Report: Assignment 1, Practical Data Science

Name: Akshay Salunke

Student Number: s3730440

Task 1: Data Preparation

The report was run on IPython in the jupyter environment. The report shows the Data Preparation and Data Exploration steps of the Automobile dataset created/donated to UCI repository by (Schlimmer , 1985).The csv file was loaded using pandas and the loaded file was closely observed and compared to the original csv. The original csv was pivoted and those values were checked with the loaded count_values() for each column. The top rows were displayed using the head function, the datatype was checked using dtypes function. The dimensions of the data was checked using the shape function which showed that there were 238 rows and 26 columns in the dataset (Week 2 lecture notes 2019).

Typos

There were two kinds of typos- one where the spelling was incorrect, other where the case was incorrect. Functions replace() and lower() were used to replace spelling errors in columns like make, aspiration and num-of-doors and case was changed to lower case in columns like make, num-of-doors and engine-location. The lower() function was added to all string columns to make sure we cover any missed values as well (Week 2 Lecture notes 2019).

Extra Whitespaces

Extra whitespaces were removed using the strip() function. Whitespaces were seen in columns like make for values volvo but there could be other undetected whitespaces as well and hence the strip() function is applied to all string columns.

Sanity Checks for impossible values

Columns like symboling, normalized-losses and price had impossible values. For all the columns the minimum and maximum value was checked to ensure that the column has values within the range. For example, the value for price was 0 for less than 2% of the data and hence they were excluded. The range of the column was specified using the isin() and range() functions. The range() function takes only integer values and the upper limit is not included which is the reason the upper limit was incremented by one while using the function.

Missing Values

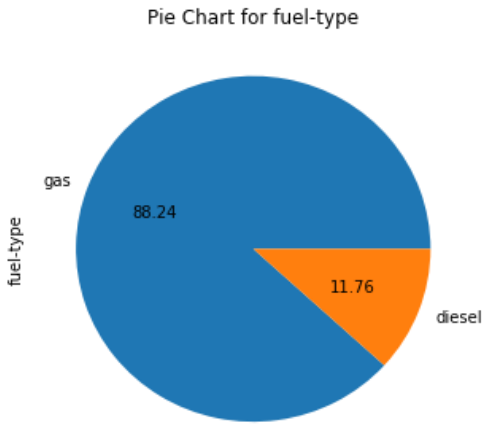
Missing values were treated in two ways- one way is where NAs were excluded and the other way where mean/median was imputed. The NAs were excluded in columns like num-of-doors, bore, stroke, horsepower, peak-rpm, price as these columns contained less than 2% NA values. The column normalized-losses contained around 20% NAs and hence exclusion was not an option, imputing with mean/median was another option. Median was chosen for imputation as the median is preferred for skewed values(Week 2 lecture notes 2019). The data was right skewed which justified the use of median.

Task 2: Data Exploration

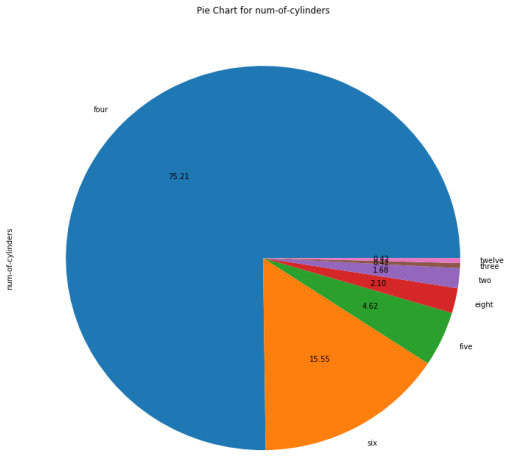
Subsection 1: A nominal, ordinal and numerical value was fuel type, number of cylinders and price respectively.

Pie chart was used to interpret the nominal variable fuel-type. About 88% of the cars had fuel-type gas and around 12% cars had diesel as fuel type. Pie chart is a good representation for a nominal variable (Week 3, lecture notes 2019).

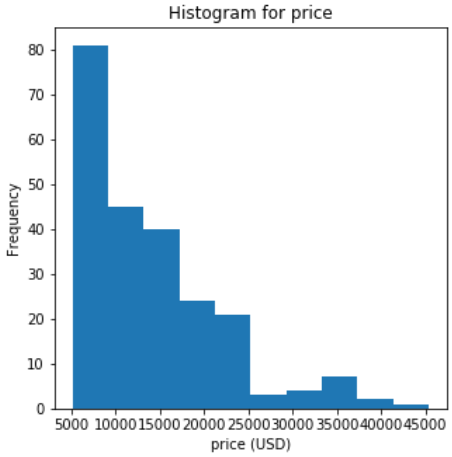
.



