Used cars dataset: Exploratory Data Analysis

Sabbella Prasanna

3/10/2021

The used cars data set has been sourced from kaggle, uploaded by a freelance data scientist who prepared this data from various sources on personal interest. https://www.kaggle.com/lepchenkov/usedcarscatalog

Scope

CONTEXT

- Carvana is an online platform that allows users to sell/trade their cars. The marketing strategists of Carvana knows that the success and popularity of their online brand highly depends on the number of customers who sell and also buy used cars through Carvana.
- It is often difficult for any customer to decide on what’s the best sale price of their car and would end up either selling it at a very lower price than at what it could have been sold or at a rather higher price than usual. In both cases either of the buying or selling customers are dissatisfied. This leaves a negating effect on the Carvana’s customer base.
- In order to attract both selling and buying customers the strategists came up with a unique idea, to invest in launching a new feature on their app/website that allows the customer to know what could be the best price to sell his/her car.

NEED

- The strategists wanted to suggest the selling price to their new selling customers in a way that attracts both buyers and sellers so that it can end up as a win-win for both parties.
- They would like to investigate on what major aspects/features of a car does a buyer really spends his money and then match these parameters with the car that is about to be listed on sale and make a rightful selling price suggestion.

VISION

- The strategists will make use of already available data history of sold out cars and investigate what features are really causing a shift in the selling price and measure the sensitivity of change for these features so that they could build a predictive model for predicting the price. This predicted price is then used as a selling price suggestion in the new feature.

OUTCOME

- The predictive model will be summarized as a report to the board members of Carvana who could further test/validate the price suggestion feature (Probably as a trial feature for one quarter) before its launch. This sheds light on how impactful the feature is on increasing the

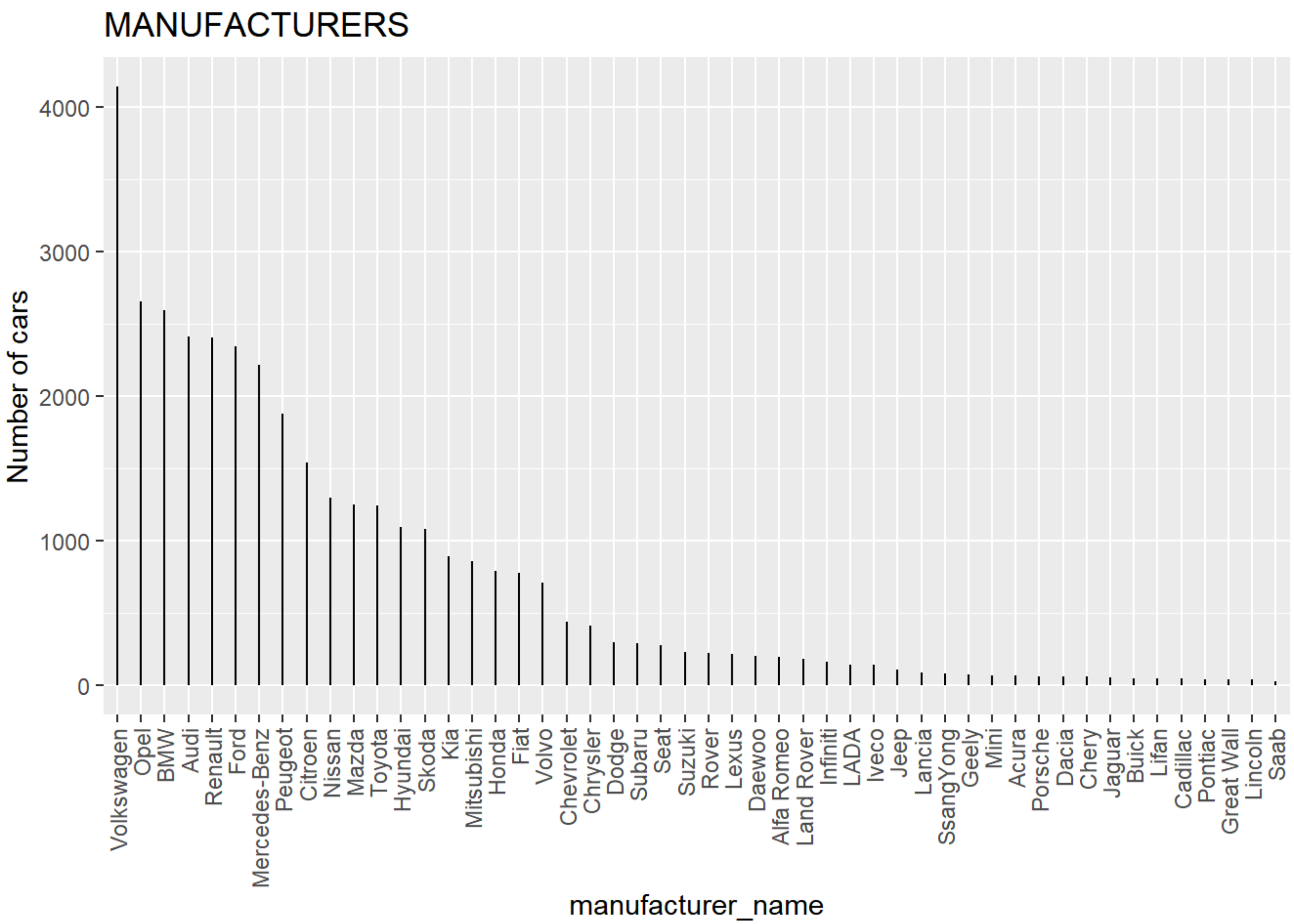
customer base of Carvana.
- Once convinced, they would invest higher volumes of budget into the project for gathering more data and make the app more reliable.

Introduction to the data variables

- Number of cars (Discrete Numeric): Number of cars on sale.
- transmission (Categorical): There are two types of transmission…automatic & manual
- color (Categorical): Color of the outer body of the car
- body_type (Categorical): Body type of the car
- engine_type (Categorical): There are two types of engines…Gasoline and Diesel
- engine_fuel (Categorical): Different types of fuels
- drivetrain (Categorical): 3 types of drives are present..Front, Rear and All wheel drive
- location_region (Categorical):There are 6 different locations where the cars are listed on sale.
- has_warranty(Categorical):Shows whether a car on sale has a warranty
- (Predictor)price_usd(usd)(continuous Numeric):Sale price of the car
- year_produced (Discrete Numeric):Year when the car got manufactured.
- duration_listed (days) (Continuous Numeric):For how long was the deal on sale?
- engine_capacity (Liters) (Continuous Numeric):Volume swept by all pistons in one engine
- odometer_value (Kilometers) (Continuous Numeric):Total distance covered by the car.
- feature_0 to 9: Features such as Alloy wheels and other accessories in the car that adds on more price to it. # Summary is attached at the end of the document.

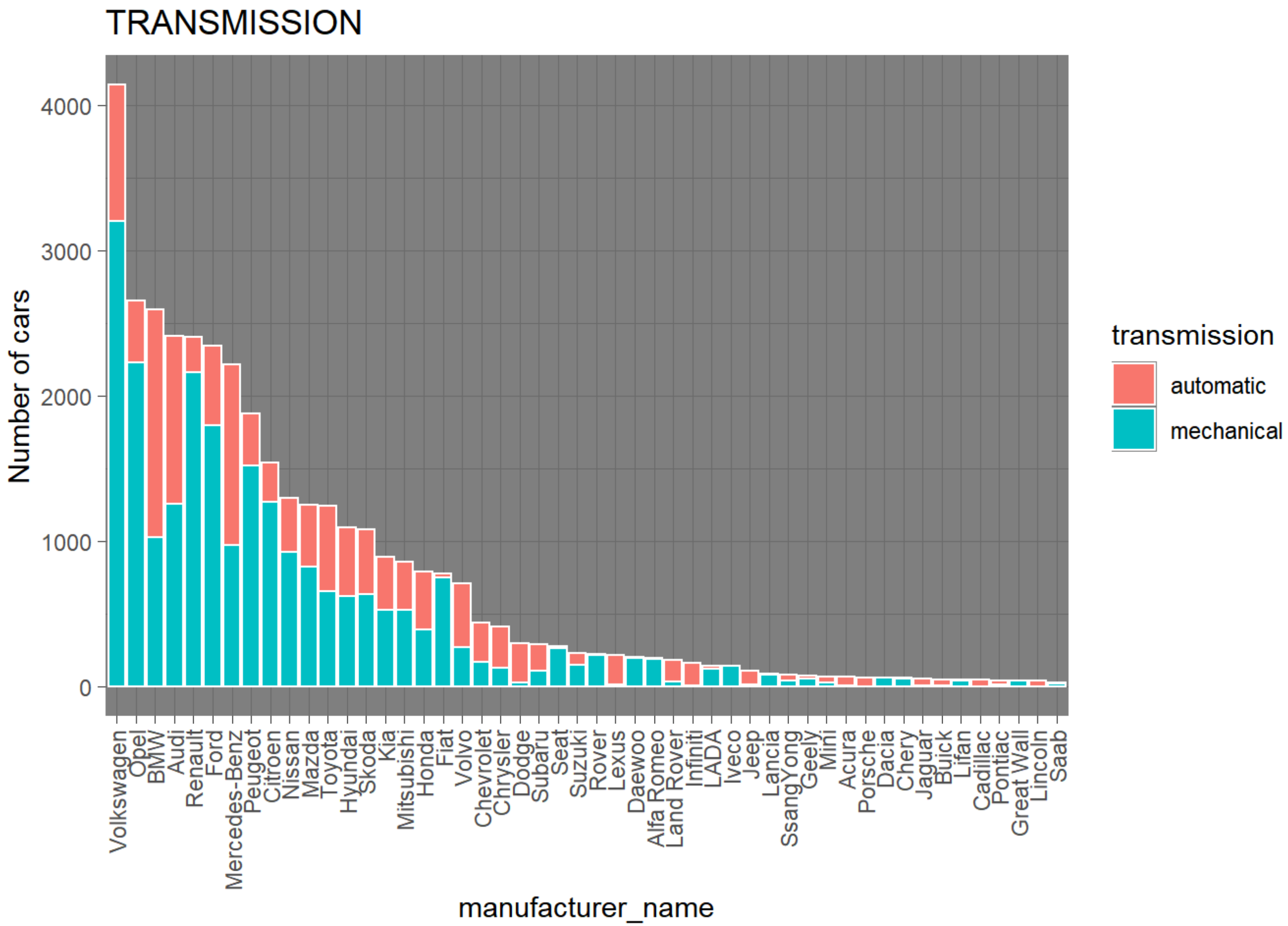
Number of cars (Discrete Numeric): Number of cars on sale.
manufacturer_name (Categorical): Brand name

- Volkswagen owns the highest share while Saab the lowest


MANUFACTURERS
Number of cars
4000
3000
2000
1000
0
manufacturer_name
Volkswagen
Opel
BMW
Audi
Renault
Ford
Mercedes-Benz
Peugeot
Citroen
Nissan
Mazda
Toyota
Hyundai
Skoda
Kia
Mitsubishi
Honda
Fiat
Volvo
Chevrolet
Chrysler
Dodge
Subaru
Seat
Suzuki
Rover
Lexus
Daewoo
Alfa Romeo
Land Rover
Infiniti
LADA
Iveco
Jeep
Lancia
SsangYong
Geely
Mini
Acura
Porsche
Dacia
Chery
Jaguar
Buick
Lifan
Cadillac
Pontiac
Great Wall
Lincoln
Saab

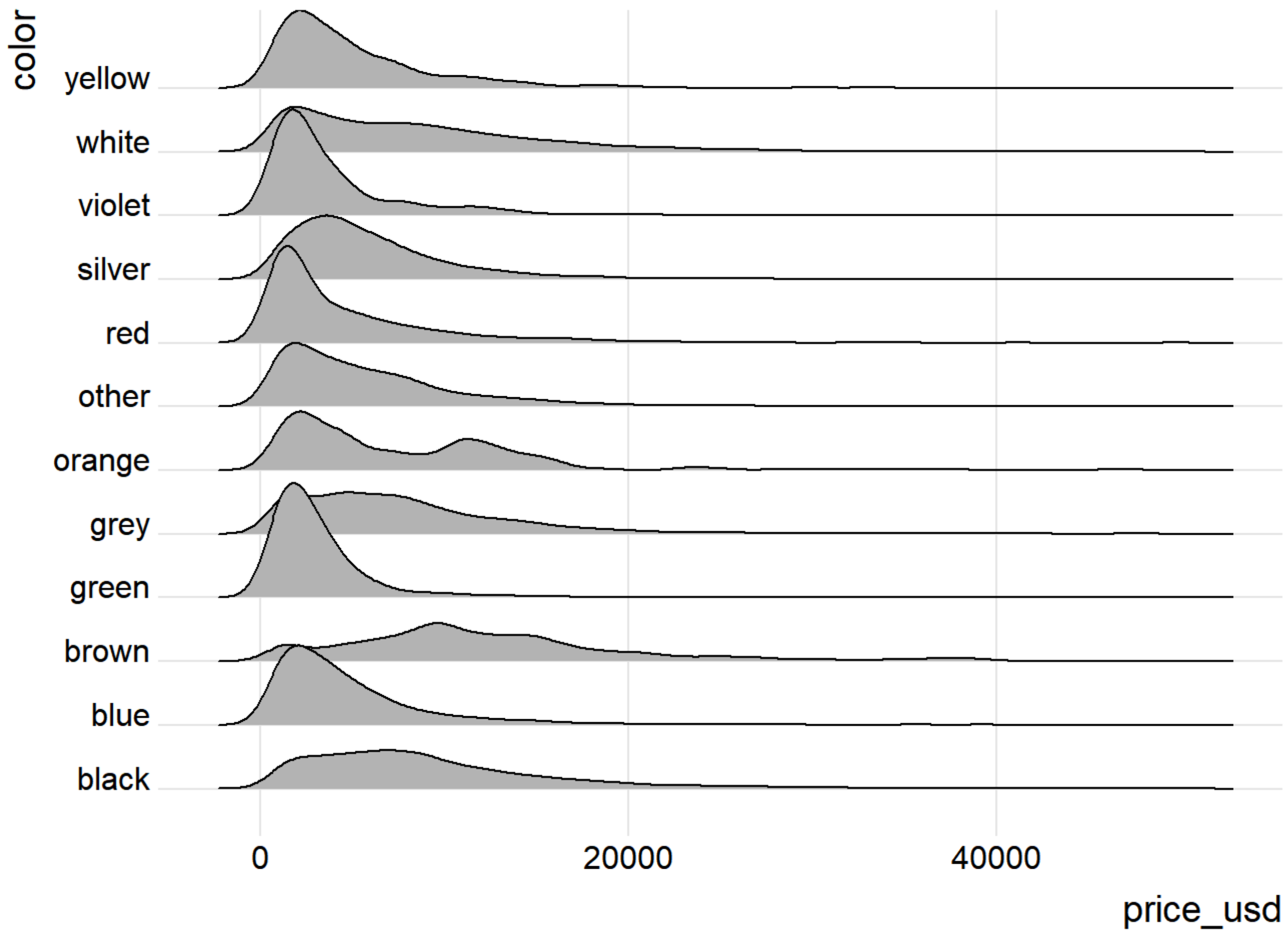
transmission (Categorical): There are two types of transmission…automatic & manual

- Most brands are giving equal importance to manual and auto trains but manual cars are more frequent



