Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

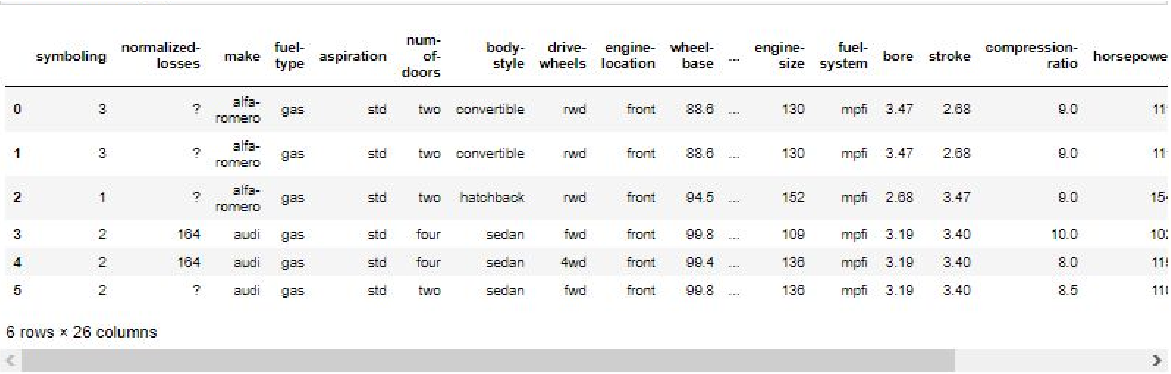
Visit our website

Introduction

Purchasing and choosing a car can be a daunting task, especially if you are concerned with car performance, fuel consumption, body shape, price, horsepower etc. This paper focuses on using exploratory data analysis (EDA), to give the consumer enough information to decide which automobile to select.

DATA CLEANING

Cleaning the automobile.txt dataset one had to look for the missing data and understand the information contained in the dataset. The dataset had a question (?) mark to represent missing data. Figure 1 below shows a normalized-loss column with '?'.



	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepowe
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	11
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	11
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	15
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	10
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	11
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19	3.40	8.5	11

6 rows × 26 columns

Figure 1. Missing values represented by the '?' mark.

The replace() method was used to replace '?' with NaN values which can be interpreted by the isnull().sum() methods to give a total of missing values in every attribute. A total of seven attributes had missing values which are as follows:

1. the price column had 4 missing values.
2. peak-rpm column has 2 missing values.
3. horsepower column has 2 missing values.
4. stroke column had 2 missing values.
5. bore column had 2 missing values.

6. num-of-doors column had 2 missing values.
7. normalized-losses column had 41 missing values.

The imputation method was used to replace the missing values with mean values of normalized-losses, bore, stroke, horsepower, peak-rpm and price columns respectively. This is due to The imputation method develops reasonable guesses for missing data, useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model. Figure 2 shows all the attributes with missing data.

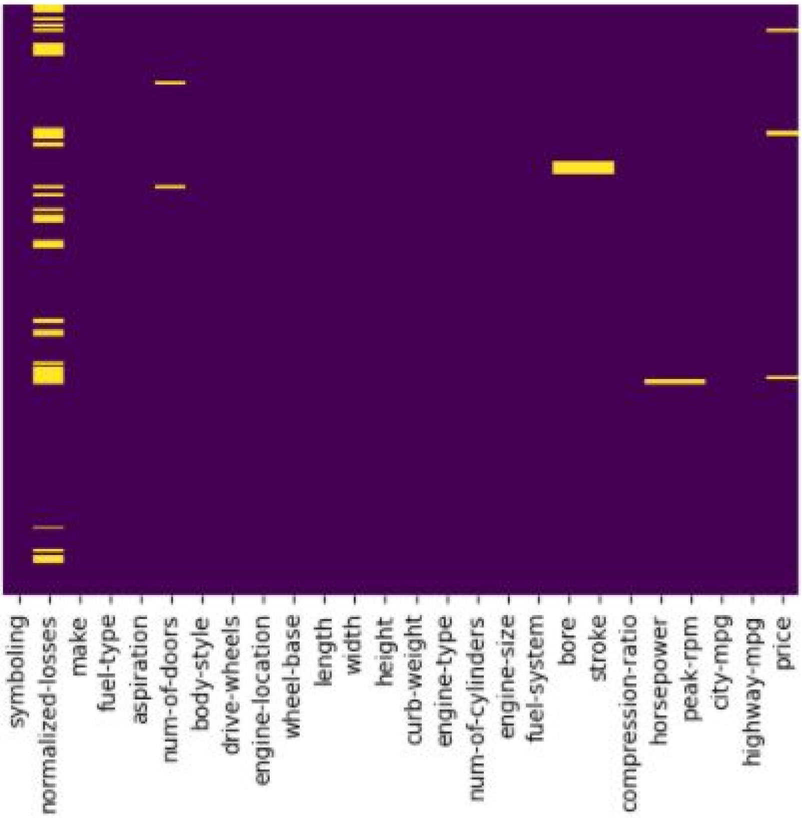


Figure2. Heatmap highlighting missing values.

The num-doors column, the largest frequency of the number of doors, was used to replace the missing values. Figure 3 illustrates the clean data set after the imputation process was performed.

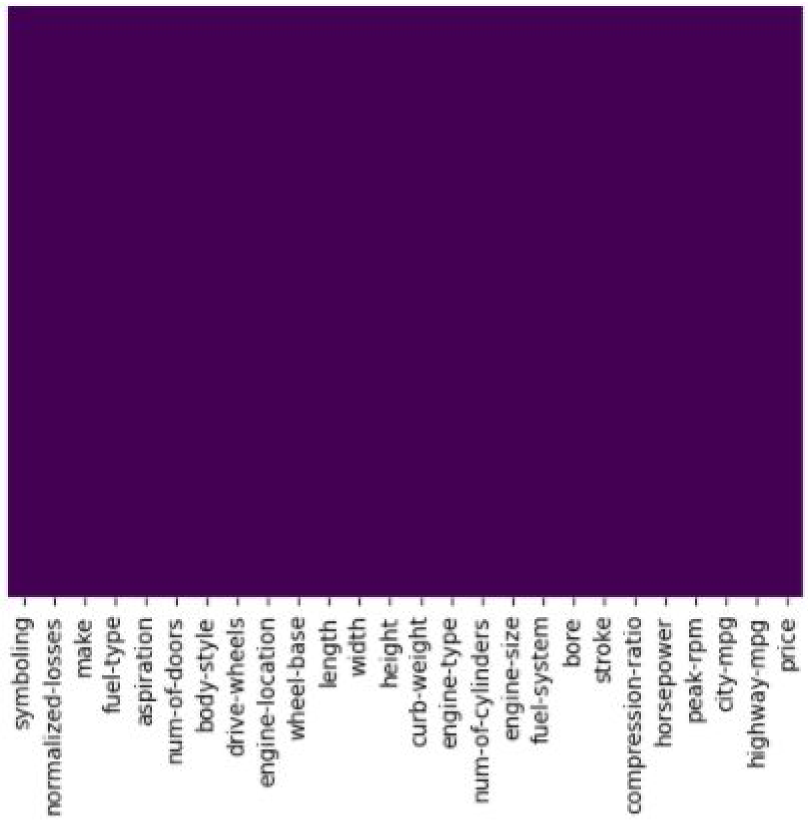
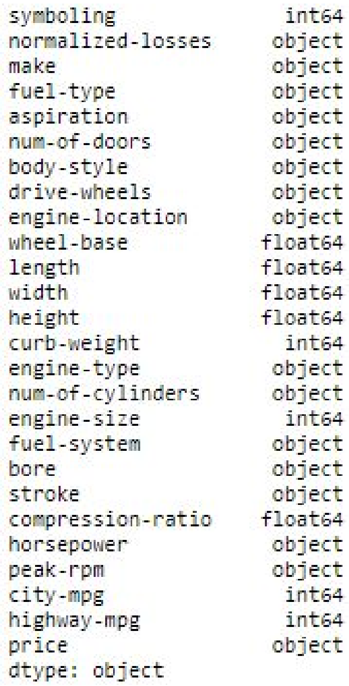


Figure 3. Dataset visualization after imputation.

