German used car market analysis

CS498 – Special Topics in Computer Science I
Computing with Data

Jahja Muratović

Computer Science and Software Engineering Department
Faculty of Engineering and Natural Sciences (FENS), International University of Sarajevo (IUS)

Sarajevo, January 2021

Table of Contents

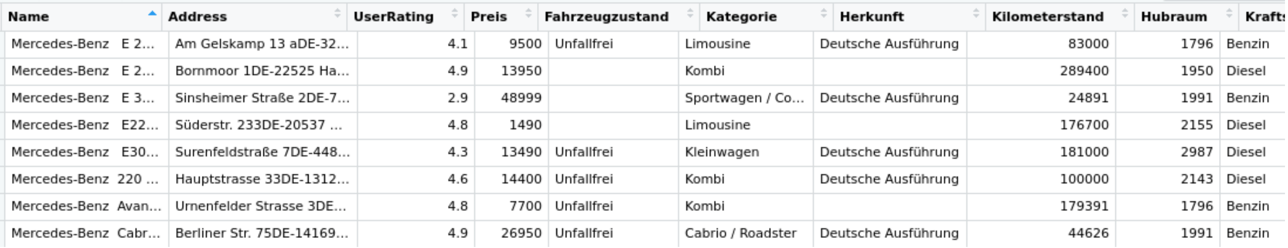
Introduction

The Mercedes-Benz was at a lot of points one of the automotive pioneers in its history (ABS and Airbag). Whatever new technology they put in the S-class, it finds its way to the E-class through the next generation. Since I have a lot of anecdotal experience with the E-class and since there is around 2700 E-classes available on the Bosnian used car market, that would be a very relatable car model to do data analysis on. This report is mainly based upon the mobile.de dataset, it contains around 23 000 rows and around 140 columns. It contains all the Mercedes E-Classes that were available on the night of 22nd December. Another dataset used was the olx.ba dataset which was scraped the same way as mobile.de dataset. The analysis tackles some points I personally found very interesting like difference between diesel and gasoline burning E-classes, differences between all-wheel-drive and rear-wheel-drive models, some comparisons between the Bosnian and German markets in terms of price, mileage and fuel used. I have also made a couple of maps of Germany which depicts presence of different E-class models mapped according to the postal area they were being sold in. I have also made some analysis on the Taxi models and what differs them from the rest and for which production years are they most numerous. The code samples and snippets can be found in the appendix.

Data Description

The data consists of all the E-classes available on mobile.de and olx.ba sites. The analysis is going to be done on both and based upon that, comparisons will be made between German and Bosnian car markets. The data was scraped using R’s rvest library then saved , cleaned, processed in R and saved in a csv file.

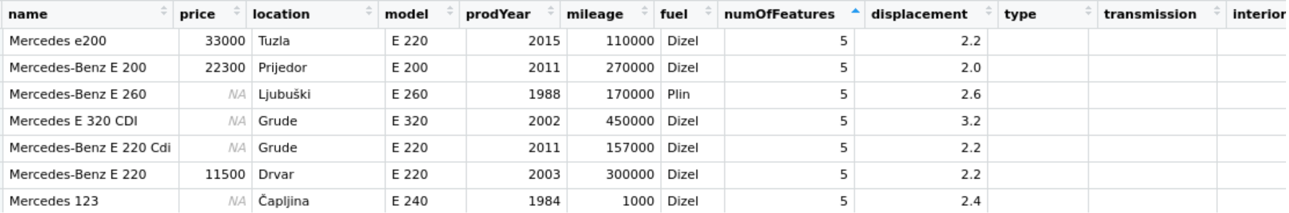
The data scraped from Mobile.de was very regular and didn’t require much work cleaning and processing it, and was also very detailed and reliable.Mobile.de dataframe has 23 675 rows and 139 columns, here is an example of the data:



Name	Address	UserRating	Preis	Fahrzeugzustand	Kategorie	Herkunft	Kilometerstand	Hubraum	Krafts
Mercedes-Benz E 2...	Am Gelskamp 13 aDE-32...	4.1	9500	Unfallfrei	Limousine	Deutsche Ausführung	83000	1796	Benzin
Mercedes-Benz E 2...	Bornmoor 1DE-22525 Ha...	4.9	13950		Kombi		289400	1950	Diesel
Mercedes-Benz E 3...	Sinsheimer Straße 2DE-7...	2.9	48999		Sportwagen / Co...	Deutsche Ausführung	24891	1991	Benzin
Mercedes-Benz E22...	Süderstr. 233DE-20537 ...	4.8	1490		Limousine		176700	2155	Diesel
Mercedes-Benz E30...	Surenfeldstraße 7DE-448...	4.3	13490	Unfallfrei	Kleinwagen	Deutsche Ausführung	181000	2987	Diesel
Mercedes-Benz 220 ...	Hauptstrasse 33DE-1312...	4.6	14400	Unfallfrei	Kombi	Deutsche Ausführung	100000	2143	Diesel
Mercedes-Benz Avan...	Urnenfelder Strasse 3DE...	4.8	7700	Unfallfrei	Kombi		179391	1796	Benzin
Mercedes-Benz Cabr...	Berliner Str. 75DE-14169...	4.9	26950	Unfallfrei	Cabrio / Roadster	Deutsche Ausführung	44626	1991	Benzin

Figure 1: Mobile.de dataframe

Olx data on the other hand, was a mess and it required a lot of cleaning and processing, I would say that mobile.de is better for getting data simply by design since it doesn’t allow characters in ex. Horspower or Mileage form field whereas olx allows it, so it spits out a lot of weird results when trying to be visualized. Olx dataframe has 18 columns and 2763 rows, here is an example of the data:



name	price	location	model	prodYear	mileage	fuel	numOfFeatures	displacement	type	transmission	interior
Mercedes e200	33000	Tuzla	E 220	2015	110000	Dizel	5	2.2			
Mercedes-Benz E 200	22300	Prijedor	E 200	2011	270000	Dizel	5	2.0			
Mercedes-Benz E 260	NA	Ljubuški	E 260	1988	170000	Plin	5	2.6			
Mercedes E 320 CDI	NA	Grude	E 320	2002	450000	Dizel	5	3.2			
Mercedes-Benz E 220 Cdi	NA	Grude	E 220	2011	157000	Dizel	5	2.2			
Mercedes-Benz E 220	11500	Drvar	E 220	2003	300000	Dizel	5	2.2			
Mercedes 123	NA	Čapljina	E 240	1984	1000	Dizel	5	2.4			

Figure 2: Olx.ba dataframe

Note the NAs in the price column, that is so since olx.ba allows “Cijena po dogovoru”.

German market analysis

Let’s start with a basic barplot:

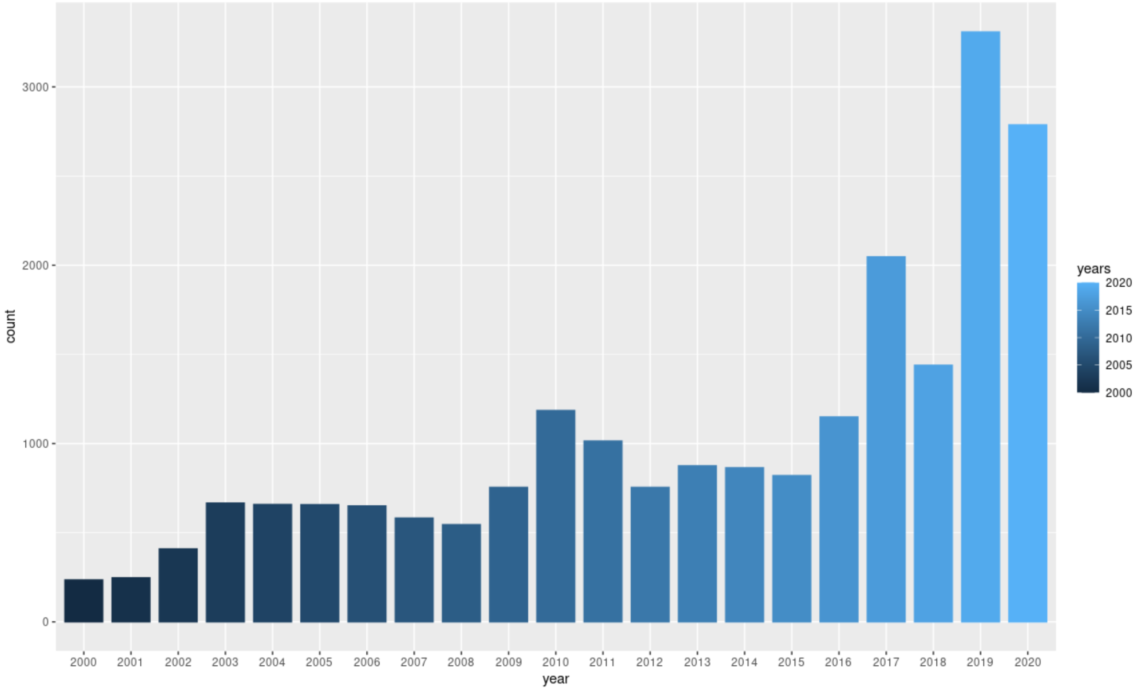


Figure 3: German E-classes by year

We can see that the data is left skewed which means most of the E-classes being offered are less than 10 years old, and as we go further back in time we can see that there are less and less cars available, which means they are mainly exported.

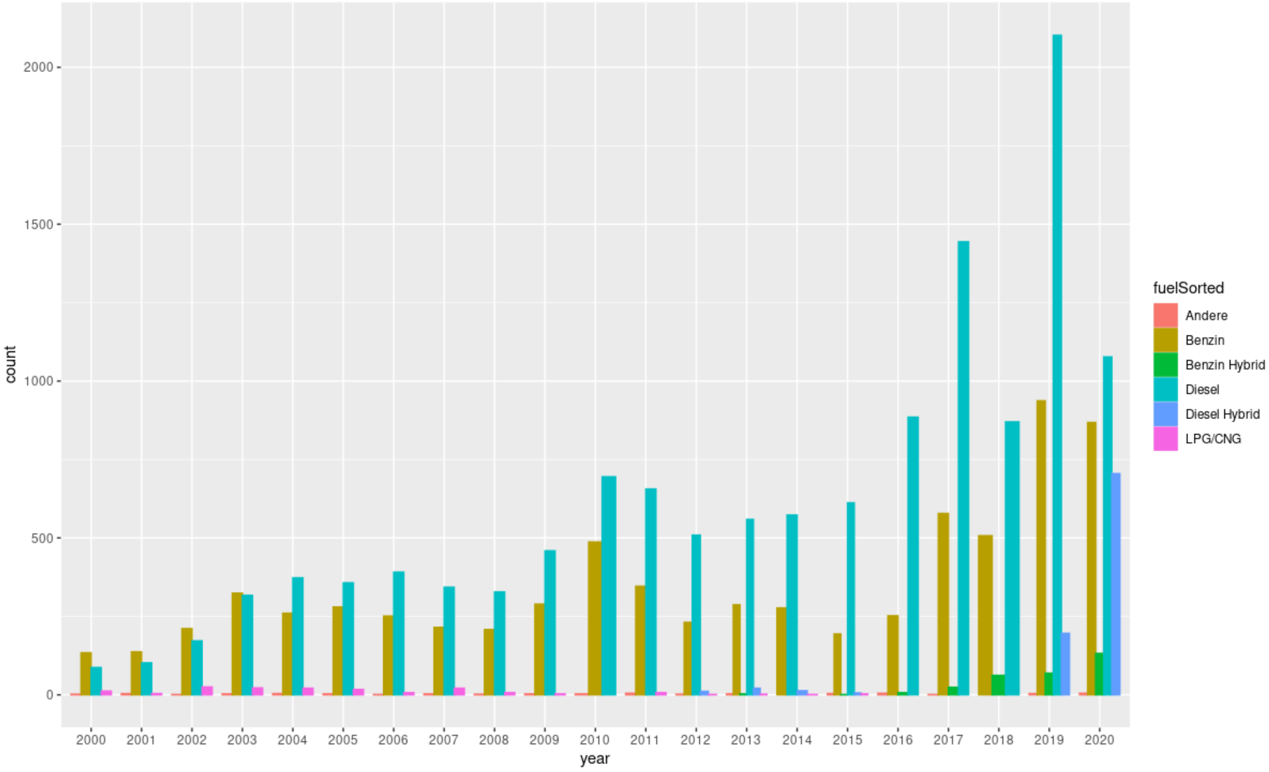


Figure 4: E-classes by fuel sorted by years

We can see in Figure 4 that most of the cars younger than 2003 are diesels and that means that diesels are the one mostly exported. That is due to very strict German laws regarding to air pollution and using diesel engines (which tend to be dirtier and more polluting than gasoline engines). Gasoline engines, however tend to hold on a lot longer, because of two reasons:

- They don’t pollute as much as diesels (German law tolerates them)
- All performance oriented editions are gasoline (AMG)

If we show the same above graph but for cars made before 2000 (Figure 5) one can see a really interesting graph, which makes as ask ourself, where did all those diesels go?

LPG was mainly fit to naturally aspirated gasoline engines, and since the newer engines were mainly turbocharged and a lot more fuel-efficient, it became nearly non-existant on newer cars.

