Report

Data Preparation

Preparing the data Automobile.csv was done using Jupyter Notebook where the data was about different fields of an automobile. Preparing the data would help cleanse and develop the data in such a way that it would help to obtain the required output and a truthful analysis. Working with the data to strip white spaces, typos and other errors helps to acquire strong representation and gives effective results *(Ren 2019)*.

Data preparation includes eliminating duplicates, white spaces, data entry error, null values, missing values, outliers and sanity checks. Data was carefully considered, with column by column checks to obtain any error that would effect the understanding of the data. They were dealt with different methods accordingly.

Duplicates

The dataset consisted of many duplicate rows when it was imported. The repetition of rows can cause incorrect statistics and erroneous records. Maintaining a good data quality would provide confidence and assurance towards the final analysis *(Cheng, n.d.)*. This would cause a biased analysis and would affect the conclusion. At the beginning of the preparation process there were 23 duplicate rows that were found and dealt with. These rows were deleted so as to satisfy requirements. After cleaning and correcting many other errors the dataset was subjected to check for more duplicate values and was found to have 6 more rows. They were later deleted using drop_duplicates() for consistent data.

White Spaces

White spaces are unwanted spaces present in your data that needs to be cleaned up. When dealing with a large dataset having a white space can affect your calculations and lead to unusual observations. Striping of white spaces helps to reduce redundancy of data and obtain an ideal dataset. The automobile dataset consisted of columns that had white spaces which had to be stripped using strip function strip() to obtain distinct values. After stripping the white spaces the dataset had no redundant data and later providing consistency.

Data Entry Errors

A large dataset such as automobile had many data entry errors that had to be dealt with. Typos were found in large numbers and had to be replaced with valid data. The value count of the columns were taken and analysed. Any data that appeared to be a typo was changed to the next immediate valid alternative that was compatible. This helped reduce mismatches and structure the dataset.

NaN and Missing Values

Missing data in a dataset can be due to loss of information or data not available for analysis. Missing values can cause bias towards the final conclusion and lead to invalidate study (Frisell, 2016). When missing values are not dealt properly it might cause inaccurate assumptions. There are many ways in which they are dealt with, and some of them are by omitting the values, setting values to null, imputing static values such as mean or 0, imputing the value from an estimated or theoretical distribution, modelling the value(nondependent) (PDS Teaching Team 2019). In the case of automobile dataset imputed static mean value was taken. Masks were used to filter out columns using relevant conditions. Missing values was filled by computing the mean of the respective column by applying masks that helped filter out columns of relevance based on certain conditions (Ren 2019). Therefore missing values were dealt with masks.

Outliers

Outliers are observations that lie at the far end from other values and provide different logic (PDS Teaching Team 2019). They can cause drastic changes to the analysis. In the automobile data set there were outliers that were present in many columns that had to be dealt with. They were found using box-plots from column to column, where it was found (using univariate analysis) that 7 columns included outliers that were found due to variance in the data. Variances these small did not make a huge difference to the expected output, and so these 7 were later ignored. There were 4 other columns that were found to have outliers with the help of bivariate analysis. Bivariate analysis discover association between two variables. Compression-ratio vary with different fuel systems and so when combined together the outliers seemed to be just minimal variance in the data (T, C.O and G.Y, 2012). Hence they were ignored. Due to the connection engine size has with both horsepower and stroke they were analysed respectively to come to the same conclusion of having variance in the data. The outliers present in these cases do not seem to be a threat towards the final analysis (Ray, 2016).

Sanity Checks

Sometimes in the dataset there might be columns that include values that have no relation and purpose. These impossible values cannot help produce an ideal analysis. In the automobile dataset we have dealt with these values using masks, replacing the 0 values in the price using the mean of the column by filtering out columns that hold relevance. This helped obtain values that gave significance to the final analysis.

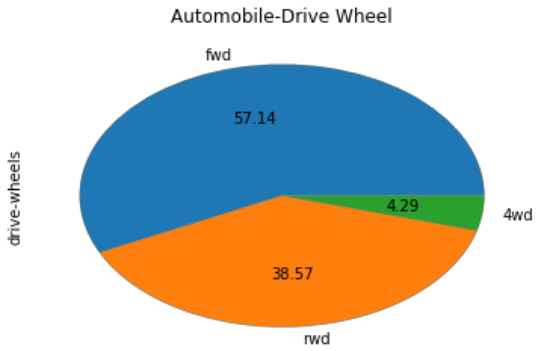
Data Exploration

Data exploration is about summarising the characteristics of a data set. It is the introduction to data analysis and help us uncover relationships between variables. Exploration of data can help find the general hypothesis the data supports (Rouse, 2019).