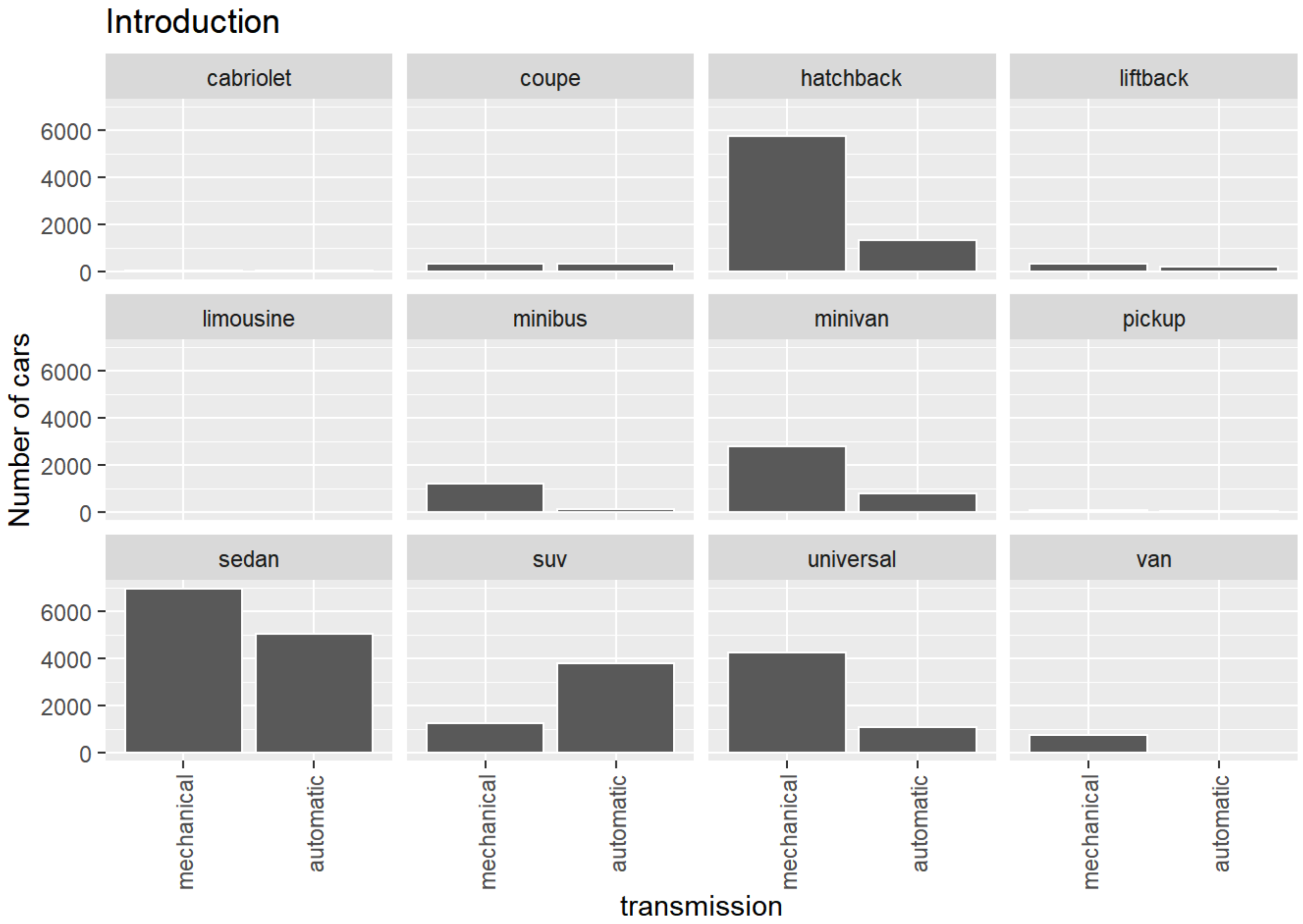
color (Categorical): Color of the outer body of the car

- More obsessed with Black, White, Silver, Blue



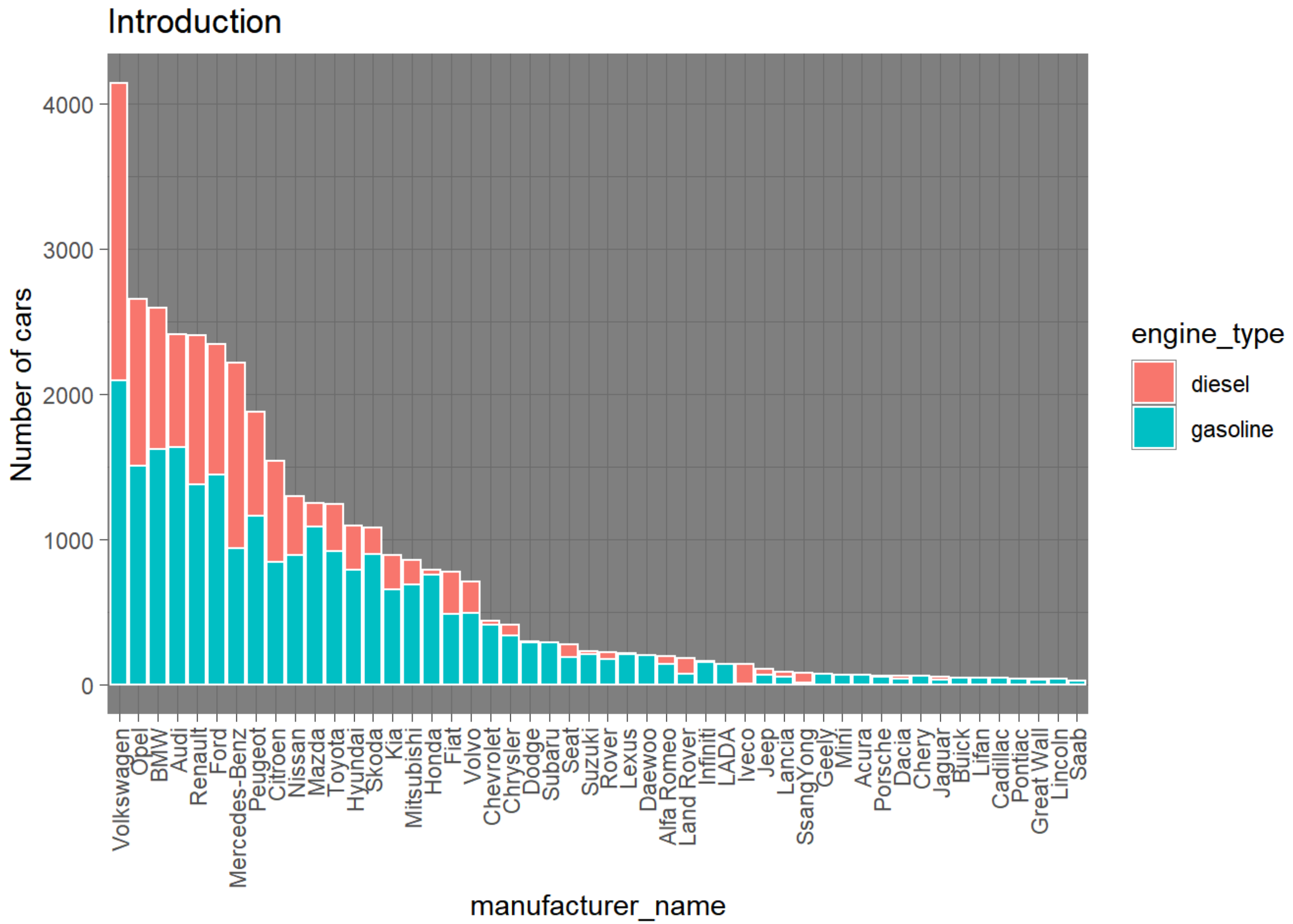
body_type (Categorical): Body type of the car

- Universal, Hatchback and Sedan are the most common



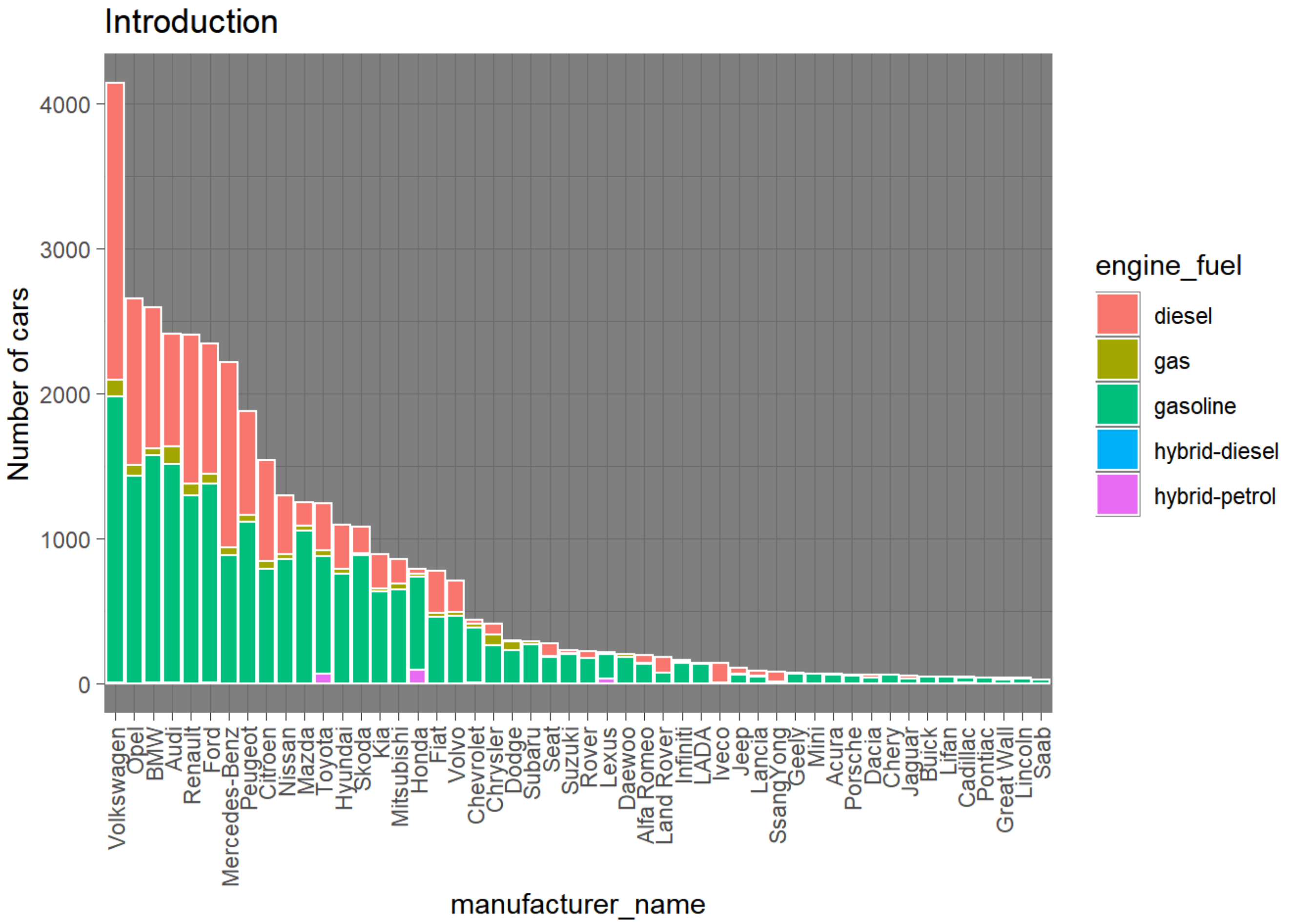
engine_type (Categorical): There are two types of engines…Gasoline and Diesel

- Gasoline engine is most frequently used

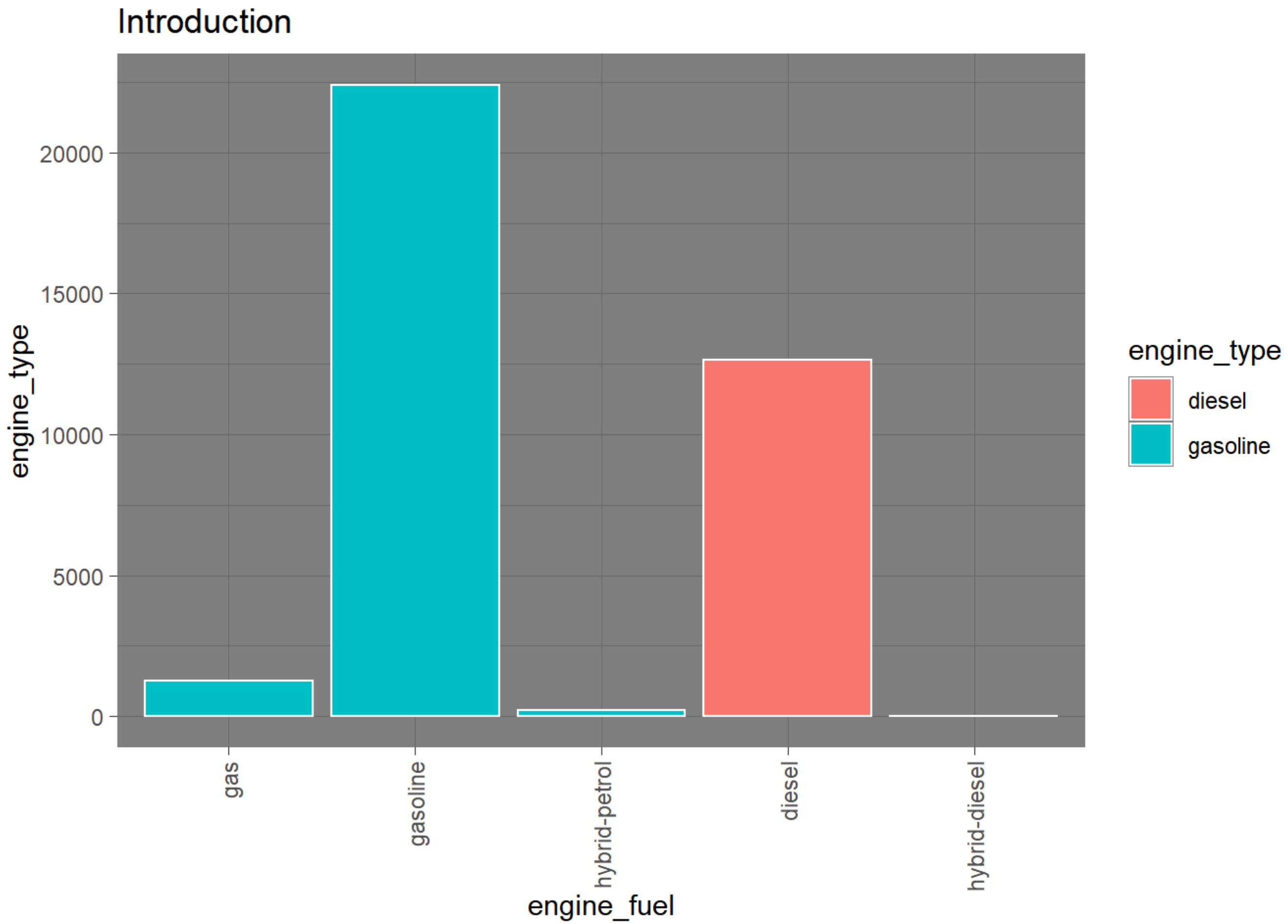


engine_fuel (Categorical): Different types of fuels

- Gasoline and Diesel are frequently consumed.

