USED CAR DATASET

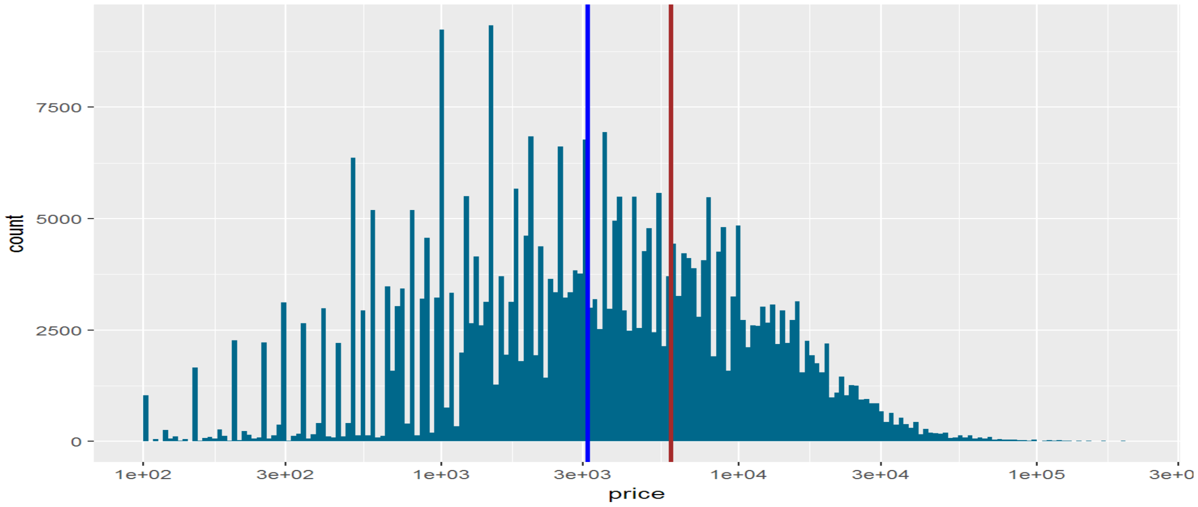
Problem Statement: To predict the price of used cars.

Content: There are 20 variables in the given dataset which are as follow as:

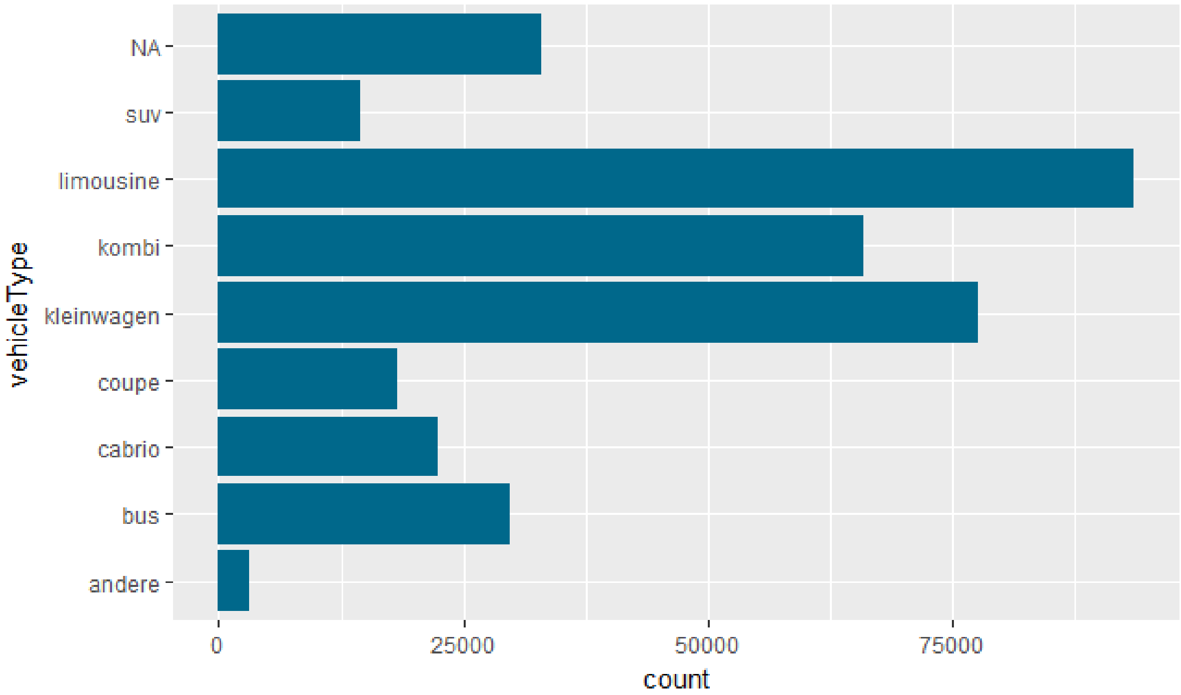
- dateCrawled : The date on which car is registered
- name : "name" of the car
- seller : private or dealer
- offerType : Angebot(proposal) (Gesuch(**request)**
- price : the price on the ad to sell the car
- vehicleType : Bus, Cabrio etc.
- yearOfRegistration : at which year the car was first registered
- gearbox : automatic, Manual
- powerPS (horse strength): power of the car in PS
- model: Niva, Navara etc.
- kilometer : how many kilometers the car has driven
- monthOfRegistration : at which month the car was first registered
- fuelType : Petrol, Diesel
- brand : Audi, BMW
- notRepairedDamage : if the car has a damage which is not repaired yet
- dateCreated : the date for which the ad at ebay was created
- postalCode
- lastSeenOnline : when the crawler saw this ad last online

Analysis of Dataset

- There are random and empty values in the price columns. I dropped observations which are above €200,000 and below €100, assuming the majority of them would be input errors.
- There are some brands which are not premium, but present in premium category (price is more than €75000. Therefore, I removed all these values.