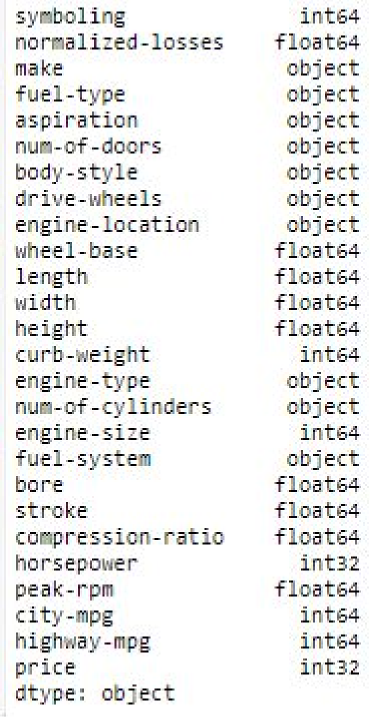
Attributes of the dataset had to be looked at before performing any analysis or visualization to be of the correct format or data type. Figure 4 shows the format or data type of all the columns in the dataset, dtypes returns the data types of each column.



```
symboling              int64
normalized-losses     object
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                  object
stroke                object
compression-ratio    float64
horsepower            object
peak-rpm              object
city-mpg               int64
highway-mpg            int64
price                 object
dtype: object
```

Figure 4. Original dataset attribute data types.

The following attributes peak-rpm, bore, stroke, normalized-losses their data types are object instead of float. Price and horsepower their data types were changed to integer as can be seen in figure 5.



```
symboling              int64
normalized-losses    float64
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                 float64
stroke               float64
compression-ratio    float64
horsepower             int32
peak-rpm             float64
city-mpg               int64
highway-mpg            int64
price                  int32
dtype: object
```

Figure 5. Formated data types.

The data types of the different attributes were changed using astype() method.

MISSING DATA

The dataframe had missing data, a total of seven attributes. The data frame had a low percentage of missing values, less than 20 percent which allowed for imputation using mean for numerical attributes and highest frequency of occurring class to replace missing values for word attributes. The num-of-doors column had two missing values after imputation and had to delete those rows that had missing values instead of the whole column, as can be seen in figure 6 below.

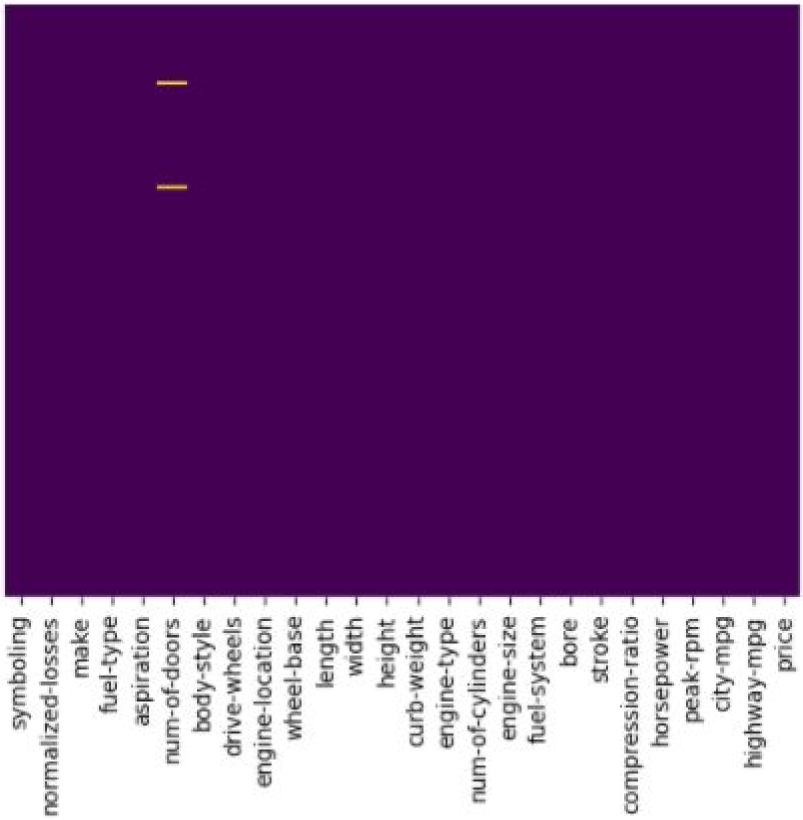


Figure 6. Nu-of-columns remaining missing values after imputation.

Data analysis and visualization of the Automobile dataset

Figure 7 below is a horizontal bar chart that shows the total number of cars available for every automobile maker. Toyota, Nissan, Mazda, Mitsubishi and Honda have the largest total respectively. Mercury, Renault, Alfa-Romeo, Chevrolet and Jaguar had the lowest total.

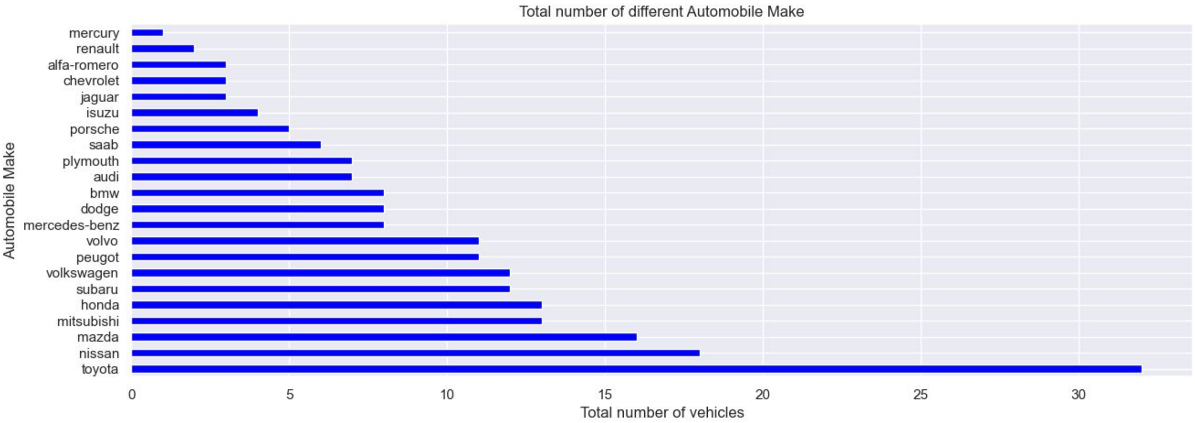


Figure 7. Make total automobile numbers.

The automobile makers and price are shown in figure 8. The box plot shows the different car models and their respective prices. Chevrolet is one of the cars with the affordable vehicles. Bmw, mercedes-benz , jaguar and porsche are the most expensive.

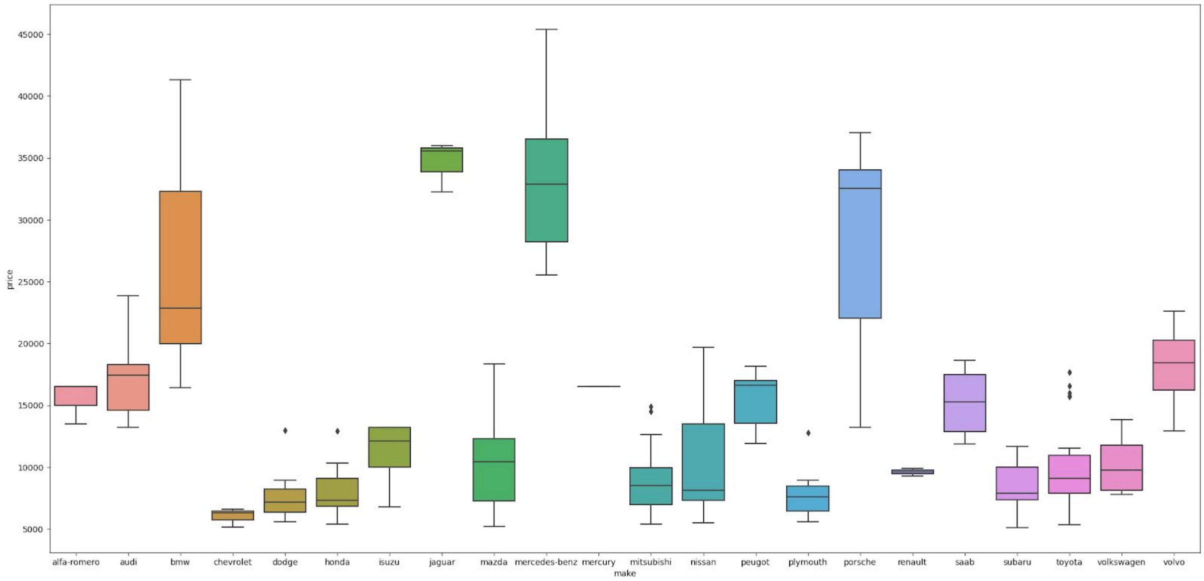


Figure 8. Average prices of different automobile makers.