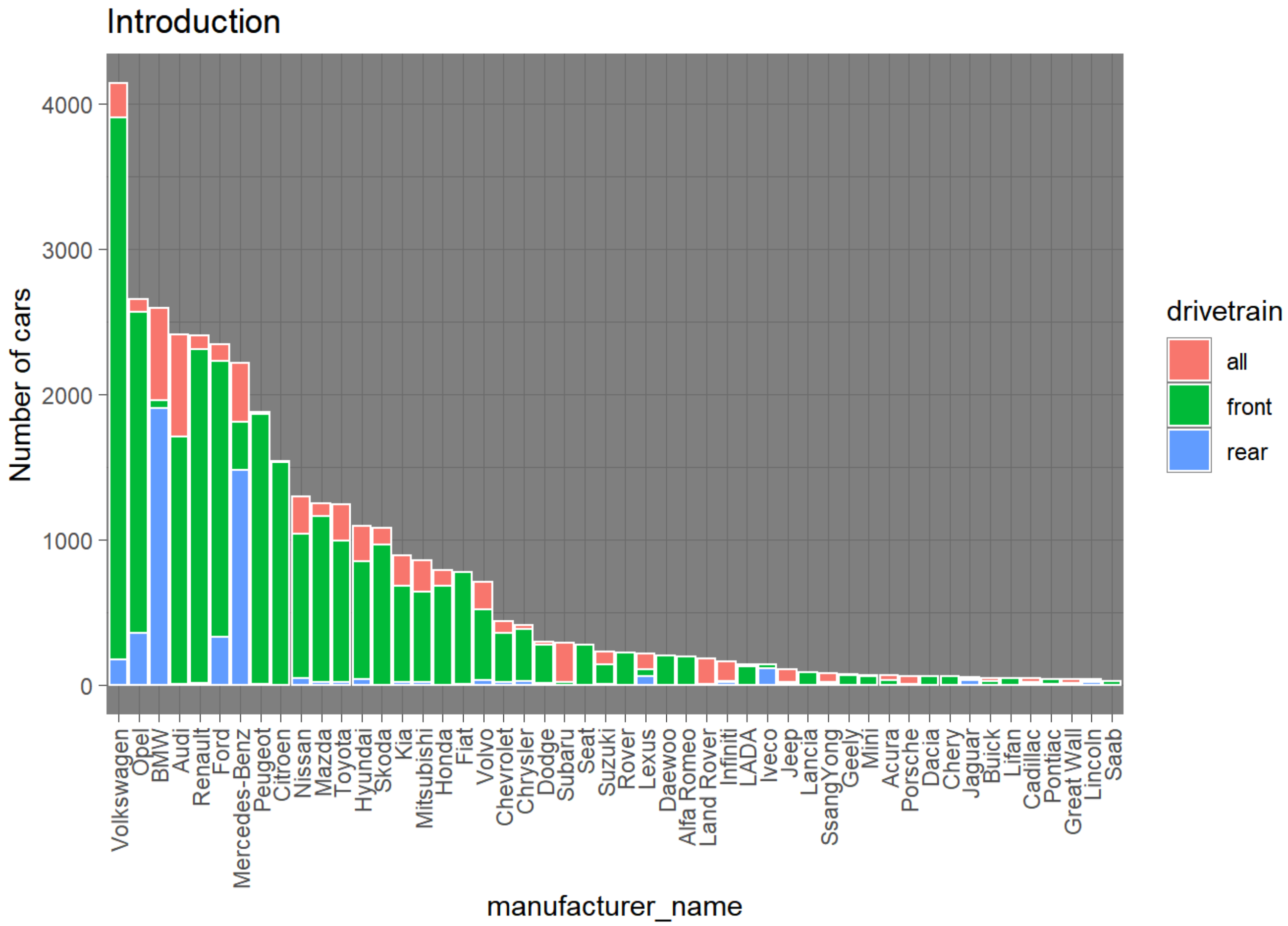


General Knowledge: What fuels are fed in what engines

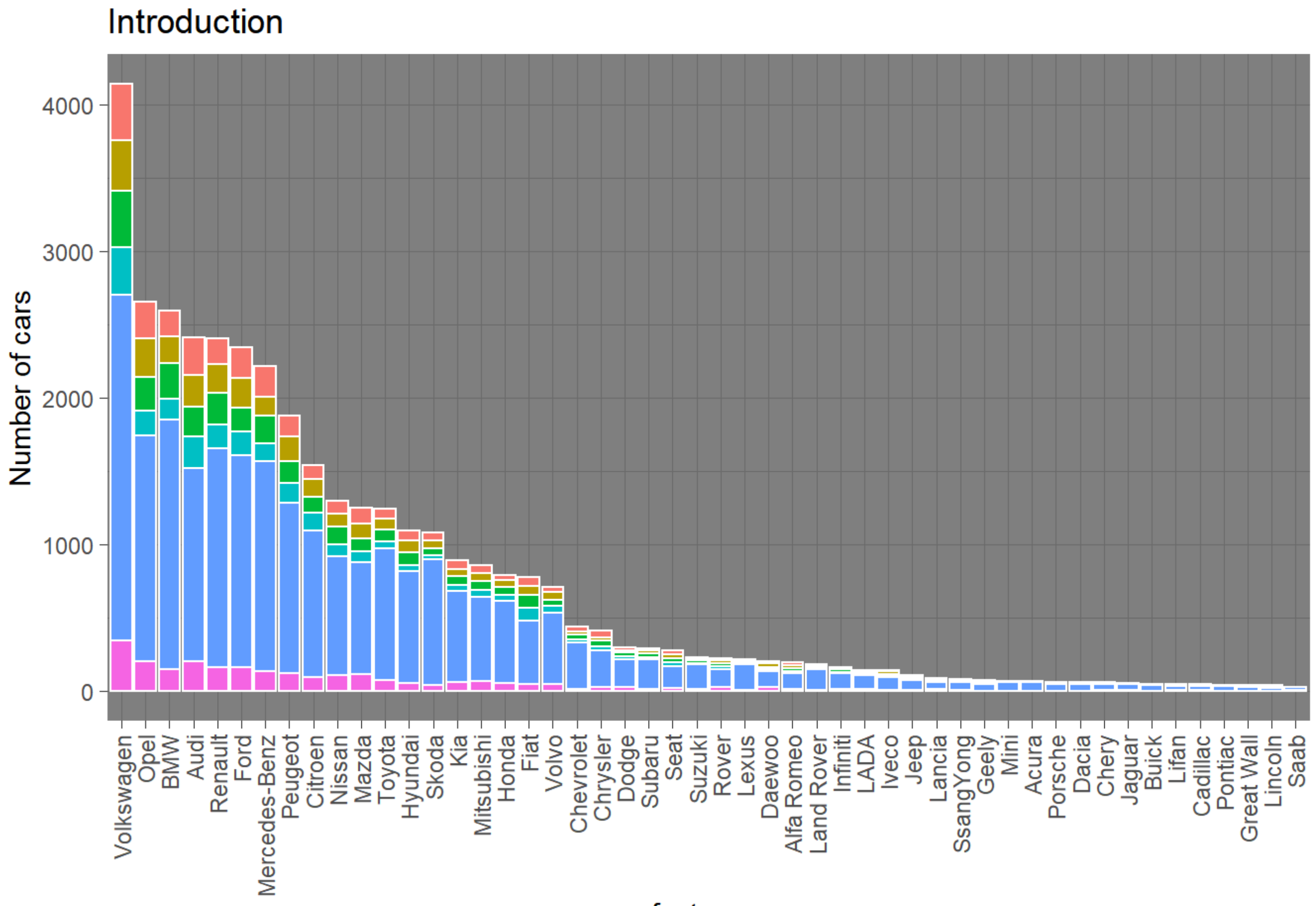


drivetrain (Categorical): 3 types of drives…Front, Rear and All wheel drive

- Front wheel drive is the most common.

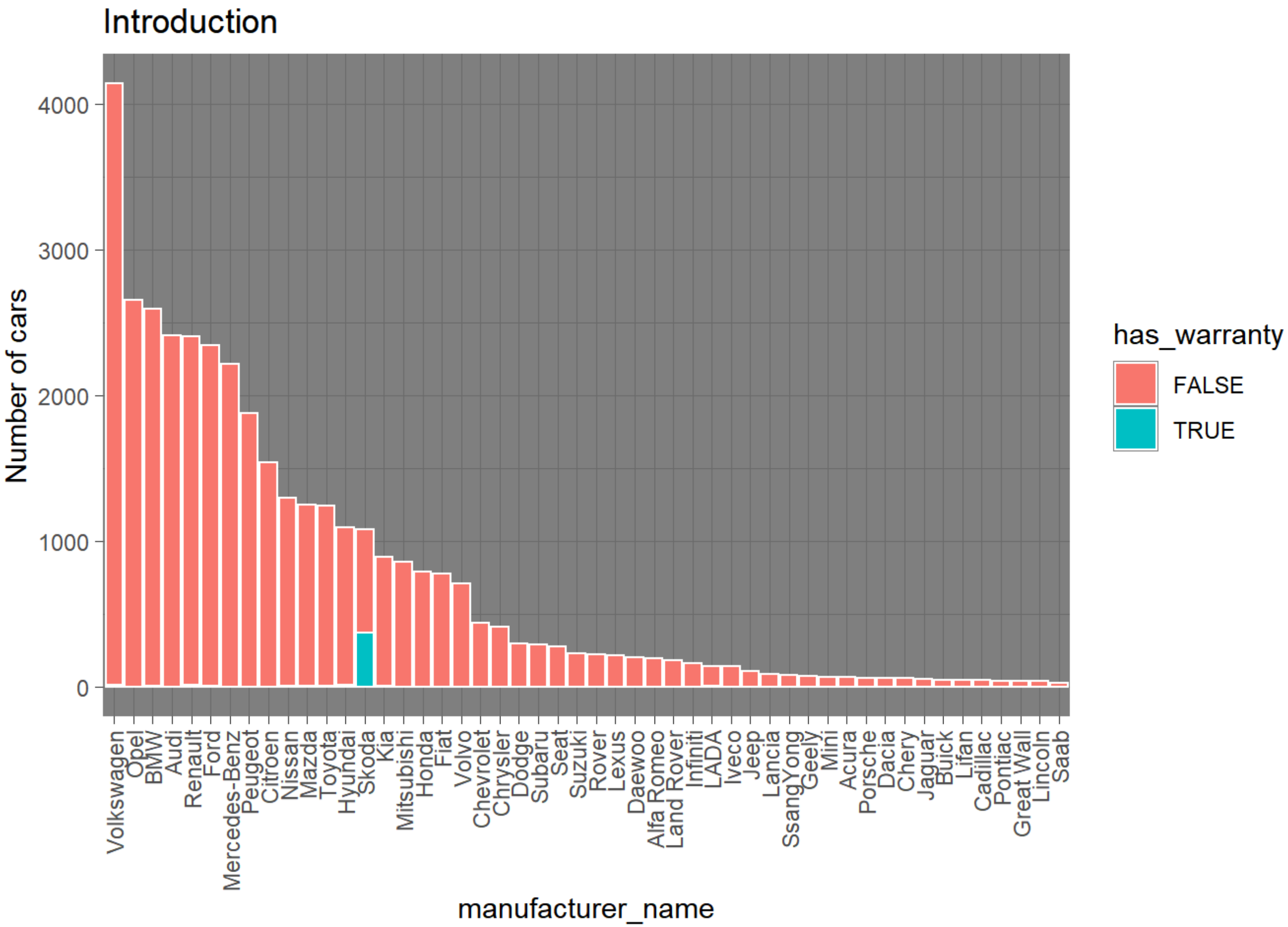


location_region (Categorical): There are 6 different locations where the cars are listed on sale.



has_warranty(Categorical):Shows whether a car on sale has a warranty.

- Most brands don’t except for Skoda



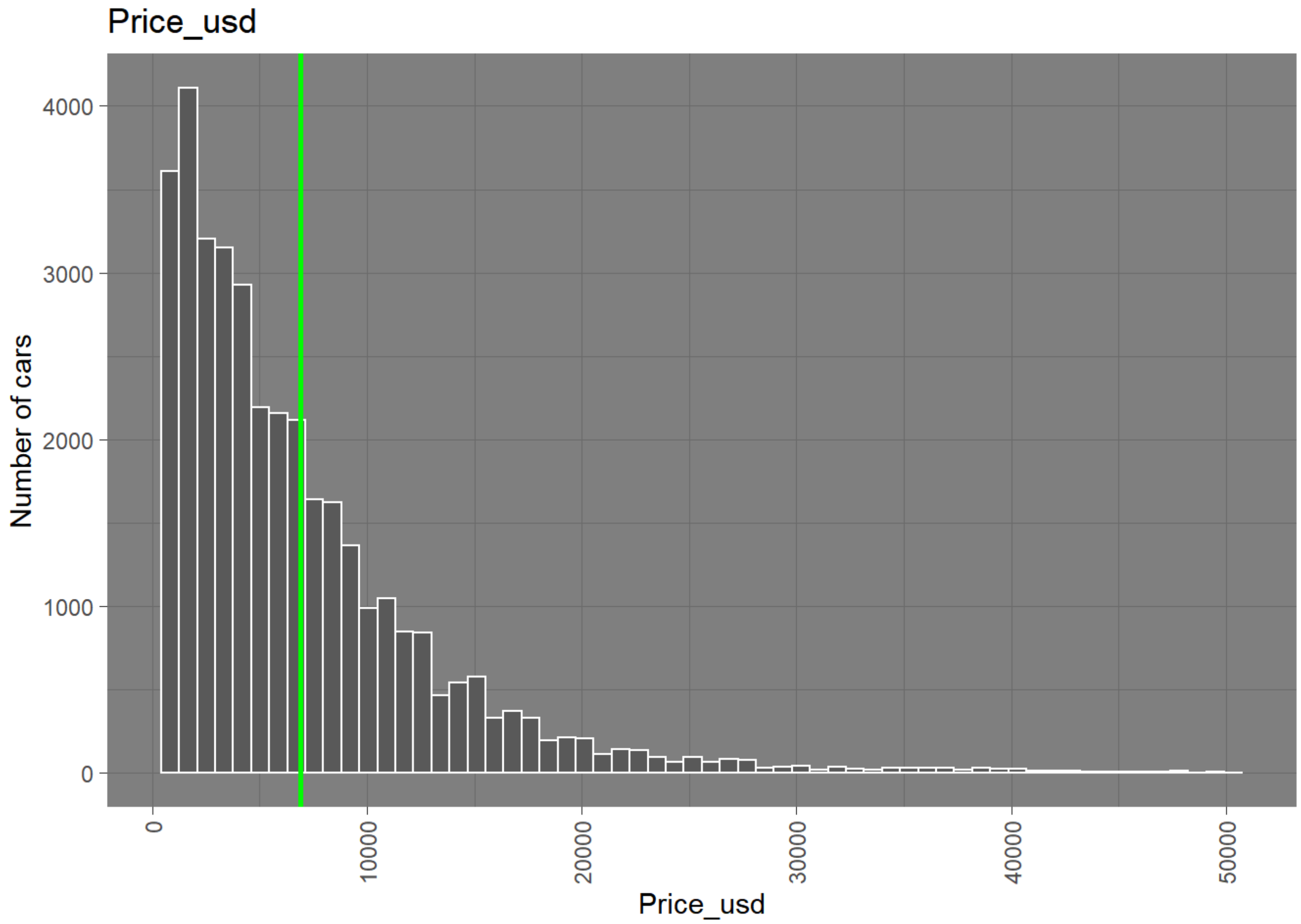
##

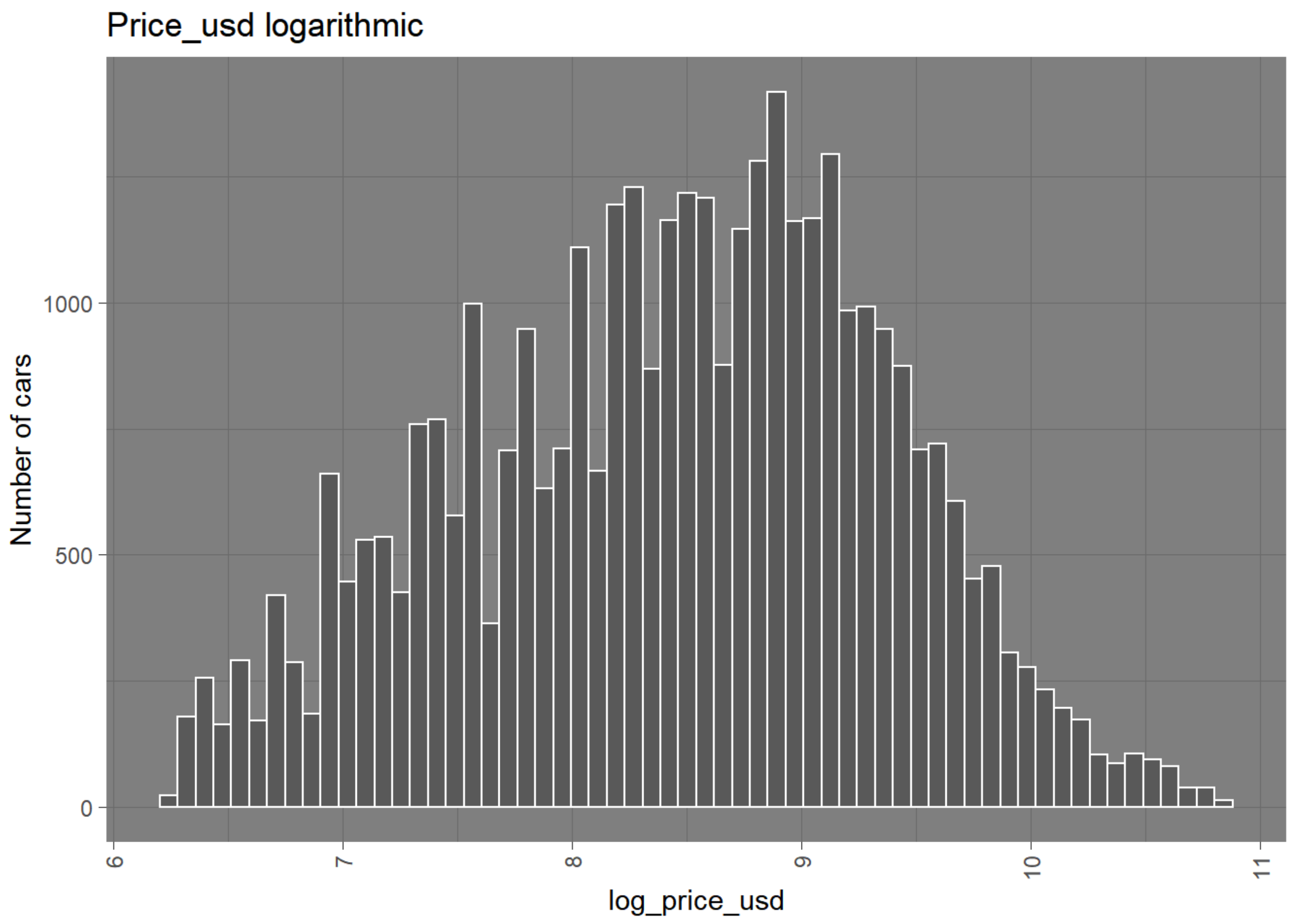
Exploring numerical variables

price_usd (usd) (continuous Numeric):Sale price of the car

- Most cars are listed below 10,000 usd.

