**Project Write Up - Part 1**

After loading in the Rain dataset the first thing I took a look at was the number of variables and the number of observations. I noticed that there were 20 variables with just over 28,000 thousand variables, so I knew this was going to be quite a large dataset. After mutating the RainTomorrow variable per the instructions, I plotted all numerical variables against the RainTomorrow Variable. The RainTomorrow Variable was located on the X axis and the numerical variables along the Y axis, this was done to see, at a quick glance which variables appeared to be significant in whether or not it was going to rain the next day. This exercise is similar to what we performed in an earlier module. The variables that stuck out the most to me from this exercise were the following: the mintemp and maxtemp variables, which represent the minimum and maximum temperature in degrees Celsius, windgustspeed, humidity9am, humidity3pm, temp3pm, as well as cloud9am and cloud3pm. What caused me to think that these were significant variables was the fact that there appeared to be significant discrepancies between the boxplots on the No and Yes, indicating that they were good predictors for whether or not it would rain the next day. What I found shocking was that cloudcoverage at 3pm seemed to be significant but not the cloud coverage at 9 am. It should be noted however, that at this point I have not imputated any values to accommodate for missingness or deleting any data at this point in time. The boxplots being performed automatically removed any missing data; however, it is possible that this should be handled in a different manner.

Before any additional changes to the dataset were made to adjust for the missing data I wanted to see which variables were correlated with one another. This was performed in R via the ggcorr function. This was important to see as I needed to know if any of the variables were highly correlated with one another. This took a look all of my numerical variables and saw the correlation level. The reason that this is important is that once I get into building the model, I will need to know if any of the variables are causing false results within the dataset artificially increasing my r-squared value. We want to try and avoid that if possible.

Now that I had a picture of what the data looked like prior to any imputations or column / row wise deleting, I utilized the vim\_plot function to see what percentage of the 28k observations were missing for each variable. The cloud coverage variables were missing around 40% of their observations, this value was too high for me to be comfortable to imputate their values. While cloud coverage at 3 pm did seem to be a predicting variable, there were plenty of other variables that I believe will help build a strong variable, which are much more complete. I removed these two variables from the dataset. The only variable I chose to imputate was the windgust speed. This was the only variable selected because out of the original 20 variables, I decided that only 6 were needed to help predict rain: mintemp, maxtemp, windgustspeed, temp3pm and the humidity levels, these seemed to be the best predictors of rain or not. I let R use the mean average to fill in the blank variables. The density plots were then compared between the imputated and original variables and the two were similar enough for me to comfortable with the imputation. This has set me up to be able to test the variables in various models within part 2.