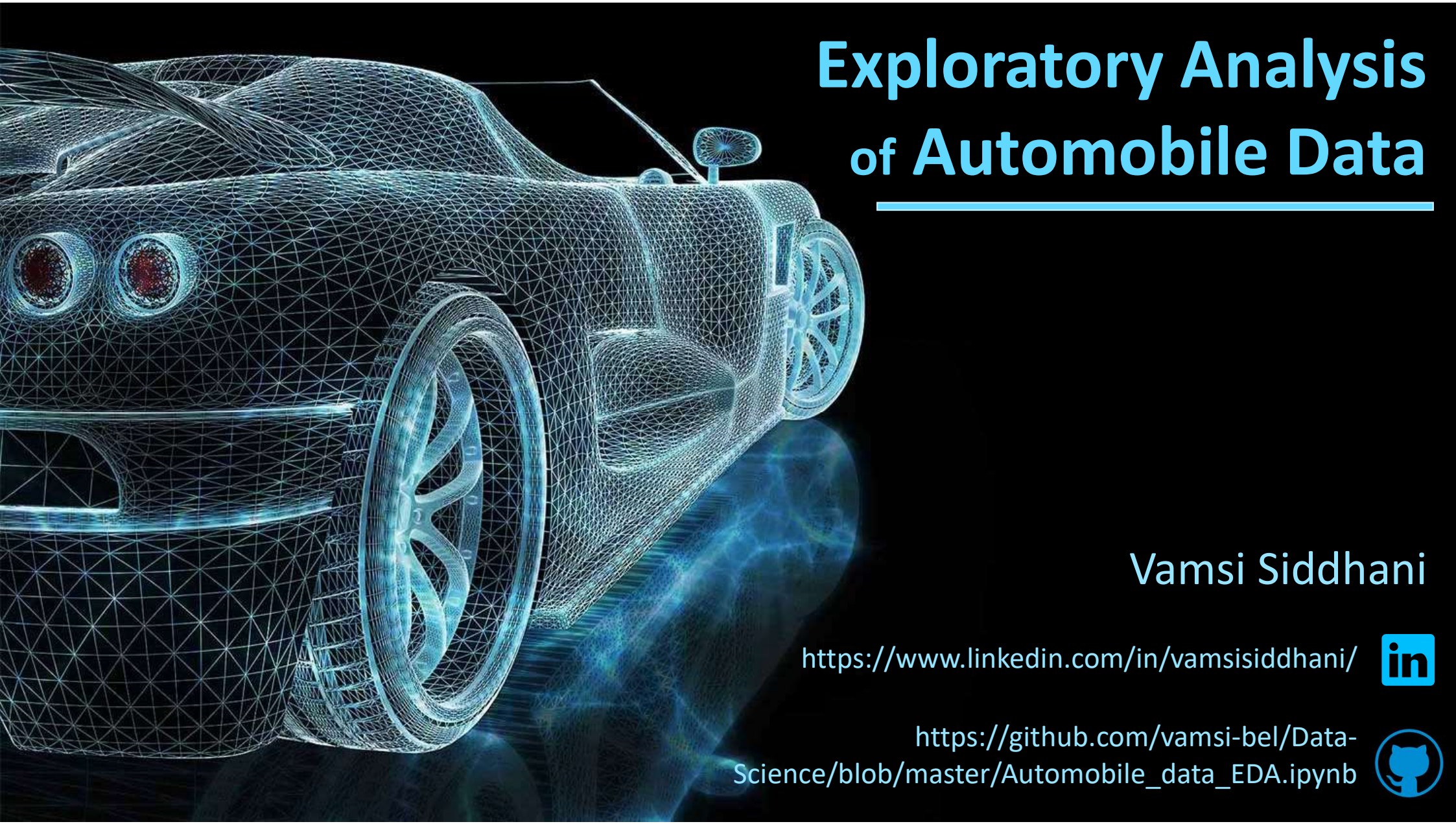


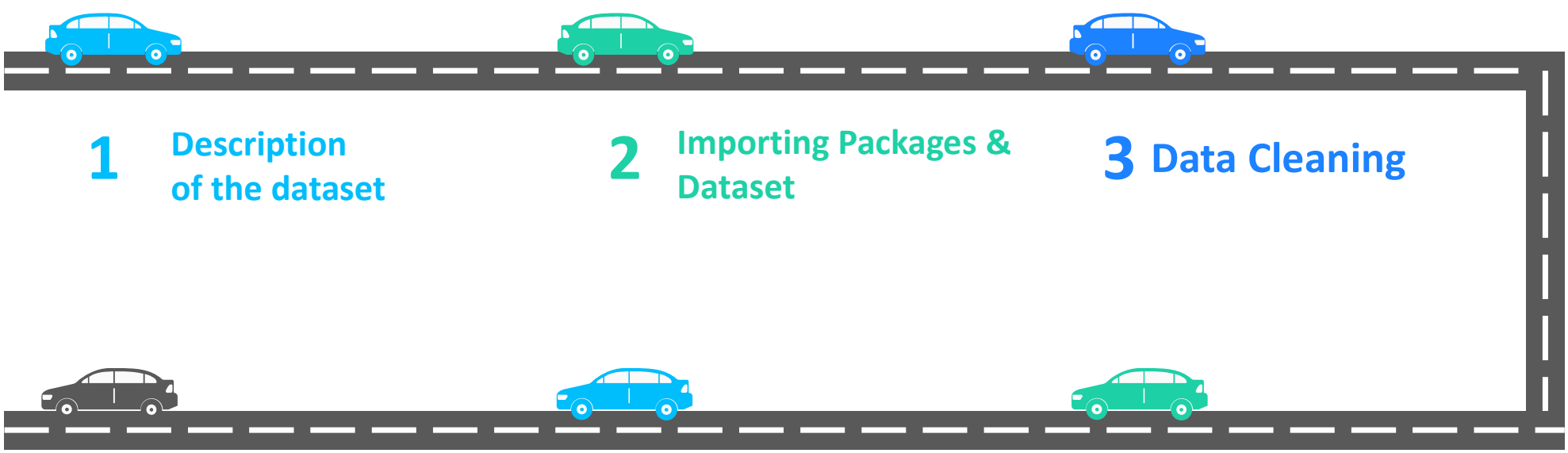
# Exploratory Analysis of Automobile Data

Vamsi Siddhani

https://www.linkedin.com/in/vamsisiddhani/

https://github.com/vamsi-bel/Data-Science/blob/master/Automobile_data_EDA.ipynb

# Steps Involved



1 Description of the dataset

2 Importing Packages & Dataset

3 Data Cleaning

4 Data Analysis

5 Correlation Analysis

6 Conclusion

# Tools & Technologies Used

***Programming language***


***Packages***


***Data visualization libraries***


***IDE***

# About the Dataset

**Contents**: Insurance risk symboling and normalized loss for each model, along with body and engine specifications, and price.

**Source**: https://github.com/insaid2018/Term-1/blob/master/Data/Projects/Automobile_data.csv

**Data Volume**: 205 records, 26 variables

**Attribute Information**

1. **symboling**: -3, -2, -1, 0, 1, 2, 3.
2. **normalized-losses**: continuous from 65 to 256.
3. **make**: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. **fuel-type**: diesel, gas.
5. **aspiration**: std, turbo.
6. **num-of-doors**: four, two.
7. **body-style**: hardtop, wagon, sedan, hatchback, convertible.
8. **drive-wheels**: 4wd, fwd, rwd.
9. **engine-location**: front, rear.
10. **wheel-base**: continuous from 86.6 120.9.
11. **length**: continuous from 141.1 to 208.1.
12. **width**: continuous from 60.3 to 72.3.
13. **height**: continuous from 47.8 to 59.8.
14. **curb-weight**: continuous from 1488 to 4066.
15. **engine-type**: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. **num-of-cylinders**: 2, 3, 4, 5, 6, 8, and 12
17. **engine-size**: continuous from 61 to 326.
18. **fuel-system**: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. **bore**: continuous from 2.54 to 3.94.
20. **stroke**: continuous from 2.07 to 4.17.
21. **compression-ratio**: continuous from 7 to 23.
22. **horsepower**: continuous from 48 to 288.
23. **peak-rpm**: continuous from 4150 to 6600.
24. **city-mpg**: continuous from 13 to 49.
25. **highway-mpg**: continuous from 16 to 54.
26. **price**: continuous from 5118 to 45400.

# Importing packages & dataset

## 1. Importing Packages

```
import numpy as np

import pandas as pd
pd.set_option('mode.chained_assignment', None)
pd.set_option('display.max_colwidth', -1)
pd.options.display.max_rows = 999

import matplotlib.pyplot as plt
plt.style.use('seaborn-whitegrid')
%matplotlib inline

import pandas_profiling

import seaborn as sns
sns.set(style='whitegrid', font_scale=1.3, color_codes=True)

from scipy import stats

import warnings
warnings.filterwarnings('ignore')
```

## 2. Importing dataset

```
automobile_data = pd.read_csv('Automobile_data.csv', encoding="ISO-8859-1")
```

Data Cleaning

# Observations from Pandas Profiling

## Before Data Cleaning

**Dataset info:**
- Number of variables: 26
- Number of observations: 205
- Missing cells: 0 (0.0%)

**Variables types:**
- Numeric 9
- Categorical 16
- Boolean 0
- Date 0
- Text (Unique) 0
- Rejected 1
- Unsupported 0

Missing Values:

| Feature | Count of '?' Values |
|---|---|
| Normalized Losses | 41 |
| Number of doors, Horse Power, Peak RPM | 2 |
| Bore, Stroke, Price | 4 |

**Count of NaN Values: 0** for all attributes in our dataset.

```
for col in automobile_data:
    val = automobile_data[col].isna().sum()
```

**Removing Duplicate Rows:**

```
automobile_data.drop_duplicates(keep=False, inplace=True)
```

**Clearing '?' Values:** *Sample code on normalized-losses*

```
a=automobile_data[automobile_data['normalized-losses']!='?']
b=(a['normalized-losses'].astype(int)).median()
automobile_data['normalized-losses']=automobile_data['normalized-losses'].replace('?',b).astype(int)
```

## After Data Cleaning

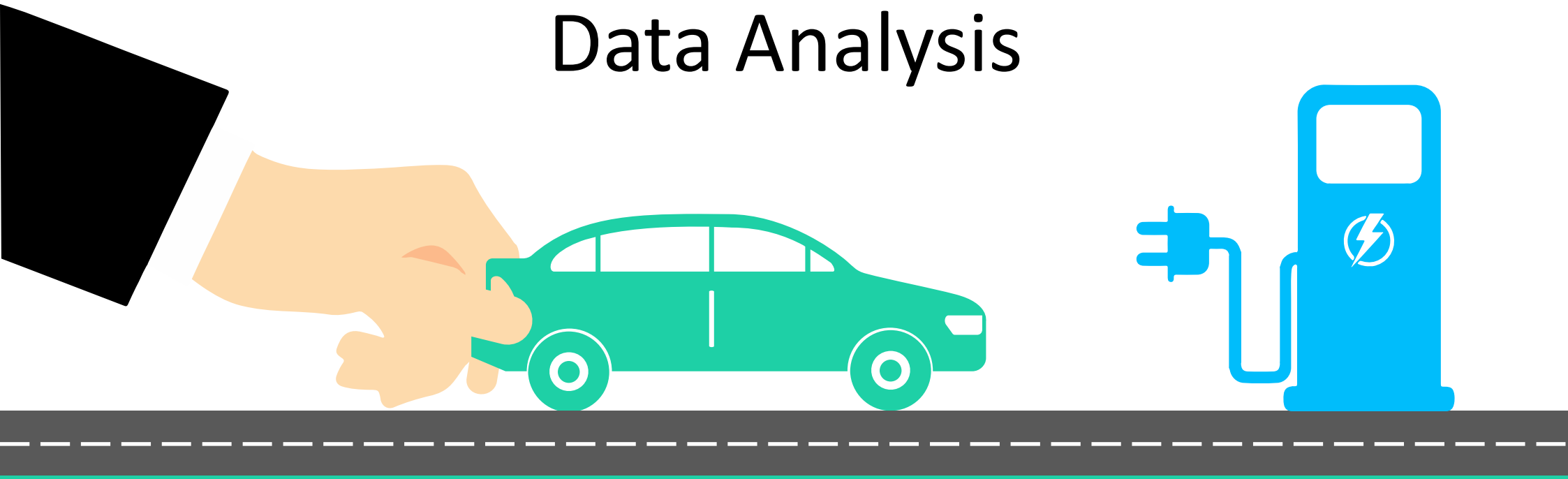
**Dataset info:**
- Number of variables: 26
- Number of observations: 205
- Missing cells: 0 (0.0%)

**Variables types:**
- Numeric = 16
- Categorical = 8
- Boolean = 1
- Date = 0
- Text (Unique) = 0
- Rejected = 1
- Unsupported = 0