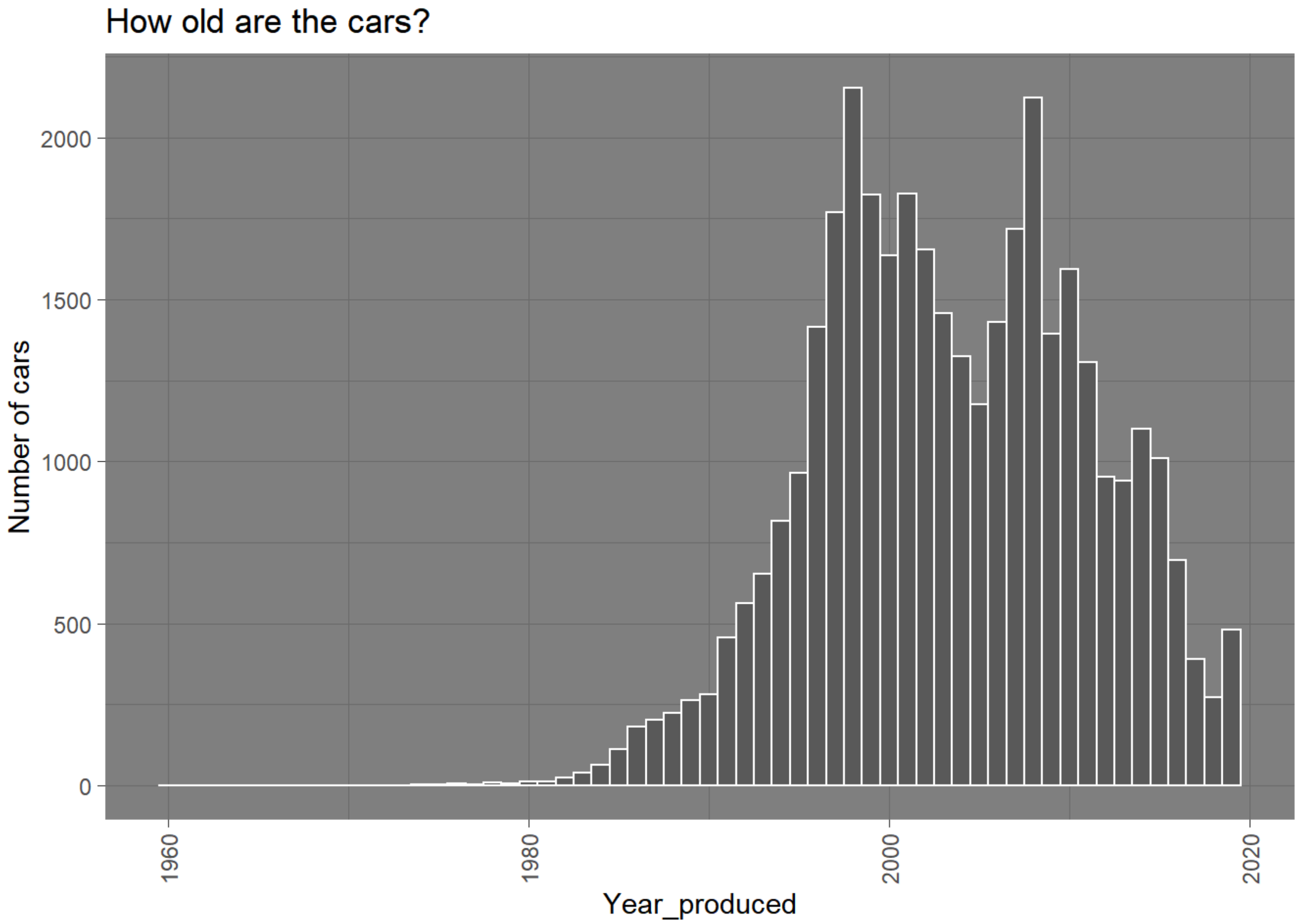
- Green line marks the mean, 6925.35 usd.



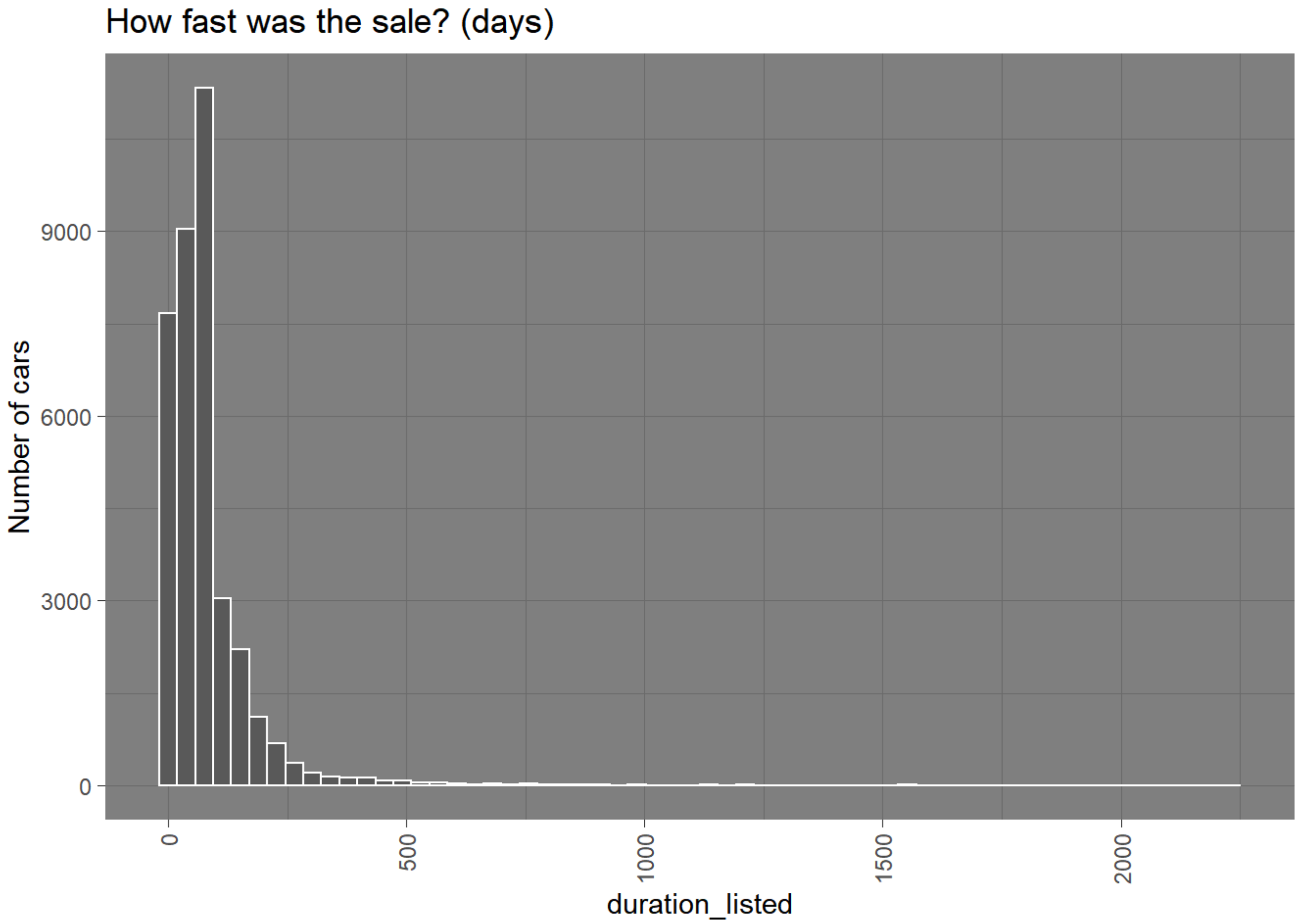


year_produced (Discrete Numeric):Year when the car got manufactured.

- The histogram follows two waves.

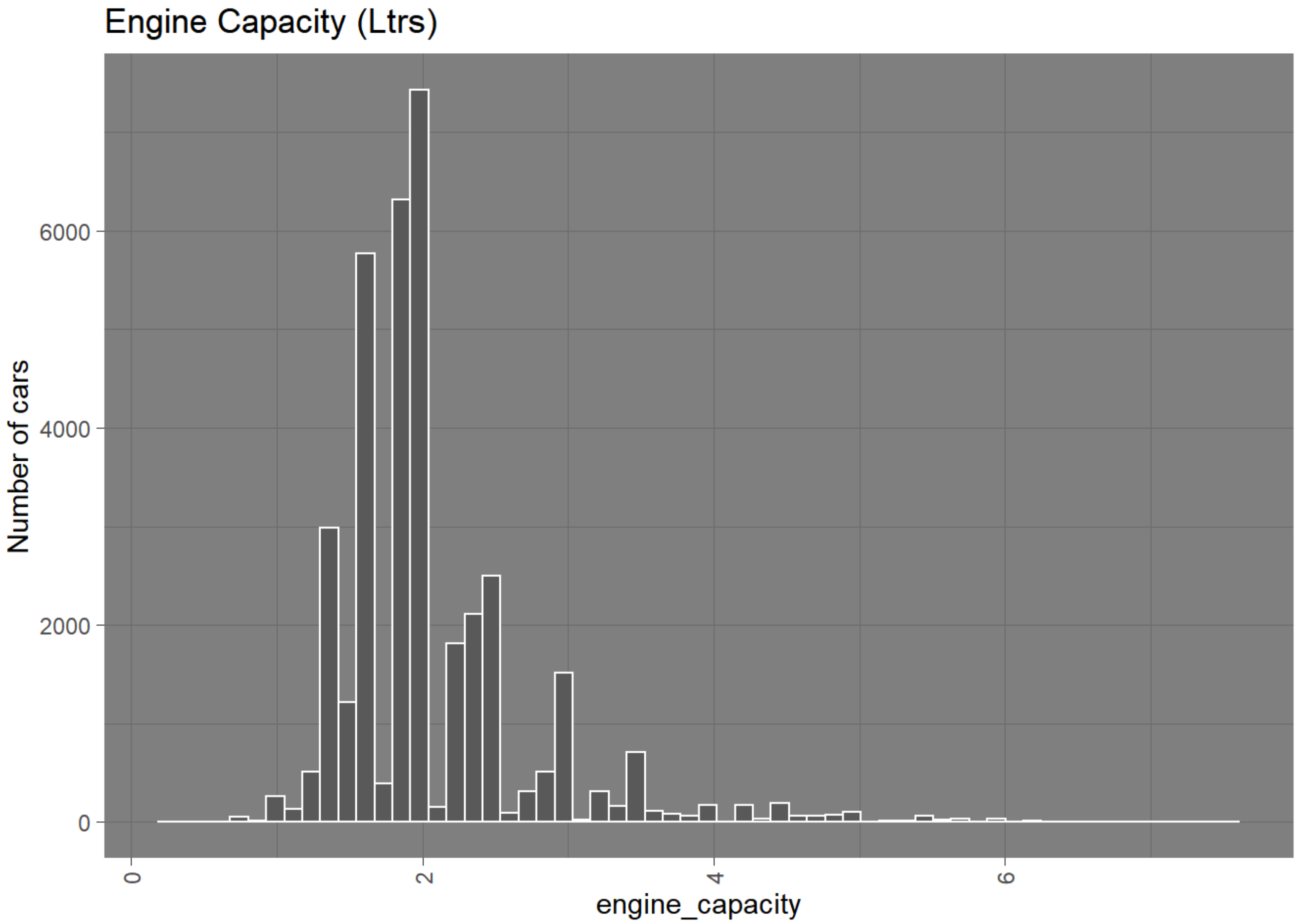


duration_listed (days) (Continuous Numeric):For how long was the deal on sale?

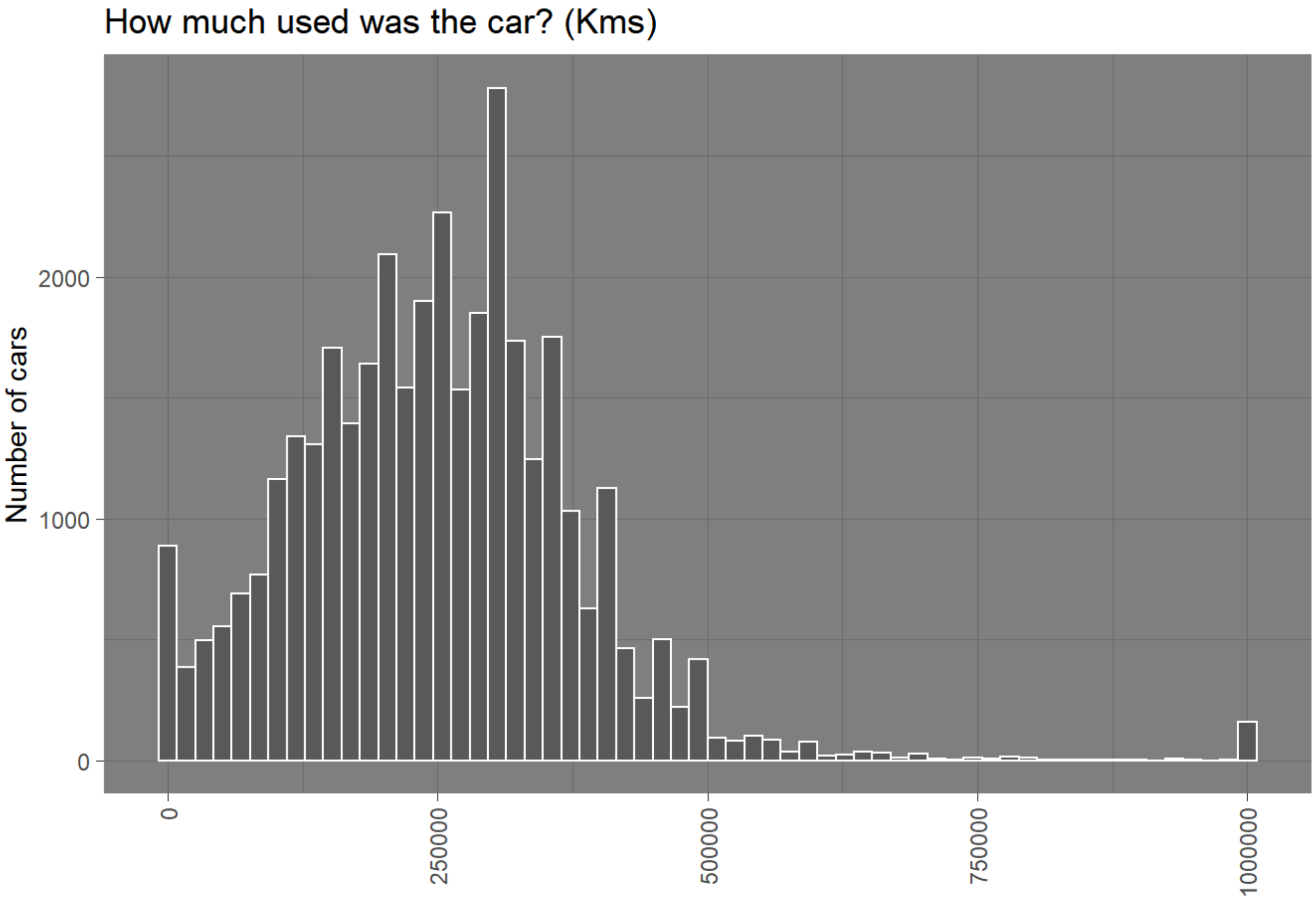


engine_capacity (Liters) (Continuous Numeric):Volume swept by all pistons in one engine

- GK: More the capacity lesser is the mileage for a particular engine.



