Practical Data Science – Assignment 1

1. Data Preparation

1.1. Background Check:

Using the data set provided, analysis into the data was performed in order to acquire background information and brief domain knowledge to properly prepare and analyse observations in each attribute. (Automobile Data Set, N/A) This data set was donated for use and values have been altered for use of Assignment 1.

The data that was within the given set encompasses attributes regarding automobiles and uses 3 entities to group the attributes. The first entity is the various characteristics of the automobile regarding physical attributes e.g. mileage, weight and price. These accompany majority of the attributes in the data set. The second entity is an attribute regarding its insurance rating and scales with its price, with values ranging from -3 to 3, with -3 being safe for its price and 3 being unsafe for its price. The third entity is an attribute called normalised losses and corresponds to the average loss payment for the automobile within its relevant grouping, classified by a multitude of physical characteristics and physical attributes of the vehicle. This third entity was interpreted as the average amount of loss payment due to issues with given automobiles with the same classification.

With basic knowledge based on the data set and what the data in all attributes convey, concrete attribute information was required to ensure the data set only contained possible data for their relevant attributes. (Auto Reports Database, 1985) This report where the data set was initially published held the attribute ranges of each attribute within the given data set. Using the given attribute all values the data will be checked and if any data violates the attribute range, it will be corrected but if not possible, will be removed and ignored from the analysis of the data set.

1.2. Loading Data

A quick analysis on the file structure of the automobile.csv file containing the dataset revealed that the data was structured with observations as rows and each attribute in every observation is separated using a '#' character also known as a delimiter. As the dataset was downloaded the data location was declared using simple navigation to the automobile.csv file and then read into a dataFrame data structure (a readable table for code) using the pandas library in python with relevant parameters such as the delimiting character in the .csv file, and an array (a list data structure) containing all attributes in order from left to right, serving as the columns of the data in the dataFrame.

An initial check was required to ensure the data loaded into the dataFrame was correct in terms of being the same data to that of the .csv file. This check was performed using a check in datatypes of each attribute or column in the dataFrame. After viewing and comparing the data types outputted and the actual data types that the dataset used for each attribute, it was concluded that the loaded data was formatted but was still the same to the original data loaded from the .csv file.

1.3. Data Checking and correction

The overall order of which possible errors were checked and handled was determined by their overall affect on the remaining data and the amount they would decrease the analysis of following checks. As redundant whitespace would increase manual analysis of possible impossible values or data entry errors to correct, it was the first to be corrected. No manual analysis of whether the whitespace was needed in the data was done due to previous analysis of attribute ranges hence code was used to completely remove all whitespace in the entire dataset without checking. Similarly, capital letter mismatches would create more work when analyzing potential

impossible values or data entry errors to correct due to there being more unique values. This was also not checked due to capital letter mismatches being easy to fix. This correction was performed by changing all capital letters within the dataset were turned into lower case letters.

The first data attributes to be checked were attributes using nominal data. This decision was based on the lack of mean and medians in nominal data making them easier to handle as correcting errors would not largely affect the greater data set. With code that counts the amount a unique nominal value was used for an attribute, the output displayed each unique nominal value used within the chosen attribute. Using this code for each attribute with nominal values, it was simple to find any errors within the attributes.

All errors discovered using this method were interpreted as typos due to the high possibility of the errors being data entry errors. Errors found were: Attribute[Make] “vol00112”, Attribute[Aspiration] “turrrbo”, Attribute[NumOfDoors] “four”. These errors are clearly data entry errors and was fixed by going to each attribute and replacing all values with the data entry error into the intended value, e.g. all values holding “vol00112” in Attribute[Make] was changed to “volvo”. With the background knowledge that Attribute[Symboling] should only contain discrete values between -3 and 3, it was decided that the same method of data checking would be used as with the other nominal data. It was discovered that there were three values of “4” within the attribute and it could have been interpreted as an impossible value but also as a data entry error due to the value being close to a valid value. It was decided that these values would be considered as a data entry error for said reason and the additional reason of preventing data loss. All data with the value “4” within Attribute[Symboling] was changed to value “3”.

Sanity checks for impossible values were then checked, as the errors could still be fixed if it was clear that the impossible value was due to a data entry error. These sanity checks were carried out through all remaining attributes individually as the remaining datasets all contained continuous values and or large ranges. Using the attribute ranges, code was constructed for each individual attribute which displayed the impossible value if it was outside the range of the respective attribute. When a value was displayed it was analysed if it was a data entry error by establishing whether it was close to the range or a simple conversion error, and if neither, the values were set to NaN to be dealt with later as the correct value could not be determined.