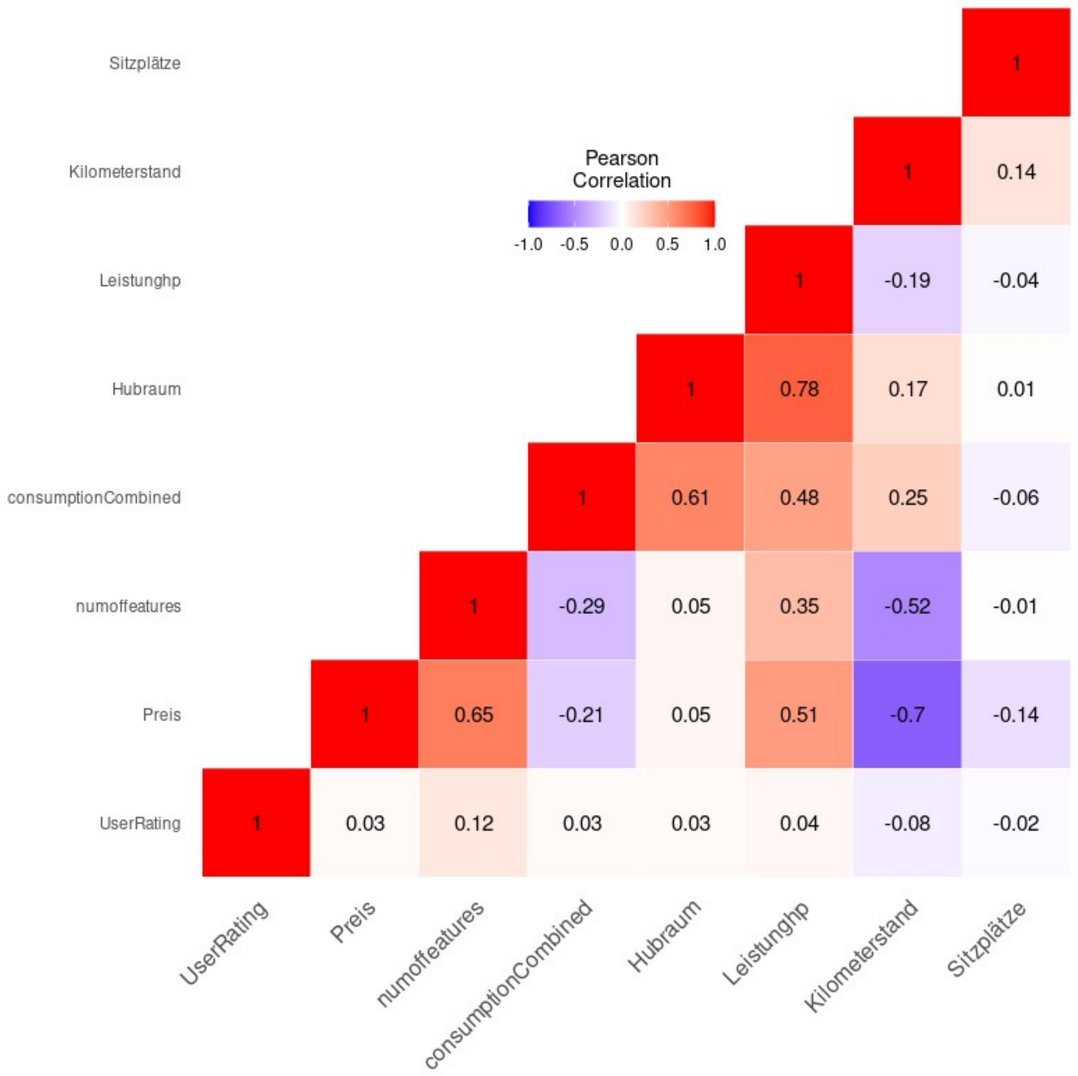


Figure 5: Correlation heatmap for the mobile.de dataset

If we look at the heatmap we’ll see that the user rating and the number of seats aren’t correlated much to any other variable, except that user rating is slightly positively correlated with the number of features and that the number of seats is slightly negatively correlated with the price. It can be explained saying that users who usually sell more cars will describe them in more detail hence the bigger correlation. As for the number of seats and price correlation, we can say that it is because some Taxis that are in estate versions have seven seats.

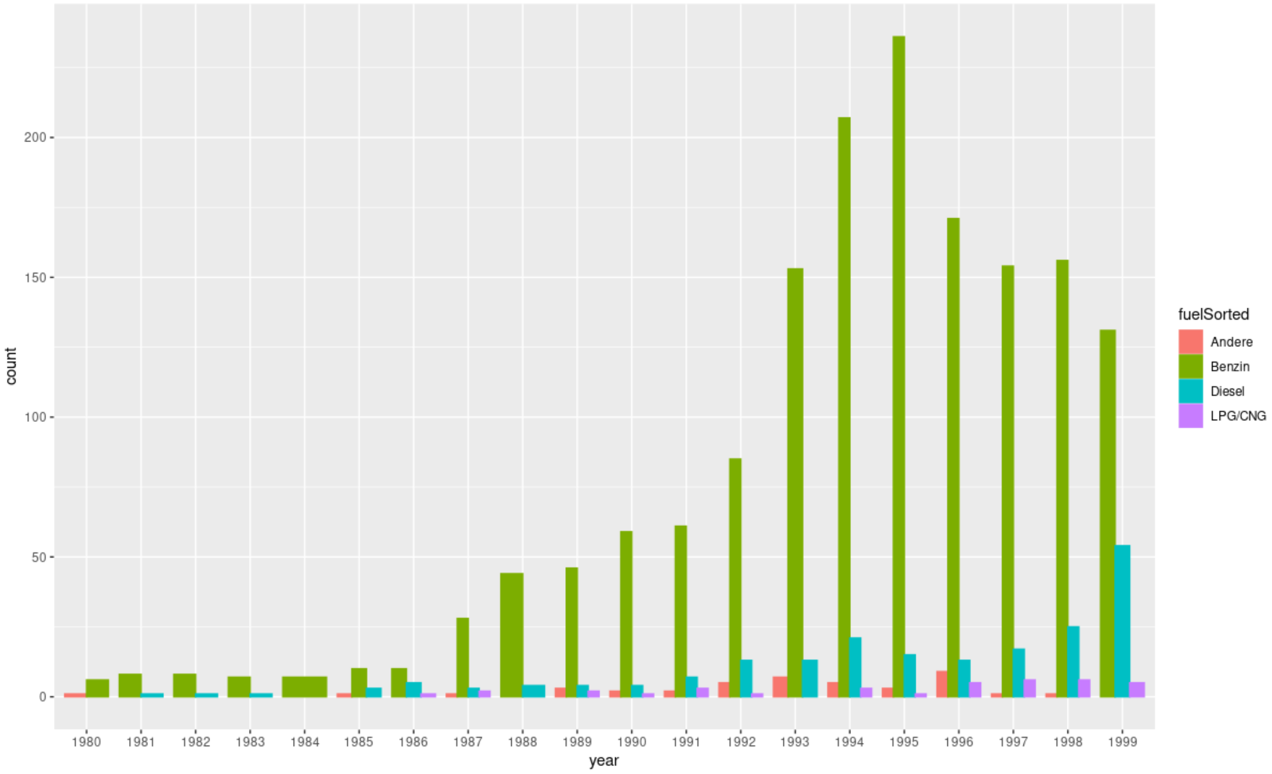


Figure 6: Cars made between 1980 and 1999 by fuel type

Peak of number of cars between 1993 and 1995 can be explained that those were the last years of legendary w124 E-class and that it was coincidentally the golden Age of AMG before the corporate acquisiton from Daimler AG, so in conclusion, many of the classics stayed in Germany(Figure 7).

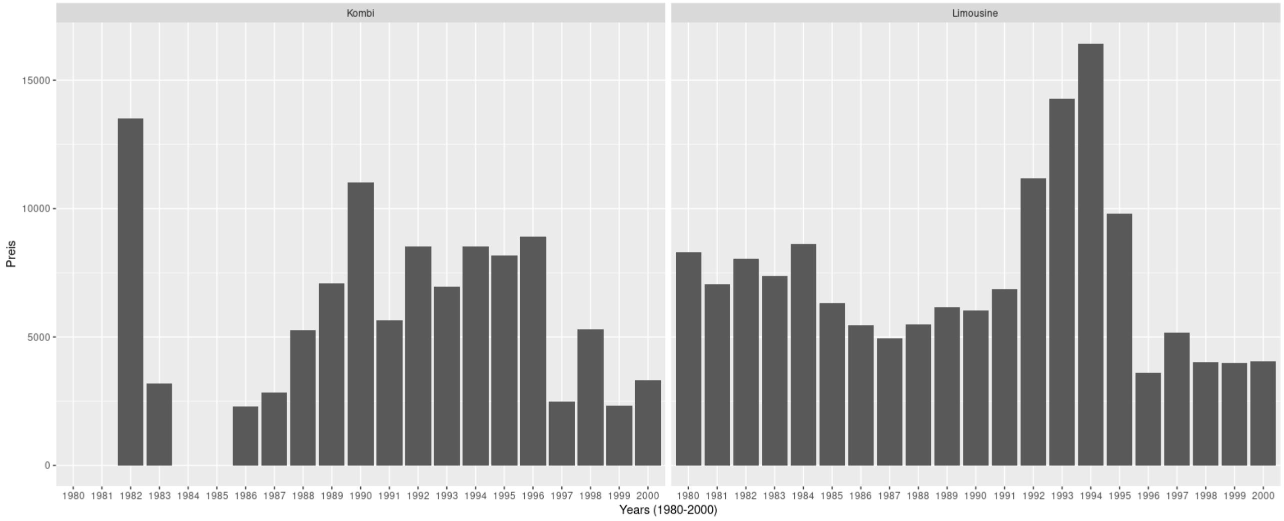


Figure 7: Price of the old-youngtimers (1980-200)

If we look at Figure 8 we can note that estate(kombi) version is more popular that the limousine for cars less than 10 years old, but for the cars that are more than 10 years old we can see that limousines become more popular or more numbered because the estates are mainly exported and a good chunk of them are taxis.

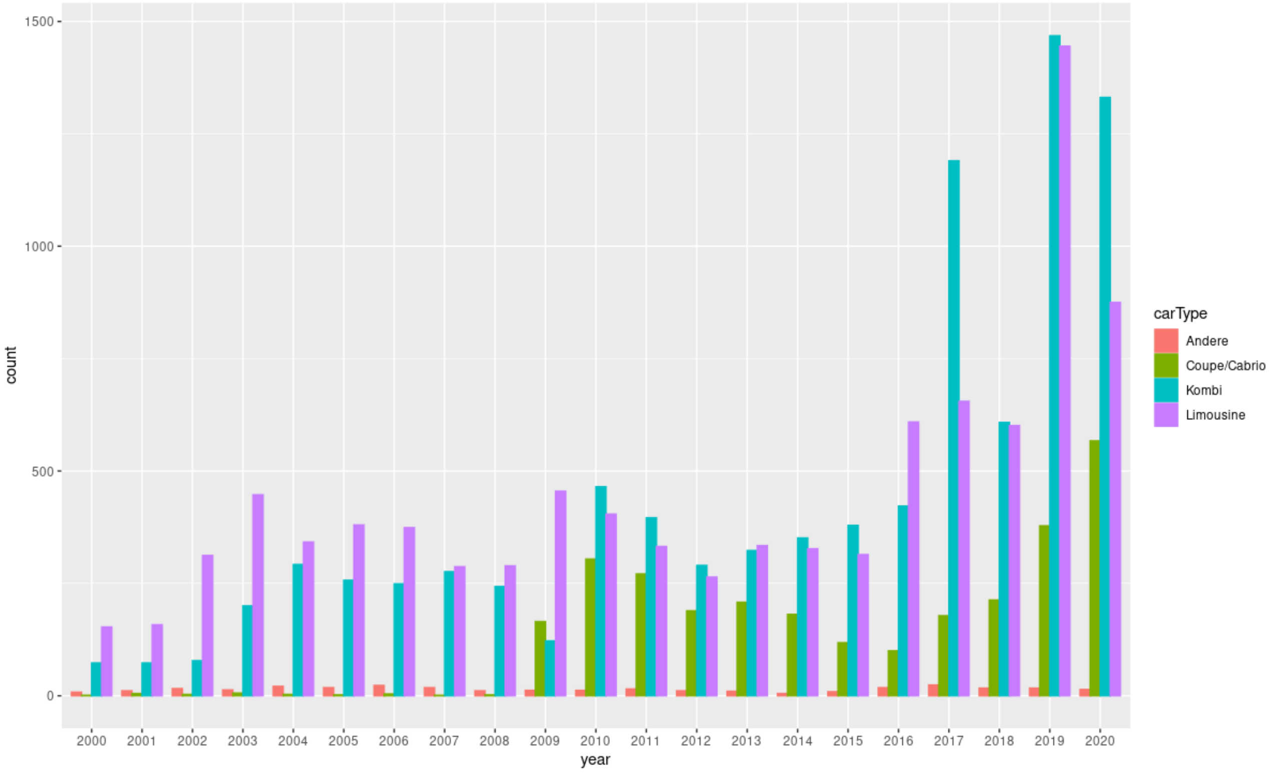


Figure 8: Cars by car type

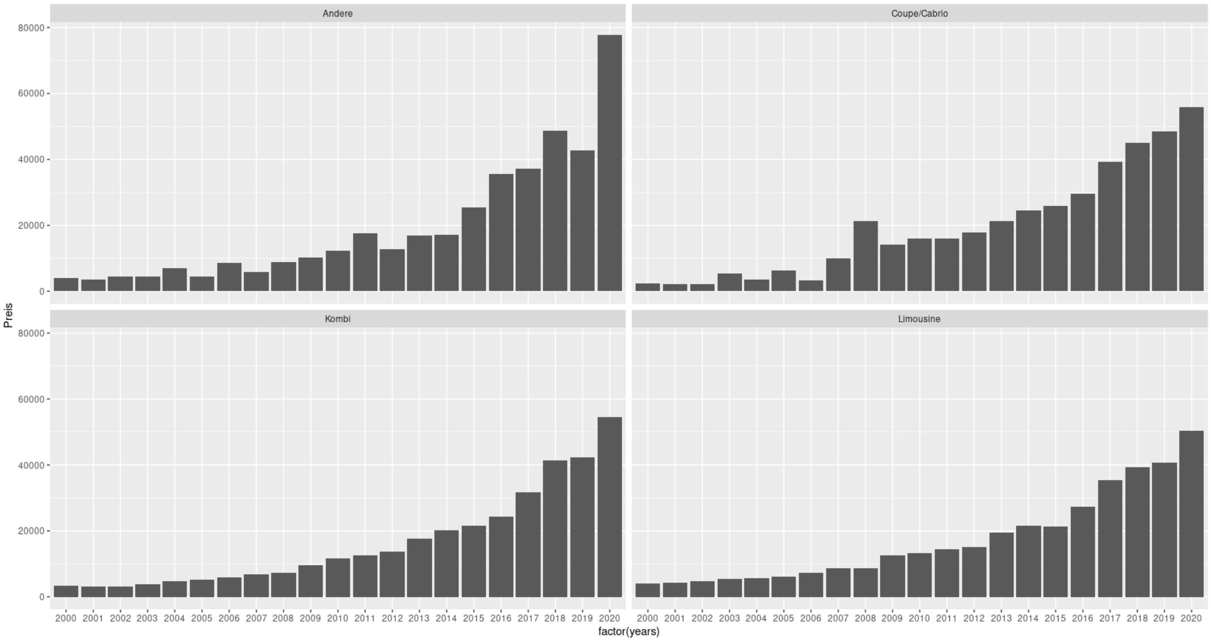


Figure 9: Car price depreciation by car type

In figure 7 we can see that nearly every car type depreciates similiarly with the sedan version holding the price slightly longer. The “andere” includes a lot of unique and one-of-a-kind vehicles that are mainly made out of estates such as E-class ambulance, funeral car (leichenwagen) etc., so it has kind of a special but nevertheless very similar price drop over the years. Coupe and cabriolet versions have also a bit different data since E-class coupe and cabrio were designated as CLK-class from 1996-2009.

One of the trends that are a bit sad at least for automotive enthusiasts is that the manual transmission is going to slowly die off. Mercedes E-class doesn’t offer manual transmission since 2016 (the new W213 line).

