

**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

The purpose of this report is to discuss the findings of the automobile data set using a preliminary exploratory data analysis (EDA) approach. The automobile data set contains a total of 26 attributes including information about prices, makes, various mechanical features and speed performances. The dimension of the data set is 205 x 26. The data set had missing information which was addressed prior the EDA. No deep analysis was conducted for investigation such as deep learning.

**DATA CLEANING**

The data was checked for any duplicate rows but none was found. For each categorical data column, the values were visually checked for any duplicates or similarities by determining the respective unique data.

MISSING DATA

Initially, the data seemed complete with no missing values. Upon further investigation, several columns contained question marks shown as “?” in the data set. This represented that there were missing information present in the data set. The question marks were replaced with null values before they were replaced with actual values. This was done to make the data easier to manipulate. The following columns had missing data:

Table 1: Data sets attributes with missing data

|  |  |
| --- | --- |
| **Categorical** | **Numerical** |
| num-of-doors | bore |
|  | stroke |
|  | horsepower |
|  | peak-rpm |
|  | price |

For the numerical data, the missing values for each attribute was replaced by its respective mean. This was done as the value of each attribute was dependent on several other attributes in the data set. Due to the scope limitations of this investigation, using the mean was though as the appropriate approach. A deeper investigation is to be done to obtain more accurate and appropriate values.

Number of doors attribute (i.e. num-of-doors) was the only categorical data attributes that had missing data. An investigation was done to determine the relationship between the automobile body style and the number of doors. The following findings was found:

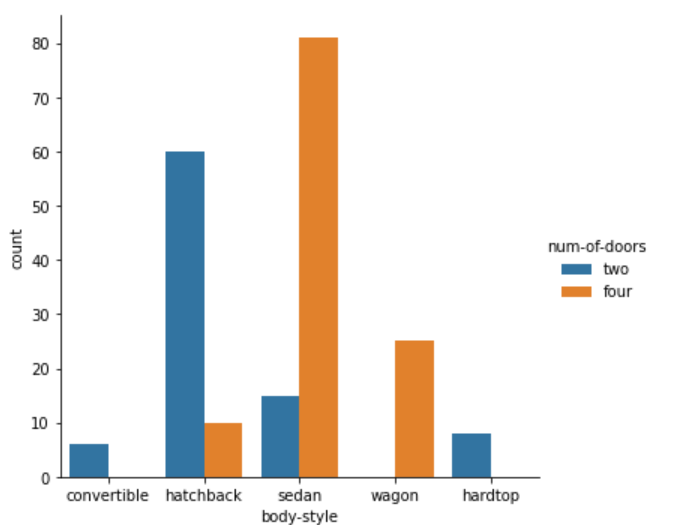


Figure 1: Relationship between body type and number of doors

Based on what was observed in Figure 1, automobiles either have two or doors. Additionally, it was found that some body types were exclusively manufactured with a finite number of doors. For example, convertibles and hardtops are only manufactured with two doors. Alternatively, some body styles are mostly manufactured with a preferred door number (Refer to Table 2). The missing values in number of doors attribute column were replaced with the values shown in the table above.

Table 2: Preferred # of doors per body style

|  |  |
| --- | --- |
| **Body Style** | **Preferred number of doors** |
| Convertible | 2 |
| Hatchback | 2 |
| Sedan | 4 |
| Wagon | 4 |
| hardtop | 2 |

DATA STORIES AND VISUALIZATIONS

This is section investigates the following:

* Automobiles categorical attributes that may effect the price
* Typical features or attributes an automobile would have based solely on the automobile’s body style.
* Correlation between the automobile’s numerical attribute data

PRICE

This section seeks to investigate the attributes which affect the price of the automobile. The following categorical attributes were investigated:

* Body style
* Fuel Type
* No of Doors
* Make
* Engine Type
* Number of cylinders
* Aspiration
* Drive-wheels
* Engine-location

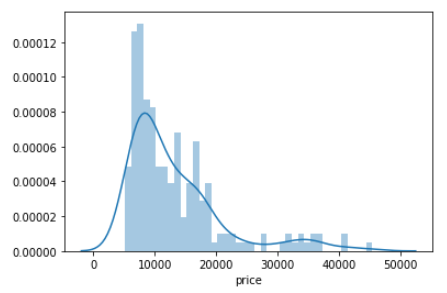


Figure 2: Price distribution of automobiles

Figure 2 displays the price distribution of the automobile prices. The distribution of the price shows that automobile generally costs between the 50000 and 120000. The lowest and highest price value is approximately 5000 and 45000 respectively.

