cars.

Equation 3 : Available Days

It is however mentioned worthy that the period of availability period as defined in this study should not be mixed up with the unreserved period of the car. While the unreserved period of the car refers to the period by which a car is.

i. parked at the station
ii. not reserved by a user
iii. and available to be booked

the available_days is a more wholistic figure that indicates

i. how long a car has been on the Bildeleringen platform
ii. includes unreserved period and reserved period.

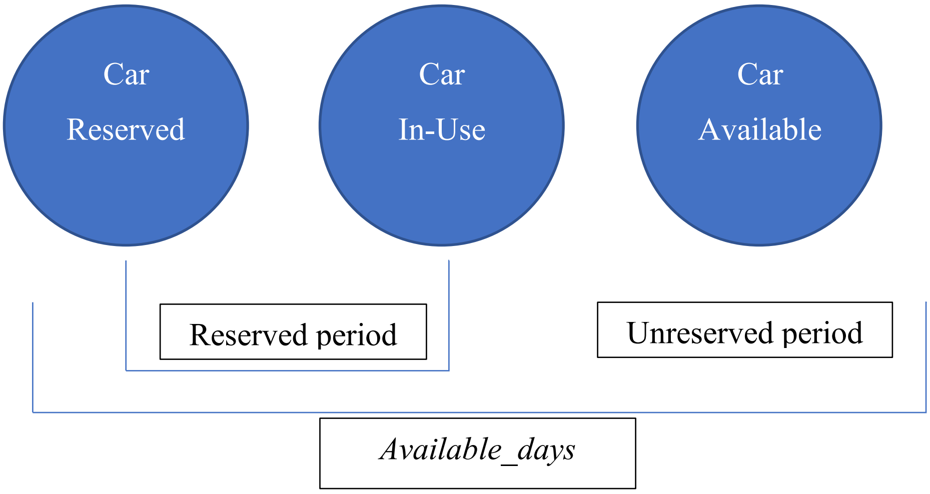

Figure 1 : Car State - Reserved vs Unreserved

b. Demand ratio: This is the usage density across its across availability days. This shows how much a car is used as a function of how long it is (was) available on the platform. This helps to make a fair comparison across cars that were recently added to the fleet and cars that were taken off the fleet in the past.
This is calculated by taking the current usage indicator (time or distance) as a ratio of the availability period (available_days) on the platform. This is explained below.

$$Demand\_ratio = \frac{Usage\_minutes}{availablity\ days}$$

$$Demand\_ratio = \frac{\sum_{t \in T} endts(t) - startts(t)\ \forall c \in C}{\sum_{a \in A} endts(a) - startts(a)\ \forall c \in C}$$

$$Demand\_ratio = \frac{Usage\_km}{availablity\ days}$$

$$Demand\_ratio = \frac{\sum_{t \in T} km\_driven(t)\ \forall c \in C}{\sum_{a \in A} endts(a) - startts(a)\ \forall c \in C}$$

where T is a set all Trips

C is a set of all the cars

and A is a set of car assignments

Equation 4 : Demand Ratio

3.3 Analysis Dataset

To perform the car usage analysis in this report, we have created a new dataset by combining variables from the tables in the Bildeleringen database and the derived variables. The variables in the new dataset are explored below.

i. Car_id: This is the unique identifier for each car in the car fleet. This is used to index cars as the analysis done in this study is car-level. For this study, 256 active cars are used in the demand analysis and demand prediction.

ii. Car_category: This is the definition of the high-level category to which a car belongs with size as the key differentiator. This is a critical part of car level demand as the suggests what people are using the cars for – from small cars for single person travel to mini vans for moving large items around.

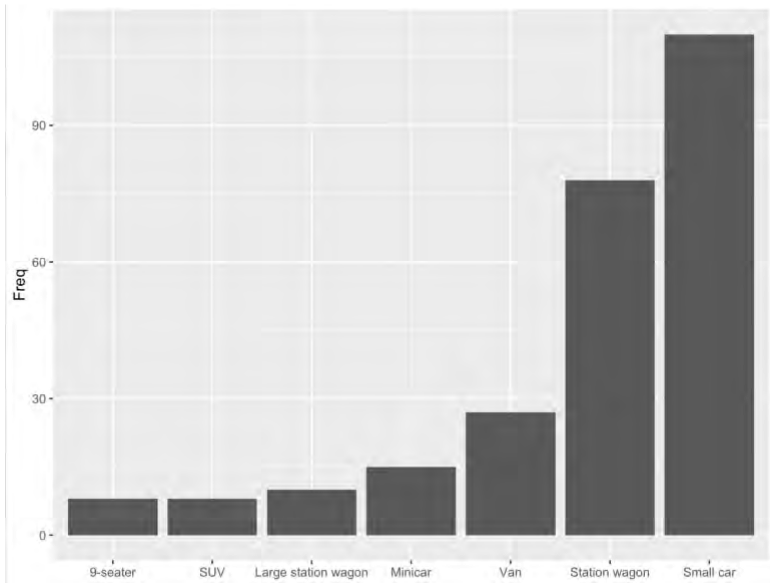

Figure 2 : Car Type Frequency Distribution

Majority of the cars on the Bildeleringen network are small cars. This type of vehicles constitutes 43% of the fleet, followed by station wagons at 30%. Vans, minicars, large station wagons, SUVs and 9-Seaters make up the rest of the fleet with 27%.

iii. Number of Seats: This is similar to the car category however it focuses only on the number of seats as the only differentiator. For example, a minivan is a three-seater vehicle, but the use case of the minivan will not be captured in this data irrespective of the fact that the minivan serves a different purpose from a minicar for example.

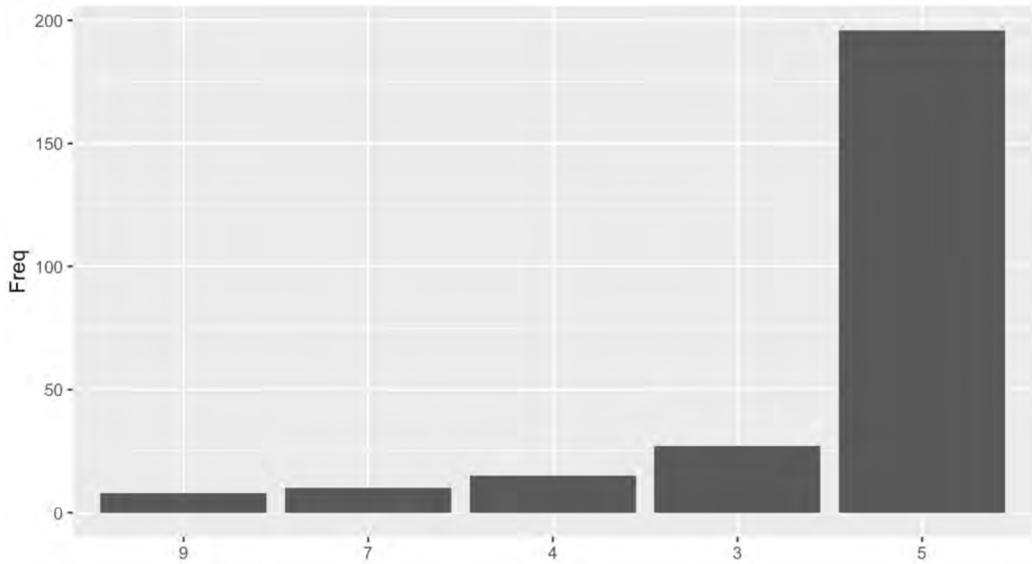

Figure 3: Number of Seats Frequency Distribution

77% of cars on the Bildeleringen fleet are 5-Seater vehicles. This is explainable by the most popular car types being small cars and station wagons that are 5-seater cars.

iv. Fuel_type: This defines the power type of the automobiles in the fleet. The cars in this fleet are of three types – Electric, Diesel and Gasoline. This is an important indicator in identifying if there are increasing customer preferences towards more eco-friendly options available.

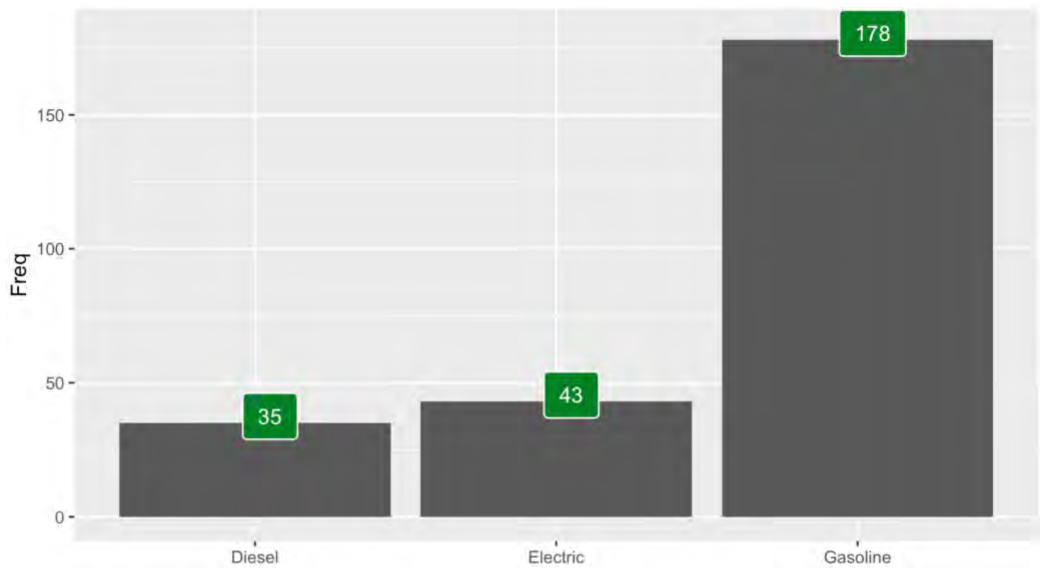

Figure 4: Power Type Frequency Distribution

Bildeleringen's fleet is majorly Gasoline-powered cars. 69% of the cars are gasoline powered with only 17% being electric.

v. Transmission: This is the gearbox transmission mode of the vehicle. This can either by manual or automatic. This can give insight to whether or not customers are using any of these types of cars based on their ability to use them. If this is the case, this can show the preference of these two choices.

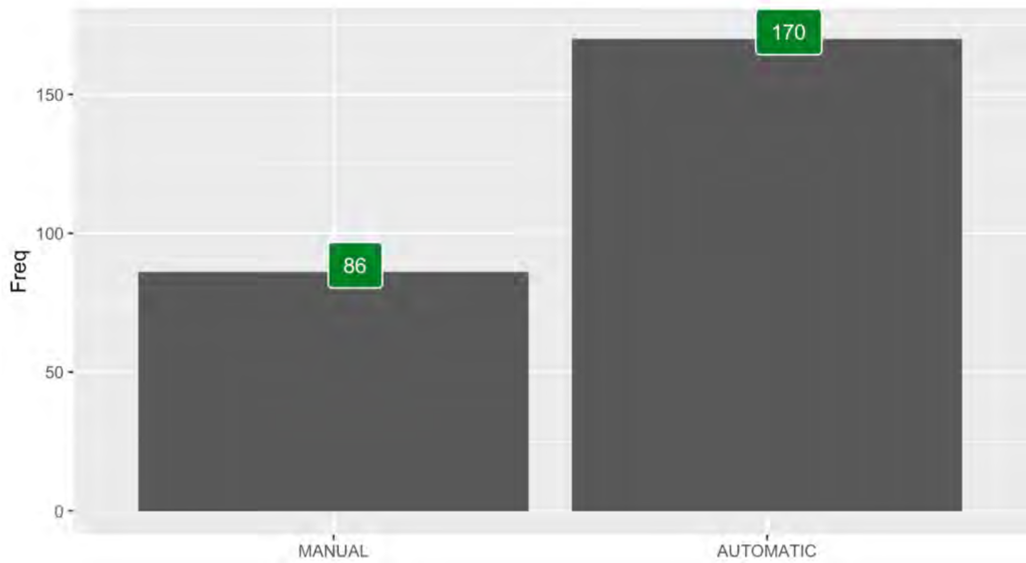

Figure 5: Transmission Type Frequency Distribution

The fleet is made up of 170 cars with Automatic transition. At 66%, this makes automatic transmission the dominant type of transmission of vehicles hosted by Bildeleringen on their platform.

vi. Wheel_Drive: This is a measure of the number of wheels in the drive system of a vehicle where the engine power is transmitted to. This can be Front, Rear or Four wheeled drive systems. In the Front wheel drive system, only the wheels in the front are the recipients of the power while in the Rear wheel drive, the two wheels behind receive the engine power. In the case of the four-wheel drive, all wheels receive power.

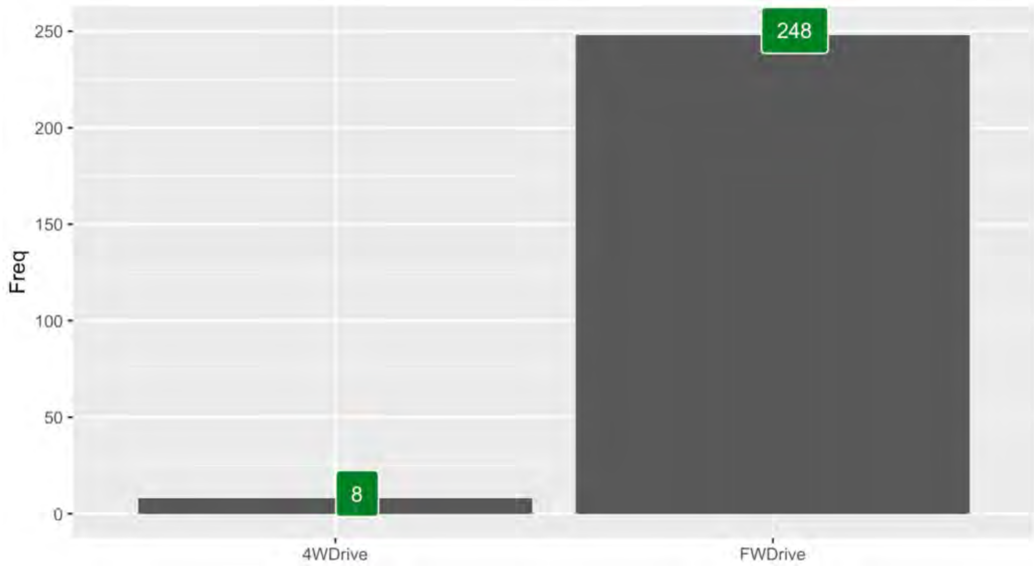

Figure 6: Wheel Drive Frequency Distribution

Bildeleringen's fleet consists of majorly Front-wheel-drive vehicles with just 3% being Four-wheel-drive. Notably, Bildeleringen has no Rear-wheel-drive vehicles.

vii. Animals_allowed: This indicates whether animals are allowed in the car or not. This can be important if a lot of customers are travelling with their animals or there is a general want for cars that allow animal travel.

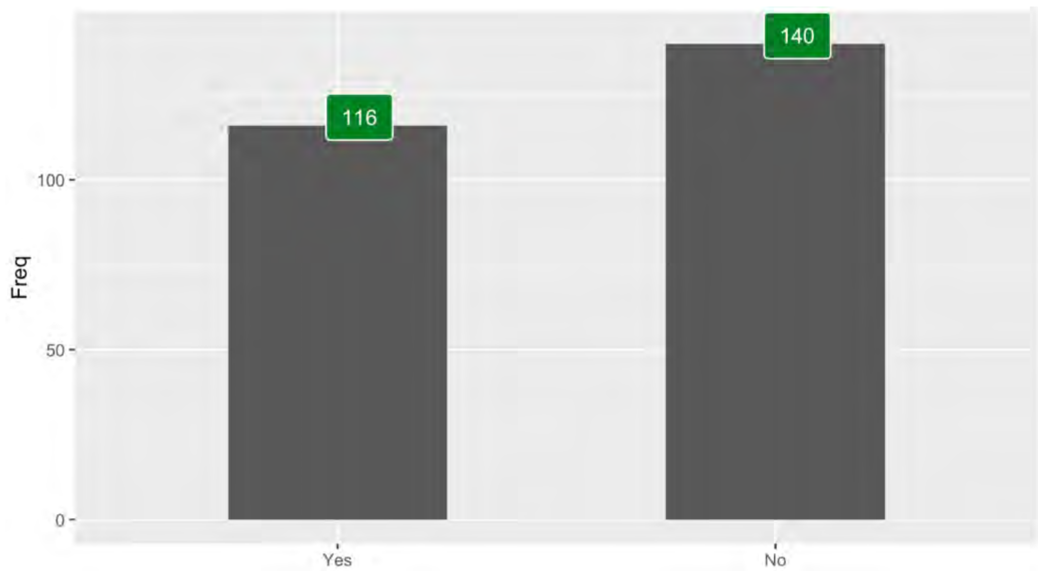

Figure 7: Animals Allowed Frequency Distribution

The vehicles on Bildeleringen's network have slightly more cars that do not allow animals on them (55%).

viii. Towing_hitch: This is an indicator of whether a vehicle has a towing hitch. The towing hitch is usually present at the rear of vehicles and is used for towing attachable objects.

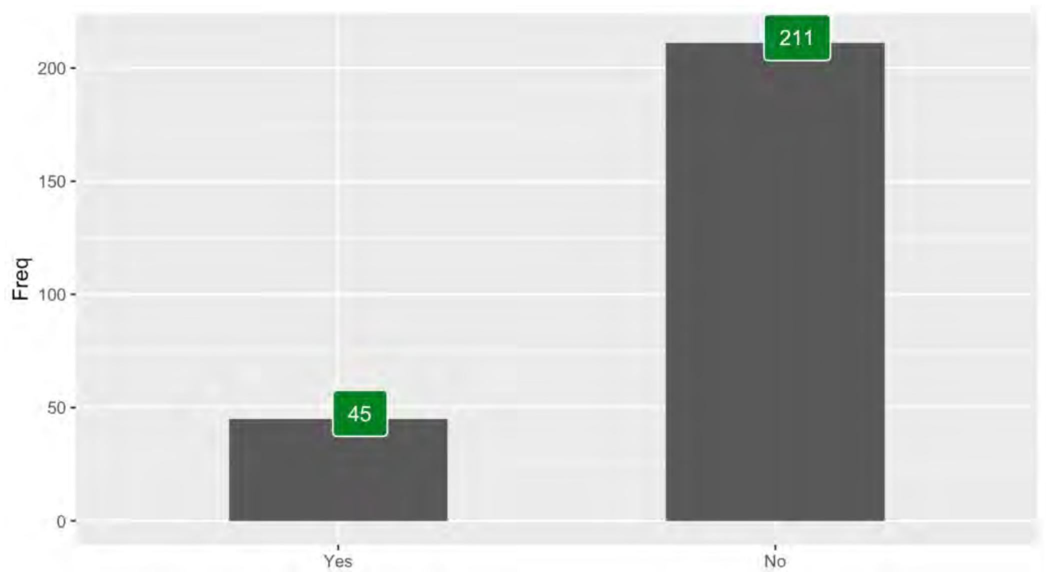

Figure 8: Towing hitch frequency distribution

Cars with Towing hitches are not very popular on the Bildeleringen network. As reflected in the bar chart above, Bildeleringen has just 18% of their fleet having towing hitches.

ix. Child_seat: This is an indicator of whether the vehicle has a child seat or not. This child seat is rated from 0-18kg.