Figure 9. Insurance risk factor for automobiles in dataset.

The second attribute “normalized-losses” is the relative average loss payment per insured vehicle year. This value is normalized for all automobiles within a particular size classification (two-door, small, station wagons, sports, specialty, etc), and represents the average loss per car per year. The values range from 65 to 256. Figure 10 shows this relationship.

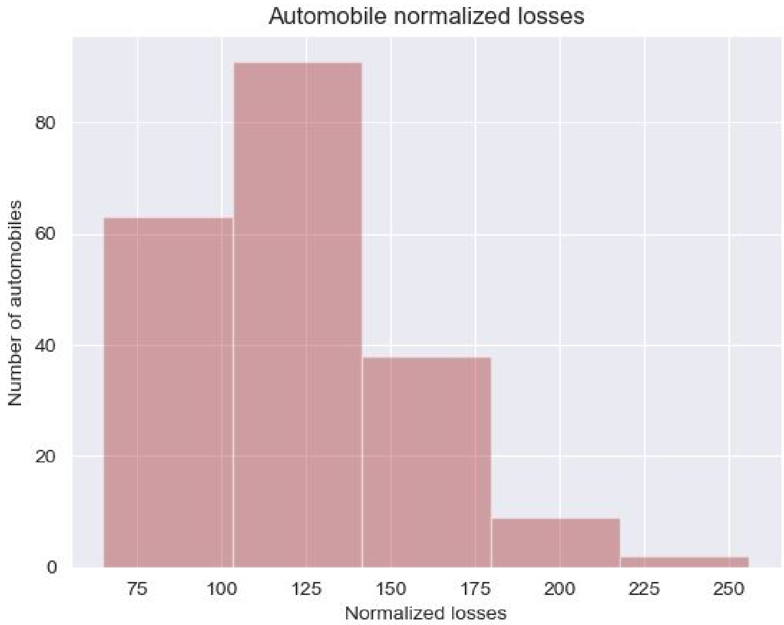


Figure 10. Automobile normalized losses.

Figure 11 below shows that gas automobiles are more affordable than diesel automobiles.

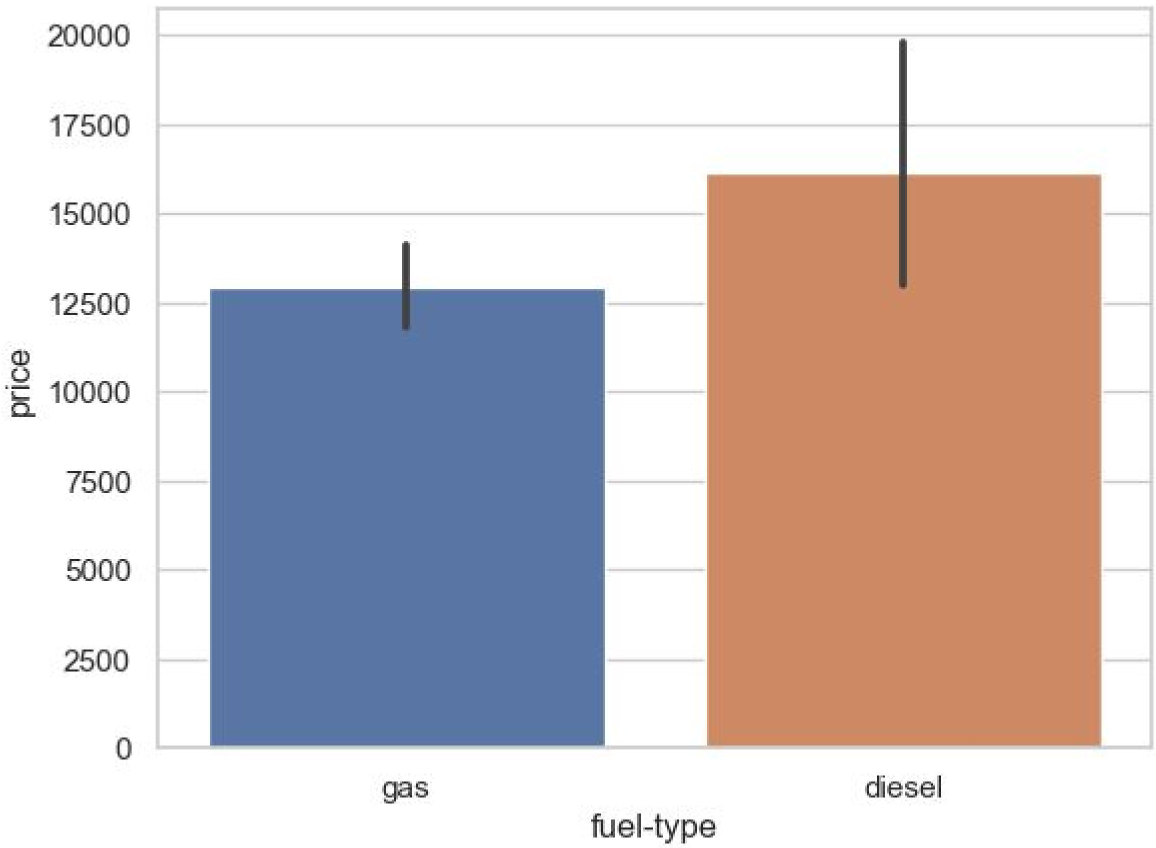


Figure 11. Automobile prices based on fuel type

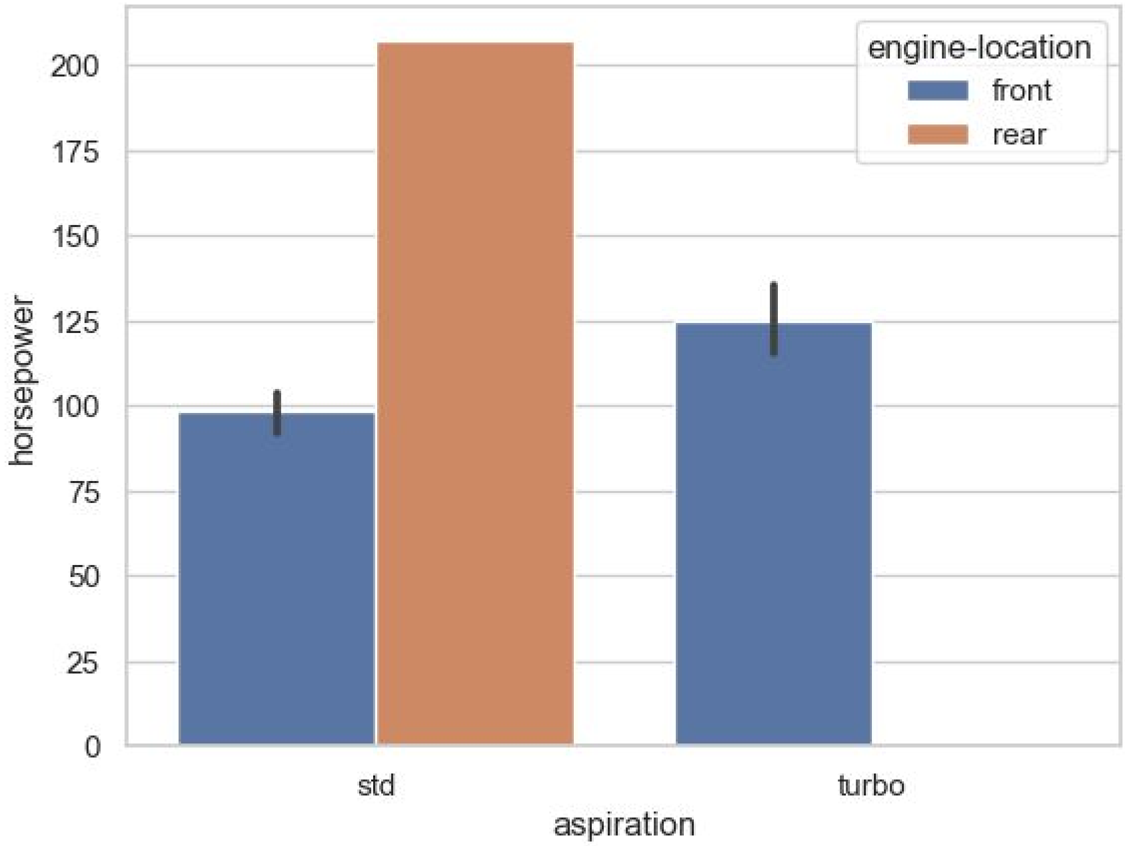


Figure 12. Automobile horsepower based on aspiration.