While dealing with the automobile dataset we used different plots to visualise and analyse different columns to create each hypothesis.

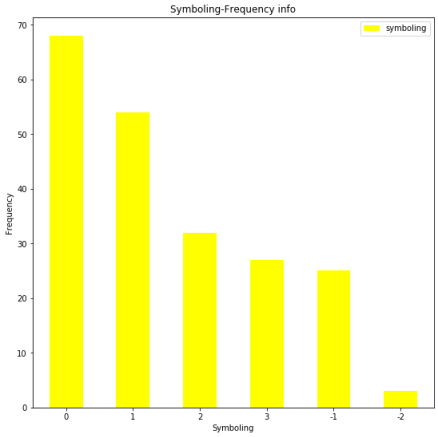
Task 2.1

Nominal Value



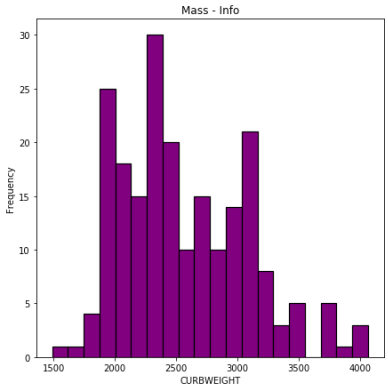
Nominal value means discrete units that have no quantitative value and order. For that reason even if there is a change in order it would not affect the meaning of that particular attribute. They cannot be quantified or handled using mathematical operators(Donges, 2018). The attribute selected for a nominal values in this particular dataset is the drive-wheel. Drive-wheel is the drive-type of the car, that is if it is a front wheel drive, back wheel drive or four wheel drive. The drive wheel consists of three categories that can be best depicted using a pie chart. The values in drive wheel was represented using a pie chart for visual comparison. It is a frequent method used to show how different parts make up the overall factor. Pie charts helps to understand the percentage of each category from the whole and that helps to enable immediate analysis of the drive wheel attribute . Using the pie chart it was easier to make the assumption that the dataset consists of more number of automobile with front wheel drive.

Ordinal Value



Ordinal values are categorical values where they are ordered. However the difference between the values does not matter. The ordinal data that was found in this particular dataset was the symboling attribute. Symboling is the insurance risk rating of an automobile. It indicates values from +3 to -3, where +3 specifies high risk and -3 safe. Each of these observations have a symboling value that indicates if the automobile has high risk rate or not. Here symboling is represented using a bar chart to picture the number of cars that fall into each state. Bar charts are easier to interpret.On using this chart we can find out how many cars belonging to the dataset have high insurance risk rating. The count of each rating can be depicted effortlessly using the bar chart. Later drawing a conclusion that more number of cars in the dataset tend to have a neutral risk rate (Bhat, 2019).

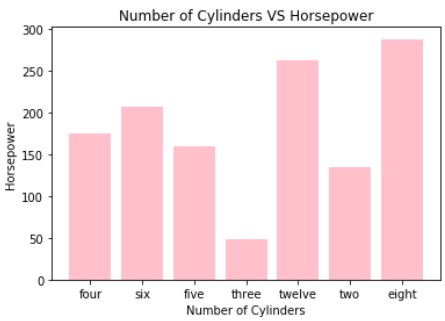
Numeric Value



Numeric value is known as measurable values. They are represented and collected in numeric form. Numeric value can perform mathematical operations. The numerical value choose for this specific dataset is the curb weight of a car which are continuous values. Curb weight is the mass of the car. Histogram was used to visualise this single numeric column with a bin size of 20. Here the frequency(count) of cars were shown on the y axis and the mass values were shown as bins in the x axis (study.com, 2019).