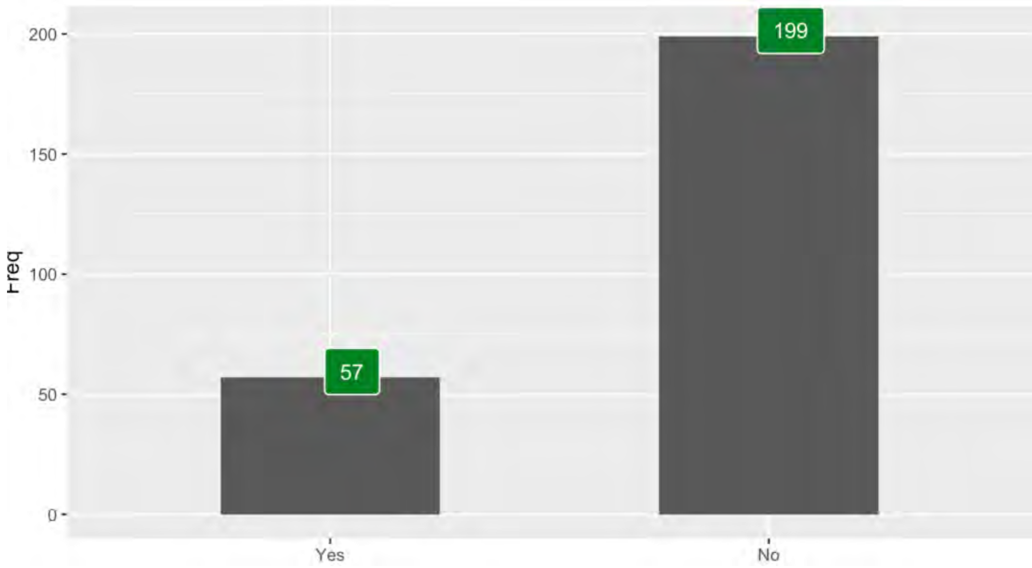

Figure 9: Child seat frequency distribution

Just 22% of cars on Bildeleringen's network have the child seat installed in them.

x. Roof_racks: This is an indicator of if the car is fitted with the racks for attaching items to the top of the car. Large items that cannot fit into the car such as bikes, sport equipment's and camping gear are usually attached to vehicles using these roof racks.

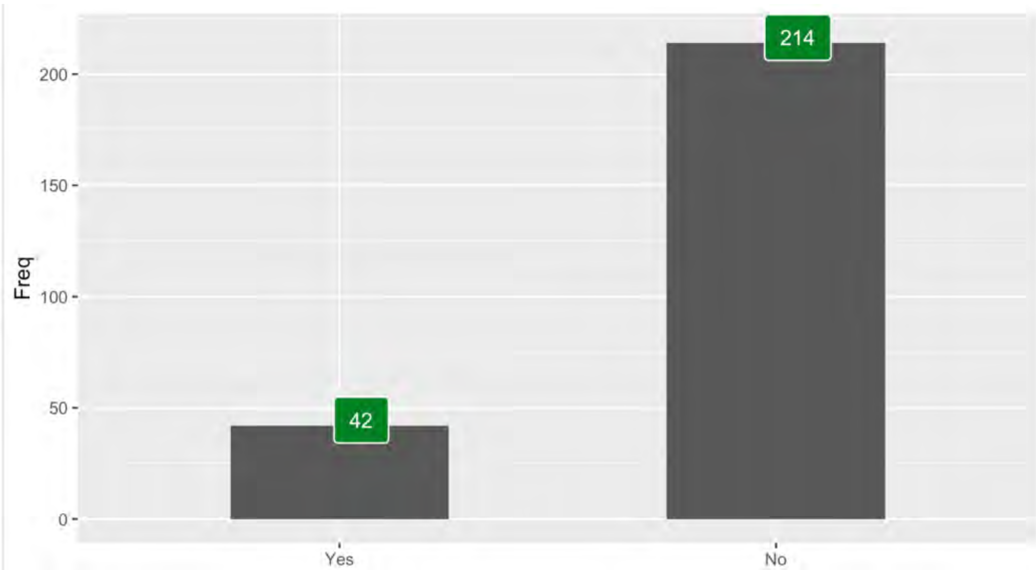

Figure 10: Roof rack frequency distribution

Roof racks are not popular in Bildeleringen's fleet of vehicles as just 42 of the cars (16%) have these roof racks.

xi. Roof_box : This is an indicator for whether or not the vehicle is fitted with roof boxes. These roof boxes are used for storing extra personal effects. This is particularly useful when travelling in a group or transporting a lot of personal effects.

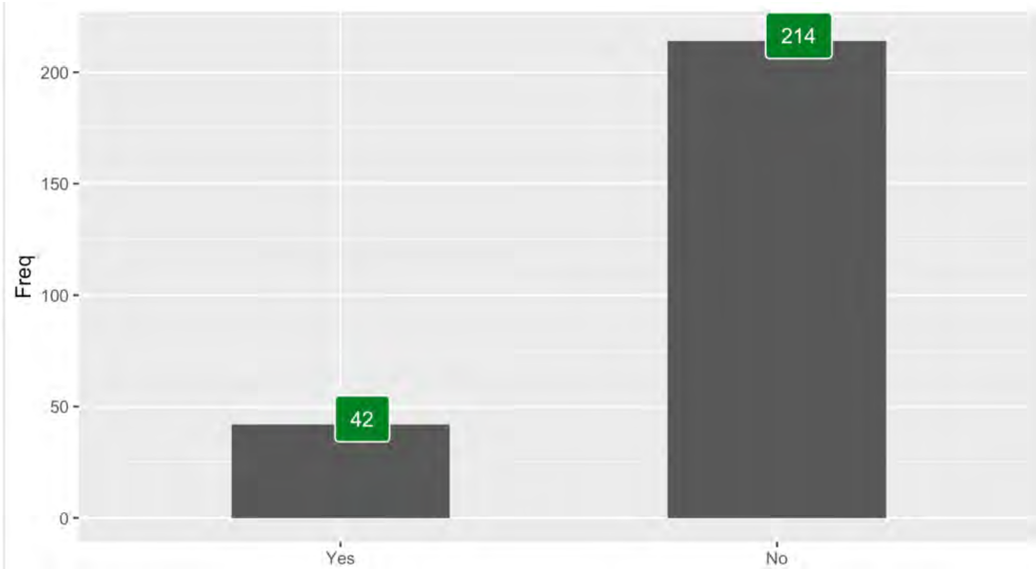

Figure 11: Roof box frequency distribution

The ratio of cars with Roof boxes is the same as the previous variable, Roof_racks.

xii. Baby_pillow: This is an indicator for whether or not the car has a baby pillow which is used for babies during travel to prevent excessive head motion.

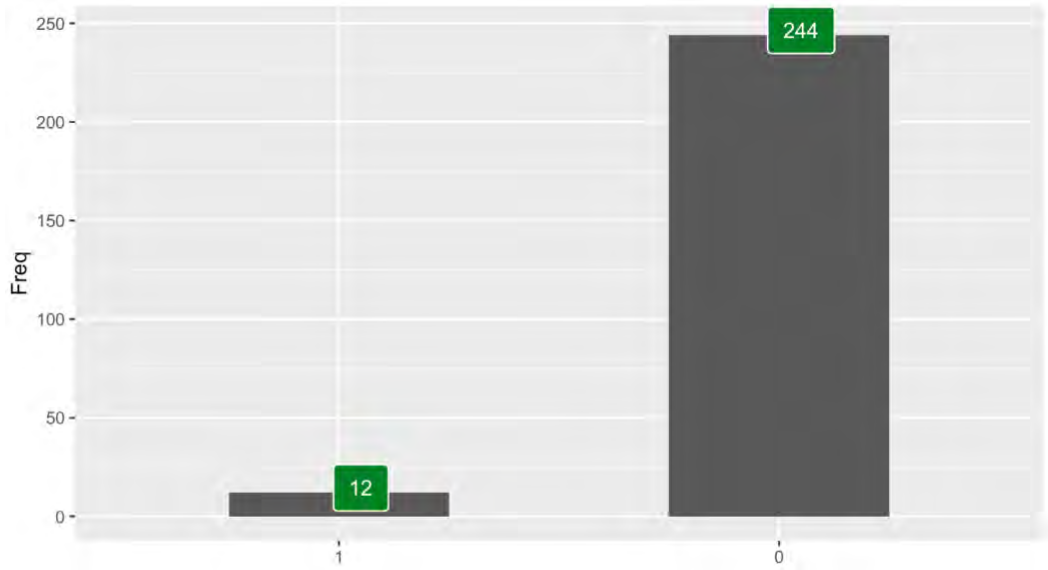

Figure 12: Baby pillow frequency distribution

This feature is not popular in the Bildeleringen fleet as just 12 out of the 256 analysed cars have a baby pillow provided in them.

xiii. Availability_Duration_Days: As defined previously, this is the total period of time for which the car has been available on the platform for booking. As the data provided in the assignment data provides the assignment start date and assignment end date which might be outside the analysis period, time ranges outside this defined period are coerced into the time range.
For example, the car with car_id 83 was onboarded on the platform on 08/03/2017, however for calculating the availability duration we use 12/02/2019 which is the start date of this analysis.

```
> summary(AvailData)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   35.0   180.0   399.0   321.3   434.0   434.0
```

Table 1: Summary of car availability

This data is quality checked by the above summary where the max is 434 days – which is the duration of the analysis period for this study.

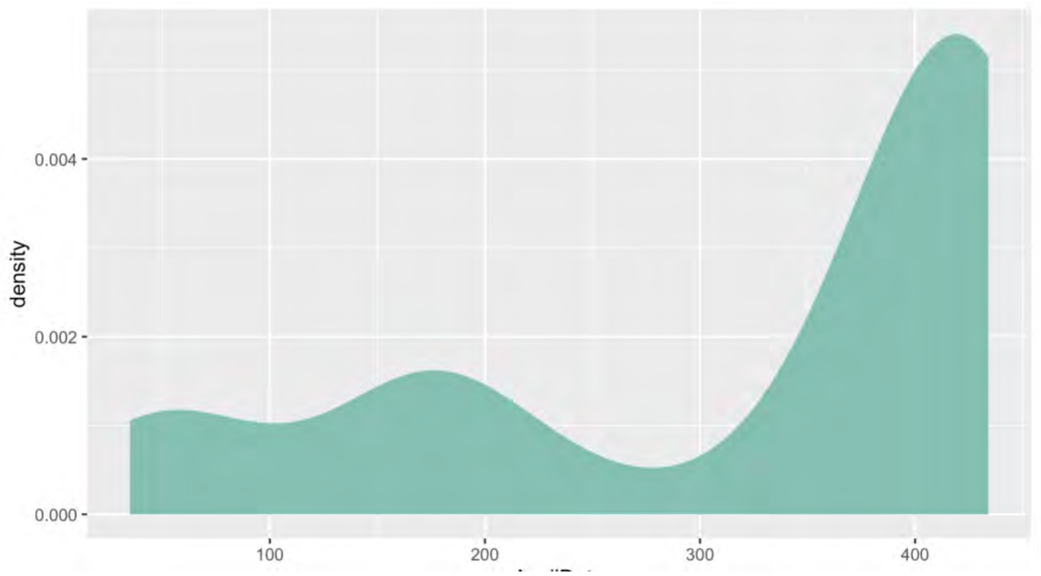

Figure 13: Density plot of car availability

xiv. Car_Usage_KM: As defined previously, this is the sum of the distance covered by the car's instances during the analysed period. Upon primary analysis of the data, we see that are some outliers in the data.

```
> summary(CarUsageData$Car_Usage_KM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    228   12691   20732   19889   24696   80390 
```

We then carry out further investigation using a boxplot to understand the outlier's range. We see from the boxplot below that there are a few observations out of the outer fence (40,000km). Upon checking the data, we notice some unusual kilometre readings that do not correspond to the trip time.

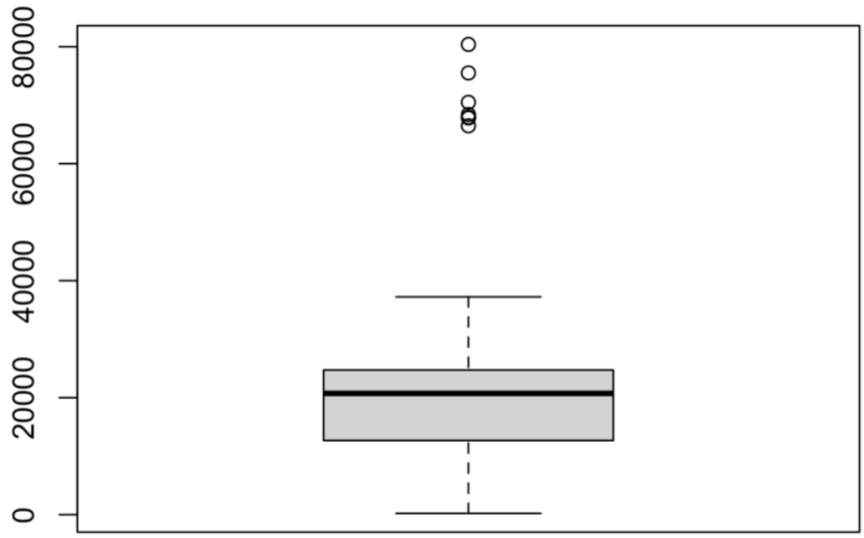

Figure 14: Boxplot for Usage in Kilometres

These outliers are treated as database write errors and deleted from the trips table.

xv. Usage_Density_KM_Day: This represents the average distance (in kilometres) that a car was driven everyday while the car was available on the network. This is measured in KM/Day.

```
CarUsageData3$Usage_Ratio_KM_Day
Min.   : 4.60
1st Qu.:46.80
Median :53.99
Mean   :55.16
3rd Qu.:64.06
Max.   :97.77
```

Table 2: Summary of Usage Density in KM/Day

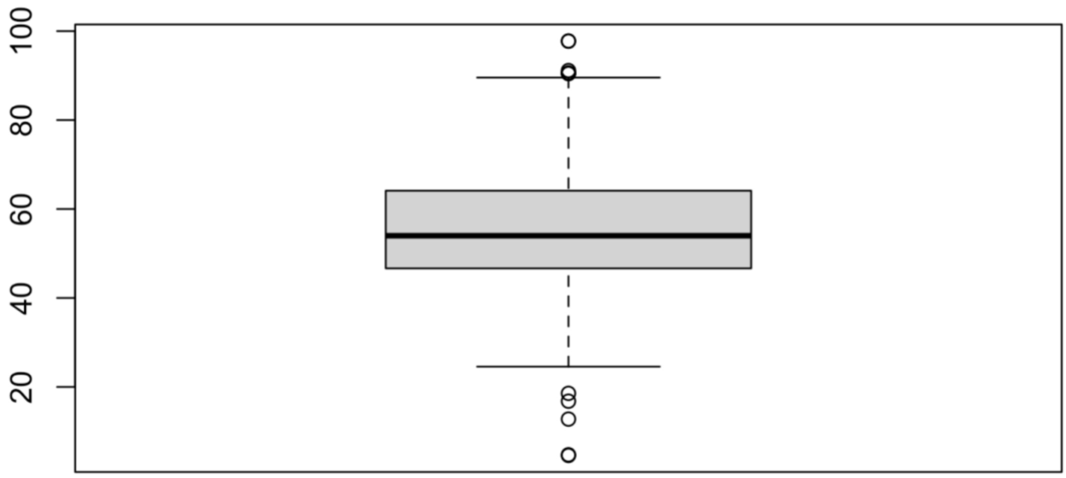

Figure 15: Box plot of Usage Density in KM/Day

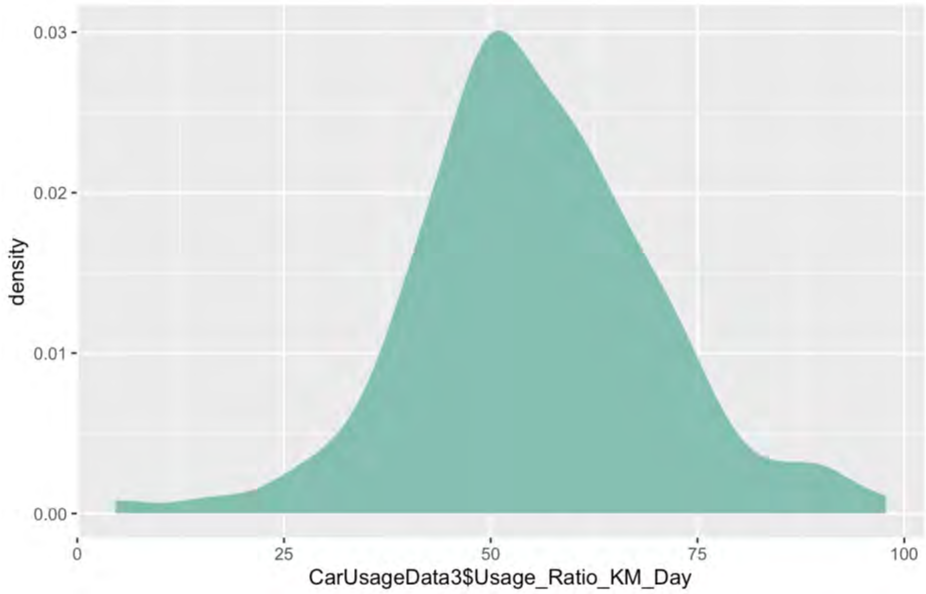

Figure 16: Density plot for Usage Density in KM/Day

xvi. Car_Usage_Mins: This represents the total amount of minutes that a car is in use during the analysed period.

```
> summary(CarUsageData2$Car_usage_mins)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1612   93939  164216  148862  203940  340006
```

Table 3: Summary of Car usage in minutes

The distribution appears to be without any outliers, and this is further investigated with a box plot.