A standard automobile with a rear engine has more power as can be seen in figure 12. The database indicates that turbo automobiles with a front engine perform much better than the standard automobile.

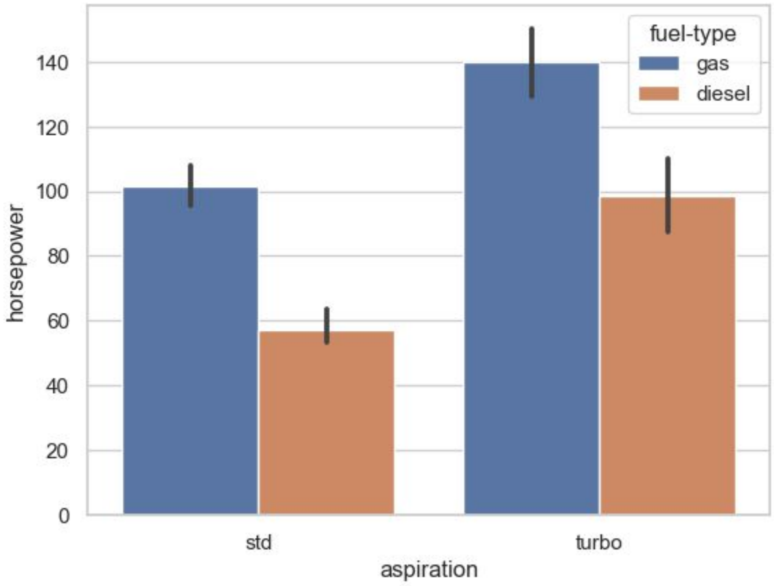


Figure 13. Automobile horsepower based on aspiration and fuel type.

Figure 13 above shows that gas automobiles have better horsepower irrespective of whether its a standard or turbo automobile. An automobile with a rear engine is expensive as it has more horsepower as can be seen in figure 14 below and figure 12 above.

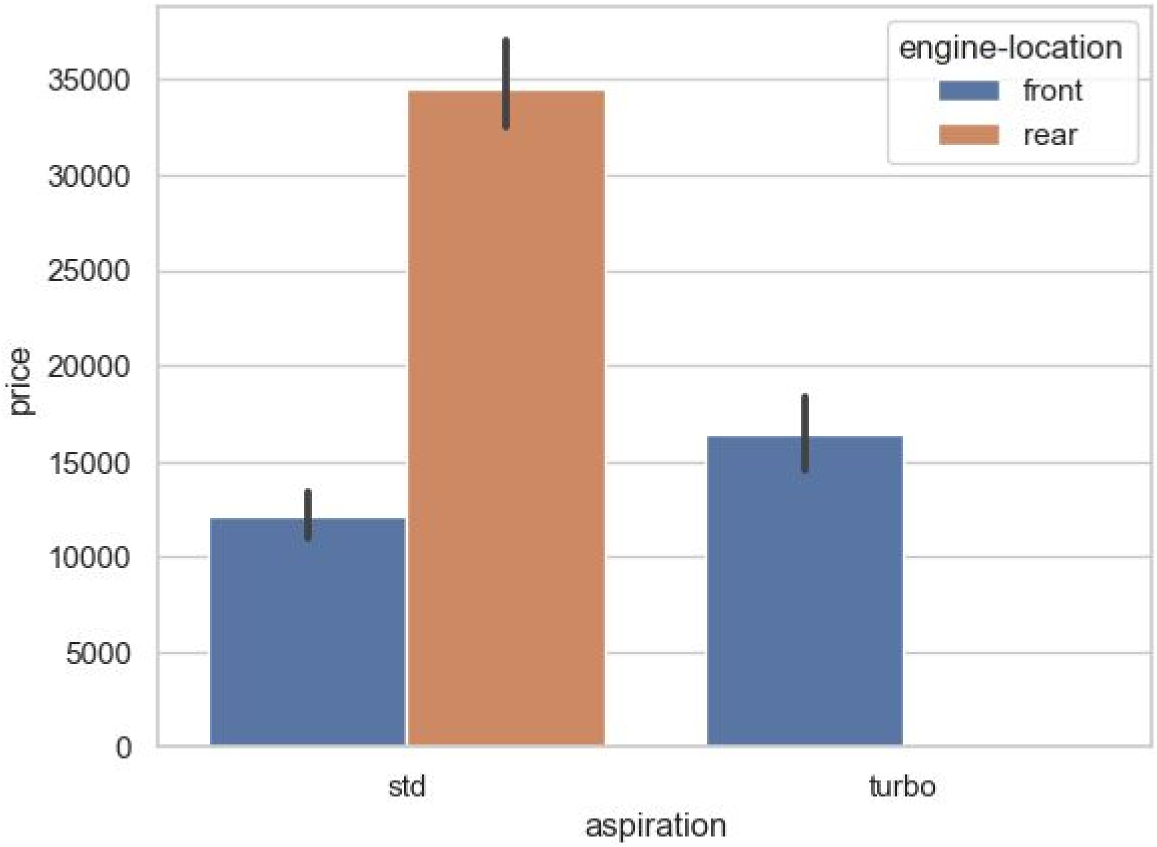


Figure 14. Automobile engine location pricing.

Figure 15 below shows that rear engine standard automobiles use less fuel on a highway compared to a front engine automobiles.

