**COSC 2670 – Practical Data Science**
**Assignment 1: Data Cleaning and Summarizing**
Tenzing Sangay Bhutia (s3680446)
s3680446@student.rmit.edu.au
Wednesday, 10th April 2019

## ABSTRACT

The aim of the report is to examine an Automobile dataset and carry out the first initial steps of data science process, including the cleaning and exploring of the data. We start with Data Curation and examine the dataset. Then we clear the dataset from all unnecessary errors such as typos, redundant spaces and missing values. Body-Style, Symboling and Horsepower is then explored with different visual representation. Relationship between Body-Style, Symboling and Price is then explored and necessary hypothesis are tested. Lastly a scatter matrix and heat map are plotted to explore the correlation between all the numerical data.

## DATASET

This Dataset consist of various characteristics of an automobile and its insurance risk rating along with its normalized losses in use as compared to other cars. In total the dataset contains 26 attributes and 238 observations or instances. The attributes present contain Categorical, Integer and Real characteristics. The range and complete information of the attributes in the dataset have been given by the experts in the industry.

## DATA PREPARATION:

- **Packages:** We start by loading relevant packages necessary for our analysis.

- **Importing Dataset:** Then, we import the raw dataset into the our Jupyter workspace. As our source file is of the type csv, we use the read csv function in pandas to help us with the importation and store it to a variable name Automobile. We provide relevant headers to the dataset as provided by the experts to get better insights on the data.
  We then use the shape and head function to check whether the loaded dataset is equivalent to the data in the source CSV file. All duplicate observations present were then dropped to prevent bias in the dataset. The datatypes of the variables were also observed for further awareness of the dataset.

- **Typos:** All typos present in the dataset were checked for each column using the value count function and a replace function was then used to manually overrule all the error values with the actual required value given by the experts. We handle these initially because Python is case sensitive and consider similar observations such as ‘gas’ and ‘Gas’ as different. This small difference can cause a huge error and problem later in our analysis and modelling and led us to the famous proverb ‘Garbage in equals Garbage out’.

- **Redundant Whitespace:** Whitespaces tend to be hard to locate but these errors cause problems like another redundant characters and need to be taken care of before analysis. A string function is used to remove all the leading and trailing whitespaces in the analysis so as to prevent any inconvenience in the future.

- **Impossible Values:** As we are provided with all the standard interval of quantitative values by the automobile experts we consider all values beyond this interval as impossible values in this analysis. We start our sanity check by using the describe function to see the max and min values present and compare it with the threshold given. In this case we see the presence of impossible values in the symbolling, normalized-losses and Price column. We take care of this impossible values differently, i.e. For Symbolling and Normalized Losses, we use the concept of capping and we replace the impossible values with the nearest threshold value. However, in the case of price we replace the

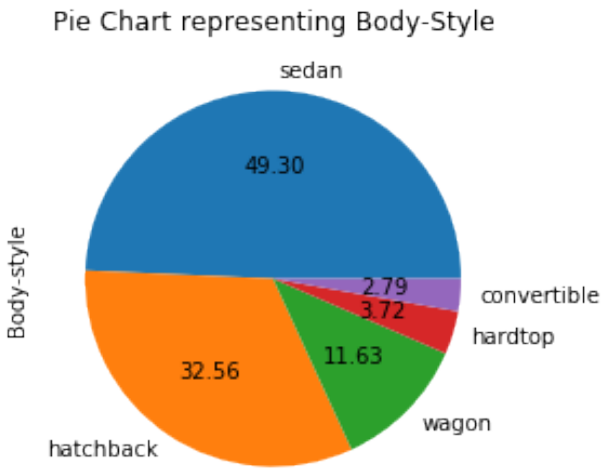
impossible value by replacing it with the mean of the standard price of the manufacturer by using the group by function. This solution is taken into consideration as companies has standard price for their vehicles.

- **Outliers:** All outliers present in the dataset has been validated and has been considered important for the analysis and not erroneous. Therefore, we do not remove the outliers in this analysis.

