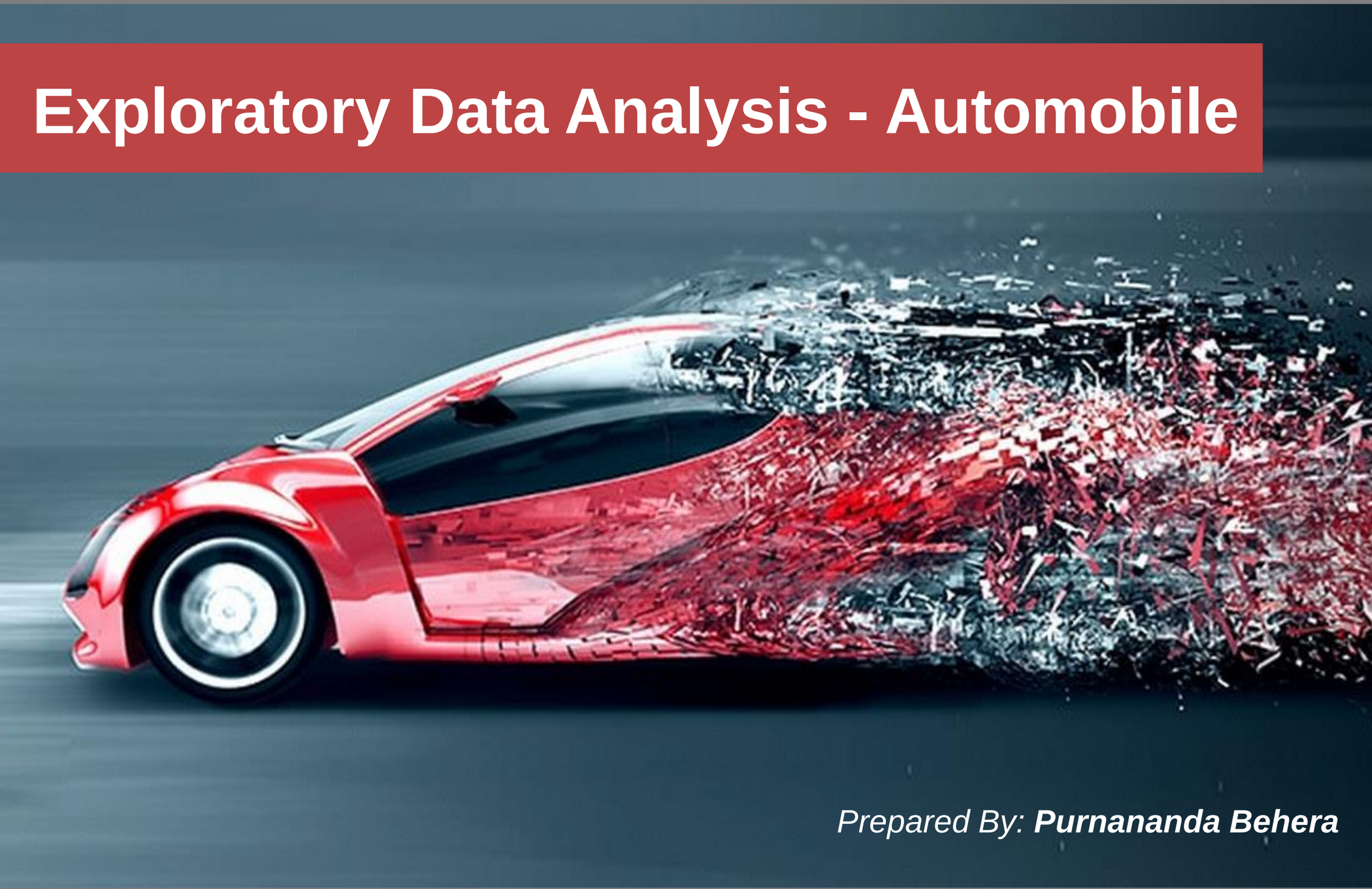
Exploratory Data Analysis - Automobile



Prepared By: ***Purnananda Behera***

EDA: *Automobile Dataset*

About the Data

Contents: Insurance risk rating, normalized losses, engine, body style, mileage, price for each car model.
Source: https://archive.ics.uci.edu/ml/datasets/automobile
Data Volume: 205 records, 26 columns

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**:** eight, five, four, six, three, twelve, two.
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.

Data Clean-up

Variable Identification

- Symbolling, Price, Mileage, and other necessary variables to support initial hypothesis

Univariate Analysis

- Explored the data for individual column. Refer notebook.

Bi-variate Analysis

- Explored the data for multiple column. Refer notebook.

Cleaning missing values

- Replaced missing values of non numeric fields with categorical mean value.

Detecting, analysing and treating outliers

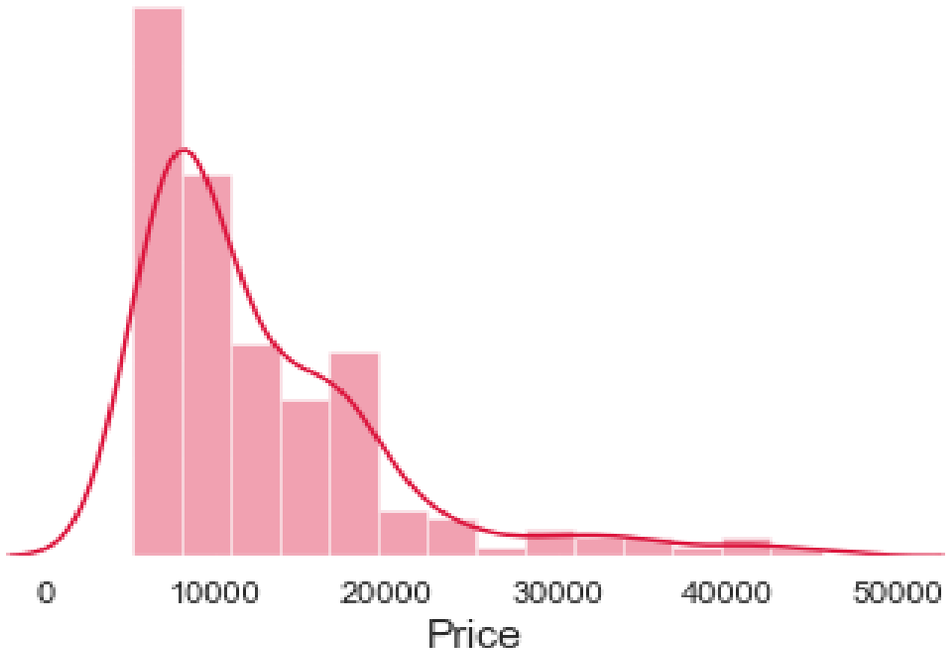
- Engine size and horsepower outliers are kept because they represent real world data.

Deriving variables

- Introduced few new variables like avg-mpg, isrisky, price-range, etc.