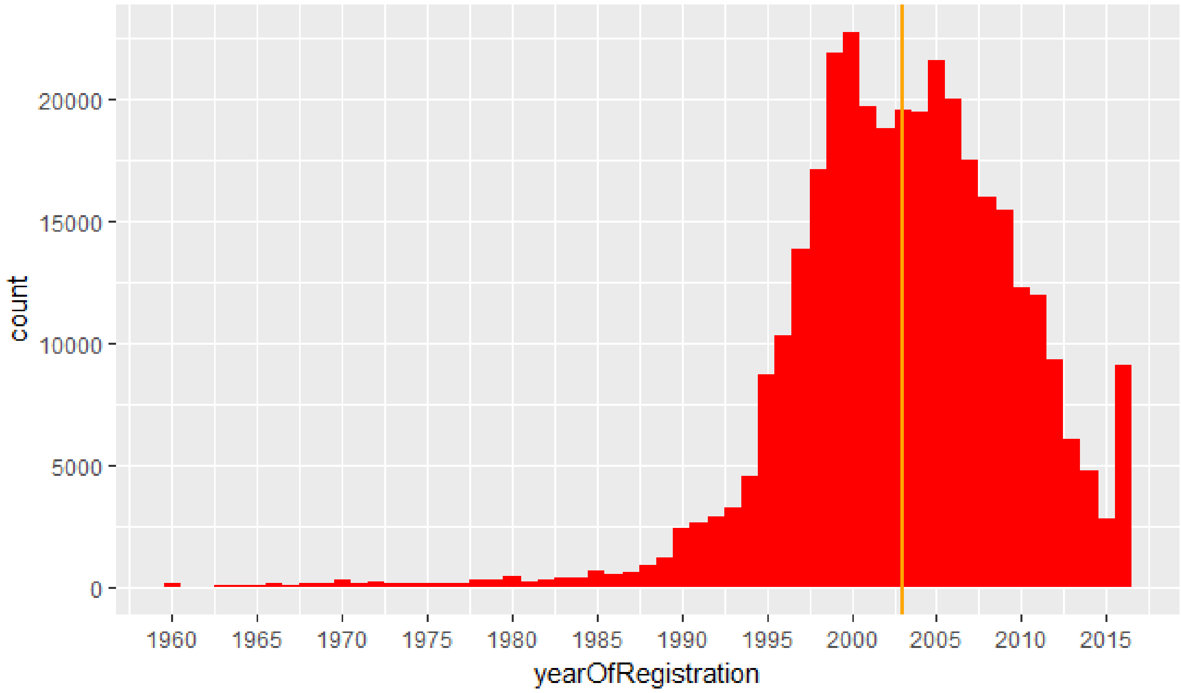


- I plotted the graph for vehicle type. We can infer from the graph that limousine is most popular and andere is least popular in Germany.

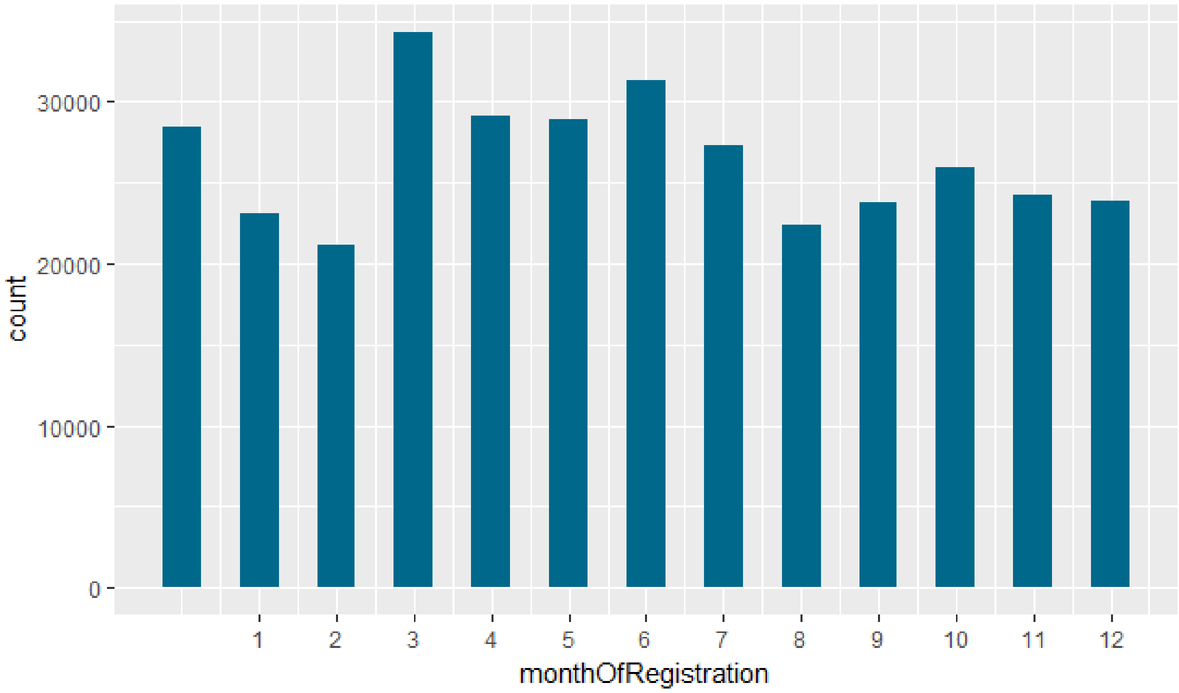


Year of registration:

- We can conclude from the summary of the "year of registration" that maximum and minimum year is 9999 and 1000 respectively which is wrong. Therefore, I dropped all the values which are not in the range of 1960 to 2016.
- From the graph, we can visualise that there are four peak years (1999, 2000, 2005 and 2006) for new car registration in the Germany.

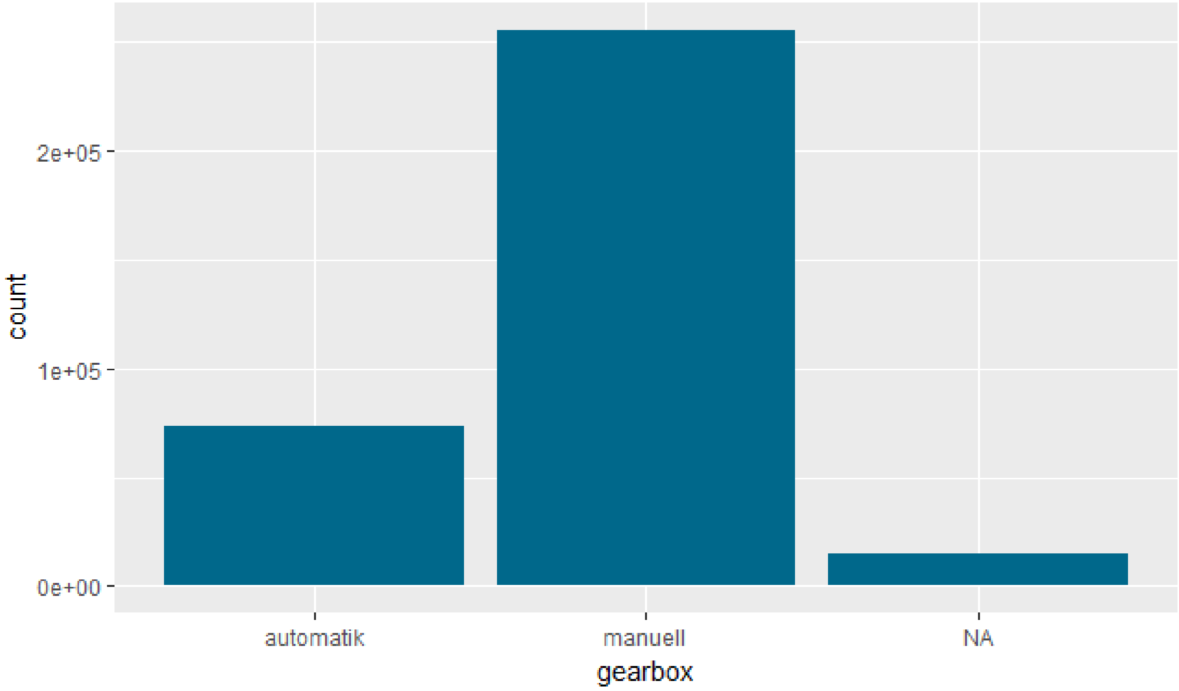


- Month of registration: we can also visualise from the graph that March and June are the strongest for vehicle registration.