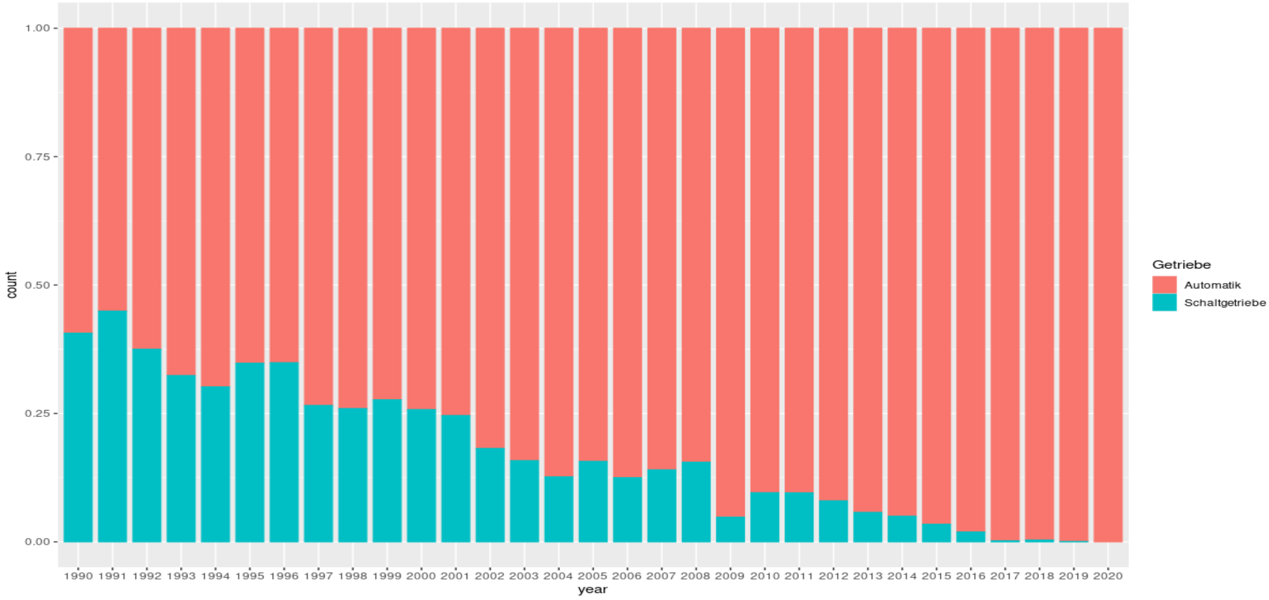


Figure 10: Manual vs Automatic transmission (1990-2020)

Comparison between diesels and others

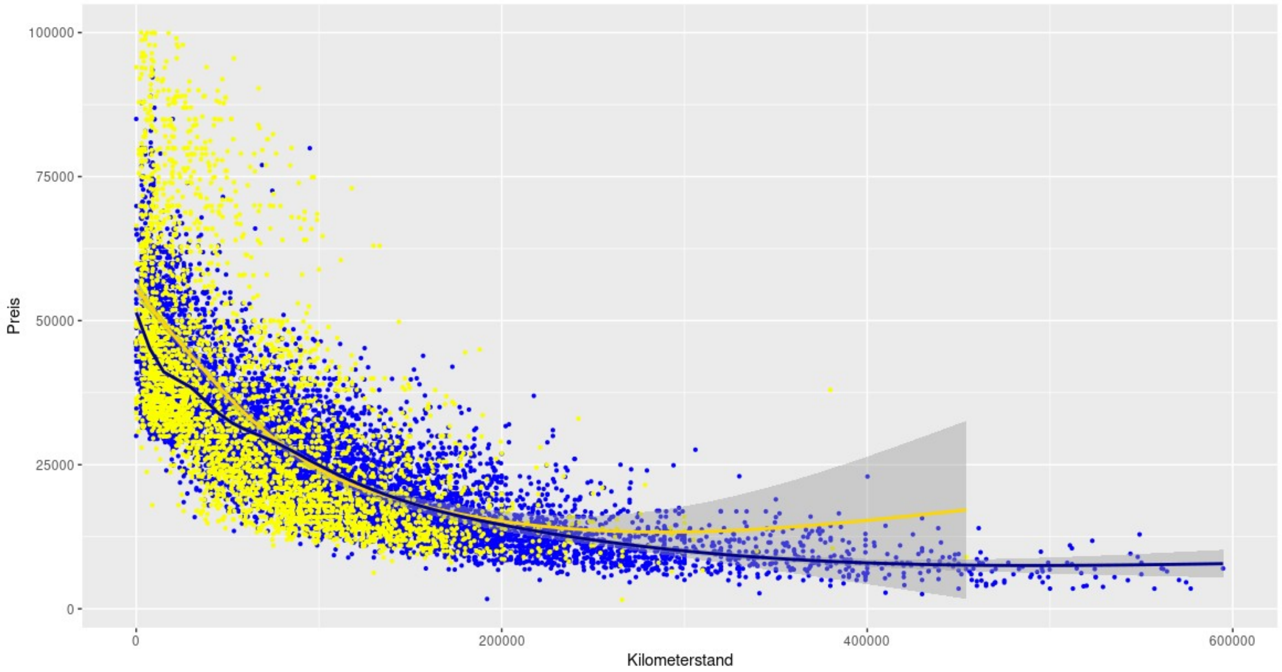


Figure 11: Price and kilometers point plot with diesel(blue) and gasoline(yellow) regression line

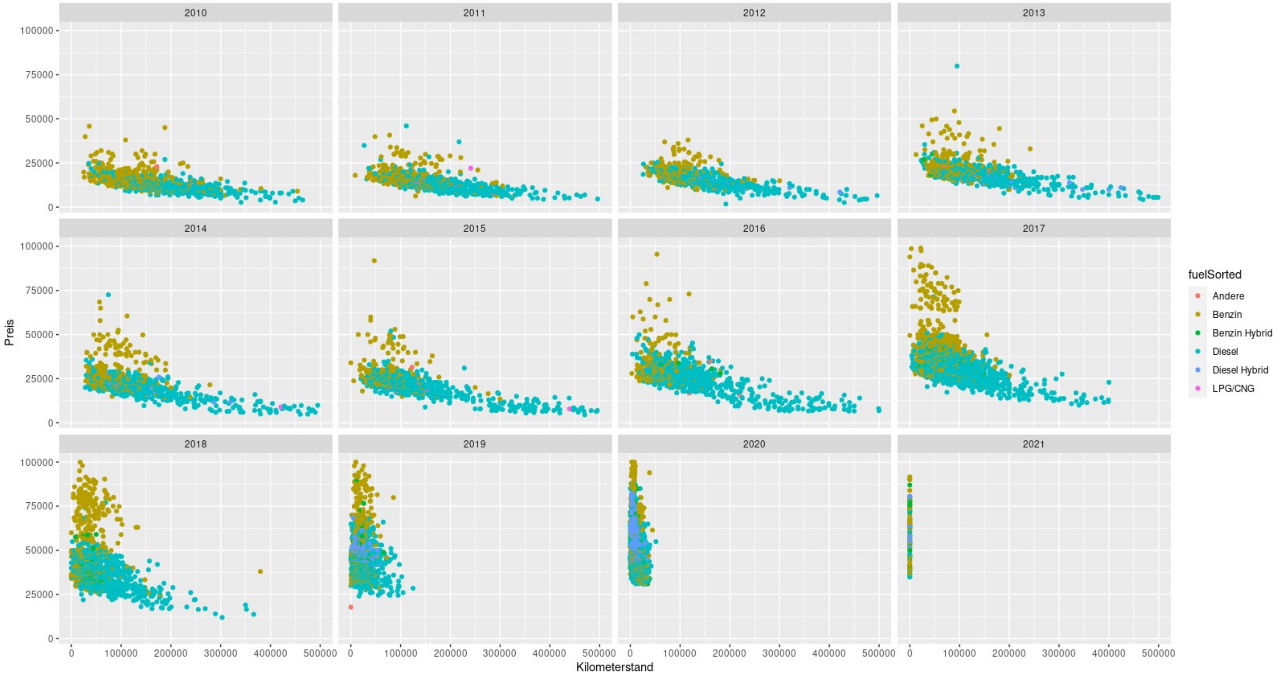
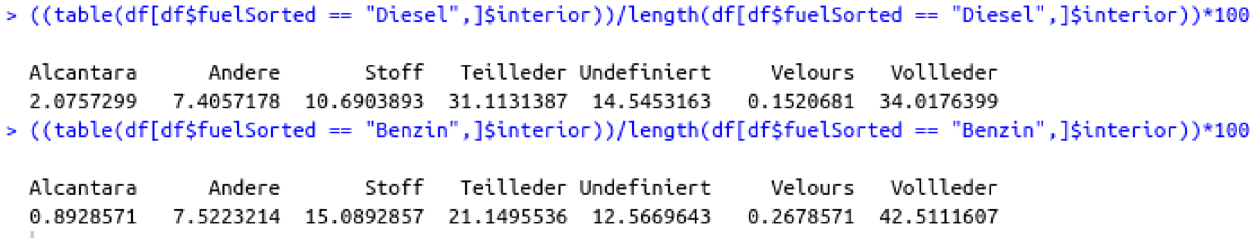


Figure 12: Car price and mileage, colored by fuel (blue-diesel, brown-gasoline)

In figure 11 and 12 we can see the price development of the E-class based upon mileage and fuel type. Overall we can note that diesels are mainly cheaper and driven longer whereas Gasolines are not driven that much and remain expensive a lot longer. That can be explained that all E-classes used comercially such as taxis, ambulances, corporate vehicles, lower public officials vehicles are diesels, since when used they must be reliable and must be able to cover great mileage in short amount of time without braking up, and that is what usually a diesel does. Even though German lawmakers are harsh on diesels, people continue to drive them even in when choosing between diesel or gasoline hybrid. It usually makes more sense to own a diesel even with the yearly expediture if the car needs to cover around 100 000km yearly.

Gasolines are usually vehicles used by urban population and AMG enthusiasts, and as such, they cover a very different space on the figures 11 and 12. They usually aren’t driven for more than 200 000 kms and they are usually have a bit more features and are kept better than the diesels. Even for more than 10-year old cars gasolines still have less mileage and very rarely have driven more than 350 000km(Figure 14).



```
> ((table(df[df$fuelSorted == "Diesel",]$interior))/length(df[df$fuelSorted == "Diesel",]$interior))*100

  Alcantara      Andere       Stoff   Teilleder Undefiniert     Velours   Vollleder 
  2.0757299   7.4057178  10.6903893  31.1131387  14.5453163   0.1520681  34.0176399 
> ((table(df[df$fuelSorted == "Benzin",]$interior))/length(df[df$fuelSorted == "Benzin",]$interior))*100

  Alcantara      Andere       Stoff   Teilleder Undefiniert     Velours   Vollleder 
  0.8928571   7.5223214  15.0892857  21.1495536  12.5669643   0.2678571  42.5111607 
```

Figure 13: Interior material percentages gasoline vs diesel

Based upon Figure 13 we can conclude that gasoline versions usually have a bit more refined interior then their diesel counterparts.

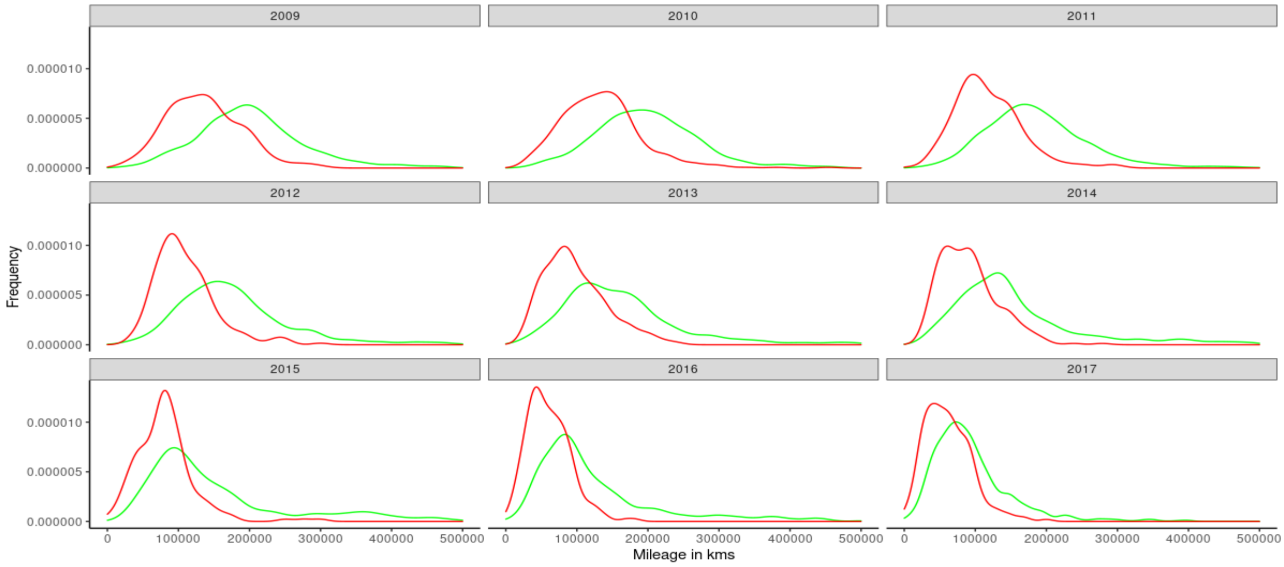


Figure 14: Mileage density plot for Gasoline(red) and Diesel(green)