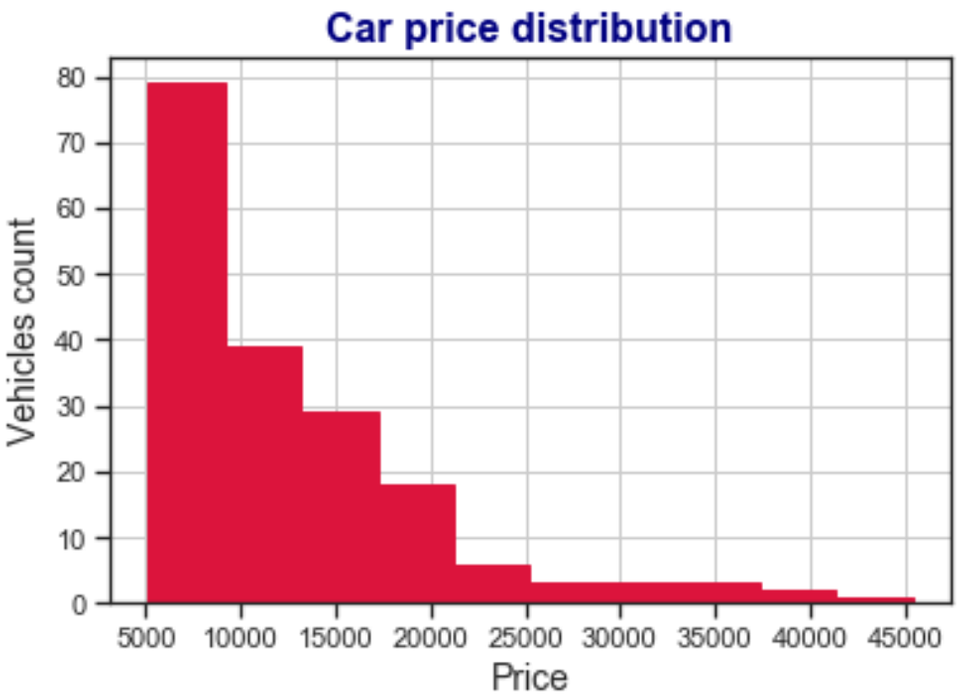
Problem 1

Do the **Body Size, Style and Engine Specification** determine the car price?



Our study focus on the relationship between car prices with body size, style and engine specification determine the car prices?

Car Price - Distribution

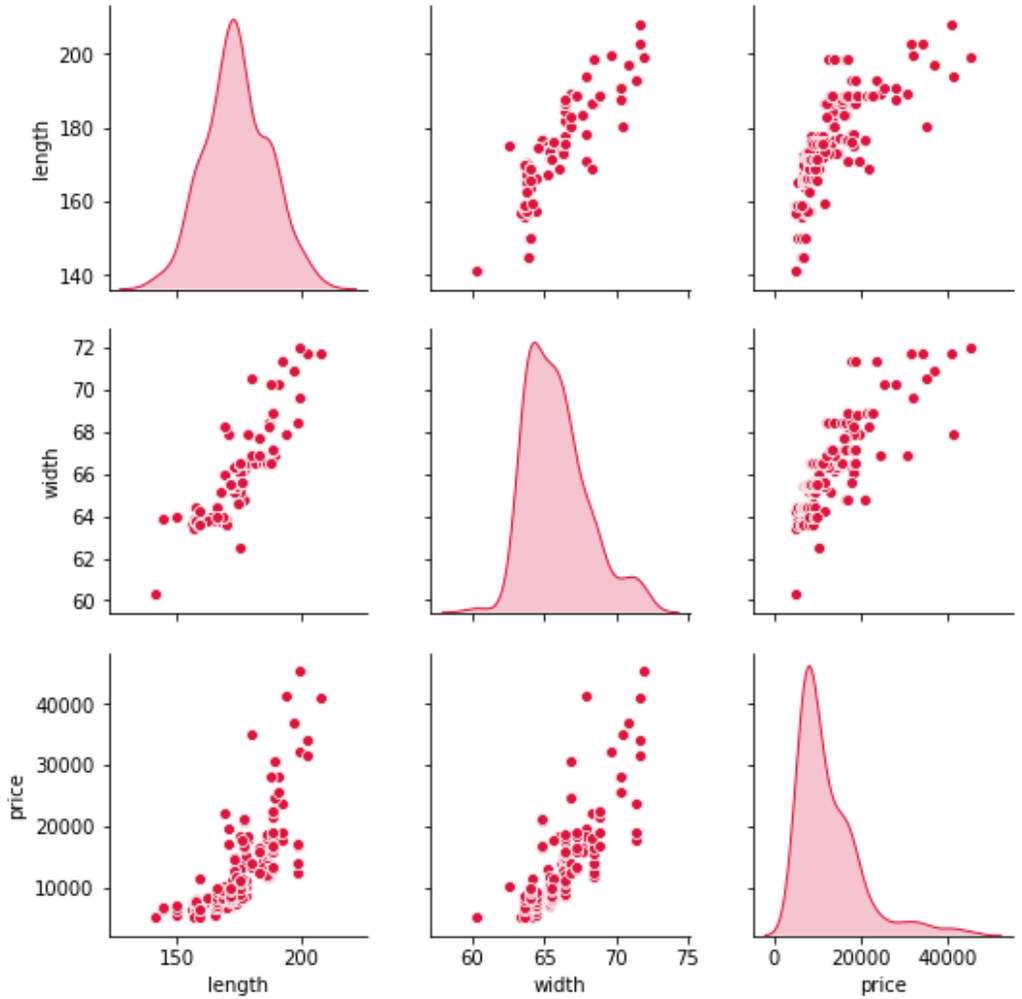


Distribution of Car Prices

- Majority of cars belongs to lower price brackets (<$20K).
- Car prices range in between $5K to $45K.
- The car prices skewed to the right (right-skewed): The mean and median are greater than the mode.

Note: *Price mentions here is in $ dollar.*

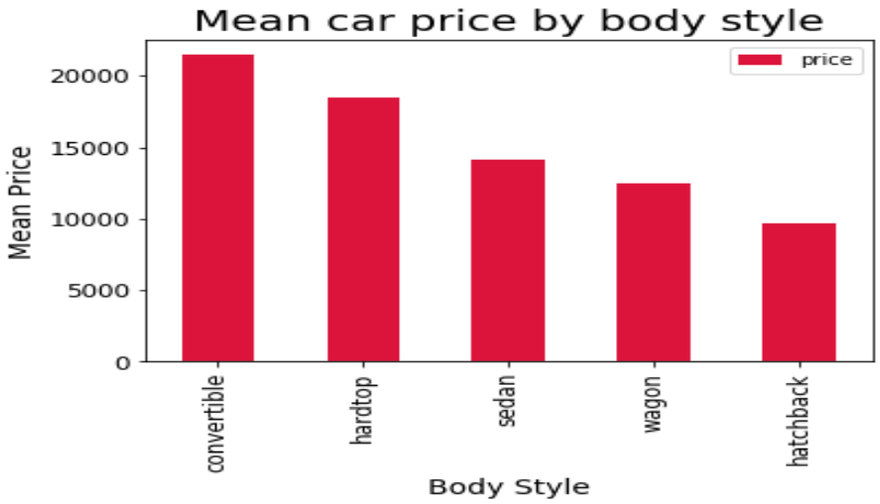
Body Size (Length, Width) vs Car Price

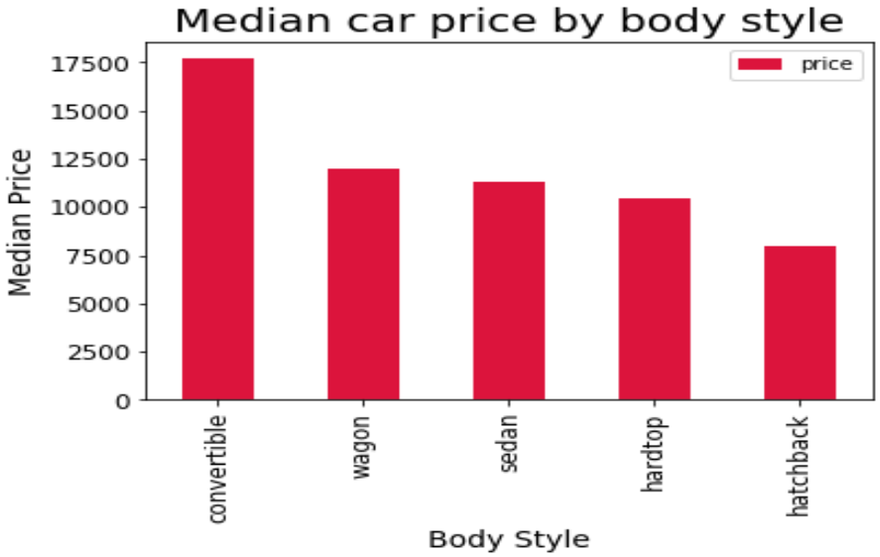


Body size (length & width) has strong correlations with price.