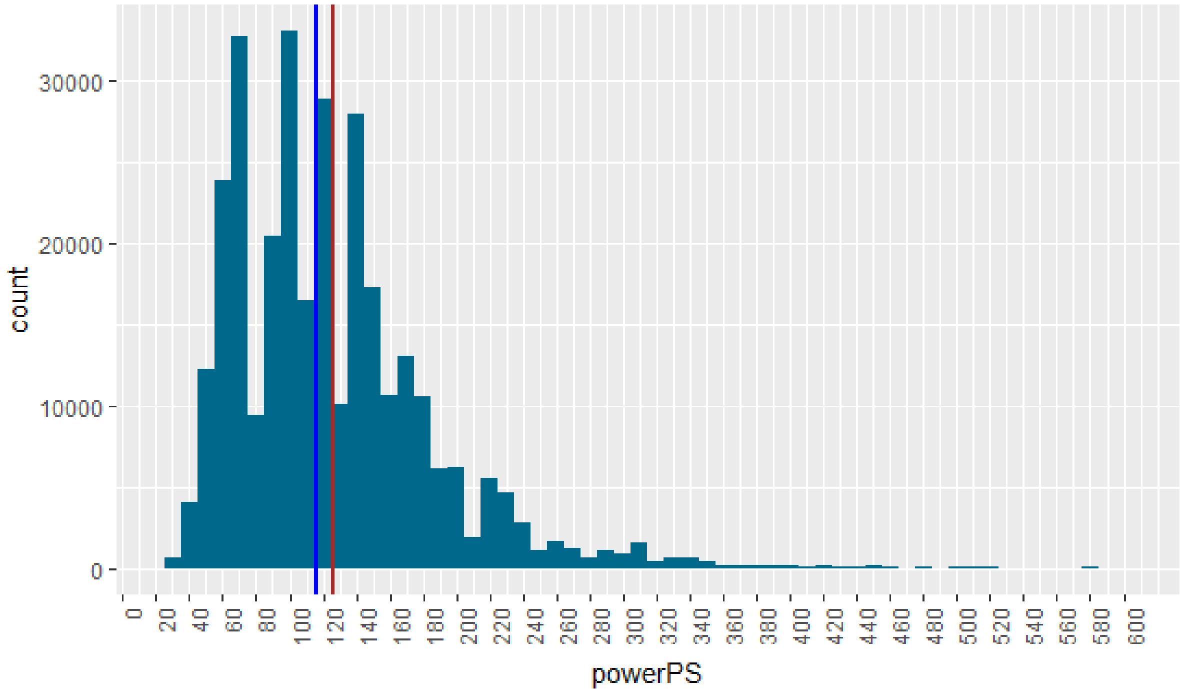
Transmission Type

- We can see from the graph that the European market is primarily a manual transmission market.



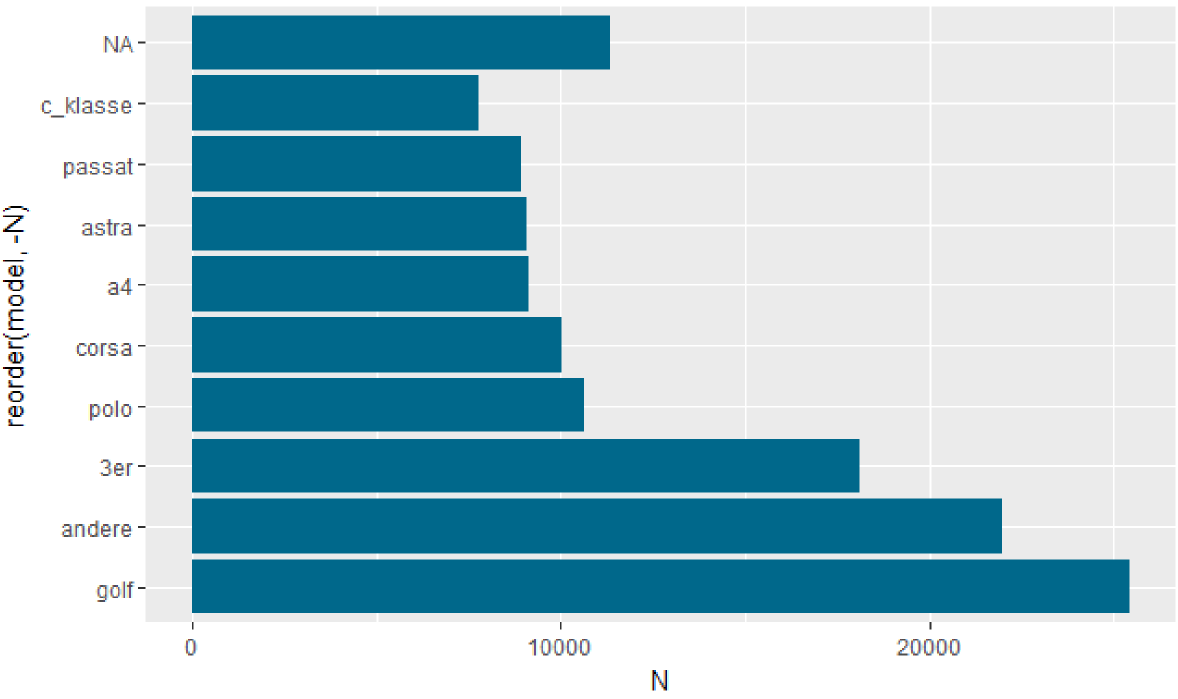
Engine Power

- There are many random values in the engine power column. I dropped all the values which are not in the range of 25 to 600 PS because; all the cars have engine power in this range.



