AIM 5001 Project 2 (Module 10) (100 Points)

Data Preparation & Feature Engineering

***** You may work in small groups of no more than three (3) people for this Project *****

When the number of explanatory variables is relatively large with respect to the number of observations contained within a data set, data science practitioners need to know how to effectively reduce the number of explanatory variables required for the intended model. Furthermore, as we've learned, the individual variables within a data set may need to be transformed prior to use within a machine learning algorithm. Additionally, we've learned that missing data values can impede the proper functioning of many machine learning algorithms. For this Project your primary task is to apply data preparation and feature engineering techniques to a data set comprised of information related to **automobile gas mileage**. The data set you will be using is sourced from the UC Irvine machine learning archive:

- https://archive.ics.uci.edu/ml/datasets/Automobile

The data set is comprised of 205 observations and 26 attributes. Please refer to the UCI web page for further details on these variables.You are to apply your data preparation and feature engineering expertise to the problems listed below. But first, get started with the Project as follows:

- Load the provided M10_Data.csv file to your AIM 5001 Github Repository.
- Then, using a Jupyter Notebook, read the data set from your Github repository and load it into a Pandas dataframe. Assign meaningful column headings to the content of the dataframe based on the information provided at the UCI web link provided above.
- Using your Python skills, perform some basic exploratory data analysis (EDA) to ensure you understand the nature of each of the variables contained within the dataset. Your EDA writeup should include any insights you are able to derive from your statistical analysis of the attributes and the accompanying exploratory graphics you create (e.g., bar plots, box plots, histograms, line plots, etc.). It is up to you as the data science practitioner to decide how you go about your EDA, including selecting appropriate statistical metrics to be calculated + which types of exploratory graphics to make use of. Your goal should be to provide an EDA that is thorough and succinct without it being so detailed that a reader will lose interest in it.

When you have completed your EDA work, use the results of that work to help you answer the following questions:

1. **(10 Points)**Which numeric variables contained within the data set appear to require the use of a feature scaling method for purposes of preparing them for use within a machine learning algorithm? Be sure to list each relevant variable and explain why you believe each variable that you've identified requires the use of some sort of feature scaling method.

2. **(15 Points)** Consider the **number-of-doors** and **price** variables: Based on your EDA work, how many missing data values occur within each of these attributes? As we've learned, missing data values can impede the proper functioning of many machine learning algorithms. To address the missing the **number-of-doors** and **price** values, you have been instructed to formulate what you believe will be an

effective data imputation approach for purposes of estimating reasonable proxies for the missing data values. Your supervisor tells you that the affected data observations **MUST** be retained within the data set, and that it would be inappropriate to use either a mean, median, or mode value for any of the missing values since doing so would increase the likelihood of introducing unwarranted bias within the data set. Describe the imputation method you would employ for each variable. Then, using your Python skills, apply your prescribed imputation methods to the variables. Be sure to include graphics and commentary that explain your approach as well as the results of your efforts.

3. **(15 Points)**Consider the **engine-size** and **stroke** variables: Describe the specific feature scaling method you would apply to each of them. Then, using Python, generate both a histogram and a boxplot for the original content of these two variables. Next, apply your prescribed feature scaling methods to the two variables and create histograms and boxplots that show the results of your feature scaling efforts. Compare your newly created plots against the plots you created for the original content of the variables. Comment on whether your feature scaling efforts improved the distribution of the data. If your feature scaling efforts did not improve the distribution of the data, explain why you believe your efforts were not effective.

4. **(15 Points**) Consider the **symboling**,**make**, and **engine-type** variables:

   A) For each variable, specify whether its content is numeric/continuous, numeric/discrete, categorical/nominal, or categorical/ordinal

   B) For each variable, describe the methodology you would employ for purposes of preparing its data values for use within a machine learning algorithm.

   C) Using your Python skills, apply your prescribed data preparation methodologies to the three variables. Be sure to show a sample of your results within you Jupyter Notebook.

5. **(15 Points)** Consider the wheel-base, length, width, height, curb-weight, engine-size, compression-ratio, horsepower, peak-rpm, and city-mpg variables. Using your dimensionality reduction expertise, use Python to reduce the dimensionality of this group of variables to a set of new orthogonal features. Be sure to include appropriate commentary explaining the dimensionality reduction method you have elected to implement and discuss the results of your efforts. For example, you should explain how many orthogonal features your approach has generated as well as how much variability is explained by each of your new features.

Your deliverable for this Project is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Data Loading (5 Points)**: This section should include the Python code used to load the data set + assign meaningful names to each column within the dataset.

2) **Exploratory Data Analysis (25 Points)**: Explain + present your EDA work including any conclusions you draw from your. This section should include any Python code used for the EDA.

3) **Data Preparation & Feature Engineering (70 Points total)**: This section should contain all Python code and explanatory commentary related to your analysis and answers for Questions 1 through 5 from above**.**

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload / submit your Jupyter Notebook within the provided Project 2 Canvas submission portal. Be sure to save your Notebook using the following nomenclature:**first initial_last name_Project2_assn**" (e.g., J_Smith_Project2_assn_). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***