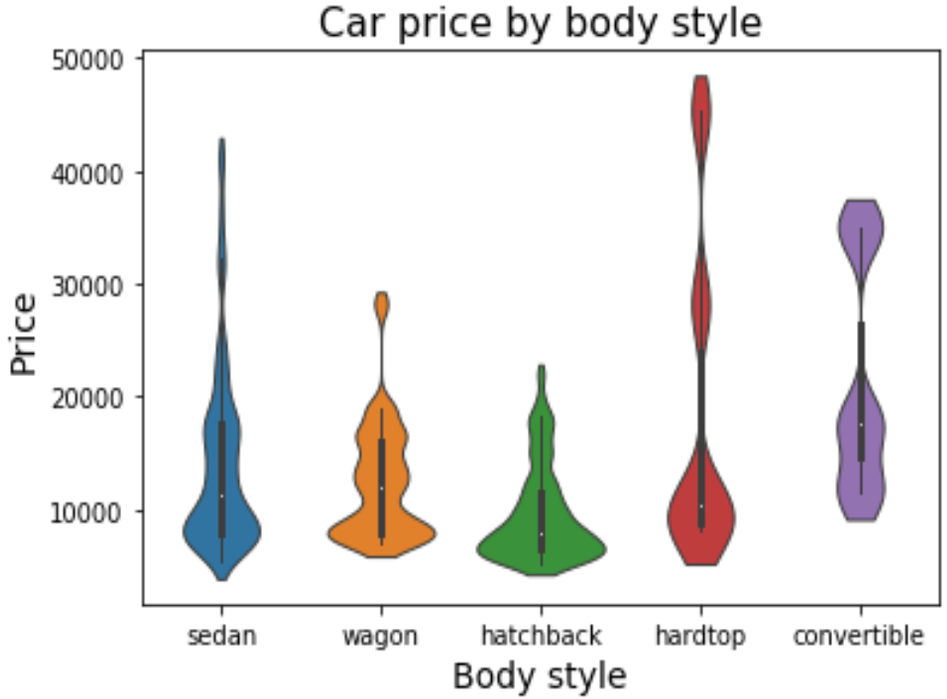
- Whenever there is increase in length or width of cars that escalate the car price irrespective of body style.

Body Styles vs Car Price



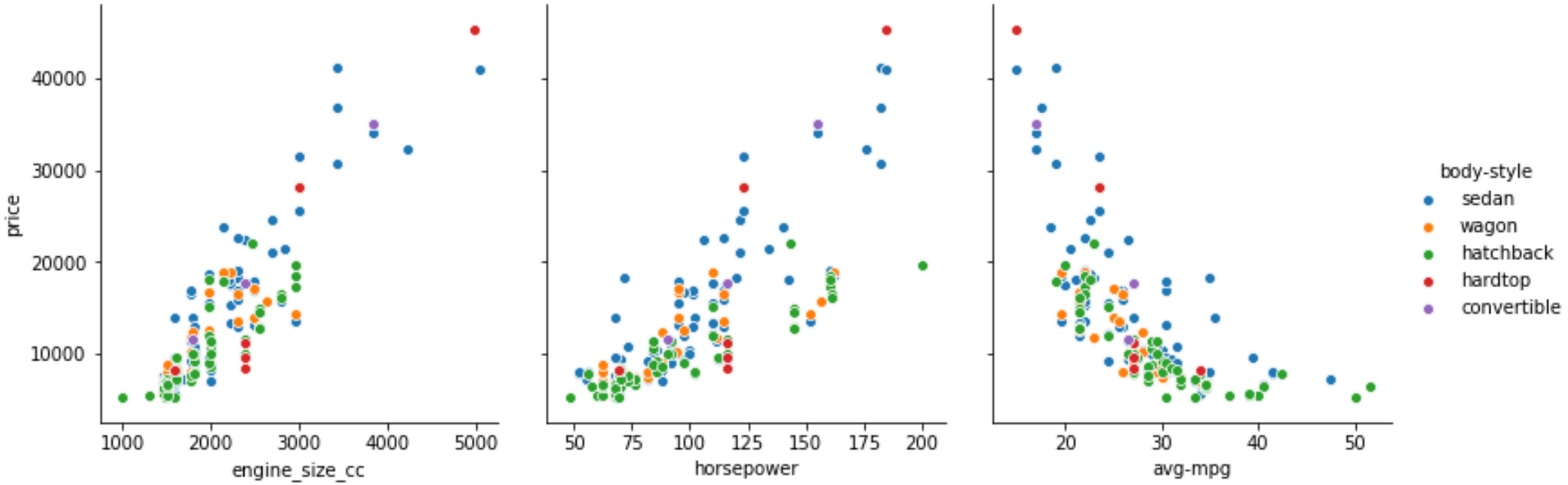


Convertibles and hardtops cars are the costliest car models.



Engine Specs vs Car Price

Engine size (CC) & horse power has strong correlations with price, where as mileage ((city-mpg + highway-mpg)/2) is negatively correlated with price.



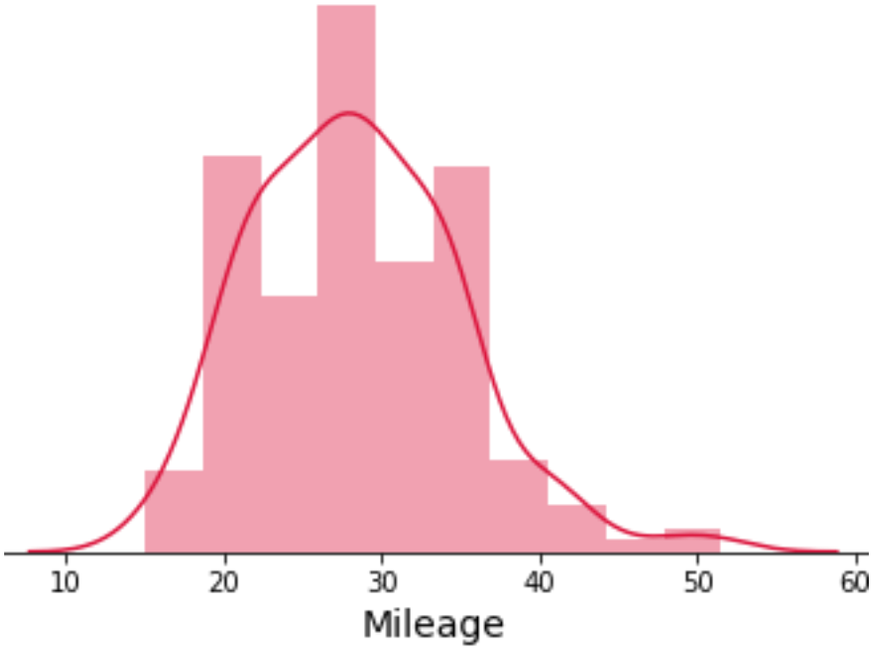
Problem 1: Summary

Do the **Body Size, Style and Engine Specification** determine the car price?