**automobile_data.info()**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
symboling            205 non-null int64
normalized-losses    205 non-null object
make                 205 non-null object
fuel-type            205 non-null object
aspiration           205 non-null object
num-of-doors         205 non-null object
body-style           205 non-null object
drive-wheels         205 non-null object
engine-location      205 non-null object
wheel-base           205 non-null float64
length               205 non-null float64
width                205 non-null float64
height               205 non-null float64
curb-weight          205 non-null int64
engine-type          205 non-null object
num-of-cylinders     205 non-null object
engine-size          205 non-null int64
fuel-system          205 non-null object
bore                 205 non-null object
stroke               205 non-null object
compression-ratio    205 non-null float64
horsepower           205 non-null object
peak-rpm             205 non-null object
city-mpg             205 non-null int64
highway-mpg          205 non-null int64
price                205 non-null object
dtypes: float64(5), int64(5), object(16)
memory usage: 41.8+ KB
```

# Data Analysis



**why data analysis?**

Data analysis helps us to unlock the information and insights from raw data. So data analysis plays an important role by helping us to discover useful information from the data, answer questions, and even predict the future or the unknown.

Assume you want to sell your car? But the problem is, you don't know how much you should sell your car for. But you also want to set the price reasonably so someone would want to purchase it.

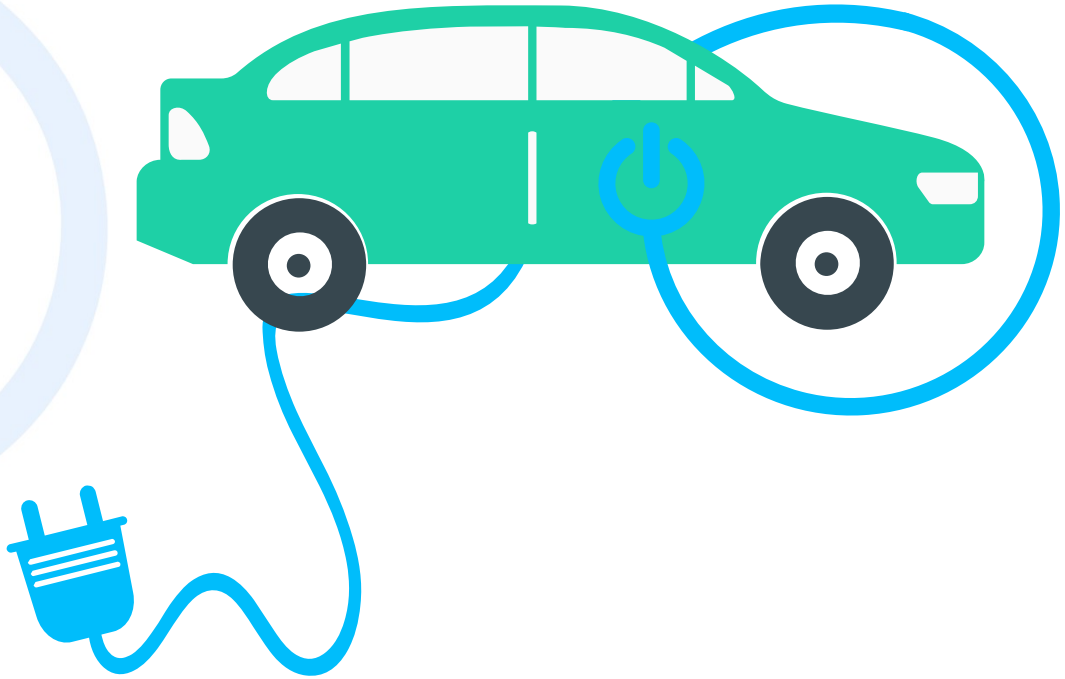
As a data analyst we should ask various questions to solve the problem. As for the problem here,

- what were the features that affect the price of the car? is it color or brand?
- Does horsepower also affect the selling price, or perhaps, something else?

**In the coming slides we'll see answers to some of such questions.**

# It consists of.....

1. Univariate Analysis
2. Multivariate Analysis
3. Risk Analysis
4. Correlation Analysis



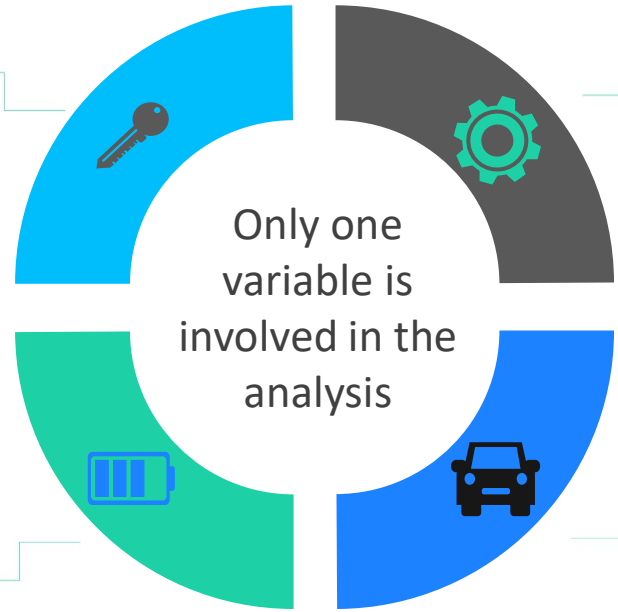
# Univariate Analysis

**1. Basic details of the vehicle**

make (brand),
symbolling (insurance risk rating),
normalized losses and
price of the car.

**2. Engine specifications**

engine type,
engine size and its location,
No. of cylinders and compression ratio,
fuel system it uses and
horsepower it produces.



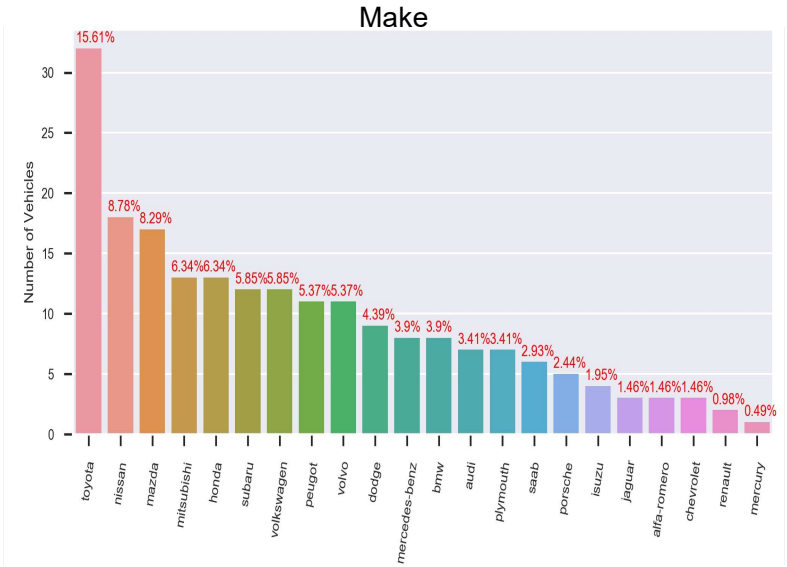

**4. Fuel & Efficiency**

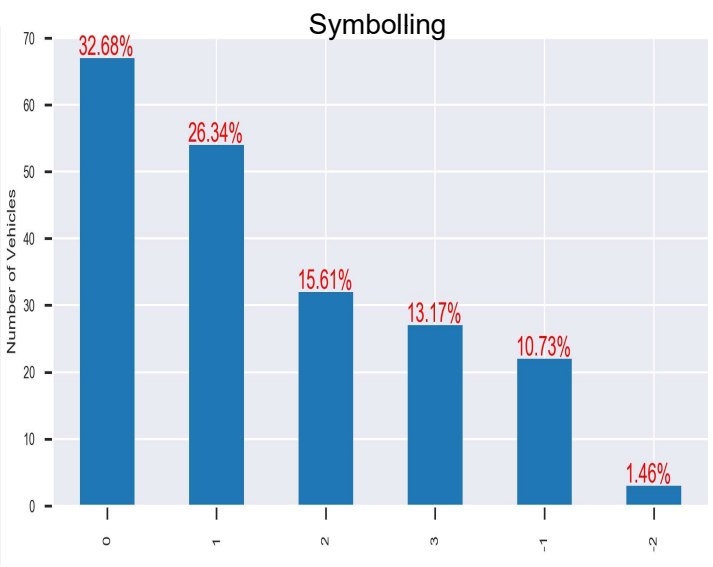
type of the fuel,
aspiration and
mileage it offers
in the city and the highway.

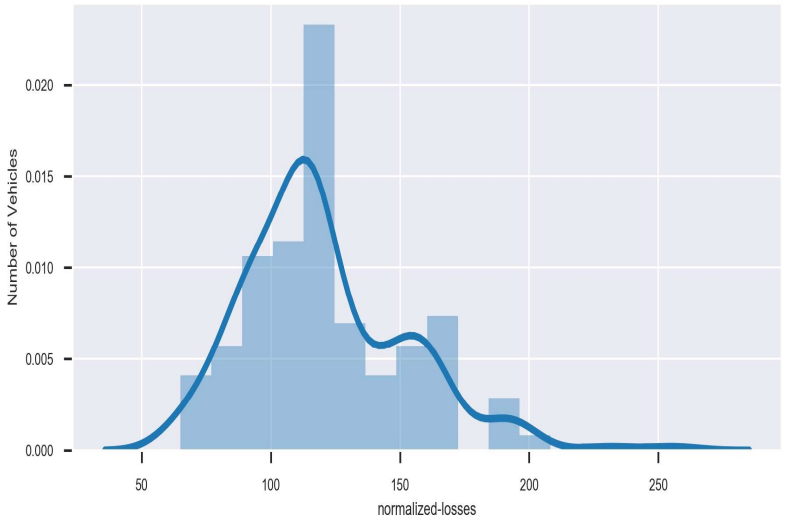
**3. Vehicle Dimensions**

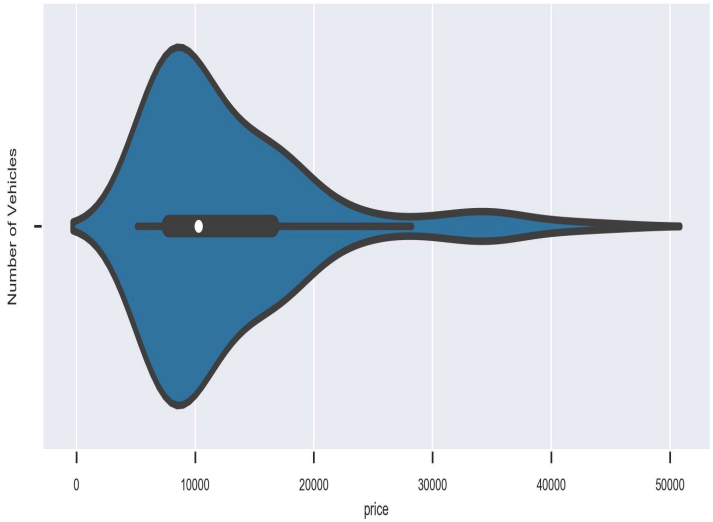
curb-weight,
wheel base,
Length X width X height,
body style,
no. of doors it has,
type of wheel drive it offers.

# 1. Basic Details




Toyota is the make of the car which has **most number of vehicles** with more than 40% than the 2nd highest is Nissan.

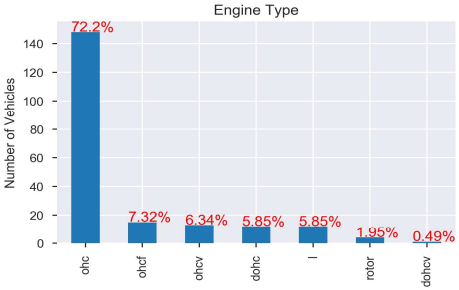
**Insurance risk rating** have the ratings between -3 and 3 however for our dataset it starts from -2. There are more cars in the range of 0 and 1.

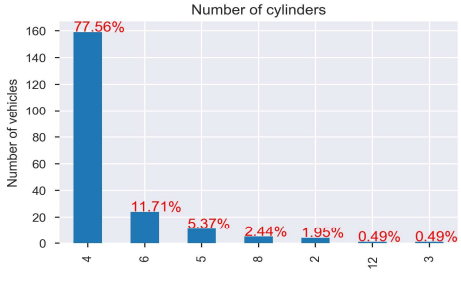