Figure 3 is a boxplot showing price distribution for each automobile body style. Additionally, the body style is further split into gas or diesel were possible. Hardtops, sedan and convertible were found to be the more expensive body styles. The cheaper automobiles were the hatchback and wagon body styles.

Convertibles are the only body style that only has one engine type i.e. gas. It could be argued that if the body style is a gas type, the price will be slightly more expensive. However, the sedan automobile body style has relatively expensive outliers thus there will be cases where sedans are more expensive.

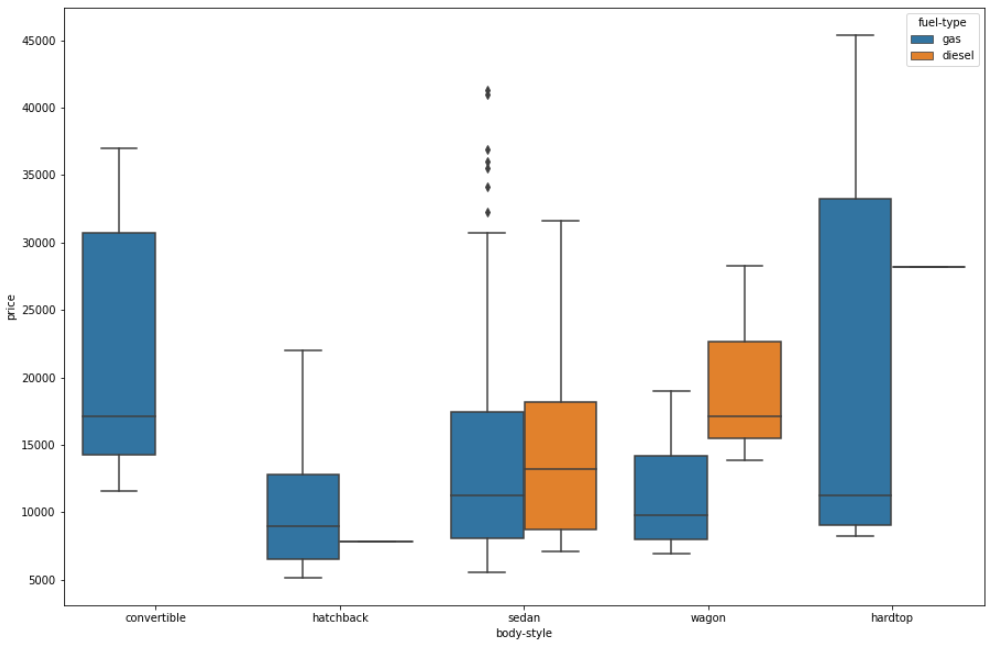


Figure 3: Correlation between price and body style and gas type

Figure 4 is a bar graph showing the correlation between price and the number of doors of the automobile. There is no significant difference between the number of doors and the price.

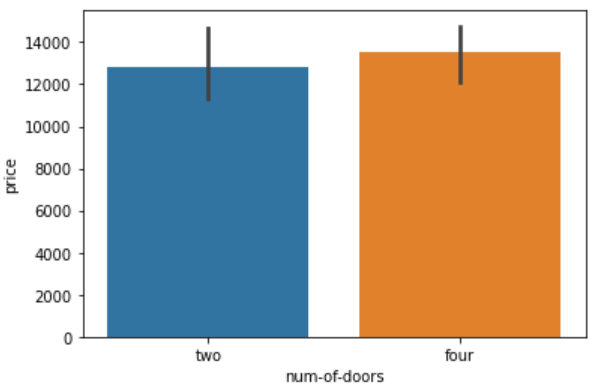


Figure 4: Correlation between # of doors and price

Figure 5 shows the correlation between the automobile make and the price. From all the different automobile makes, Mercedes-benz, BMW, Jaugar and Porsche make the most expensive cars Based, on Figure 3, these expensive automobile makes are mostly likely hardtops, sedans or convertibles. Automobile makes that tend to make the cheapest cars are Chervrolet, Dodge and Plymouth.