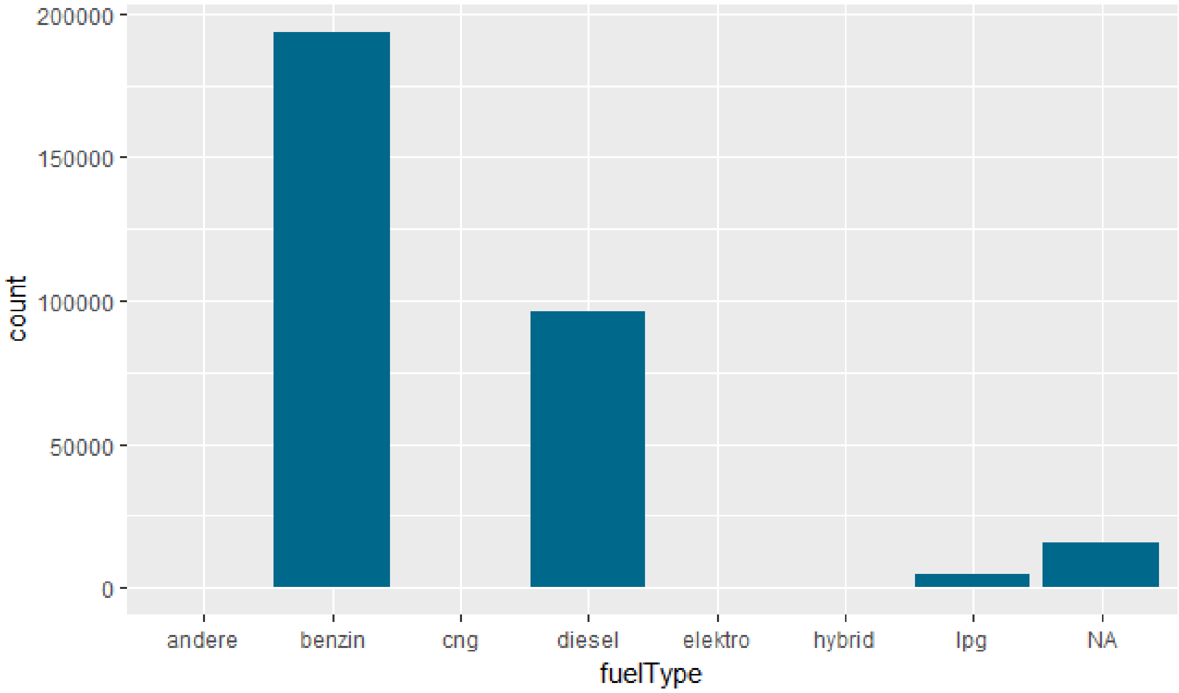
Model

- There are 251 models, but I plotted the graph of top 20 based on popularity. We can infer from the graph that golf model is popular in Europe.



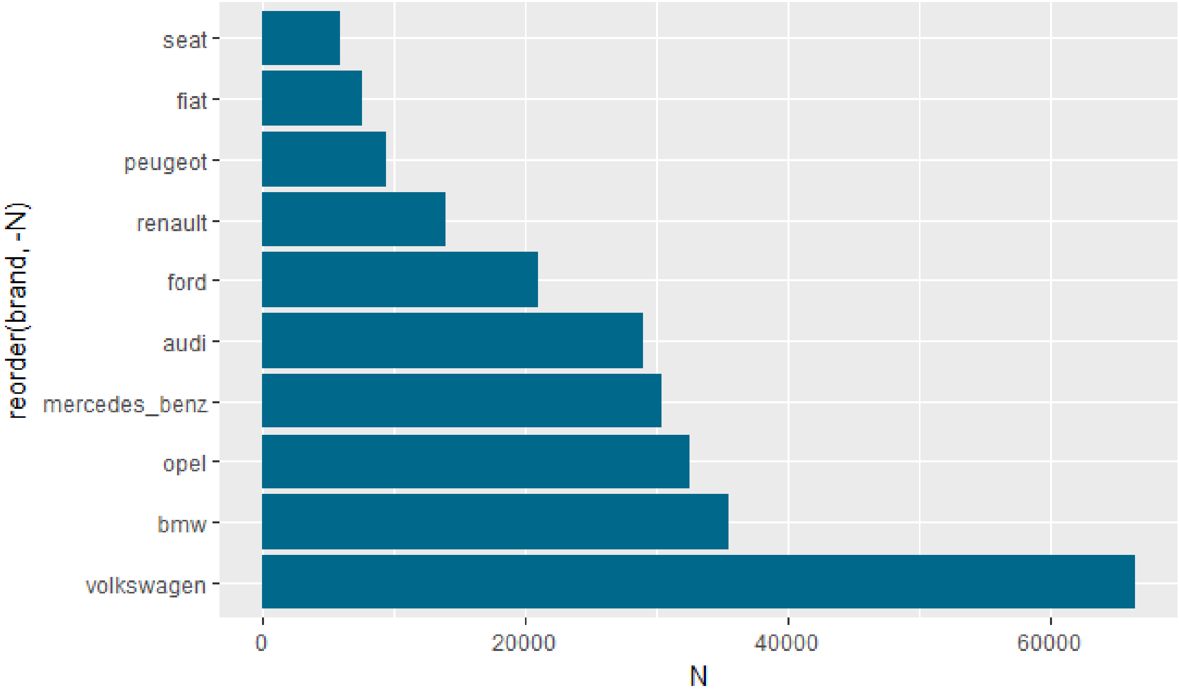
Source of Energy

- Alternative sources of energy are almost negligible in this dataset, which is not surprising considering that over 50% of the vehicles were 12 to 13 years old when this data was extracted.
- Petrol is roughly twice as prominent as Diesel