**Normalized losses** which is the average loss payment per insured vehicle per year has more number of cars in the range between 175 and 200.

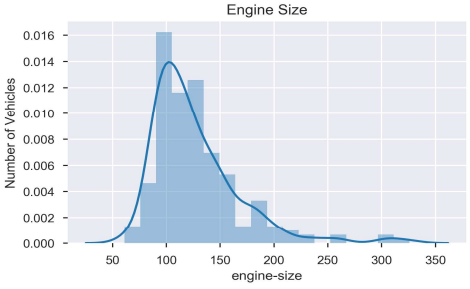
The violin plot shows an outlier on the higher value side.

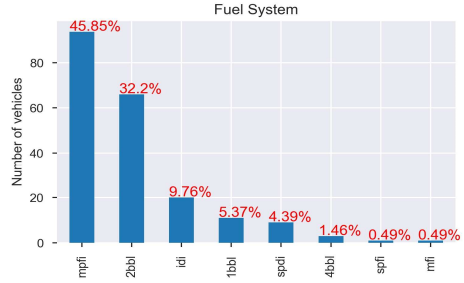
Most of the cars has the **price** of $20k and less.

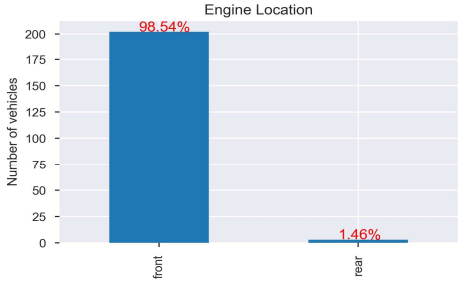
# 2. Engine Specifications

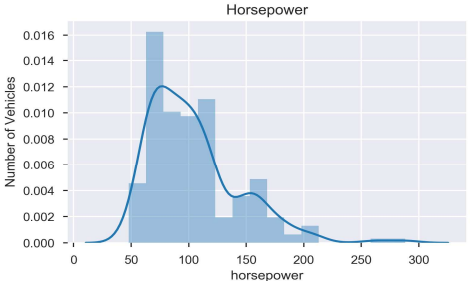


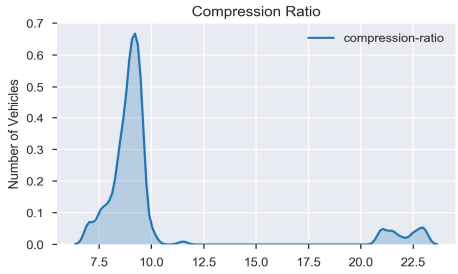



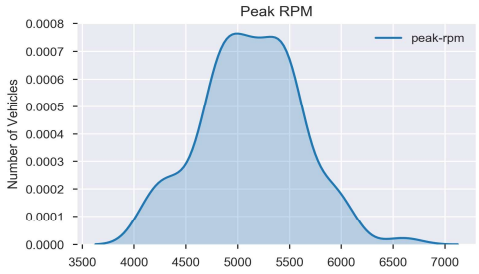



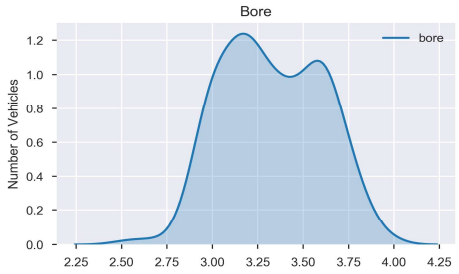



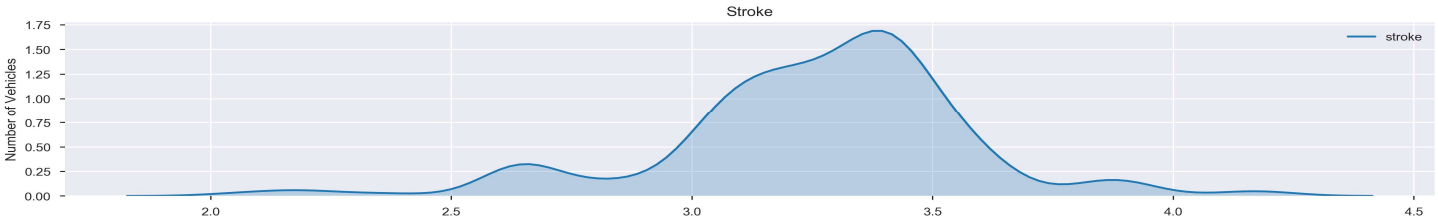



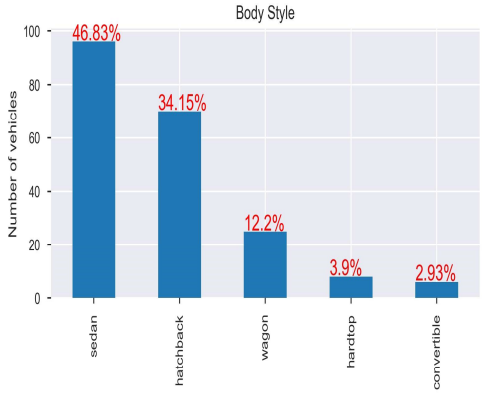



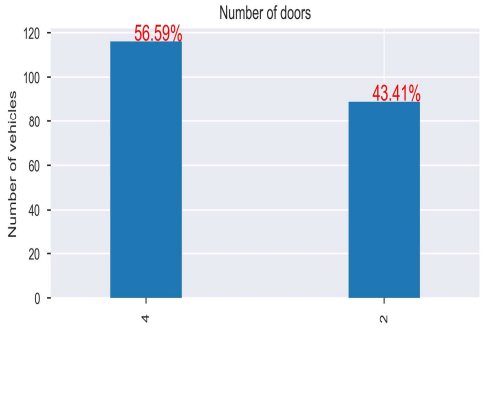



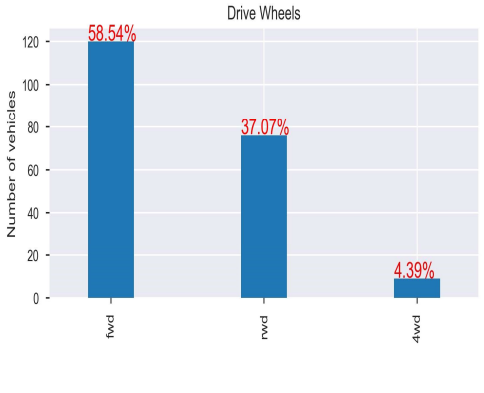

Most cars in our dataset have the following Engine parameters ranges:

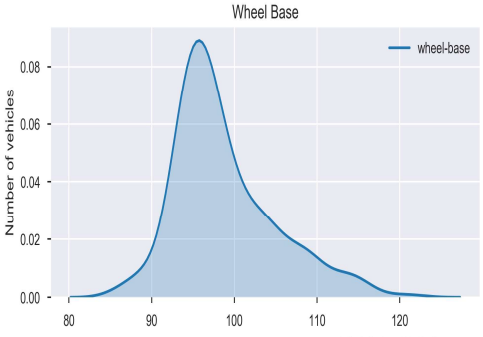
- Engine Type - Ohc
- No. of cylinders - 4
- Engine size - 90 to 160 cc
- Fuel System - mpfi
- Engine Location - Front
- Horsepower - 50 to 200 hp
- Compression Ratio - 8:1 to 10:1
- Peak RPM - 4500 to 6000 rpm
- Bore - 2.75 to 4 inch
- Stroke - 2 to 4 inch

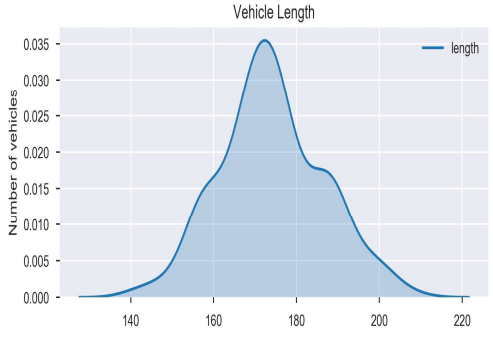
# 3. Vehicle Dimensions

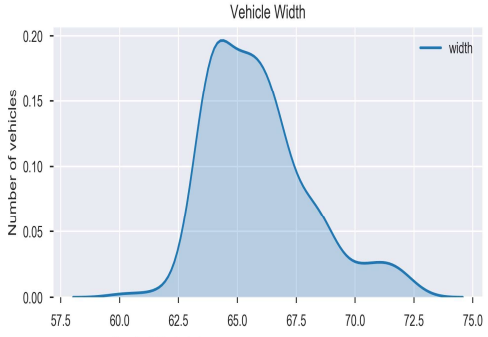


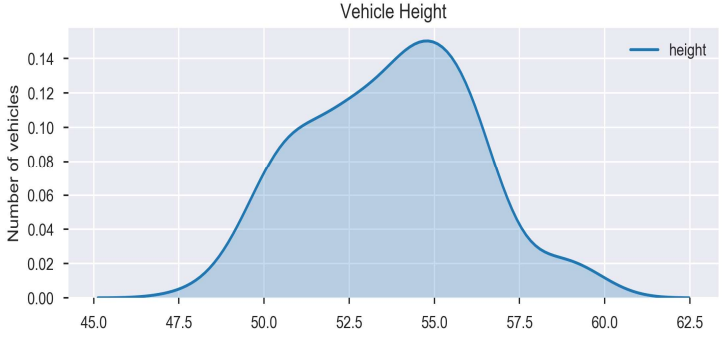



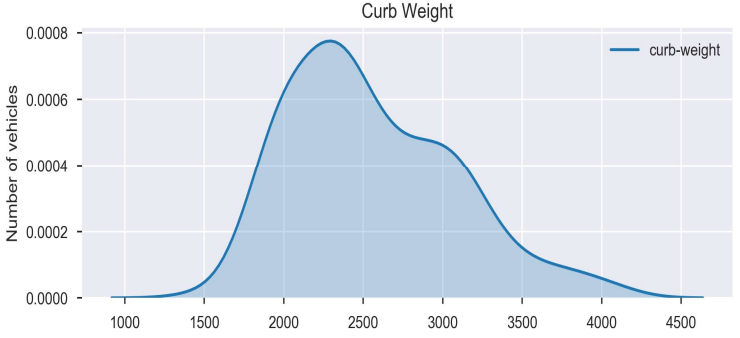



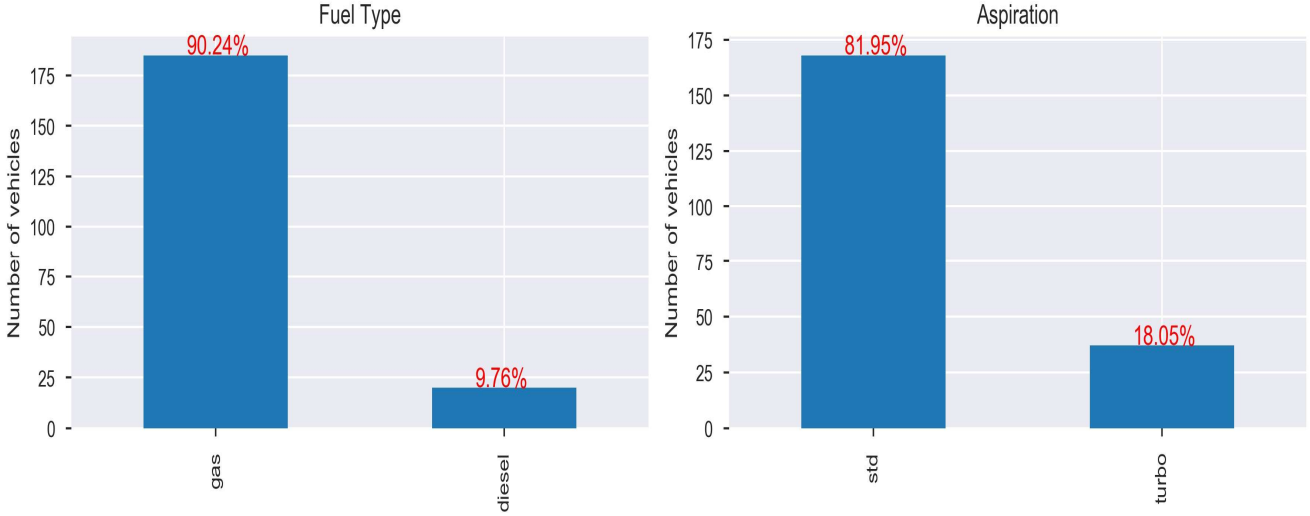

**Sedan** type cars having **4** doors with **forward wheel drive** has the majority share in the market.

For drive wheels, front wheel drive has most number of cars followed by rear wheel and four wheel. There are **very less** number of cars for four wheel drive.

Most cars have the following Vehicle dimensions:

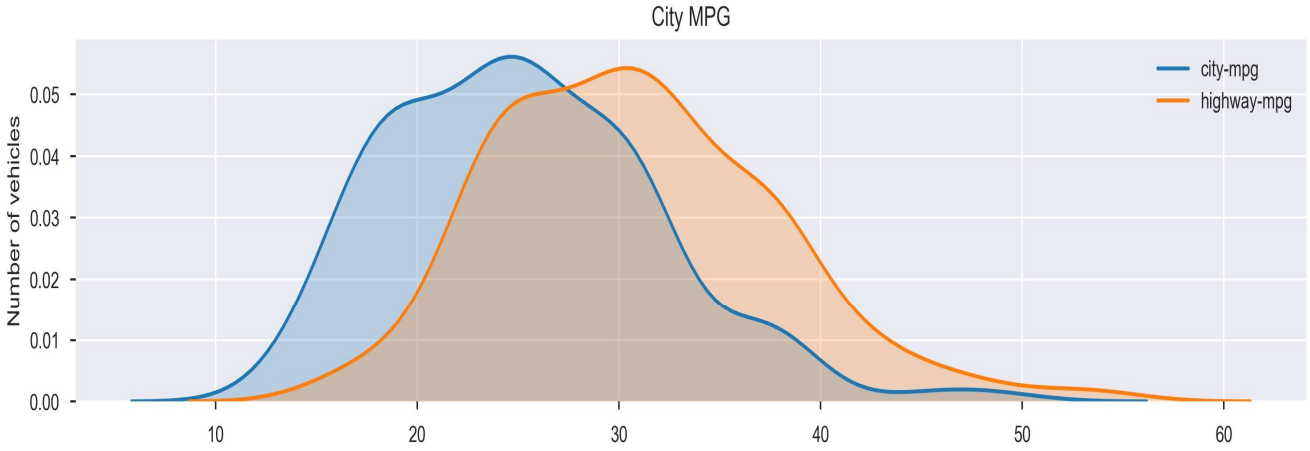
- Wheel base - 90 to 110 inch
- Vehicle length - 150 to 200 inch
- Vehicle width - 62.5 to 70 inch
- Vehicle Height - 48.0 to 60 inch
- Curb Weight - 1.5 to 3.5 ton

# 4. Fuel & Efficiency










Most preferred **fuel type** by the customers is **gas** and the **aspiration type** is **standard** against turbo having more than 80% of the choice.

Highway mileage is more than the city mileage for all the vehicles.

# Multivariate Analysis

**1. Analysis based on price**

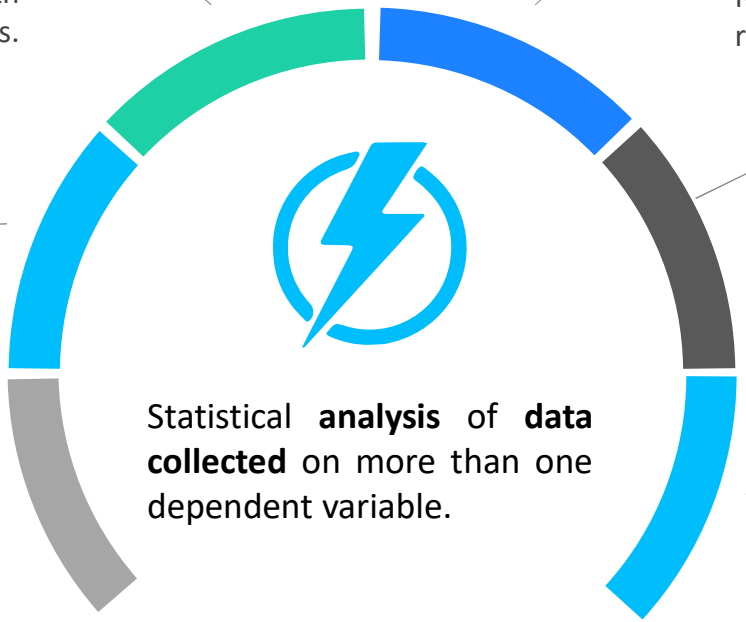
Price variations are observed with respect to other attributes.

**2. Analysis based on normalized losses**

Normalized losses are analyzed with respect to brand and body styles..

**3. Analysis based on fuel**

Most preferred fuel, average mileage, effect of drive wheels & curb weight on mileage are analyzed here.



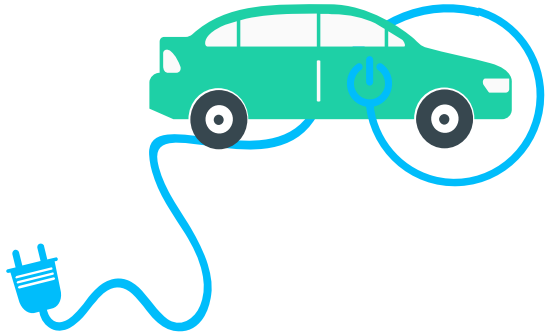
Statistical **analysis** of **data collected** on more than one dependent variable.

**4. Analysis based on engine & vehicledimensions**

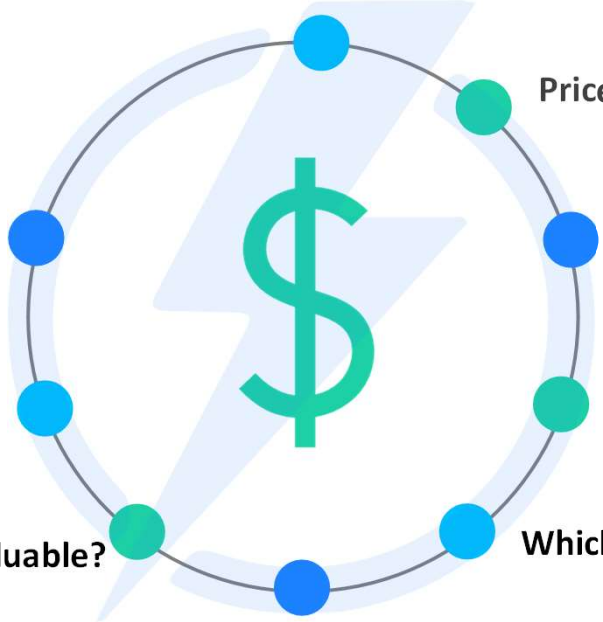
Horsepower, no. Of cylinders, curb-weight and engine size relation, mileage as per engine type are analyzed here. Also, classification of cars as per body size we'll have a look.

**5. Analysis based on efficiency parameters**

# 1. Analysis based on price



We'll explore the following queries for price analysis.



Price vs continuous attributes

Price vs categorical attributes

How the price is distributed among different body sizes?

What is the price of the cars by body-style and engine wheels type?

What is the correlation between drive-wheels and body-style as per the price?