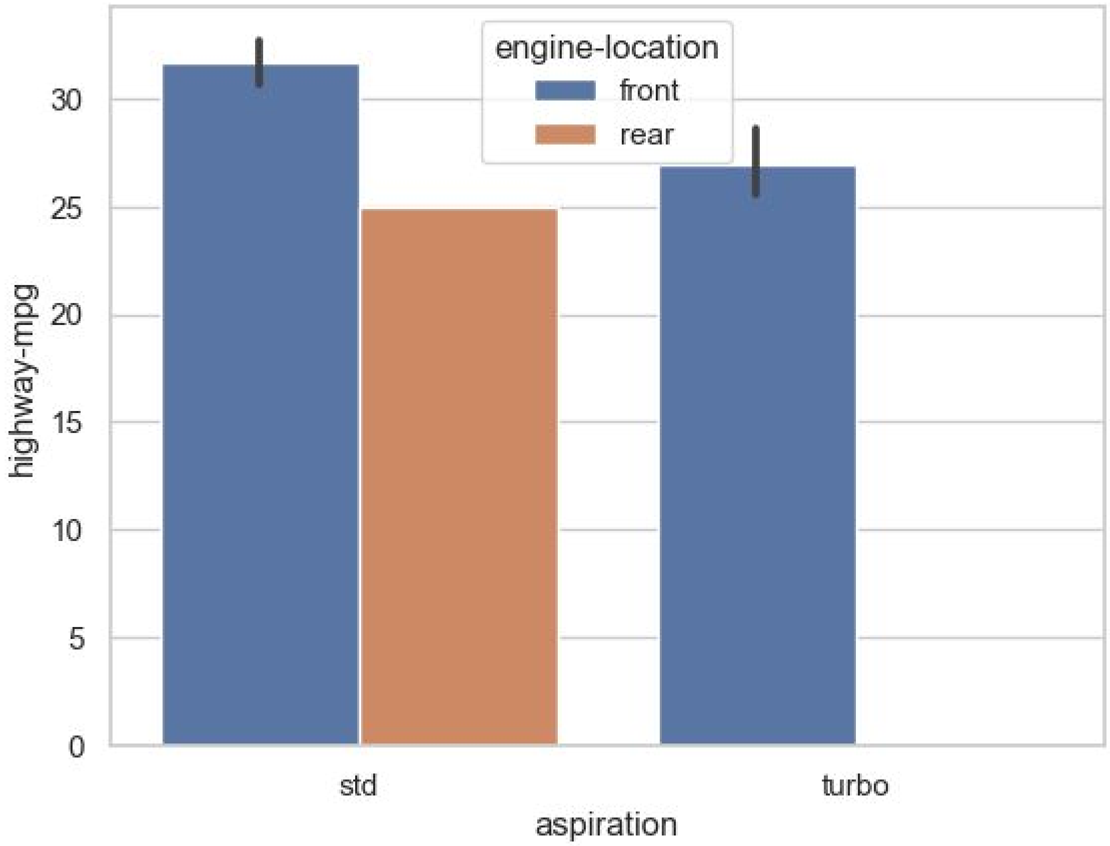


Figure 15. Automobile fuel consumption based on engine location.

Figure 16 demonstrates that rear wheel drive cars have more horsepower based on the body shape of the different automobiles. Four wheel drive performs better except for a wagon body shaped automobile.

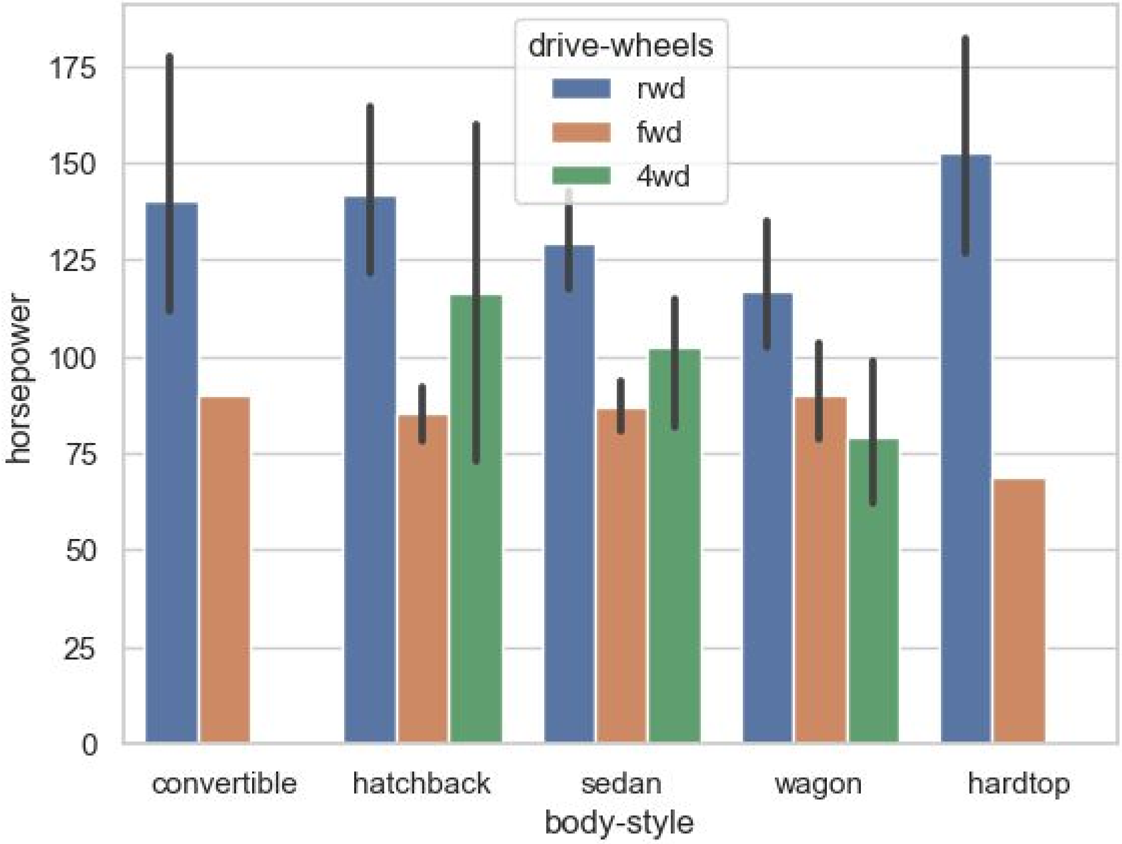


Figure 16. Automobile horsepower performance vs body style.

Figure 17 shows how drive wheels horsepower and fuel system compare attributes compare..

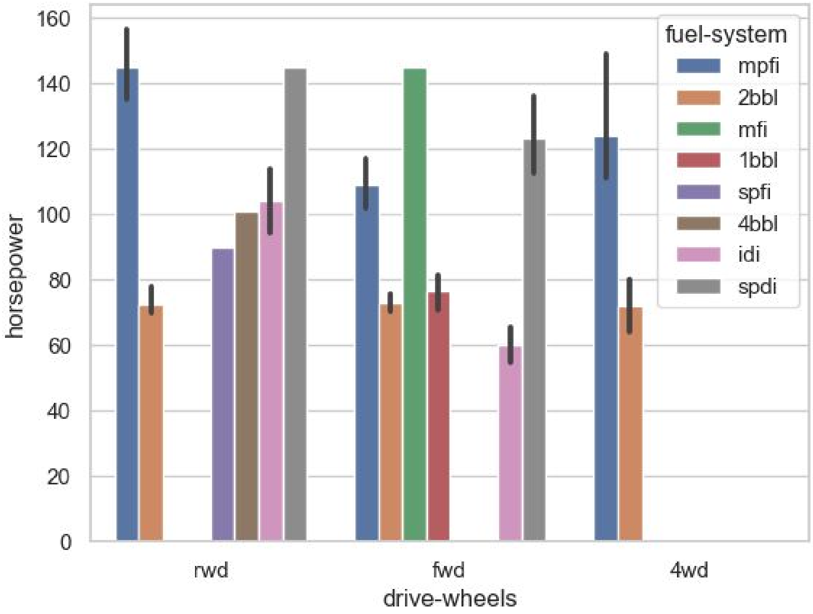


Figure 17. Automobile horsepower, fuel system and drive wheels.

The boxplot shown in figure 18 confirms the automobiles in this dataset, gas automobiles perform better than diesel automobiles.

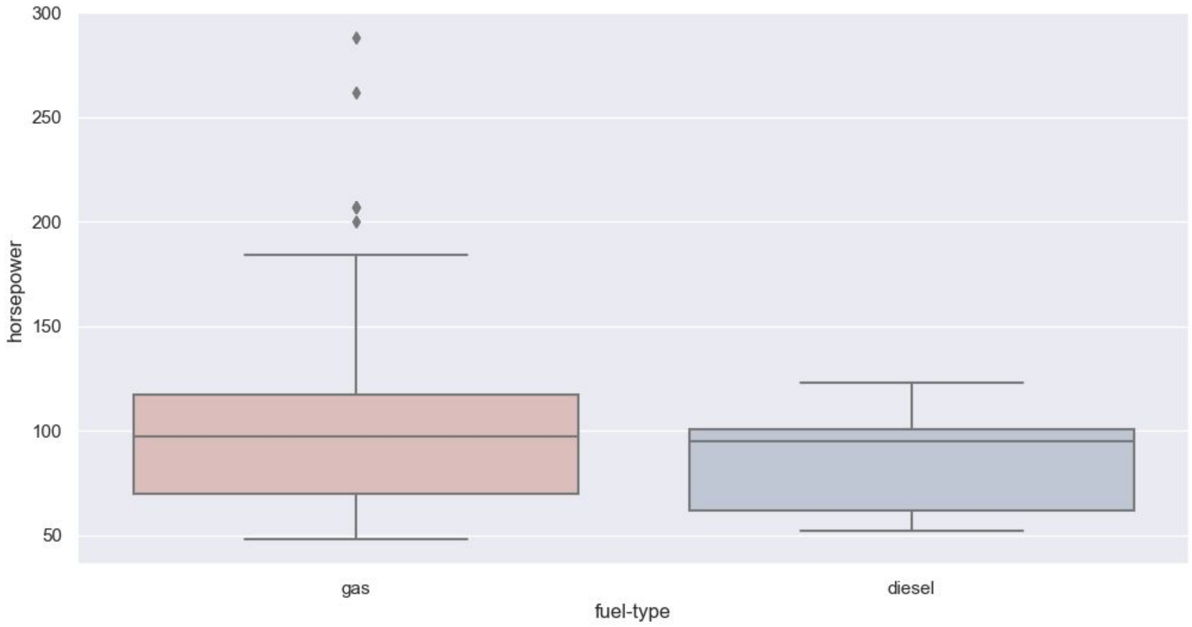


Figure 18. Boxplot for fuel type vs horsepower.