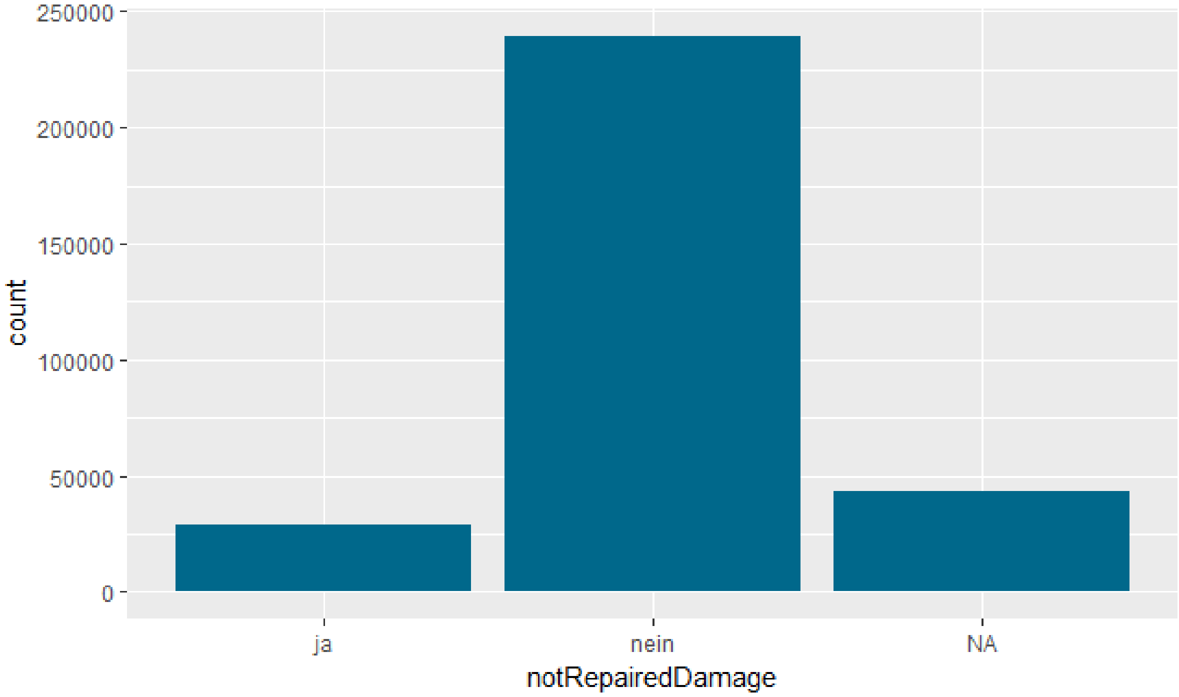
Brand

- There are 40 brands but, I plotted top 10 brands. The top 5 brands are German. The number 6 belongs to Ford, which is in Europe and is largely perceived as German as it has its European headquarters in Cologne and many of its European products are actually designed and built in Germany. The next two brands are French, and then Fiat is Italian. Seat is Spanish but it is actually part of the VW Group and their cars share almost all their components with VW products. In other words, German manufacturers are hugely dominant on their home turf.



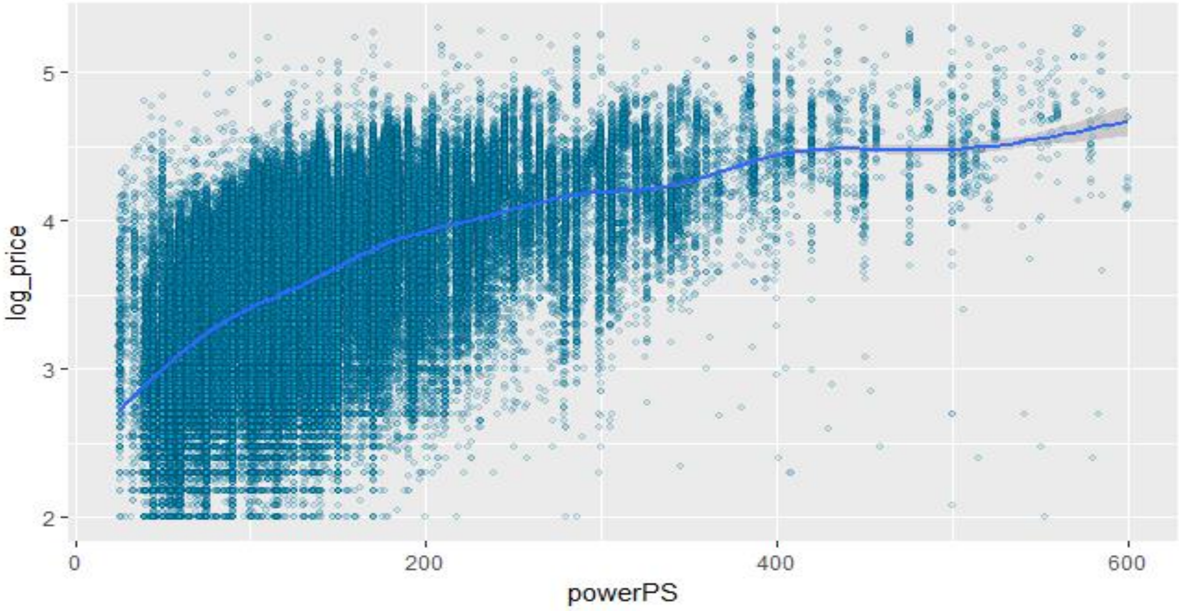
NotRepairedDamage

- The variable “notRepairedDamage “can only take two values: “yes” or “no”. But it does have NAs – about 72,000, which is twice as many as the number of “yes”. It does not seem like this is a mandatory field. As far as my understanding, it refers to potential unrepaired damage on the vehicle being sold.