Pie chart was used to interpret the ordinal variable num-of-cylinders. About 75% of the cars had four cylinders and around 16% cars had six cylinders. Pie chart is a good representation for a nominal variable(Week 3, lecture notes 2019).



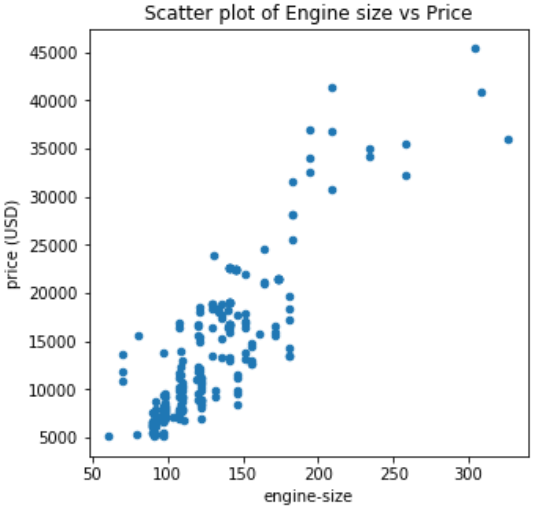
Histogram was used to interpret the numerical variable price. More than 80 cars out of 238 which was nearly 34% of the data had the price ranging from 5000-10000 (USD). The number of cars reduced as the price increased as was visible in the histogram(Tutorial 3, 2019).



Subsection 2: One scatter plot and two boxplots and were made for this section.

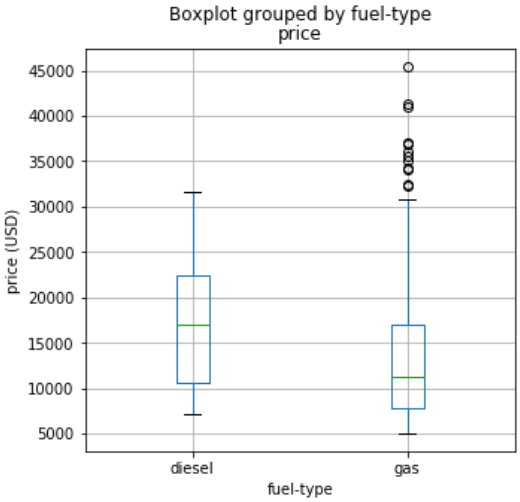
Engine-size and price were used to make the scatterplot. The relationship between two numerical variables was shown. The hypothesis to be tested was: The price of cars decreases as the engine size decreases.

- There was a positive relationship between engine-size and price. As engine-size decreased the price decreased as well. A positive correlation was seen(Week 3, lecture notes 2019)..



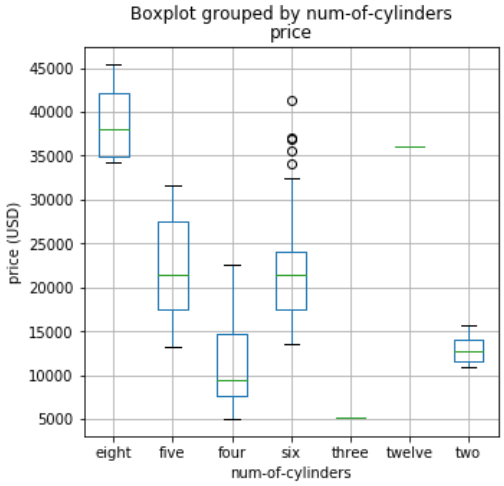
Fuel-type and price were used for the first boxplot. The hypothesis to be tested was: The price for gas cars is less than the price of cars with fuel type diesel. In the first subsection, two individual graphs were made for fuel-type and price and these two attributes were chosen to understand the relationship between them. This explained the relation between a numerical variable and a nominal variable.

- The cars using diesel turned out to be in the higher price range whereas those using gas were less costly.
- Some gas using cars were found to be costly and that may be due to some other factors like the make or the engine of the car which was not captured in the below plot.