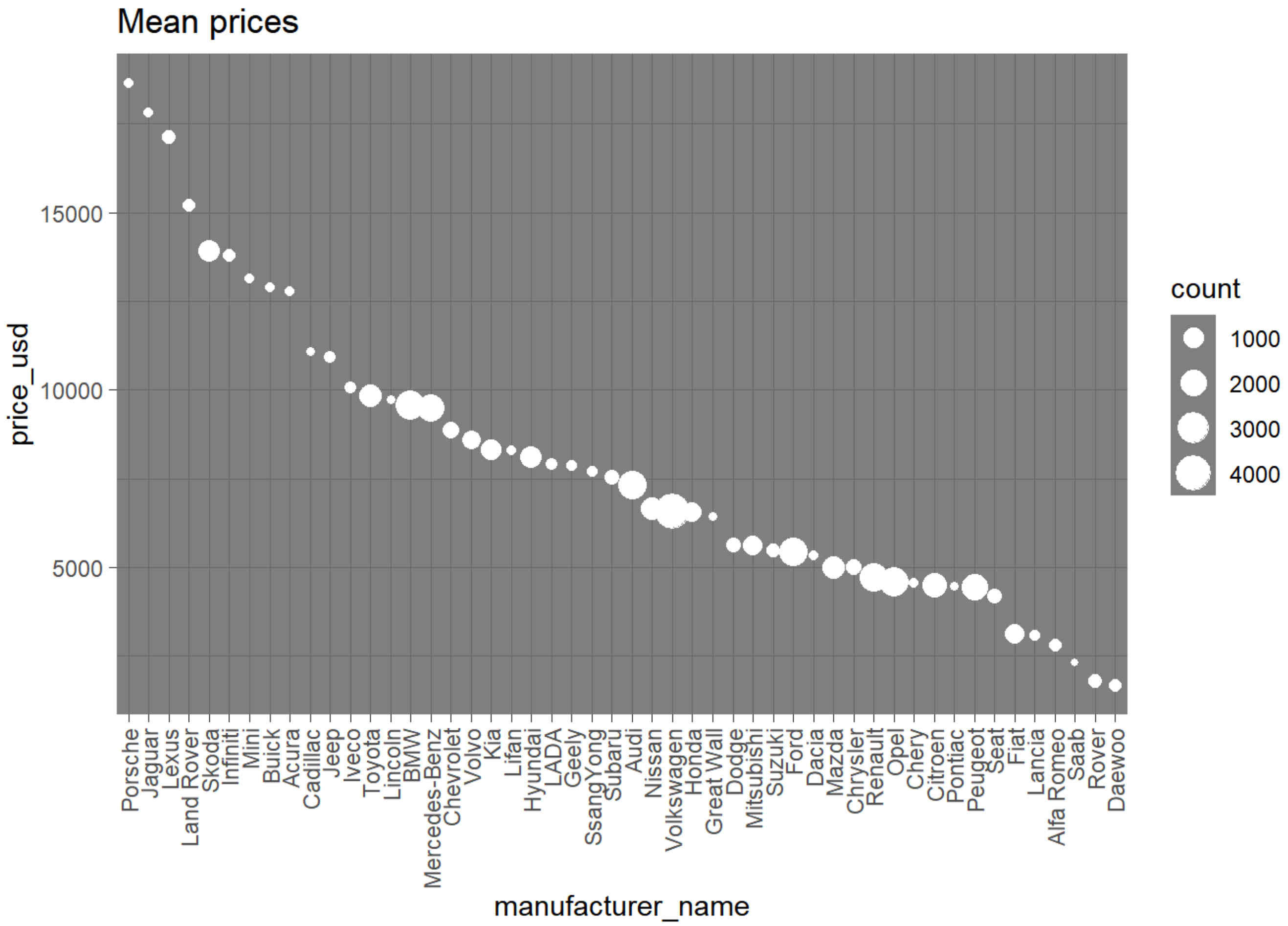
odometer_value (Kilometers) (Continuous Numeric):Total distance covered by the car.

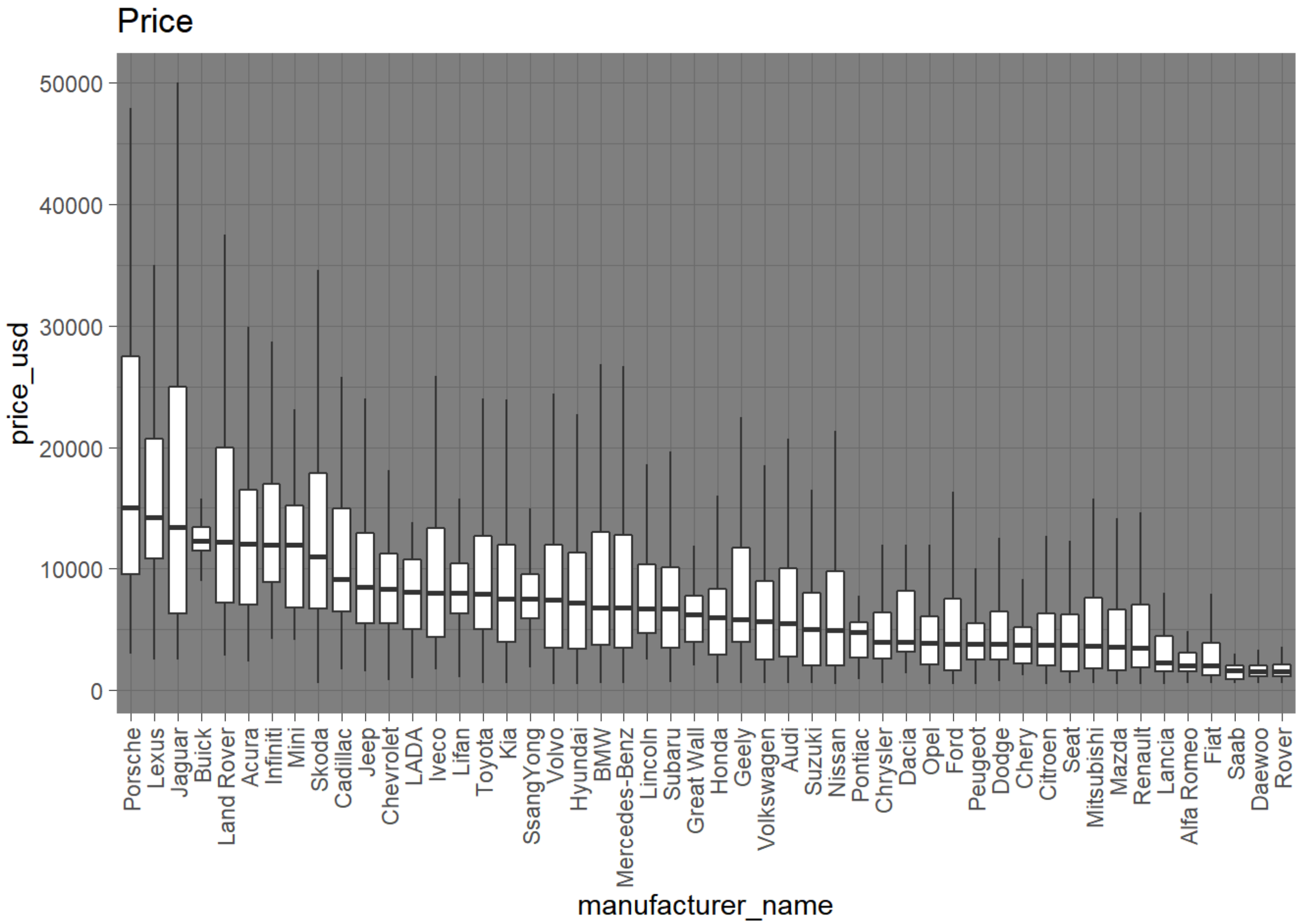


##Diving more into the

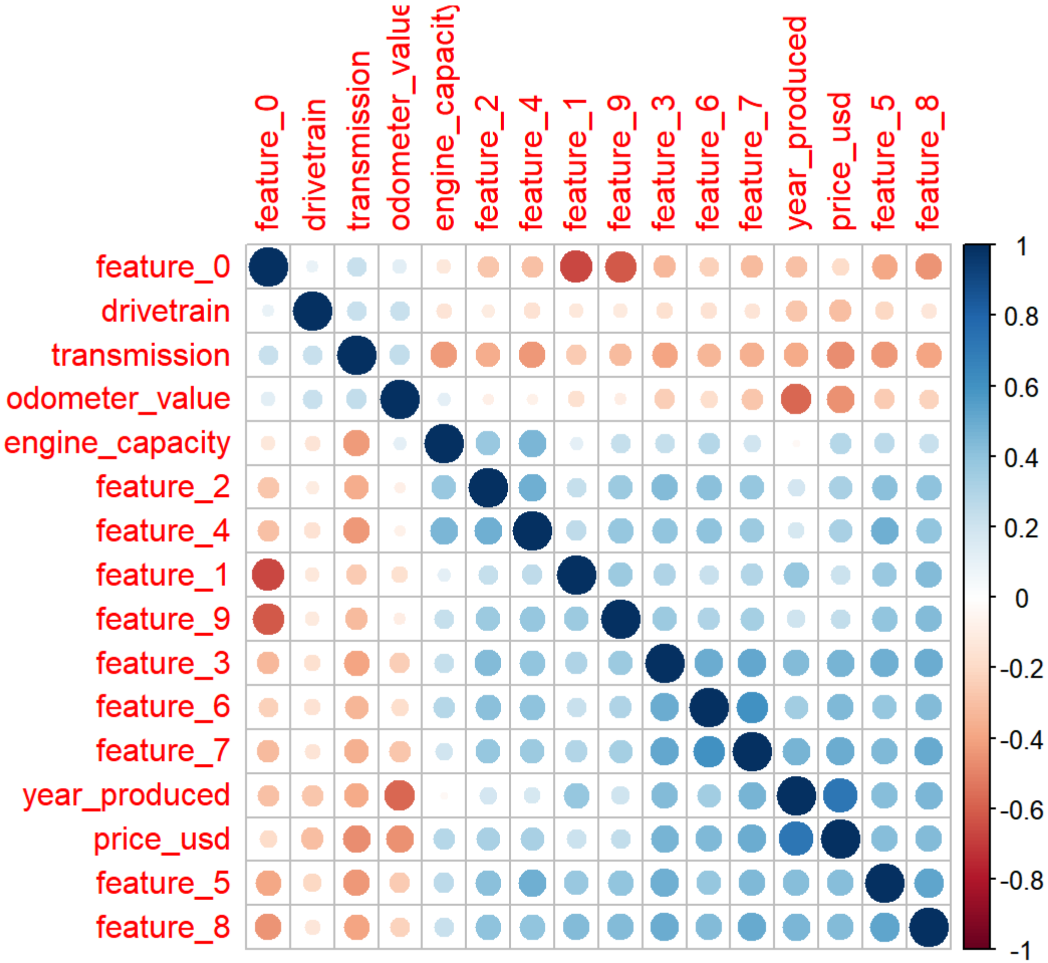
price_usd variable

Price distribution

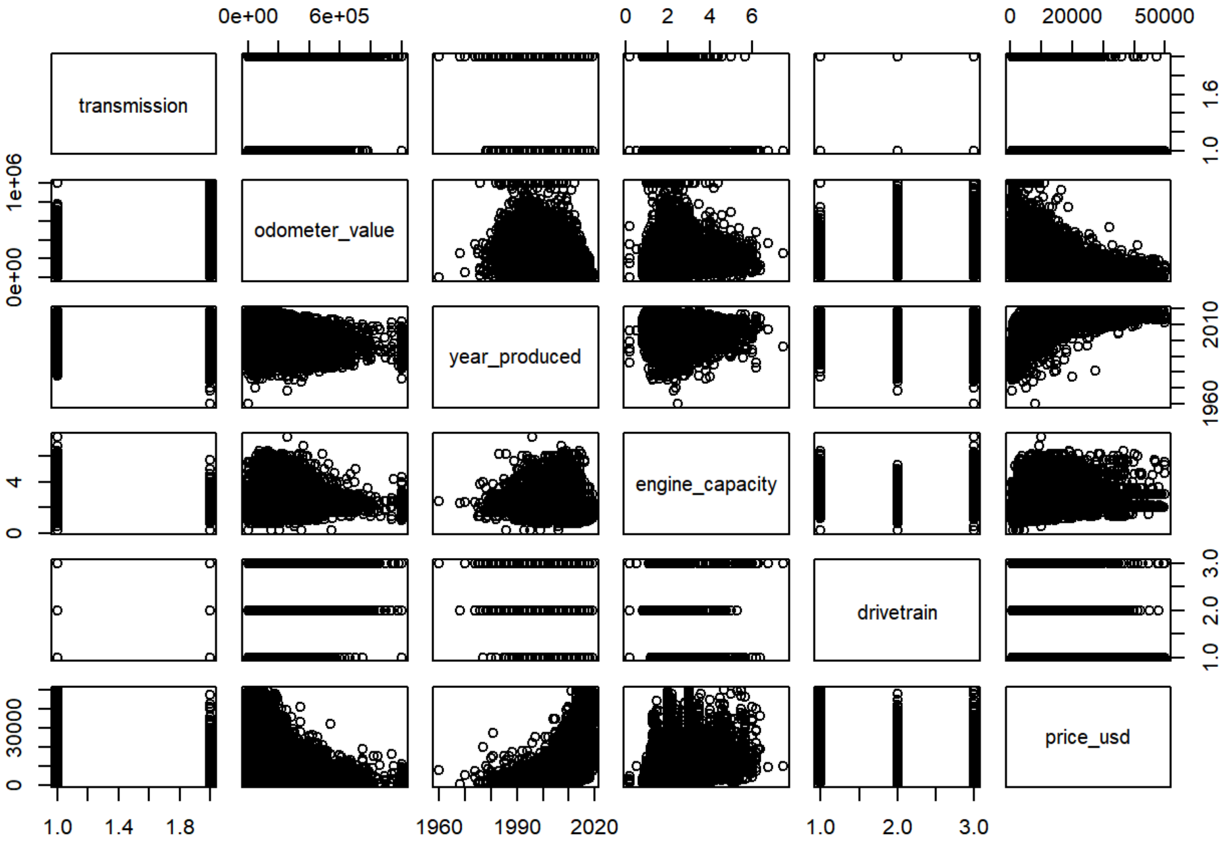



Price
price_usd
50000
40000
30000
20000
10000
0
Porsche
Lexus
Jaguar
Buick
Land Rover
Acura
Infiniti
Mini
Skoda
Cadillac
Jeep
Chevrolet
LADA
Iveco
Lifan
Toyota
Kia
SsangYong
Volvo
Hyundai
BMW
Mercedes-Benz
Lincoln
Subaru
Great Wall
Honda
Geely
Volkswagen
Audi
Suzuki
Nissan
Pontiac
Chrysler
Dacia
Opel
Ford
Peugeot
Dodge
Chery
Citroen
Seat
Mitsubishi
Mazda
Renault
Lancia
Alfa Romeo
Fiat
Saab
Daewoo
Rover
manufacturer_name

Plotting for correlation.

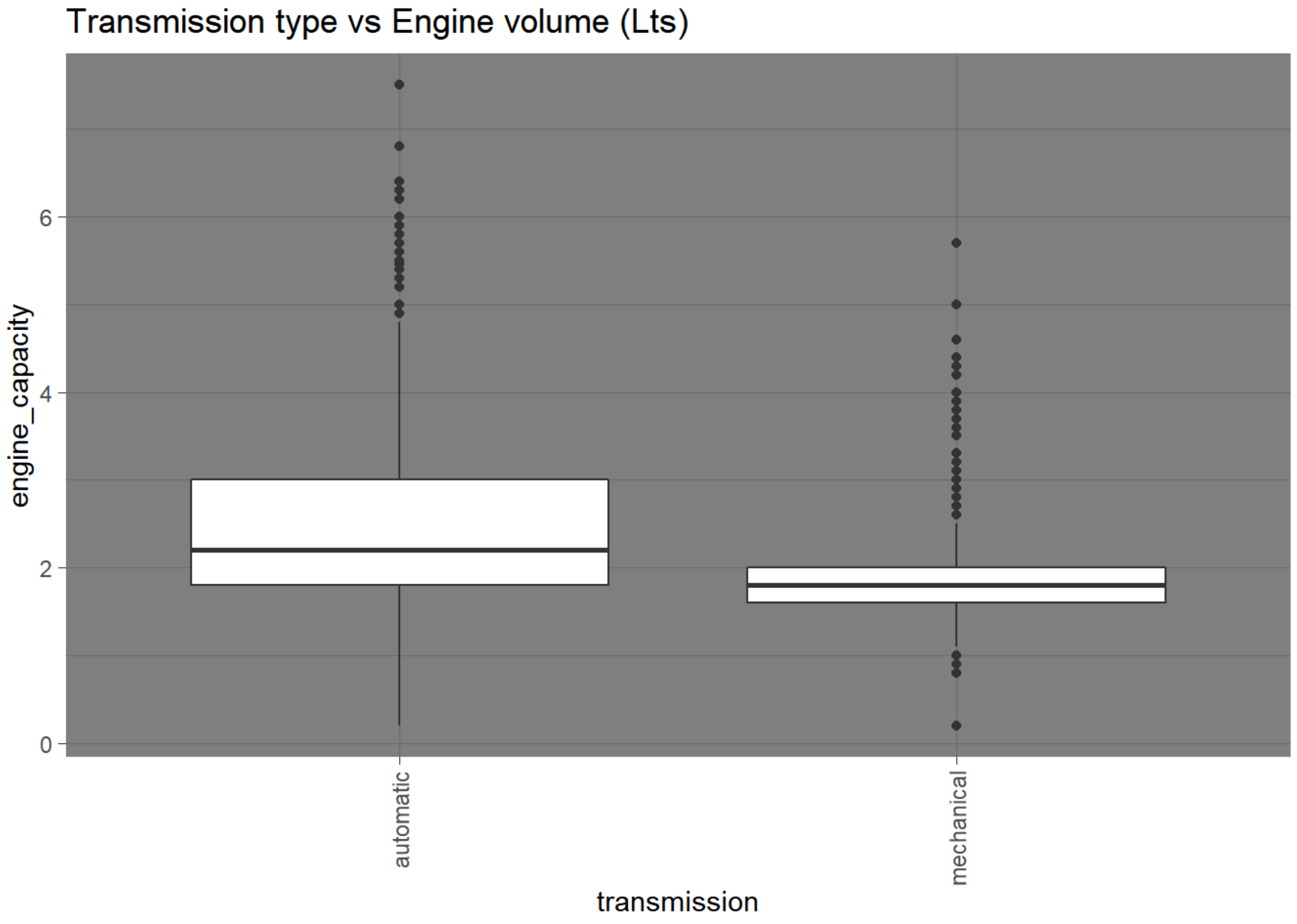


Pairs plot for potential predictors.



