Final Project

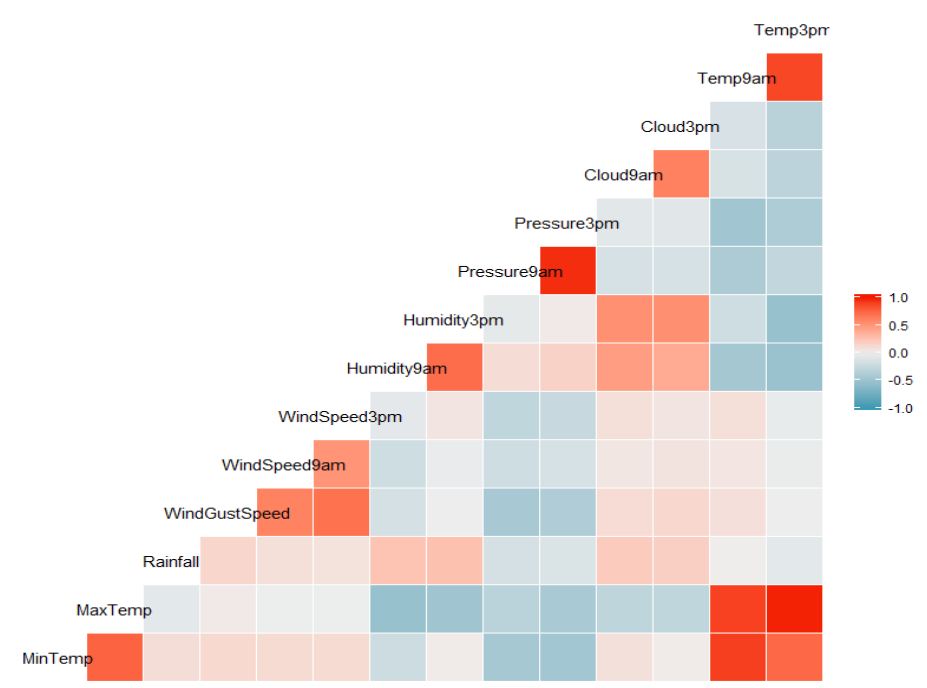
Part 1 - Data Exploration, Preparation, and Visualization

Cynthia Nickerson

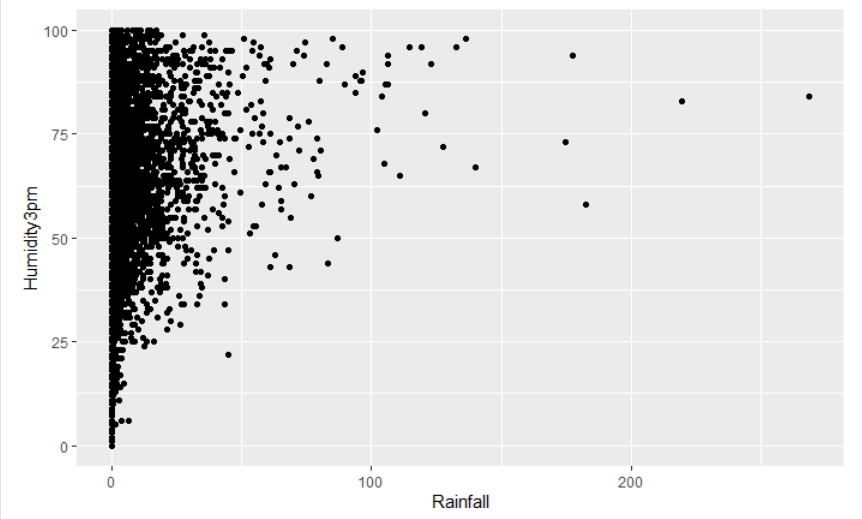
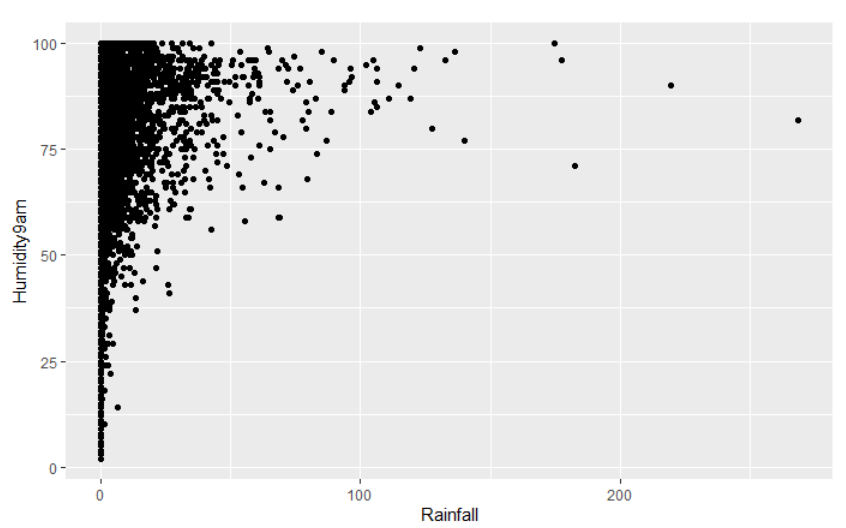
For this assignment, we are analyzing historical weather data to try and predict rain. I began the assignment by loading in a variety of packages that could be helpful, both now and during phase two. I then imported the datafile provided entitled “rain.csv”. This data had 28,003 observations with 20 variables.

Variables include date; location of the record; max and min temp; rainfall; wind gust direction and speed; wind direction and speed at 9am and 3pm; humidity at 9am and 3pm; pressure at 9am and 3pm; clouds at 9am and 3pm; temperature; and rain today and tomorrow.

Upon first review, it was evident that the file was missing data. A review of the missing data via a summary of rain and the vim\_plot function showed data was missing in several different fields. While I considered imputing some of the data fields, I ultimately decided to remove the rows with NA data for two reasons. One was that several fields were characters; the second reason was that when another student struggled with this, the consensus was that it was simpler to remove the rows. After removing the rows, I was confident I still had enough data to move forwards with 13,887 observations and 20 variables.

I next began to look for relationships within the data. This was first achieved via the ggcorr function: 

It is not unsurprising that there is a relationship between the 9am and 3pm temperatures and the max and min temperatures. It’s also unsurprising that there is a relationship between windspeed (9am – 3pm), humidity (9am – 3pm), and pressure (9am – 3pm) as the day passes. These seem like obvious relationships; I’m not sure how much they will contribute to predicating rain. Instead, I’d like to focus on the relationship between rainfall and humidity.

To look at this relationship, I started by creating ggplot visuals for both 9am and 3pm; see below.

We can roughly see that the higher humidity, the better the chances are of rainfall. At first glance this appears more so in the 9am data.