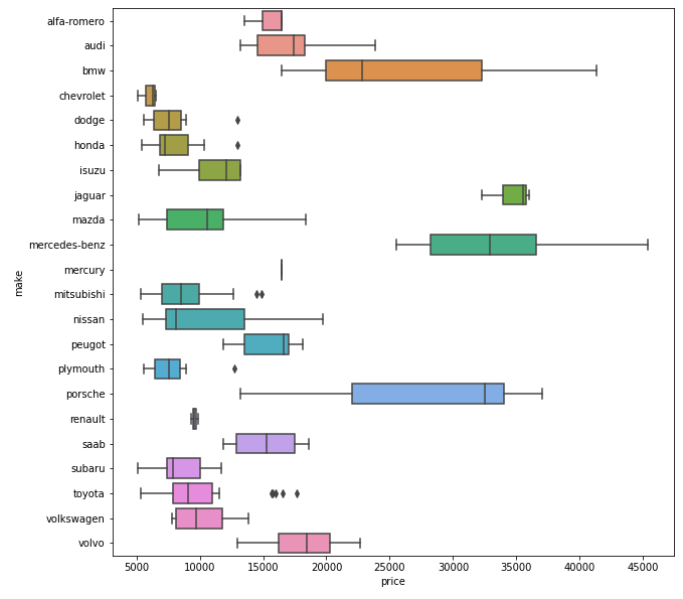


Figure 5: Correlation between automobile make and price

Figure 6 is a bar graph showing the relationship between the number of cylinders in automobile and price. Excluding the two cylinders, generally, the price is directly proportional to the increase in the number of cylinders. This trends only starts when the number of cylinders are three to twelve. The exception lies when the number of cylinders are two.

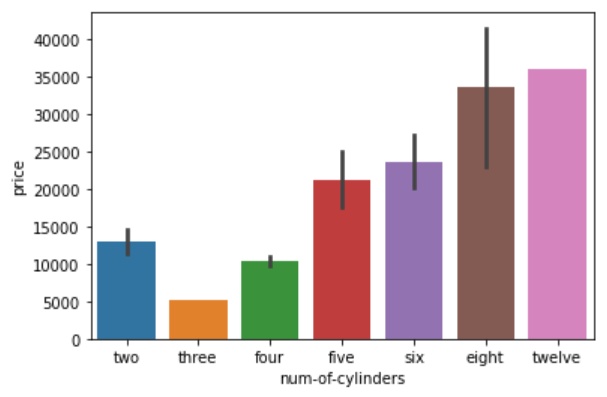


Figure 6: Correlation between # of cylinders and price

Upon future investigation, automobiles with two cylinders are gas fuelled hatchbacks manufactured by Mazda. The automobile with 12 cylinders is a gas fuelled, sedan automobile manufactured by Jaguar.