- Automatic transmission is more suitable for cars with big engines.
    - In general higher the engine capacity, higher the brake horse power and lower is the mileage

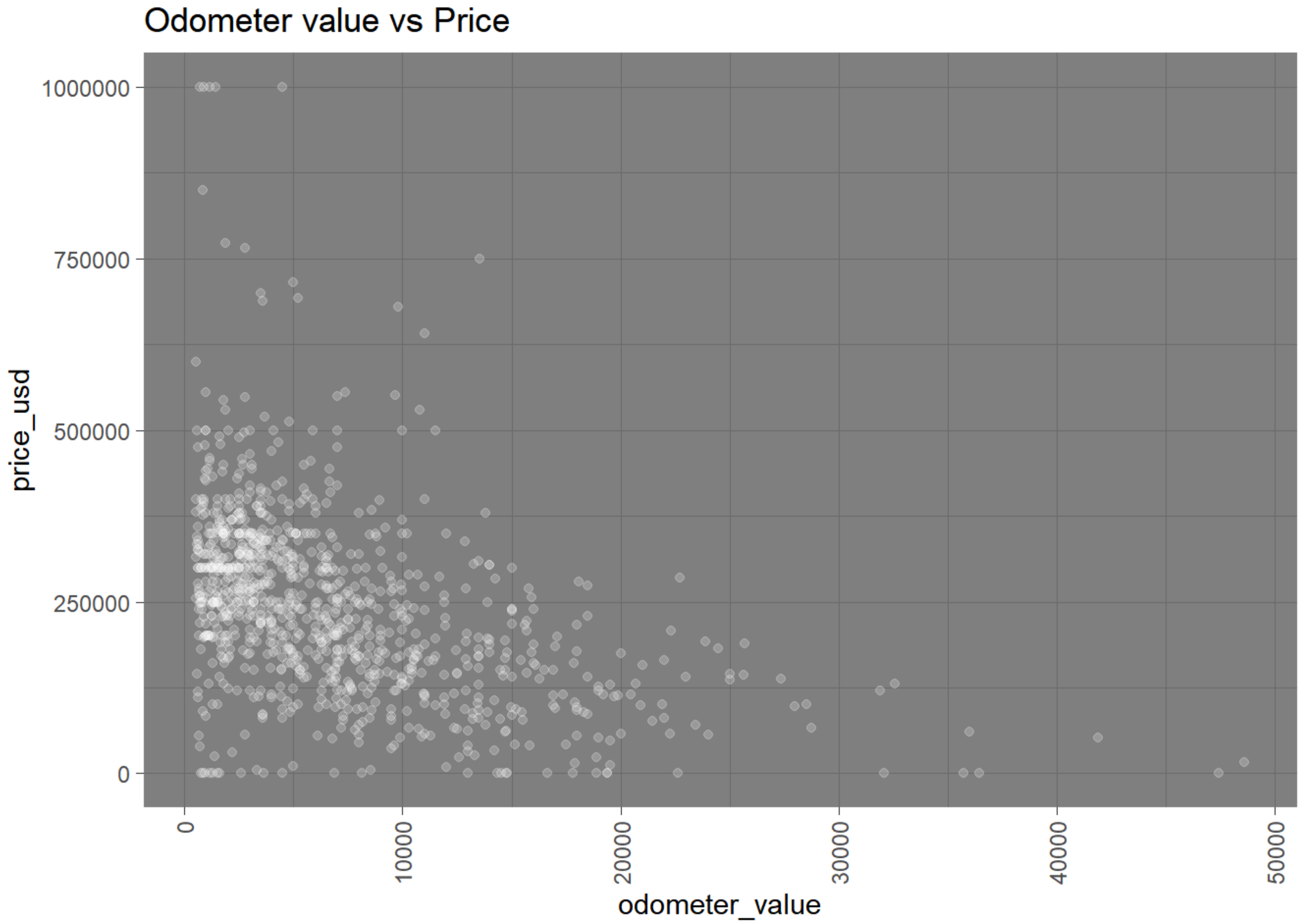
- Automatic transmission is more fuel efficient than manual transmission hence preferred with high capacity engines as to save some miles.



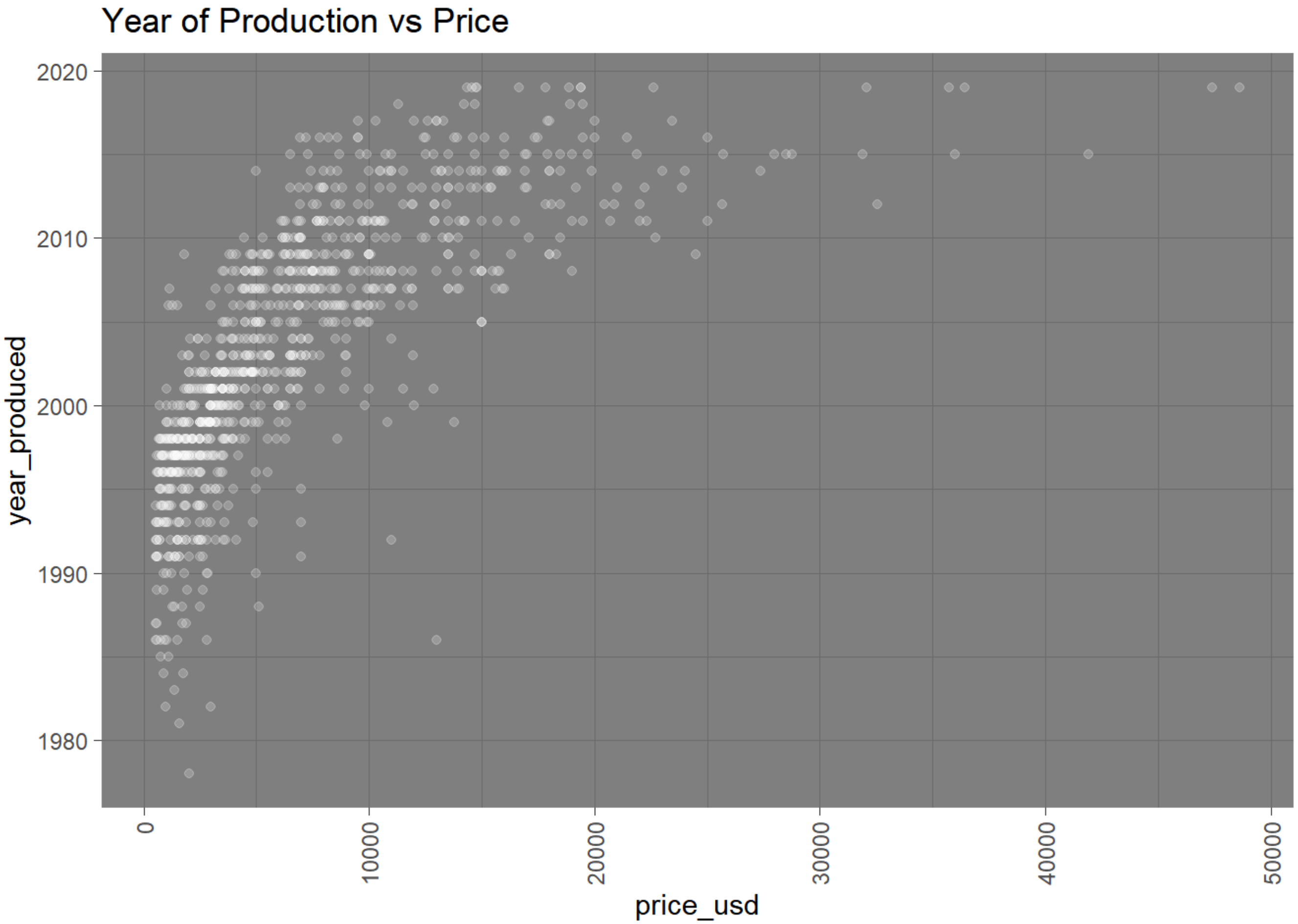
What makes a car expensive?

A quest for root causes and interrelations within the Potential predictors.