Figure 19 gives the prices of different automobile body styles, hatchback and hardtop are the least to most expensive based on their mean values.

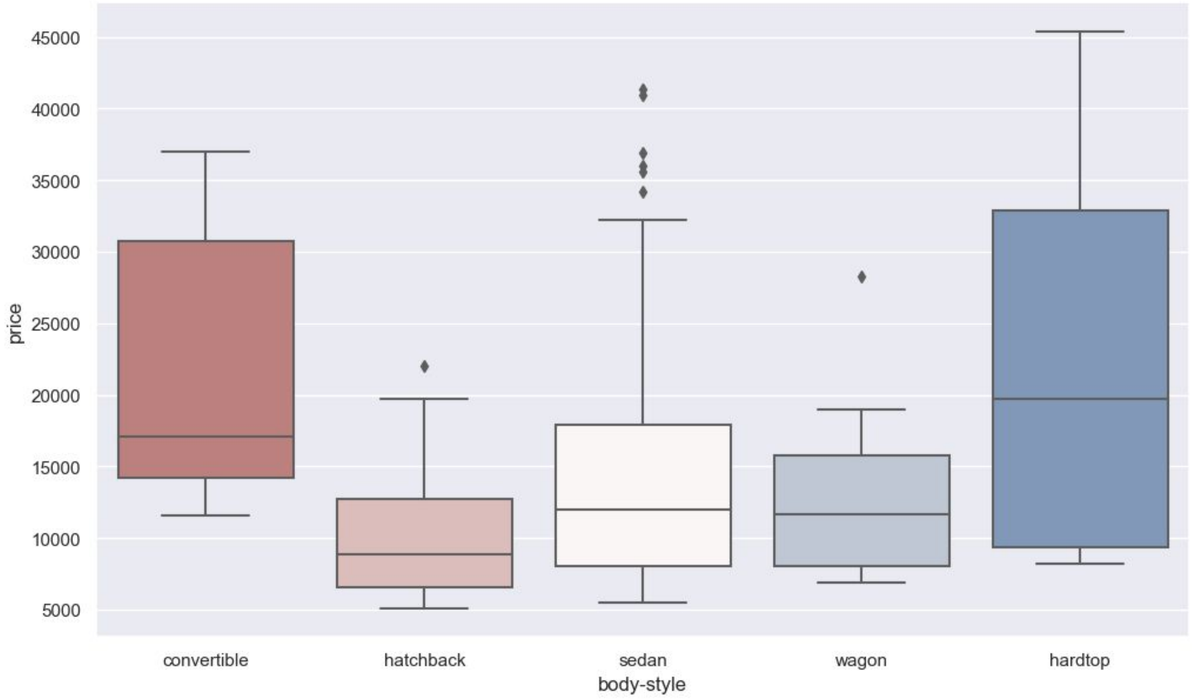


Figure 19. Body-style automobile pricing.

The gas automobile of engine type chcv has the highest price range compared to the other engine types, price range as can be seen in figure 20.

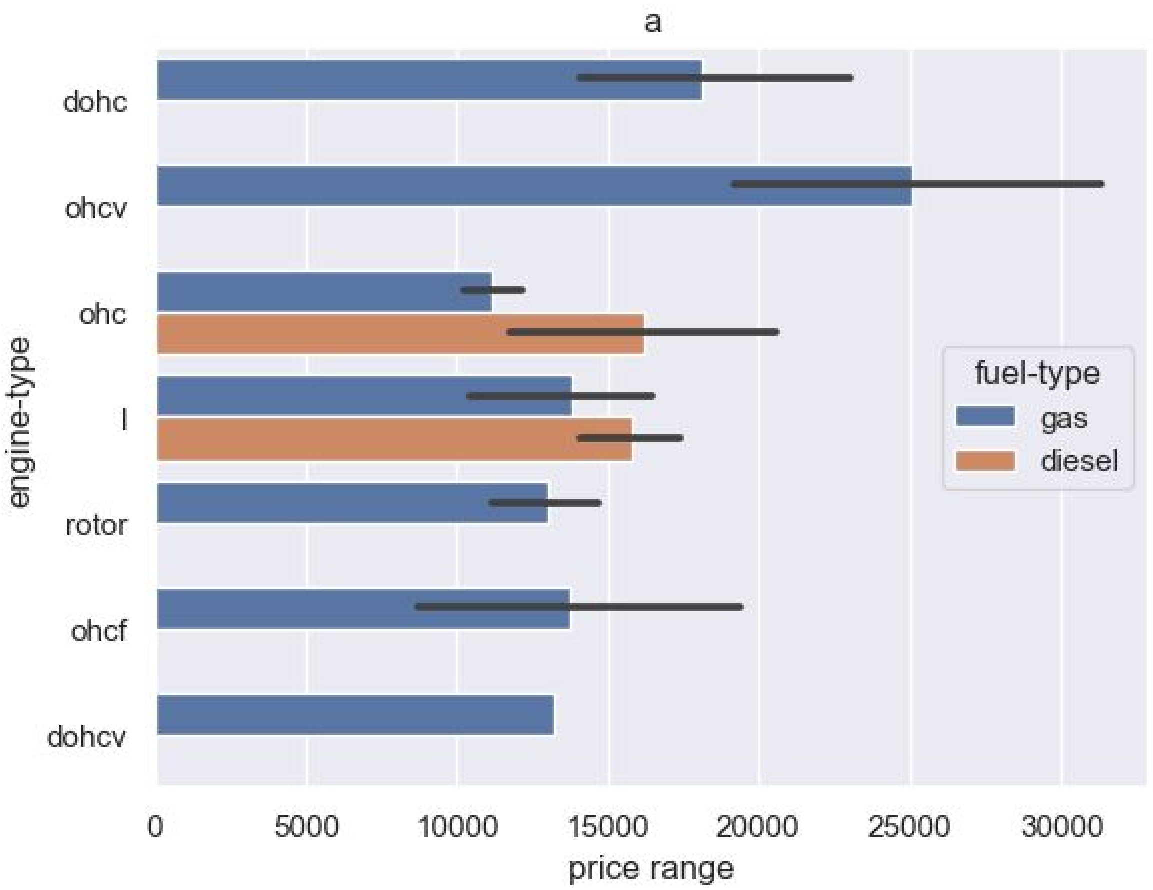


Figure 20. Automobile engine types.

Figure 21 below gives automobile price range based on horsepower.

Figure 21. Price range based on horsepower.

Figure 22 gives the automobile price range based on the normalized losses..