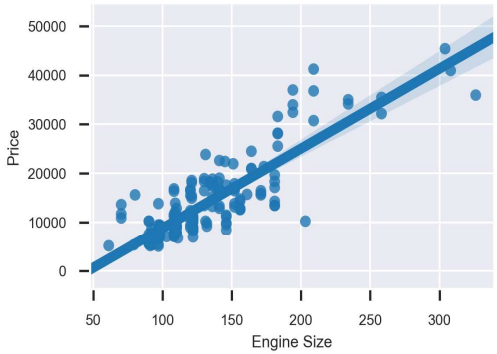
What is the price of cars by engine type and drive wheels type?

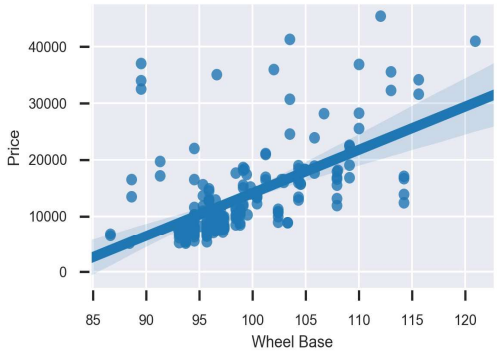
which type of drive wheel is most valuable?

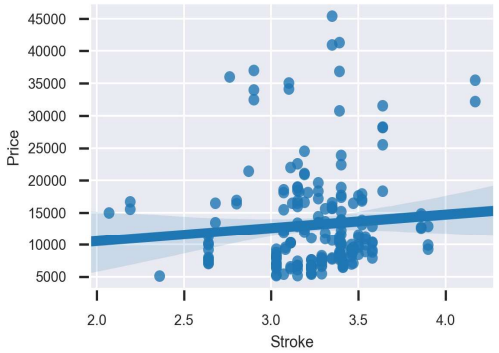
Which price segment cars sold the most ?

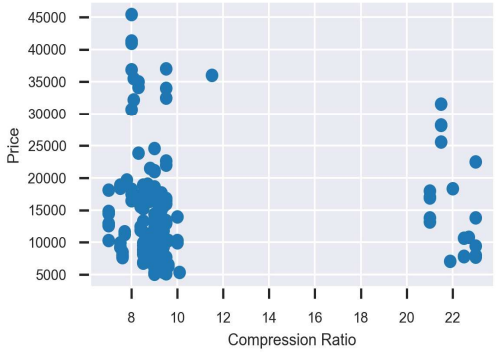
What are the price variations across different brands?

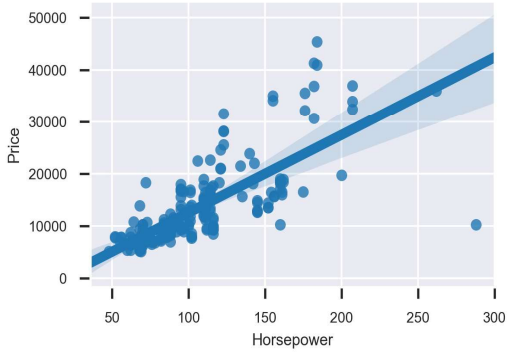
# Price vs Continuous attributes

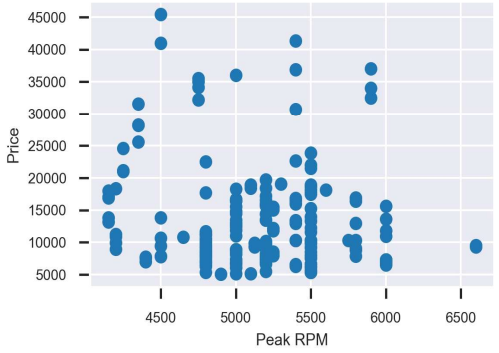


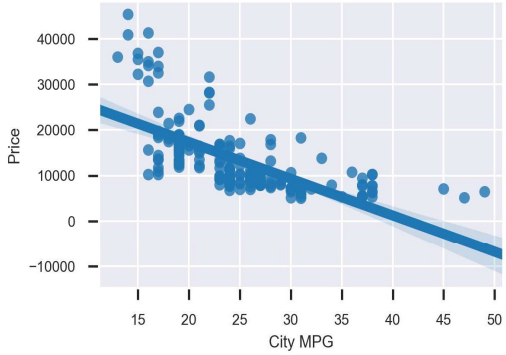



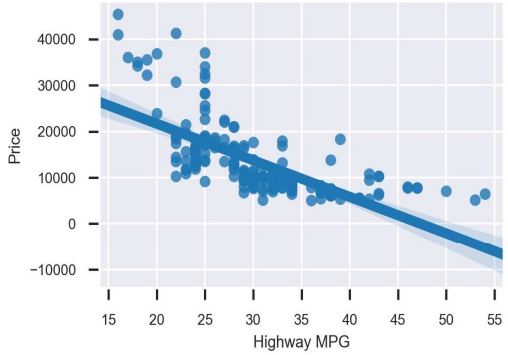




Price increases with the increase in engine size, wheel base, Horsepower.

Engine size is directly proportional to price of the vehicle. As the size of the engine increases the price also increases.

Most of the vehicles are in the price range of 20k USD and most of the vehicles has engine sizes of up to 200 cc.

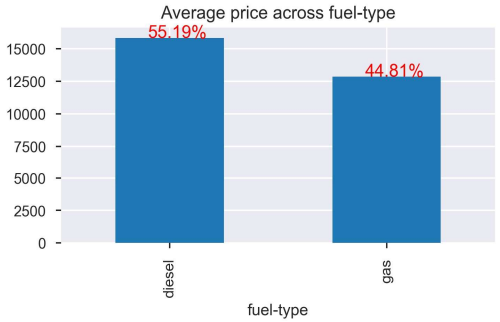
Vehicle with high price have low mileage. This because high priced vehicles go into luxury segment which are meant for speed, high performance and running cost is not very important in this segment.

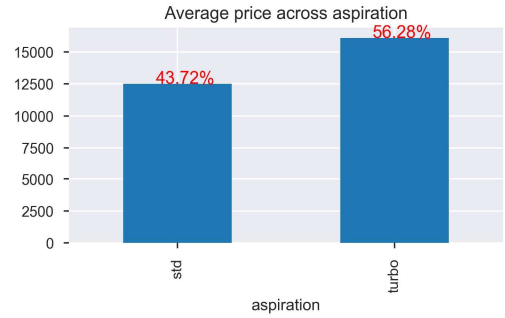
Engine-size, horsepower and mileage are having high correlation with price.

Fuel Efficiency has a negative correlation with price.

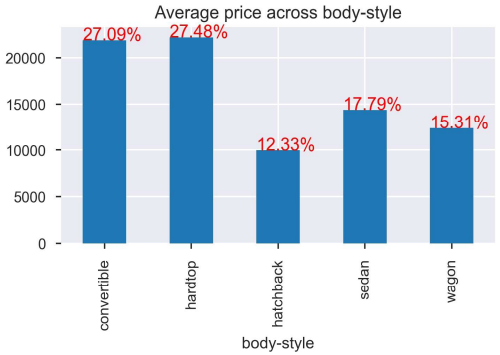
People who look for highly fuel efficient car will normally be budget conscious. Hence it is probable that those cars are made for lower price brackets.

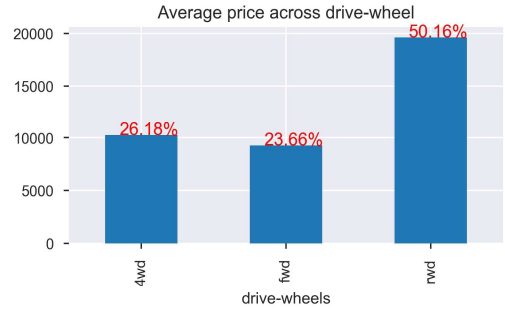
# Price vs Categorical attributes

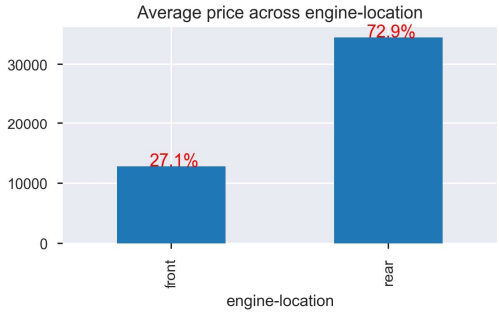


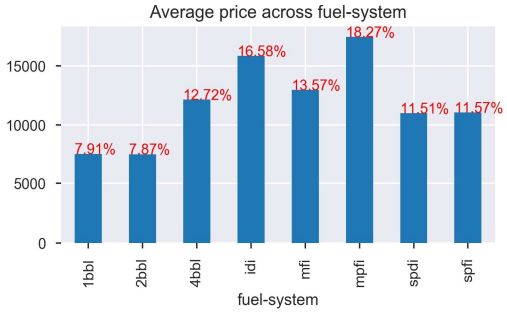



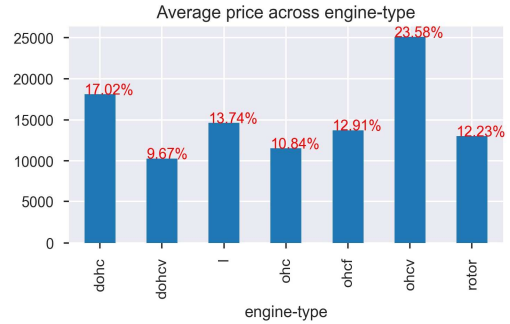


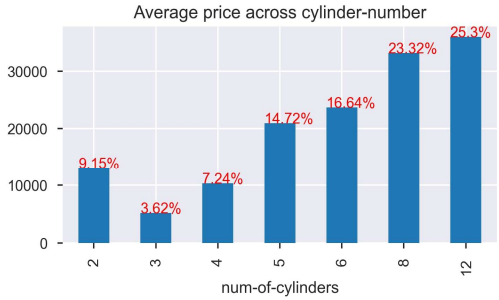



Price is more for diesel type. That's why if went to see univariate analysis of Fuel & Efficiency, there most people opted for gas. Same is the case with aspiration, engine location and engine type.

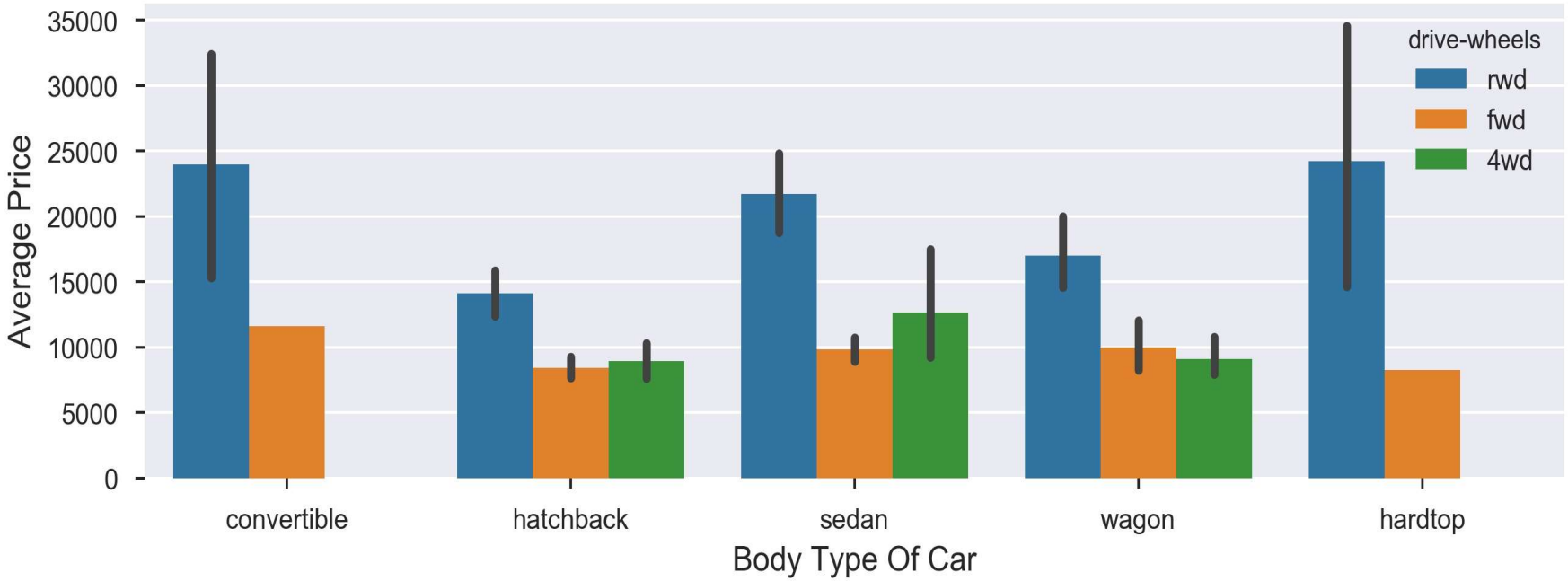
Even though the hatchback cost is less, most people are going for sedan vehicles.

Rear wheel drive type cars are the most expensive.

Even though the price of mpfi engine is more, people are going for it because of the following reasons.

- Engine vibrations from MPFI equipped engines are very less, hence the life of MPFI system equipped engines is high.
- This system is very responsive in case of sudden acceleration or deceleration.
- Lower fuel consumption leads to better mileage. The volumetric efficiency of MPFI is high.

# What is the price of cars by body style and drive wheels type?



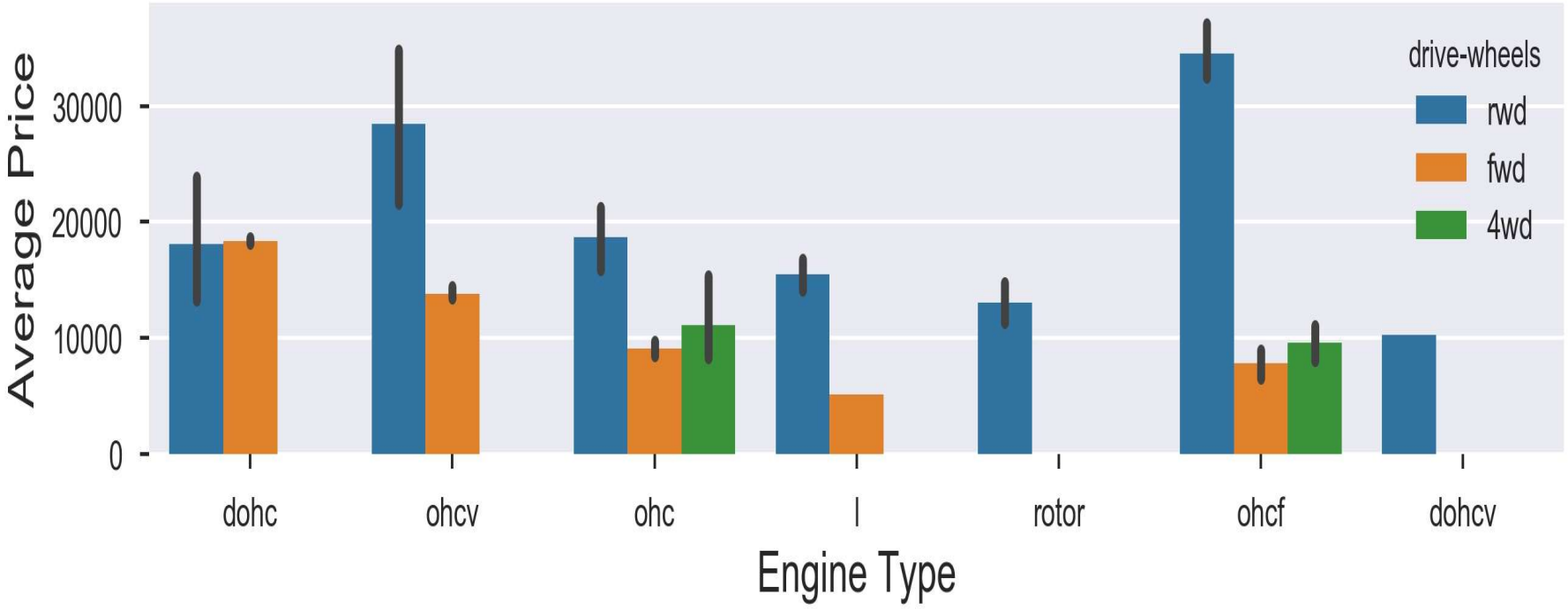

Price is more for rear wheel drive type for all the body types. Most of the cars present under price of $10K USD. But none of them are rear wheel drive type.

# which type of drive wheel is most valuable?

| | price | | | | |
|---|---|---|---|---|---|
| body-style | convertible | hardtop | hatchback | sedan | wagon |
| drive-wheels | | | | | |
| 4wd | 0.0 | 0.000000 | 8949.000000 | 12647.333333 | 9095.750000 |
| fwd | 11595.0 | 8249.000000 | 8396.387755 | 9828.754386 | 9997.333333 |
| rwd | 23949.6 | 24202.714286 | 14125.000000 | 21711.833333 | 16994.222222 |