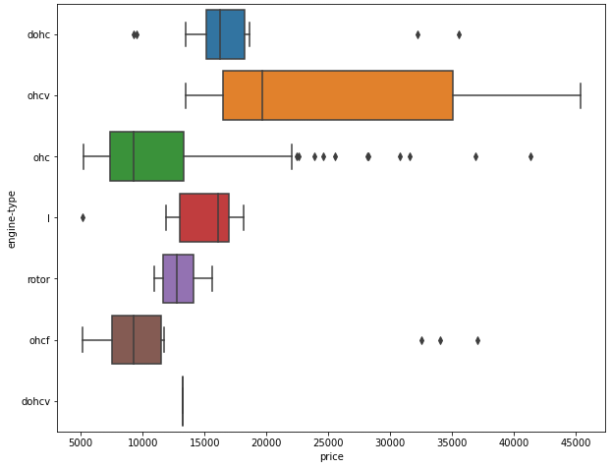


Figure 7: Correlation between engine type and price

Figure 7 is a boxplot showing the correlation between engine types and price. Ohcv engine types have the largest price distribution and has most expensive automobile. Upon further investigation, a variety of different car types contain ohcv engines. The rotor engine has the shortest price distribution. Upon further investigation, gas-fuelled Mazda hatchbacks contains are the only automobiles that contain rotor engine types. Finally, only one automobile has the dohcv engine which is a gas-fuelled Porsche hatchback.

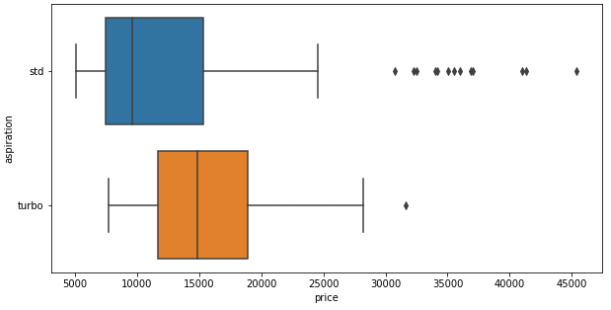


Figure 8: Correlation between price and aspiration

Figure 8 is a boxplot showing the price distribution of the different types of aspiration in automobiles. Having a turbo in a car allows it to generally make the car faster and more efficient. Otherwise the car’s speed naturally aspirated engine i.e. ‘std’. As expected, having a turbo, will generally be more expensive than having a standard aspiration. However, there are cases where the naturally aspirated engines will be more expensive as in the outliers.

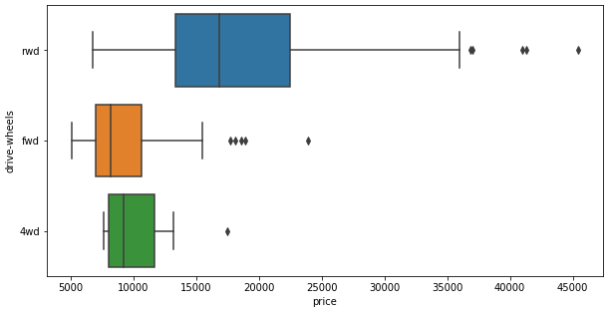


Figure 9: Correlation between price and drive wheels

Figure 9 is a boxplot showing the price distribution of the different types of drive wheels. Rear-wheel drive i.e. ‘rwd’ has the widest distribution suggesting that this type has in a variety of cars. However, a further investigation was done to see if there is any specific car type when looking at front-wheel drive (“fwd”) and four-wheel drive (“4wd”) and nothing significant was found.

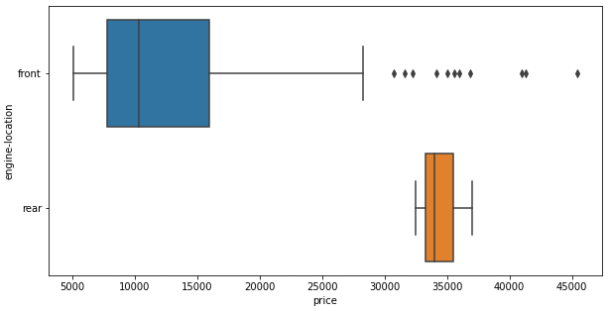


Figure 10: Correlation between engine location and price

Figure 10 is a boxplot showing the price distribution of the different types of engine locations. Automobiles with engines located in the front have a wider price distribution in comparison to rear positioned engines. Rear positioned engines are generally more expensive compared to the font positioned engines but have a smaller distribution. Upon further investigation, automobiles with rear positioned engines are gas-fuelled Porsche automobiles which are either hardtops or a convertible.

