***BE 1600***

***Introduction to***

***Programming and Computation***

***Python***

**Project- Due Date: Friday December 16**

120 points

Our goal with this project is for you to have the chance to apply what you learned in this course to solve a problem.

You are given a sample data set from [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/datasets/Auto+MPG). In this project you should create an end-to-end program using a main function and multiple functions as needed to address the followings:

1. Read the dataset from the provided csv file.
2. Data preparation:
   * Identify and remove/replace the outlier values.
   * Identify and remove/replace the null values.
3. Interpret the dataset using the statistics metrics.
4. Provide a set of plots presenting pair-wise correlation for variables/columns.

You can use matplotlib.pyplot library.

import matplotlib.pyplot as plt

1. Provide a pairwise correlation between variables using the Pearson correlation formula.
2. Define a new set of variables and test them by checking their correlation w.r.t. MPG
   * Displacement on Power
   * Weight on cylinder
   * Acceleration on power
   * Acceleration on cylinder

Like items 3 and 4, provide their correlation and plots.

1. Implement a function that marks each vehicle with a label as high or low fuel efficiency for each data entry, use the vehicle fuel efficiency (MPG) to identify the label and add the labels as a new column.
2. Identify the most important variables that would contribute to predict the vehicle fuel efficiency, justify your answer using the pair-wise correlation and plots from the previous item.
3. Finally, write a report explaining your approach to address the above items; how did you clean the data. What kind of outliers did you identify. And interpret the obtained statistics and correlation. (1-2 pages).