**Rear wheel drive is the most valuable under hardtop body** model car and it would price an average of **24203 USD**.

# What is the price of the cars based on engine type and drive wheels type?

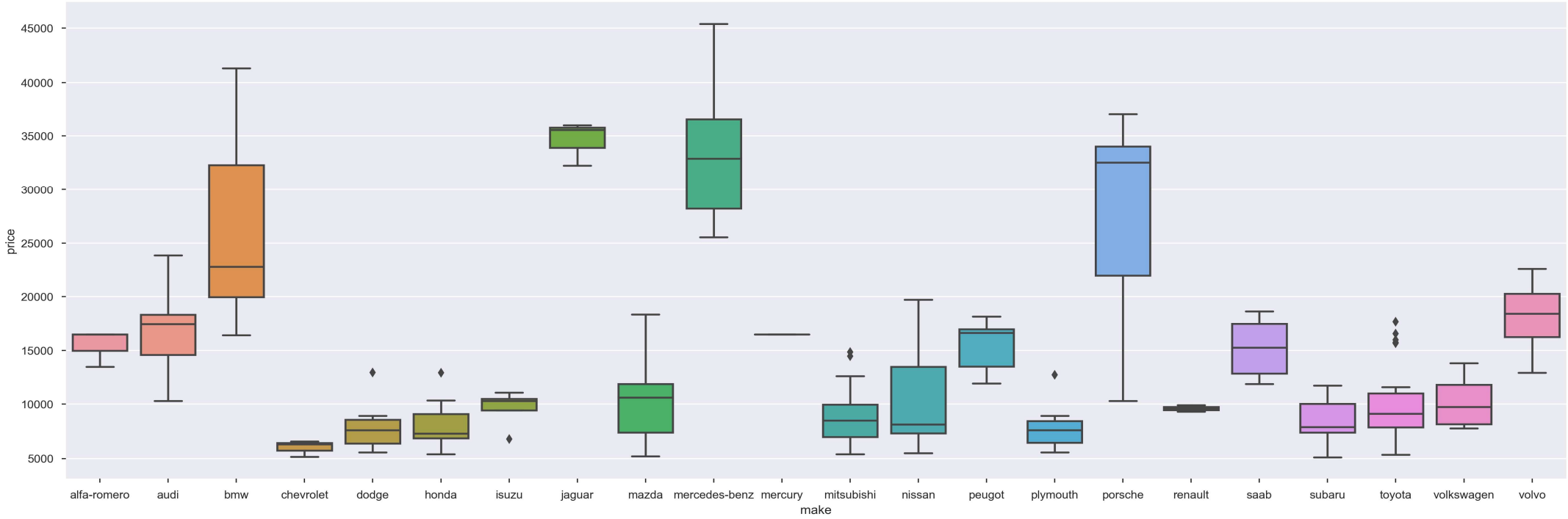



Price is more for **rear** wheel drive type for all the body types..

# Which price segment cars sold the most?

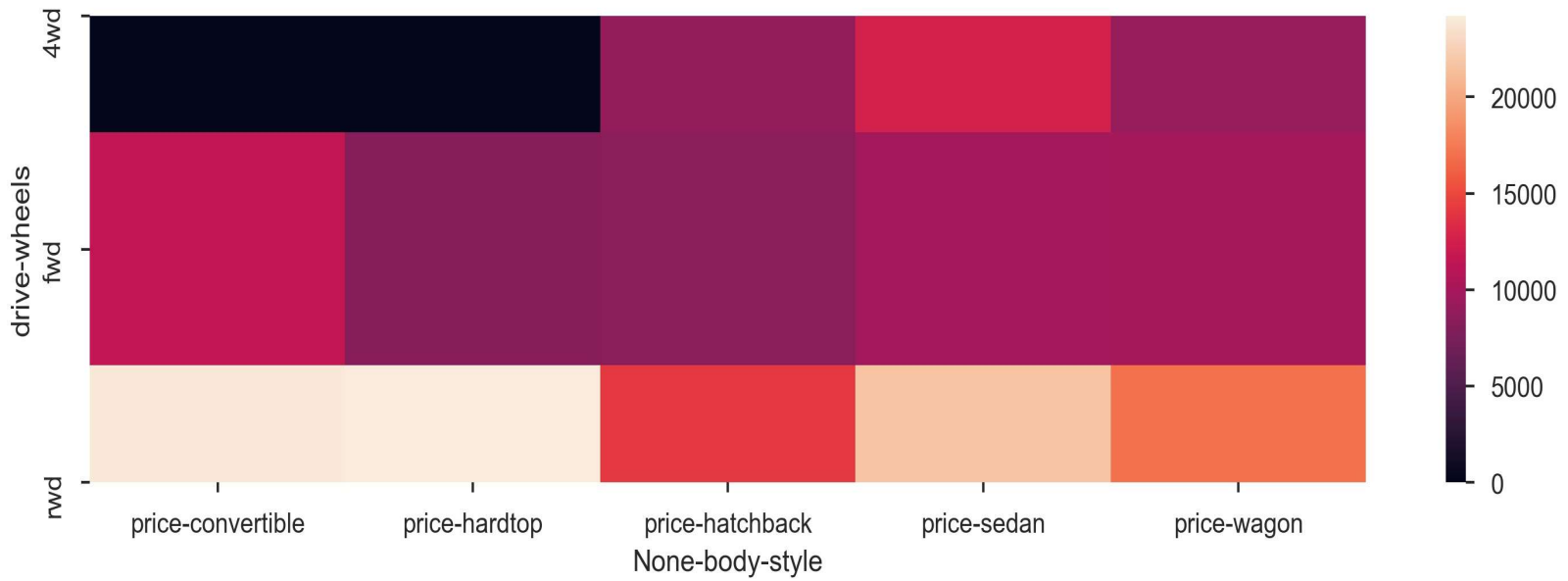

From this bar diagram we can think that, most of the cars belongs to **low budget** range which are used for daily commute. There are very less number of people (**1.46%**) purchasing luxury cars.

# What are the price variations across different brands?



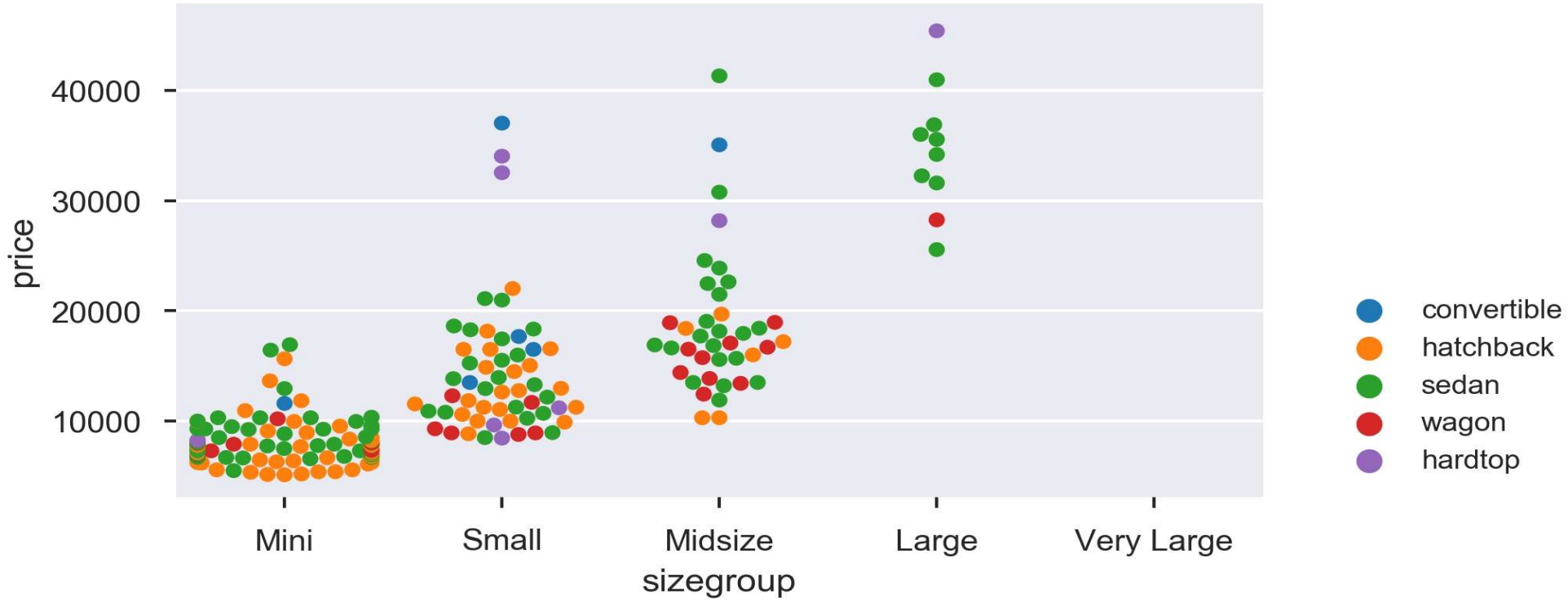

- The **most expensive car** is from **Mercedes Benz** and the **least expensive car** is from **Chevrolet.**
- The **premium** cars costing more than 20000 are BMW, Jaquar, Mercedes Benz and Porsche.
- **Less expensive** cars costing less than 10000 are Chevrolet, Dodge, Honda, Mitsubishi, Plymouth and Subaru.
- Rest of the cars are in the **midrange** between 10000 and 20000 which has the highest number of cars.

# what is the correlation between drive-wheels and body-style as per the price?




- **Rear** wheel drive with **hardtop** body style is the most expensive car
- 4 wheel drive type, don't have convertible and hardtop models.
- In rear wheel type, hatchbacks are cheap.
- In forward wheel drive, convertibles are cheap and in 4 wheel drive, sedans are cheap.

# How the price is distributed among different body sizes?




- For mini segment cars, price is below $20K USD.
- Average price of small segment cars is $15K USD.
- Sedan type vehicles are present in all the segments.
- Midsize segment has very large price variation. It has almost all price variants

# 2. Analysis based on normalized losses

**Risk Rating**

We’ll analyze normalized losses by considering risk rating and body style.

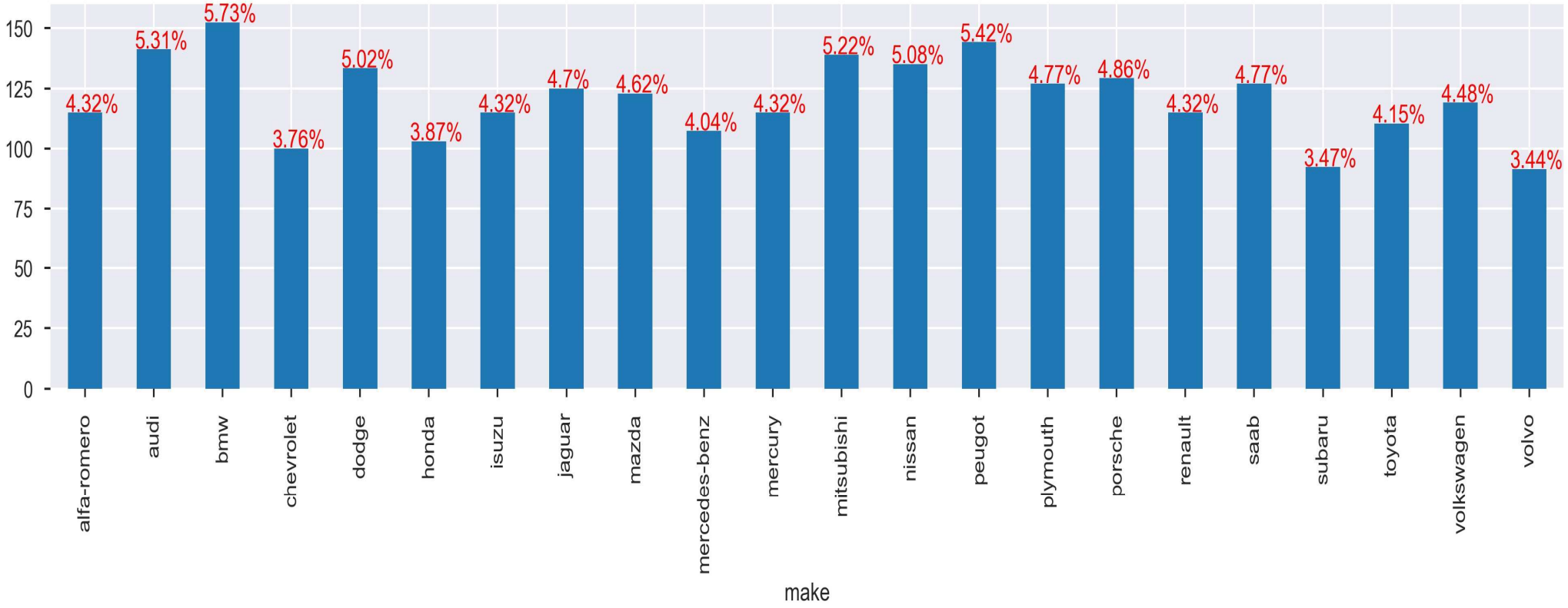

**Make (Brand)**

Here, we'll analyze normalized losses across different car makers. For our analysis purpose we'll find which car maker has maximum percentage of normalized losses.

**Body Style**

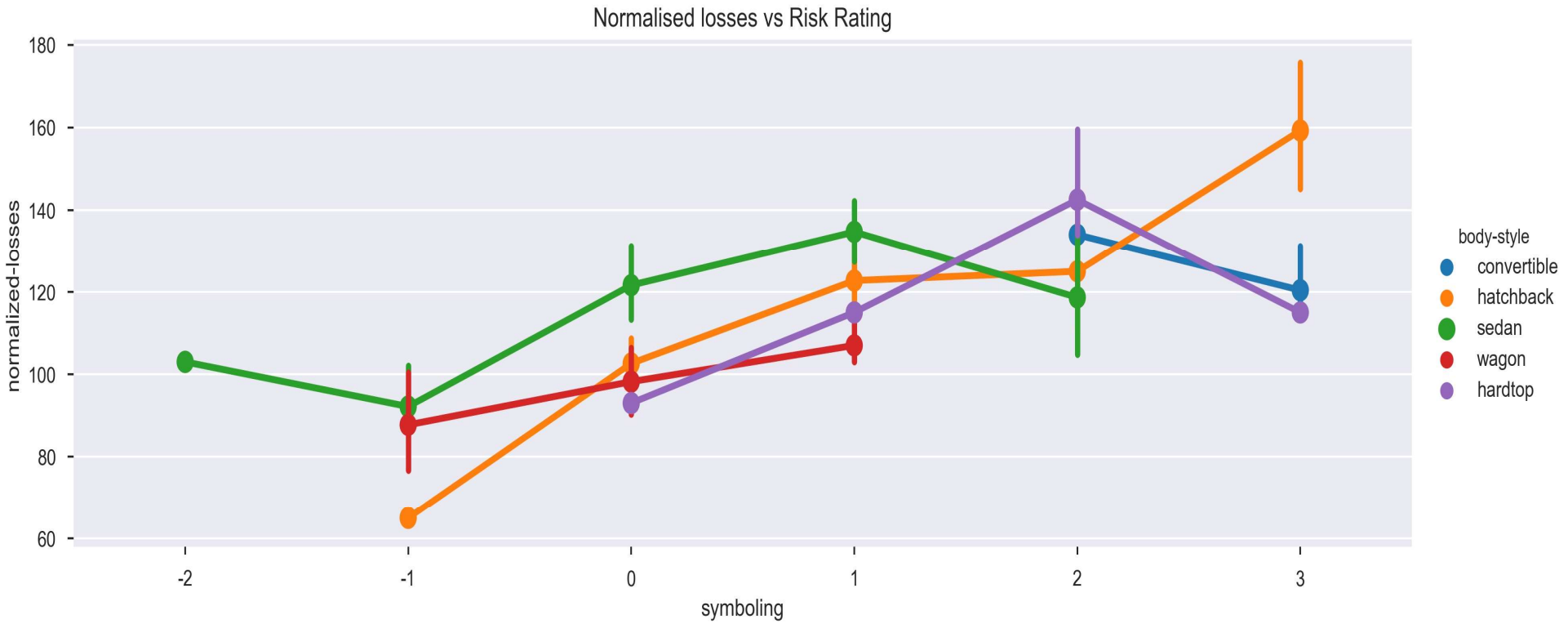
We'll analyze normalized losses as per body style depending on number of doors.

# Which company has the highest normalized losses?




**BMW** has the most normalized losses having **5.73%** of overall losses among all the makers.

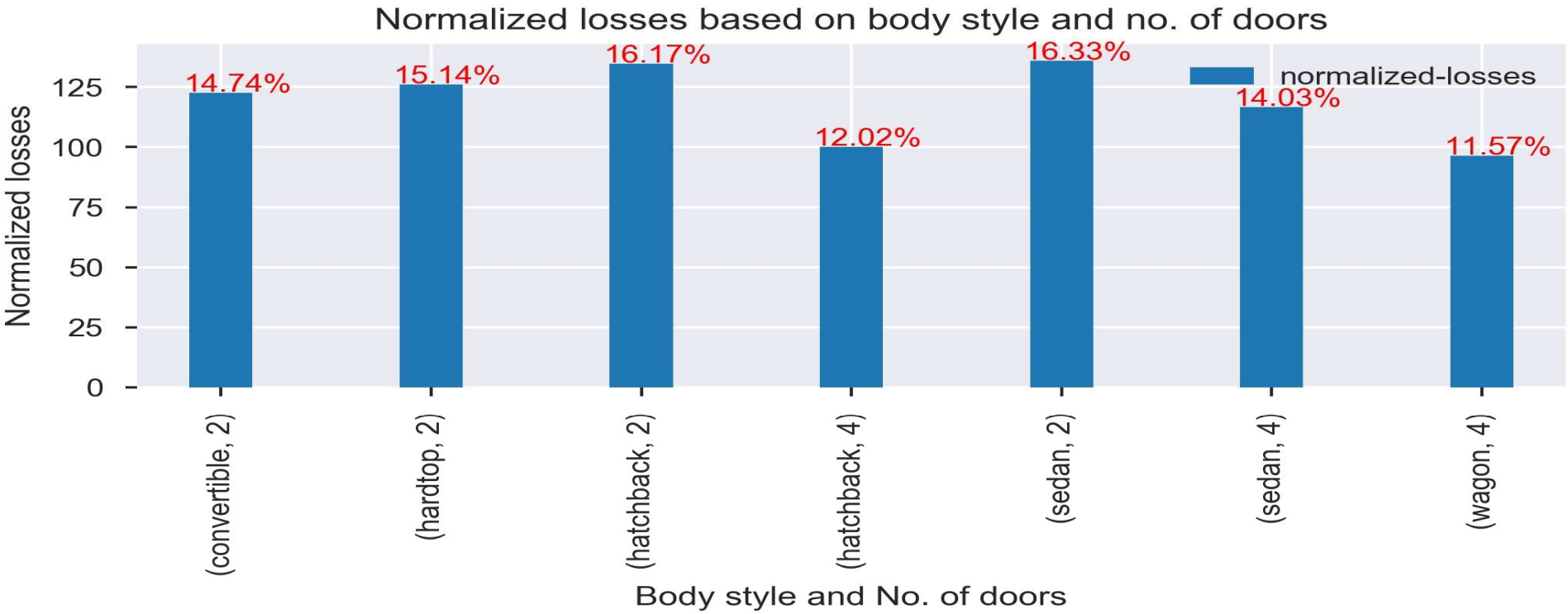
# How Normalized losses are related to Risk Rating across different body-styles?




- Normalized losses linearly increases with the increase in risk rating.
- Convertible and hardtop cars have more losses with risk rating above 0.
- Hatchback cars have highest range of losses with risk rating 3.
- Sedan and Wagon car have losses even with less risk rating.

- Sedan cars have the high range of normalized losses at risk rating **2**.
- **Convertibles** have **only 2 and 3** as risk rating.

# What are the Normalized losses across different body styles?




- If we compare 2 door and 4 door variants, **2 door** variants have **more** no. of losses for all vehicle types.
- Sedan with 2 doors is having more normalized losses compared to others.

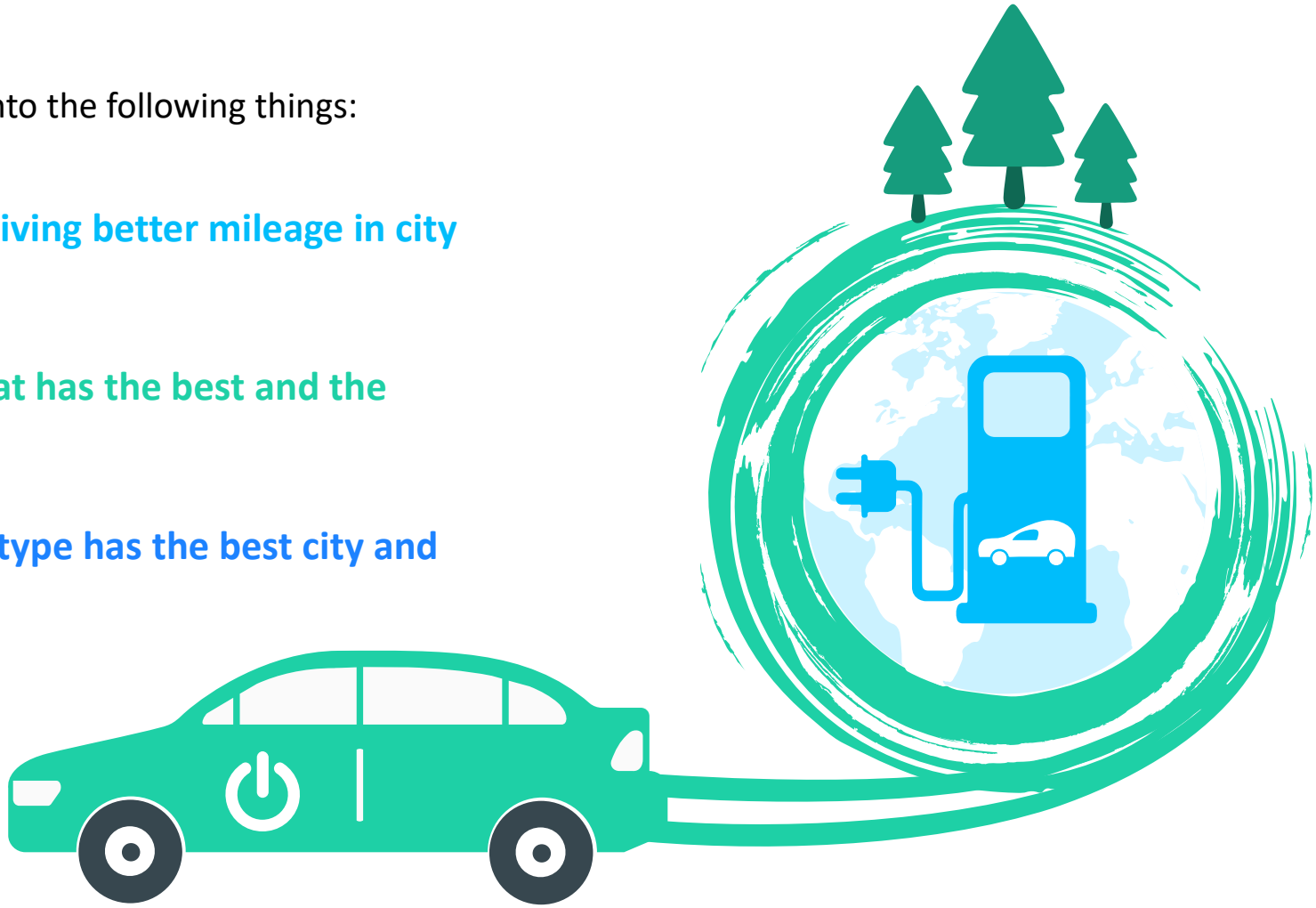
# 3. Analysis based on fuel

In this section we'll look into the following things:

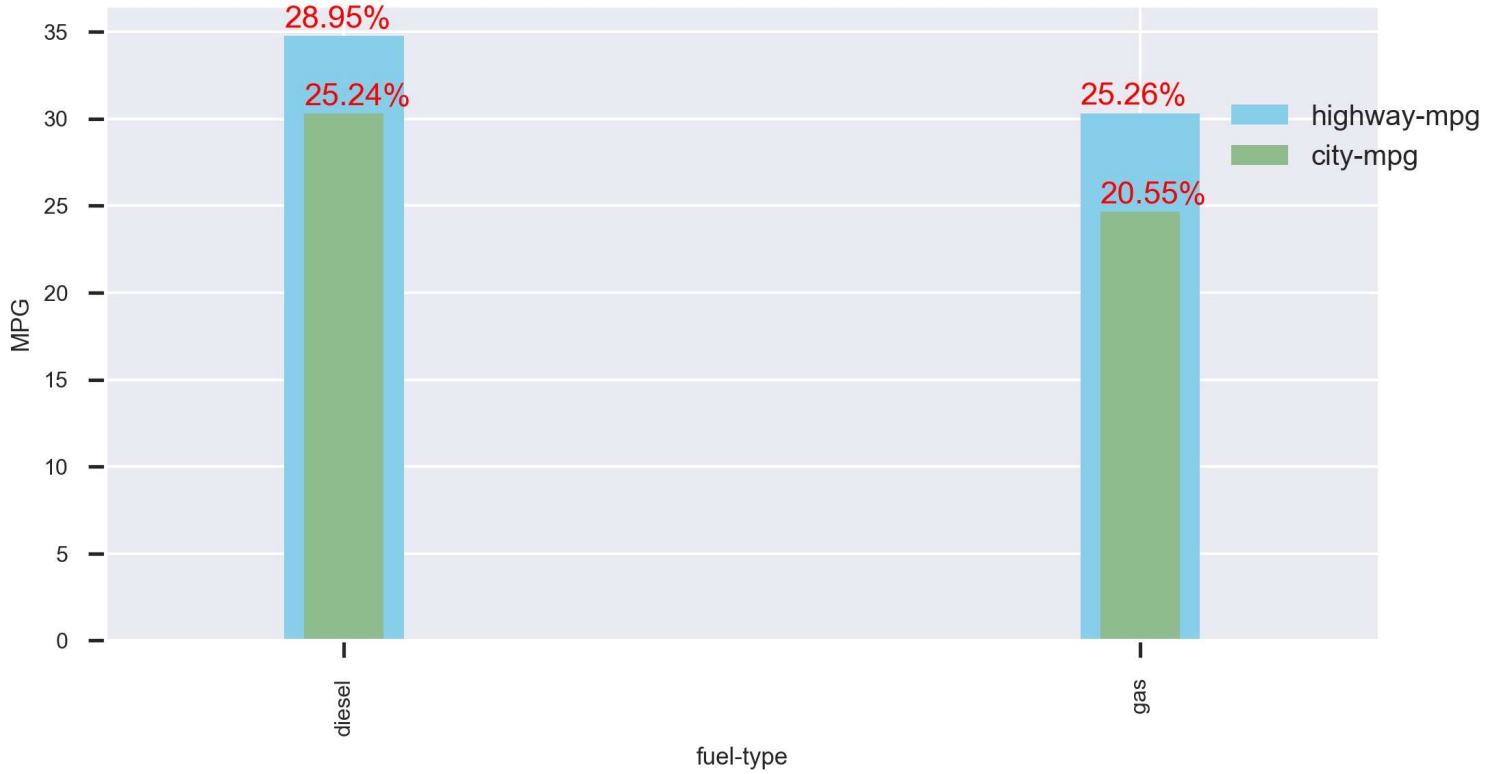