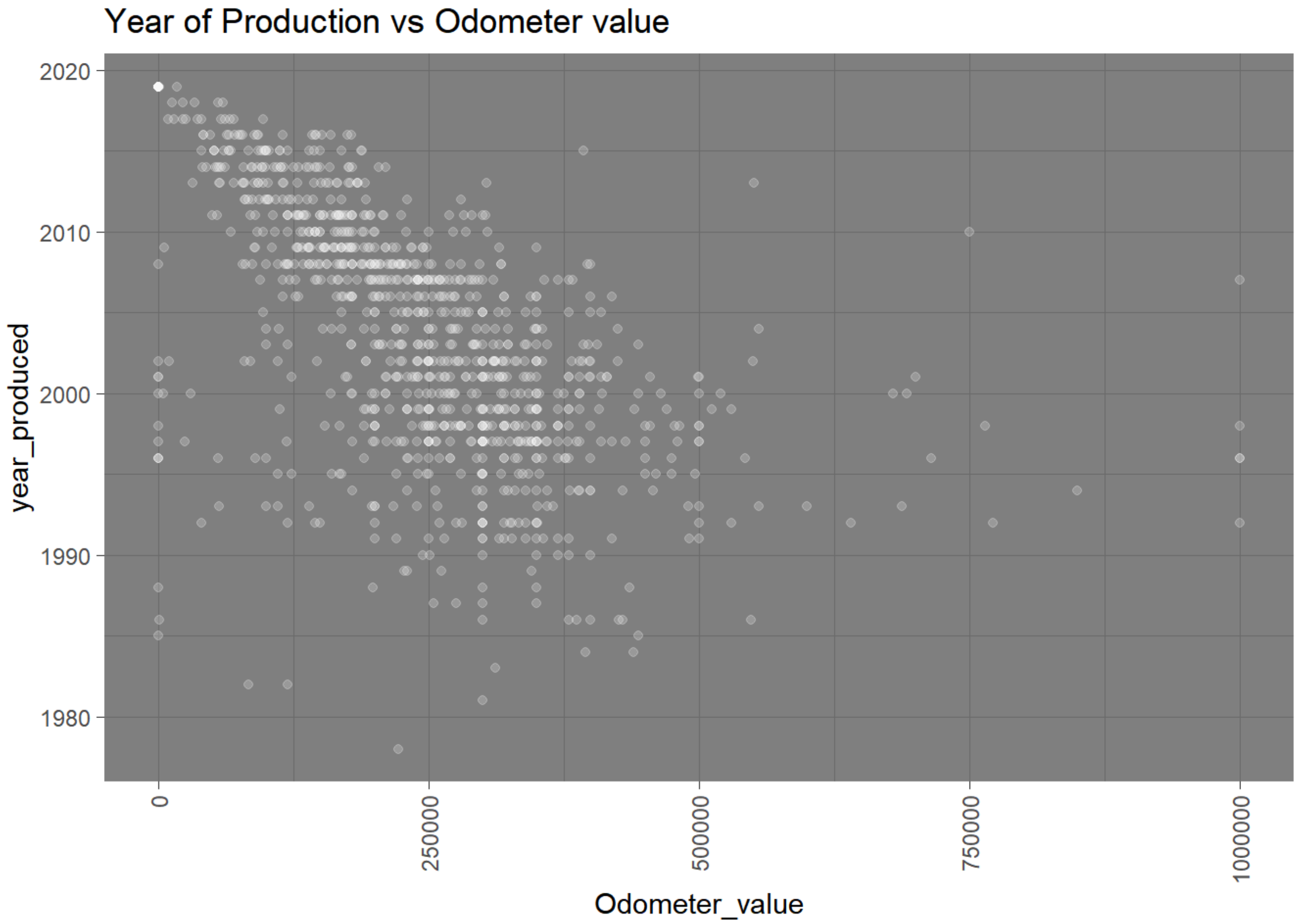
- Highly driven cars are more cheaper

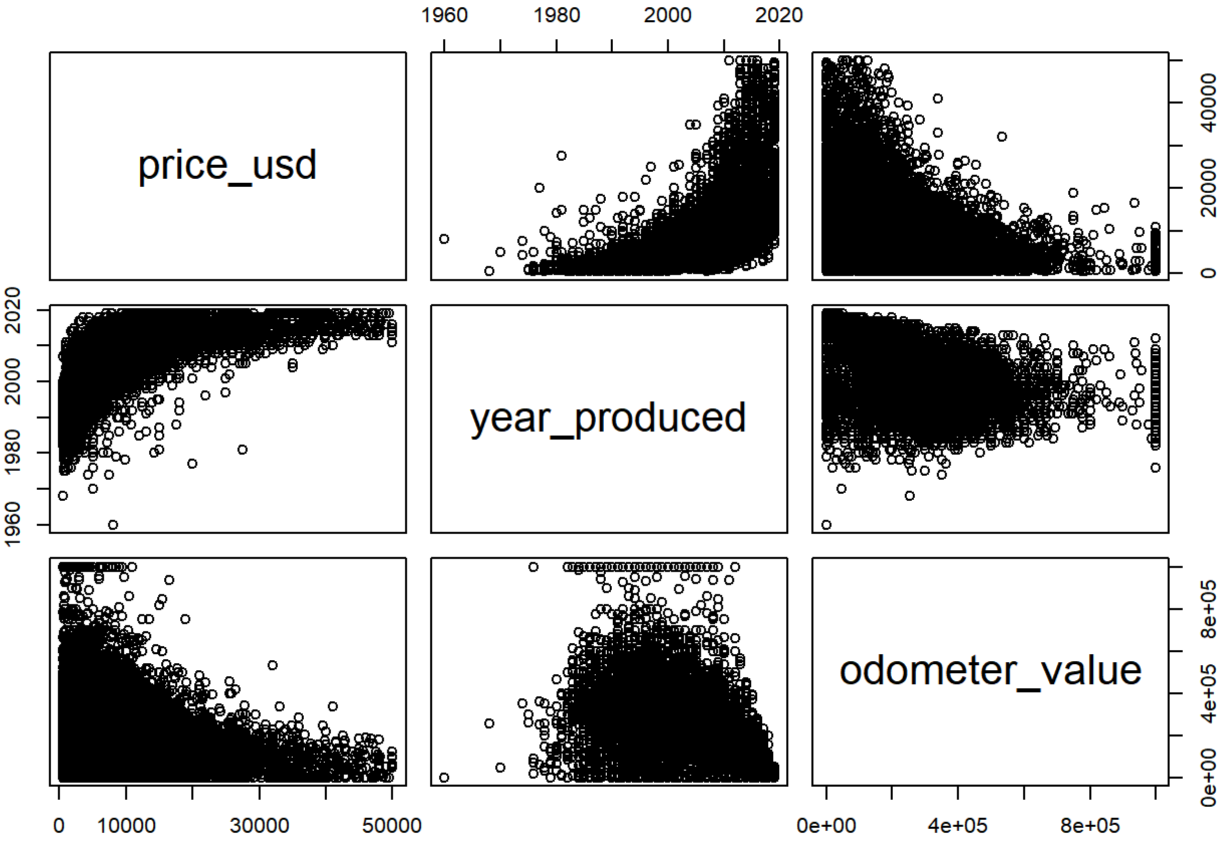


- Older cars are cheaper

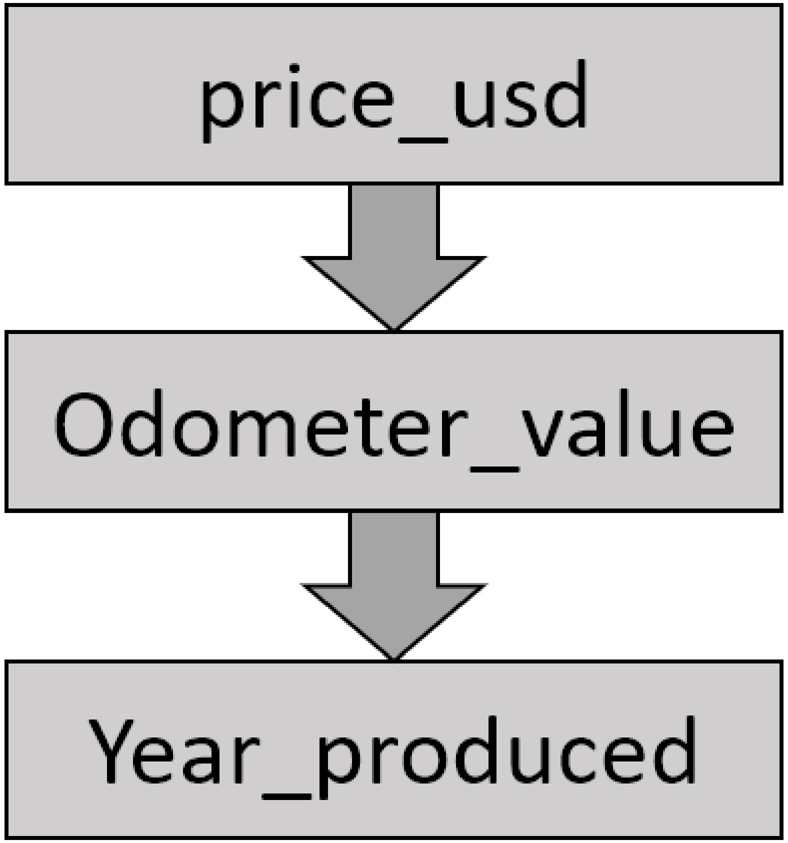


- Older cars are more driven



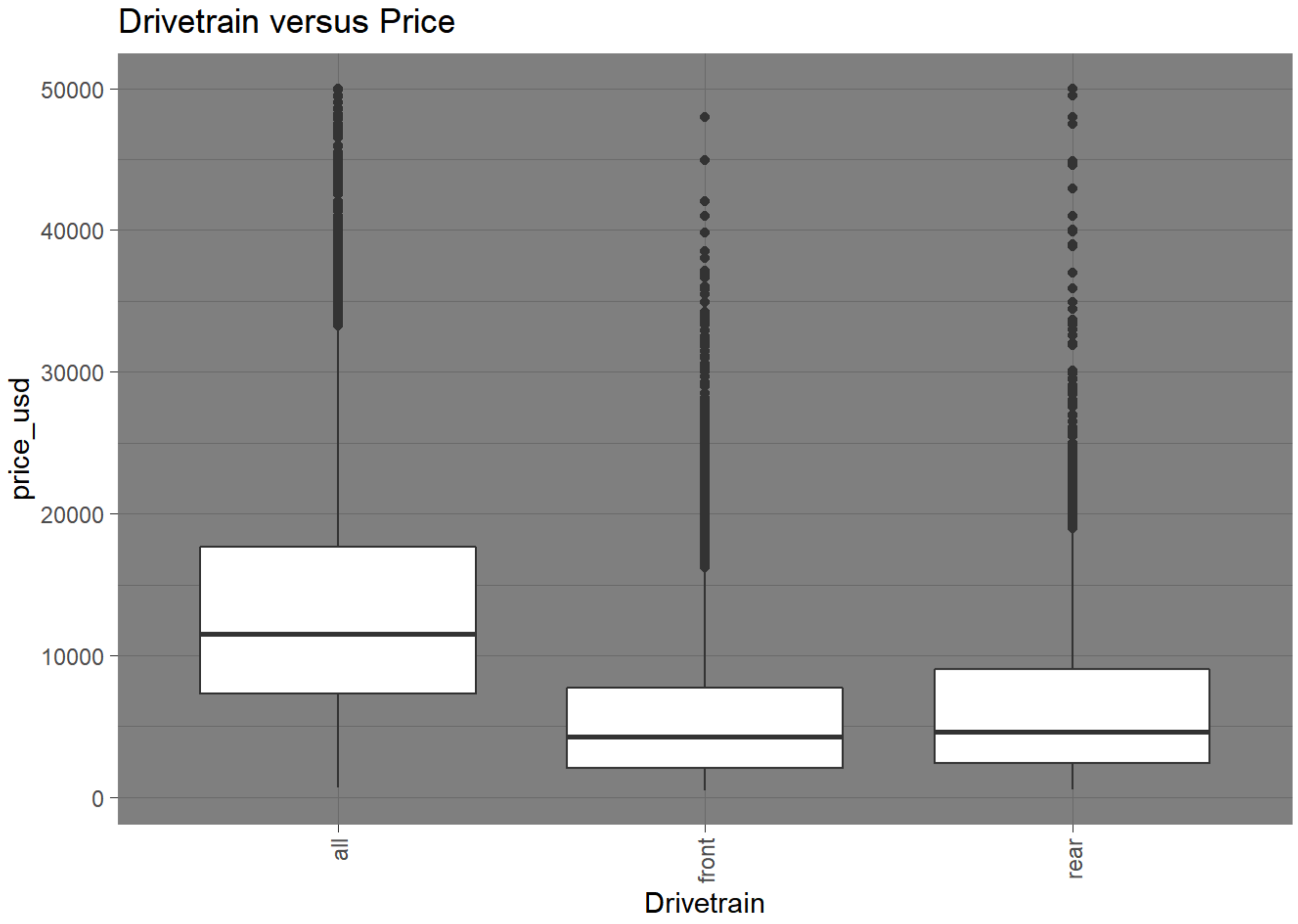

price_usd
year_produced
odometer_value
1960
1980
2000
2020
0
20000
40000
1960
1980
2000
2020
0e+00
4e+05
8e+05
0
10000
30000
50000
0e+00
4e+05
8e+05

- Early produced cars are highly driven showing high odometer readings which makes the car cheaper.

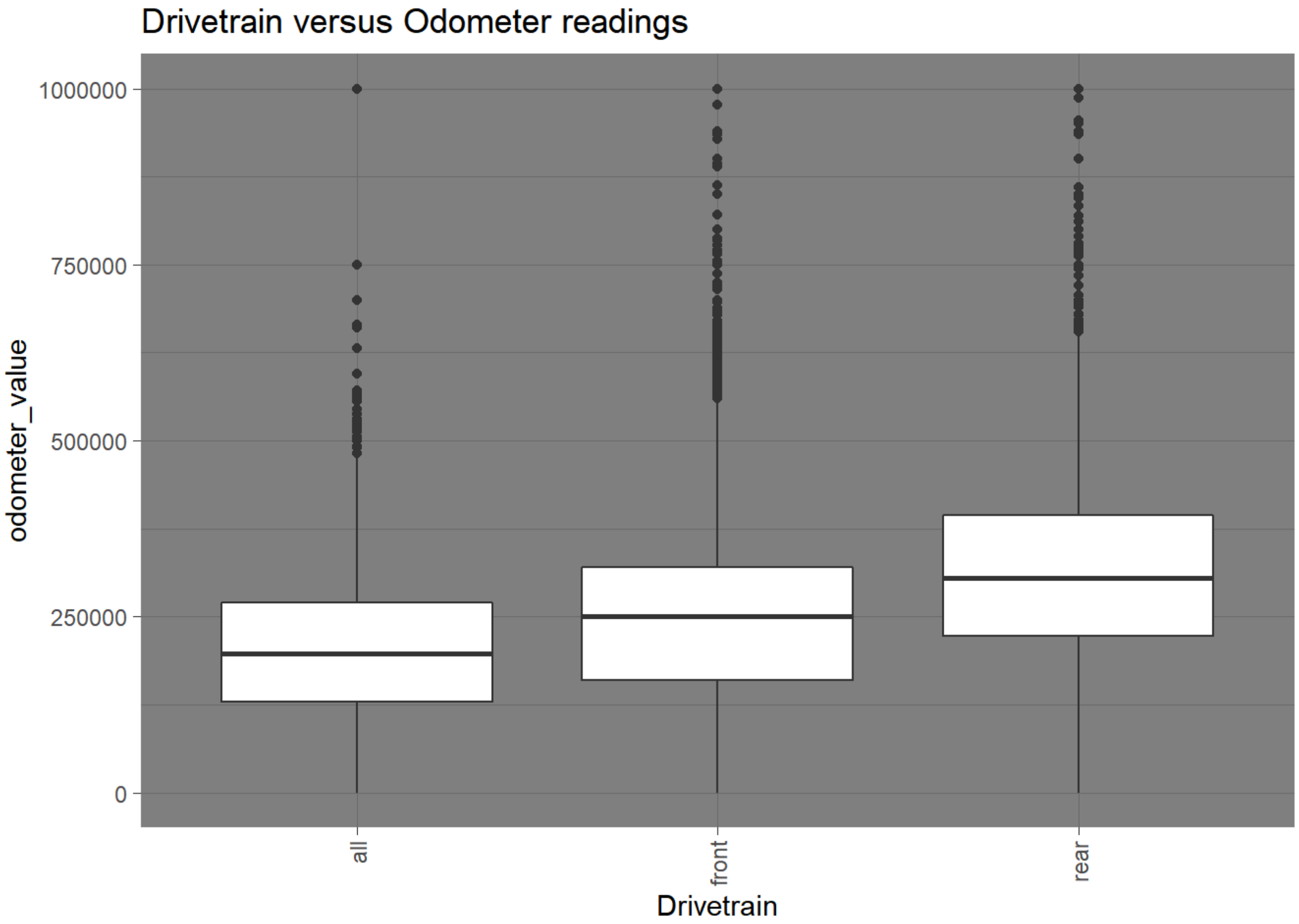


Relationship of drivetrain with price, year of production and odometer reading

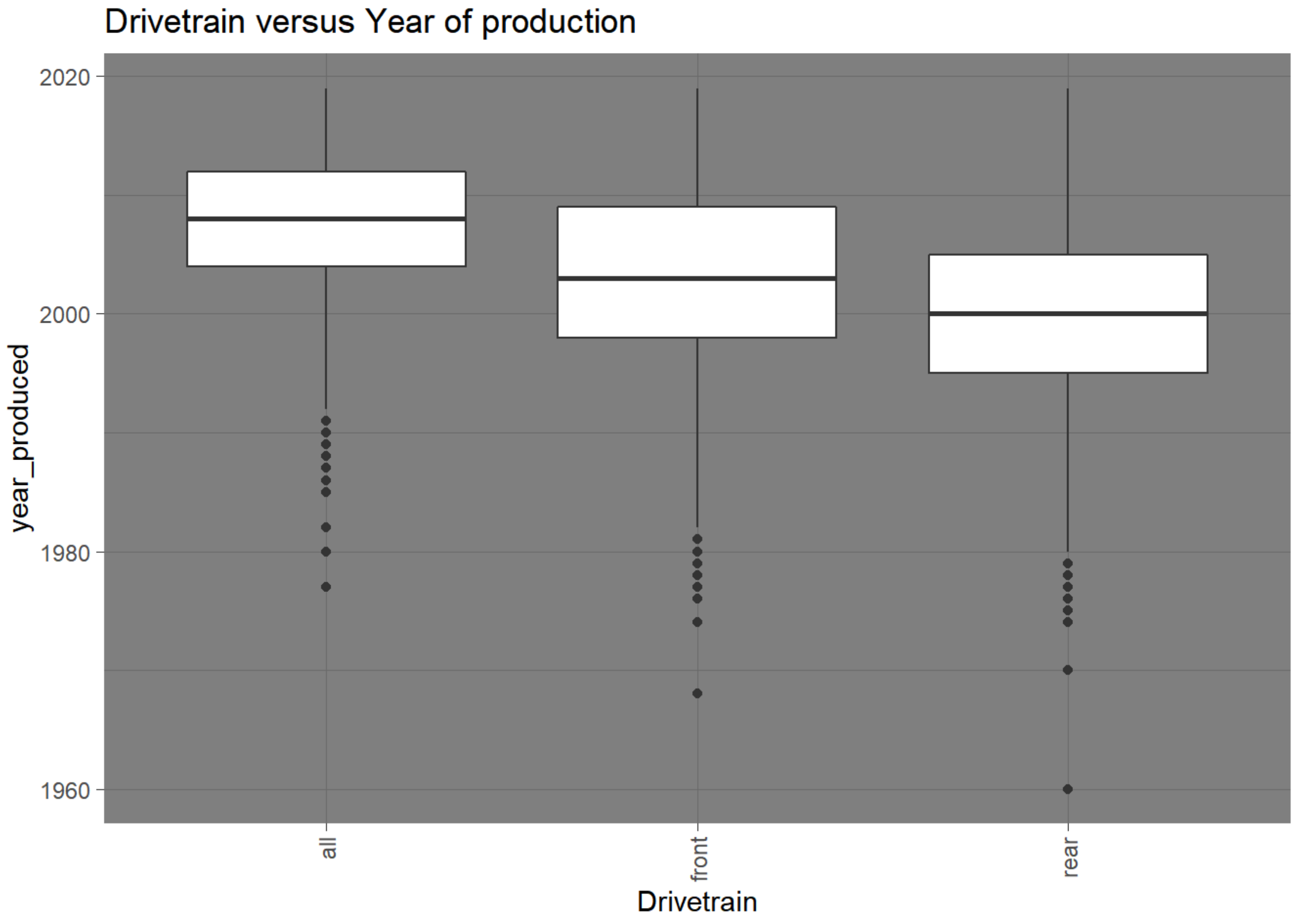
- All wheel drive cars are more expensive.



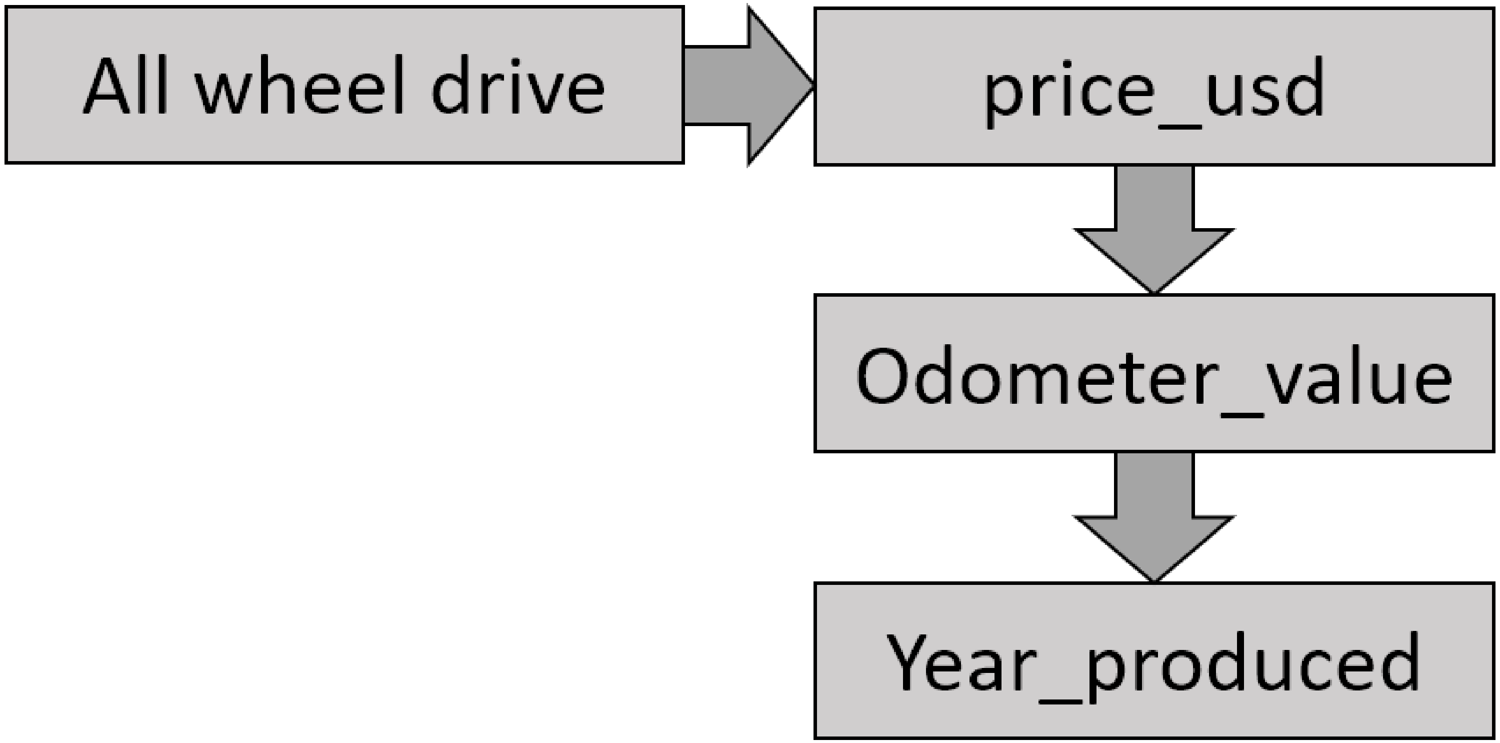
- All wheel drive cars are less driven.