Comparison between all-wheel-drive models and others

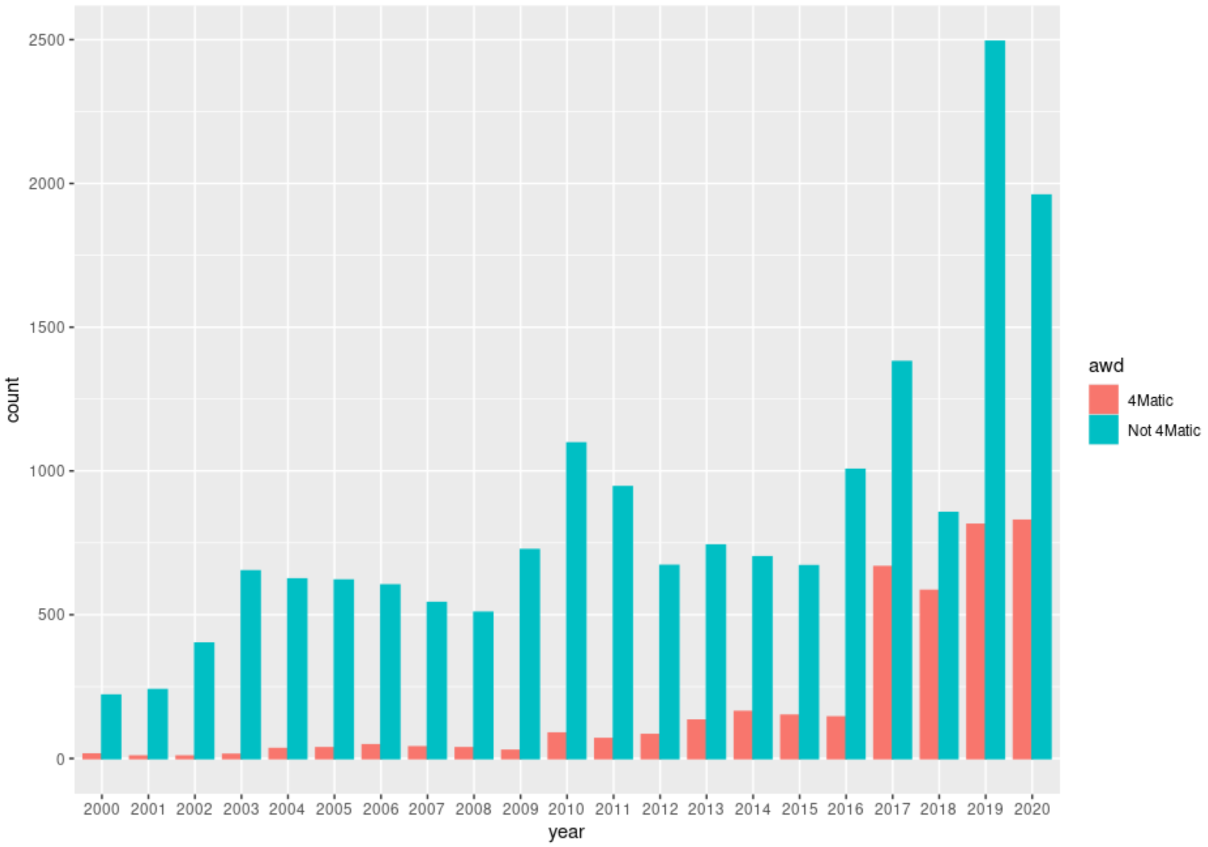
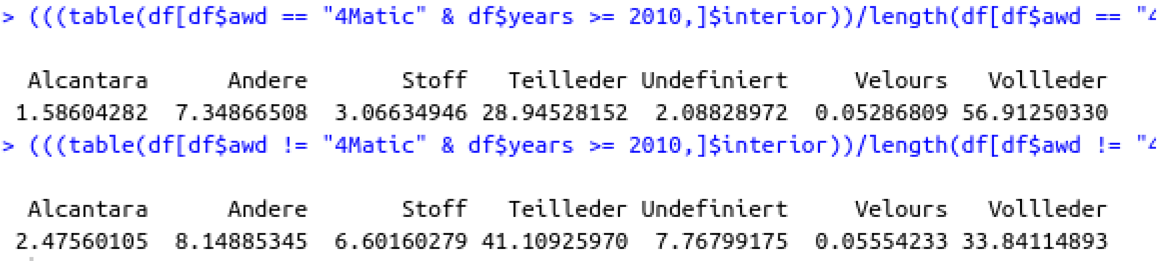


Figure 15: Number of AWD(red) vs RWD(blue) E-classes

The Mercedes all-wheel-drive system (4Matic) was introduced back in 1987, but it wasn't significantly represented in sales or numbers up to recently. In the Figure 15 we can see that the proportion of 4Matics wasn't significant until 2012 (11%) after that it only grew in proportion, up to 30% in 2020. We can surely say that all-wheel-drive, as the automatic transmission in the 1980s is going to become a standard in the upcoming decades, at least for the E-class. The 4Matic models ocupate a more premium space in the car market and their price stays higher than comparable standard models even 10 years after the production (Figure 17). Also they cover less mileage than their standard counterparts and usually have more features and signifcantly better interior (Figure 16).



```
> (((table(df[df$awd == "4Matic" & df$years >= 2010,]$interior))/length(df[df$awd == "4

  Alcantara      Andere       Stoff   Teilleder Undefiniert     Velours   Vollleder 
 1.58604282  7.34866508  3.06634946 28.94528152  2.08828972  0.05286809 56.91250330 
> (((table(df[df$awd != "4Matic" & df$years >= 2010,]$interior))/length(df[df$awd != "4

  Alcantara      Andere       Stoff   Teilleder Undefiniert     Velours   Vollleder 
 2.47560105  8.14885345  6.60160279 41.10925970  7.76799175  0.05554233 33.84114893 
```

Figure 16: Interior materials used awd(above) vs rwd(below)

If we look at Figure 18 we’ll see that the density plot of the 4Matic always occupies more premium space, even for every year, having more features by default. From figure 17 we can conclude that 4Matics aren’t that much abused as the RWD counterparts since they are rarely used as commercial or corporate vehicles. Even after 300 000 kms covered up 4Matic Estates have a price 25% higher than their RWD counterparts, as for limousines the difference is more than 30%. The Viermatics usually have the stronger engines such as V6 Diesel or Petrol and bigger.

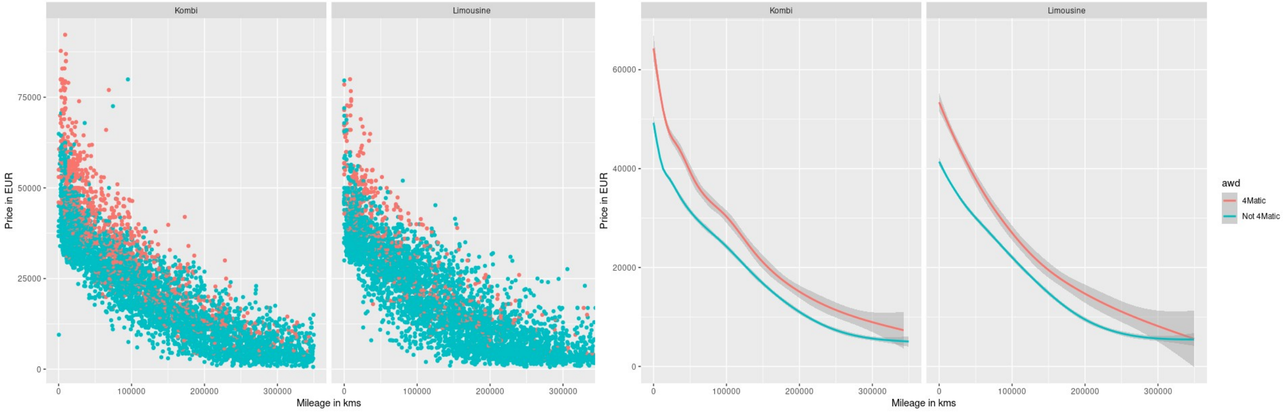


Figure 17: Scatterplot and regression line of awd and rwd models by price and mileage(sorted in estate and limousine models)

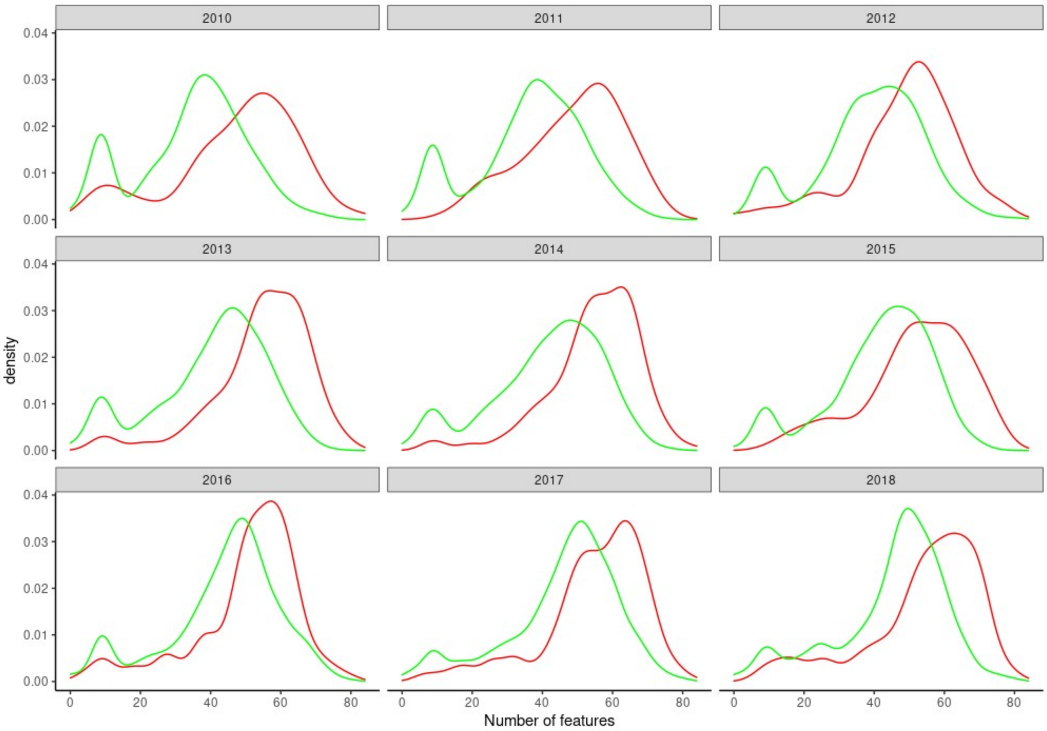


Figure 18: Number of features for 4Matic(red) and non-4Matic(green) models

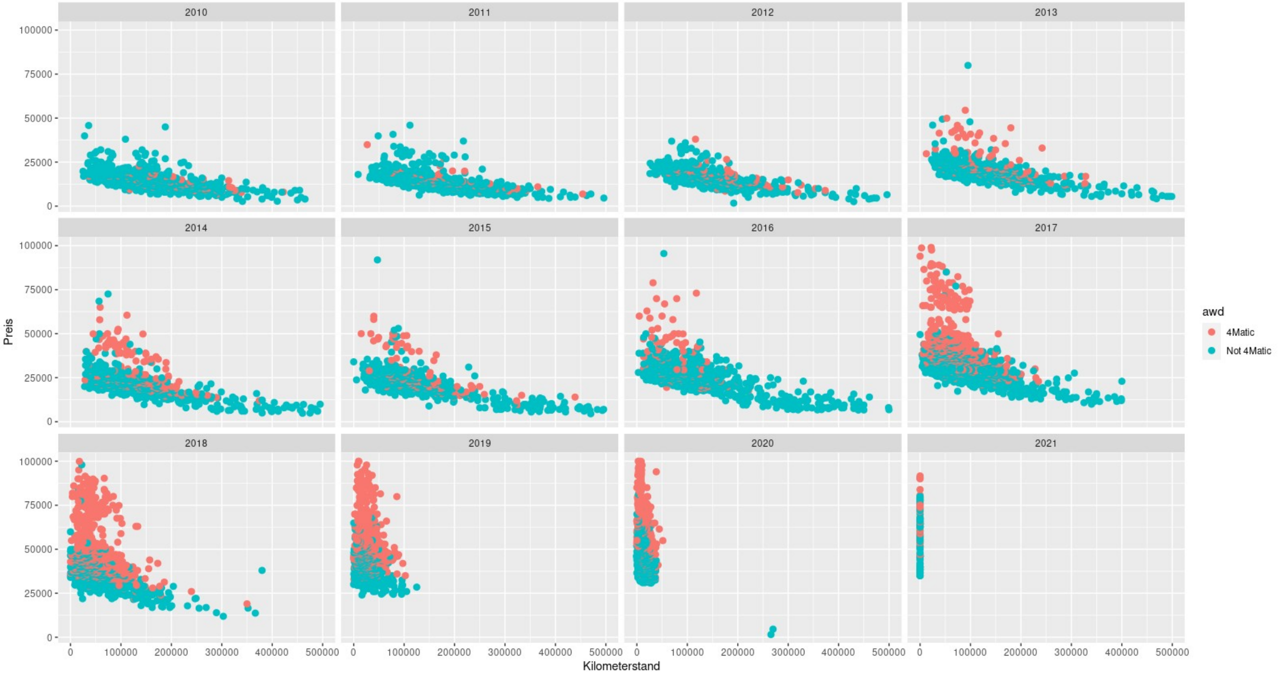


Figure 19: Mileage and Price scatterplot by year for 4Matic and non-4Matic models

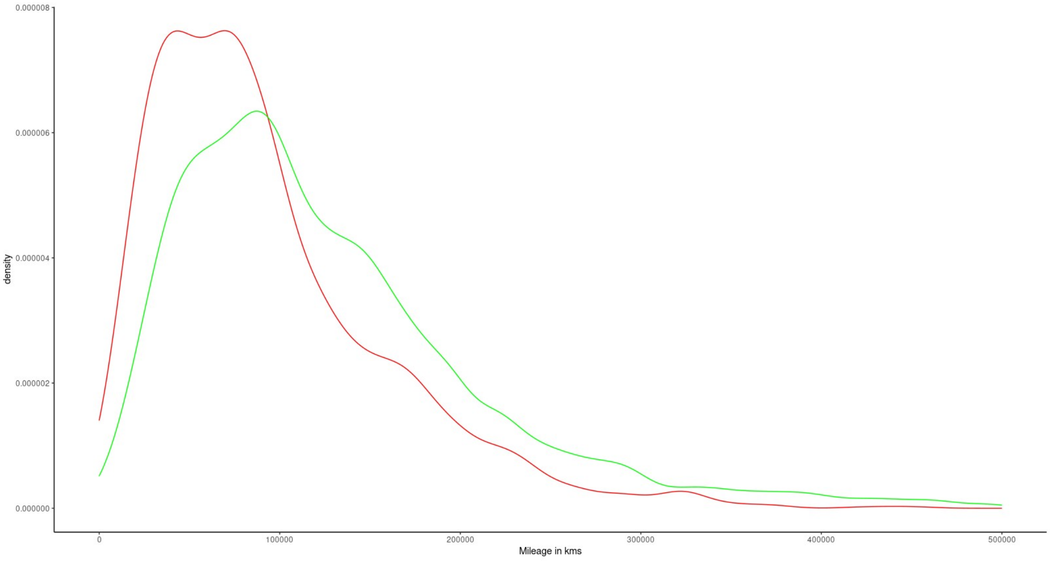


Figure 20: Mileage density plot for 4Matic(red) and non-4Matic(green) models

Comparison of the Bosnian and German markets

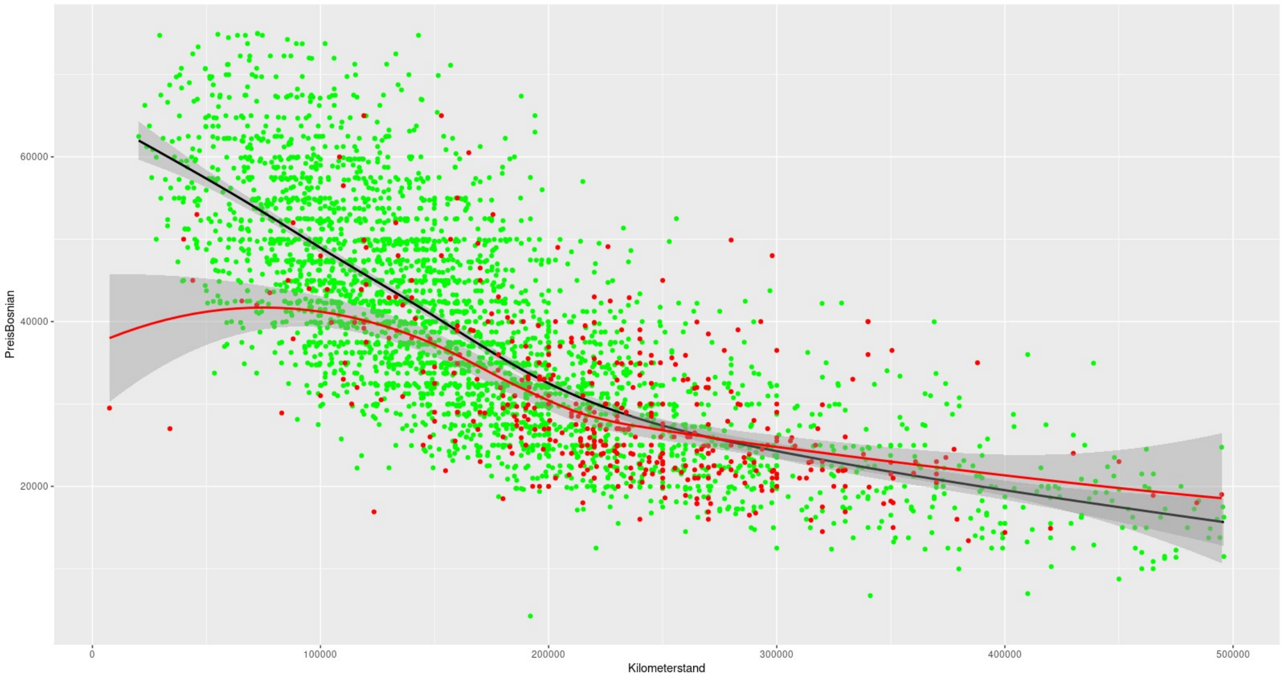


Figure 21: Scatterplot of base models E200 & E220 from Mobile.de(green) and Olx.ba(red) with the regression lines (cars made between 2010 and 2020).

