Part 1 – Data Exploration, Preparation, and Visualization

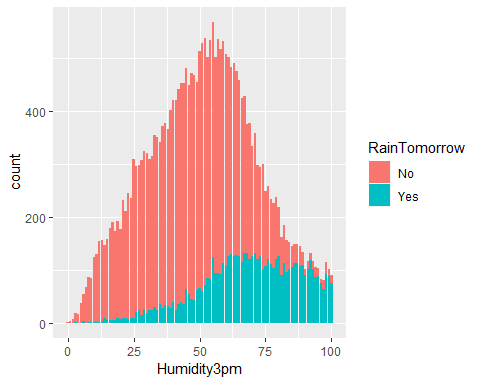
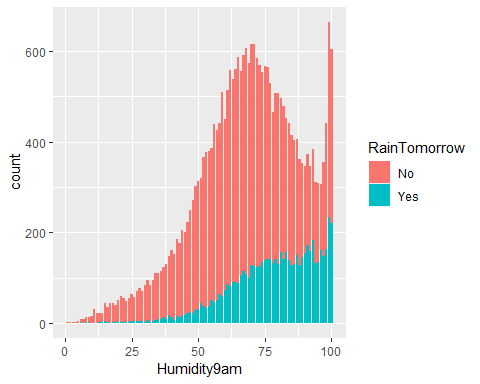
After viewing the “rain” data file for this project, I have concluded what I think might be some of the significant variables in determining the response variable might be. Through examining the data and visualizing the relationships between the variables, I think the significant variables in the dataset are "MinTemp", "MaxTemp", "Humidity9am", "Humidity3pm", "WindSpeed9am", and "WindSpeed3pm.”

The first thing I did in order to help determine which variables were significant is view the structure and summary of the dataset. From this, I was able to determine which variables had missing data and which variables needed to be converted to factors. I converted the “RainToday” and “RainTomorrow” variables to factors as they both only had two levels and factor conversions can make the data easier to work with as opposed to leaving them as characters. I left the variables “Date,” “WindGustDir,” “WindDir9am,” and “WindDir3pm” as characters because the dates are all different and would have created to many levels and the other three variable also would have had a large amount of levels and were later determined to be pretty insignificant variables.

As far as the missing data goes, all of the numeric variables had missing data or rows with NA’s but the most notable numbers were in the variables “Cloud9am” and “Cloud3pm.” The variable “Cloud9am” had 10,673 NA’s and the “Cloud3pm” had 11,341 NA’s in the data. With the dataset having 28,003 observations, these variables proved to be missing a substantial amount of data. To further visualize the missing data and decide how to proceed, I used the aggr function in the VIM package. This visualized the scale of missingness in the data in both the “Cloud9am” and “Cloud3pm” variables. For these two variables, I decided to use column-wise deletion as I concluded that row-wise deletion would be too destructive to the dataset and there were too many rows of missing data for imputation to be effective. Once deleting these two variables, I used the aggr function again to visualize the missingness in the remaining variables. Again, there proved to be multiple variables with a significant amount of missingness and I decided it would be best to use column-wise deletion for these variables as well because they were all missing over 60% of their data. I used the aggr function again to see the proportion of missingness left in the dataset but did not delete anymore columns.

Next, using the most recent plot of missingness I selected 8 variables (including the response variable) based on the proportion of missingness that remained. The variables chosen were "MinTemp", "MaxTemp", "Humidity9am", "Humidity3pm", "WindSpeed9am", "WindSpeed3pm", "RainToday", and "RainTomorrow.” One variable I left out even though the missingness of data was comparable to the others, was “Rainfall.” When considering this variable and the response variable, it made the most sense to leave this variable out as the amount of rainfall does not really impact whether or not it will rain. A few of the variables had a bit more missingness but their counterparts showed a low volume of missing data and I determined it would be best to keep them in the selection. After looking at the structure and summary of the data again, there were still variables with missing data that needed to be dealt with.

In order to deal with the remaining missing data, I chose to use imputation. Once imputing the data set and viewing the summary, it can be seen that the only variable not included in the imputation was the response variable, “RainTomorrow.”

 Once the missing data was dealt with, I moved on the visualizing the relationships between the remaining variables and the response variable. The first function I used was ggpairs so I could get an understanding of all the variables and their relationships with each other. Based on the results of the boxplots, it appeared that the variable with the best relationship with the response variable was “Humidity3pm.” In favor of getting a better sense of the relationships, I created a bar graph plotting each of the remaining variables with the response variable. The variable that appeared to be a strong predictor of “RainTomorrow” was in fact “Humidity3pm” and “Humidity9am” seemed to follow close behind.

The variable that appeared to be a pretty weak predictor of the response variable was “RainToday.” The “RainToday” variable does not seem to impact whether it will rain tomorrow as the graph indicates there is a very similar chance of rain either way.