- Bigger vehicles are priced above the smaller ones.
- Convertibles and hardtops cars are priced above the hatchback, sedan and wagon.
- Cars having bigger engines and more horse power are priced higher.
- Also less mileage cars are priced higher, and may be when body size and power get increase which impact to mileage and it get declined.

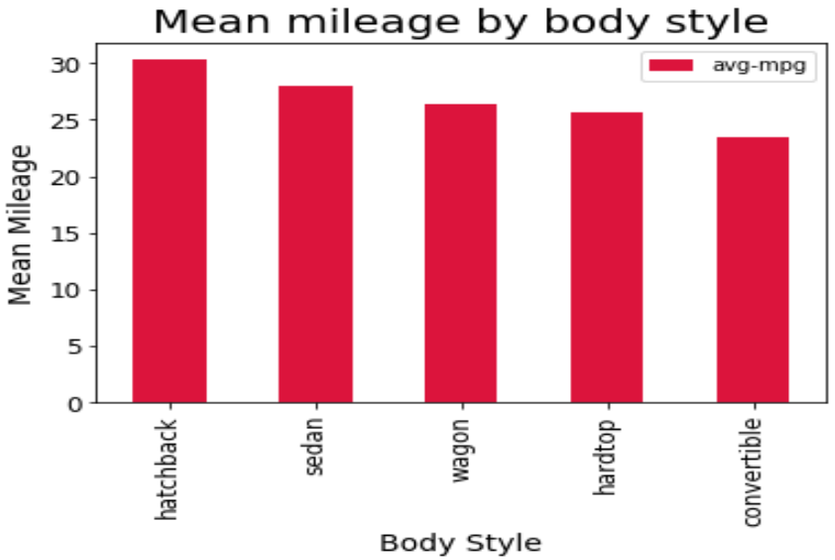
Problem 2

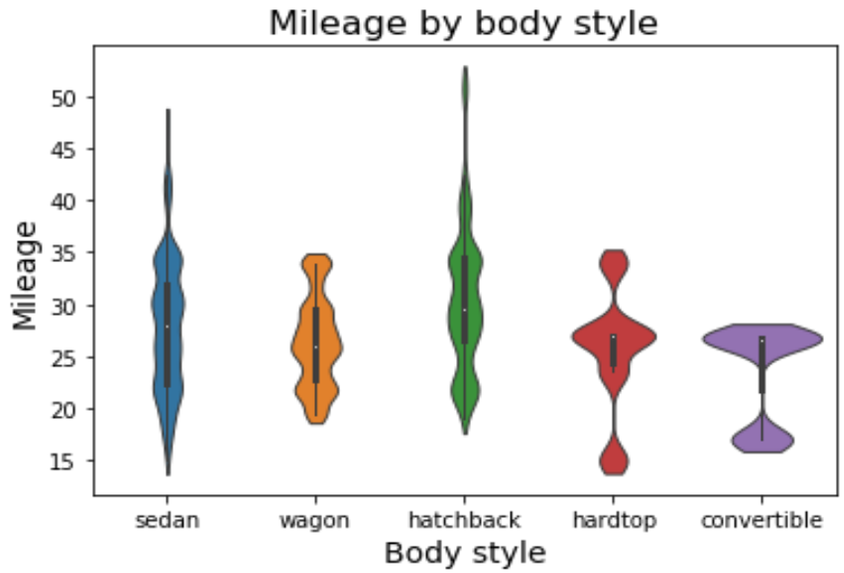
Which type of cars are better in terms of **mileage**?



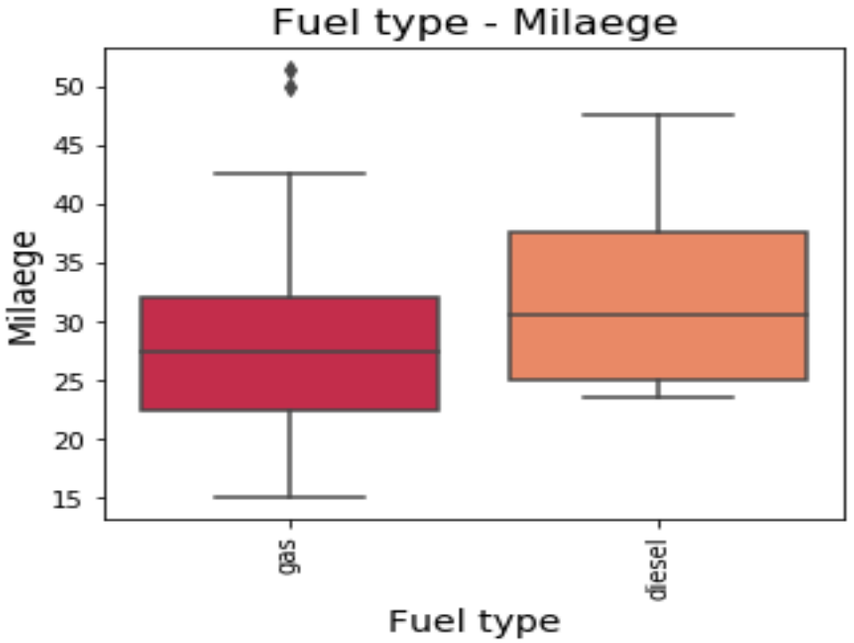
Our study focus on the relationship between mileage with body style, fuel type and engine specs to determine the mileage efficient car?