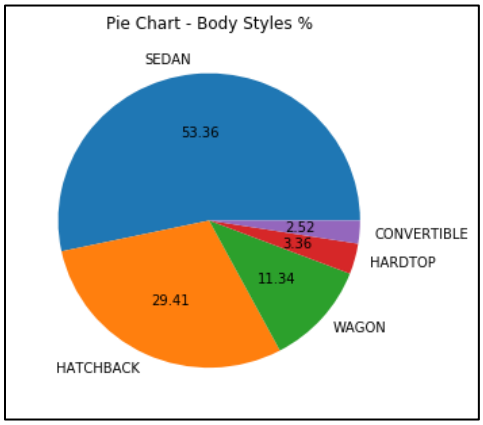
Missing values or NaN values for numerical data were the next the be checked and corrected and is last due to missing values being unable to be corrected but estimated hence skewing the overall dataset slightly. Searching through each numerical attribute, the code outputted the amount of NaN values within the attribute. Attributes with missing values: normalised losses, bore, stroke, horsepower, peak RPM, price. If the amount of the missing values in an attribute were less than 20% of the amount of observations (21 values) then a mean or median was imputed depending on the attribute. All attributes other than normalised losses contained less than 21 missing values hence medians were imputed to the missing values due to the possibility of large outliers which should not be changed in the dataset but simply ignored as a range has already been given, meaning the outlier is still valid data which can and should be interpreted.

Handling the large number of missing values in the normalised losses attribute is slightly more complicated as if the mean or median was the data imputed it would alter the overall data for normalised losses significantly due to the large amount of missing values. To fix this data was imputed from an estimation due to the analysed relationship between the symbolling and normalised losses attributes. Grouping all observations using their symbolling values, the mean value of normalised losses in that symbolling group was imputed into the missing values of the normalised losses to their respective symbolling group. This reduced the amount that the data was skewed as the missing data was more accurately predicted from a relationship within the attributes.

The final check is towards missing values within all categorical attributes. The same check for missing values was performed to the categorical attributes with the only missing values being two values within the number of doors. As this attribute in truth is a discrete numerical value with only two possible values, the error is corrected by imputing the highest frequency value of the two possible values into the missing values.

2. Data Exploration

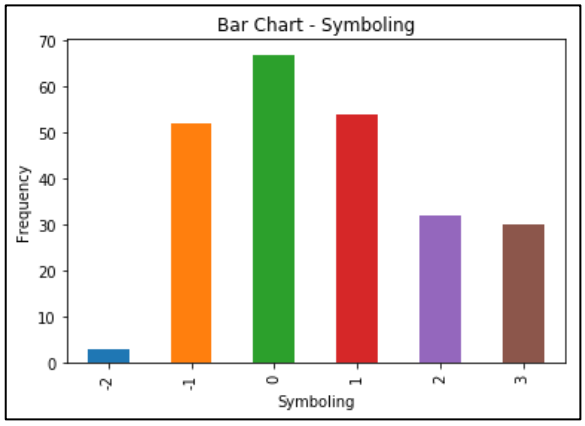
2.1. Nominal, ordinal, numerical graphs and exploration



To effectively represent the comparison between the chosen column of body styles, a pie chart was chosen. The decision was based on the desire to compare the frequency of one body style to another. Doing a comparison with a pie chart displays how one value measures up in terms of frequency to another. Additionally, by displaying the percentage of all values a certain value takes up, the graph presents an accurate and concrete value to collect and manipulate.

With the column of body styles, the sedan is displayed to be over 50% of all the observation body styles. Convertibles and hardtop automobiles on the other hand only represent a small portion of the data set. Using this knowledge, it is presumed that sedans are the most popular automobile choice followed by hatchbacks and wagons, whilst convertibles and hardtops are not popular.