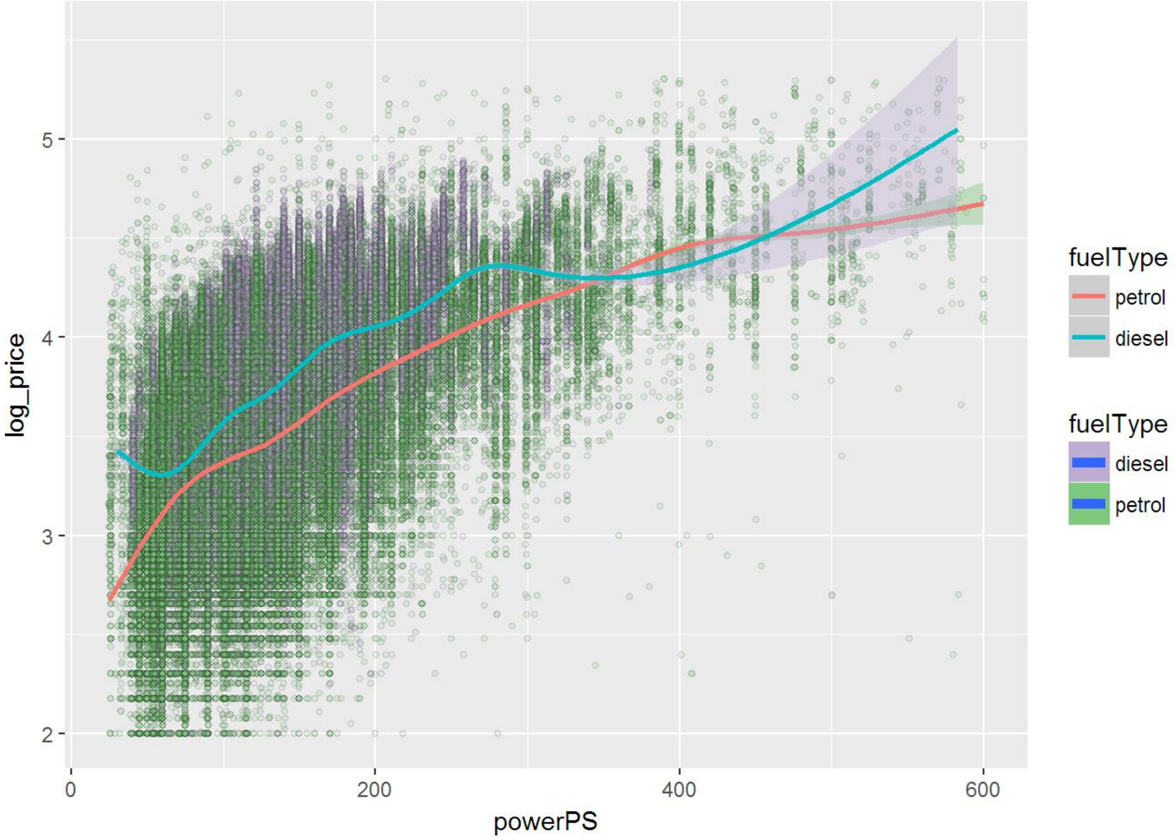
Price vs Power

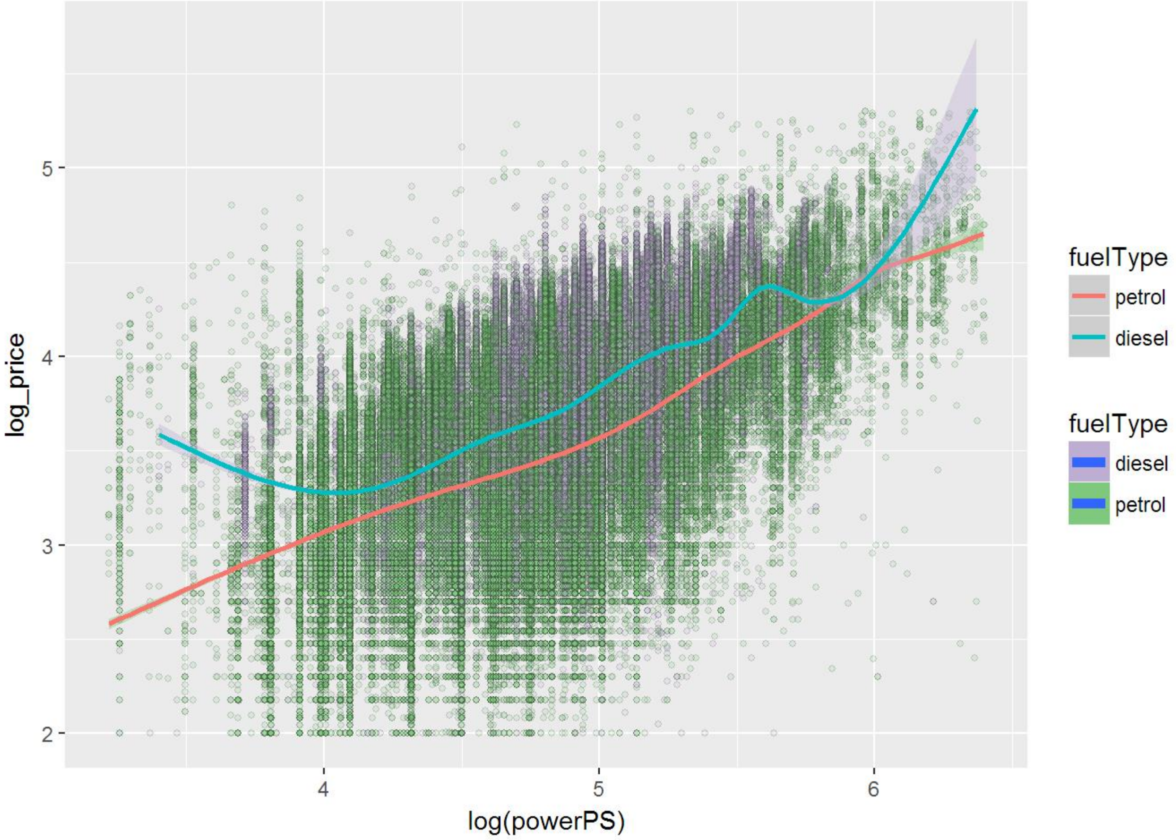
- I plotted the log price vs PowerPs. From the graph, we can conclude that there is correlation between log price and PowerPs along with lot of noise.



Price vs Power vs Fuel

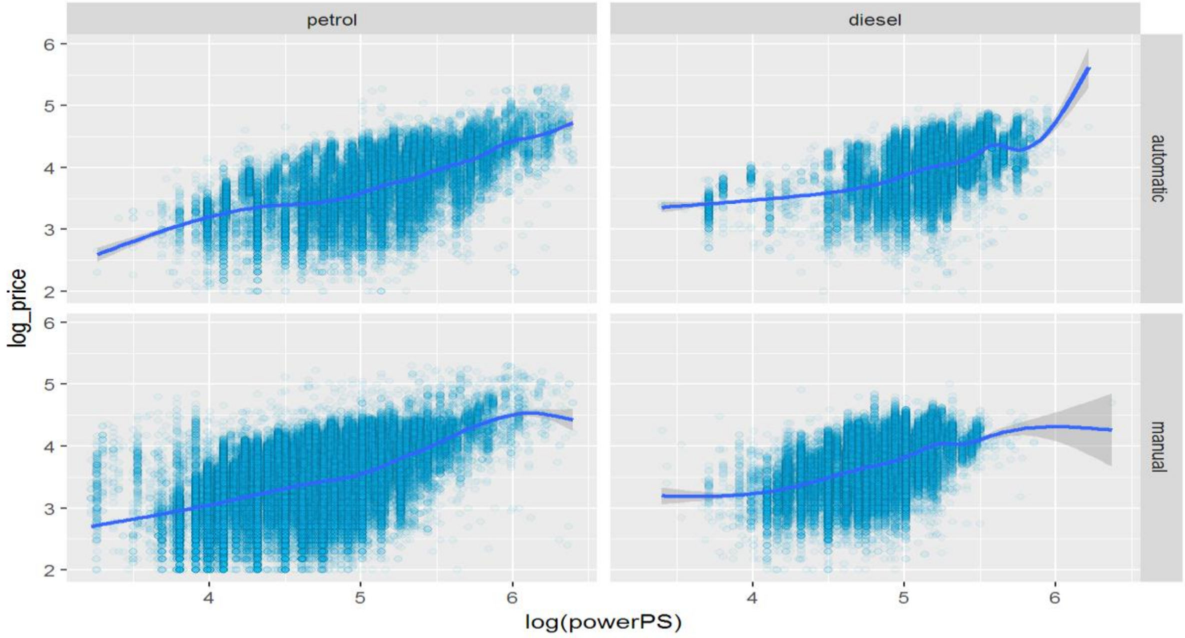
- As we know that diesel engine are more expensive than petrol, keeping power as constant. When I plotted the graph, I got the same result.





