PYTHON FOR PRACTICAL DATASCIENCE – PART 1

S3702041

Ritwick Dev

2019

**Task 1: Data Preparation**

In this report we will see how examine the Automobile dataset created/donated to UCI repository by (Schlimmer, 1985) using the preliminary steps of Data science process. The code output generated is from version2.7 iPython using juypter notebook.

The comma separated values (.csv file) was imported using the pandas library and column names were assigned. Using the head function, we appropriated the top column as header. Validation steps were taken to check if the csv file importation was successful with all observations accounted for there were 238 rows and 26 columns in the csv file for the assignment as well as the loaded one. Data types of all the variables were checked using function dtypes.

* **Typos**

Using functions like replace() and lower() we were able to handle typos. Two kinds of typos were encountered-

* one where the spelling was inconsistent and
* other were the casing (lower-case/upper-case) was different.

The columns with inconsistent spellings were make, aspiration and num-of-doors whereas the columns with different case alphabets were make, num-of-doors and engine-location. In accordance with tutorial 2, lower() function was used on all string columns to avoid any unacknowledged alphabet in upper case

*Week 2 Lecture notes 2019*

* **Extra Whitespaces**

Using strip() function, extra whitespaces were rendered. Although, clear extra whitespaces were clearly observed in some columns like “make”, but the strip() function was used on all string columns in order to avoid whitespaces in the data.

* **Sanity Checks for impossible values**

Symboling, Normalized-losses and Price are some of the columns that had impossible values. For instance, range for symbolling is [-3 , 3] but we saw value of 4 while using the function value\_counts().

Numerical variables were checked to confer within that all the values fall within the range. The impossible values were excluded for the data set as those values would have just acted as noise in the data analysis and would not have added value in the data science process ahead. Using isin() and range() function impossible instances were omitted. (Week 2 lecture notes 2019).

* **Missing Values**

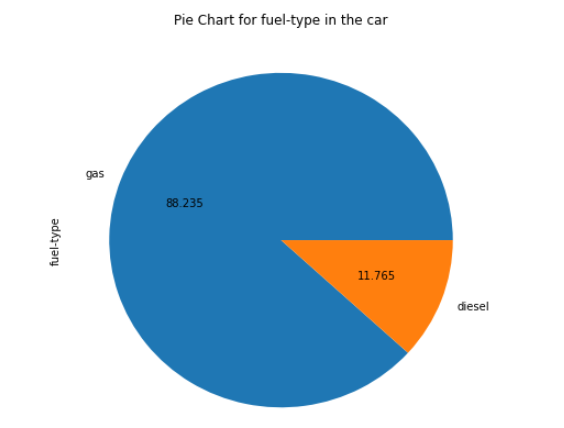
Missing values treatment was done to handle all the NA, NaN, etc in various columns across data set. Missing values were handled either by excluding those values while performing statistical analysis or imputations were performed on those values like mean, median & mode based on the type of variable in picture and the number of NA values as well. Variables like normalized-losses, bore, stroke, horsepower & price had missing values in them. Except normalized-losses, all other columns had less than 2% of missing values (i.e insignificant amount of missing values and can be excluded from analysis). Using dropna() function all those NA values were excluded. For the normalized-losses variable there were around 20% NA in the data set and these cannot be excluded as they are significant for analysis. Median based imputation were performed for these cases as the variable in picture has a right skewed data and for skewed data median is a better option for imputation than mean. Using fillna() function NA values are imputed with median