As a picture says 1000 words so do Figure 21 and 22. In Figure 22 we can see that Bosnian and German markets are basically in a opposite state, since since E-classes produced 2000-2003 are the most numerous in the Bosnian market, but quite oppositely the least numerous in the German market and vice versa. That is depicting very vividly the purchasing power of the Bosnian market and what would happen if there were no legal restrictions on the car import like the most recent one (ban on cars with less than Euro 5). But nevertheless the import of newer cars is still slow and it is not substantial, which, in turn, leads to overpriced 20 year old cars in a market that can't afford better. If we look at Figure 21 we will see some anomalies which should be there and are in some way quite obvious. Firstly if we look at the regression lines from 0 to 150000kms we can see that the German one goes declines steadily while the Bosnian price regression line does some weird curve. Well, this can mean a couple of things, either Bosnians import substantially cheaper cars than the usual market prices, or cars that have more mileage are being reset to a smaller value. The Olx.ba regression line between 150000 and 250000 can be explained simply with the fact that most E-classes imported are base models and not so rich in features, hence the price. Another way it can be explained is that there is enough space below the line in the German marketplace(that means there is enough cheap E-classes on Mobile.de to explain that line).
But why would the Olx regression line be above the Mobile.de one after 250000kms?

That could be only explained by the fact that maybe that is the only segment where the odometer was not touched and where it is maybe sold as is, without any modification upon importing.

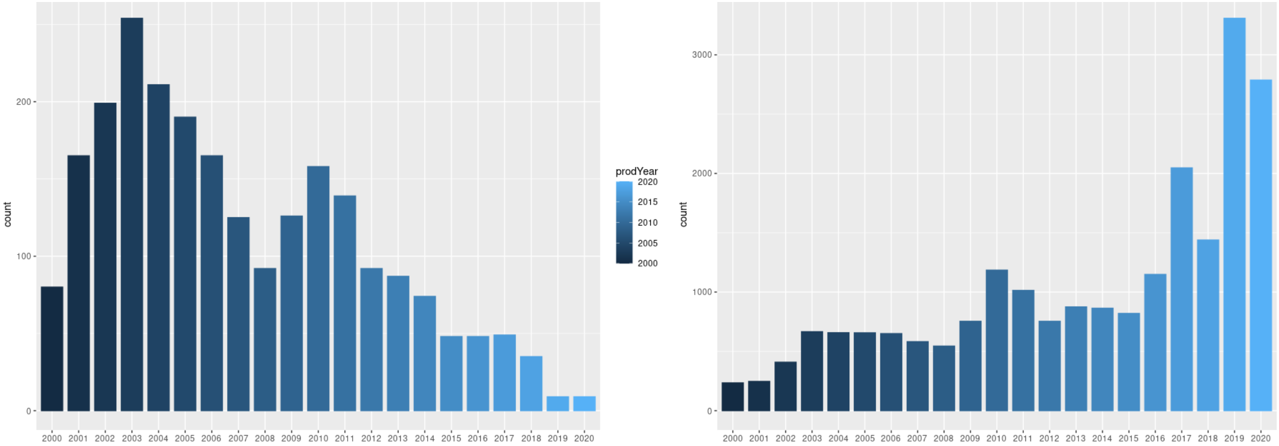
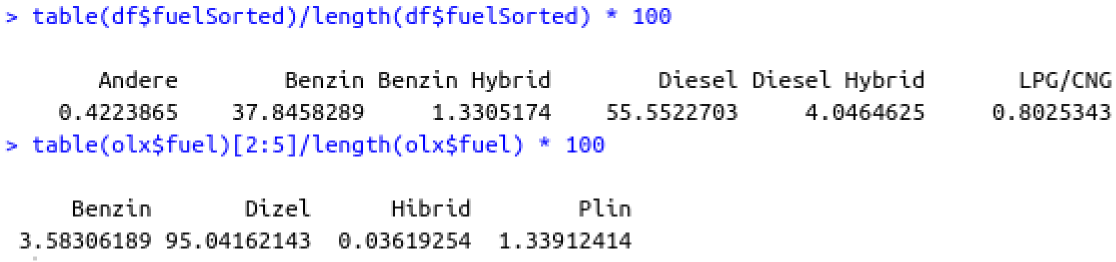


Figure 22: Number of E-classes of all levels on Olx.ba(left) and mobile.de(right), sorted by years from 2000 to 2020. Note the graph is not to scale as the highest bar on the left is ~250 cars, whereas the highest bar on the right is ~3200 cars.

Another interesting fact to note is that 95% percent of all E-classes on Olx are diesels! The hybrids are basically nonexistant, a gasoline E-class is somewhat of an exotic vehicle on the Olx.ba marketplace. On the other Hand on mobile.de we can already see the advent of hybrids which make around 5.3% (diesel and petrol combined), and the ongoing domination of diesel, in spite of the harsh German laws pointed at diesel burning cars (inner cities diesel bans).



```
> table(df$fuelSorted)/length(df$fuelSorted) * 100

      Andere        Benzin Benzin Hybrid        Diesel Diesel Hybrid       LPG/CNG
   0.4223865    37.8458289     1.3305174    55.5522703     4.0464625     0.8025343
> table(olx$fuel)[2:5]/length(olx$fuel) * 100

     Benzin       Dizel      Hibrid        Plin
 3.58306189 95.04162143  0.03619254  1.33912414
```

Figure 23: Percentages of different fuels used in Mercedes E-class at Mobile.de(above) and at Olx.ba(below)

Taxis

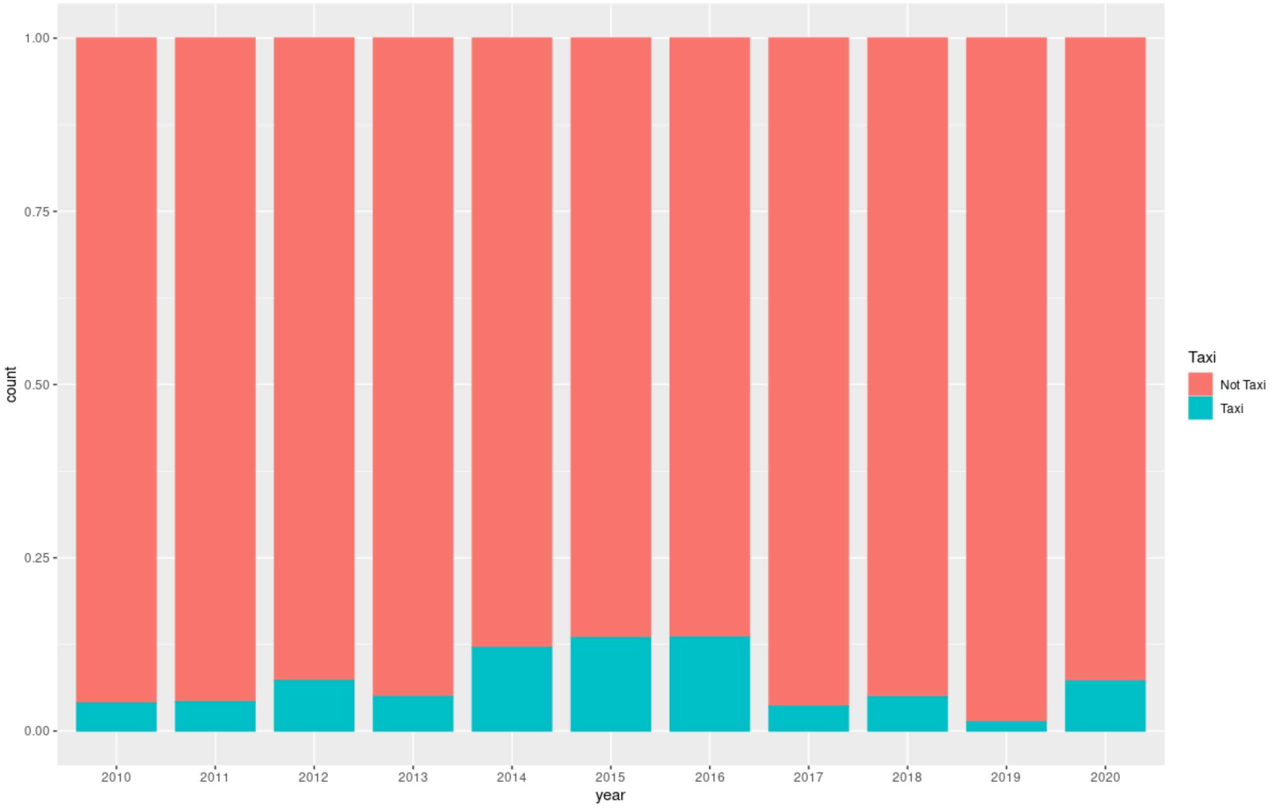


Figure 24: Percentage of taxis between 2010 and 2020 among the E 200 and E 220 models

Taxis are very important topic regarding the E-class, since historically taxis in Germany have mostly been E-classes and that legacy, to a large extent continues today. We can see in the Figure 23 that nearly 14% of all E 200 and E 220 models made in 2015 and 2016 are taxis. The number drops after that to around 4% which means they are mainly being exported and possibly repurposed in other countries. We could say that there are many hidden taxis in the base models in Bosnia if we look at the Figure 25 which were once used as taxis in Germany and then painted in a different colour upon arrival in Bosnia. That claim is supported by the irregularities shown in the Figure 21. If we look at the Figure 26 we can see that the taxis are basically sold either as new vehicles or as cars that have already done 300 000 km (there are taxis in between those mileages but very few when compared to the rest of taxis). The mean price of a new taxi is 37 326 Euro, whereas for the same non-taxi models the mean price is 44 967 Euro which is 17% difference, but after 4 years those mean prices change to 14213 for taxis and 25301 for non-taxi base models which makes 44% price difference. This process is partially visualized in the Figure 27. As we can see taxis actually have higher regression line, since they attain those mileage numbers faster than any other type of E-class, but after 500 000 km the taxi price regression line goes lower than the other regression line since taxis with that mileage are mainly done with their service and likely have some severe mechanical issues that pull their price down. For. ex. Taxi that is 4 years old likely has around 350-450 000 km whereas the same non-taxi model E-class car has 100-150 000 km or less. If we look at Figure 30 we will see that, quite suprisingly, taxis have kind of better interior trim than other base models in terms of materials used (leather).