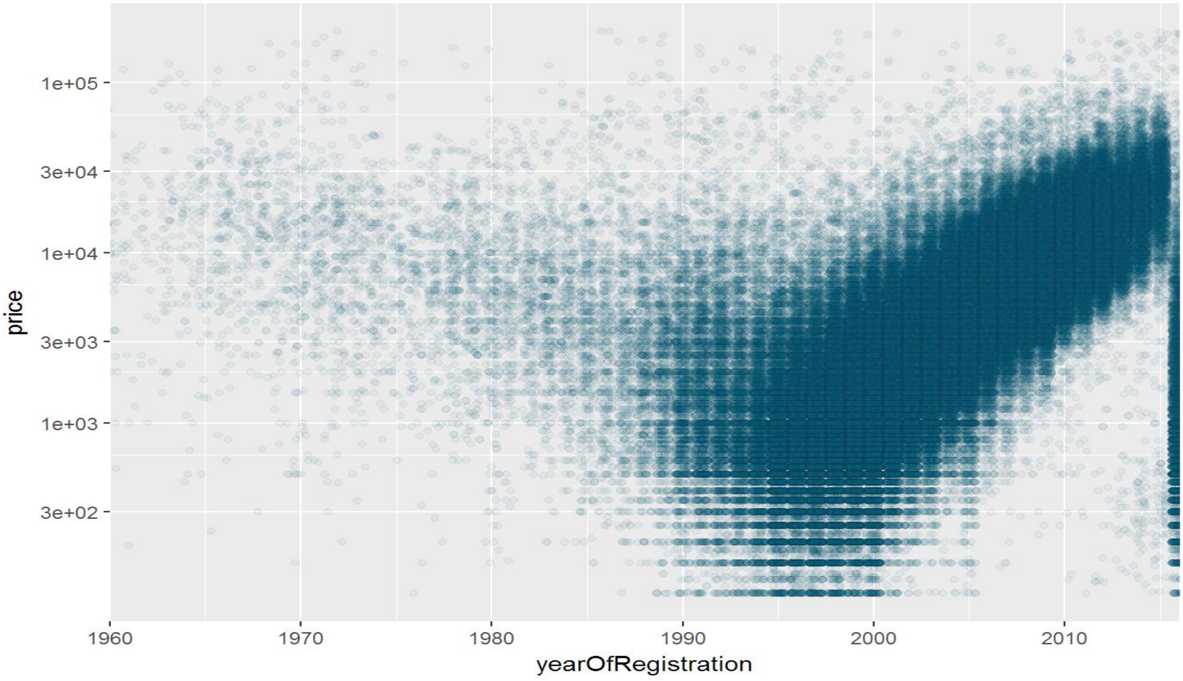
Price vs Power vs Fuel Type vs Gearbox

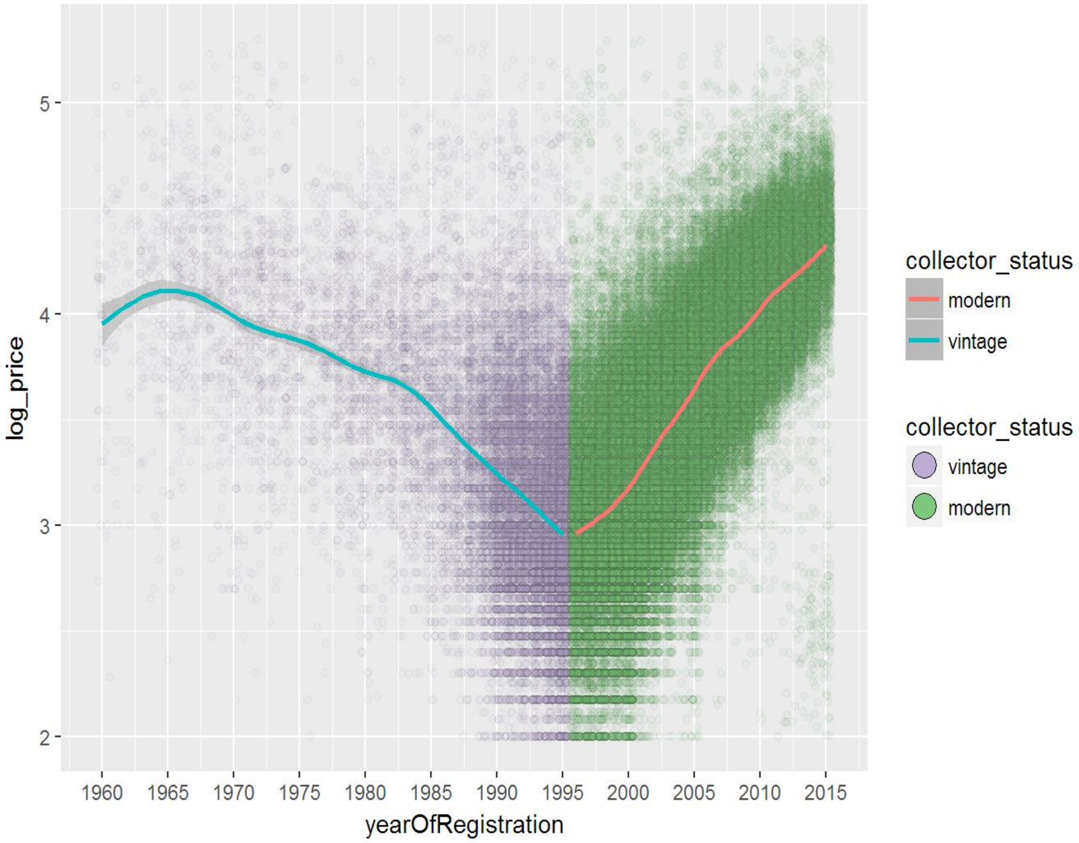
- We can see from the graph that petrol cars have more variance than Diesel cars in terms of price and power.
- We can also see that automatic cars are more powerful and expensive than Manual cars.



Price vs Year of Registration

We can see from the graph that correlation is good after 1992.but; there is one outlier which is 2016.We can also infer that cars are more expensive before 1990 than after.