**Task 2: Data Exploration**

* **Subsection I, exploring variables.**

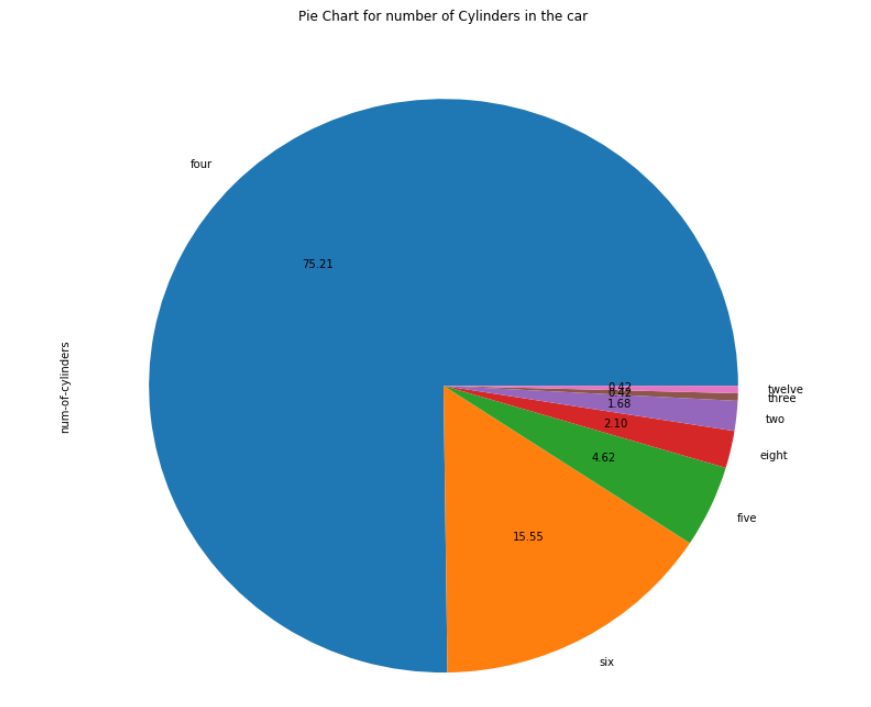
*Nominal variable* Fuel type

Using pie chart, we interpreted the variable fuel-type which has two levels i.e Gas and Diesel. From the visualization it is evident that almost 88% vehicles had a fuel-type of gas.



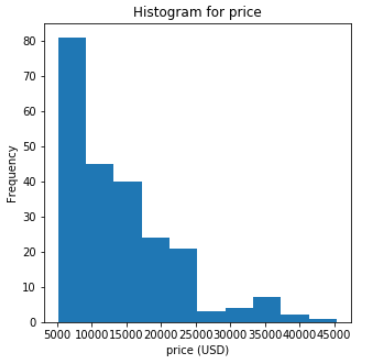
*Ordinal variable* Number of Cylinders

Pie chart indicates that nearly 75% of the cars had 4-cylinders engines while around 16% had 6-cylinders and only 0.5% cars had 12-cylinder engines.



*Numerical variable* Price

Using histogram, we interpreted the variable price and it was observed that nearly 34% i.e. (about 80 observations out of 238) of cars had a price range from $5000 to $10000. Moreover, it was also seen that the frequency of car fell almost exponentially as the prices increased.

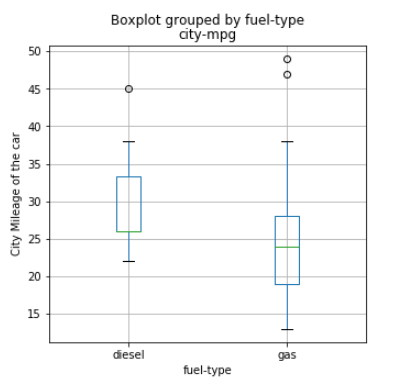


When visualizing a nominal or an ordinal variable, pie chart provides a clear distinction between various levels and their values visually whereas histogram provides a fair interpretation of the numerical variables.

* **Subsection II, exploring relationships between variables.**

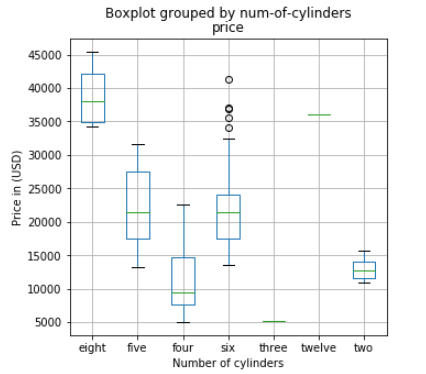
*a) City mileage and Fuel-type*

Using the boxplot between *City mileage and Fuel-type,* we are trying to test the hypothesis that Gas car have a higher city average than Diesel cars.