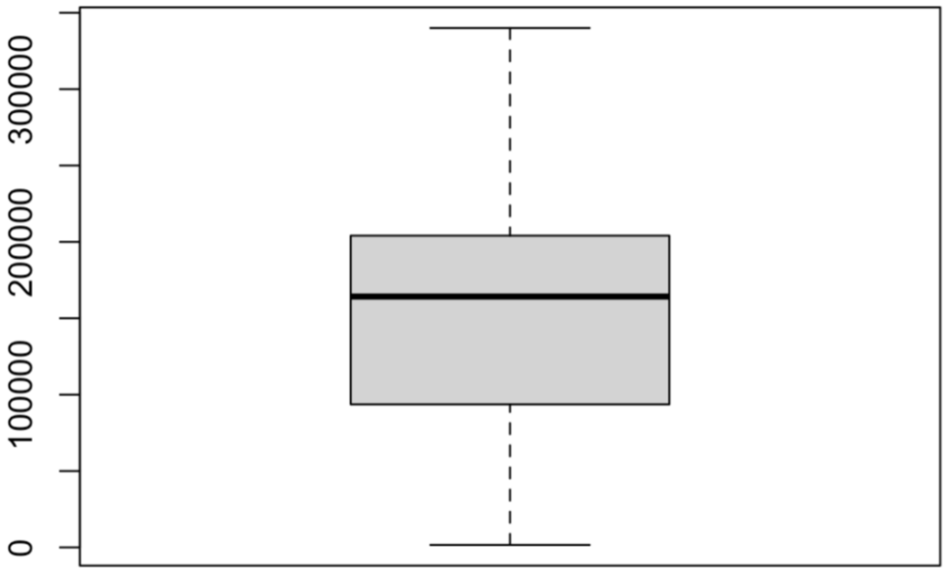

Figure 17: Boxplot of Car usage in minutes

As earlier hypothesised, there are no outliers in this variable data.

xvii. Usage_Density_mins_day: This represents the average time (in minutes) that a car was driven everyday while the car was available on the network. This is measured in Min/Day.

```
> summary(UsageratioMin)
 CarUsageData3$Usage_ratio_mins_day
 Min.   : 33.58
 1st Qu.:374.17
 Median :452.95
 Mean   :449.05
 3rd Qu.:529.37
 Max.   :783.42
```

Table 4: Summary of Usage Density in Min/Day

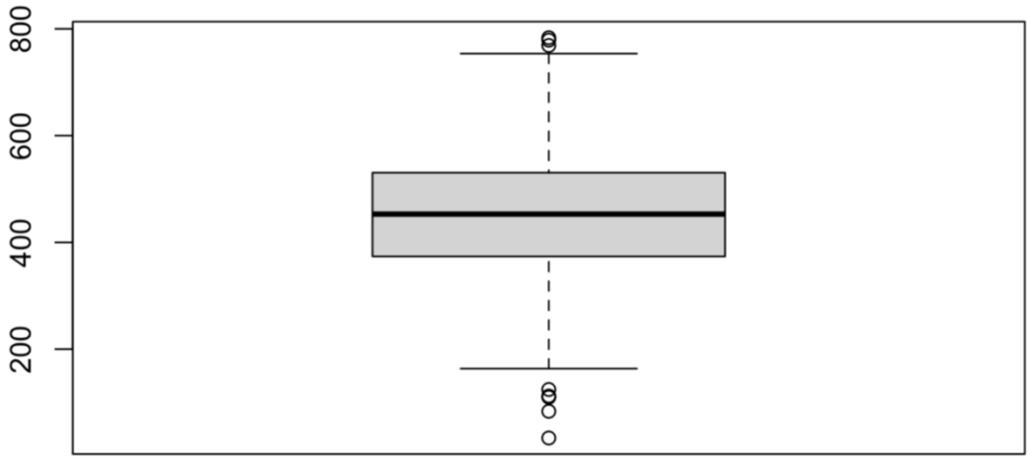

Figure 18: Boxplot of Usage Density in Min/Day

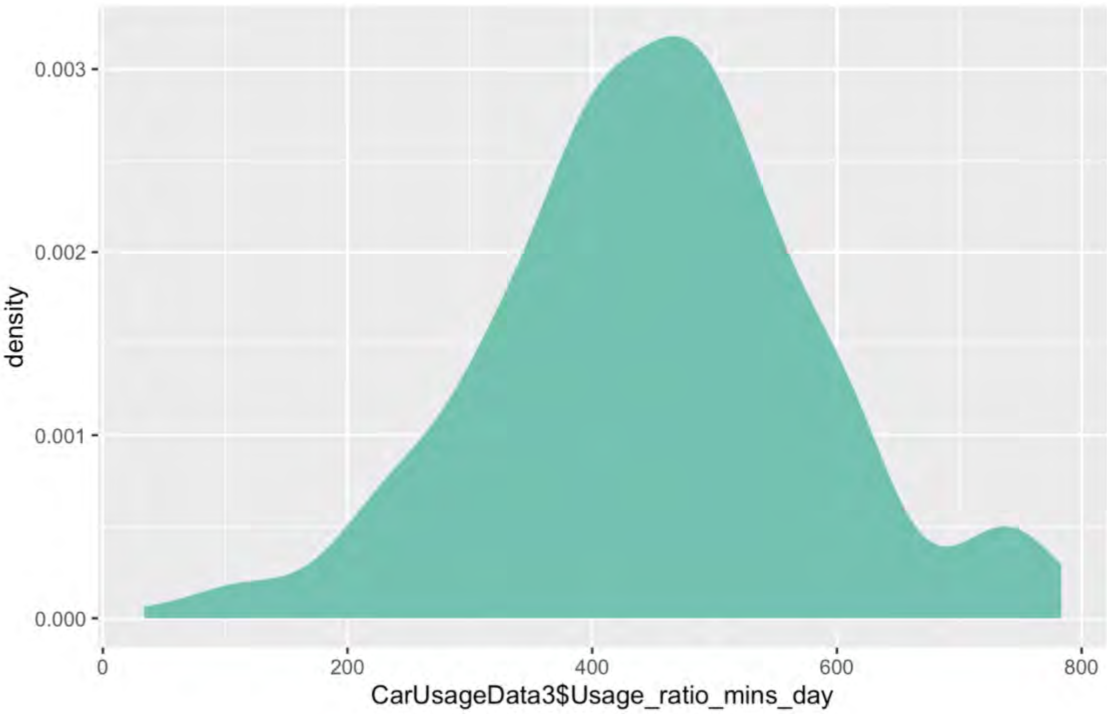

Figure 19: Density plot of Usage Density in Min/Day

4. METHODOLOGY

4.1 Preliminary Analysis

The preliminary analysis of the dataset that will be used in this study will be carried out using R programming language. R studio IDE RStudio 2021.09.1 will be used for this analysis which will involve visually exploring the variables that will be used in modelling the car usage data for Bildeleringen during the analysed period.

We have selected the Usage_ratio_min_days as the measure of usage for this study. This is due to the business model of Bildeleringen where the time (price per hour) constitutes a higher ratio of the billing. We will be exploring the data types, data range and highlighting some promising causal relationships between the dependent variable (Usage) and the dependent variables.

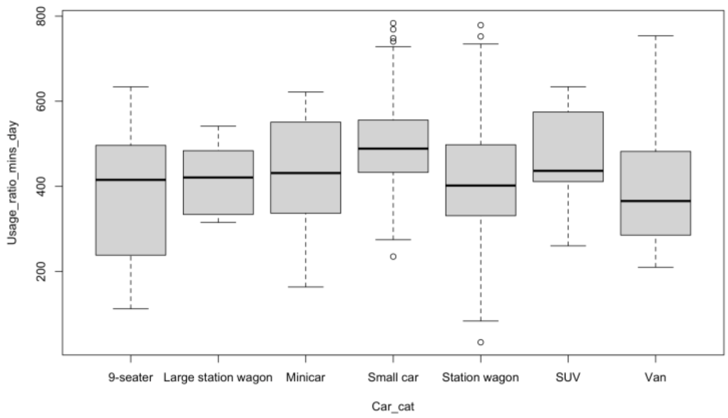

Figure 20: Boxplot of average usage ratio by car category

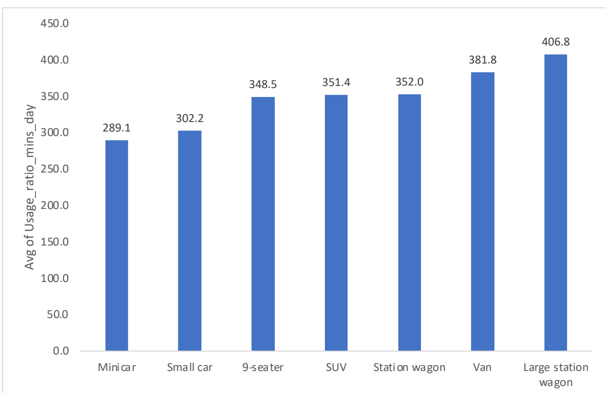

Figure 21: Bar chart of average usage ratio by car category

The mean usage of the larger cars (Vans and Large station wagons) is higher than that of the smaller cars (Minicars and small cars). This suggests a trend in the customer behaviour: car sharing is still being used mostly for moving equipment or group travels.

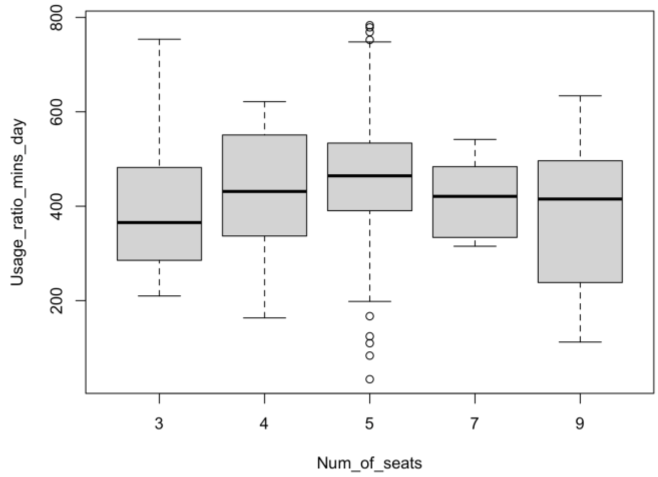

Figure 22: Boxplot of average usage ratio by number of seats

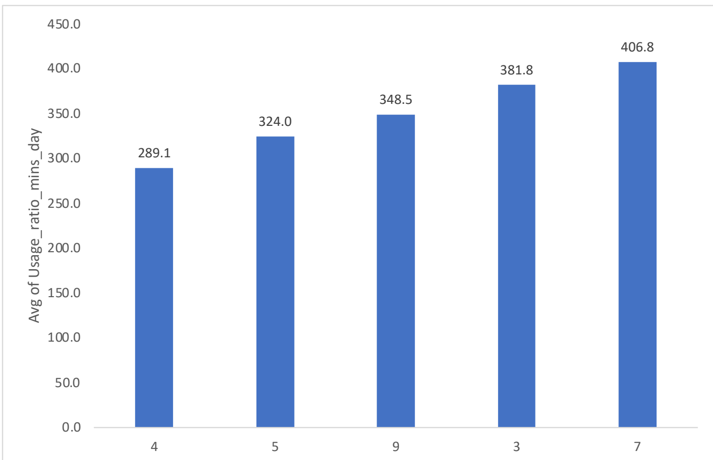

Figure 23: Bar chart of average usage ratio by number of seats

As suggested in the previous hypothesis, the current trend of usage is that most of the demand is for group travels or transporting large quantities of items. The trend in Figure 23 supports this hypothesis as Large station wagons (7-seater vehicles), and Vans (3 seaters) have the most usage.

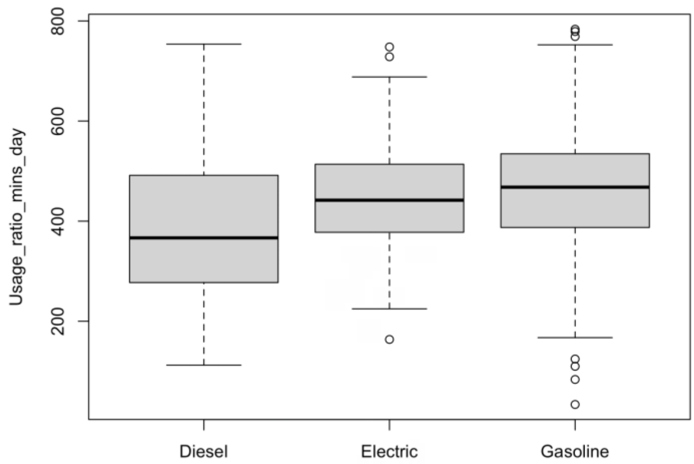

Figure 24: Boxplot of average usage ratio by fuel type

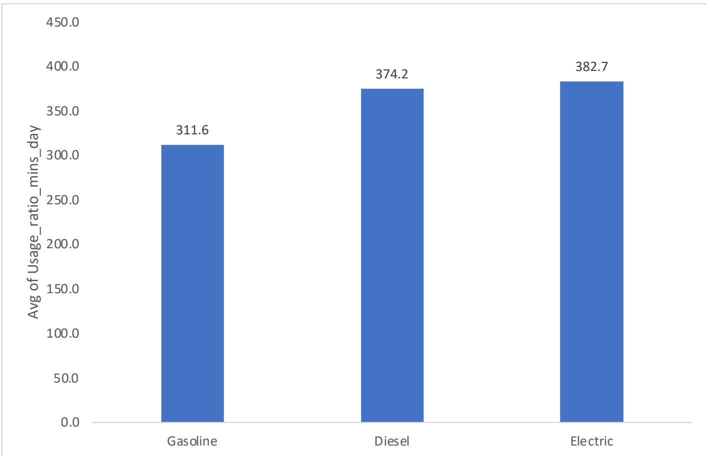

Figure 25: Bar chart of average usage ratio by fuel type

The trend in Figure 25 suggests that Electric cars have a higher usage than the other power types (Gasoline and Diesel). The gasoline powered cars have the least usage by a large quantity, this suggests customer preferences lie away from these kinds of cars.

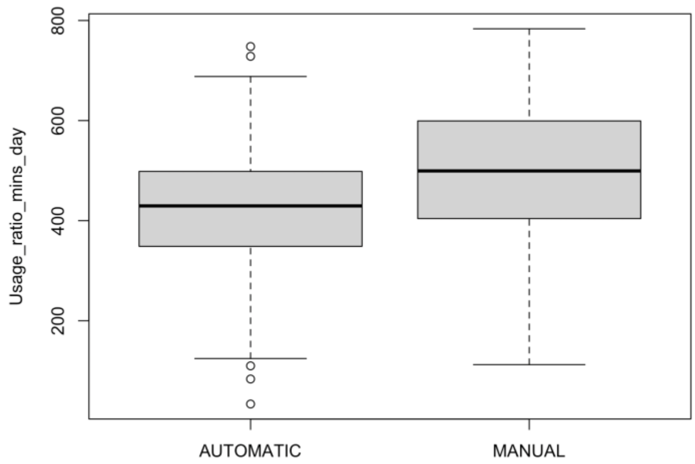

Figure 26: Boxplot of average usage ratio by transmission type

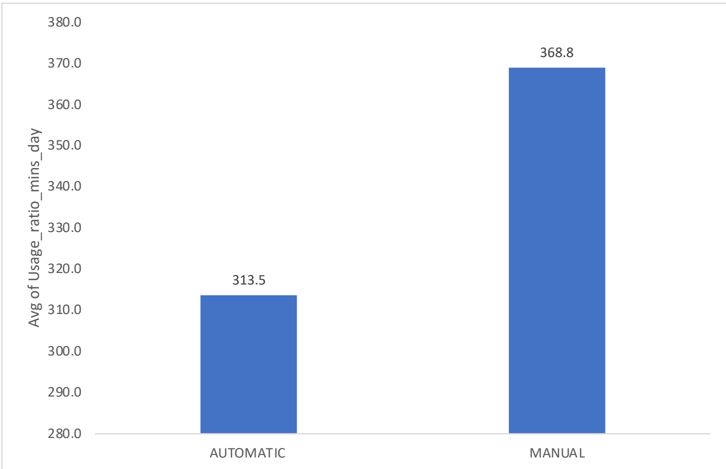

Figure 27: Bar chart of average usage ratio by transmission type

Figure 27 suggests that there is a clear preference for manual transmission vehicles over automatic transmission vehicles.

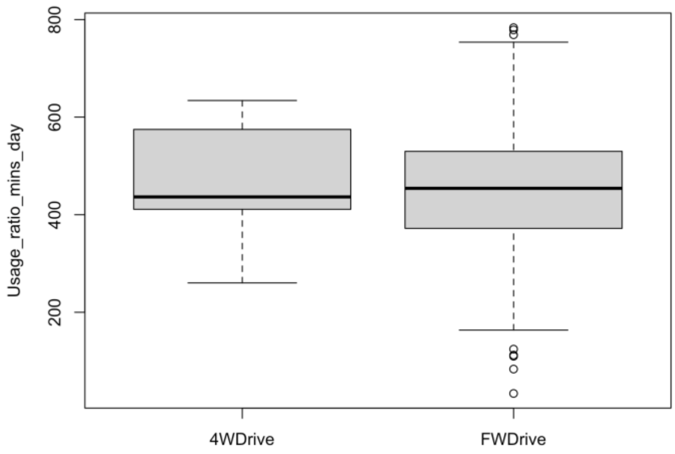

Figure 28: Boxplot of average usage ratio by wheel drive system

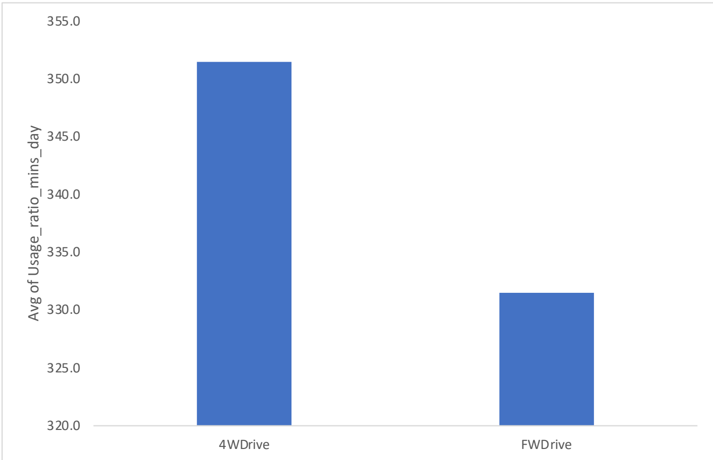

Figure 29: Bar chart of average usage ratio by wheel drive system

The wheel drive information on cars is not immediately observable by prospective users before ordering the vehicles and this might not affect the car demand. This inference is also visible from Figure 29 where we see very little difference in the usage for these two systems.

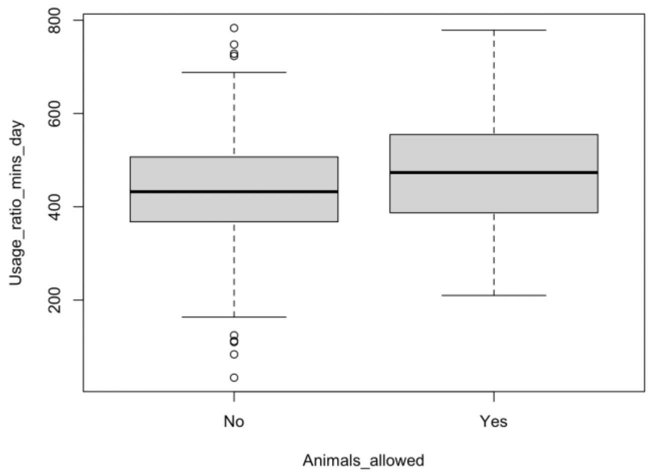

Figure 30: Boxplot of average usage ratio by animals allowed

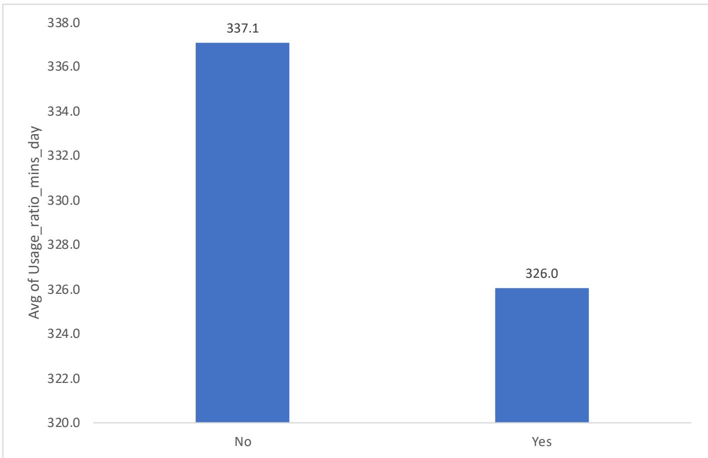

Figure 31: Bar chart of average usage ratio by animals allowed

There is no obvious difference in the average usage across cars that allow animals on them and do not. This will suggest that there is no real use case for animal travel.