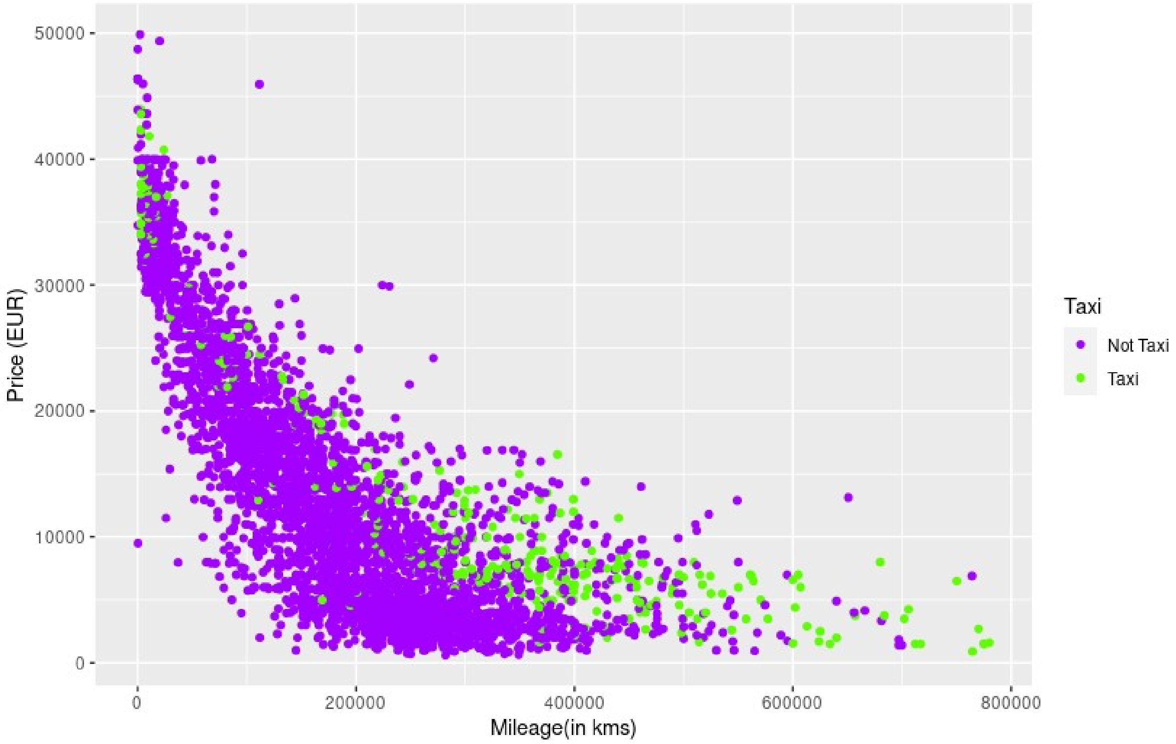


Figure 25: Price and Mileage scatterplot of taxis and non-taxis

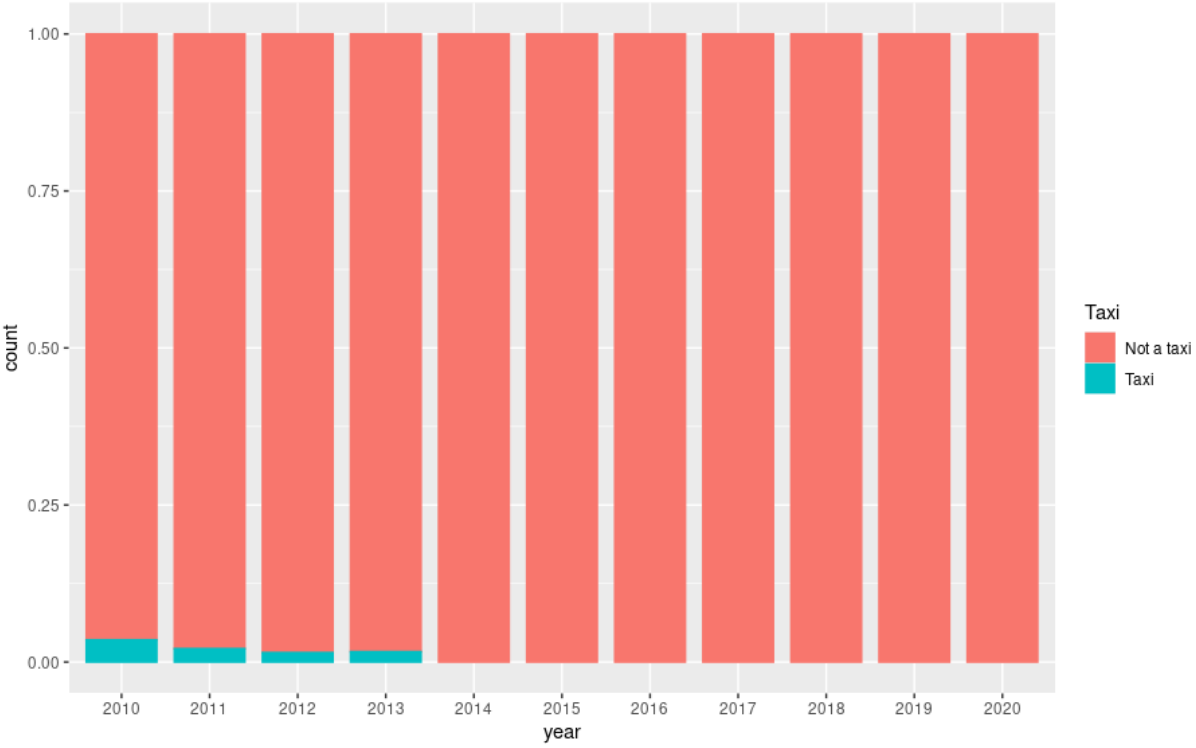


Figure 26: Taxis on Olx.ba
Note: This graph might be misleading since most people don't fill up the car data correctly, eg. it is taxi and it has the taxi beige color but that isn't written anywehere on the olx.ba page of the specific car, so there are many more taxis if not a couple of times more than shown on this graph.

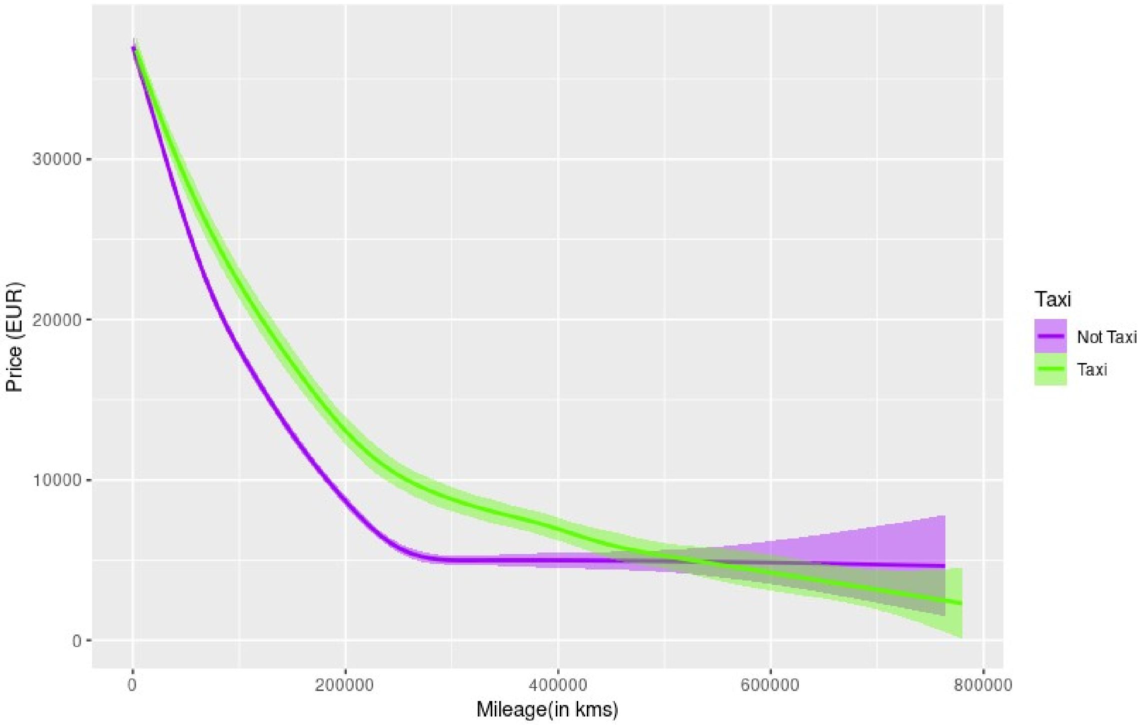


Figure 27: Regression lines of price and mileage for taxi and non taxi models.

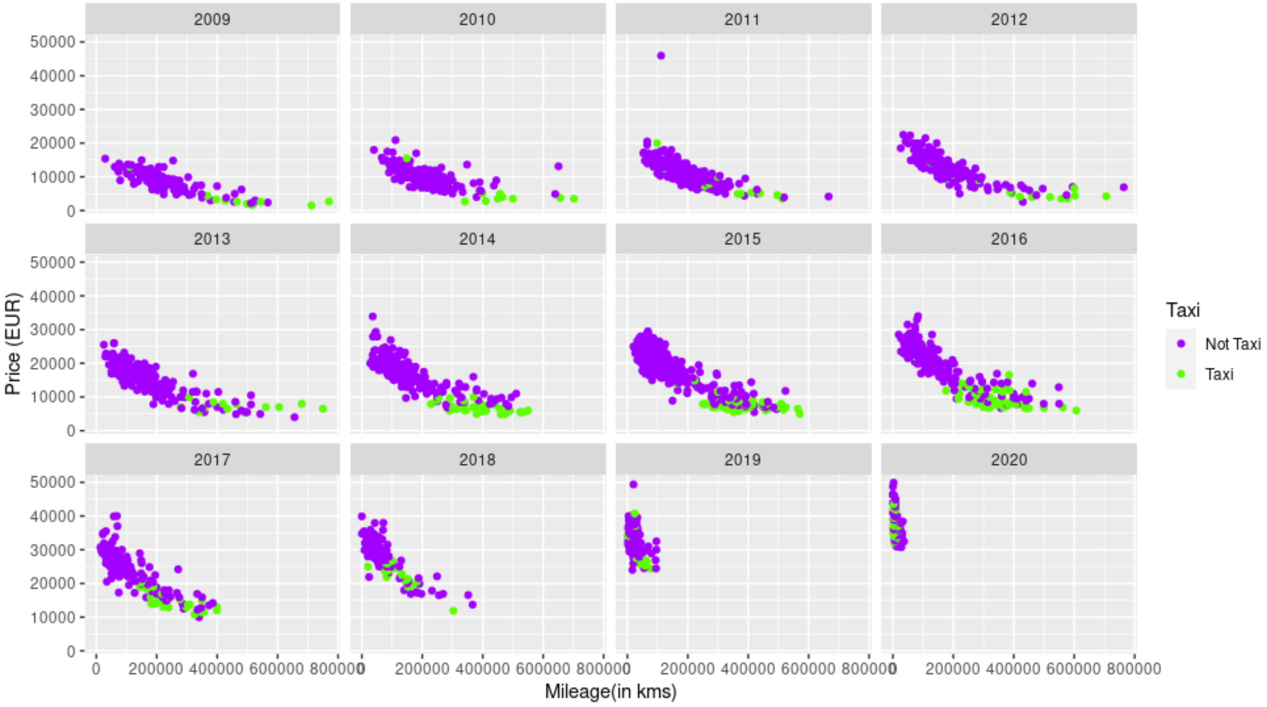


Figure 28: Scatterplot of price and mileage from 2009 to 2020

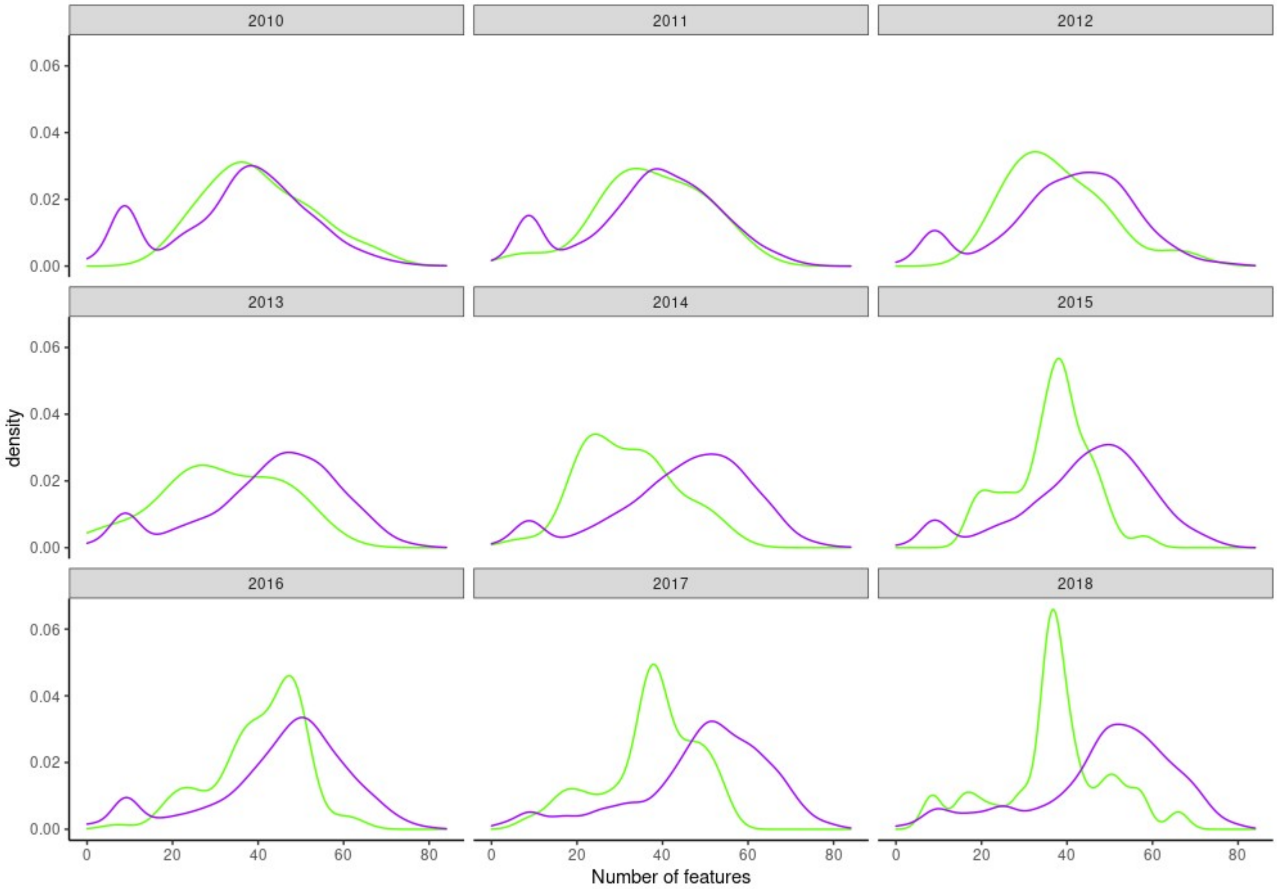
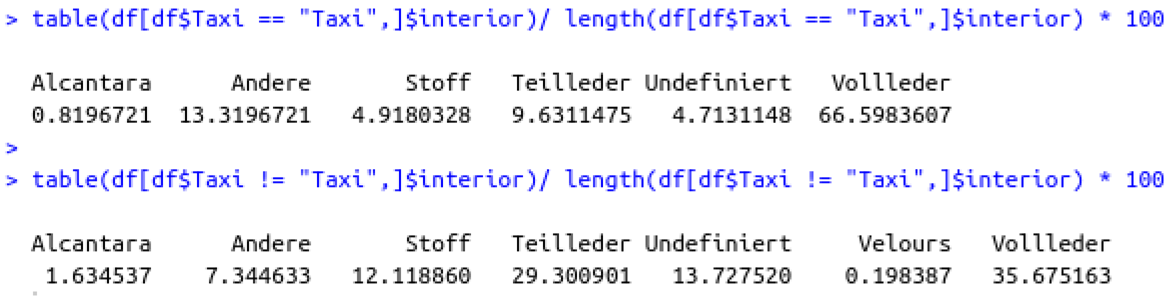


Figure 29: Density plots of number of features of taxi and non taxi models.