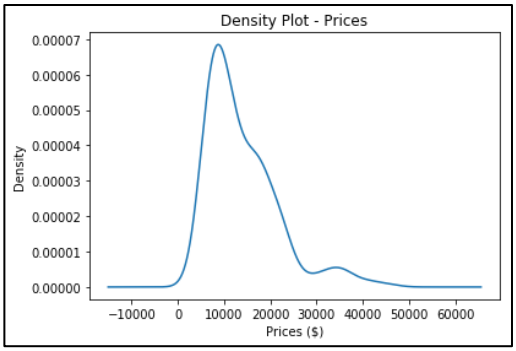
Using a bar chart creates an effective visualisation to compare frequencies but also see the frequencies change across a set of ordered unique values. With symbolling using a rating between -3 and 3 it is classified as ordinal data, hence using a bar chart enables for a visualisation of the most frequently used value. In addition, the visualisation also displays the overall change in frequency between one value to another, hence showing the change in frequency within the order of set values.



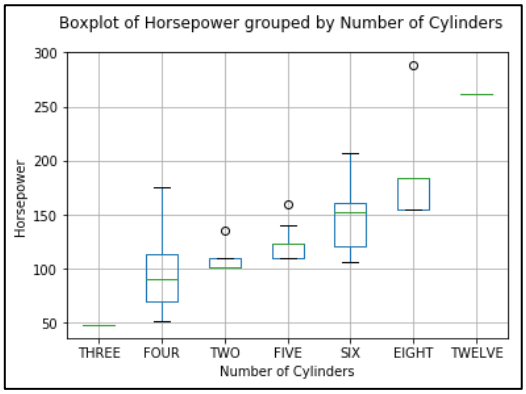
In this visualisation of the symbolling column, the most frequent ranking within the data set is 0 meaning it has a moderate insurance risk rating. The frequency of values peak at 0 and taper off to -1 and 1 meaning that the automobiles which had their data collected were mostly moderately safe give or take. Although there is a considerable difference between the higher and lower ends of the scale, with 2 and 3 having a considerably larger frequency than -2 and the non-existent -3. This shows that from the observations collected, there are more automobiles considered a higher risk than a low risk.

To effectively display the continuous numerical data of column prices, a density plot was used. This decision was due to the desire to visualize the amount of values revolving around a price range. Using a density plot for prices shows the peaks and troughs of prices that automobiles are sold at.

Within this visualisation, most automobile prices are around $10,000 and slowly declines afterwards. This slow decline shows how the observed automobiles are more likely to have a higher price of over $10,000 rather than a lower price due to its steep incline in density and slow decline after the peak. Another interesting aspect of the visualisation shows a small peak at around $35,000. This spike could be due to multiple other factors regarding the automobile, although it is certain that an attribute is affecting the prices of automobiles creating a second spike.



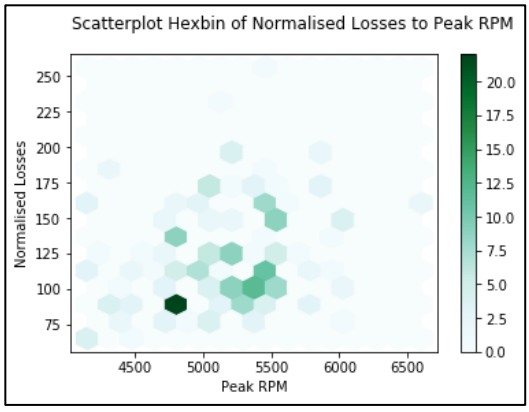
2.2. Relationships between columns of the dataset



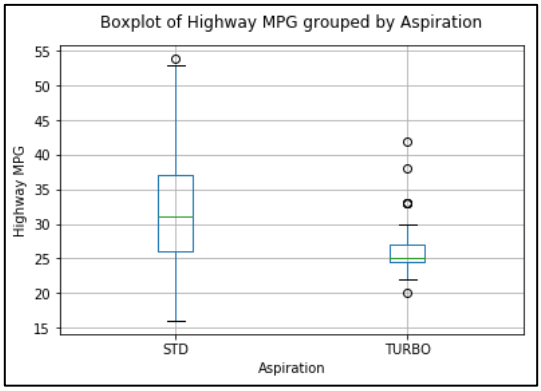
(How number of cylinders affect performance, 2015) Using information gathered, it is presumed that the number of cylinders affects the amount of horsepower the automobile can output. This assumption came from each cylinder each outputting a torque to the overall output of the engine measured as the horsepower, hence more cylinders should result in a higher horsepower.