AUTOMOBILE BODY STYLE

This section seeks to investigate typical categorical features or attributes an automobile would have based solely on the automobile’s body style. Currently, the typical number of doors and price range for the automobile body style. The following categorical attributes were investigated:

* Drive wheels
* Aspiration
* Fuel system
* Highway-mpg
* Make

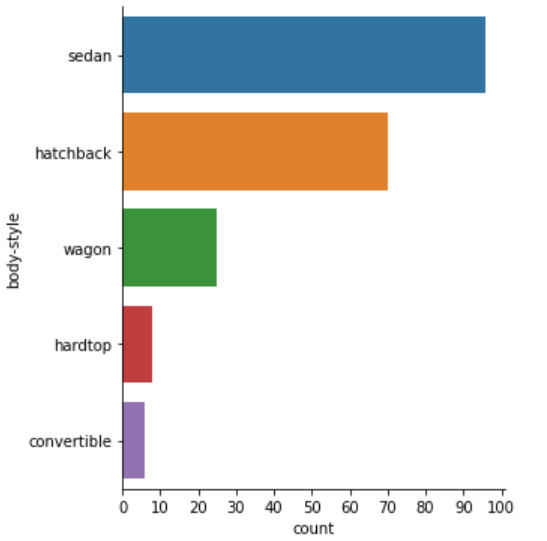


Figure 11: Count of various automobile body styles in data set

Figure 11 is a bar graph showing the total number of various body styles in the data set. The graph will allow to better understand whether an attribute is significant by comparing the value to the total number of the respective body style. Sedans and convertibles are the most and least common body styles in the data set.

Figure 12 is a bar graph showing the total number drive wheels for each respective body style. Both Convertibles and hardtops are built as rear-wheel drives. There are no convertibles and hardtops which is a four-wheel drive. For the other vehicles body styles, generally are manufactured as front-wheel drive or rear-wheel drive by leaning more towards front-wheel drive. From a drive wheel perspective, four-wheel drive type automobiles are not present in convertibles and hardtops

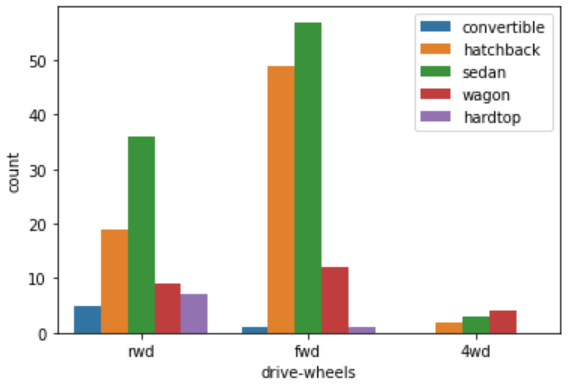


Figure 12: Correlation between automobile body type and drive wheels

Figure 13 is a bar graph showing the total number of aspiration types for each respective body style. For all body styles, the automobile are mainly manufactured with or do not include a turbo. For example, there are no convertibles with turbos in this data set.