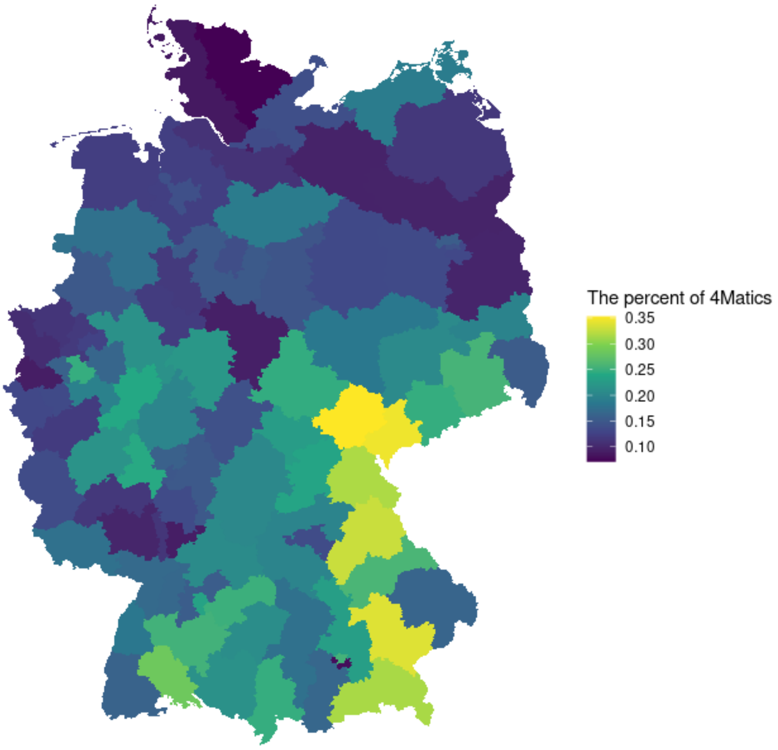


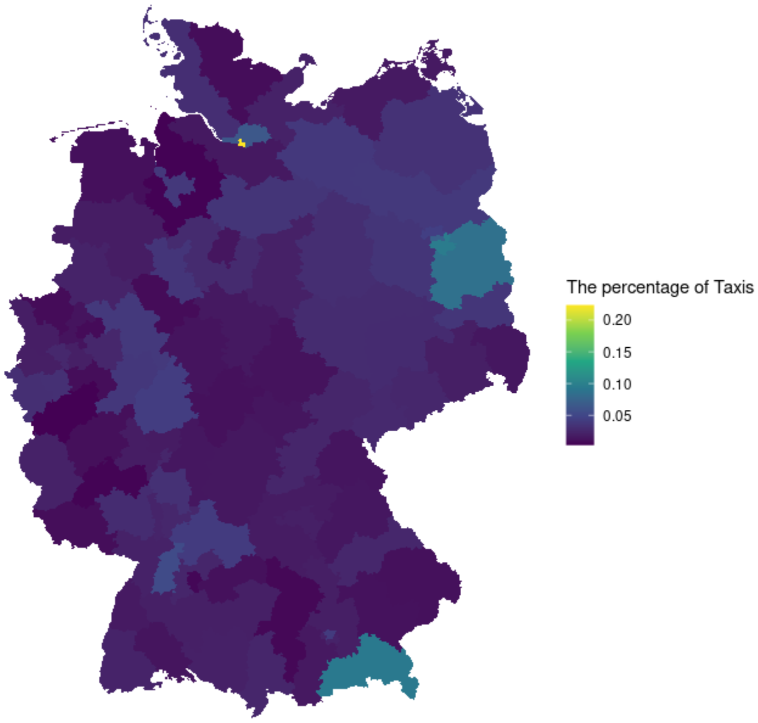
```
> table(df[df$Taxi == "Taxi",]$interior)/ length(df[df$Taxi == "Taxi",]$interior) * 100

  Alcantara      Andere       Stoff   Teilleder Undefiniert   Vollleder 
  0.8196721  13.3196721   4.9180328   9.6311475   4.7131148  66.5983607 
> 
> table(df[df$Taxi != "Taxi",]$interior)/ length(df[df$Taxi != "Taxi",]$interior) * 100

  Alcantara      Andere       Stoff   Teilleder Undefiniert     Velours   Vollleder 
   1.634537    7.344633   12.118860   29.300901   13.727520    0.198387   35.675163 
```

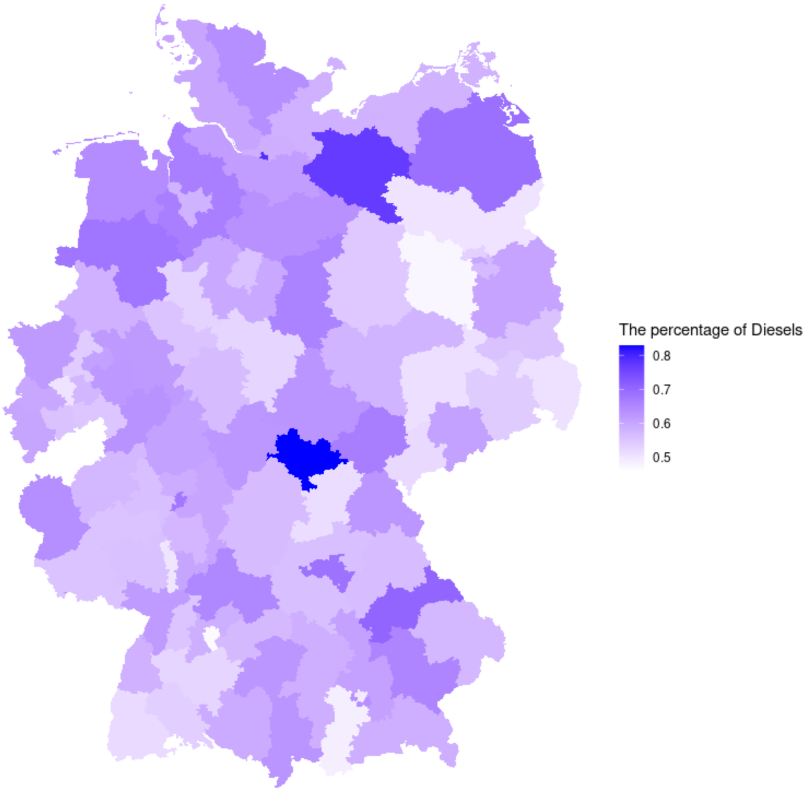
Figure 30: Interior material of taxi(above) and non-taxi(below) models

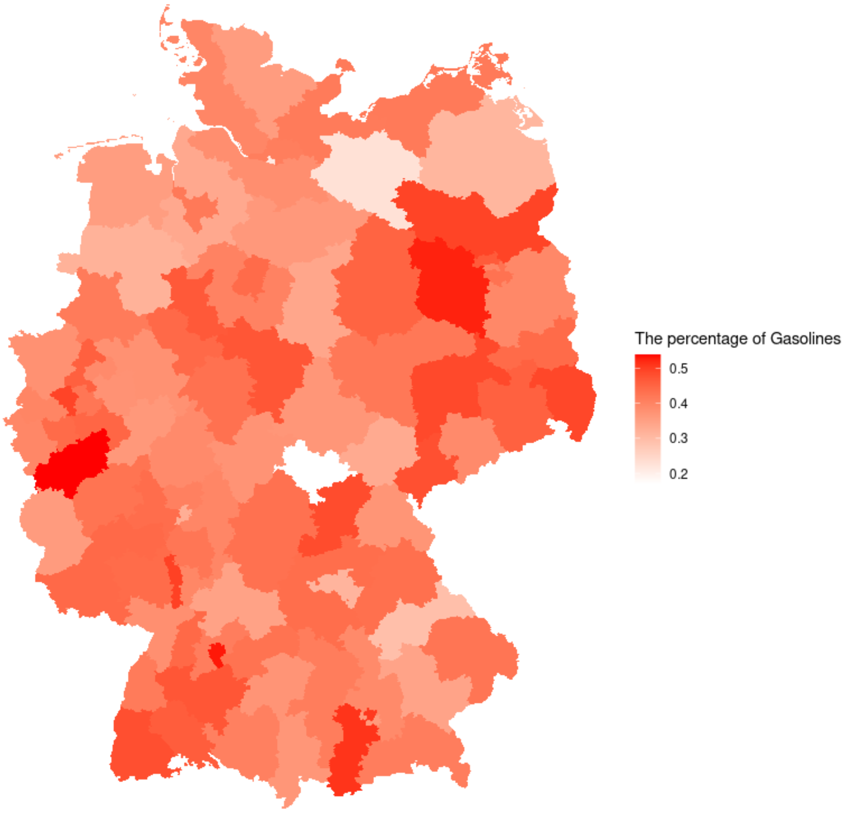
Geospatial visualization of the dataset

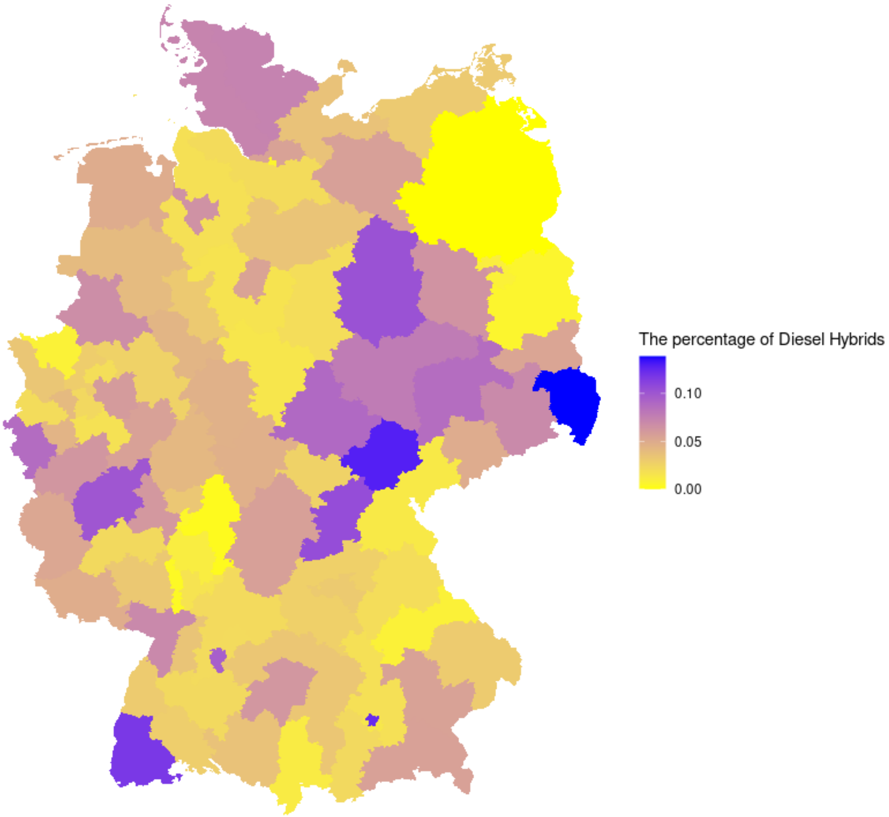
Below are maps of Germany made with the help of ggplot2. They depict the data according to the postal code of cars, that’s the reason why borders aren’t same as the inner German state borders.