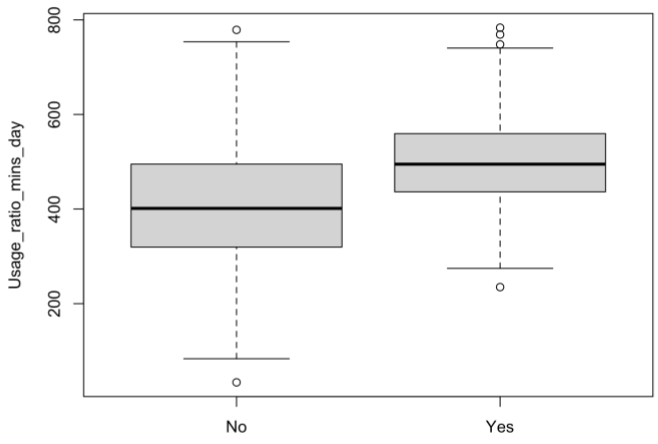

Figure 32: Boxplot of average usage ratio by baby pillow available

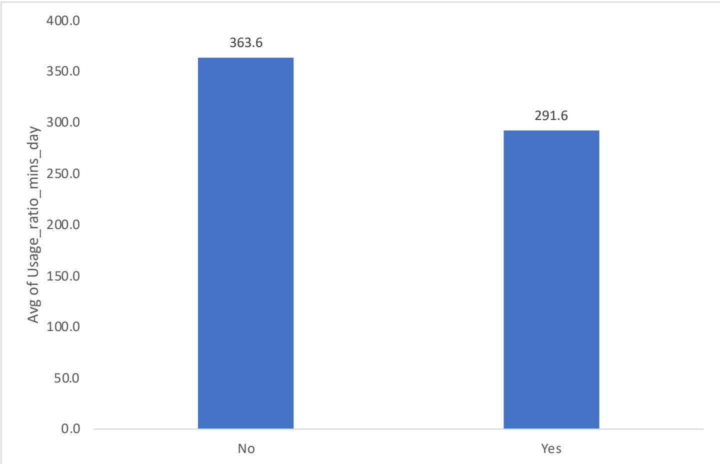

Figure 33: Bar chart of average usage ratio by baby pillow available

Figure 33 shows a considerable difference in the average usage for cars that have a pillow and those that do not. Very interestingly, the cars that do not have baby pillows have higher usage that do not.

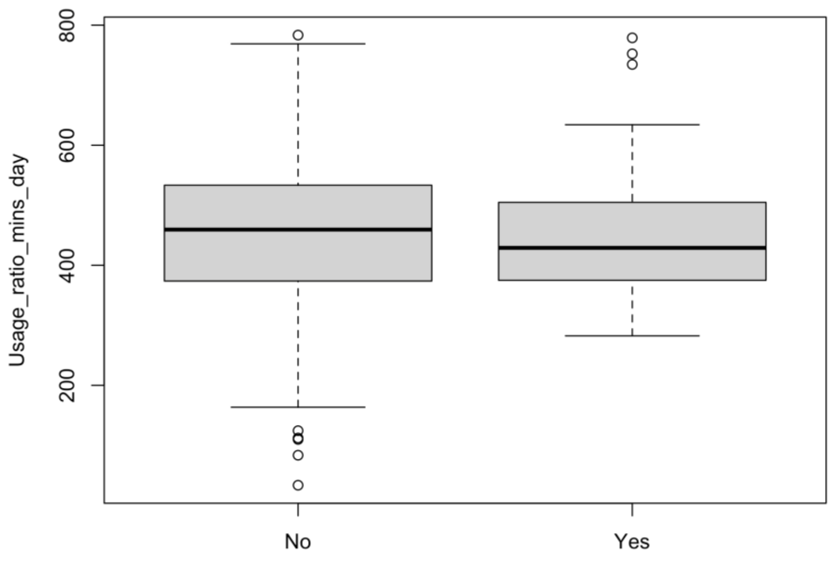

Figure 34: Boxplot of average usage ratio by child seat installed

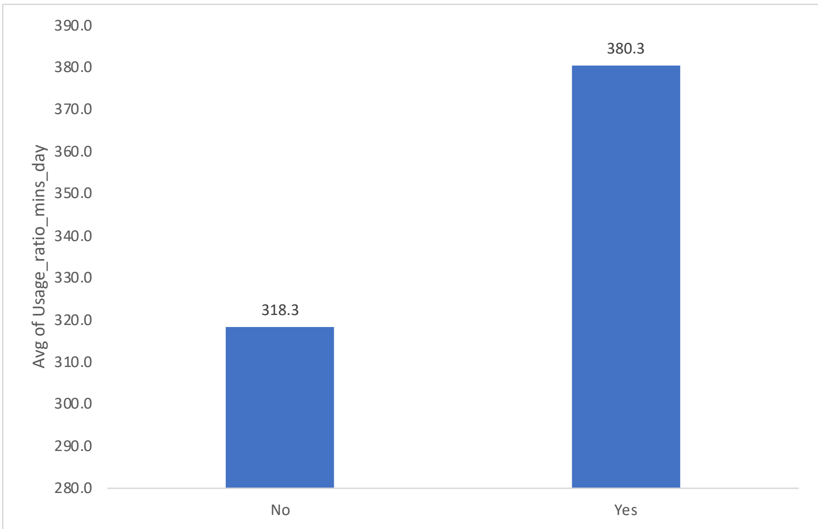

Figure 35: Bar chart average usage ratio by child seat installed

There is a large difference in the usage of cars that have child seat as compared to those that do not have, as seen in Figure 35.

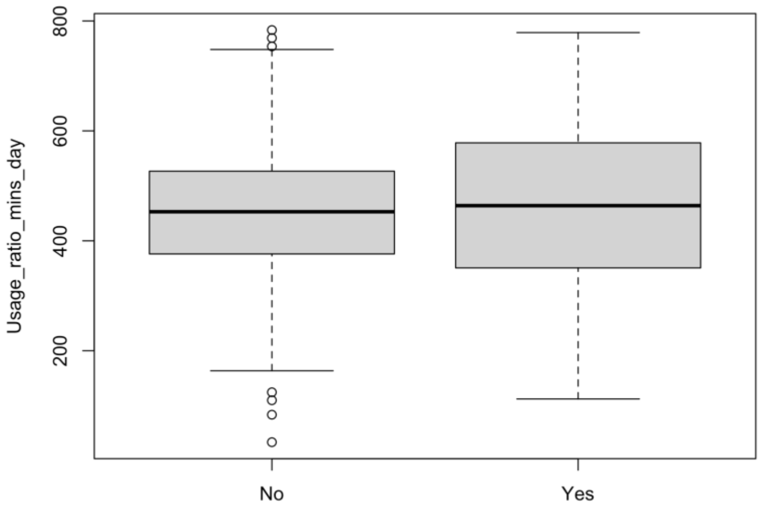

Figure 36: Boxplot of average usage ratio by towing hitch installed

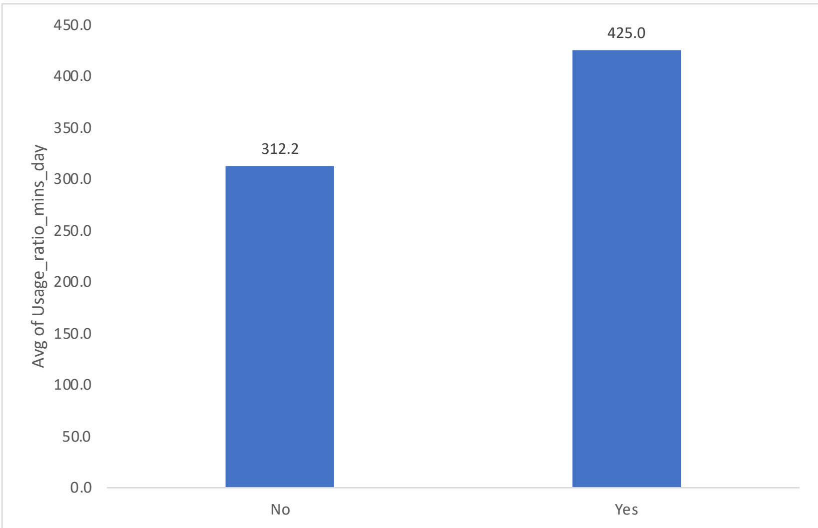

Figure 37: Bar Chart of average usage ratio by towing hitch installed

Also, Towing hitches seem to be prioritized by the users of the Bildeleringen platform, we see in the Figure above that the cars with the towing hitch have a much larger average usage than does that do not.

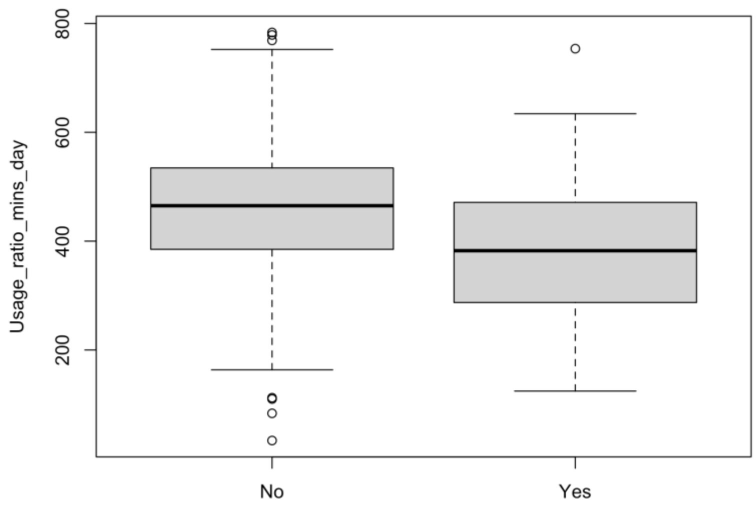

Figure 38: Boxplot of average usage ratio by roof rack installed

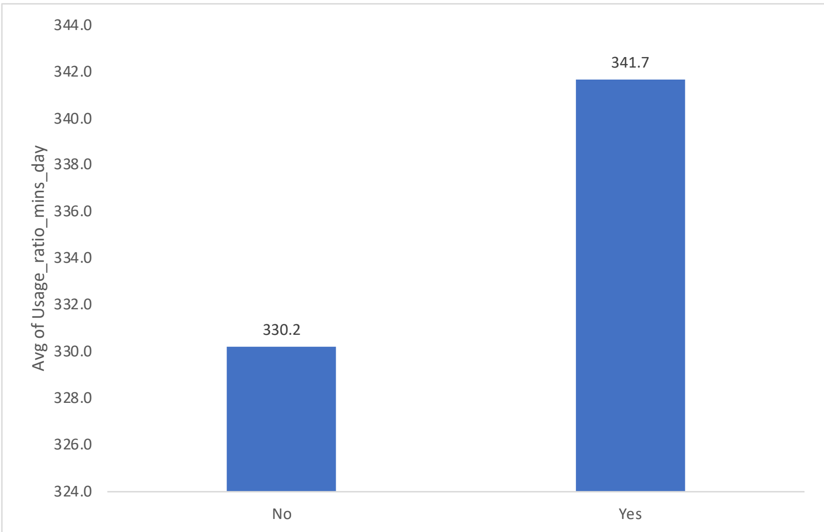

Figure 39: Bar chart of average usage ratio by roof rack installed

The average usage of the cars with and without the roof racks are very similar. Hence, this car feature is not important in evaluating the usage of cars.

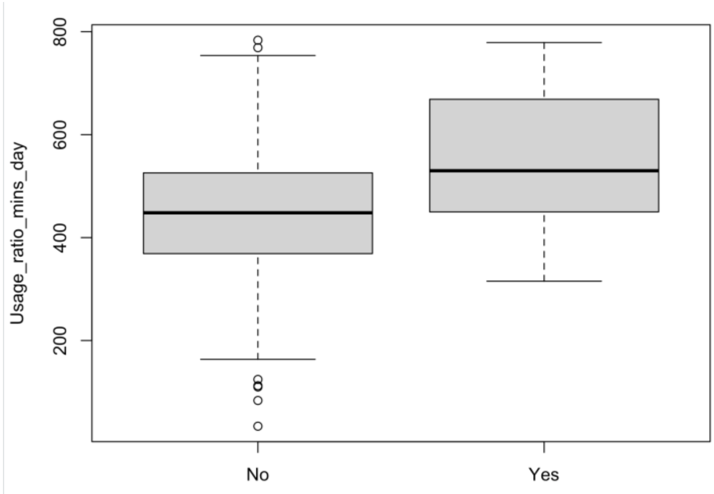

Figure 40: Boxplot of average usage ratio by roof box installed

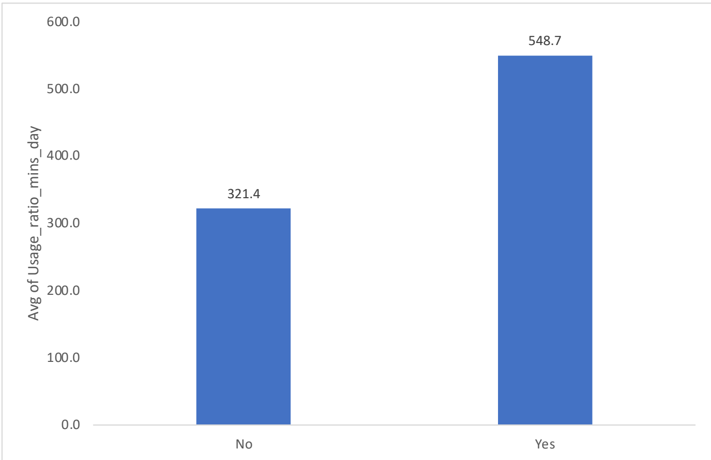

Figure 41: Bar chart of average usage ratio by roof box installed

With the largest difference in the car features discussed so far in this analysis, Figure 41 suggests that the Prescence or absence of roof racks is a very critical factor in the decision to use a car on the car sharing network.

4.2 Drivers of Demand

To determine which independent variables would be included or excluded from the regression model, we reviewed works of literature and built hypotheses based on theoretical support.

Hypothesis development and theories

1. Transmission – automatic or manual

In Europe, people prefer to drive with manual rather than automatic transmissions (Jen McCaffery, 2022). There are several reasons. First, automatics are more utilitarian that allow drivers to multi-task when they are driving (Jen McCaffery, 2022). However, it is not a culture in Europe where people would like to be more engaged with their cars and pay attention to their surroundings (Jen McCaffery, 2022). Second, Europeans prefer to drive manual because it is more fuel-efficient, and petrol is expensive in Europe. Third, from a driving licensing perspective, in Europe, if people have passed a driving test on an automatic car, they are only allowed to drive automatics. Nevertheless, if people have been given a driving test on a geared car, they receive a full license to drive automatic and manual vehicles ("Getting a Driving Licence in the EU," 2019). Therefore, Europeans learn how to drive and feel more comfortable driving a geared vehicle.

H1: There is a negative relationship between automatic cars and demand.

2. Power type – electric, gasoline, or diesel

Besides vehicles that run on gasoline and diesel, electric vehicles as new technology has been offered in the car-sharing market for some years. In recent years, governments and public authorities around the globe have increased energy efficiency targets and restricted the amount of C02 emissions (Glerum, Stankovikj, Thémans, & Bierlaire, 2014). Therefore, both car manufacturers have introduced electric vehicles into the market on a large scale and the demand for electric vehicles is increasing (Glerum, Stankovikj, Thémans, & Bierlaire, 2014). When opposed to gasoline or diesel, electric cars have significant advantages in that they generate no carbon dioxide or greenhouse gases (Glerum, Stankovikj, Thémans, & Bierlaire, 2014). However, they have limitations: their driving range is restricted, a full charge of the battery may take up to eight hours (until quick charges are available), and there are currently few charging stations accessible (Glerum, Stankovikj, Thémans, & Bierlaire, 2014). Hence, the driving range is one of the major choice determinants when people purchase a vehicle or decide the type of vehicle to use in car-sharing services (Giansoldati, Danielis, Rotaris, & Scorrano, 2018). Fortunately, Norway has great incentives for the adoption of electric vehicles such as registration tax exemption, free parking, access to bus lanes, exemption from toll roads, development of dense charging stations in cities, and so on (Zhang, Qian, Sprei, & Li, 2016). Research has suggested that technological improvements in electric vehicles, space, toll waivers, and charging stations density have a positive impact on the demand for electric cars, and therefore, Norway has the highest market penetration rate for electric vehicles in 2014 and has continued to undergo a significant growth (Zhang, Qian, Sprei, & Li, 2016; Cyriac & Erik Julsrud, 2018). As Norway is a world leader in terms of electric car support mechanisms, infrastructure, and user acceptance, as well as people are more environmentally conscious and are more willing to purchase and drive with electric vehicles, the demand for cars with different fuel technologies can be significantly impacted (Cyriac & Erik Julsrud, 2018).

H2: There is a positive relationship between electric cars and demand.

3. Wheel drive – front-wheel drive (FWD) or all-wheel drive (4WD)

The primary difference between front-wheel drive (FWD) and all-wheel drive (4WD) is the number of wheels to which power may be applied. An FWD vehicle can solely power the front wheels while a 4WD car can power all four wheels. One major benefit of 4WD over FWD is that it enables more control and mobility in rough terrain and severe weather such as snow, mud, boulders, and so on, where FWD can easily lose traction. In addition to the snowy weather and landscape diversity in Norway, crossovers, SUVs, and hybrid and electric cars are becoming more popular, hence, the demand for vehicles will 4WD also increase (Loveday, 2022).

H3: There is a negative relationship between FWD and demand.

4. Baby pillow and child seat

In 2016, 630 children were dead because of road traffic collisions in Europe ("ETSC | European Transport Safety Council," 2021). Child road deaths account for around 2.5% of total road deaths and approximately 6% of all major road traffic injuries in Europe ("ETSC | European Transport Safety Council," 2021). In Norway, by legal regulations, drivers must ensure that children under four and a half feet (135cm) must wear weight-appropriate kid restraints (“Trygg Trafikk • Trafikksikkerhet for Alle - Barn, Unge Og Voksne,” 2010). The child seat and baby pillow have complemented each other. When using car-sharing services, families with children although able to bring their own baby seat, prefer to rent it from the car rental companies for convenience (“Car Hire Price Comparison | Car Booker,” 2021).

H4: There is a positive relationship between child seats, baby pillows, and demand.

5. Roof box

A roof box is a storage space attached to a vehicle’s roof. The global roof rack market size for vehicles was USD 2.11 billion in 2020 and is expected to grow from $2.18 billion in 2021 to $3.39 billion in 2028 (“Automotive Roof Rack Market Size, Growth & Forecast [2028],” 2020). Road trip is being more popular nowadays and car-sharing is an option, the demand for more luggage storing space as a result of tourism, recreational activities, and people moving around cities all drive the demand for car racks.

H5: There is a positive relationship between roof box and demand.

6. Car category

When choosing a car to be rented for car-sharing, there are different types of rental cars. Small and medium-sized cars are more popular in Europe as they account for more than half (51%) of total EU passenger car sales (“New Passenger Cars by Segment in the EU,” 2021). In addition, depending on the trip, it can affect the car rental categories to be picked. For example, SUVs are more suitable for long trips and rural roads while mini-cars are more suitable for city and short trips. Bildeleringen offers 7 vehicle types and the demand for each can vary.

H6: There is a relationship between car category and demand.