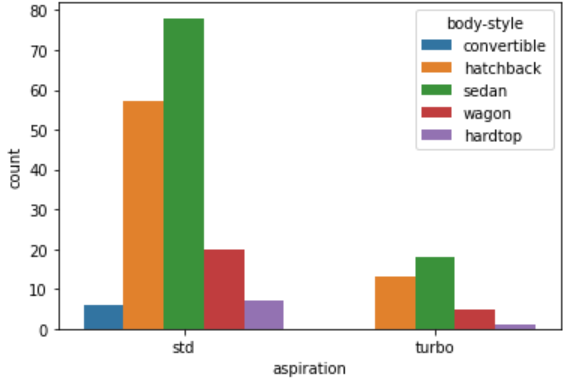


Figure 13: Relationship between aspiration and automobile body style

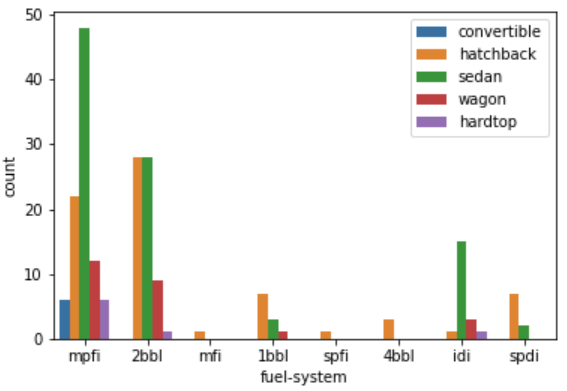


Figure 14: Relationship between fuel system and body style

Figure 14 is a bar graph showing the total number of different engine types for each respective body style. Not all body styles contain every engine type. The data shows that each body style is only manufactured with a finite number engines. The exception lies with hatchbacks where can have been manufactured using all the engine types in the data sets.

Generally, automobiles are manufactured with either mpfi or 2bbl type engines. Convertibles are an exception to this observation. Convertibles are only manufactured with mpfi type engines.

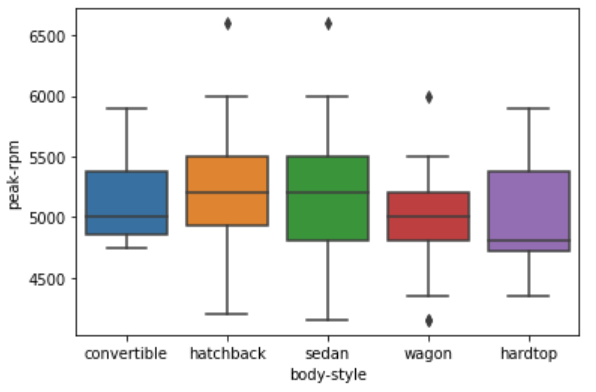


Figure 15: Relationship between highway-mpg and body style

Figure 15 is a box plot showing the peak-rpm distribution for each respective body style. No significant difference could be observed. Peak-rpm has a high linear correlation with high-rpm (Refer to following section) thus a similar trend can be observed.

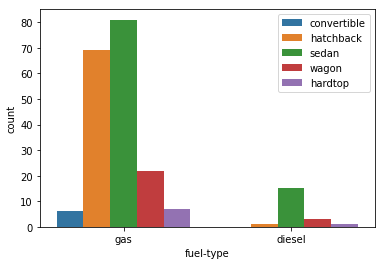


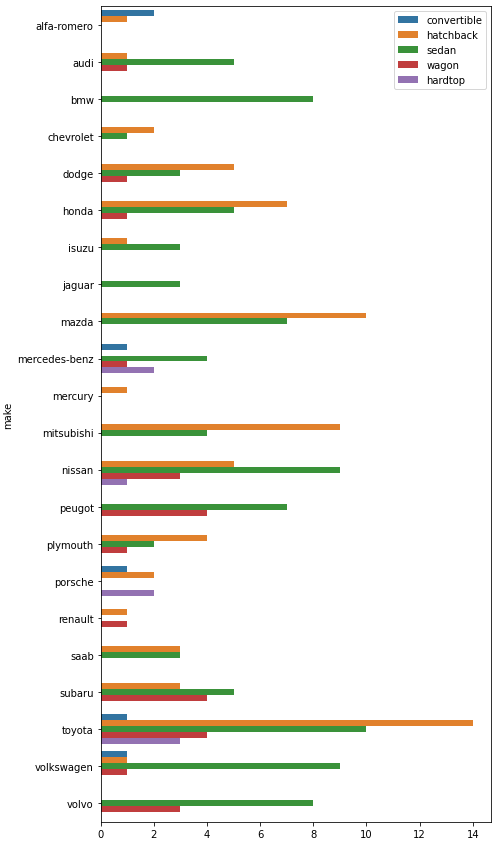
Figure 16: Correlation between fuel type and body type

Figure 16 is a bar graph showing the total count of the different fuel types for the different body styles. Most of the vehicles in the data are gas-fuel types. As mentioned in a previous section, there are no gas-fuelled convertibles.