**1. Which Fuel-Type is giving better mileage in city & in highway?**

**2. Name the brands that has the best and the least fuel economy?**

**3. Which drive wheels type has the best city and highway mileages**

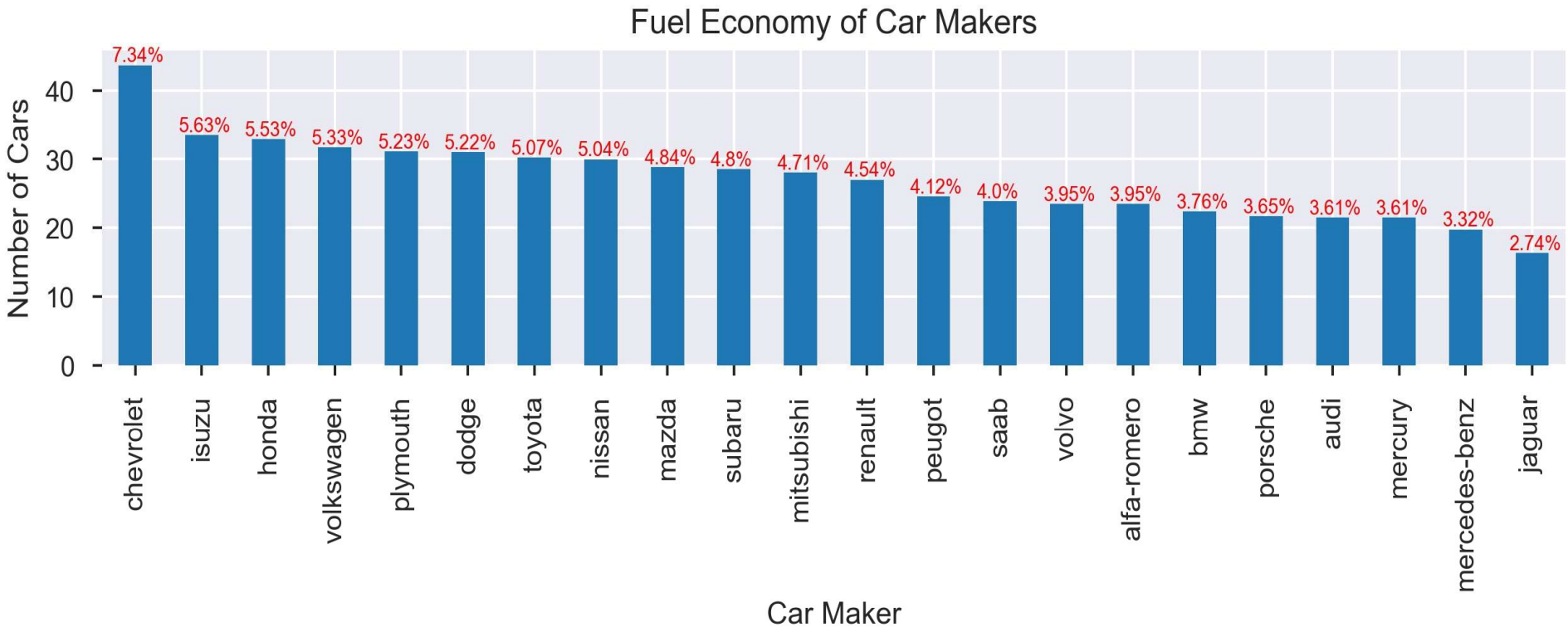


# Which Fuel-Type is giving better mileage in city & highway?



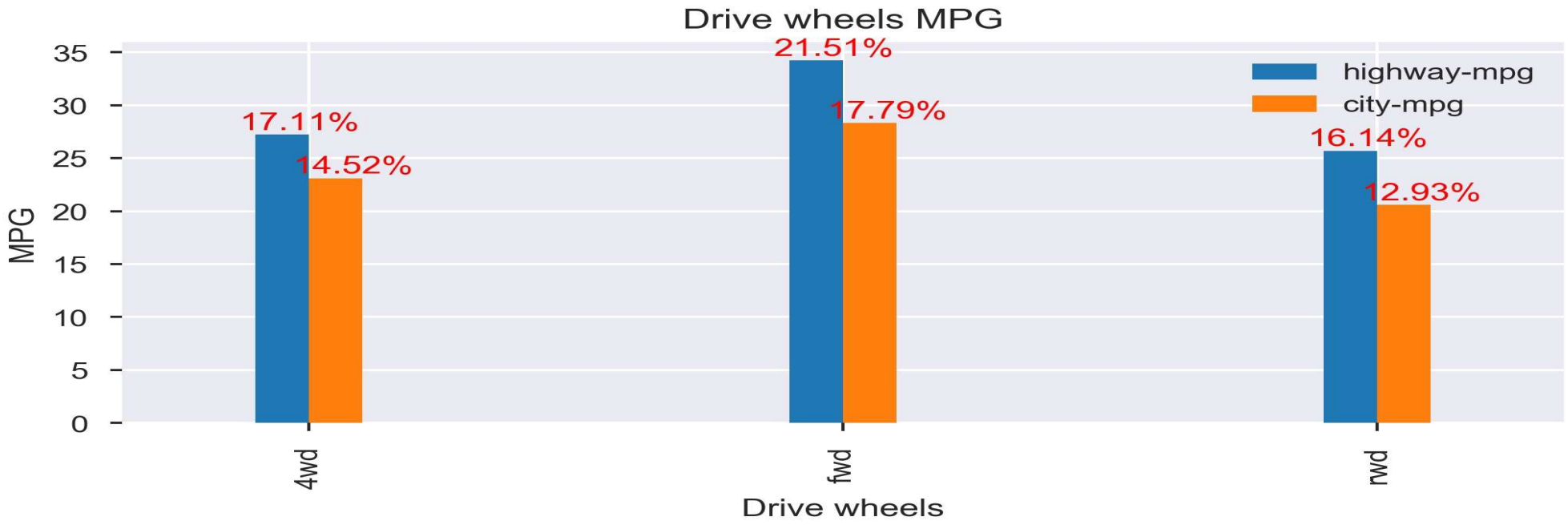

**Diesel** cars has the best mileage in both city & highway traffic scenarios.

# Name the brands that has the best and the least fuel economy



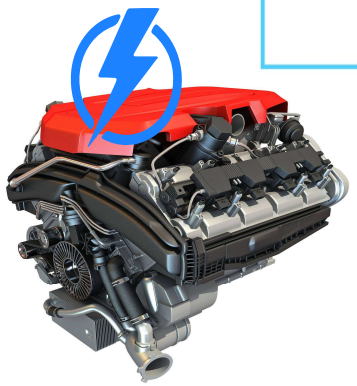

- **Chevrolet** has the **best** fuel economy with most number of cars (**7.34%** of cars).
- **Jaguar** has the **least** fuel economy(**2.74%** of cars)

# Which drive wheels type has the best city and highway mileages




Forward drive wheel type is giving the mileage in both city and highway traffics.

# 4. Analysis based on Engine size & Vehicle dimensions



How does horsepower is related to no. of cylinders?

Which engine type is giving more mileage in the city and what is the type of the fuel it uses?

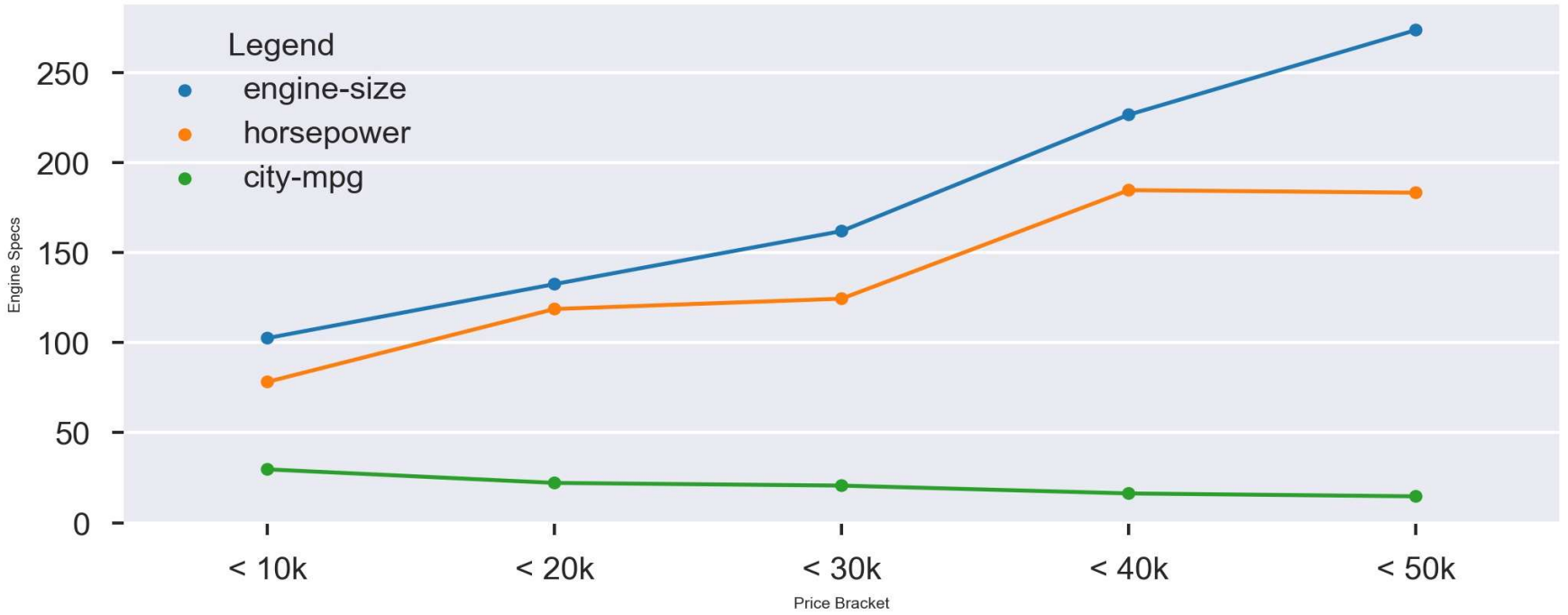
1 2 3 4 5

Continuous attributes of engine specifications and their variation in price?

What is the effect of curb-weight on engine size?

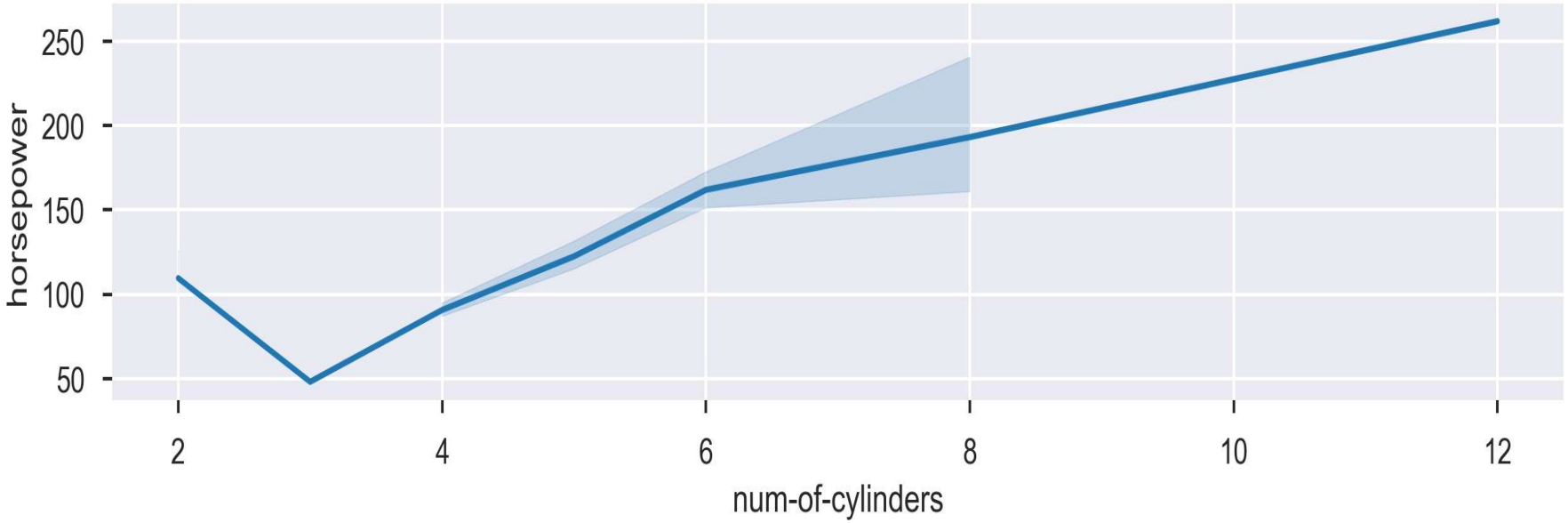
Classify the cars based on its body size

# Continuous attributes of engine specifications and their variation in price?



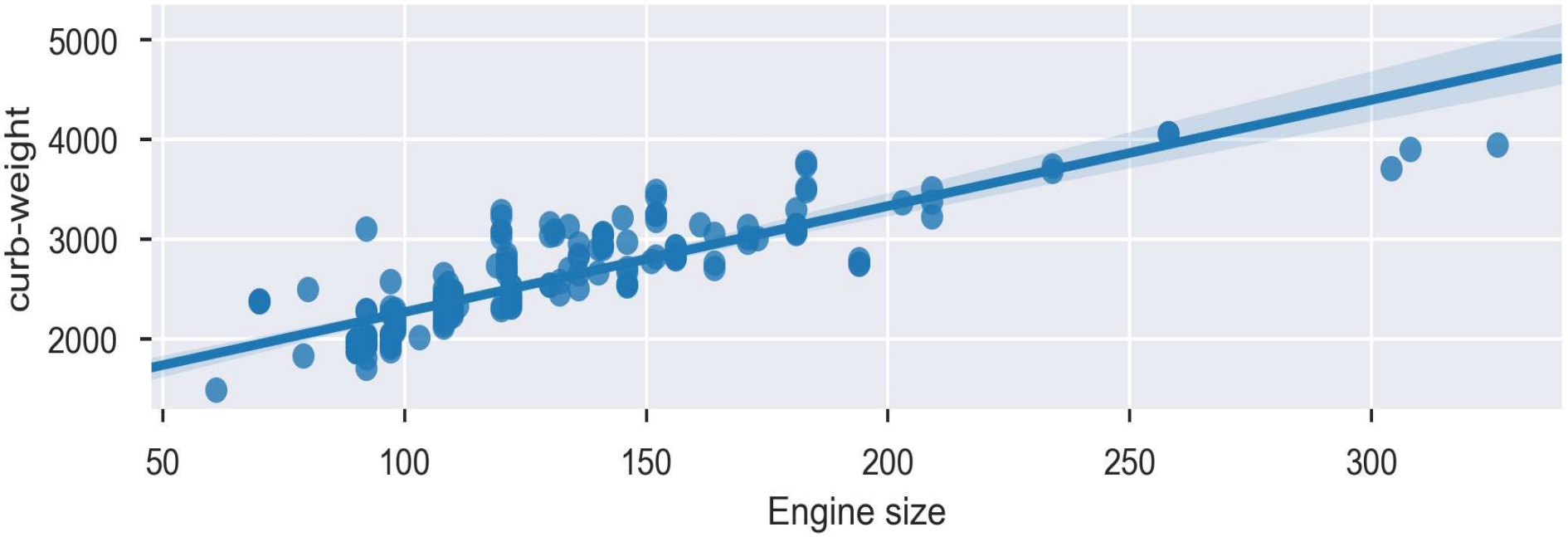

Price is increasing with the increase in engine size and horsepower but it is not the case with city mileage. The reason for this decrease in the mileage is that high price vehicles belongs to luxury segment there they are not bothered about the cost. Instead, they want speed and high performance.

# How does horsepower is related to no. of cylinders?



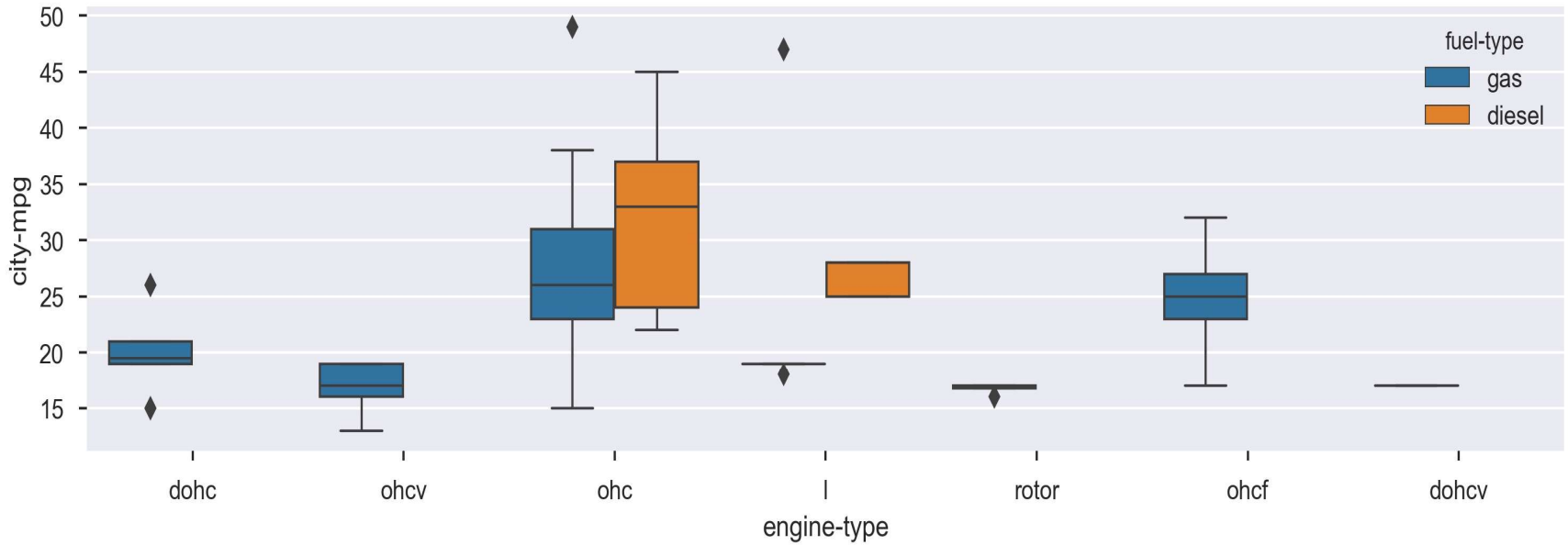

As the number of cylinders **increases**, horsepower also **increases**. We can see that the range of power output from 8 cylinder engine is very high. It seems, engines with higher number of cylinders can give a bigger range of power output. But no. of cylinders at 10 and 12 have very less number of samples. so, there it is not showing range of power output.

# What is the effect of curb-weight on engine size?



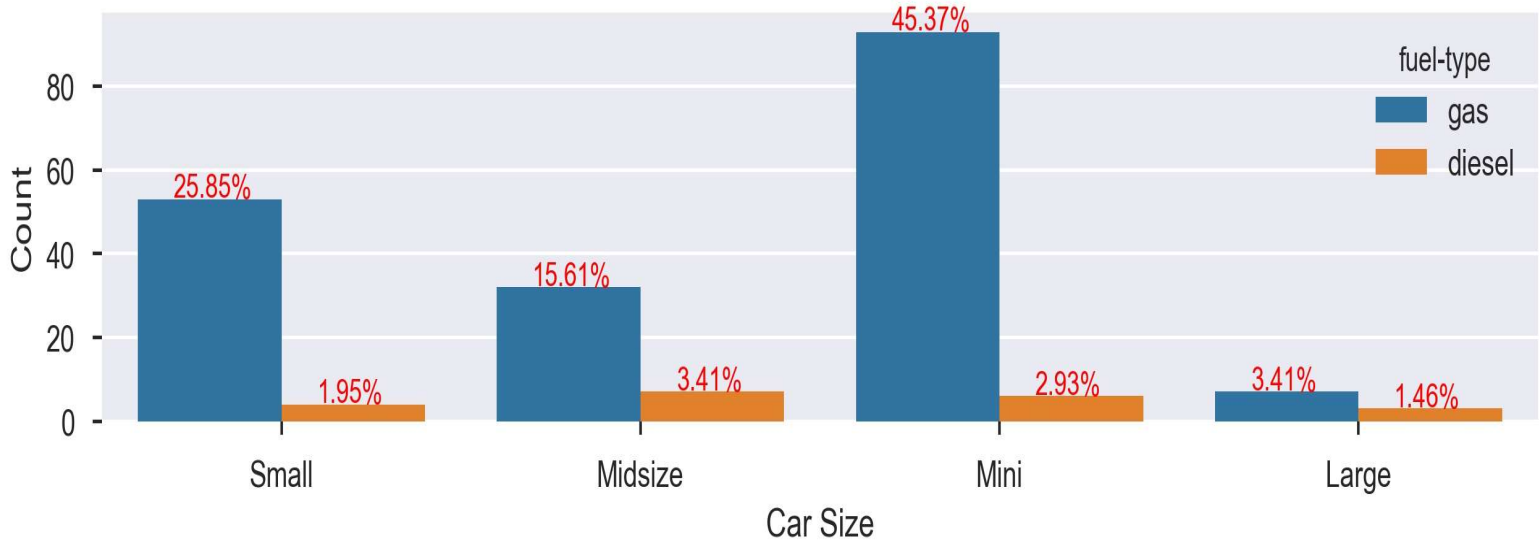

As the engine size is **increasing**, curb-weight is also **increasing** and both are **strongly** correlated.

# Which engine type is giving more mileage in the city and what is the type of the fuel it uses?




**Ohc** type engine is giving more mileage in the city and the fuel type is **diesel**.

# Classify the cars based on its body size




- Most of the cars belongs to **Mini** size and majority of them are using gas as fuel.
- **Midsize** segment vehicles has more number of diesel cars.

# Analysis based on efficiency parameters

## In this section, we'll see 2 main efficiency parameters

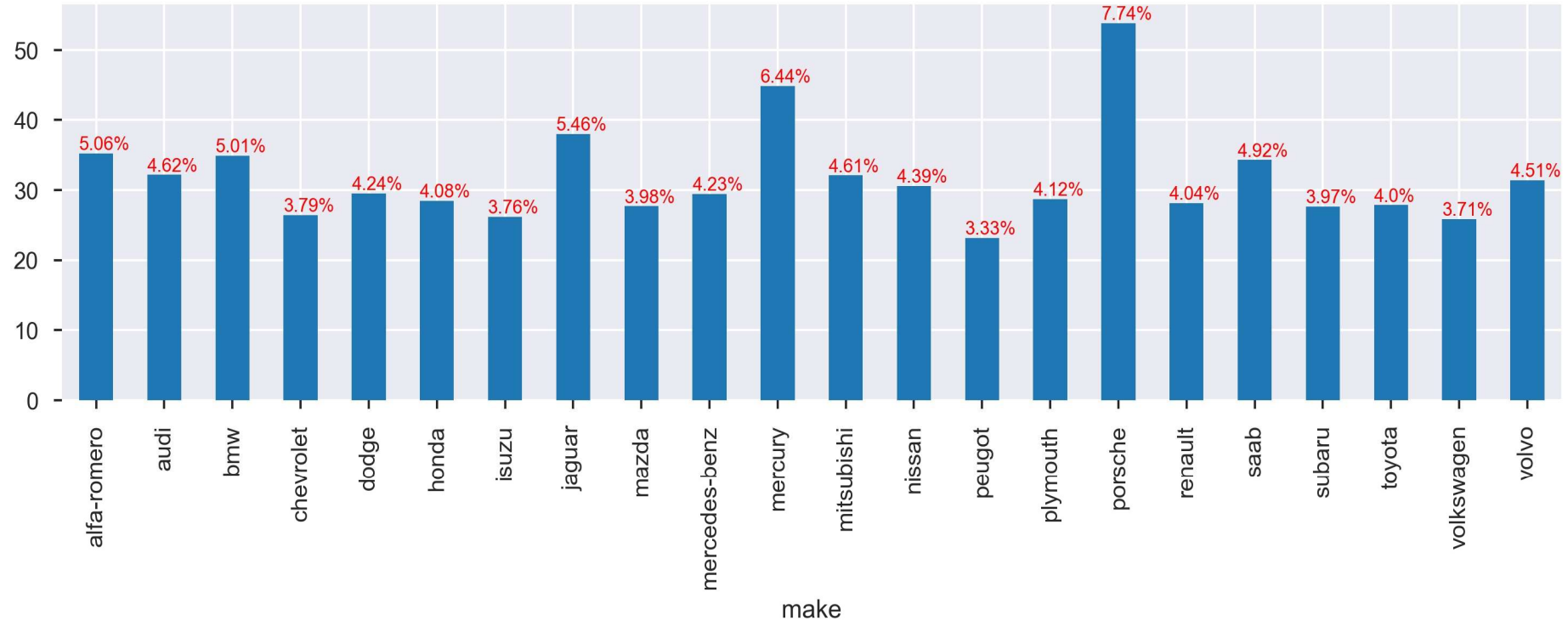
### Power-to-weight ratio

The higher the number, the better your car is going to be in terms of performance.

### Stroke-to-bore ratio

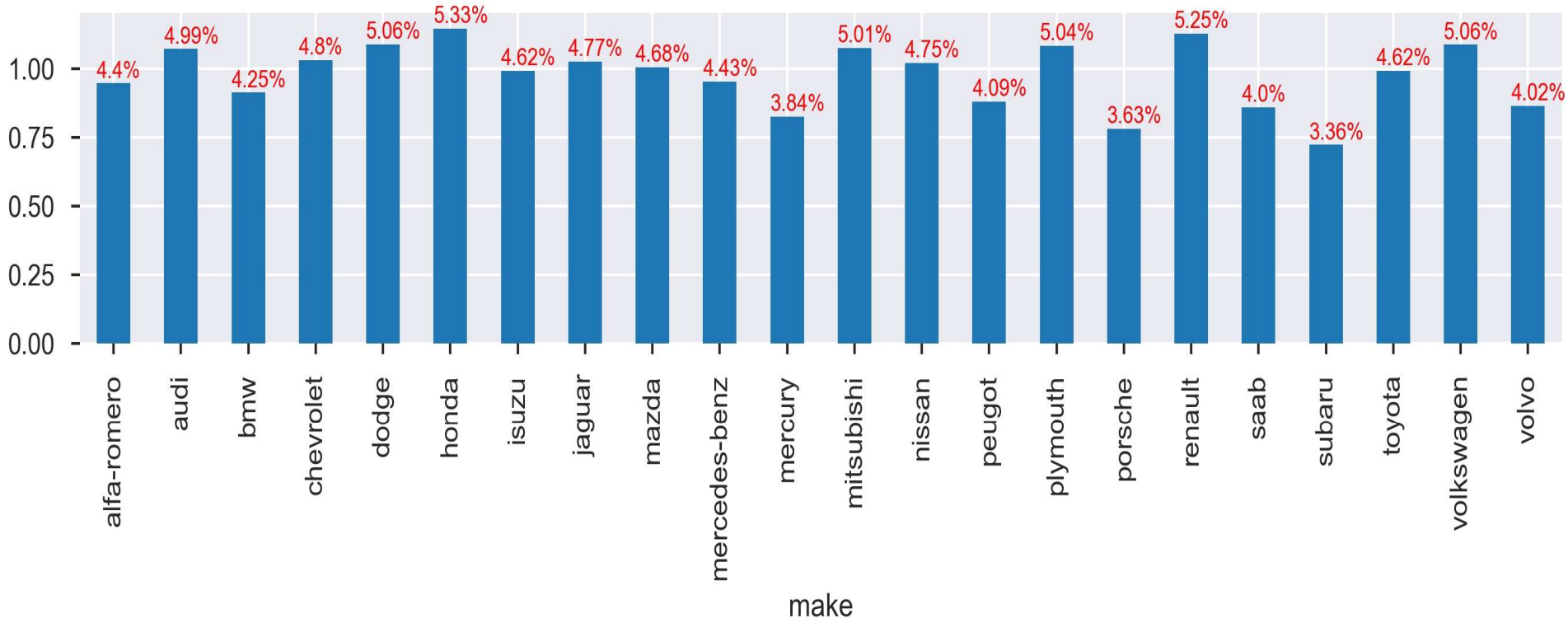
As the stroke-length is short, the piston has to travel a shorter distance. Hence, this design tends to produce higher engine speed and is typically used in high-speed cars & bikes.

