Nickhil Tekwani
CS6140 Assignment 1
1/29/2023

1. Reproduce Exploratory Data Analysis as provided in the specimen. Discover an interesting aspect of the data that is not shown in the specimen. The specimen is using R language but you will reproduce the analysis in Python using libraries of your own choice.
    a. One insight that can be found that is not already shown in the specimen is that the average fuel efficiency of vehicles (measured by UCity) is positively correlated with the vehicle class, with compact and subcompact vehicles having the highest average fuel efficiency, and larger vehicles such as SUVs and pickups having lower average fuel efficiency. This suggests that consumers who prioritize fuel efficiency may benefit from choosing smaller vehicle classes.
    b. See attached ipynb file for analysis in python
2. The UCity variable is treated as dependent/target variable in the specimen. Discuss your approach to build a predictive model. Is it going to be a classification model or regression model. Why?
    a. UCity is a continuous variable, therefore it is a target variable that would require a regression model to predict. In a regression model, the goal is to predict a continuous value for the target variable based on the values of the input variables. In this case, UCity is the target variable, and the input variables can be any other variable in the dataset that may have an impact on UCity, such as vehicle class, engine cylinders, fuel type, etc.
3. Discuss which variables you will not consider as inputs to the model. Why?
    a. When building a predictive model, it is important to carefully select the input variables that will be used to make predictions. Some variables may not be useful or appropriate to include as input variables, and may even negatively impact the performance of the model.
    b. These variables will not be as useful as inputs to the model:
        i. 'year' : as we know that the data is from the year 1984 to 2018, so the year variable may not be informative in the model.
        ii. 'model' : the model of the vehicle can be inferred from other variables such as vehicle class, fuel type, transmission, etc.
        iii. 'barrels08', 'co2TailpipeGpm' : These variables are correlated with the target variable 'UCity' which means that these variables are redundant.
        iv. 'fuelCost08' : this variable is not directly related to the target variable, rather it is a monetary value.
        v. 'trans_dummy' : as we know that most common transmission type is Automatic 4 and 4 Speed, this variable may not provide any additional information to the model.

    c. These variables may not be useful as inputs to the model because they are either highly correlated with other input variables, not directly related to the target variable, or do not provide any additional information to the model.
4. How will you evaluate your model to avoid over-fitting/under-fitting?
    a. Splitting the data into training and test sets: The training set is used to fit the model, while the test set is used to evaluate the performance of the model. By comparing the performance of the model on the training set and the test set, we can determine if the model is overfitting or underfitting. If the model performs well on the training set but poorly on the test set, it is likely overfitting. If the model performs poorly on both the training and test sets, it is likely underfitting.
    b. Another approach to avoid overfitting is to use a regularization method, such as L1 or L2 regularization, which adds a penalty term to the cost function to discourage the model from fitting the noise in the data.