Model development process

We have applied a backward stepwise regression approach where we first included all the selected variables to build a full regression model, then gradually removed variables from the regression model at each step to get the best-fitting reduced model. Using a significance level of 0.05, we removed one variable at each step (the least significant variable – the variable with the largest p-value) until we obtained a model that contain all significant variables. The following is our final model.

```
Call:
lm(formula = Usage_ratio_mins_day ~ Transmission + Wheel_drive +
    Baby_pillow + Child_seat + Roof_box, data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max
-298.44  -79.08   -4.21   71.85  345.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    544.01      42.54  12.787  < 2e-16 ***
Transmission   -76.62      15.03  -5.097 6.83e-07 ***
Wheel_drive   -135.37      42.13  -3.213  0.00148 **
Baby_pillow    143.14      16.68   8.580 1.02e-15 ***
Child_seat      80.76      19.67   4.106 5.47e-05 ***
Roof_box       137.61      34.24   4.019 7.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112.6 on 250 degrees of freedom
Multiple R-squared:  0.3084,	Adjusted R-squared:  0.2946
F-statistic:  22.3 on 5 and 250 DF,  p-value: < 2.2e-16
```

Code Snippet 1: Linear Regression

Findings

Although we have selected 7 relevant independent variables with theoretical support (transmission, power type, wheel drive, baby pillow, child seat, roof box, and car category) in our initial model, it turns out that only 5 of them are significant (transmission, wheel drive, baby pillow, child seat, and roof box), the 2 variables with a p-value greater than 0.05 are car category and power type. Overall, the p-value of the regression model is less than 0.05, and the model is significant. The regression has an R square of 30.84%, which indicates that these 5 significant variables explained 30.84% of the variability in the demand. All coefficient signs are consistent with our hypotheses. Transmission and wheel drive have negative signs, if the transmission is automatic or has an FWD wheel, the Usage_ratio_mins_day will be decreased by 76.72 and 135.37 respectively. On the other hand, baby pillow, child seat, and roof box have positive coefficient signs. If a car has a baby pillow, child seat, or a roof box, the Usage_ratio_mins_day will be increased by 143.14, 80.76, and 137.61 respectively.

4.3 Demand Prediction

Model development process

For our regression problem, we have imported the data into R from the CSV file: *CarDataVar.csv.* We have 256 car observations across 13 variables. The first variable Car_id is the identifier variable for the observations in the data set. Since this does not contribute to the model, for the rest of the analysis this will be deleted.

```
> head(CarReg)
# A tibble: 6 × 13
  Car_id Car_cat Num_of_seats Power_type Transmission Wheel_drive Animals_allowed
   <dbl>   <dbl>        <dbl>      <dbl>        <dbl>       <dbl>           <dbl>
1     39       1            5          1            1           1               1
2     40       1            5          1            1           1               1
3     41       1            5          1            1           1               1
4     42       1            5          1            1           1               1
5     16       2            4          3            2           1               0
6     37       2            4          3            2           1               1
# … with 6 more variables: Baby_pillow <dbl>, Child_seat <dbl>,
#   Towing_hitch <dbl>, Roof_racks <dbl>, Roof_box <dbl>,
#   Usage_ratio_mins_day <dbl>
```

Code Snippet 2: Car Reg Data Head

We then proceed to divide the data into the training and test data, using a 50:50 split of randomized samples. In other to ensure reproducible results, we use the set.seed function.

```
> set.seed(123456)
> n=nrow(CarReg)
> ind=sample(1:n,size=n/2)
> train=CarReg[ind,]
> test=CarReg[-ind,]
```

Code Snippet 3: Transform data frame to matrix

We also transform the data into matrices for the shrinkage methods whose functions accept predictors and dependent variable and observation data in the matrix form and not in data frame format.

4.3.1 Multiple Linear Regression

In this method, we run a regression function on the entire data set and all of the regressors in the training data using the least squares. The prediction performance of the model is then measured on the test data, outputting the Mean Squared Error as the indicator.

```
> summary(MRegr)

Call:
lm(formula = Usage_ratio_mins_day ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-303.289  -74.662    8.165   61.799  261.154 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      668.839     97.870   6.834 4.06e-10 ***
Car_cat          -16.125     17.191  -0.938 0.350199    
Num_of_seats       7.575     14.685   0.516 0.606964    
Power_type        -9.305      9.912  -0.939 0.349795    
Transmission    -109.443     29.693  -3.686 0.000348 ***
Wheel_drive     -167.983     71.375  -2.354 0.020278 *  
Animals_allowed   31.029     22.566   1.375 0.171764    
Baby_pillow       36.469     54.800   0.665 0.507057    
Child_seat        52.030     34.677   1.500 0.136216    
Towing_hitch     -88.160     39.407  -2.237 0.027186 *  
Roof_racks         4.897     40.591   0.121 0.904189    
Roof_box         176.808     52.129   3.392 0.000950 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 115.9 on 116 degrees of freedom
Multiple R-squared:  0.2843,	Adjusted R-squared:  0.2164 
F-statistic: 4.188 on 11 and 116 DF,  p-value: 3.25e-05
```

Code Snippet 4: Kitchen Sink regression on all variables

As discussed in the section 4.2 (Drivers of demand), the multiple regression identifies Transmission, Wheel_drive, Towing_hitch and Roof_box as the significant variables (regressors). We then take make predictions based on this model using the test data and calculate the MSE in the next step.

```
> ErrMRegr = PredMRegr - test$Usage_ratio_mins_day
> MSEMRegr = mean((ErrMRegr) ^2)
> MSEMRegr
[1] 13027.07
```

Code Snippet 5: Calculation of MSE for Multiple Regression

4.3.2 Linear Regression-Backward Stepwise selection

Next, we take into consideration only a subset of the predictors in our linear regression model. This helps to ensure that we are only taking into consideration the best subset of predictors in order to improve the model prediction performance. This helps to eradicate the effect of noise from the insignificant variables.

There are two major procedures to choosing the best subset of variables: best subset and stepwise model selection. The difference between these two methods is that while best subset selection method fits a regression for all possible combination of variables, the stepwise model selection systematically selects the possible combination eradicating some possibilities early on.

For our analysis, we run the best subset selection process first using the regsubsets function from the leaps library. The algorithm for the subset selection is divided into a 3-step process by James et al (2021).

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \dots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Figure 42: Best subset selection logic (James et al, 2021)

The best regression models per model size for the best subset regression is then viewed using the summary () function.

```
1 subsets of each size up to 11
Selection Algorithm: exhaustive
          Car_cat Num_of_seats Power_type Transmission Wheel_drive Animals_allowed
1  ( 1 )  " "     " "          " "        "*"          " "         " "
2  ( 1 )  " "     " "          " "        " "          " "         " "
3  ( 1 )  " "     " "          " "        "*"          " "         " "
4  ( 1 )  " "     " "          " "        "*"          " "         "*"
5  ( 1 )  "*"     " "          " "        "*"          "*"         " "
6  ( 1 )  "*"     " "          " "        "*"          "*"         "*"
7  ( 1 )  "*"     " "          " "        "*"          "*"         "*"
8  ( 1 )  " "     " "          "*"        "*"          "*"         "*"
9  ( 1 )  "*"     " "          "*"        "*"          "*"         "*"
10  ( 1 ) "*"     "*"          "*"        "*"          "*"         "*"
11  ( 1 ) "*"     "*"          "*"        "*"          "*"         "*"
          Baby_pillow Child_seat Towing_hitch Roof_racks Roof_box
1  ( 1 )  " "         " "        " "          " "        " "
2  ( 1 )  "*"         " "        " "          " "        "*"
3  ( 1 )  "*"         " "        " "          " "        "*"
4  ( 1 )  "*"         " "        " "          " "        "*"
5  ( 1 )  " "         " "        "*"          " "        "*"
6  ( 1 )  " "         " "        "*"          " "        "*"
7  ( 1 )  " "         "*"        "*"          " "        "*"
8  ( 1 )  "*"         "*"        "*"          " "        "*"
9  ( 1 )  "*"         "*"        "*"          " "        "*"
10  ( 1 ) "*"         "*"        "*"          " "        "*"
11  ( 1 ) "*"         "*"        "*"          "*"        "*"
```

Code Snippet 6: Summary of Subset regression output

By default, a maximum of 8 variables subset models are evaluated by the regression function. In order to consider larger sizes, the nvmax term can be used to specify a smaller or larger number. Each column represents the number of variables in the model (1 - 8), and the asterisk (*) serves as the indicator of whether or not the variable is included in the model. For this study, we have used a size 12, which is the total number of variables in our dataset.

In order to select the best subset model, we can examine the following model indicators; R^2, RSS, adjusted R^2, Cp, and BIC. These can be retrieved using the summary () function or plotted using the plot function from the regsubsets library. In this study, we will use the adjusted R^2 as the factor in determining the best subset. We can make this comparison by plotting the models, ranked by the adjusted R^2.

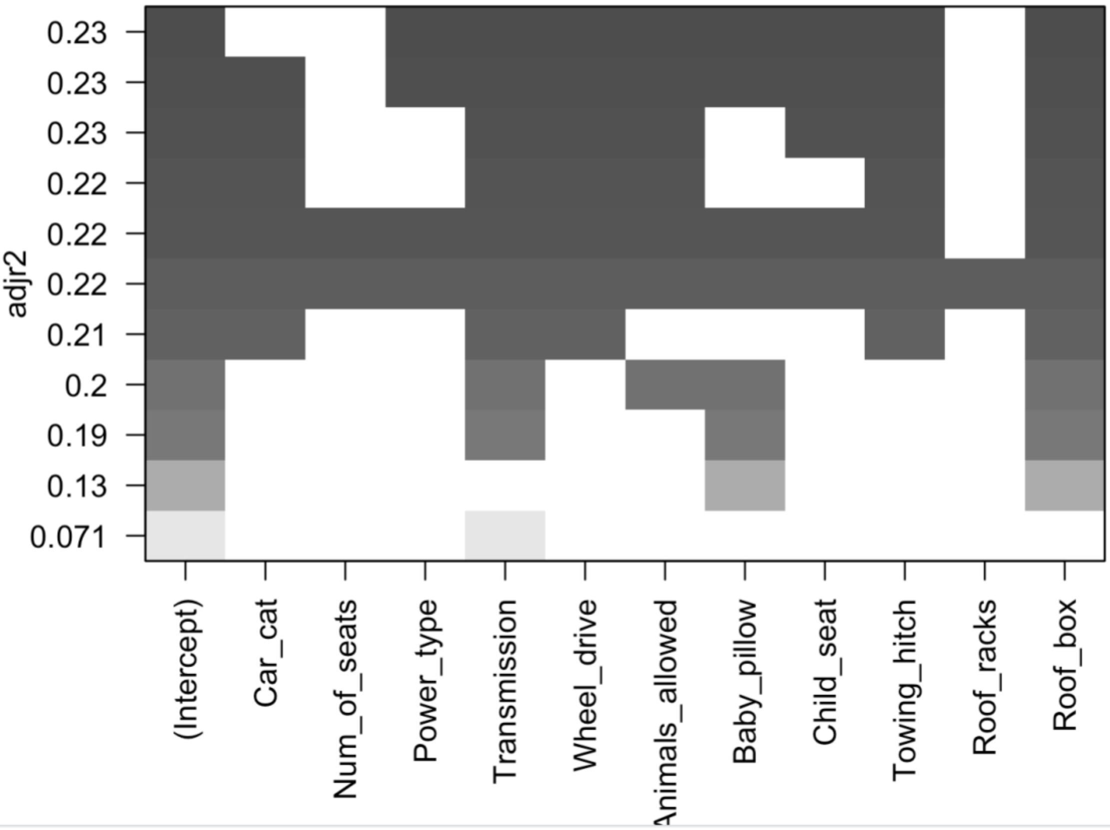

Figure 43: Plot of variables by adjusted Rsquare

This shows that the best subset of variables to be used for the regression is all of the variables except the Car_cat, Num_of_seats, and Roof_racks.

The next step for us is to check if using the backward and forward selection methods give us a varying subset of variables to be used for the regression. To change the subset selection method to forward or backward selection method, the method term is added to the regsubset function.

```
forwardSel=regsubsets(Usage_ratio_mins_day~.,data= train, nvmax = 12,
                       method = "forward")
plot(forwardSel,scale = "adjr2",col = gray.colors(1000))
forwardsum=summary(forwardSel)
```

Code Snippet 7: Forward subset selection

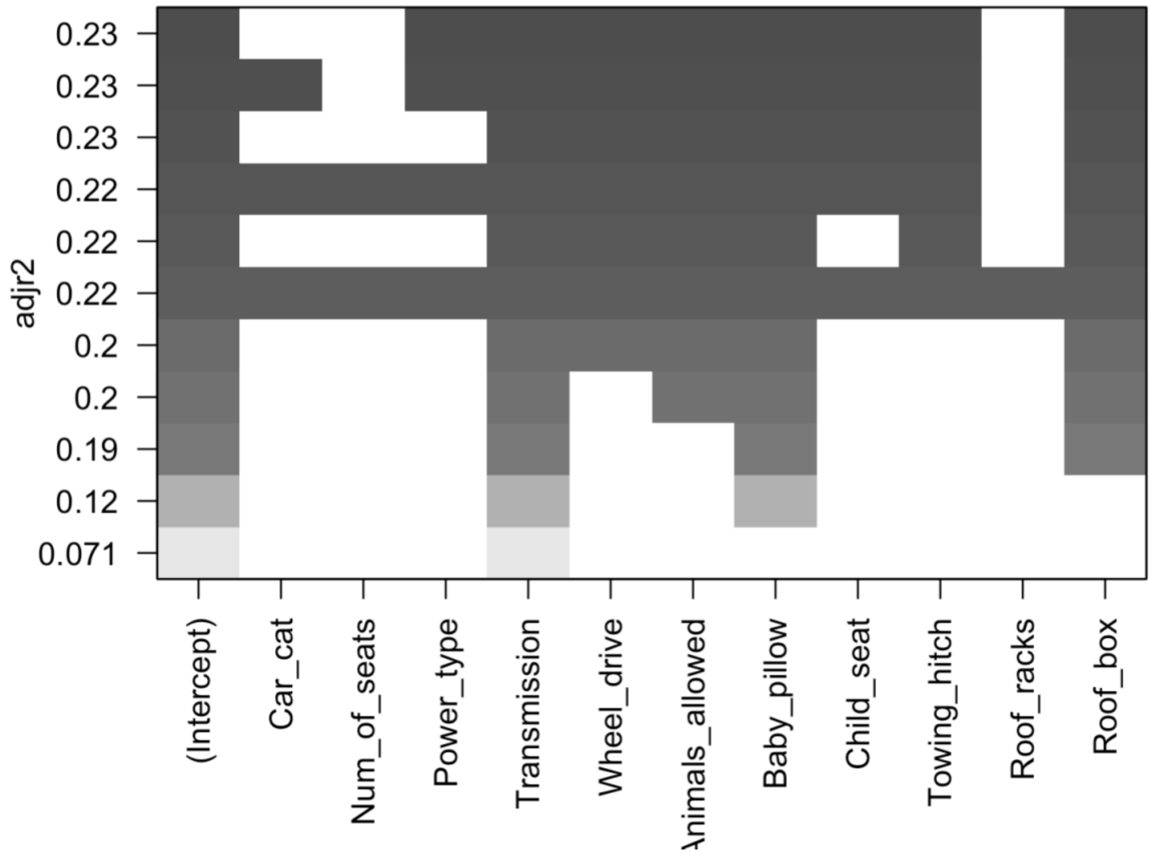

Figure 44: Plot of variables by adjusted RSquare for Forward subset selection

The result of this regression shows the same output as the best subset selection, with the same 8 variable subset being the best based on the adjusted R^2. We repeat the same process for the backward selection process as well using the function below.

```
> backwardSel=regsubsets(Usage_ratio_mins_day~.,data= train, nvmax = 12,
+                       method = "backward")
> plot(backwardSel,scale = "adjr2",col = gray.colors(1000))
> backwardsum=summary(backwardSel)
> backwardsum
```

Code Snippet 8: Backward subset selection

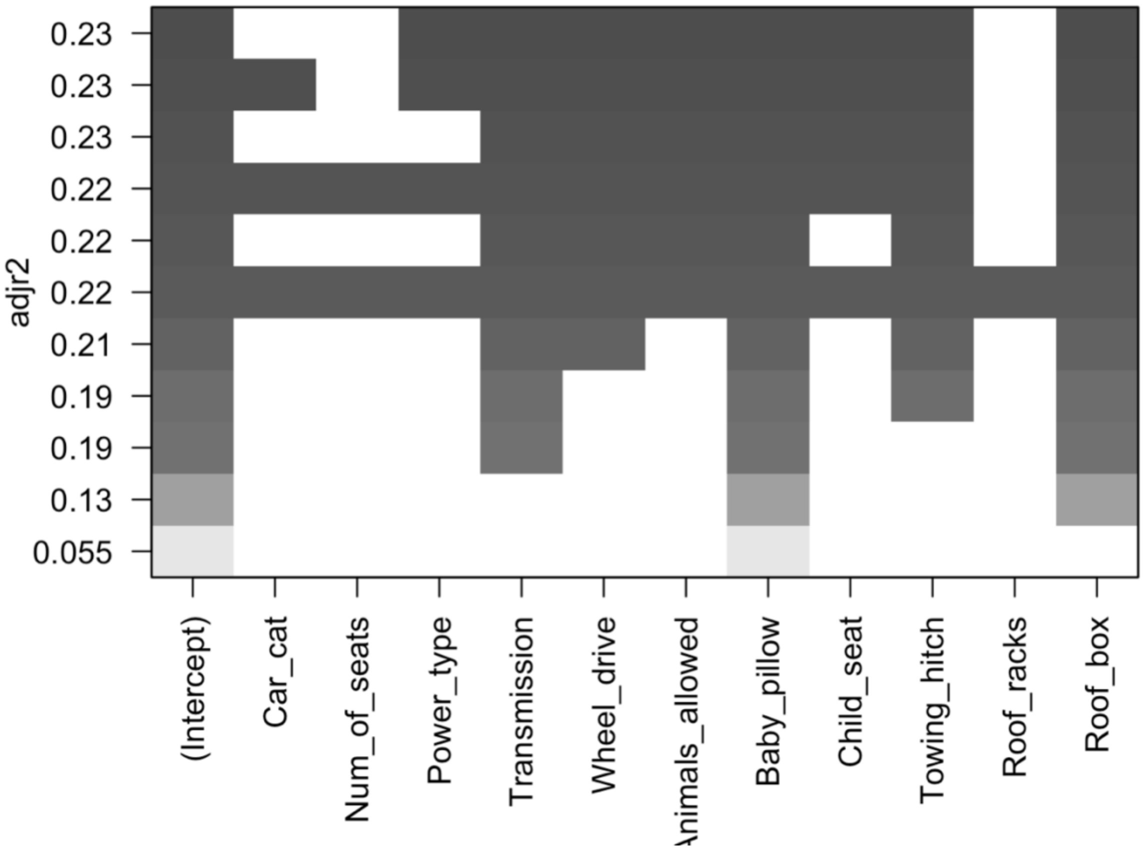

Figure 45: Plot of variables by adjusted RSquare for Backward subset selection

The result is also identical. This proves beyond reasonable doubt that this is the best subset, and we can go ahead to run a multiple linear regression based on these variables.

```
> PredSelRegr=predict(SelectRegr, newdata = test)
> ErrSelRegr = PredSelRegr - test$Usage_ratio_mins_day
> MSESelRegr = mean((ErrSelRegr) ^2)
> MSESelRegr
[1] 12924.7
```

Code Snippet 9: Calculation of MSE for Subset selection

The test MSE of this prediction model obtained by cross validation is marginally lower than the kitchen sink multiple regression model using all of the variables (12,924.7)

4.3.3 Shrinkage Methods

Shrinkage methods are defined as regularisation methods that involve fitting a regression model using all predictors, under some constraint on the size of their estimated coefficients (Andreis, 2017). As opposed to the previous methods used (Regression and Piecewise regression) where we used only the significant predictors, we use all of the predictors albeit in varying proportions.