Body styles and Fuel type vs Mileage

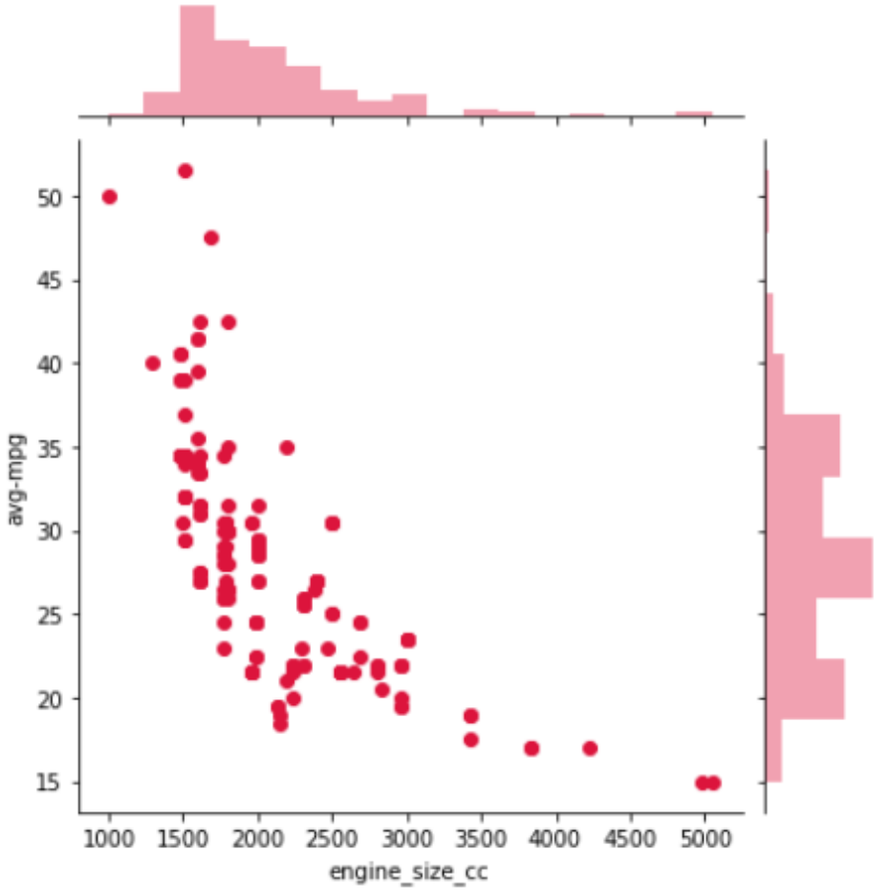


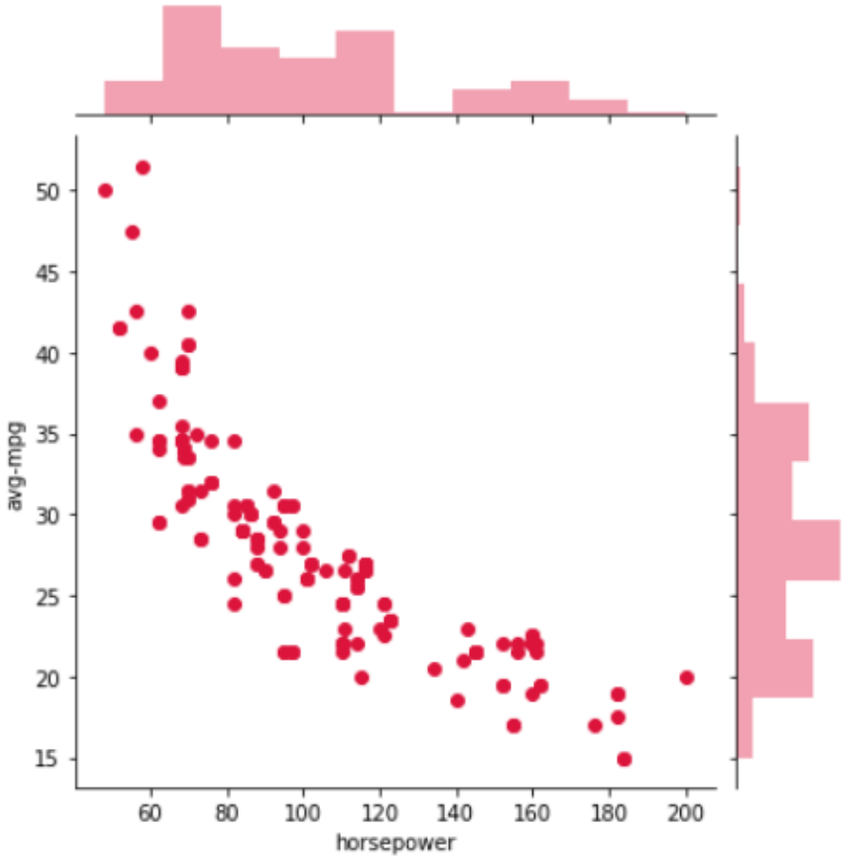


Hatchback cars has the highest mileage followed by sedan class. Also Diesel based cars gives more mileage then the Gas based cars.



Engine specs vs Mileage





Engine specs (Engine size & Horsepower) has negative correlations with mileage.

- Whenever engine size and horsepower of cars get increase that impact the mileage of the cars and it get declined irrespective of body style.

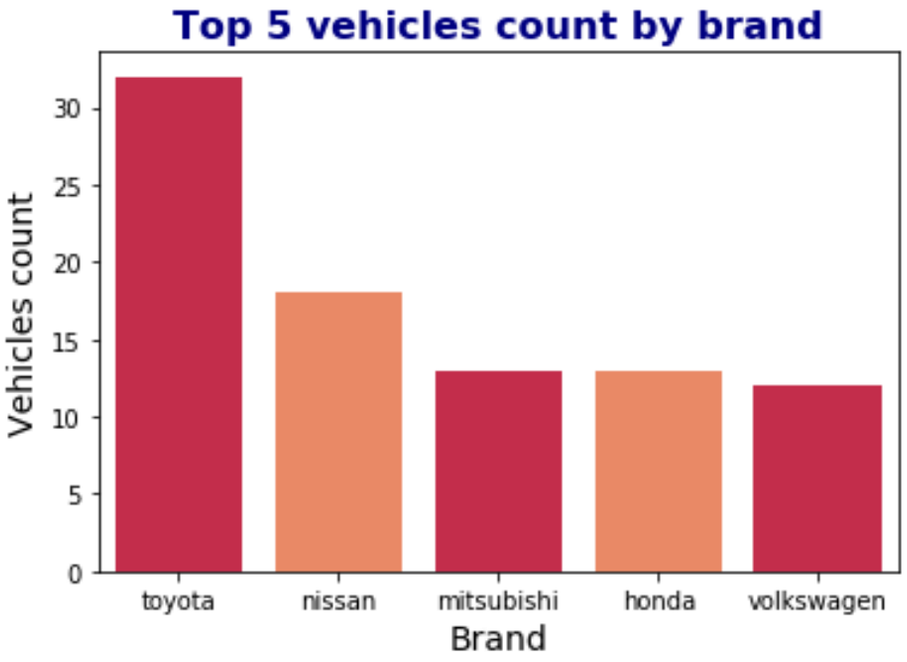
Problem 2: Summary

Which type of cars are better in terms of **mileage**?

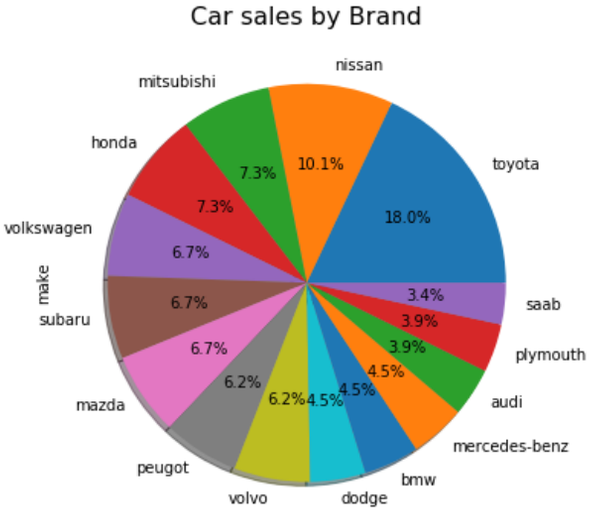
- Hatchback and sedan cars produce better mileage.
- Cars having smaller engines and less horse power produce better mileage.
- Diesel cars produce better mileage compare to cars with Gas fuel type.

Problem 3

Which are highest selling cars based on **brand**, **body style** and **price slab**?