- All wheel drive is the latest

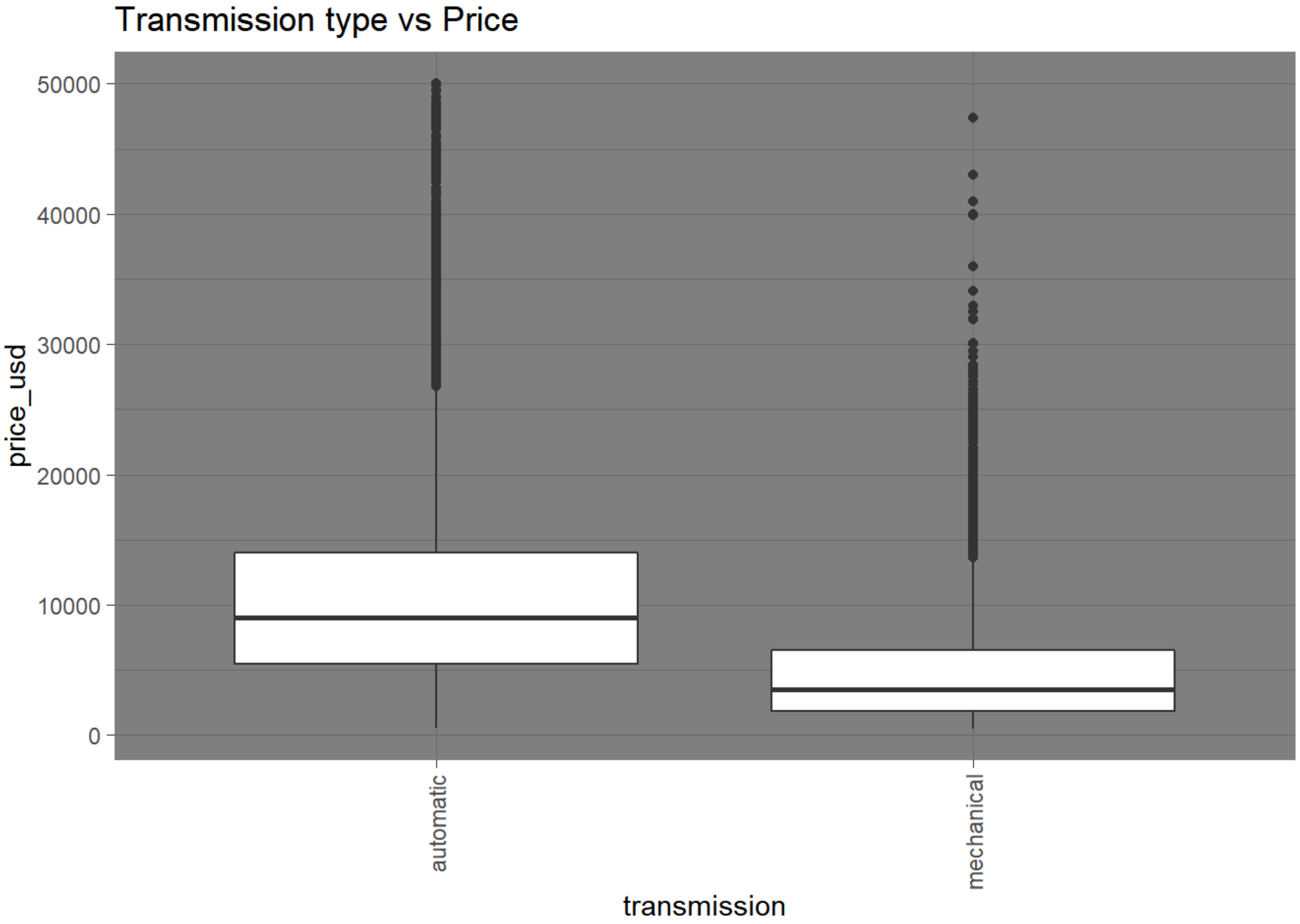


- All wheel drive cars were produced lately, having lower odometer value which makes the car more expensive.

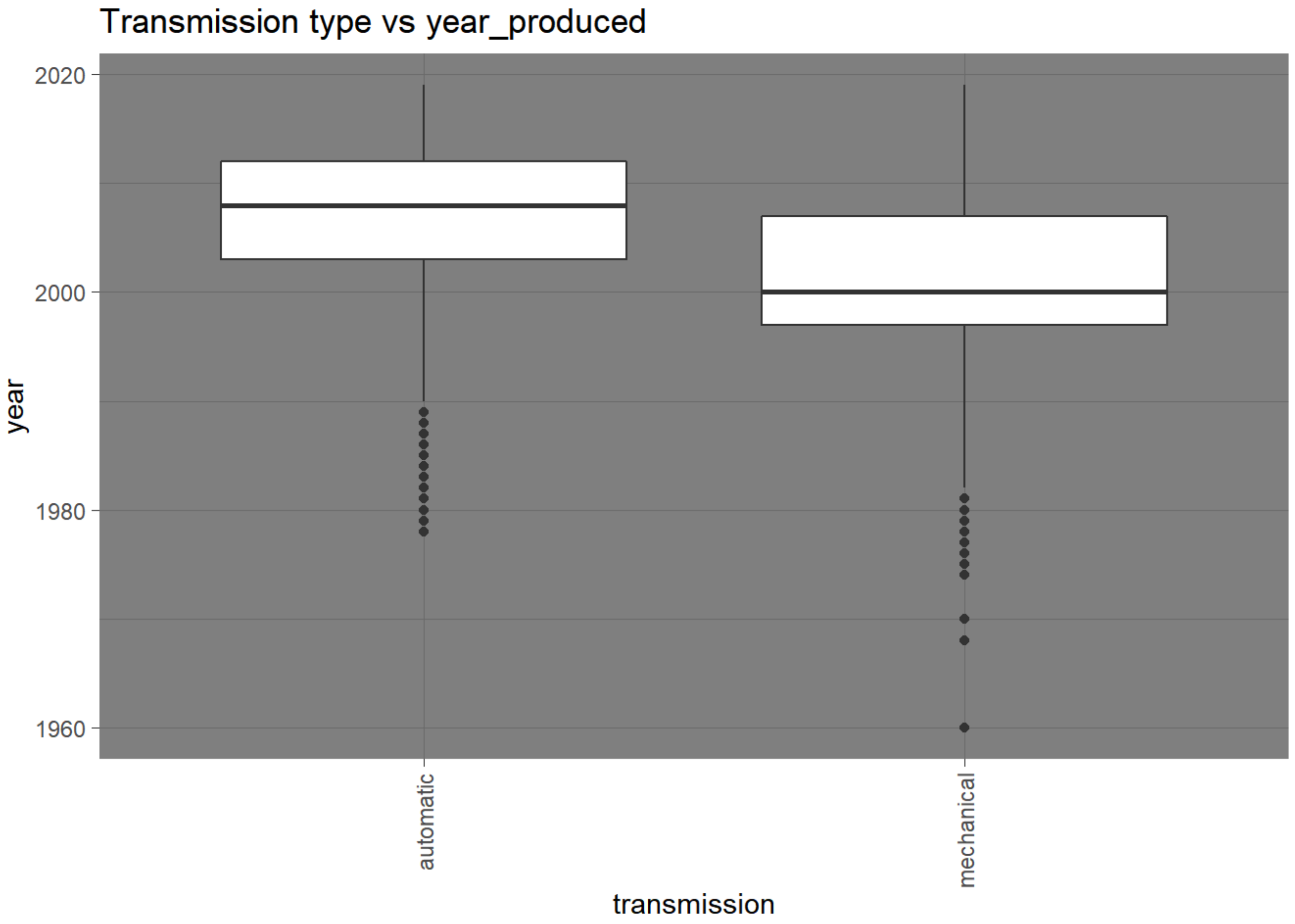


Relationship of transmission with price, year of production and odometer reading.

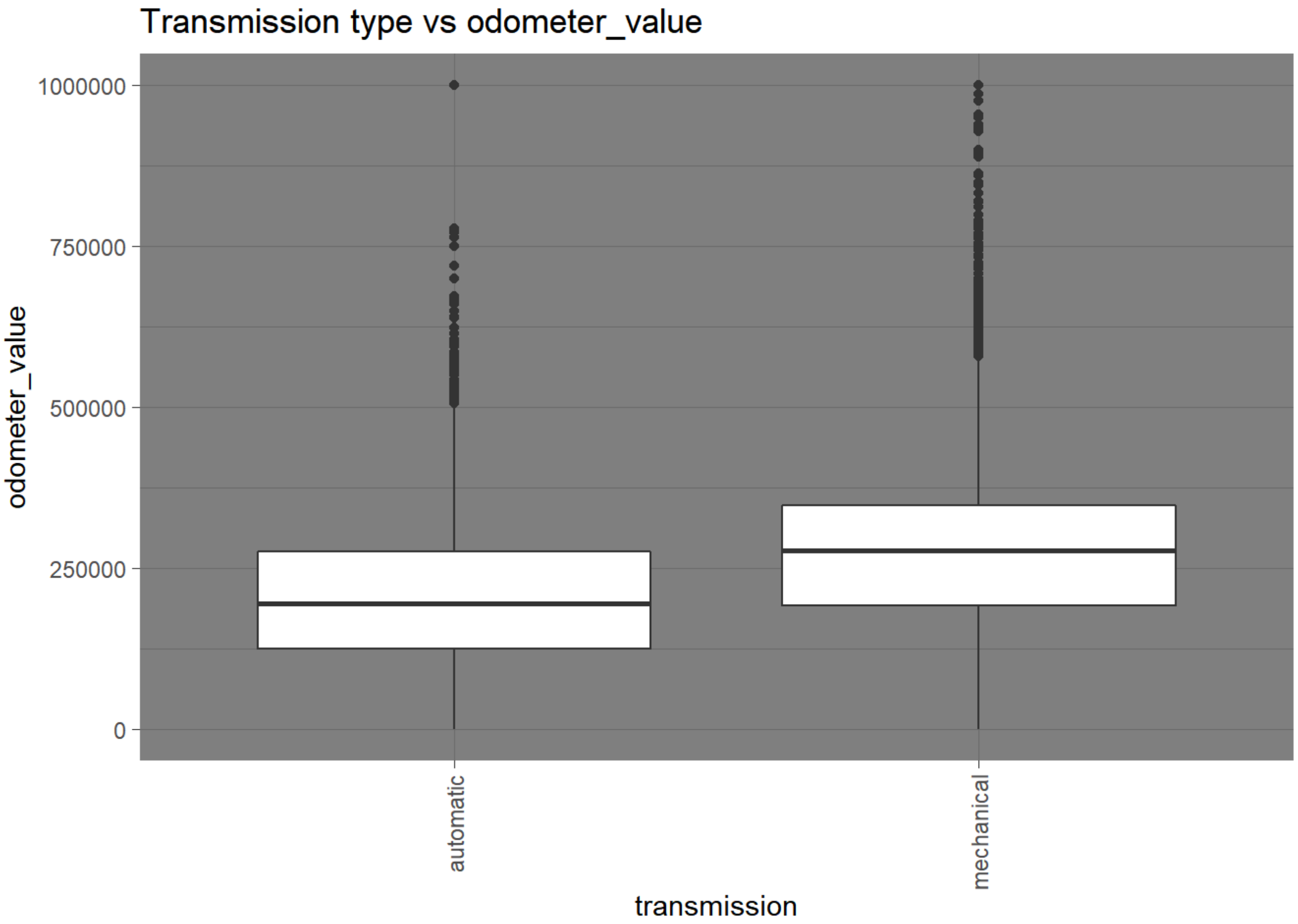
- The automatic transmission is an assistance feature and these cars are usually expensive than manual transmission cars.
    - Older cars do not come with automatic transmission.



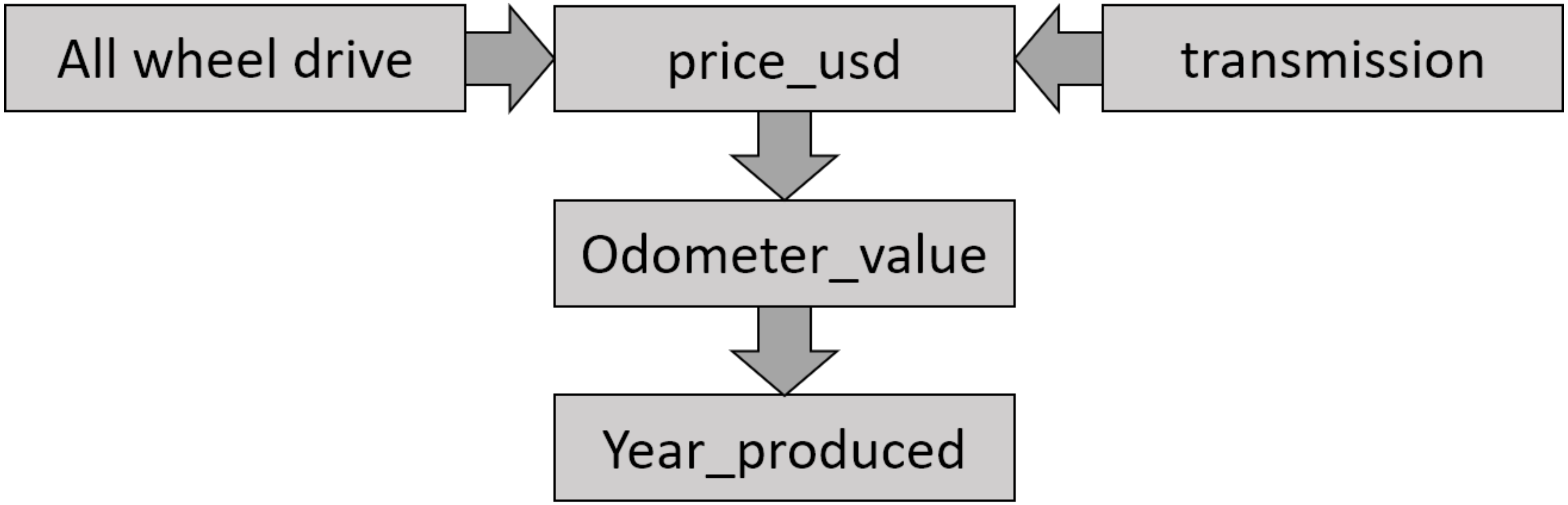
- Unlike manual transmission, automatic cars are not produced earlier. They are the latest.



- Automatic cars are also less driven



- Automatic transmission cars were produced lately, having lower odometer value which makes the car more expensive.



Summary of Numerical Variables

```
## # A tibble: 2 x 4
##    mean `max(price_usd)` `min(price_usd)` `range(price_usd)`
##   <dbl>            <dbl>            <dbl>              <dbl>
## 1 6925.            50000             502.               502.
## 2 6925.            50000             502.             50000
```

```
## # A tibble: 2 x 4
##      mean `max(odometer_value)` `min(odometer_value)` `range(odometer_value)`
##     <dbl>                 <dbl>                 <dbl>                   <dbl>
## 1 249135.               1000000                     0                       0
## 2 249135.               1000000                     0                 1000000
```

```
## # A tibble: 1 x 2
##   `max(year_produced)` `min(year_produced)`
##                  <dbl>                <dbl>
## 1                 2019                 1960
```

```
## # A tibble: 2 x 3
##   `max(engine_capacity)` `min(engine_capacity)` `range(engine_capacity)`
##                    <dbl>                  <dbl>                    <dbl>
## 1                    7.5                    0.2                      0.2
## 2                    7.5                    0.2                      7.5
```

Difficulties while preparing the data for analysis

- 36533 rows of data made the scatter plots overwhelming, slicing of the dataset has been done and made sure that the trends in subset data plots are no different to that of the parent data set .
- Most of the variables are categorical which were then converted into numeric format for plotting several plots.
- Many temporary data sets were created using group feature to help make plots easier.

After performing EDA, I believe no other datasets are required at this point but opinion may change in the future.