- **Missing Values:** Now we handle the missing values which is indicated as Nan. Missing values are first detected with the appropriate function and are replaced with more meaningful values. The missing values in Bore, Stroke, Horsepower and Peak-rpm is replaced by taking column mean value because the percentage of missing value in these columns are less than 5%. In the case of Normalized losses there are 44 missing values which is about 18.48% of the observations in the column. Directly replacing the missing value with the mean value in this case can lead to false estimation in the model in the future and can also lead to biased dataset. We overcome the missing value in this case by taking the standard normalized loss mean of the manufacture using the group by and transform function. Lastly the missing values in the qualitative column of Number of doors is done by replacing it with the mode value.

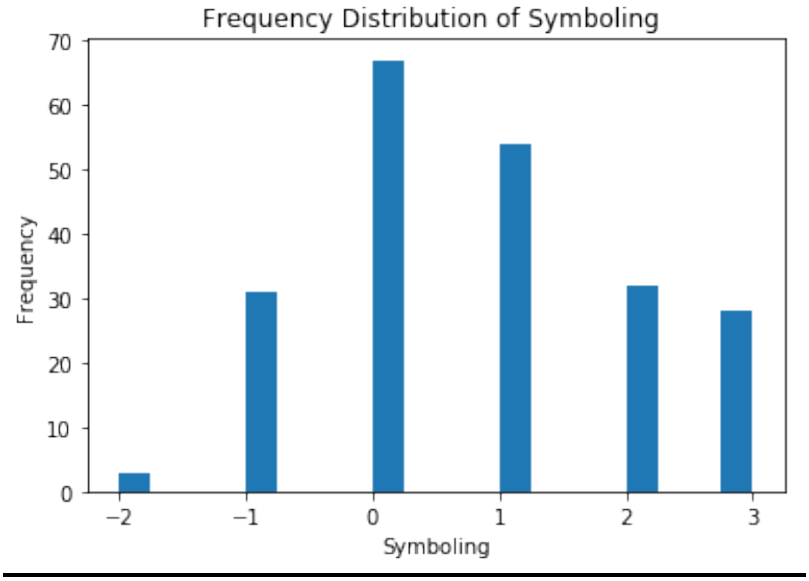
**DATA EXPLORATION**:

**Nominal Value - [Body Style]**



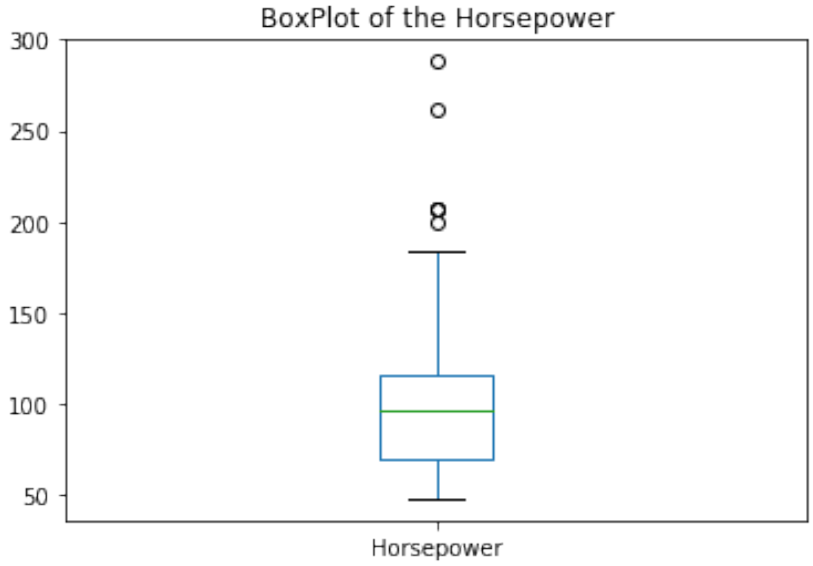

The pie chart is used to represents the nominal variable. It shows the distribution of different body styles in vehicles in the dataset. From the pie chart it is clear that the majority of the body style in the automobile industry is the sedan and the hatchback covering over 80% in the representation. A wagon body-style accounts for about the 11.63% and the hardtop for about 3.72%. The least style present is the convertible which accounts for about 2.79% which is considerate, considering its price and demand in the industry.