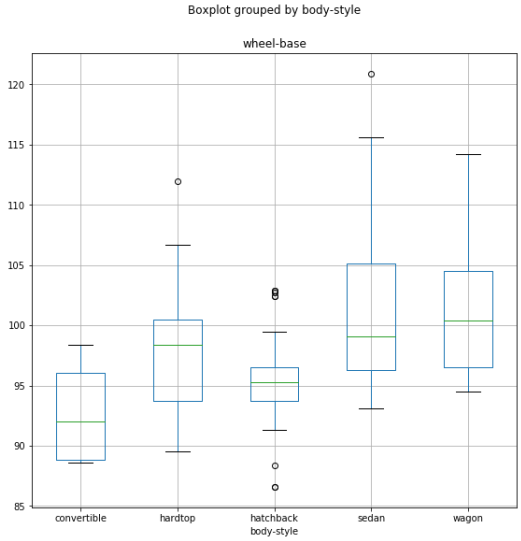
Task 2.2

Horsepower VS Number of Cylinders



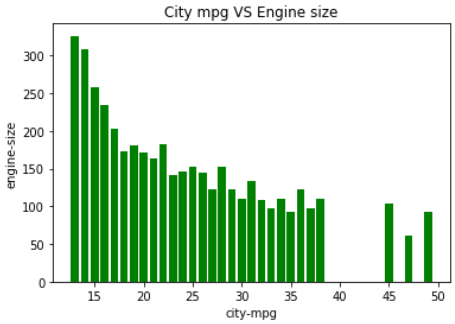
In the automobile dataset two attributes horsepower and number of cylinders were analysed together to obtain a plausible hypothesis. The reason to choose both these attributes for a relation is due to the variation it causes amongst each other. Automobiles do provide more power by the number of cylinders . To prove the hypothesis, a bar chart is plotted to understand the difference in horsepower accordingly with the number of cylinders. In this case the automobiles with 8 and 12 cylinders tend to have more horsepower than the ones that contain 2 or 3 cylinders. However, an interesting observation was found between 4 cylinder and 5 cylinder automobiles that tend to have similar horsepower. This could be a slight difference due to the data present in the dataset. The graph could be concluded by saying, more power is obtained from automobiles that have more number of cylinders thus proving the hypothesis.

Wheel base VS Body style



Automobile dataset consists of wheel base of cars that is the distance between the centre point of the front and back wheel. Comparing the wheel base to each model style of an automobile helps understand the difference of how each model style's comfort and efficient drive. A box-plot is chosen to depict the relation and prove the hypothesis between these two attributes to obtain which of these models are preferred for more comfort and efficiency. In the graph it is shown that inter quartile range of a sedan (approximately 96 -105) with a maximum value (approximately 116) is greater compared to every other model. Wagon closely behind with an inter quartile range (approximately 97-104) with maximum value(approximately 114) which prove that these two models tend to have better comfort and drive efficiency than the rest. The outliers present in the hatch back ,hardtop and sedan model does not affect the hypothesis as the values do not differ vastly. The values for hatchback model (approximately 103-104) does not change the assumption taken for the hypothesis. Thus, the graph concludes by proving the hypothesis that sedan with higher wheel base value tend to have better comfort and drive than other models.