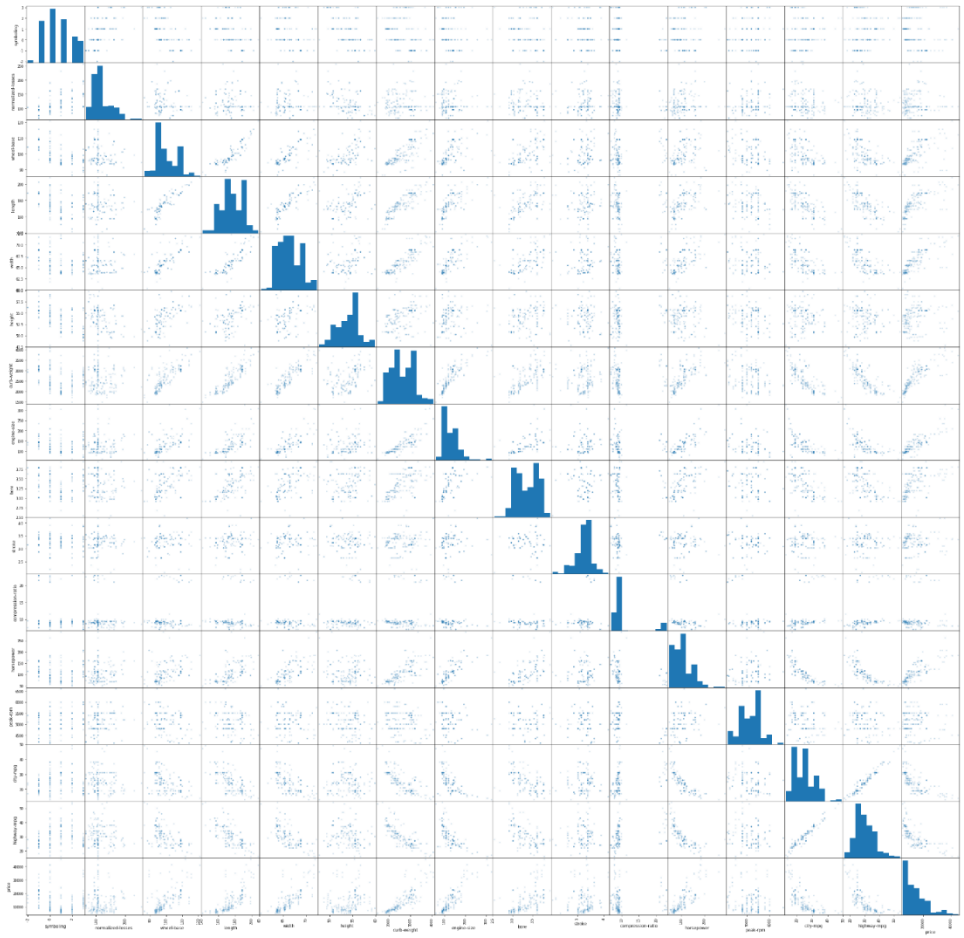
Num-of-cylinders and price were used for the second boxplot. The hypothesis to be tested was: The price for eight cylinder cars is more than the price of cars with four cylinders. In the first subsection, two individual graphs were made for num-of-cylinders and price and hence these two attributes were chosen to understand the relationship between them. This explained the relationship between an ordinal variable and a numerical variable.

- The costliest cars had eight cylinders.
- The cheapest cars had 4 cylinders.



Subsection 3: A scatter matrix for all numerical columns was plotted.

- The mileage decreased as curb weight increased.
- The curb weight increased with the engine size.
- As mileage increased engine size decreased, shown by the above two points and from the scatter matrix as well.
- curb-weight, engine size, length, horsepower and weight were seen to be positively related with price. As the engine- size increased price increased as well.
- city-mpg and highway mpg are negatively corelated with price. As price increased the city mpg and highway mpg decreased.
- As the horsepower increased the city mpg and highway mpg decreased. The mileage decreased but the price increased.
- This also showed that the horsepower increased with the price.



References

Schlimmer , J, C, 1985, *Automobile Data Set*, viewed 14 April 2019, <http://archive.ics.uci.edu/ml/datasets/Automobile?ref=datanews.io>

Dr. Yongli Ren; 2019,'Practical Data Science: Data Curation', PowerPoint slides, COSC 2670, RMIT University, Melbourne.

Dr. Yongli Ren; 2019,' Practical Data Science: Descriptive Statistics and Visualisation', PowerPoint slides, COSC 2670, RMIT University, Melbourne.

Tutorial 2, Tutorial 3; 2019,' Practical Data Science', RMIT University, Melbourne.