Analysing this visualisation of horsepower grouped by number of cylinders, the hypothesis is considered false. The order of three, four and two cylinders should be in ascending order hence the overarching hypothesis can not be verified as true with the data set. It can be presumed that the hypothesis only holds for cylinders above the amount of five, although it can also be assumed that another attribute is altering the amount of horsepower for all observations with cylinders under five. Although looking at the overall data set, it is presumed that there is insufficient data held for observations with three and two cylinders, possible giving only extreme outlying horsepower values for that given number of cylinders but is unverifiable without more data.

(How does RPM affect the Engine, 2016) Using previous and acquired background knowledge on the mechanics of an engine, it is hypothesised that a higher RPM will cause more damage to the engine resulting in higher maintenance cost and normalised loss. This is due to a higher RPM meaning more explosions in the engine per minute, wearing down the structural integrity of the engine. Although, engines that have a higher RPM are expected to have sturdier engines to reduce wear. It is expected that a there should be a group of observations with moderate RPM with high normalised losses and a group of high peak RPM observations with low normalization losses.



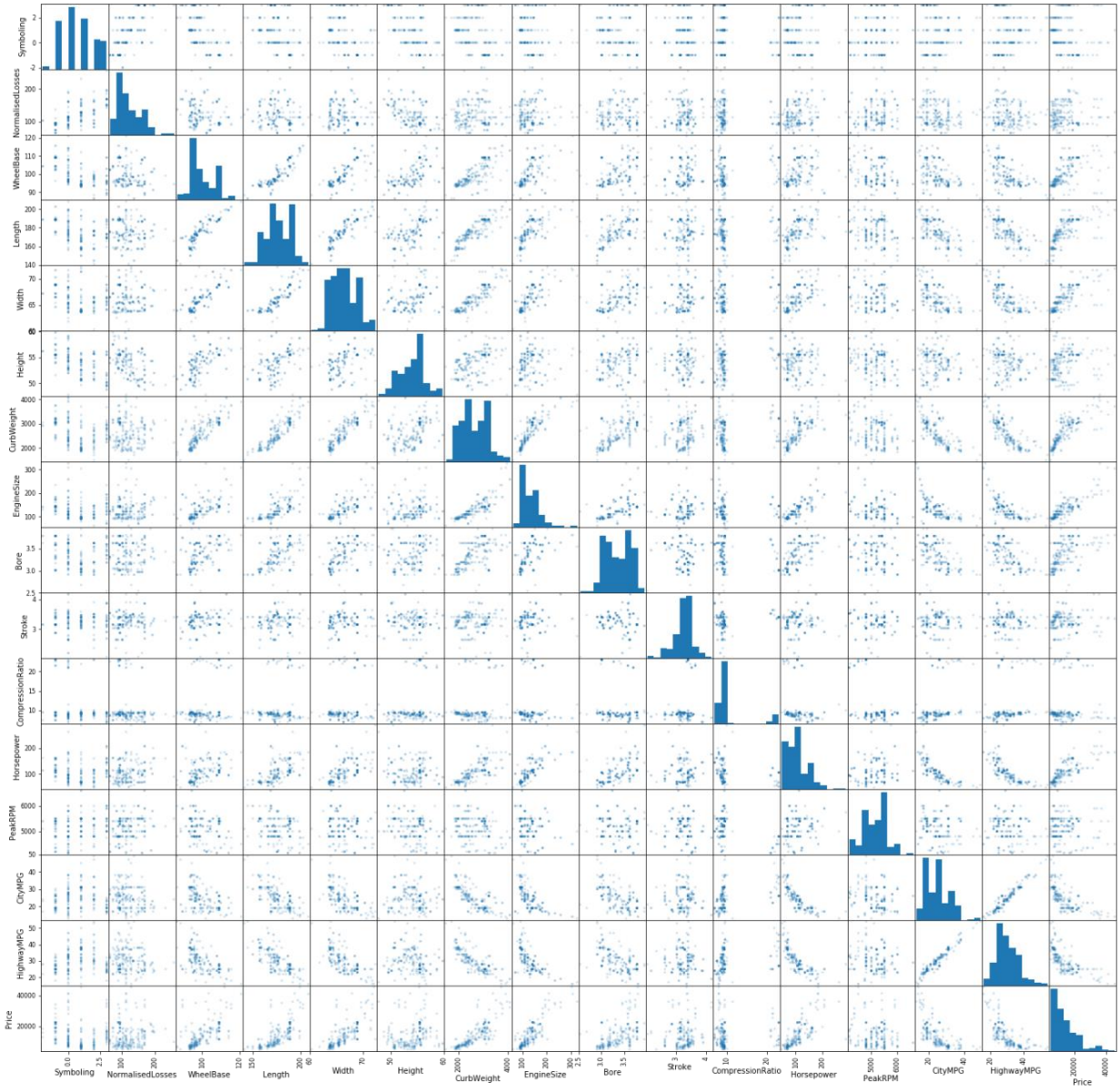
Observing the hexbin visualisation the overarching hypothesis is proven false. This is due to the visualisation not displaying two small clusters, but one irrelevant large cluster. Using the dataset, the only distinguishable fact between the relation of normalised losses and peak RPM is that there is no relationship between them. It can be argued that the basis hypothesis that higher RPM results in larger wear on the engine can still be valid. This argument is due to the drivers not using the automobile at the peak RPM, making it so the peak RPM does not become a factor when comparing to the use of the automobile and in this case normalised losses of the vehicle.