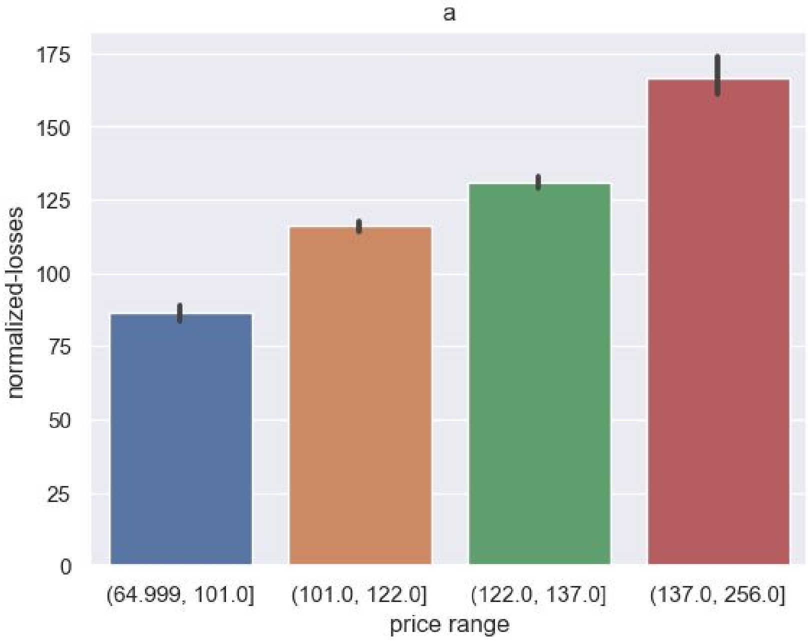


Figure 22. Price range based on normalized losses.

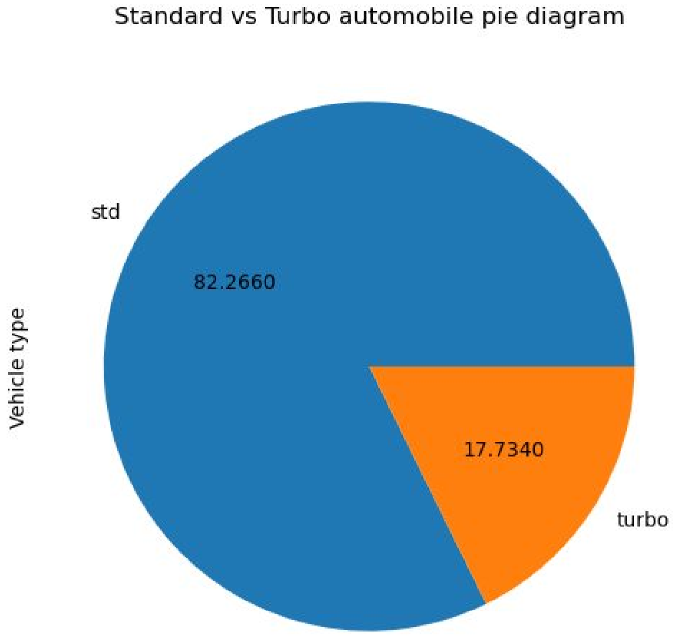


Figure 23. Automobile dataset aspiration makeup.

The pie chart in figure 23 shows that the dataset has mostly standard automobiles than turbo automobiles.

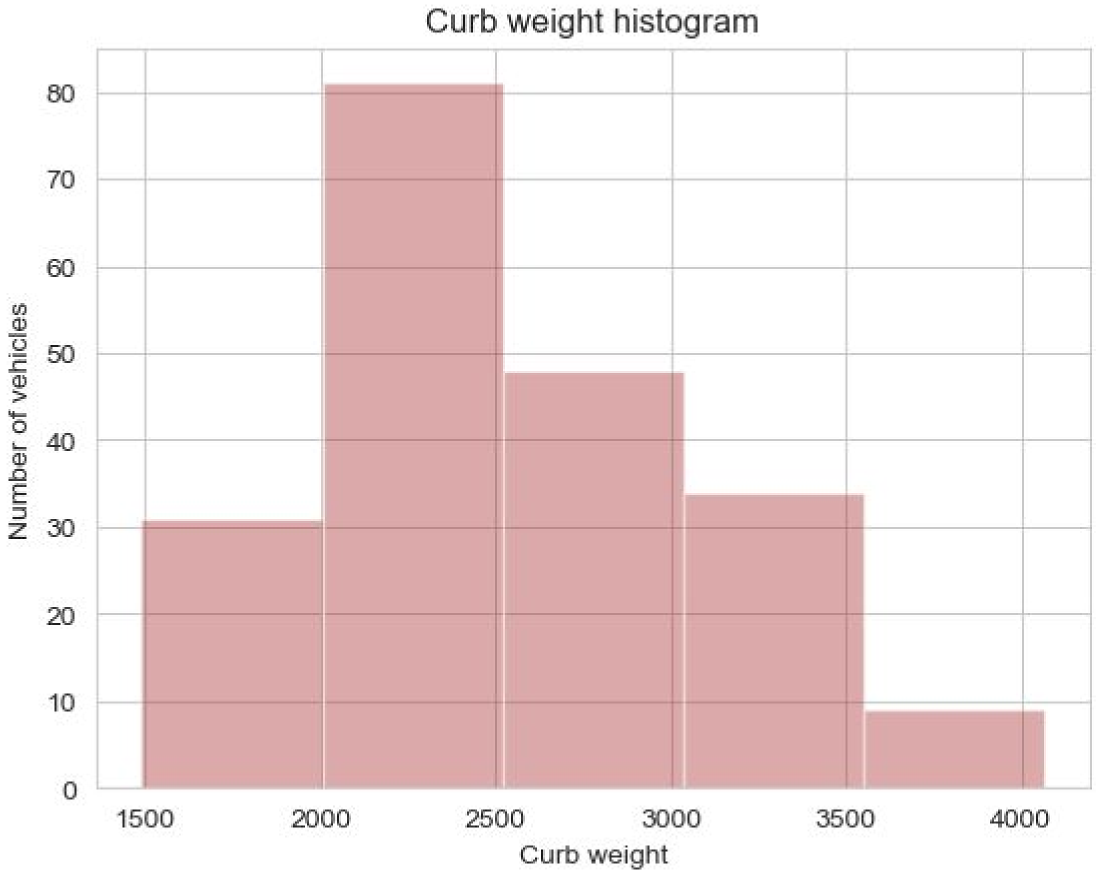


Figure 24. Total number of vehicles based on curb weight.

Figure 24 gives information on the number of cars in the dataset based on their curb weight.

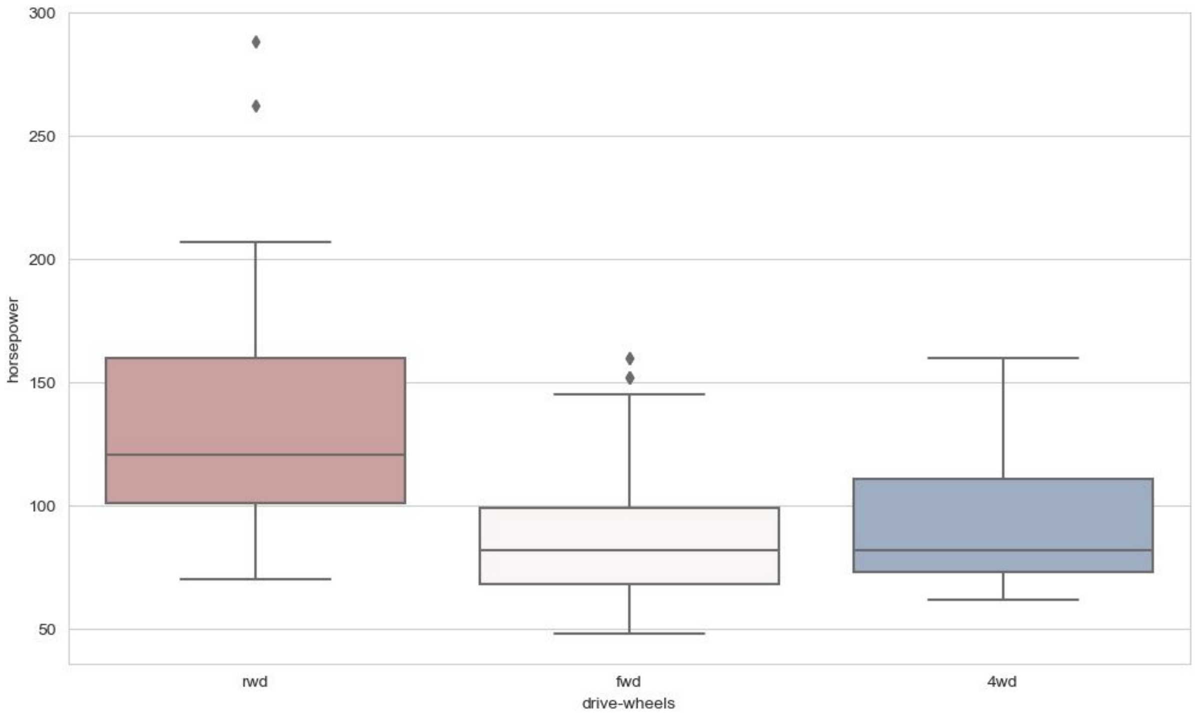


Figure 25. Boxplot horsepower based on drive wheels.