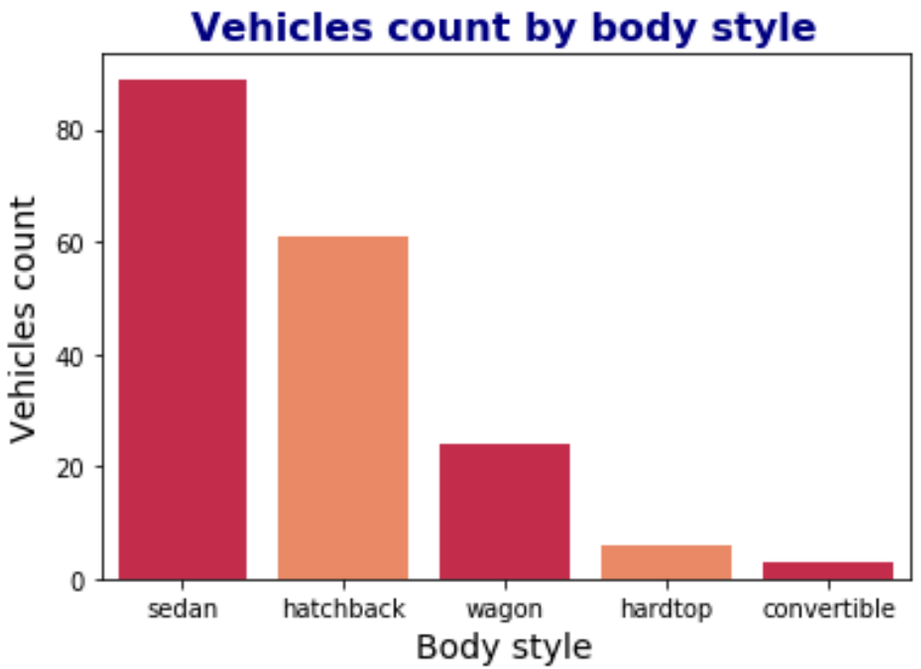


Toyota lead the list with highest number of cars selling followed by Nissan.

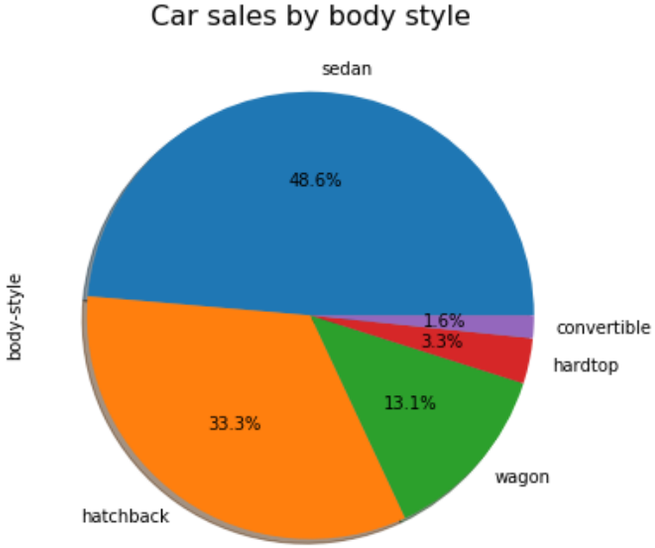


Our study focus on extracting popular cars through various fields.

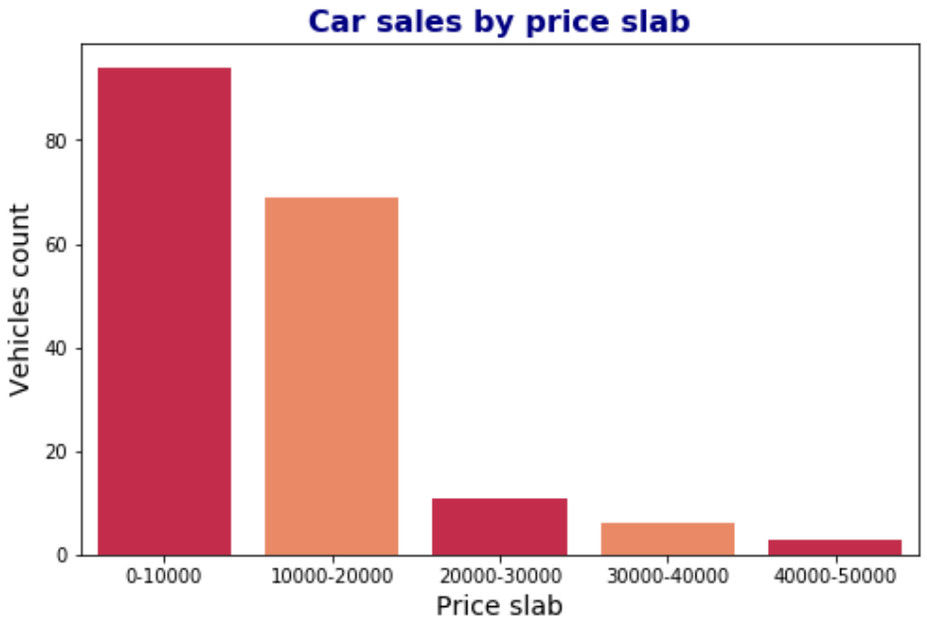
Car sales with body style



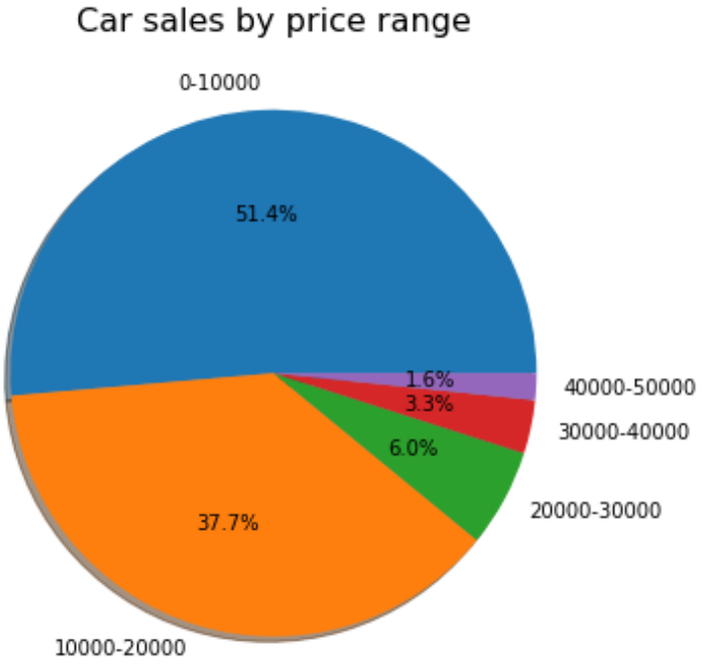
Sedan segment lead the list with highest number of cars selling followed by hatchback followed by body style.



Car sales with price range



Cars under $10k lead the list with highest number of selling followed by Cars under $10k to $20k.