Using background knowledge in engines it is presumed and hypothesised that a standard aspiration can have higher overall mileage to that of a turbo. This is due to turbo aspirations being used to force air into the engine to increase the RPM and hence amount of fuel used.

With the visualisation it can be verified that turbo aspirations have a lower overall mileage is true, supporting the hypothesis. Observing the graph, standard aspirations has a smaller lowest value, although this can be due to the large amount of observations with a standard aspiration. Regardless turbo aspiration highest value falls below the mean and within or lower than the 25th percentile of the standard aspiration, hence proving the hypothesis.

2.3. Analyse the scatter matrix



On a standard practice, a scatter matrix should only be used when handling a small number of variables otherwise the graph will clutter and there can be potential issues conveying relationships. Although, there is a requirement within the assignment that requires the creation of a scatter matrix of all numerical values.

Viewing all relationships within prices, there is a clear correlation between prices and all size attributes associated with automobiles. These size attributes include length, wheel base, width, height, curb weight, engine size, (Bore and Stroke, 2015) bore. This supports the assumption that larger automobiles require more materials and hence a larger price.

In terms of overall MPG of the automobile, it seems to also be linked directly to the overall size of the automobile. For all the previously mentioned size attributes, as each attribute increases the overall millage decreases for the vehicle due to the increase in weight, supporting the idea that the heavier the object the harder to move the object. Incidentally, as lower priced vehicles are smaller, these cheaper vehicles tend to have larger MPG.

As horsepower increases, highway and city MPG decreases. This result can be derived from the previously hypothesised relationship between horsepower and number of cylinders where horsepower increases with the number of cylinders. With an increase of cylinders, an increase of fuel is required at any given time resulting in lower MPG. Conclusively a larger horsepower tends to result in more cylinders and hence more fuel needed and lower MPG. This is connection is also further cemented with the larger engine size due to the presumed larger number of cylinders.

Surprisingly there are no relationships found with any of the numerical values between the normalised losses attribute. It was expected multiple aspects would affect the size of normalised losses, e.g. a higher RPM would result in higher normalised losses due to possibility of crashes. Although, considering that all variables expected to affect normalised losses would result in crashes rather than natural automobile failure, data from crashed automobiles would not be collected. Considering this possibility, the dataset would not contain data that would show a relationship between normalised losses and other attributes.

References:

- Michael Galarnyk. (2018). *Understanding Box Plots.* Available: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51. Last accessed 08/04/2019.
- Caitlin Garrett. (2014). *Handling Outliers.* Available: https://www.rapidinsight.com/handle-outliers/. Last accessed 08/04/2019.
- Graham Williams. (2014). *Mean/Median/Mode.* Available: http://datamining.togaware.com/survivor/Mean_Median_Mode.html. Last accessed 08/04/2019.
- Jeffrey C. Schlimmer. (1985). *1985 Auto Imports Database.* Available: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names. Last accessed 08/04/2019.
- Jay Safford. (2015). *How number of cylinders affect performance.*Available: https://www.yourmechanic.com/question/how-do-the-cylinder-numbers-impact-vehicle-performance-and-reliability. Last accessed 08/04/2019.
- Sunil Kumar. (2004). *How does a high RPM affect the engine.*Available: https://www.bobistheoilguy.com/forums/ubbthreads.php?ubb=showflat&Number=644131. Last accessed 08/04/2019.
- Unknown. (2015). Bore and Stroke. Available: https://www.grc.nasa.gov/www/k-12/airplane/stroke.html Last accessed 07/04/2019.