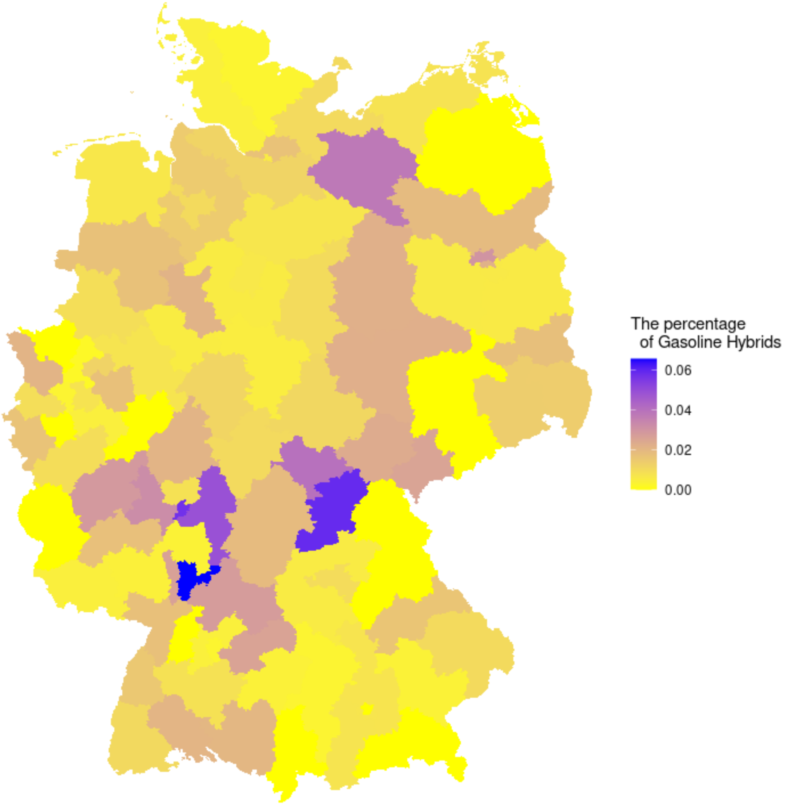


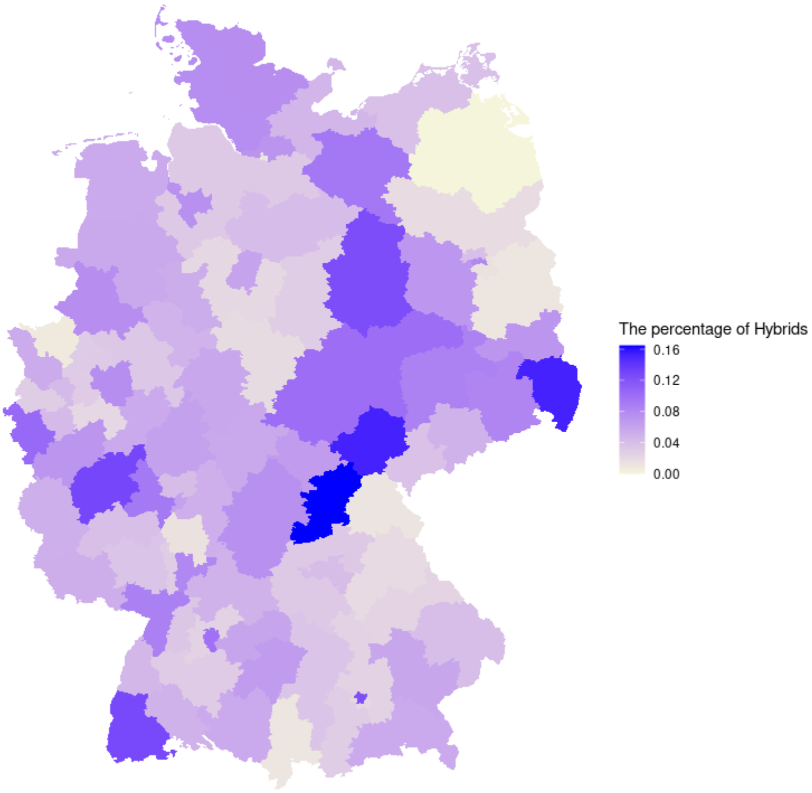


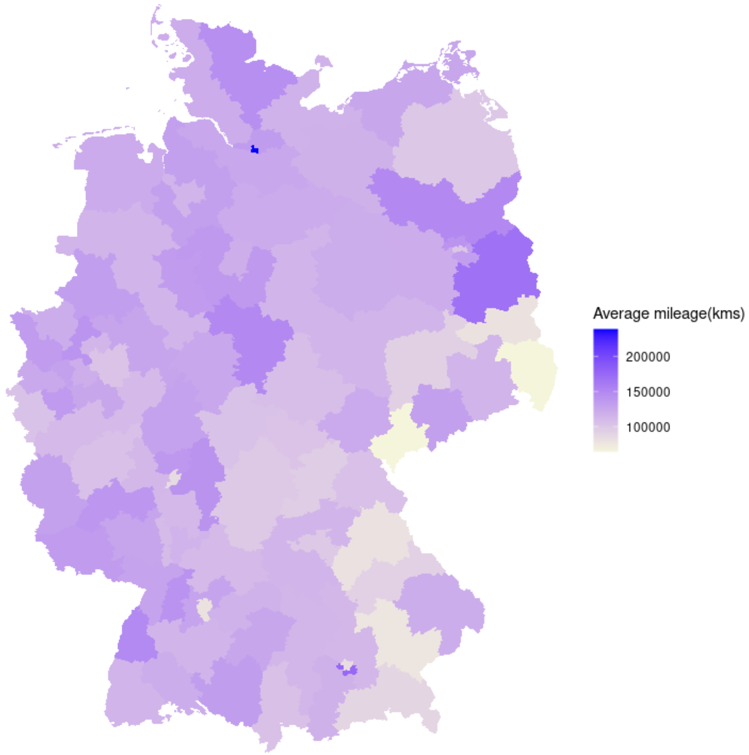

The percentage of Diesels
0.8
0.7
0.6
0.5

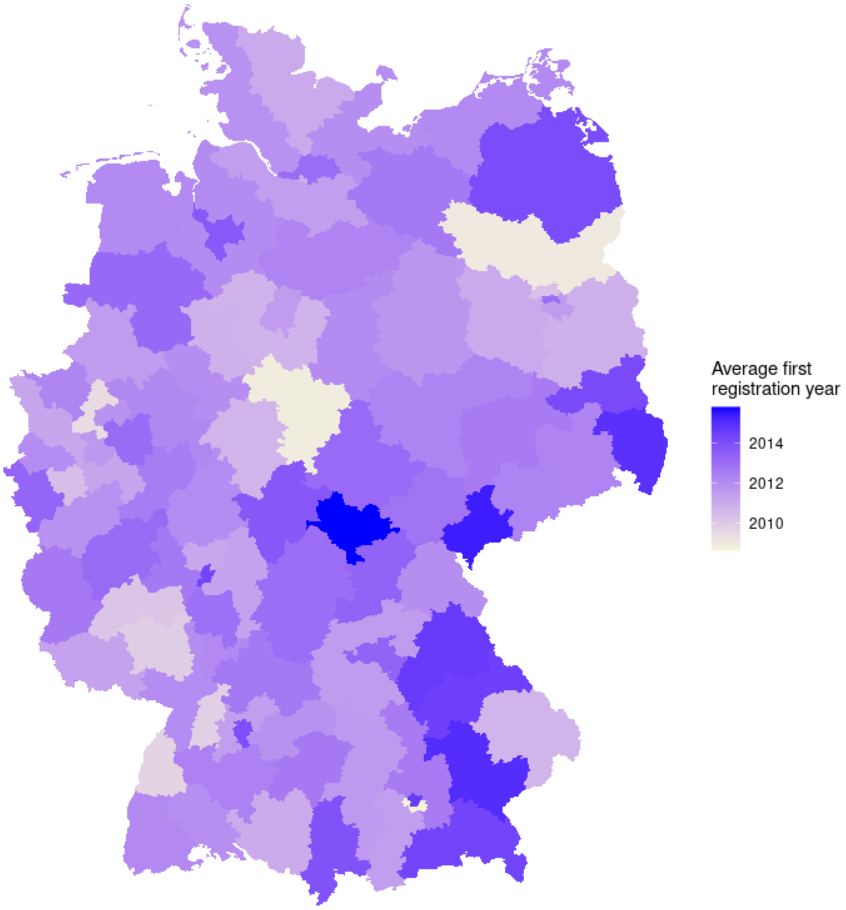

The percentage of Gasolines
0.5
0.4
0.3
0.2

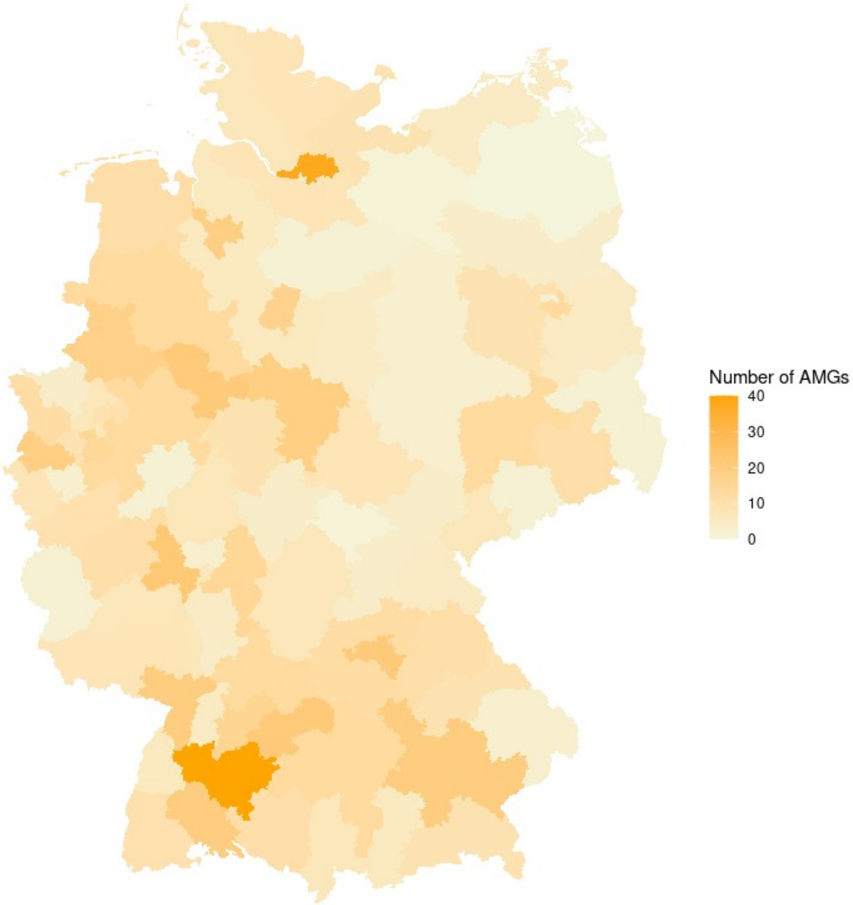

The percentage of Diesel Hybrids
0.10
0.05
0.00


The percentage
of Gasoline Hybrids
0.06
0.04
0.02
0.00


The percentage of Hybrids
0.16
0.12
0.08
0.04
0.00


Average mileage(kms)
200000
150000
100000


Average first
registration year
2014
2012
2010


Number of AMGs
40
30
20
10
0

Conclusions

As expected the most interesting analysis and graphical representation done so far was the one between the Bosnian and German markets. It vividly depicts the irregularities in the Bosnian used German car market. So one conclusion for those buying their E-class in the Bosnian market, based upon that analysis would be to avoid the base models, especially if they have less than 250 000 km on the odometer, and avoid the least powerful diesel model E 200. We can also see that the hybrids are basically ignored in the Bosnian market since nobody actually buys them or imports them even for the cars only a couple of years old (W213). The safest thing a Bosnian buyer could do when buying a used E-class would be to buy that car in Germany if in any way possible. We can also conclude that the 4Matic models hold their price significantly longer and in the long-term they price will not go below 3500km even if the car is more than 30 years old (olx.ba). Regarding the interior trim 4Matics have a lot better materials utilized in the interior when compared to the RWD models and usually have significantly more features. The 4Matics are going to be a lot more numerous in the future since they make around 30% of models made in 2020. Manual transmission “Schaltgetriebe” is a thing of the past, at least for Mercedes E-class, and most probably all other German car makers are going to follow up on this trend. Taxis quite suprisingly, have better interior materials (leather standard in more than 66% of taxis) which are made sturdier and stronger than the usual leather in the other models. Also very interestingly we can see that there is a “taxi peak” at years 2015 and 2016 which implies that taxis are used for 4-5 years then sold and mainly exported to (south) east Europe. According to the fuels used in German E-classes we can safely say that the car drivetrain electrification and hybridization is at the doorstep in the next decade, since 30% of the E-classes produced in year 2020 are Hybrids (Diesel or Gasoline). As for the old-timer market we can see that the most valuable and sought after models are W124 models made between 1993 and 1995 since those were the last years of production for that model and quite coincidentally that was the golden era of AMG tuning company. We can also note that gasolines are mainly driven in the urban areas(with the odd exception of Hamburg) while diesels are driven in more rural parts of Germany, especially former East Germany.

Appendix

Nearly every graph and figure was made using some variation of the following code snippets:

1. R Code for getting the cars by transmission and by year(Figure 10):

```
ggplot(df[
  df$years >= 1990 & df$years < 2021 &
  !is.na(df$Getriebe) &
  df$Getriebe != "" &
  df$Getriebe != "Halbautomatik"
  ,]) +
  geom_bar(aes(x = as.factor(years), fill = Getriebe, color= Getriebe), position = "fill", width =
0.8)+
  labs(x = "year")
```

2. R code for the scatterplot Price in Euro and mileage in KM 4matic vs others (Figure 17):

```
ggplot(data = df[
  df$Kilometerstand < 350000 &
    df$Preis < 100000 &
    df$years >= 2000 &
    (df$carType == "Kombi" | df$carType == "Limousine") &
    (df$fuelSorted == "Diesel")
  , ]) +
  aes(x = Kilometerstand, y = Preis, color= awd) +
  geom_point(size = 2.5)+
  #geom_smooth() +
  labs(y = "Price in EUR", x = "Mileage in kms") +
  facet_wrap(~carType)
```

3. R code for maps(it differs a bit according to which data is being mapped):

```
germany_map <- read_sf('./map/plz-2stellig.shp', 'plz-2stellig')

for (i in 1:length(germany_map$plz)) {
  dax <- NULL
  print(i)
  dax <- df[
    df$germanPostalCodeTwoDigits == germany_map$plz[i],
    ]$fuelSorted
  val <- table(dax)["Diesel Hybrid"]/length(dax)
  if (!is.na(val)) {
    germany_map$viermaticsPRC[i] <- val
  } else {
    germany_map$viermaticsPRC[i] <- 0.00
  }
  print(paste0("Postal code: ", germany_map$plz[i], " average ",
germany_map$viermaticsPRC[i]))
}

ggplot() +
  geom_sf(aes(fill=viermaticsPRC), color = 'transparent' , data = germany_map)+
#
  scale_fill_gradient(low = "beige", high = "blue", na.value = NA)+
  labs(fill = "Average mileage(kms)")+
  theme_void()
```

4. R code for the Pearson correlation heatmap:

```
getCorMapYear <- function(df){
  library(ggplot2)
  library(reshape2)
  cordf <- df[, c(3, 4, 8, 9, 13, 130, 133, 145)]
  cormatrix <- as.matrix(na.omit(cordf))

  col<- colorRampPalette(c("blue", "white", "red"))(20) ##plot(table(df[(df$years > 2000) &&
(df$Kraftstoffart == "Diesel"),]$Hubraum))

  cormat <- round(cor(cormatrix),2)

  melted_cormat <- melt(cormat)
  head(melted_cormat)

  # Get lower triangle of the correlation matrix
  get_lower_tri<-function(cormat){
    cormat[upper.tri(cormat)] <- NA
    return(cormat)
  }
  # Get upper triangle of the correlation matrix
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }

  reorder_cormat <- function(cormat){
    # Use correlation between variables as distance
    dd <- as.dist((1-cormat)/2)
    hc <- hclust(dd)
    cormat <-cormat[hc$order, hc$order]
  }

  # Reorder the correlation matrix
  cormat <- reorder_cormat(cormat)
  upper_tri <- get_upper_tri(cormat)

  # Melt the correlation matrix
  melted_cormat <- melt(upper_tri, na.rm = TRUE)

  # Create a ggheatmap
  ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
    geom_tile(color = "white")+
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") +
    theme_minimal()+ # minimal theme
    theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                     size = 12, hjust = 1))+
    coord_fixed()

  # Another refinement
  mainGraph <- ggheatmap +
    geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      panel.grid.major = element_blank(),
      panel.border = element_blank(),
      panel.background = element_blank(),
      axis.ticks = element_blank(),
      legend.justification = c(1, 0),
      legend.position = c(0.6, 0.7),
      legend.direction = "horizontal")+
    guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
  return(mainGraph)
}

g <- getCorMapYear(df)
```

Some codes by which I cleaned and prepared the data for visualization (all other codes are very similar looking with heavy usage of grepl(), gsub() and str_split()):

1. Making the “fuelSorted” column in df dataframe (mobile.de) dataframe:

```
dieselOrGasoline <- function(it) {
  res = ""
  if(
    grepl("Diesel", it, fixed=TRUE) &&
    grepl("Hybrid", it, fixed=TRUE)) {
    res = "Diesel Hybrid"
  } else if (grepl("Diesel", it, fixed=TRUE)) {
    res = "Diesel"
  } else if(
    grepl("Benzin", it, fixed=TRUE) &&
    grepl("Hybrid", it, fixed=TRUE)
  ) {
    res = "Benzin Hybrid"
  } else if(
    grepl("Benzin", it, fixed=TRUE)
  ) {
    res = "Benzin"
  } else if(
    grepl("LPG", it, fixed=TRUE) ||
    grepl("CNG", it, fixed=TRUE)
  ) {
    res = "LPG/CNG"
  } else {
    res= "Andere"
  }
  return(res)
}

# the fucntion argument is the Kraftstoffart
# column of mobile.de dataframe

newFuelVec <- function(vec){
  newVec <- ""
  for(i in 1: length(vec)){
    newVec[i] <- dieselOrGasoline(vec[i])
  }
  return(newVec)
}
```

2. Code for the “Taxi” column in the dataframe:

```
# This function takes the whole
# dataframe as argument and returns
# a vector containing whether a
# car at a particular index is a
# taxi or not
taxiColor <- function(df){
  res <- NULL
  for(i in 1 : length(df$Name)) {
    print(i)
    if (grepl("hellelfenbein", df$FarbeHersteller[i], fixed=TRUE))
{ #HELLELFENBEIN
      res[i] = "Taxi"
    } else if (grepl("HELLELFENBEIN", df$FarbeHersteller[i], fixed=TRUE)) {
      res[i] = "Taxi"
    } else if (grepl("Beige", df$Farbe[i], fixed=TRUE)) {
      res[i] = "Taxi"
    } else if (grepl("Taxi", df$Name[i], fixed=TRUE)) {
      res[i] = "Taxi"
    } else {
      res[i] = "Not Taxi"
    }
  }
  return(res)
}
```

3. Code for cleaning the mileage data right after scraping the mobile.de dataset

```
# Since if there is an empty string in
# the km field then it is a new car
# 0 kms hence
df[df$Kilometerstand == "",]$Kilometerstand = "0 km"
cleanKilometersData <- function(kms) {
  for(i in 1 : length(kms)){
    # if it is NA let it be an empty string
    if(!is.na(kms[i])) {
      last_two_chars = substr(kms[i], nchar(kms[i])-2+1, nchar(kms[i]))
      last_char = substr(kms[i], nchar(kms[i]), nchar(kms[i]))
    } else {
      last_two_chars = ""
      last_char = ""
    }
    if (last_two_chars == "km") {
      kms[i] = substr(kms[i], 1, nchar(kms[i])-3)
      kms[i] = gsub(".", "", kms[i], fixed = TRUE)
    } else {
      if(last_char == "k") {
        kms[i] = substr(kms[i], 1, nchar(kms[i])-2)
        kms[i] = gsub(".", "", kms[i], fixed = TRUE)
      } else {
        kms[i] = NA
      }
    }
  }
  return(kms)
}
df$Kilometerstand <- cleanKilometersData(kms = df$Kilometerstand)
df$Kilometerstand <- as.integer(df$Kilometerstand)
```