*b) Price and Number of Cylinder*

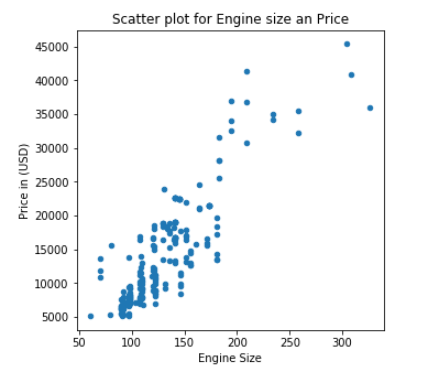
Using the boxplot between *Price and Number of Cylinder,* we are trying to test the hypothesis that 4-cylinder engine cars have an overall lower price than 8-cylinder. These two attributes were chosen to see how much variation in present in price with respect to the number of cylinders.



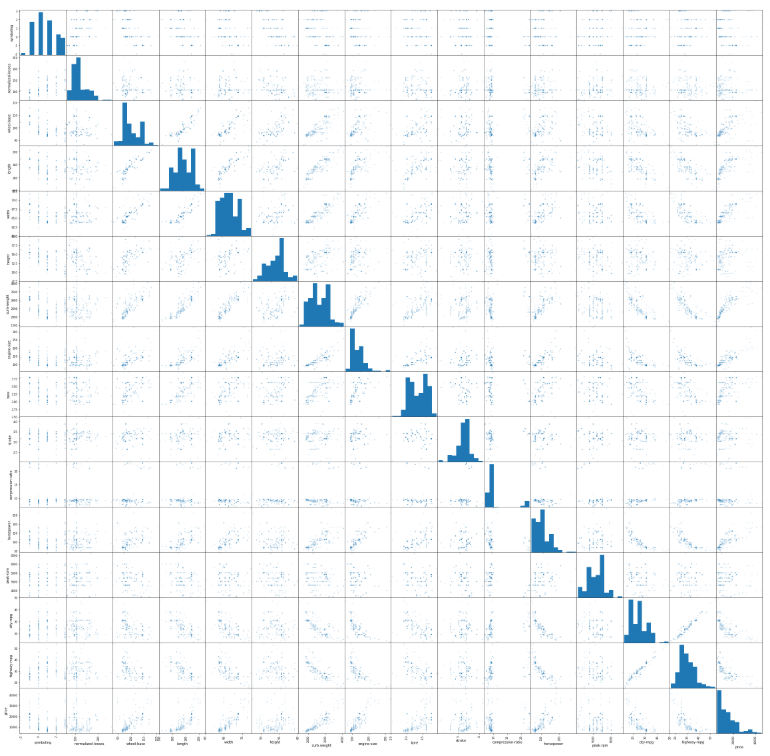
* 8-cylinder engine cars were the most expensive ones.
* 4-cylinders engine cars were the most affordable.
* Boxplot somewhat provides credibility to the interpretation that showed nearly 34% cars had price ranging from $5000-$10000.

*c) Engine-Size and Price*

Using Scatterplot between *Engine-Size and Price,* we are trying to test the hypothesis that as the engine size increases price of the car also increases and assessing how closely price and engine size are correlated.



* A positive linear correlation relationship exists between engine size and price.
* As engine-size increases the price for the car increases.
* **Subsection III, exploring all numerical variables using a scatter plot.**



* Variables that depict a positive linear correlation with Price variable.
  1. Engine size
  2. Curb weight
  3. Horsepower
  4. Length and Weight
* Variables that depict a negative linear correlation with Price variable.
  1. City mileage
  2. Highway mileage
* As the horsepower increases the city mpg and highway mpg decreased.
* As curb weight increases the mileage decreases.
* The curb weight increased with the engine size.
* From the scatter matrix as well as above points, as engine size increases mileage of the car is decreased

**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 and Descriptive Statistics and Visualisation ', PowerPoint slides, COSC 2670, RMIT University, Melbourne.
* Data Preprocessing and Introduction to statistics module notes.