# Power-to-weight ratio




**Porsche (7.74%)** is having better performance than the other vehicles as it is having maximum percentage of cars with high value of power-to-weight ratio. Least is **Peugeot (3.33%)**.

# Stroke-to-bore ratio




**Subaru (3.36%)** has the best performance in terms of stroke-to-bore ratio as it is having least percentage of stroke-to-bore ratio. Least is **Honda (5.33%)**.

# Risk Analysis

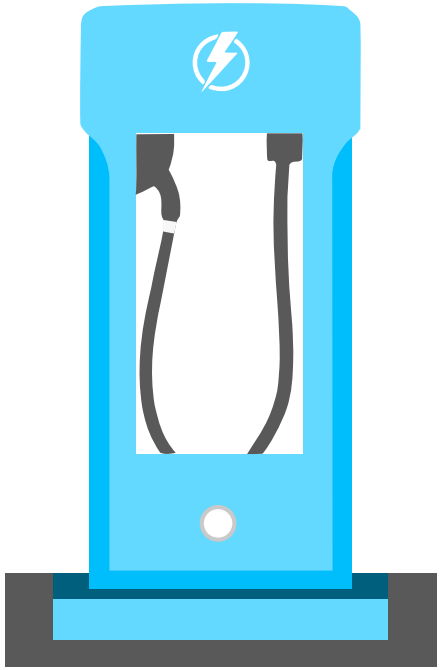
## About Symbolling

Symbolling value shows how risky or safe a vehicle is, from an insurer's perspective. It can range from -3 to +3.

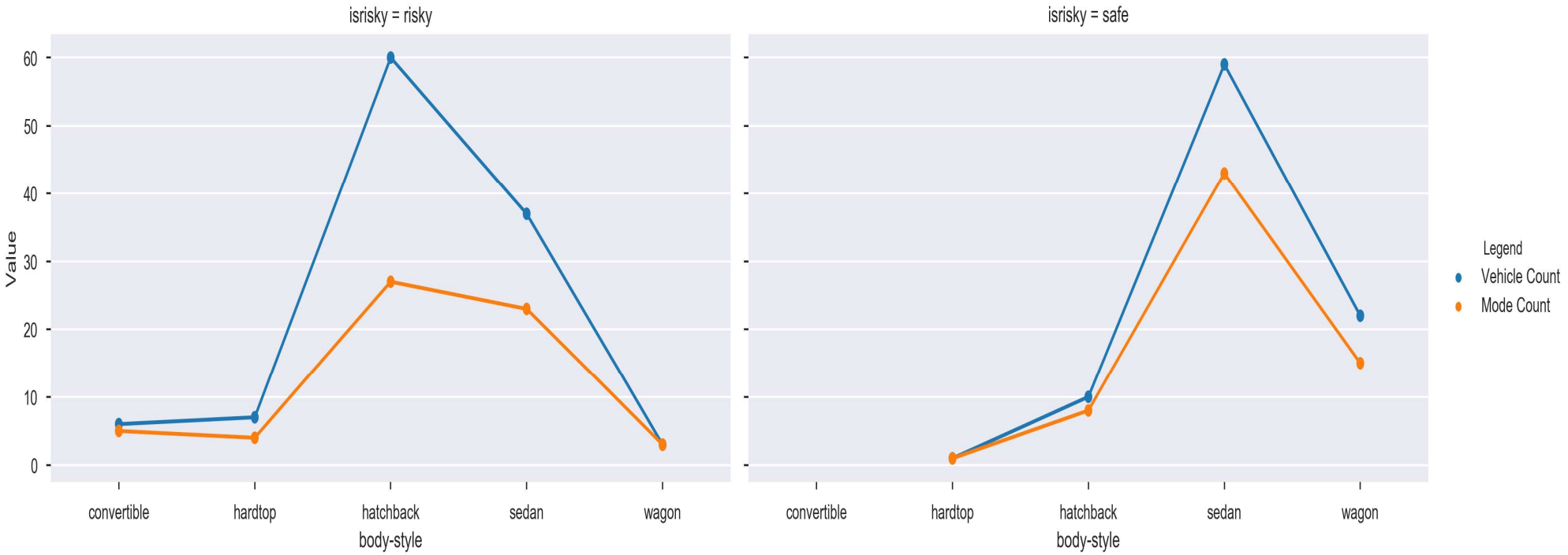
**-3** indicates a safe car while **+3** denotes a risky one.

## We'll examine the following things for our analysis

1. Which body styles are most risk prone?
2. Factors contributing to risk:
    a. Wheelbase
    b. Height
    c. Comparison of risk due to Wheel Base & Height.
    d. Number of doors
    e. Volume

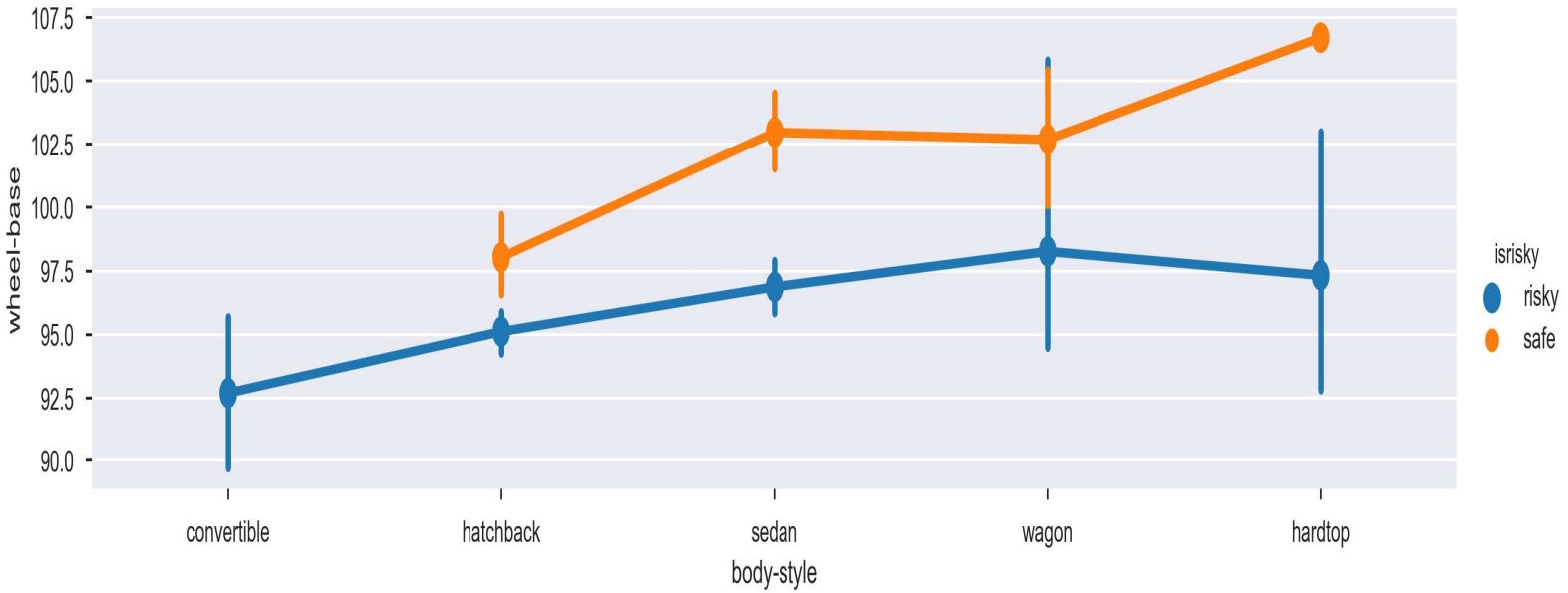


# Which body styles are most risk prone?



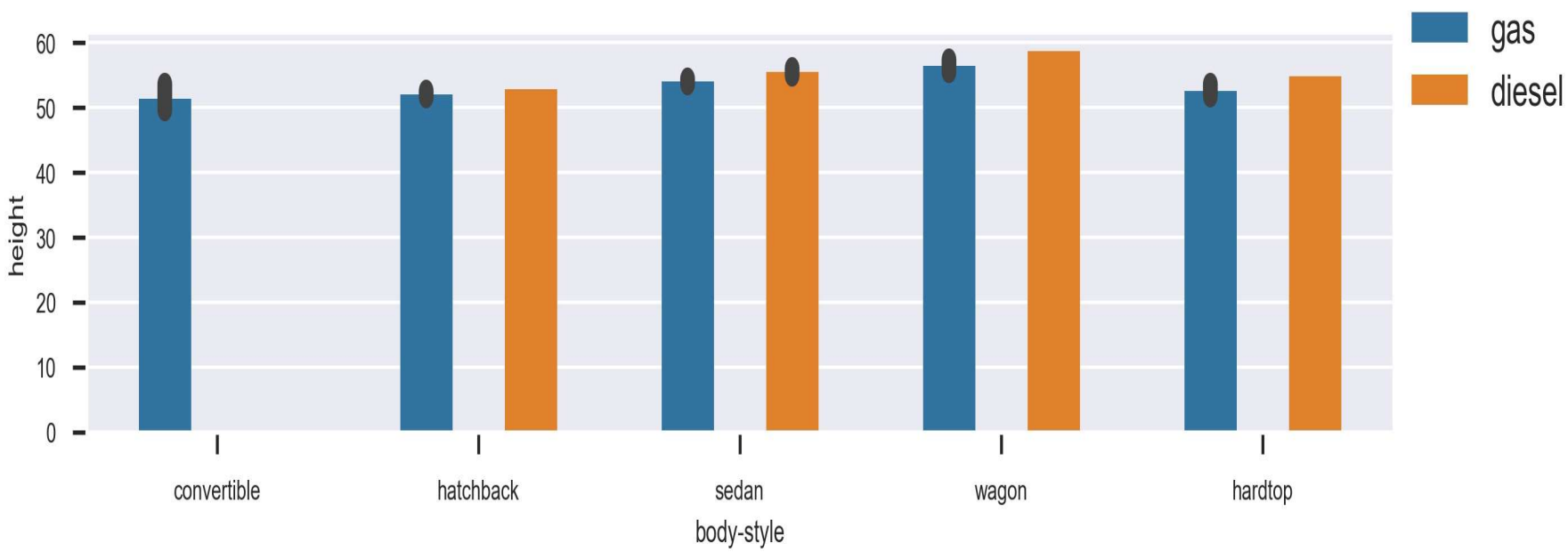

This clearly shows that convertibles and hardtops are riskier compared to sedans and wagons. Hatchbacks on the other hand does not show a clear trend with respect to risk.

# a. Wheel Base



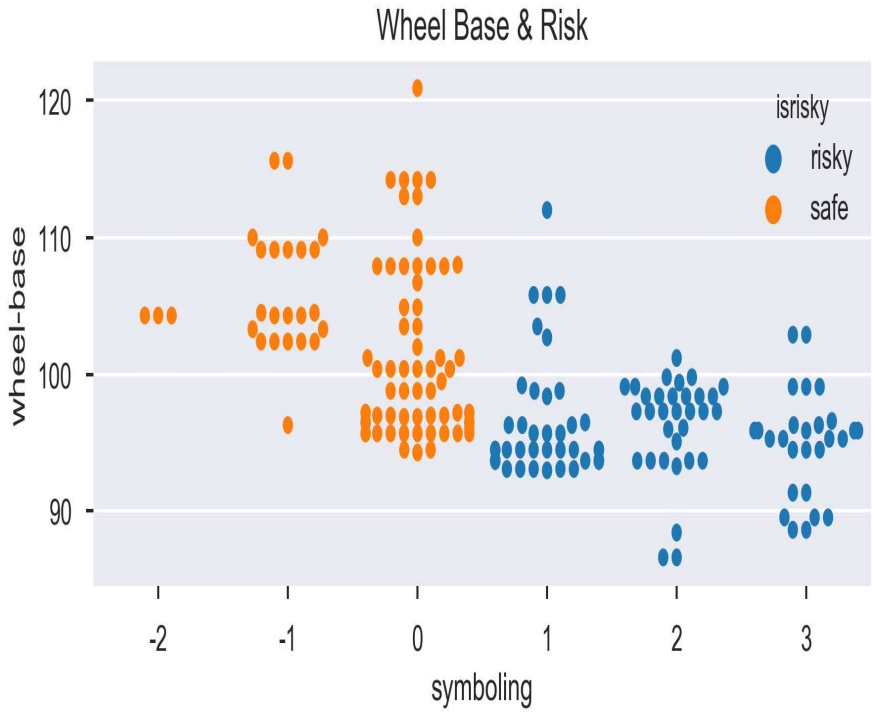

Within each body style, risky vehicles has lesser wheel bases compared to riskier ones. Conertibles and hardtops has significantly less wheel base compared to other body styles. So, it is safe to purchase vehicles with longer wheel bases.

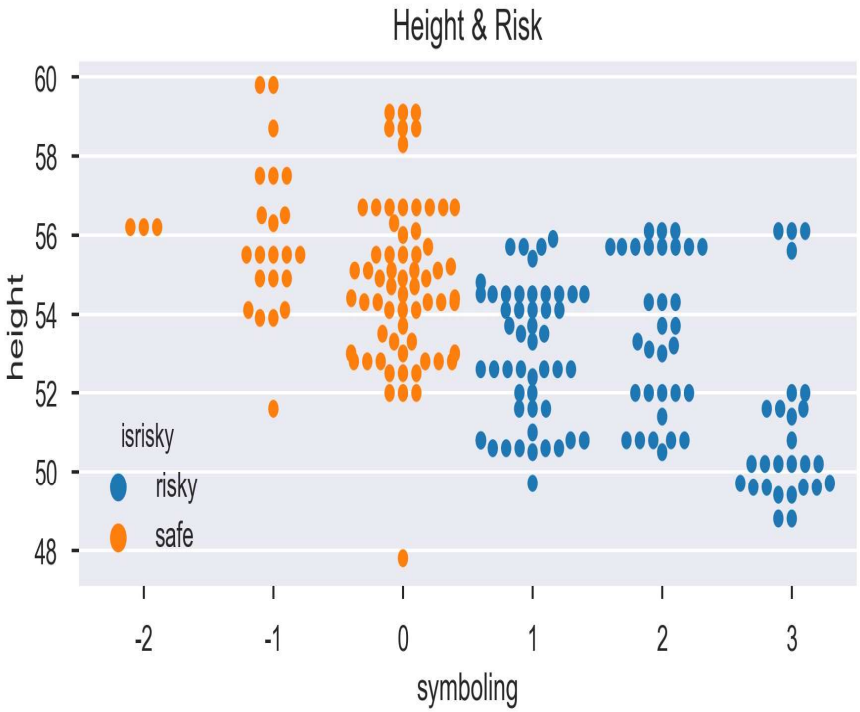
# b. Height




As the height of the vehicle reduces, ingress and egress becomes increasingly difficult for passengers. Here wagons have the maximum height for both the fuel types.

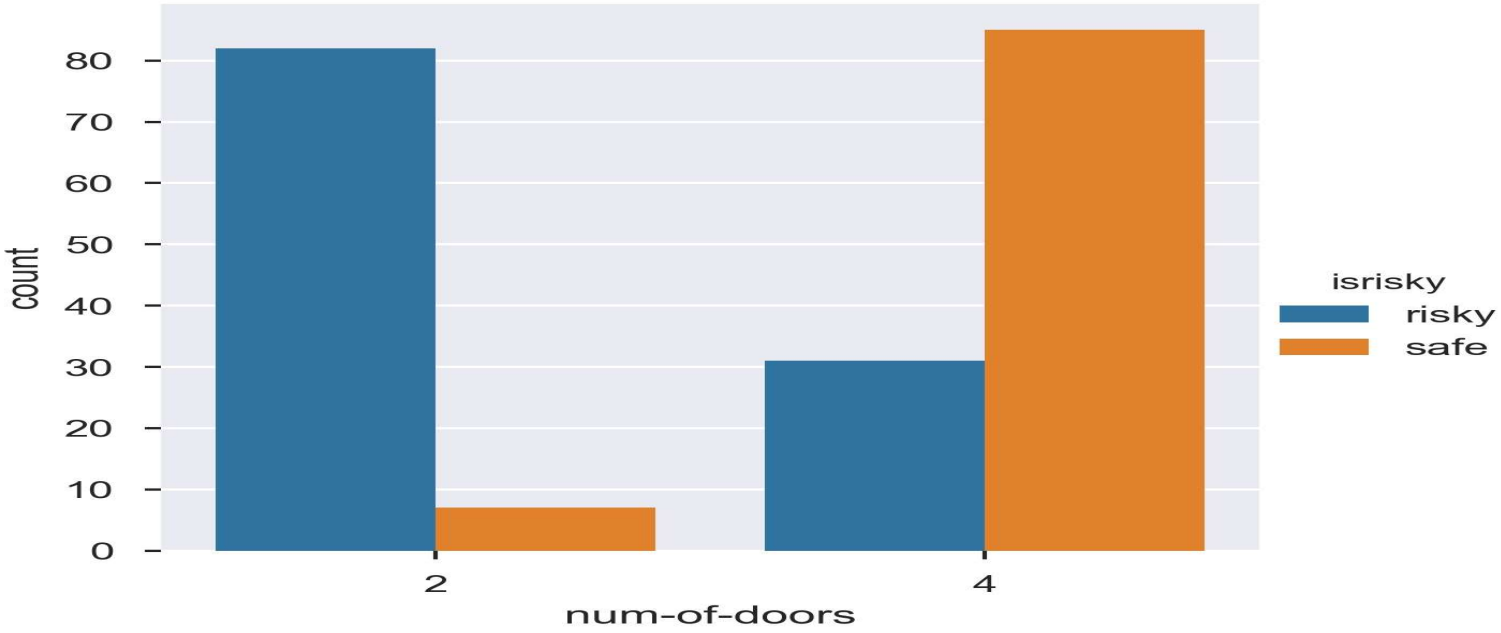
# c. comparison of risk due to Wheel Base & Height



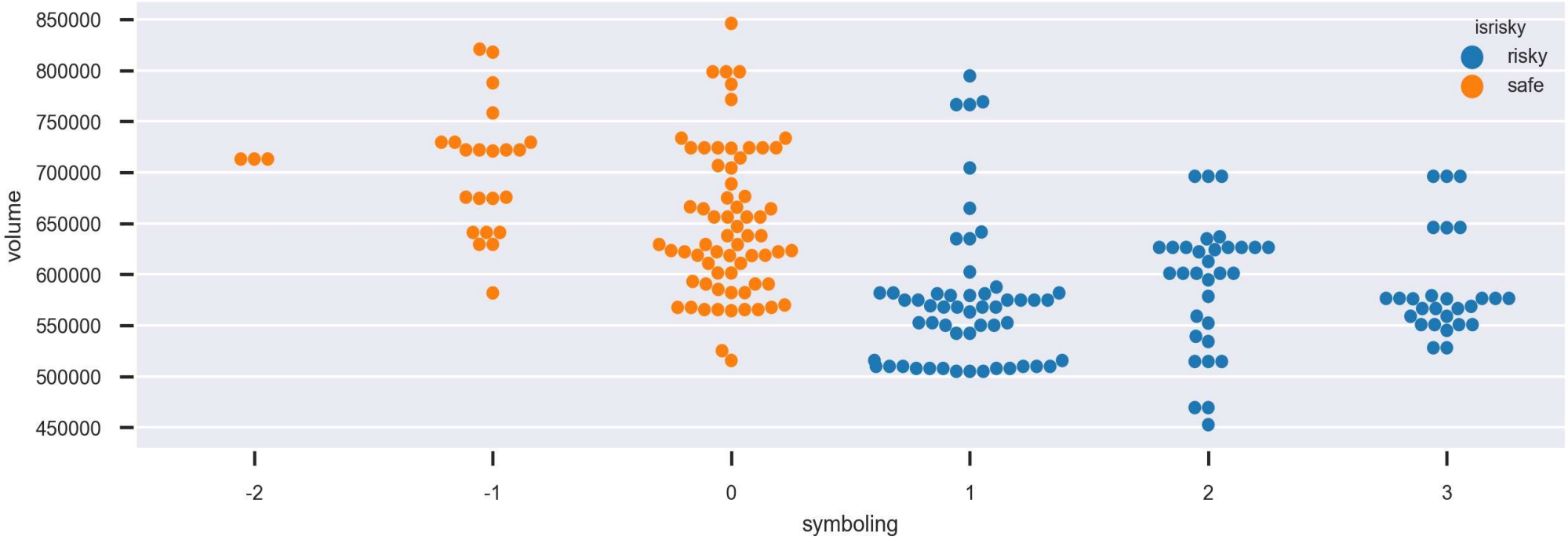

•Safe vehicles within each bodystyle are having more length than riskier ones. They have longer wheel bases.
•Reduction in vehicle height reduces the center of graity of the car and hence improved stability especially while doing high speed cornerings. So, when a manufacturer reduces the height of a car, they are clearly aiming the product for car enthusiasts and not for regular commuters.

# d. Number of doors



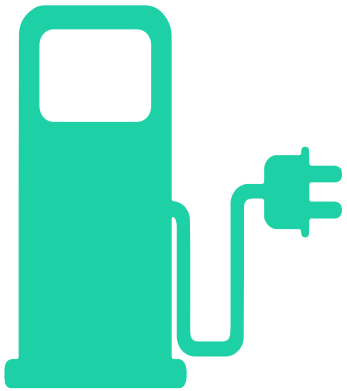

**Number of doors does not determine the risk factor of a car directly**. Having 4 doors does not ensure a place in the safe category, but then, we can see that safer cars are always found among 4 door category.

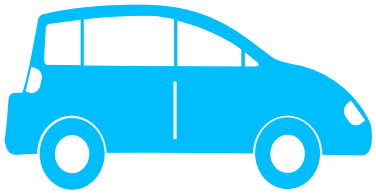
# e. Volume




As the vehicle volume reduces, symbolling values increases, indicating an increase in risk.

# Correlation Analysis





It is a measure of the extent of interdependence between variables. In other words, when we look at two variables over time, if one variable changes, how does this effect change in the other variable?

Here we used the following 2 graphical representations for our analysis :

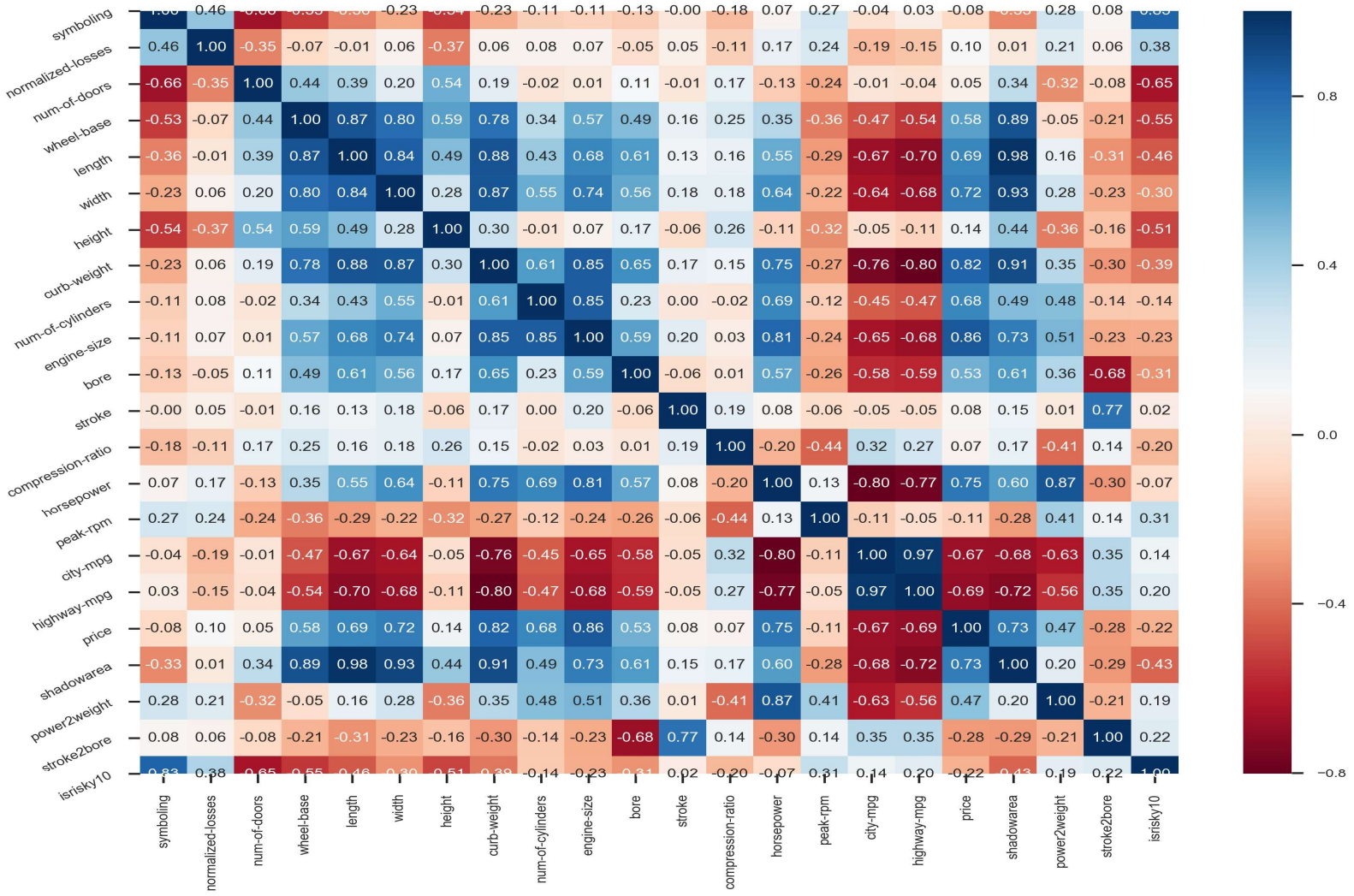
**1** **Heat Map**

It is a two-dimensional representation of data in which values are represented by colors. We use color to communicate relationships between data values that would be much harder to understand if presented numerically in a spreadsheet.