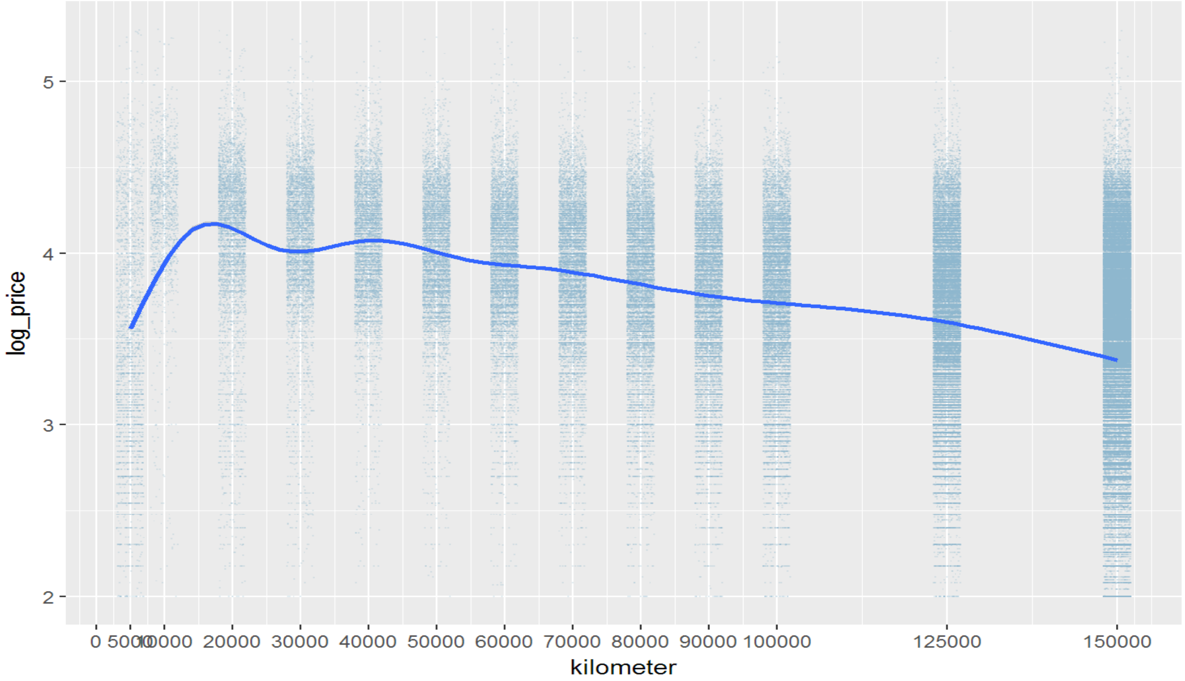




- We can also see that the correlation is lower in vintage car data and higher in modern car data.

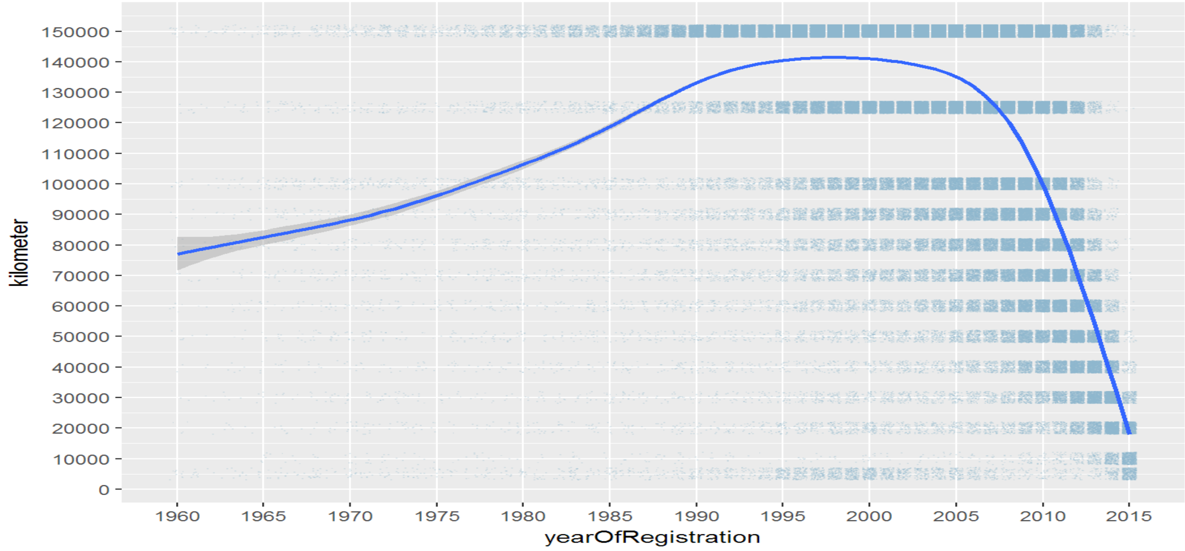
Price vs Mileage

- We can see from the graph that 5000km data is highly suspicious.



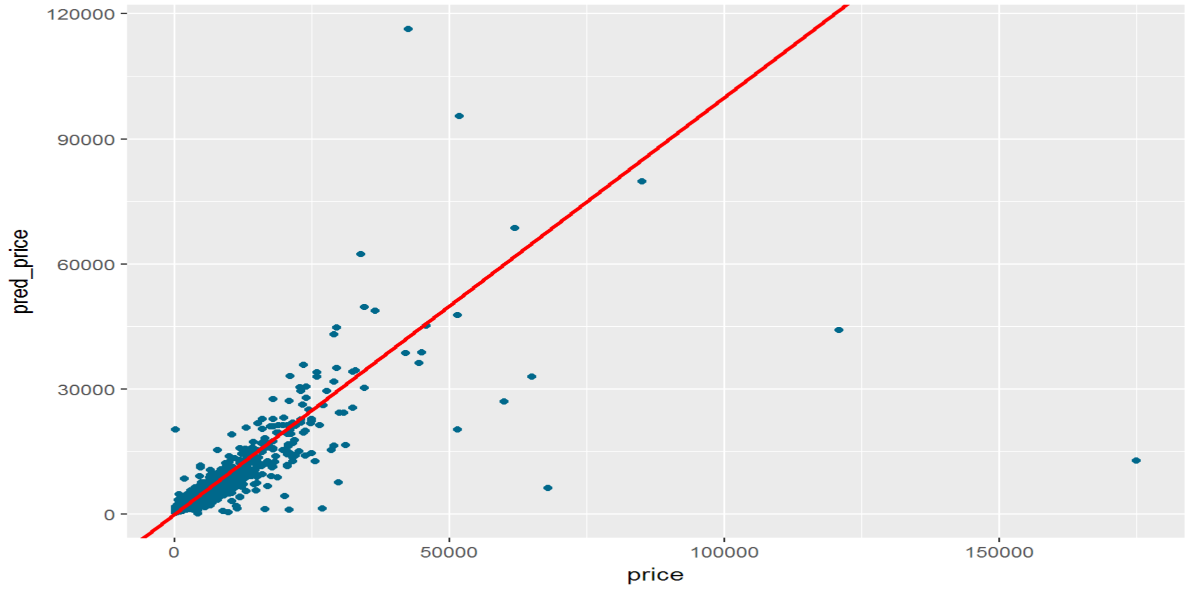
Year of registration vs Mileage

- We can see from the graph that kilometres tend to be positively correlated to the age of car till 15-20 years. Then after, there is negative correlation.
- We can also infer that vintage cars don't tend to run as much as household's main car. Modern cars generally have a higher mileage as they get older.



Linear Regression

- R- Square value is 0.75.
- I selected few data at random and treated as test data. I plotted the graph between predicted price and original price. We can see from the graph that there is more error when price is high.


pred_price
price