The boxplot in figure 25 gives a better visualization based on which wheel drive is better.

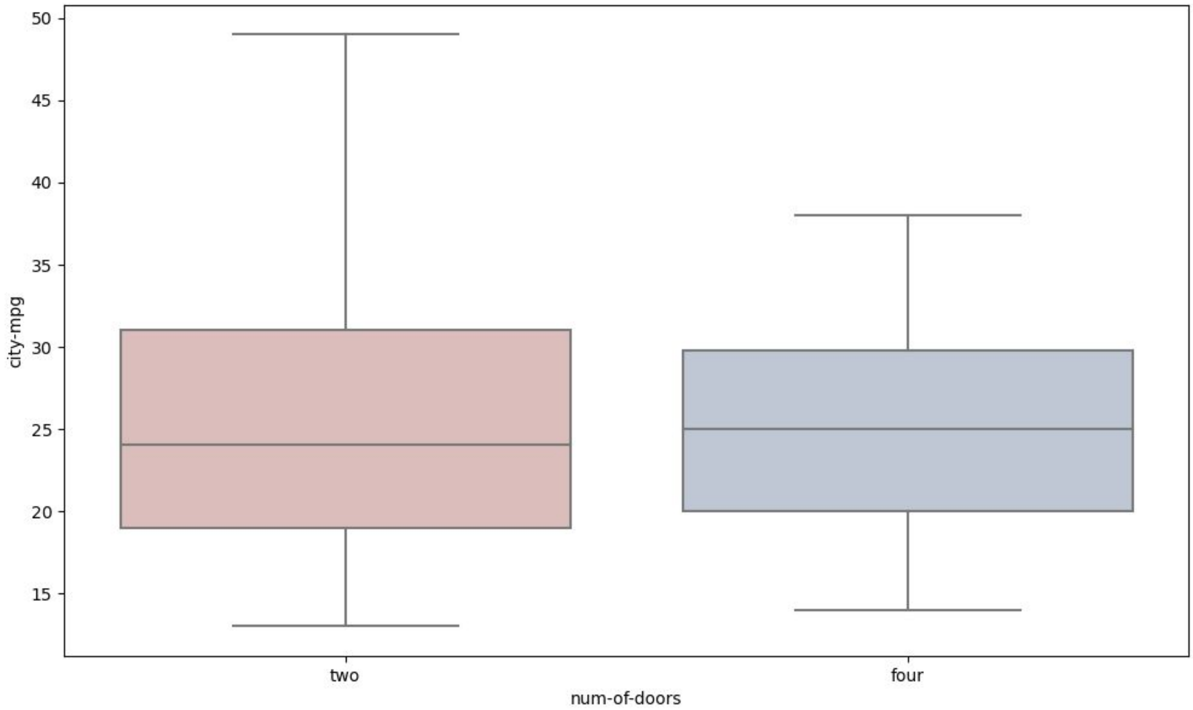


Figure 26. Num-of-doors based on city mpg.

Figure 26 above shows there is no much difference in terms of fuel consumption in the city, whether the car has four doors or two.

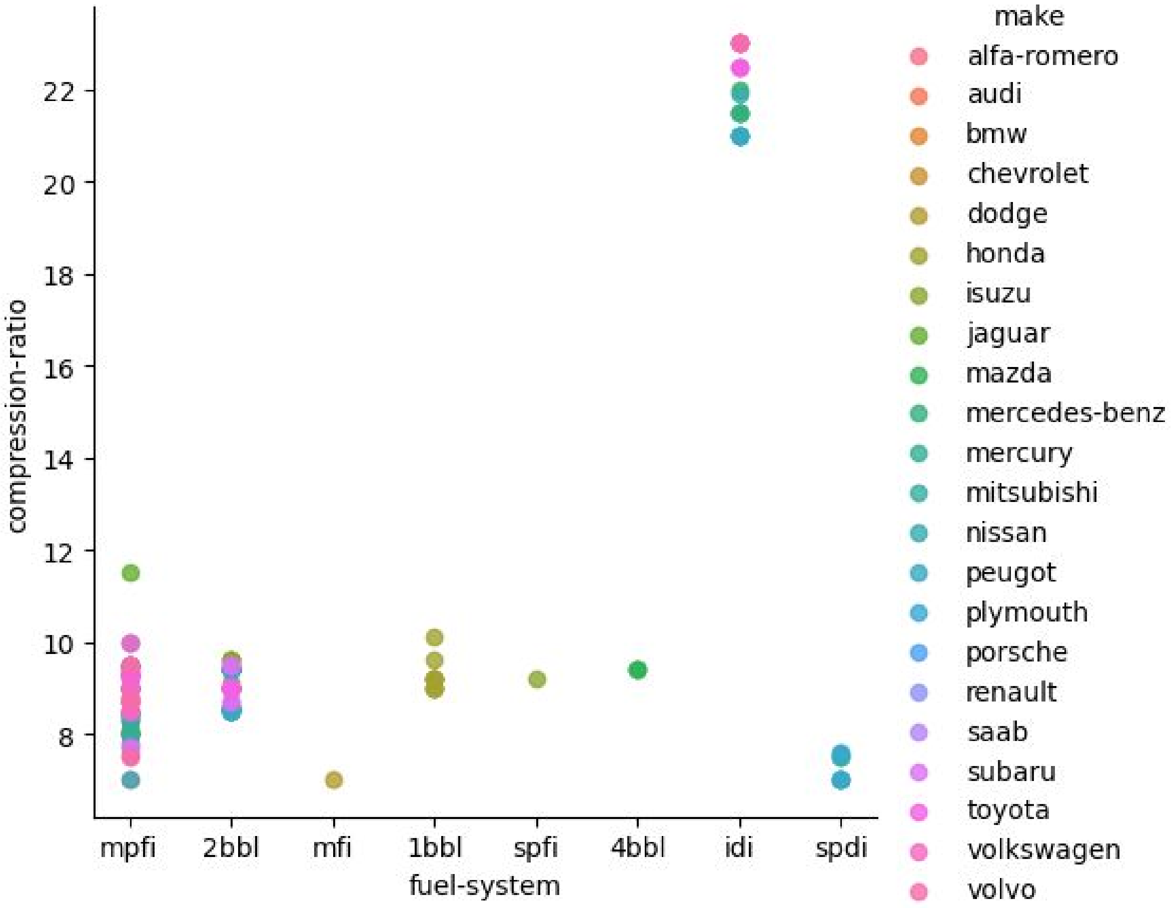


Figure 27. Compression ratio based on fuel system and make.

Figure 27 shows how compression ratios and fuel type influence the price of the different automobiles. Compression ratios usually range from 8:1 to 10:1. A higher compression ratio ranges from 12:1 to 14:1, would mean higher combustion efficiency. Higher compression ratios and combustion efficiency mean more power with less fuel, and fewer exhaust gases.

Conclusion

The most important thing to consider when purchasing an automobile that is contained in the data set, would be horsepower, fuel type, drive-wheels and pricing. These factors can help you narrow your selection when it comes to the type of automobile you want. The visualization of the automobile dataset helps the consumer to select the best option. The visualization helps compare automobiles based on different factors, than just looking for a particular brand of automobile.

References

1. EDA for Automobile Dataset, viewed 30 November 2020, <https://www.kaggle.com/toramky/eda-for-automobile-dataset>

2. Sriram, 2018, Let us do Data Analysis with Python, viewed 25 November 2020,<https://medium.com/@sriramselvank/let-us-do-data-analysis-with-python-db2cb6eca43f>

THIS REPORT WAS WRITTEN BY : LEHLOHONOLO VICTOR SEBAETSE