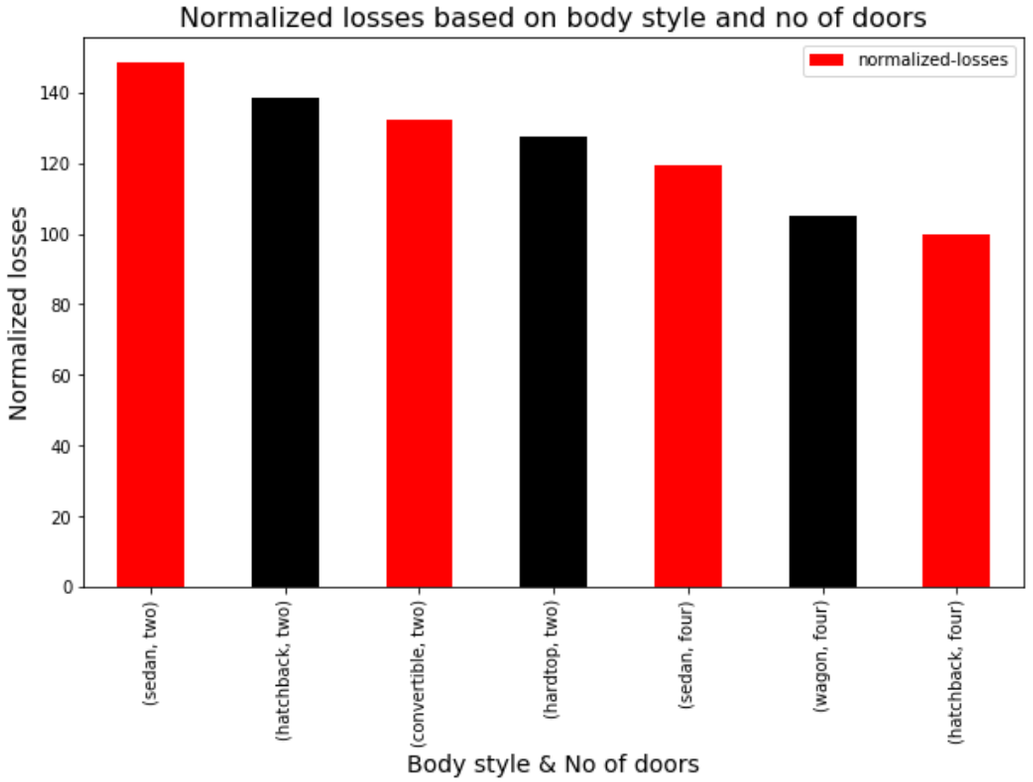
Problem 3: Summary

Which are highest selling cars based on **brand**, **body style** and **price slab**?

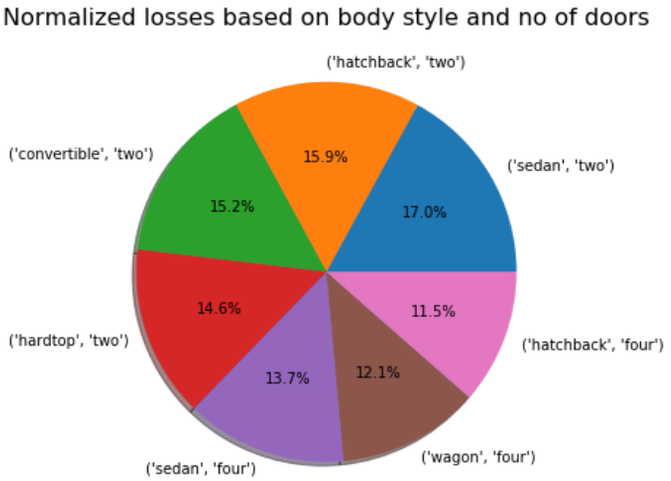
- **Toyota** sells highest number of cars followed by **Nissan** in brand category.
- **Sedan** segment sells highest number of cars with **48%** followed by hatchback.
- Cars **under $10k** lead the list with highest number of selling followed by Cars in between **$10k to $20k** on price range basis.

Problem 4

Which are highest normalized loss reported cars based on **body style** and **no of doors**?



Sedan with two doors cars reported highest average losses followed by hatchback two doors.



Our study focus on normalized losses reported by cars based on body style and no of doors.

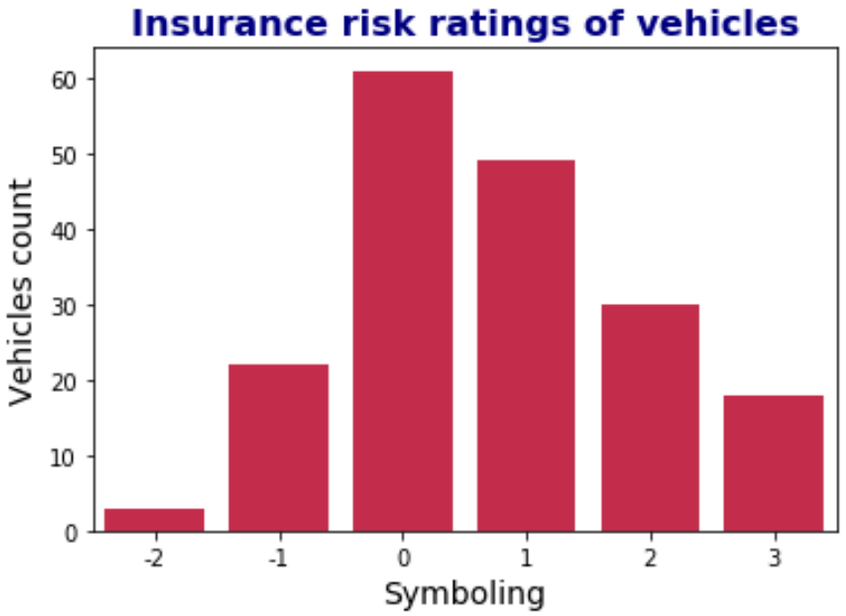
Problem 4: Summary

Which are highest normalized loss reported cars based on **body style** and **no of doors**?

- **Sedan** with **two doors** cars reported highest normalized losses followed by **hatchback** with **two doors** cars.
- **Two doors** cars has more number of losses than the **four door** cars.

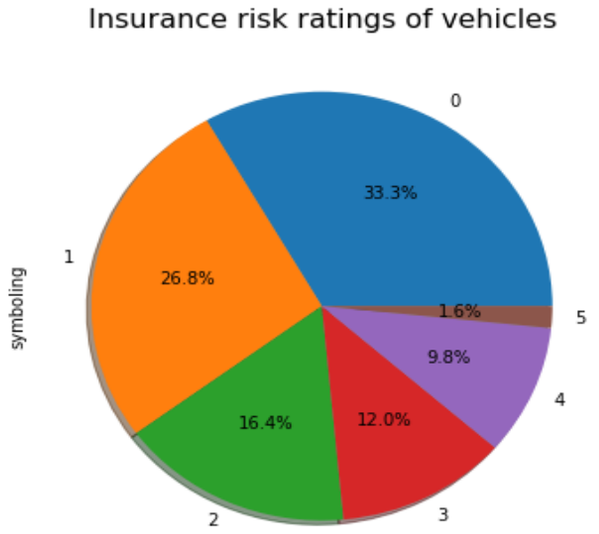
Problem 5

Does **body size** influence **symbolling**?



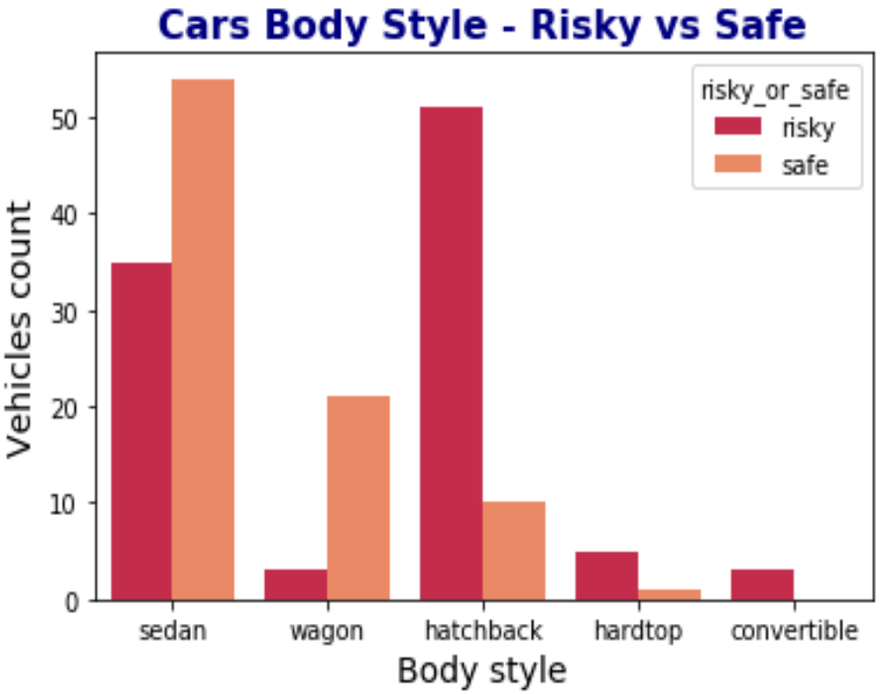
Symboling value shows how risky or safe a vehicle is, from an insurer's perspective. It can range from -3 to +3. -3 indicates a safe car while +3 denotes a risky one.