Figure 17 is a bar graph showing the total count of the different makes for the different body styles. The data shows a trend where that each body style is only manufactured with a finite number engines. Some body style are manufactured by only a select few makes. This can be observed for the convertible and hardtop body types. For Convertibles are manufactured by only Alfa-Romero and Toyota and Volkswagen. Hardtops are only made by Mercedes-Benz, Nissan, Porsche and Toyota. Juxtapose to this, there are body styles that manufactured by almost all the makes. This can be in seen in the Hatchbacks and Sedans body types.

From make perspective, not all body styles are manufactured by each make. For examples, Mercury and BMW only manufacture hatchbacks and sedans respectively. In contrast, Toyota manufactures all the automobile body styles.

In conclusion, one cannot can entirely determine the make of the car solely based on the body style. More information would be required.

Figure : Correlation between automobile make and body style

AUTOMOBILE BODY STYLE

This section seeks to investigate the significance of the linear correlations of the automobile’s numerical attributes. Figure 18 is a heat map showing linear correlations between the numerical attributes. The numerical value within each block denotes the strength of linear correlation. The closer to 1 or -1 the value is, the more significant the linear correlation either positive or negative respectively. The closer the value is to 0, the less significant the correlation is.

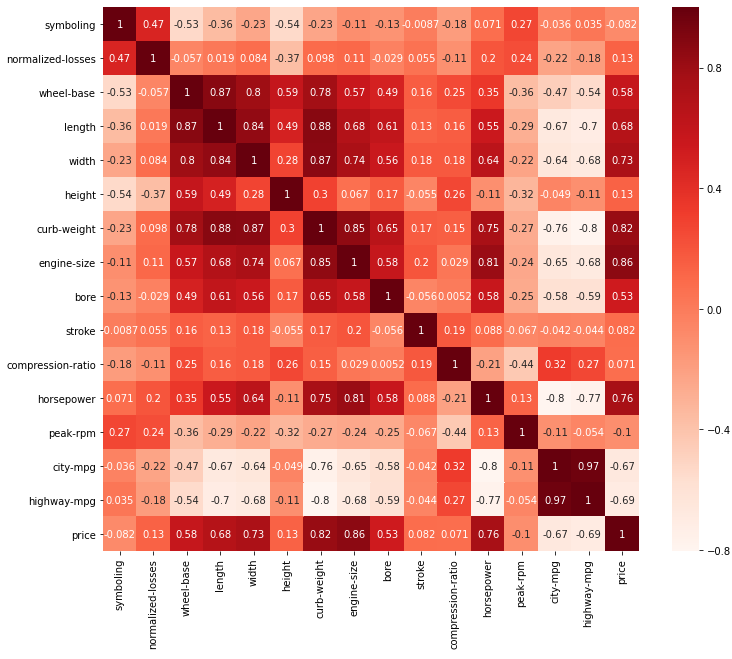


Figure 18: Linear correlation of numerical automobiles attributes

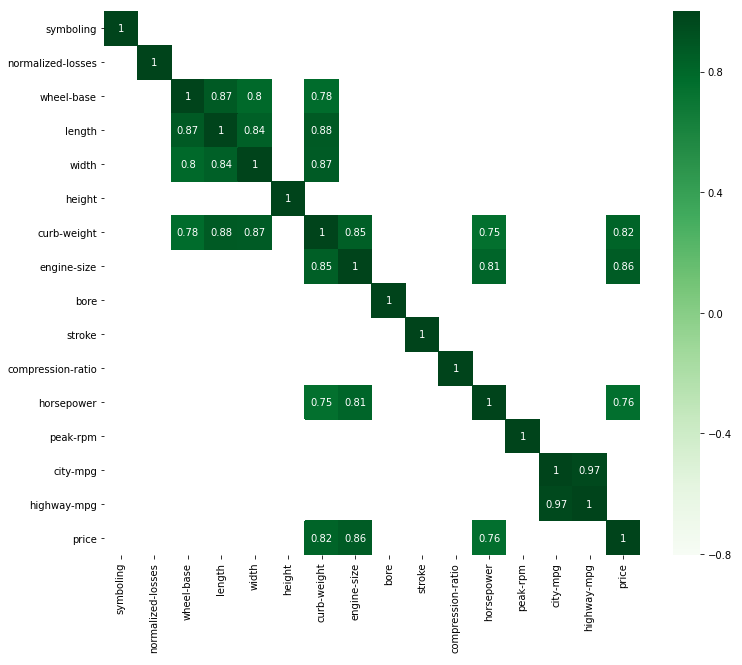


Figure 19: Automobile numerical attributes with significant positive linear correlation

Figure 19 is a heat map showing the significant positive linear correlations between numerical automobile attributes. It was assumed that is the value was greater 0.75, it would be categorised as a significant positive correlation. Several positive correlation can be observed. Most notably, city- and highway-mpg having a very high positive correlation of a value of 0.97.

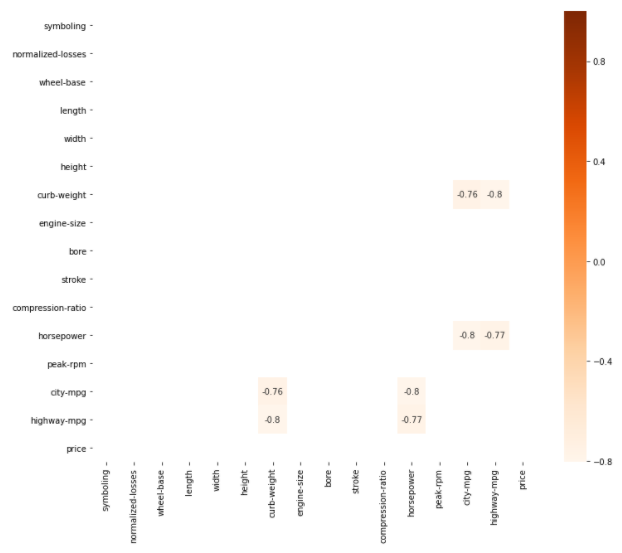