**Ordinal Value – [Symboling]**




Symboling is the ordinal variables selected as it indicates the level of safety (+3 indicates high risk and - 3 indicates safe). The bar chart represents the frequency of level of safety. Considering the level of safety, it is surprising to observe no car with complete level of safety. However, the maximum number of cars seem to listed in-between at zero which is about 65 cars. The next majority seem to be listed at 1 with about 55 automobiles and about 3 cars seem to have a maximum safety level of -2.

**Nominal Value – [ Horse Power]**

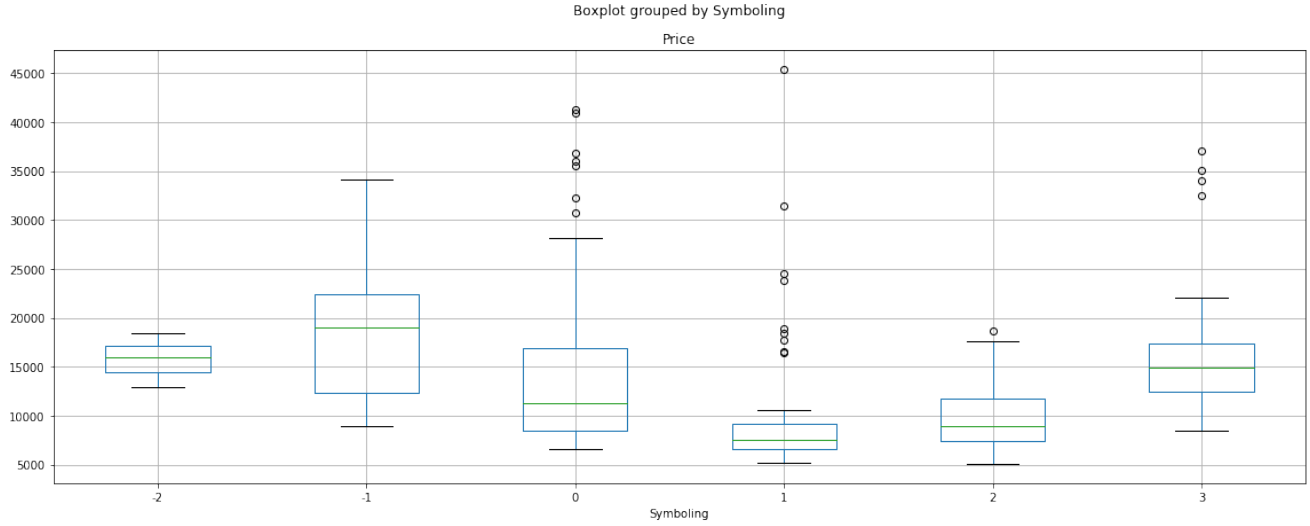



We use a boxplot to represent the nominal variable (horsepower) as it gives us a better understanding on its descriptive statistics. We see the median value of the horsepower to be 97hp, 1st Quartile is at 70hp and the 3rd Quartile is at 116hp. We see about 4 outliers in the box-plot.