In order to avoid over fitting from using all of the predictors, Shrinkage methods are employed to constrain the impact of regressors in the model. This is achieved by varying the coefficient estimates of the predictors in the regression. To do this, the method applies a penalty term to the loss function of the model (in the equation below). This penalty term is defined as Lambda (λ).

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

Equation 5 : RSS Formula

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

Equation 6: Incorporation of Lambda for Shrinkage

In order to constrain the regressors, the shrinkage methods shrink the coefficient estimates of the regression towards zero (Copas, 1993). As λ increases, the coefficients shrink towards 0.

For the different shrinkage methods, this is achieved in different ways. In this study, we will consider only the Ridge and Lasso. This serves as the basis for the difference between the Lasso and Ridge methods. For Ridge, the method tries to shrink the coefficient estimates towards 0, however the coefficients are never zero. However, the Lasso method can shrink the coefficient estimates to 0.

Shrinkage methods are best known to maximize the performance of the model by reducing the variance of the regressors and reducing the loss function. It has however been noted that this is not the case in all scenarios. Especially in cases when the dataset is not large, penalization (shrinkage) methods can be unreliable, owing to the unknown shrinkage and tuning parameters which are estimated with large uncertainty (Riley et al., 2021).

Ridge Regression

This is the first shrinkage method which will be used for this prediction problem. This is very similar to the Ordinary Least Square (OLS) method, with the major difference being that Lambda, λ, tries to fit the coeeficeints towards Zero. Hence the coefficient estimates are usually closer to Zero when compared to the OLS results.

Ridge regression considers all of the predictors, and this sometimes is a downer to the model. This problem is tackled by the Lasso regression (to be discussed in the next section).

Lasso Regression

The problem highlighted by critics of the shrinkage methods is that it includes all of the regressors in the regression function. This is addressed with the Lasso regression, which the penalty term can be shrinked to zero for some coefficients in a bid to achieve the optimal bias-variance trade off. The Lasso regression uses the L-1 penalty which takes the absolute value of the coefficient (Ridge regression takes the square of the coefficient).

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

Equation 7: Lasso regression

The glmnet function in the glmnet package is used to run the ridge and lasso regressions. Before running the regressions however, we must determine the optimal λ that reduces the MSE.

```
#Solve for lamba for Ridge
lambdamin=cv.glmnet(X,y,alpha=0,nfold=n)$lambda.min
lambdamin
```

```
> lambdamin
[1] 25.8195
```

Code Snippet 10: Calculation for Lambda

Using the λ = 25.8195, we then fit the model using the Ridge regression and we obtain the results below.

```
> #Create Ridge model
> ridgemin=glmnet(X,y,alpha=0,lambda=lambdamin)
> ridgemin

Call:  glmnet(x = X, y = y, alpha = 0, lambda = lambdamin)

  Df  %Dev Lambda
1 11 26.52  25.82
```

Code Snippet 11: Ridge regression

We use a similar method for the Lasso, with a slight change in the glmnet function. We change the alpha term to 1 from 0 used for the Ridge regression.

```
> lasso

Call:  glmnet(x = X, y = y, alpha = 1, lambda = lassolambda)

  Df  %Dev Lambda
1  4 19.91   9.95
```

Code Snippet 12: Lasso Regression

The Lambda used for the shrinkage here is significantly lower than the value for Lambda in the Ridge regression. We also notice a lower deviation in the Lasso regression signifying greater accuracy in this model.

Next, we compare the coefficients of the two regression functions with that of the ordinary least squares to observe the evolution of the coefficient estimates of the regressors across the different regression methods.

```
> cbind(coef(ols),coef(ridgemin),coef(lasso))
12 x 3 sparse Matrix of class "dgCMatrix"
                                    s0        s0
(Intercept)     127.108811 230.815707 342.11124
Car_cat         -15.746488  -7.918205   .
Num_of_seats      7.950783   3.440846   .
Power_type      -17.074999  -7.582968   .
Transmission    110.528273  73.421300  47.97115
Wheel_drive     169.410478  95.335961   .
Animals_allowed  30.448813  29.190781  17.35365
Baby_pillow      39.968766  47.229866  55.55102
Child_seat       56.150190  33.986032   .
Towing_hitch    -90.986324 -45.867453   .
Roof_racks        7.060761  -7.905983   .
Roof_box        179.306152 125.404696  87.82070
```

Code Snippet 13: Comparison of model coefficients

In the table above, the 2nd column shows the coefficients using the ordinary least squares, and the 3rd column showing the coefficient estimates of the ridge regression. We notice here that the coefficients of the regressors are tending towards Zero for all but one of the regressors (roof_racks excluded).

When we examine the 4th column, we notice that all of the insignificant regressors have been tended to 0. Only the significant regressors – Transmission, Animals allowed, Baby Pillow and roof box are included in this model.

In order to measure the accuracy of these models, we run a prediction using the test data and run a cross validation across the 3 models.

```
> cbind(MSEOls, MSERigde,MSELasso)
       MSEOls MSERigde MSELasso
[1,] 24424.02 13312.39 14565.14
```

Code Snippet 14: Comparison of MSE for the Shrinkage models

We notice that the OLS has the largest Mean Square Error (24,424.02) which makes it the least accurate of these models. The Ridge regression on the other end seems to have produced the best prediction with a slightly lower Mean Square Error than the Lasso regression.

4.3.4 Decision Trees

Decision Trees is a method that involves dividing the prediction logic into several this or that components using the predicting variables. The mean of the predictied value is used as the response value based on the training observations. Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as decision tree methods (James et al., 2021).

Decision trees are used very commonly to solve classification and regression problems, however for the regression problems, the accuracy is hypothesized to not be as good as that of the shrinkage and linear regression methods applied earlier in this study. Another low point of decision tree regression is that in a bid to reduce the MSE, we can have an over fitting problem. The model can continue to split the data set to the point where there are too specific subsets. The way to tackle this problem is by specifying the minimum number of records that a node (subset) can have.

```
> treedata = tree(Usage_ratio_mins_day~.,data=train)
> summary(treedata)

Regression tree:
tree(formula = Usage_ratio_mins_day ~ ., data = train)
Variables actually used in tree construction:
[1] "Transmission" "Baby_pillow"  "Car_cat"      "Child_seat"   "Roof_racks"
[6] "Towing_hitch" "Power_type"
Number of terminal nodes:  9
Residual mean deviance:  12920 = 1537000 / 119
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-260.800  -78.070    1.426    0.000   69.880  250.800
```

Code Snippet 15: Regression Tree

In this regression model, the variables Transmission, Baby Pillow, Car_cat, Child_seat, Roof racks, Towing Hitch and Power type are the components of the selected regression model. These 11 regressors produce 9 terminal modes which are reflected in the tree table below.

```
> plot(treedata)
> text(treedata, pretty = 0)
```

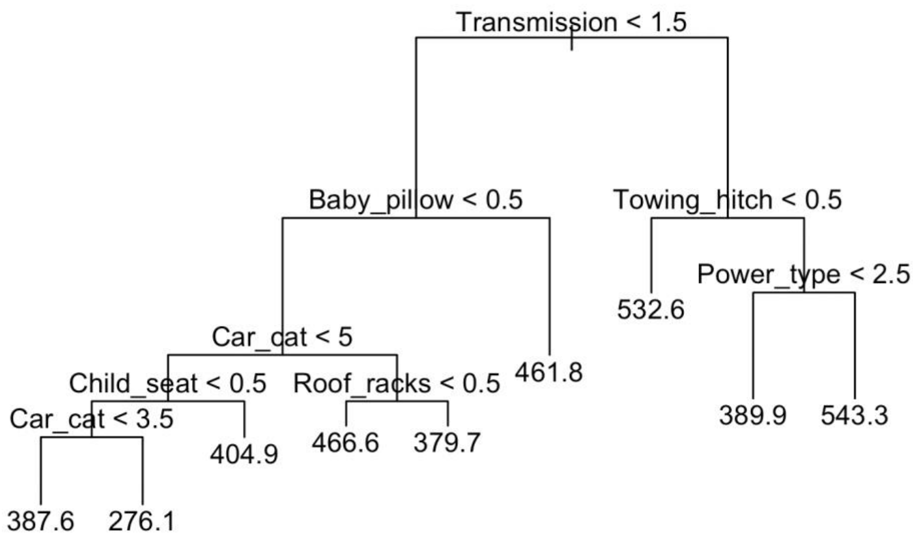

Figure 46: Output of Decision table regression

To ensure that the model has selected the optimum size of the decision tree, we check the plot of the possible tree sizes against the deviation.

```
> cvtreedata <- cv.tree(treedata)
> plot(cvtreedata$size , cvtreedata$dev, type = "b")
```

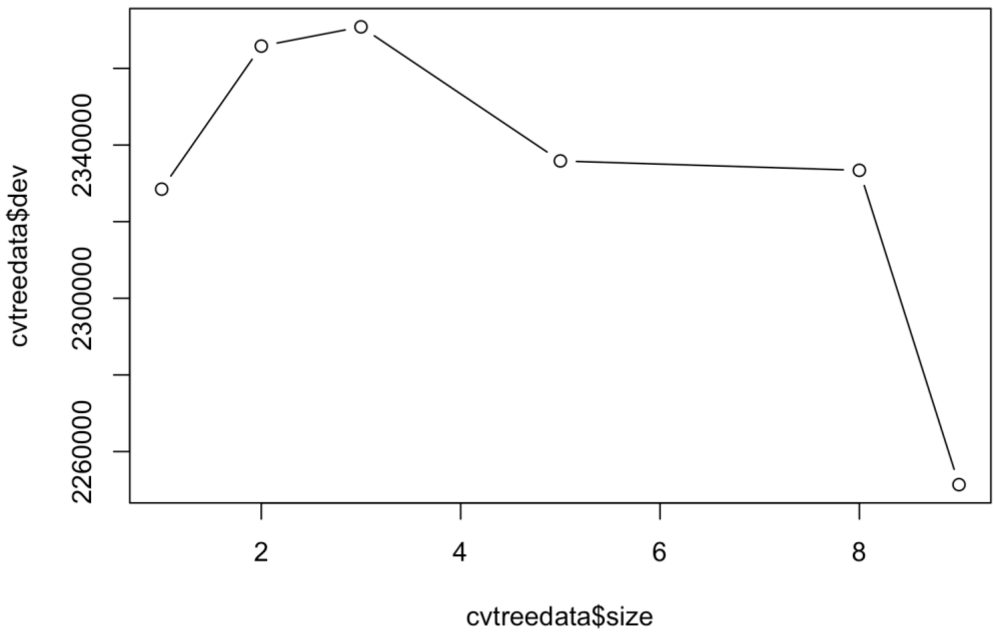

Figure 47: Plot of the deviation per tree size

From the plot above, we see that the model with tree size 9 which was used by the tree regression function has the lowest deviation. However, to form a basis for comparison, we check the next best size which is tree size = 8.

```
> prunetreedata = prune.tree(treedata, best = 8)
> summary(prunetreedata)

Regression tree:
snip.tree(tree = treedata, nodes = 9L)
Variables actually used in tree construction:
[1] "Transmission" "Baby_pillow"  "Car_cat"      "Child_seat"   "Towing_hitch"
[6] "Power_type"
Number of terminal nodes:  8
Residual mean deviance:  13000 = 1560000 / 120
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-260.800  -78.070    1.115    0.000   70.100  250.800
```

Code Snippet 16: Pruned decision tree

In the new pruned tree, the variable roof racks has been excluded from the mode. Hence this leaves the model with 8 terminal nodes. This is shown in the figure below.

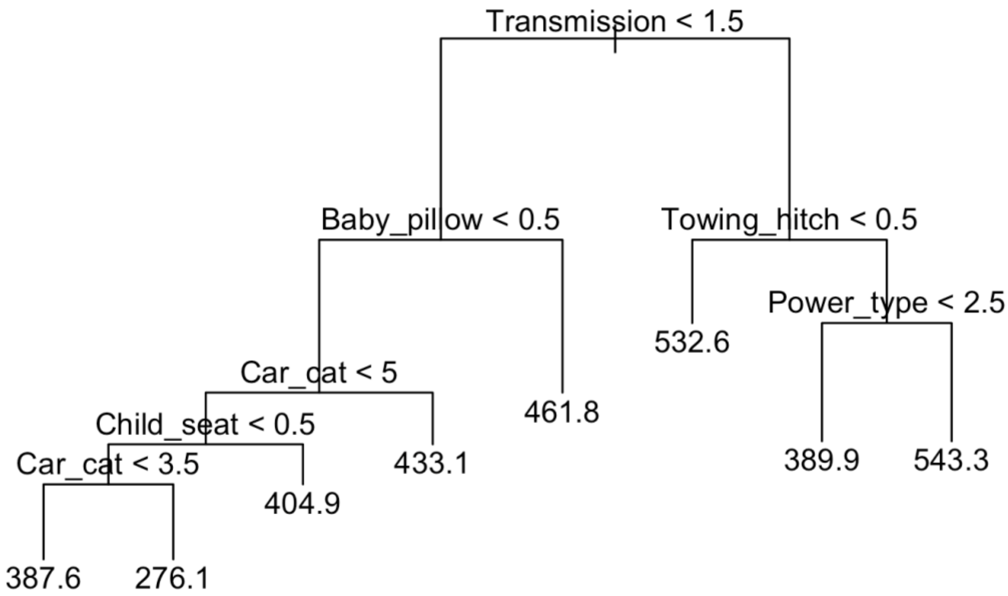

Figure 48: Output of pruned decision table regression

Next, we compare the accuracy of these 2 regression method using the cross validation to obtain the Mean Squared error of the 2 models.

```
     MSETree MSEPrtree
[1,] 12844.88  12471.96
```

Code Snippet 17: MSE calculation for Decision trees

It is noteworthy that the Pruned tree with 8 nodes performed better despite having a slight larger variance than the default tree selected by the regression function. This is a case of how the variance-bias trade-off can affect models in their out of training performance due to over fitting.

4.3.5 Boosting

Boosting is used as a technique to increase the prediction performance of regression trees. The difference here is that contrary to traditional regression trees that produce just one single model, the boosted regression trees combines a large number of simple regression trees to increase the prediction performance (Leathwick et al. 2006, 2008). This technique involves growing the trees in a sequential manner, making new trees based on previously generated trees. New trees are fit based on a modified version of the original data (James et al., 2021).

The boosted regression trees regression is executed using the gbm function in the gbm library and the boosting process is controlled by 3 parameters.

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Equation 8: Boosted regression

1. The number of trees: This is the size of the number of trees built in this boosted regression trees. This helps to control the risk of over fitting, which is likely if the number of trees is large. This is defined as the *B* in regression function and n.*trees* in the gbm function.

2. The shrinkage parameter: This is a small positive number defined as λ. This parameter slows the boosting process in a bid to help the model learn better. This controls the rate at which boosting learns. Smaller λ values require using a very large number of trees B in order to achieve good performance. This is represented by the *shrinkage* term in the gbm function.

3. Interaction Depth: This is the term that controls the complexity of the boosted trees.

```
> boostpred <- gbm(Usage_ratio_mins_day~.,data=train,
+                    distribution = "gaussian", n.trees = 5000,
+                    interaction.depth = 4)
> summary(boostpred)
                           var    rel.inf
Car_cat                Car_cat 23.5827743
Transmission      Transmission 19.9132393
Animals_allowed Animals_allowed 17.1936361
Power_type          Power_type 17.0929390
Towing_hitch      Towing_hitch  7.0670716
Num_of_seats      Num_of_seats  5.9647055
Baby_pillow        Baby_pillow  4.5388537
Child_seat          Child_seat  4.1331872
Roof_racks          Roof_racks  0.5135932
Wheel_drive        Wheel_drive  0.0000000
Roof_box              Roof_box  0.0000000
```

Code Snippet 18: Boosting prediction output

The summary of the boosted trees model shows that Car Category has the highest relative influence on the model. This is followed by the Transmission type, animals allowed and power type. The relative influence of variables is defined as the measure of the improvement made by each variable across all the trees that the variable is used. The comparison across all of the variables in the model is plotted graphically below for better insight.

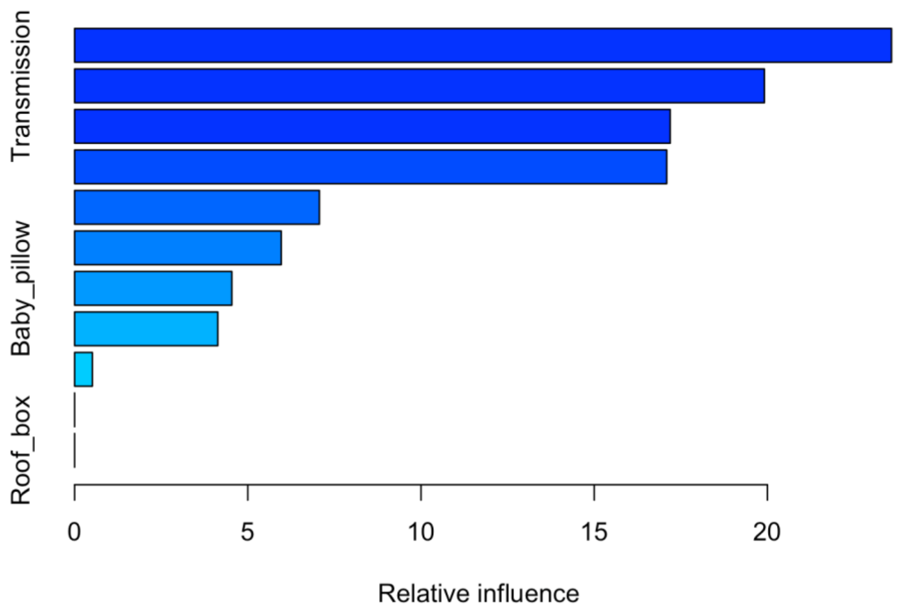

Figure 49: Relative influence of variables

We can also visually represent how the most relevant variables in the boosted trees regression affect the model. We do so below for the first 3 variables - car category, transmission type, and power type.

