Data Mining I
Homework 1

40 points

Directions: Submit all source codes with write up. All code must be submitted as a jupyter notebook and a *html saved jupyter notebook with output.

1. (10 points) Consider the "airquality" data in R
   >data(airquality)
   a) What is the dimension of this data?
   b) How many solar measurements are missing?
   c) What are the averages for: ozone, solar, wind and temp. Calculate this in two different ways.
   d) Eliminate all observations with missing solar data.
   e) With the modified dataset in Part D, recompute the averages in part C.

2. (10 points)
   Starting with the original "airquality" dataset:
   a) Create datasets for each Month.
   b) Save these Monthly datasets into a list.
   c) Save the list from part B.

3) (10 points) Consider the "Auto" dataset in the ISLR2 package. Suppose that you are getting this data in order to build a predictive model for mpg (miles per gallon). Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed. Pre-process this data and justify your choices in your write up. Submit the cleaned dataset as an *.RData file.

4) (10 points) Perform a multiple regression on the dataset you pre-processed in question three. The response variable is mpg. Use the lm() function in R.
   a) Which predictors appear to have a significant relationship to the response.
   b) What does the coefficient variable for "year" suggest?
   c) Use the * and : symbols to fit some models with interactions. Are there any interactions that are significant? (You do not need to select all interactions)