**2** **Pair plot**

It plots a pairwise relationships in a dataset. The pair plot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.

# Heat Map



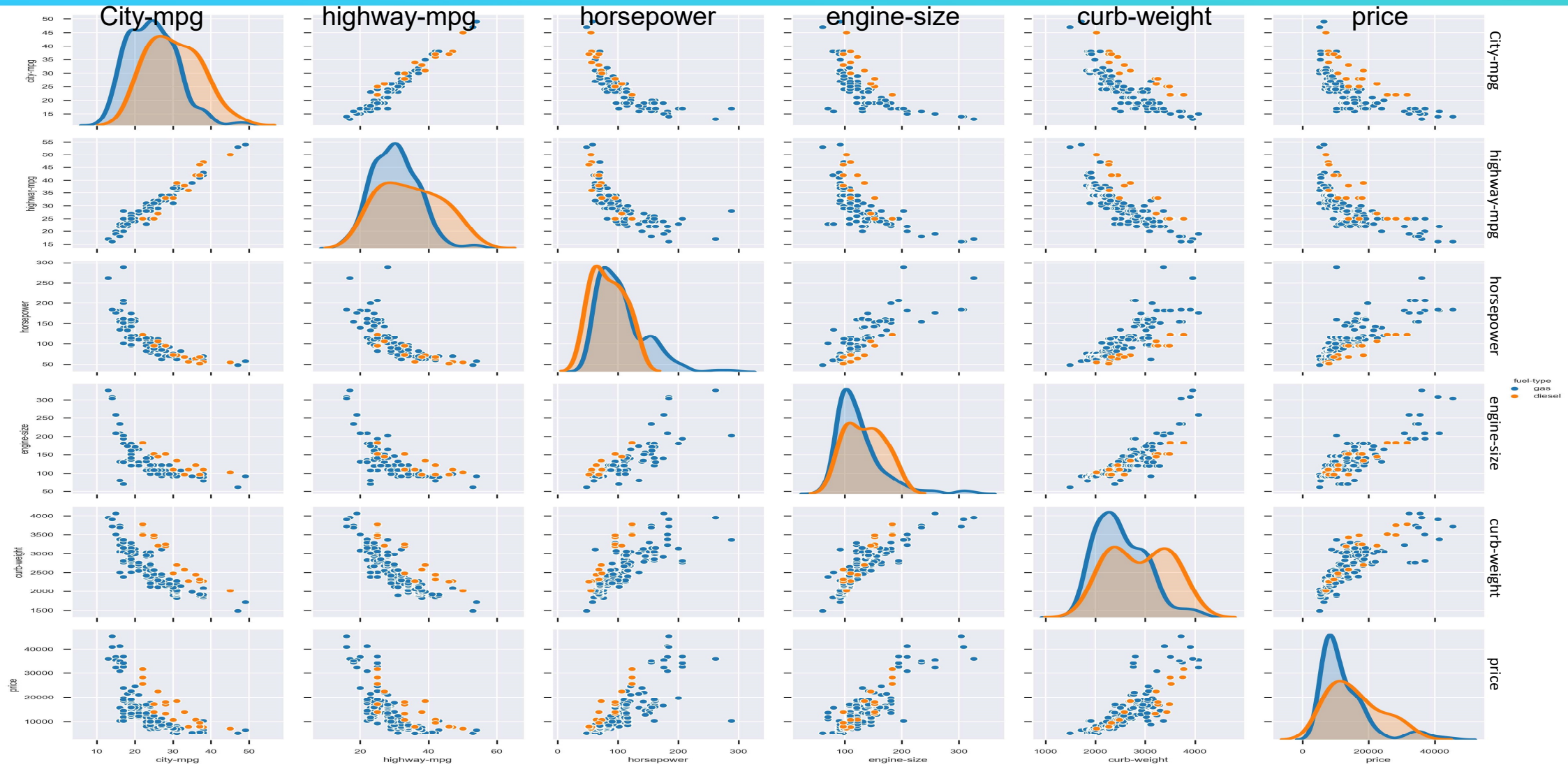

Price is more correlated with engine size and curb weight of the car.

Curb weight is mostly correlated with engine size, length, width and wheel-base which is expected as these adds up the weight of the car.

Wheel base is highly correlated with length and curb-weight of the car.

Symbolling and normalized car are correlated than the other fields.

# Pair plot


City-mpg
highway-mpg
horsepower
engine-size
curb-weight
price
fuel-type
gas
diesel

# Pair plot observations

Vehicle **Mileage** is **inversely proportional to Horsepower, engine-size and Curb Weight.**

Horsepower increases, with the increase in engine size.

There is a **strong** correlation exists between **curb weight and engine size** as Curb weight increases with the increase in **Engine Size** or **horse power**.

Vehicle with **high price have low mileage**. This because high priced vehicles go into luxury segment which are meant for high performance and running cost is not very important in this segment.

As the engine power(horse power) increases, the vehicle price also **increases**. More horse power also means bigger engine size.

High **curb weight increases price** of the vehicle and **decreases** the **mileage** of the vehicle.

# Summary

| Price | |
|---|---|
| Majority of cars | belongs to the lower price brackets (< 20K) even though there are cars that go up to 45K |
| High correlation with | Engine-size, wheel base, curb weight, horsepower |
| Expensive for | Diesel type, mpfi engine<br>Mercedes Benz<br>Rear wheel drive,<br>Convertibles and hardtop body |
| - ve correlation with | Mileage |

| Normalized losses | |
|---|---|
| highest | BMW |
| Directly proportional to | risk rating |
| more losses in | convertible car and hardtop, hatchback<br>sedan and Wagon (even with less risk rating),<br>2 door cars |

| Fuel | |
|---|---|
| preferred | gas |
| aspiration type | standard |
| More mileage | Diesel in highway,<br>Forward drive wheels,<br>OHC type engine, |
| Best fuel economy | Chevrolet |
| Inversely proportional to | Horsepower, engine-size and Curb Weight. |

| Risk Rating | |
|---|---|
| Risky | convertibles and hardtops,<br>less wheel base, Less height, Less volume, 2-door vehicles, |

| Efficiency | | |
|---|---|---|
| Power-to-weight | Best | Porsche |
| | Least | Peugeot |
| Stroke-to-bore | Best | Subaru |
| | Least | Honda |

# Conclusion

- It is analyzed that the attributes in automobile dataset are categorized as basic, engine, dimension and fuel & efficiency and based on this different chart are plotted.
- The most important inference drawn from all this analysis is, I get to know what are the features on which price is highly positively and negatively correlated with.
- This analysis will help me to choose which machine learning model we can apply to predict price of test dataset in later terms and projects.
- Now have to move into building machine learning models to automate our analysis, feeding the model with variables that meaningfully affect our target variable will improve our model's prediction performance. We just end up here with the basic pre-processing and data analysis.


THANK YOU