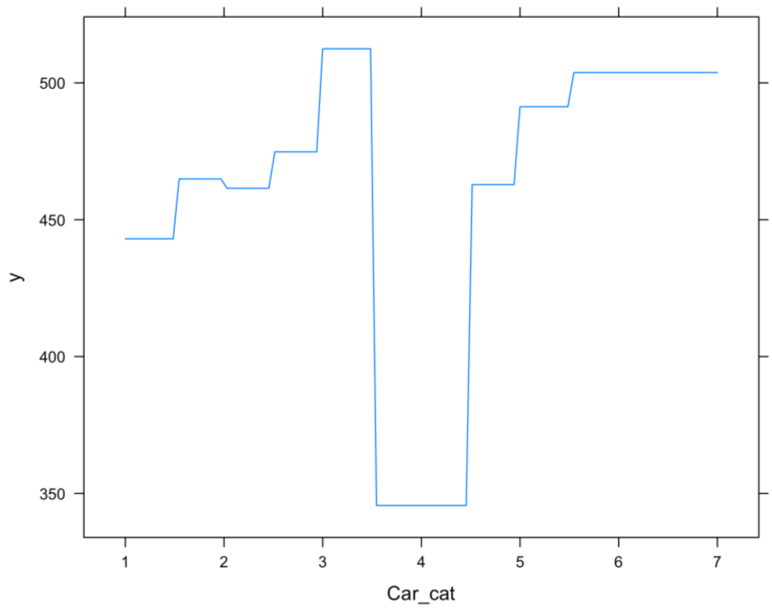

Figure 50: Car category impact

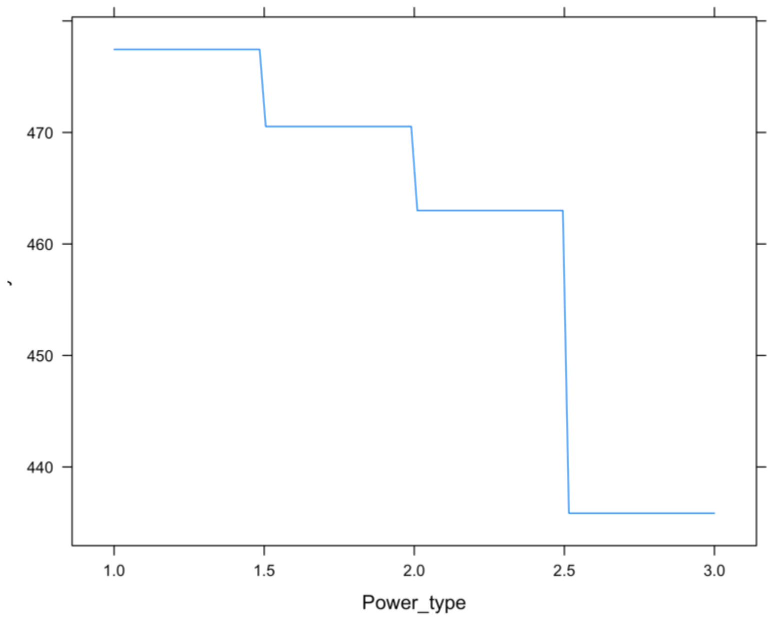

Figure 51: Power type impact

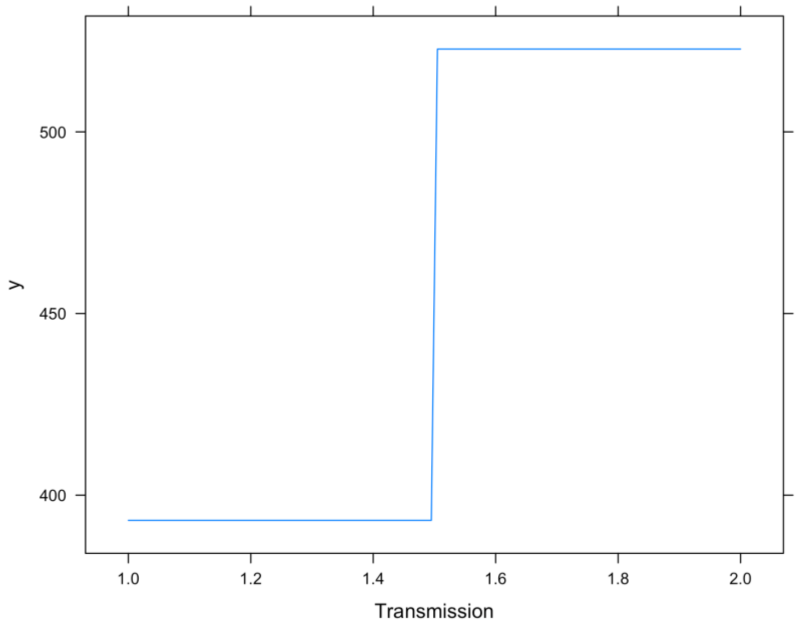

Figure 52: Transmission Impact

In the boosted trees regression above, we did not specify the shrinkage value to control the regression to avoid overfitting. Whenever the coefficient is not specified; the function automatically assigns 0.01. In order to demonstrate the effect of this variable, we will alter the regression function to include a shrinkage value. Here we use the shrinkage value = 0.01 while keeping the number of trees as 5,000 and the interaction depth as 4.

```
> boost2pred <- gbm(Usage_ratio_mins_day~.,data=train,
+                    distribution = "gaussian", n.trees = 5000,
+                    interaction.depth = 4, shrinkage = 0.01, verbose = F)
> summary(boost2pred)
                           var    rel.inf
Car_cat                Car_cat 25.8644504
Transmission      Transmission 19.4090730
Power_type          Power_type 15.9487593
Animals_allowed Animals_allowed 15.8806398
Towing_hitch      Towing_hitch  7.4201528
Num_of_seats      Num_of_seats  5.3692215
Child_seat          Child_seat  5.0197974
Baby_pillow        Baby_pillow  4.7241634
Roof_racks          Roof_racks  0.3637424
Wheel_drive        Wheel_drive  0.0000000
Roof_box              Roof_box  0.0000000
```

Code Snippet 19: Shrinked boosting

The summary of the regression model shows very similar results in terms of the order of variables by relative influence, except for power_type and animals_allowed which have switched positions. It is however noteworthy that the difference in the relative influence for these two variables is negligible across the two models.

Furthermore, we compare the coefficients of the two models (the models with and without the shrinkage coefficient specified) in the table below.

```
> cbind.data.frame(summary(boost2pred),summary(boostpred))
                            var    rel.inf             var    rel.inf
Car_cat                 Car_cat 25.8644504         Car_cat 23.5827743
Transmission       Transmission 19.4090730    Transmission 19.9132393
Power_type           Power_type 15.9487593 Animals_allowed 17.1936361
Animals_allowed Animals_allowed 15.8806398      Power_type 17.0929390
Towing_hitch       Towing_hitch  7.4201528    Towing_hitch  7.0670716
Num_of_seats       Num_of_seats  5.3692215    Num_of_seats  5.9647055
Child_seat           Child_seat  5.0197974     Baby_pillow  4.5388537
Baby_pillow         Baby_pillow  4.7241634      Child_seat  4.1331872
Roof_racks           Roof_racks  0.3637424      Roof_racks  0.5135932
Wheel_drive         Wheel_drive  0.0000000     Wheel_drive  0.0000000
Roof_box               Roof_box  0.0000000        Roof_box  0.0000000
```

Code Snippet 20: Relative influence of variables for Boosting

In the next step, we compare the performance of the two models on the test data to check for predictive performance of the models.

```
> cbind(MSEBoost,MSEShrink)
     MSEBoost MSEShrink
[1,] 14644.98  13491.49
```

Code Snippet 21: MSE calculation for Boosting

We see here that reducing the shrinkage value optimized the performance of the boosting model by reducing the shrinkage value to 0.01. The MSE for the new model is significantly lower than the older one.

4.3.6 Bagging

Bagging is a technique that is also very similar to boosting, the primary difference being that bagging continuously fits new trees based on the initial data while boosting fits new trees based on modified data based on the previously fitted trees. The fitting is performed on different subsets of training data, generated randomly with replacement. An average of all the derived decision trees is then used in fitting the model, which generally provides more accuracy than single decision trees that are generally believed to have high variance (James et al., 2021).

The major advantage of Bagging over Boosting regression trees is that it does not have the tendency to overfit the model. The average of the predictions obtained from the trees outputted by the different models is used in this model which is hypothesised to be better. The major disadvantage cited for bagging methods is that it does not give precise values for the models owing to the fact that the final predictions are based on the mean prediction from the component subset trees (Garg, 2018). Also, the general disadvantage of ensemble methods (bagging and boosting) is that they cannot be interpreted in detail as simple as single decision trees.

To summarize the differences between dingle regression trees, bagging regression trees and boosting regression trees, an infographic can be used as retrieved from Quantdare (2016).

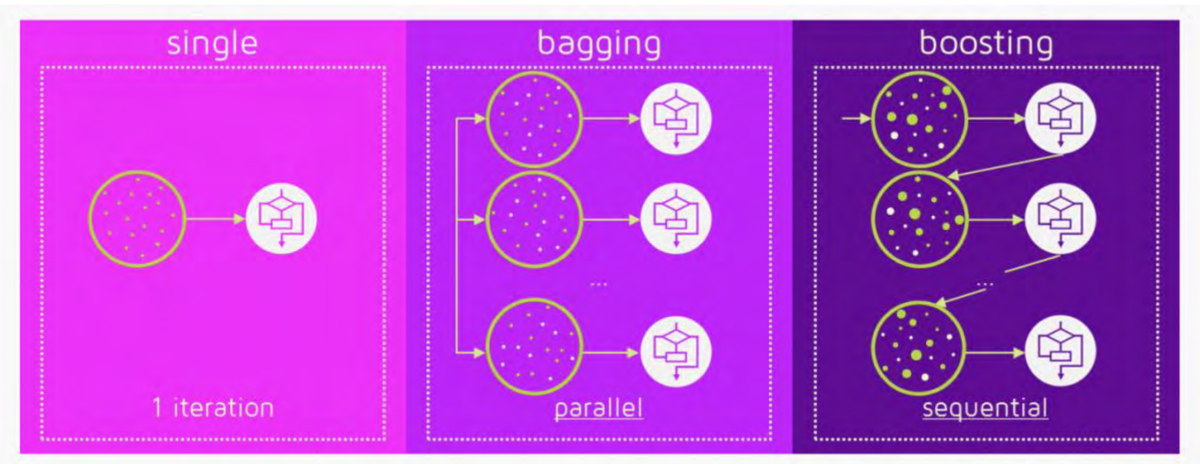

Figure 53: Comparison of data selection

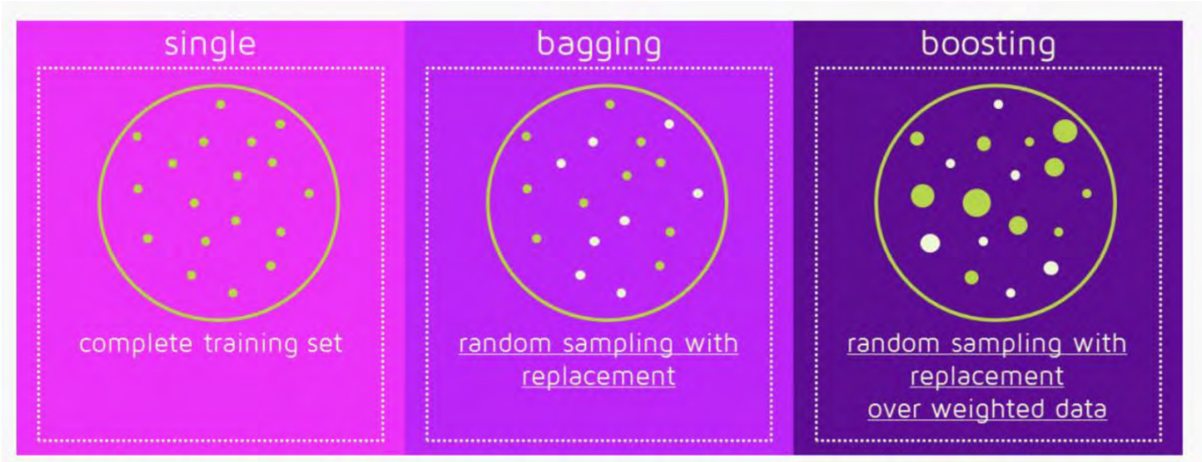

Figure 54: Comparison of set sampling

Random forests

Random forests is a variation of bagging regression trees that randomly selects subsets of features in the different subsets of data used to generate the individual regression trees. In summary, Random Forest uses bootstrap resampling (same as the bagging technique) and random variable selection. It is popular practice to specify the number of variables selected (m) as the square root of the total number of variables (p). In this model, the individual splits can only have and use the m number of variables in fitting the subtree (James et al., 2021).

Bühlmann (2011) hypothesises that random forests gives better predictions than the original bagging regression technique. This is achieved by decorrelating the trees, ensuring that the strong regressors do not have a too strong effect on the prediction derived by averaging the prediction from the sub trees. This is in most circumstances leads to a reduction in the Mean Squared Error of the regression model.

The random forest and bagging methods are executed in R using the randomForest function in the randomForest library. The two methods use the same function, the difference being the number of variables specified to be used in the model. The number of the variables is specified using the "mtry" term. For Bagging, we used the total number of predictor variables, which is 11.

```
> Bagpred = randomForest(Usage_ratio_mins_day~.,data=train ,
+                 mtry = 11, importance = TRUE)
> Bagpred

Call:
 randomForest(formula = Usage_ratio_mins_day ~ ., data = train,      mtry = 11, importan
ce = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 11

          Mean of squared residuals: 15853.26
                    % Var explained: 6.73
```

Code Snippet 22: Bagging

Unlike single regression trees, we cannot model the trees to interpret how the different predictors have an effect on our final model. However, we can investigate the relevance of the different predictors in the final average model. This is done using the varImpplot () function which outputs all of the regressors mapped against two indicators - %incMSE and incnodepurity.

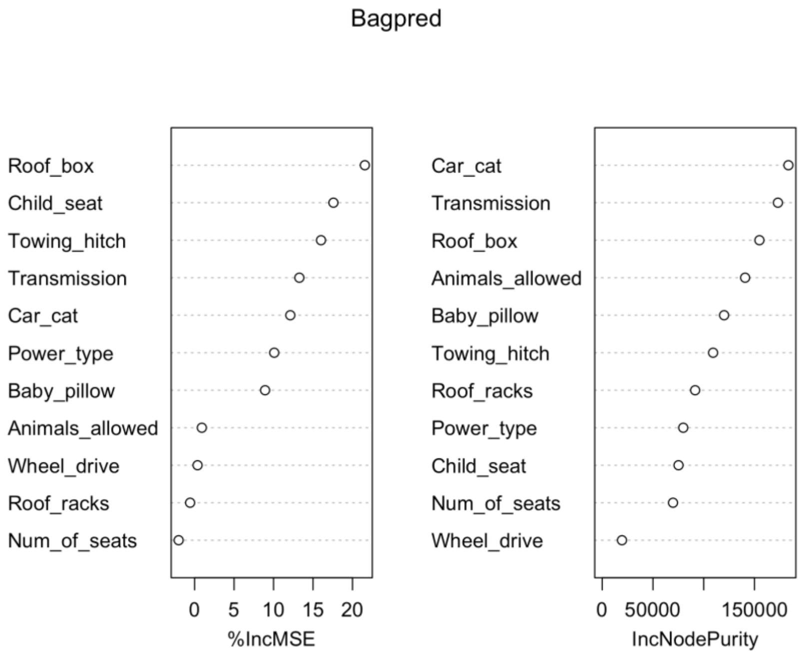

Figure 55: Variable importance plot

To run the random forest model, we make an alteration in the regression function, specifying the maximum size of the subset of variables that can be used by each sub model is 4. We have chosen 4 which is the estimate of the square root of 11 (the total number of predictor variables).

```
> forestpred = randomForest(Usage_ratio_mins_day~.,data=train ,
+                         mtry = 4, importance = TRUE)
> forestpred

Call:
 randomForest(formula = Usage_ratio_mins_day ~ ., data = train,      mtry = 4, importanc
e = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 15067.51
                    % Var explained: 11.35
```

Code Snippet 23: Random forest

Next, we also investigate the importance of the variables in the aggregate regression model achieved by the random forest tree regression.

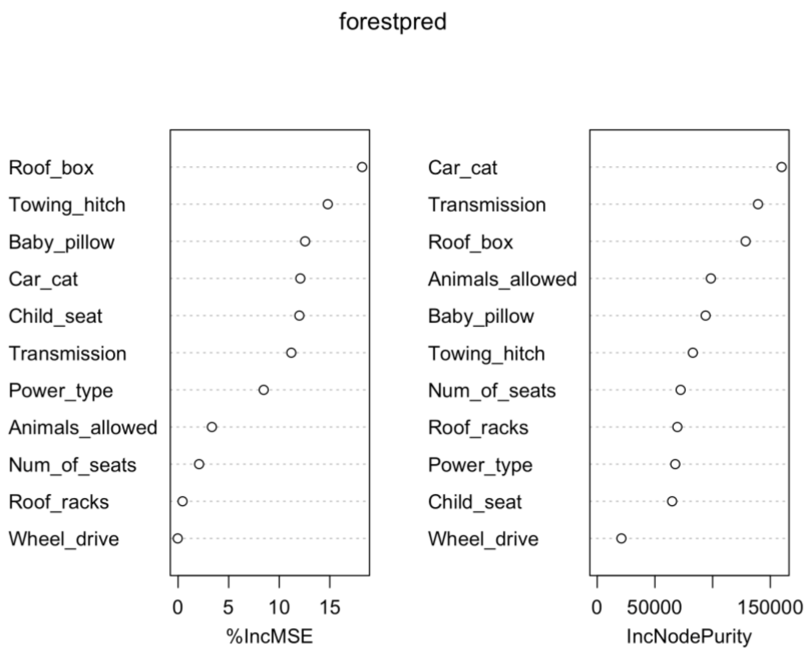

Figure 56: Variable importance for Random forest

The most important variables here are Roof_box, Towing_hitch, baby_pillow, car_cat and transmission.

The next step is to check if the use of the variable subset selection had an effect on our prediction. We do this by calculating the mean squared error of the two models using the reserved test data.

```
       MSEBag    MSERF
[1,] 14536.26 13743.13
```

Code Snippet 24: MSE calculation for Bagging techniques

The random forest methos has a lower MSE than the ordinary bagging method. This shows that this method helped address the overfitting problem by limiting the influence of the strong predictors on the final model.

5. RESULT ANALYSIS

In order to determine the best prediction model for predicting the car usage based on the car-specific features for Bildeleringen, we ran 10 different prediction models across 5 techniques (linear regression, shrinkage, decision trees, boosting and bagging).

The results of these models are summarised in the table below, highlighting the number of variables used in the model and the corresponding test mean squared error which is obtained by cross validation of the predicted usage against the holdout data (test data). We will select the best model using the test MSE and reference the number of variables to check if there is a trend in the number of variables and the predictive performance.

	Model Name	# of Variables	TestMSE
1.	Multiple Regression	11	13027.07
2.	Best Subset Selection	8	12924.7
3.	Ridge Shrinkage	11	13312.39
4.	Lasso Shrinkage	4	14565.14
5.	Decision Trees	7	12844.88
6.	Pruned Decision Tree	6	12471.96
7.	Boosted Tree	11	14644.98
8.	Boosted Tree w/ shrinkage	11	13491.49
9.	Bagging	11	14536.26
10.	Random Forest	11	13743.13

Table 5 : List of models with indicators

Ranking the results of the models based on the test mean squared error (testMSE), we deduce that the pruned decision tree with 6 variables is the model with the best performing model. The next best model is the original decision tree with 7 variables. There is only a marginal difference in the performance of these models (3%).

To further visualize the analysis of the results from the prediction models, we compare the testMSEs across the models to the number of predictive variables in the model.