Figure 20: Automobile numerical attributes with significant negative linear correlation

Figure 20 is a heat map showing the significant negative linear correlations between numerical automobile attributes. It was assumed that is the value was less than -0.75, it would be categorised as a significant negative correlation. There are fewer negative than positive linear correlations. Interesting to note that when an attribute has significant correlation highway-mpg it will have a correlation with city-mpg and verse visa. This could be due to city- and- highway-mpg together having a significant positive linear correlation.

CONCLUSION

The purpose of this report is to discuss the findings of the automobile data set using a preliminary exploratory data analysis (EDA) approach. Data was cleaned and missing data was field before in the investigation. The investigation identified trends with regards to automobiles attributes that effect the price, typical attributes an automobile would have based solely on the automobile’s body style and correlation between the numerical attribute data.

In conclusion, there were clear trends observed in correlation in the various automobile’s attributes and the price. In certain cases, when identifying the most common attributes for a body make some attributes can be generalised. If not, additional information make may be required. Finally, clear significant linear correlations both positive and negative could be observed among the automobile’s numeric attributes.

REFERENCES

https://medium.com/analytics-vidhya/a-beginners-guide-to-exploratory-data-analysis-eda-c96b780f744d

https://en.wikipedia.org/wiki/Drive\_wheel

https://stackoverflow.com/questions/24775648/element-wise-logical-or-in-pandas

https://stackoverflow.com/questions/59241052/ordering-a-python-seaborn-barplot-by-ascending-or-descending

https://www.codegrepper.com/code-examples/typescript/pandas+add+multiple+lists+to+dataframe

https://stackoverflow.com/questions/19900202/how-to-determine-whether-a-column-variable-is-numeric-or-not-in-pandas-numpy

https://www.kaggle.com/gayatrirao98/exploratory-data-analysis

https://www.w3schools.com/python/python\_lists\_loop.asp

THIS REPORT WAS WRITTEN BY: Nabeel Omar