## 2. Identifying the relationship between Columns

### Price vs Symboling

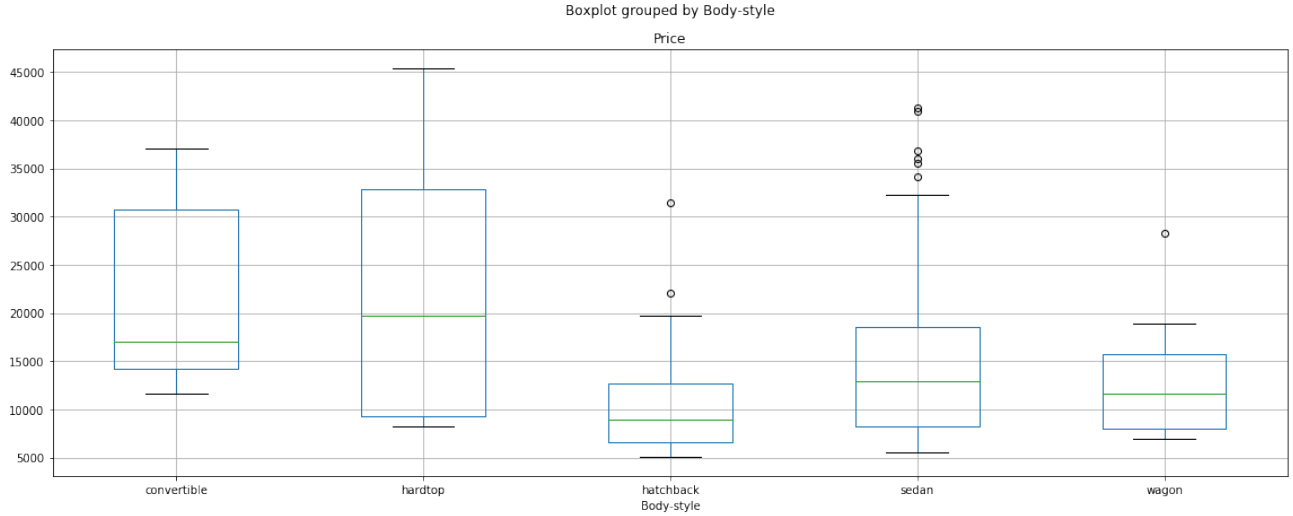
**Hypothesis:** High priced vehicles have high safety standards.




We use a boxplot to see the relationship between the ordinal and numerical variable. As we can clearly observe that the median price for the symbolling -1 is the highest and we don't see any significant relationship among the plots we can say that the hypothesis is not true. Higher median level for high risk cars in comparison to 0,1,2 symboling values also prove the hypothesis to be flawed.

### Price vs Body Style

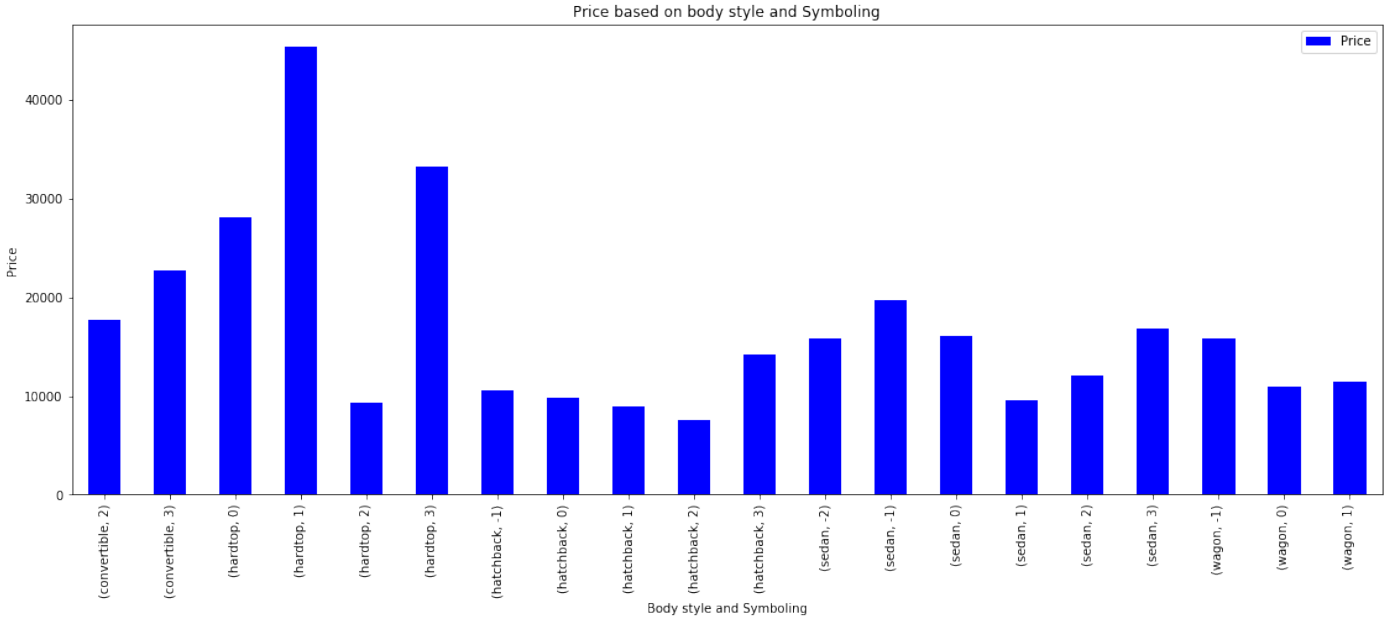
**Hypothesis:** Convertible Vehicles are the most expensive.