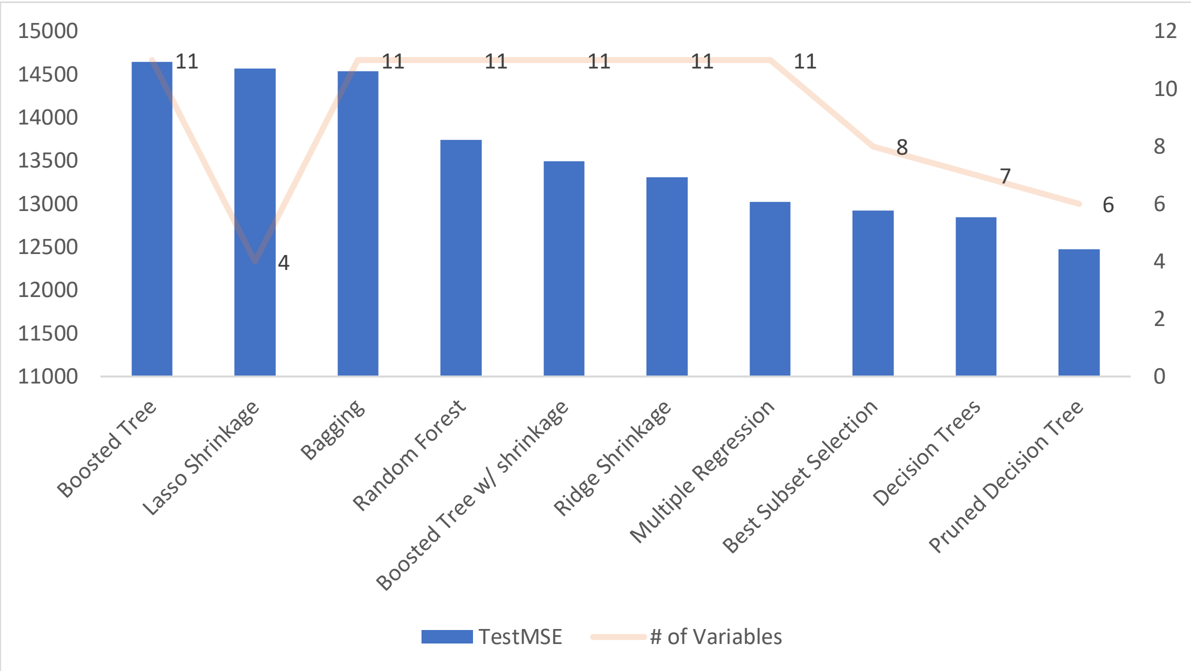

Figure 57: Plot of Models with indicators

From the plot above, we see that the test MSE reduces with number of variables in the regression model, suggesting that the presence of other variables creates some noise in the model. This in turn leads to greater difference in the prediction and the actual observed figures due to reducible error.

It is also noteworthy that the lasso shrinkage method which reduced the coefficients of all the regression variables except 4 of them did not follow this trend. Our hypothesis here is that this method performed poorly as it ignored some predictor variables that are causal to the usage variable.

We also notice that performance ranking is also in clusters based on the technique. We see that the decision tree models perform best followed by the linear regression models (multiple regression and subset selection). The decision tree ensembles methods (bagging, boosting and random forests) and regularization methods (lasso regression and ridge regression) belong to the least efficient cluster based on the MSE.

We hypothesize that the regularization methods – Lasso regression and Ridge regressions have not worked for our data as we do not have a high level of collinearity in our data set amongst the prediction variables. The shrinkage methods are used to address the problems caused by multicollinearity in datasets by introduce a bias to negate the variance.

To confirm this hypothesis, we use the vif () function from the car library to test for multicollinearity using Variance Inflation Factors (VIF). The Variance Inflation factors indicates if there is any correlation between the prediction variables. The results are plotted with all VIF values less than 5 indicating there is no severe collinearity present.

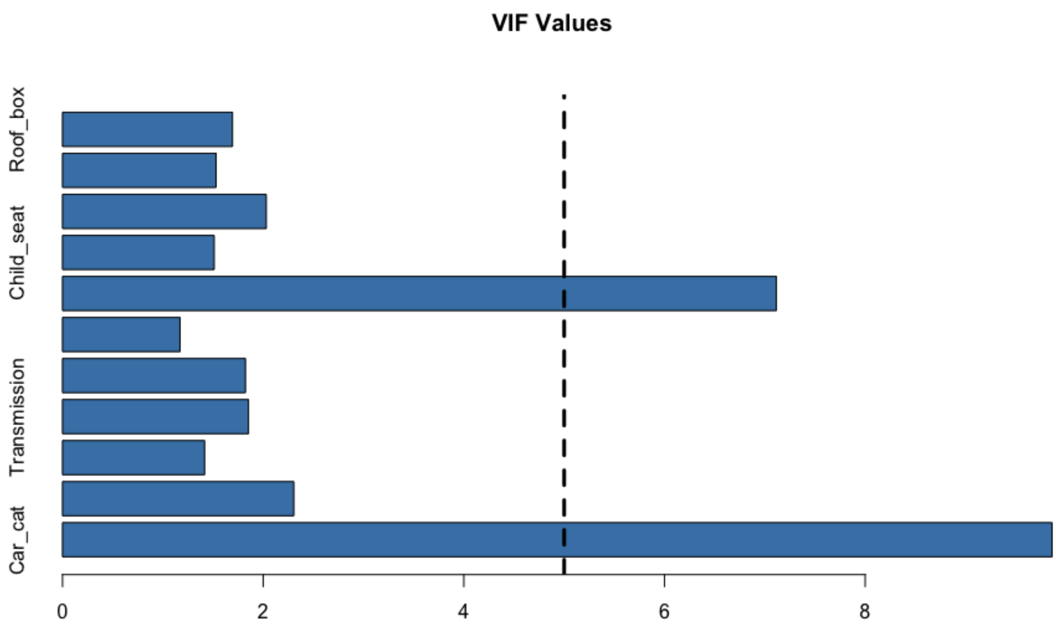

Figure 58 : Plot of variables against the VIF

This function shows that severe multicollinearity was recorded in just two variables – car_cat and baby_pillow. We hypothesize that a lot of car specific features are dependent on the car category, and that explains the critical multicollinearity present in this variable. For baby_pillow, we do not have a strong hypothesis on why multicollinearity exists here.

On the other hand, the Ensemble Methods - Bagging and Boosting, focus on reducing either the variance or bias of the models. We hypothesize that the technique might not have had the best prediction performance because these methods have not worked optimally on our dataset because it does not have high variance or high bias.

Finally, we hypothesize that the relationship between the independent and dependent variables is non-linear and complex. This is the reason why the decision tree outperforms the multiple linear regression model.

6. CONCLUSION

After our analysis and model formulation in the preceding chapters, we have discovered that the best model to predict the future usage of cars is the Decision trees. This is after conducting predictions using linear regression, shrinkage, decision trees, boosting, and bagging. We have then compared them using cross-validation of the predicted usage against the holdout data (test data), we have found that the pruned decision tree with 6 variables and a test mean squared error of 12471.96 is the best (it has the lowest MSE).

The 6 significant and relevant variables in this model are transmission, baby_pillow, car_cat, child_seat, towing_hitch, and power_type. Hence these are the features that determine the amount of appeal that are car has to potential car users on the Bildeleringen platform. The usage (measured by Usage_ratio_mins_day) is maximized if a car is large (large station wagon or van), has an automatic transmission, does not have a baby pillow, has a towing hitch, and uses gasoline.

we have found that there is severe multicollinearity in the variables car_cat and baby_pillow. The multicollinearity is most likely the outcome that car features are dependent on the car_cat. This multicollinearity was something that we did pay attention to throughout the analysis.

Recommendations to company

Bildeleringen can provide more small cars with automatic transmissions. Also, when displaying the type of vehicles available for rental on their website, Bildeleringen can highlight the accessories offered such as towing hitch, baby pillow, and child seat are available. Currently, Bildeleringen is operated in a two-way system, which means customers have to rent and return the vehicle in the same location. As the car-sharing market is increasingly competitive, Bildeleringen may consider operating in a free-floating system, in which the members can return the rented vehicles to any parking to improve the convenience. Lastly, Bildeleringen should consider collecting more information on location and customer characteristics to improve the analysis of the managerial decisions of the company.

Study limitations

This study has potential limitations. First, we only considered one-year data (the year 2020), which is two years from the time we performed this study. Car-sharing is a rapidly growing market, and the demand could have been affected by macro-environment such as demographic, physical, natural, economic, technological, political, legal, and socio-cultural conditions. For example, coronavirus in 2020 has seriously affected the travel and tourism industry, and as a result, has significantly affected the demand for car-sharing, especially among travellers or tourists. Second, the demand for car-sharing may be affected by seasonality, location, and the characteristics of car-sharing members, which we did not manage and were able to obtain relevant data. Lastly, it is superior to predict the demand for each car category on each day or on an hourly basis. However, due to insufficient data, we are only able to predict the demand in terms of Usage_ratio_mins_day.

Cases for future study

For future studies, it is nice to gather recent data across locations with more than one year to improve the accuracy of predictions of the car-sharing demand. For example, there was research studying people's usage patterns and service demand for car-sharing systems (Alencar et al., 2021). The research evaluated three car-sharing types, which were available in Vancouver, Canada, and the surrounding urban region (Alencar et al., 2021). The analysis was based on data collected over the course of a year from the Modo, Evo, and Car2Go car-sharing services — a two-way, a one-way, and a free-floating service, respectively (Alencar et al., 2021). The research studied the demand and usage patterns of vehicles from these different types of car-sharing services and provided contributions twofold: (1) characterized three major car-sharing paradigms, and (2) modelled the demand for vehicles, providing statistical distributions that describe their busy and idle periods (Alencar et al., 2021). This study is helpful in highlighting specific situations where car-sharing services are appealing and in uncovering trends and mobility patterns when combined with data from other forms of transportation (Alencar et al., 2021).

Table of Figures

References

References

Alencar, V. A., Rooke, F., Cocca, M., Vassio, L., Almeida, J., & Vieira, A. B. (2021). Characterizing client usage patterns and service demand for car-sharing systems. *Information Systems*, *98*, 101448. https://doi.org/10.1016/j.is.2019.101448

Andreis, Federico. (2017). Shrinkage methods (ridge, lasso, elastic nets).

Automotive roof rack market size, growth & forecast [2028]. (2020). Retrieved from Fortune Business Insights website: https://www.fortunebusinessinsights.com/automotive-roof-racks-market-103569

Bhutani, A., & Bhardwaj, P. (2018, April). Car sharing market size by model (P2P, station-based, free-floating), by business model (round trip, one way), by application (business, private), industry analysis report, regional outlook (U.S., canada, UK, germany, france, italy, spain, russia, turkey, china, india, japan, south korea, taiwan, malaysia, singapore, australia, brazil, mexico, GCC, south africa), growth potential, competitive market share & forecast, 2018 – 2024. Retrieved from Gminsights.com website: https://www.gminsights.com/industry-analysis/carsharing-market

Bildeleringen: Norges smarteste bilkollektiv. (2020). Retrieved from Bildeleringen website: https://bildeleringen.no/

Boldrini, C., Bruno, R., & Conti, M. (2016). *Characterising Demand and Usage Patterns in a Large Station-based Car Sharing System*.

Bühlmann, P. (2011). Bagging, Boosting and Ensemble Methods. *Handbook of Computational Statistics*, 985–1022. https://doi.org/10.1007/978-3-642-21551-3_33

Burkhardt, J. E., & Millard-Ball, A. (2006). Who is attracted to carsharing? *Transportation Research Record: Journal of the Transportation Research Board*, *1986*, 98–105. https://doi.org/10.1177/0361198106198600113

Car hire price comparison | car booker. (2021). Retrieved from car-booker.com website: https://car-booker.com/uk/#/searchcars

CarSharing association code of ethics and standards of practice. (2011, January). Retrieved May 2022, from https://carsharing.org/wp-content/uploads/2011/02/CarSharingAssociation_CodeofEthics.pdf

Chun, Y.-Y., Matsumoto, M., Tahara, K., Chinen, K., & Endo, H. (2019). Exploring factors affecting car sharing use intention in the southeast-asia region: A case study in java, indonesia. *Sustainability*, *11*, 5103. https://doi.org/10.3390/su11185103

Clewlow, R. R. (2016). Carsharing and sustainable travel behavior: Results from the san francisco bay area. *Transport Policy*, *51*, 158–164. https://doi.org/10.1016/j.tranpol.2016.01.013

Copas, J. B. (1993). The Shrinkage of Point Scoring Methods. *Applied Statistics*, *42*(2), 315. https://doi.org/10.2307/2986235

Cyriac, G., & Julsrud, E. (2018). *The development of organised car sharing in Norway: 1995-2018: historical trends and potential impacts*. Transportøkonomisk Institutt (TØI).

Efthymiou, D., Antoniou, C., & Waddell, P. (2013). Factors affecting the adoption of vehicle sharing systems by young drivers. *Transport Policy*, *29*, 64–73. https://doi.org/10.1016/j.tranpol.2013.04.009

ETSC | european transport safety council. (2021). Retrieved from etsc.eu website: https://etsc.eu/

Fang, J., Cesar Góez, J., & Guajardo, M. (2020). *Analysing the demand for car- sharing in Bergen An empirical approach using car-specific and demographic data*.

Febbraro, D., Sacco, N., & Saeednia, M. (2012). One-way carsharing. *Transportation Research Record: Journal of the Transportation Research Board*, *2319*, 113–120. https://doi.org/10.3141/2319-13

Fujii, S., & Kitamura, R. (2003). What does a one-month free bus ticket do to habitual drivers? An experimental analysis of habit and attitude change. *Transportation*, *30*, 81–95. https://doi.org/10.1023/a:1021234607980

Garg, R. (2018, February 19). A Primer to Ensemble Learning – Bagging and Boosting. Retrieved January 18, 2021, from Analytics India Magazine website: https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/#:~:text=Bagging%20is%20a%20way%20to

Getting a driving licence in the EU. (2019). Retrieved from Your Europe - Citizens website: https://europa.eu/youreurope/citizens/vehicles/driving-licence/get-driving-licence/index_en.htm

Giansoldati, M., Danielis, R., Rotaris, L., & Scorrano, M. (2018). The role of driving range in consumers' purchasing decision for electric cars in Italy. *Energy*, *165*, 267–274. https://doi.org/10.1016/j.energy.2018.09.095

Glerum, A., Stankovikj, L., Thémans, M., & Bierlaire, M. (2014). Forecasting the demand for electric vehicles: Accounting for attitudes and perceptions. *Transportation Science*, *48*, 483–499. https://doi.org/10.1287/trsc.2013.0487

Hartl, B., & Hofmann, E. (2021). The social dilemma of car sharing – The impact of power and the role of trust in community car sharing. *International Journal of Sustainable Transportation*, 1–24. https://doi.org/10.1080/15568318.2021.1912224

HealthyChildren.org - from the american academy of pediatrics. (2019). Retrieved from HealthyChildren.org website: https://www.healthychildren.org

Heinrichs, M., Krajzewicz, D., Cyganski, R., & von Schmidt, A. (2017). Introduction of car sharing into existing car fleets in microscopic travel demand modelling. *Personal and Ubiquitous Computing*, *21*(6), 1055–1065. https://doi.org/10.1007/s00779-017-1031-3

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *INTRODUCTION TO STATISTICAL LEARNING: with applications in r.* S.L.: Springer-Verlag New York. Retrieved from https://www.statlearning.com/

Kikuchi, R., & Miwa, H. (2021). Dynamic pricing method for one-way car sharing service to meet demand and to maximize profit under given utility function. *Advances in Internet, Data and Web Technologies*, 324–332. https://doi.org/10.1007/978-3-030-70639-5_30

Leathwick, J., Elith, J., Francis, M., Hastie, T., & Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, *321*, 267–281. https://doi.org/10.3354/meps321267

Loveday, S. (2022, February). U.S. news & world report. Retrieved from Usnews.com website: https://cars.usnews.com

Marsden, G. (2022). Social benefits of shared mobility: metrics and methodologies 26 th ACEA SAG REPORT. Retrieved from https://www.acea.auto/files/Social_benefits_of_shared_mobility-26th_ACEA_SAG-report.pdf

McCaffery, J. (2022, March). Why do americans drive Automatic—But most of europe drives manual? Retrieved from Reader’s Digest website: https://www.rd.com/article/why-americans-drive-automatic/

Millard-Ball, A., Murray, G., Schure, T., & Fox, C. (2005). *Car-sharing : where and how it succeeds*. Transportation Research Board Of The National Academies.

New passenger cars by segment in the EU. (2021, February). Retrieved from ACEA - European Automobile Manufacturers' Association website: https://www.acea.auto/figure/new-passenger-cars-by-segment-in-eu/

Nijland, H., & Meerkerk. (2017). Mobility and environmental impacts of car sharing in the Netherlands. *Environmental Innovation and Societal Transitions*, *23*, 84–91. https://doi.org/10.1016/j.eist.2017.02.001

Quantdare. (2016, April 20). What is the difference between Bagging and Boosting? ⋆ Quantdare. Retrieved from Quantdare website: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, *132*, 88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005

Rooke, F., Aquiles, V., Vieira, A. B., Almeida, J. M., & Drago, I. (2019). Characterizing usage patterns and service demand of a two-way car-sharing system. *Communications in Computer and Information Science*, 3–17. https://doi.org/10.1007/978-3-030-11238-7_1

Schaefers, T. (2013). Exploring carsharing usage motives: A hierarchical means-end chain analysis. *Transportation Research Part A: Policy and Practice*, *47*, 69–77. https://doi.org/10.1016/j.tra.2012.10.024

Schaefers, T., Lawson, S. J., & Kukar-Kinney, M. (2015). How the burdens of ownership promote consumer usage of access-based services. *Marketing Letters*, *27*, 569–577. https://doi.org/10.1007/s11002-015-9366-x

Schmöller, S., & Bogenberger, K. (2014). Analyzing External Factors on the Spatial and Temporal Demand of Car Sharing Systems. *Procedia - Social and Behavioral Sciences*, *111*, 8–17. https://doi.org/10.1016/j.sbspro.2014.01.033

Trygg trafikk • trafikksikkerhet for alle - barn, unge og voksne. (2010, February). Retrieved from Trygg Trafikk website: https://www.tryggtrafikk.no

Tsuji, K., Kurisu, K., Nakatani, J., & Moriguchi, Y. (2020). Evaluation of environmental impact of car sharing in consideration of uncertainty of influential variables. *International Journal of Automation Technology*, *14*, 975–983. https://doi.org/10.20965/ijat.2020.p0975

Wagner, C., & Shaheen, S. (1998). Car sharing and mobility management: Facing new challenges with technology and innovative business planning. *The Journal of World Transport Policy and Practice*, *4*. Retrieved from https://escholarship.org/uc/item/37n6c9mg

Wang, N., Guo, J., Liu, X., & Fang, T. (2020). A service demand forecasting model for one-way electric car-sharing systems combining long short-term memory networks with Granger causality test. *Journal of Cleaner Production*, *244*, 118812. https://doi.org/10.1016/j.jclepro.2019.118812

Yu, D., Li, Z., Zhong, Q., Ai, Y., & Chen, W. (2020). Demand Management of Station-Based Car Sharing System Based on Deep Learning Forecasting. *Journal of Advanced Transportation*, *2020*, 1–15. https://doi.org/10.1155/2020/8935857

Zhang, Y., Qian, Z. S., Sprei, F., & Li, B. (2016). The impact of car specifications, prices, and incentives for battery electric vehicles in Norway: Choices of heterogeneous consumers. *Transportation Research Part C: Emerging Technologies*, *69*, 386–401. https://doi.org/10.1016/j.trc.2016.06.014

Zhang, Y., Zhong, M., Geng, N., & Jiang, Y. (2017). Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China. *PLOS ONE*, *12*, e0176729. https://doi.org/10.1371/journal.pone.0176729