Based on the Symboling distribution, majority cars fall under 0 or 1 risk category which is less risky or less safer.



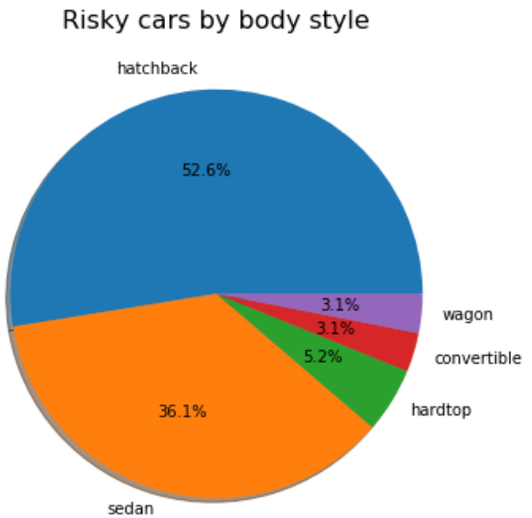
Our study focus to find relationship between body size and symbolling.

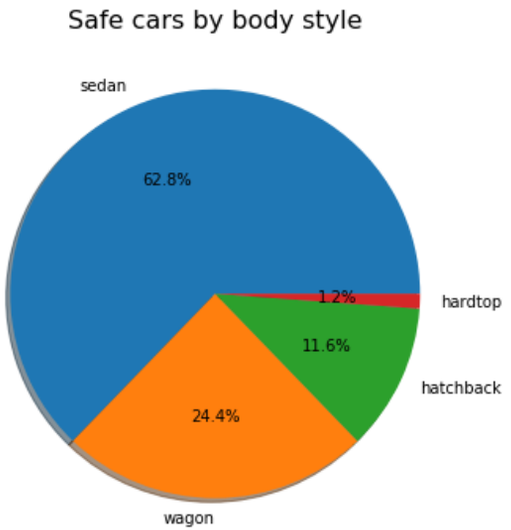
Cars Body Style – Risky / Safe



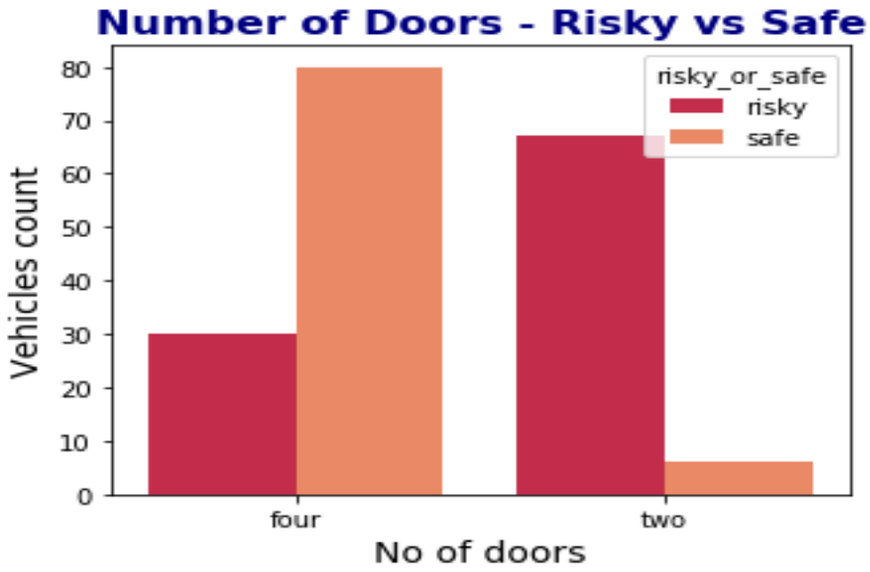
Introduced new column to classify risky or safe.

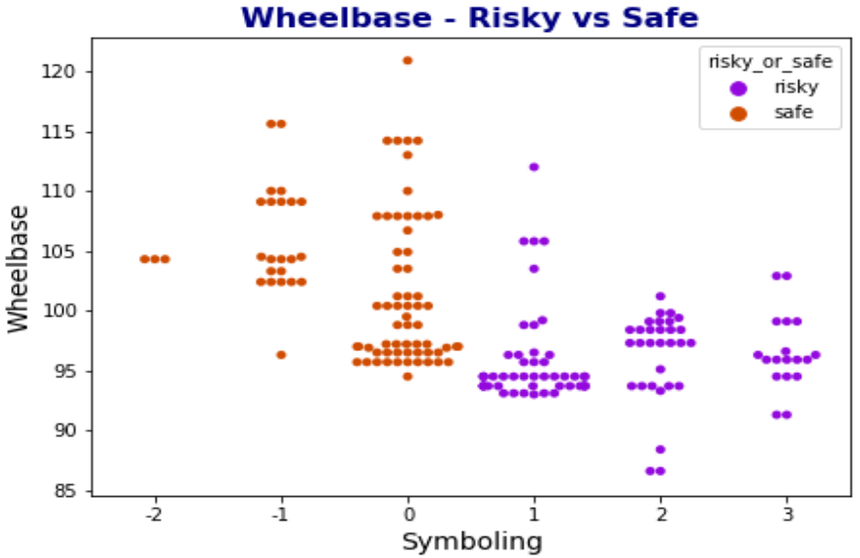
Majority of cars in hatchback, hardtop and convertible are in risky category.





Wheelbase, No of Doors – Risky / Safe





A two door cars is more risky compare to the four doors cars. May be two doors cars are design for drive enjoyment, due to that cars may be driven by an enthusiast.

More the wheelbase, more stable the car is. Declining in wheelbase is making risky the cars.

Problem 5: Summary

Does **body size** influence **symbolling**?

- Based on the Symboling distribution, majority cars fall under **0 or 1** risk category which is **less risky** or **less safer**.
- Majority of cars in **hatchback, hardtop and convertible** are in risky category.
- A **two door cars** is more risky compare to the four doors cars. May be two doors cars are design for drive enjoyment, due to that cars may be driven by an enthusiast.
- **More the wheelbase**, more **stable** the car is. Declining in wheelbase is making risky the cars.

Thank You

GitHub Notebook:

1. https://github.com/purnananda/Automobile_EDA/blob/master/Automobile_EDA.ipynb
2. https://github.com/purnananda/Automobile_EDA/blob/master/Automobile_EDA_Data_Cleaning.ipynb

References:

1. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/automobile
2. Different Types of Engine: https://www.mechanicalbooster.com/2016/08/different-types-of-engine.html