City mpg VS Engine Size

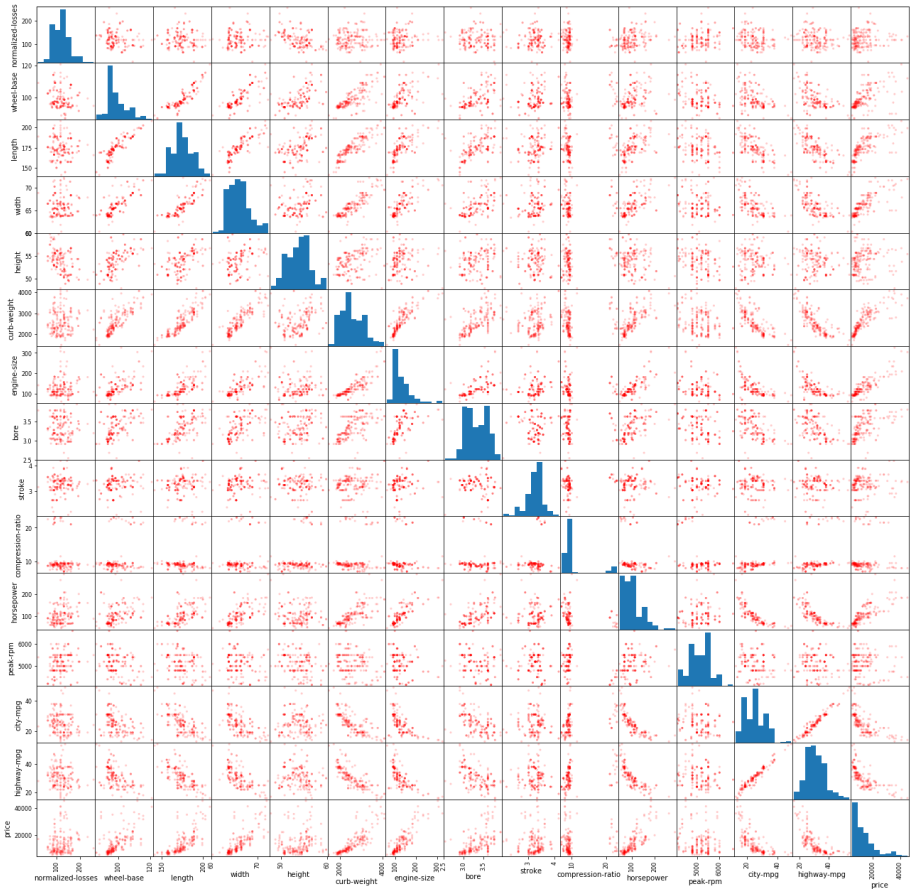


The automobile dataset has given the city mpg for every vehicle present in the dataset. The hypothesis to prove using the bar chart is that ,mileage of a car depends on the engine size. Using the chart we can show the variation of mileage to that of engine size. As the engine size decreases the city mpg increases, giving us a

graph skewed to the right. The engine size decreases towards the right while the city mpg increases. An automobile that has an engine size approximately 350 has a city mpg approximately 10, while a vehicle with engine size approximately 70 has a city mpg approximately 47.Thus the plot helps prove the hypothesis that city mpg decreases with increase in engine size.

Task 2.3

Scatter Matrix



The scatter matrix was used to build all numeric columns in the automobile dataset. There are 15 numeric columns that were plotted against each other. There are many inferences that can be obtained from the scatter matrix.

1. The wheel base of the vehicle, has a straight linear graph with length and width that proves that the wheel base varies along with length and width attribute in the dataset.
2. When compression ratio is plotted against every other attribute they come under a single value range.
3. The mass of the car is represented as the curb weight in the dataset and so it is correlated with wheel base, engine size , length and width of the automobile.
4. The engine size is dependant on the curb weight and price of the vehicles present in the dataset.
5. When normalised losses are compared to every other attribute it provides a constant and uniform graph.

Reference

1. Cheng, A. (n.d.). *The Causes, Impact and Detection of Duplicate Observations*. [online] https://www.lexjansen.com/pharmasug/1998/CODERS/CHENG.PDF. Available at: https://www.lexjansen.com/pharmasug/1998/CODERS/CHENG.PDF.
2. Frisell, T. (2016). *SP0187 Why Missing Data Is A Problem, and What You shouldn't Do To Solve It*. [online] https://ard.bmj.com/content/75/Suppl_2/45.4. Available at: https://ard.bmj.com/content/75/Suppl_2/45.4.
3. Ren, Y. (2019). Data Curation - 2019r.
4. PDS Teaching Team (2019). Practical Data Science - Tute 2/ Week 3.
5. T, A., C.O, F. and G.Y, P. (2012). Influence of compression ratio on the performance characteristics of a spark ignition engine. *Department of Mechanical Engineering, Ahmadu Bello University, Zaria, Nigeria*, [online] p. 1915. Available at: http://www.imedpub.com/articles/influence-of-compression-ratio-on-the-performance-characteristics-of-a-sparkignition-engine.pdf.
6. Ray, S. (2016). *A Comprehensive Guide to Data Exploration*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/.
7. Rouse, M. (2019). *data exploration*. [online] Search Business Analytics. Available at: https://searchbusinessanalytics.techtarget.com/definition/data-exploration.
8. Donges, N. (2018). *Data Types in Statistics*. [online] Towards Data Science. Available at: https://towardsdatascience.com/data-types-in-statistics-347e152e8bee.
9. Bhat, A. (2019). *ORDINAL DATA: DEFINITION, ANALYSIS AND EXAMPLES*. [online] QuestionPro. Available at: https://www.questionpro.com/blog/ordinal-data/.
10. study.com. (2019). *What is Numerical Data? - Definition & Examples*. [online] Available at: https://study.com/academy/lesson/what-is-numerical-data-definition-examples-quiz.html.