As we know convertible cars are cars whose roof line can be removed and refitted as required. As it is normally present in high end luxury vehicles we consider the price of this body type to higher that the other type of vehicles.
We use a boxplot to confirm this hypothesis as it provides a better understanding of the summary statistics of the different type. From the plot we see that the median price for hardtop to be the highest which is then followed by convertible. In this case the hypothesis is again flawed however deeper analysis is needed to be considered in this case as many luxury vehicles type now days have a combination of both convertible and hardtop body type.

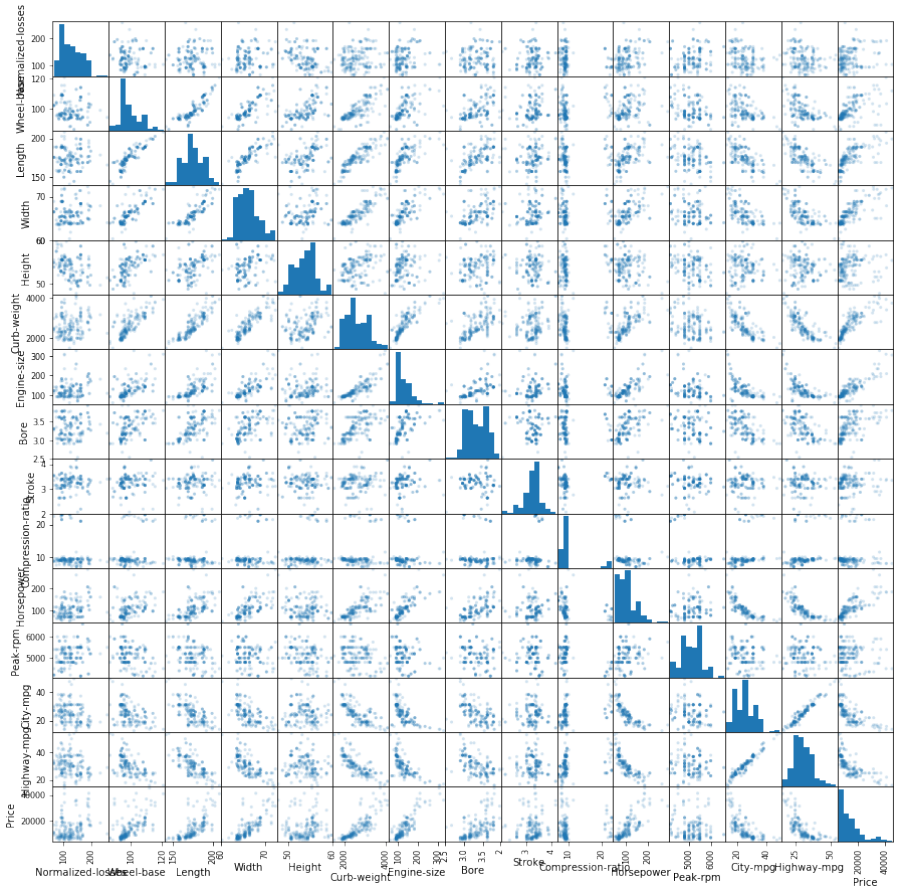
**Price vs Body Style and Symboling**

**Hypothesis:** Sedan is the best priced Vehicle in terms of Safety.




As we know that the price is based on the level of safety and the body style. Here, we observe that a fairly priced sedan provides the best safety standards in the graph therefore we can say that the hypothesis is true. However, this also signifies that the prices are not proportional to the safety standards

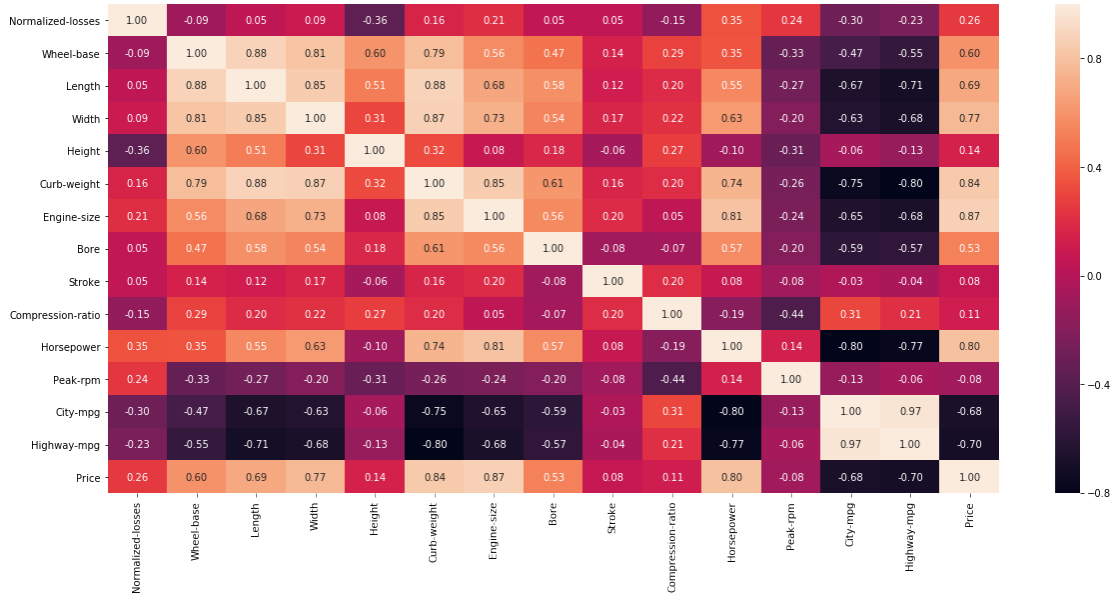
**3.Scatter Matrix of all the Numerical Variables**




A scatter plot matrix is used to represent the relationship between all the numerical variables in the dataset.

We can observe the following for the scatter plot:

- Peak-rpm show no correlation with the other parameters as the plot is randomly distributed.
- Compression Ratio also seem independent to other variables as we always see a low compression ratio value.
- We observe high positive correlation between City-mpg and Highway-mpg.
- We see positive correlation between the parameters length, width, height, Wheelbase, Curb-weight.
- Price show positive correlation between all the parameters except the mpg values where the correlation is low-negative.
- The histogram in the diagonal of the matrix represents the frequency distribution of the variable.
- I have plotted a heat map as well to provide a better understanding.




**References:**


- RMIT Course Material.
- https://www.tutorialspoint.com/python/python_matplotlib.htm
- https://en.wikipedia.org/wiki/Main_Page
- https://www.saedsayad.com/categorical_categorical.htm
- http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/DataPresentation/DataPresentation7.html