### BAN502-803: Summer 1 Gregory Moore

### Course Project - Phase 1 6/20/2019

**Data Exploration, Preparation, and Visualization**

**Intro**

While working with the data within the rain.csv source file, it became immediately apparent that some decisions would have to be made if we were to use any of the predictive analytics techniques taught to us during this semester. Some of these choices would be what to do with the columns that have a large number of missing values, what columns could have correlation or significance to our designated response variable (i.e. RainTomorrow), and how to best manipulate the data so it could be easily processed by the tools within the R statistical programming language.

**Missing Data**

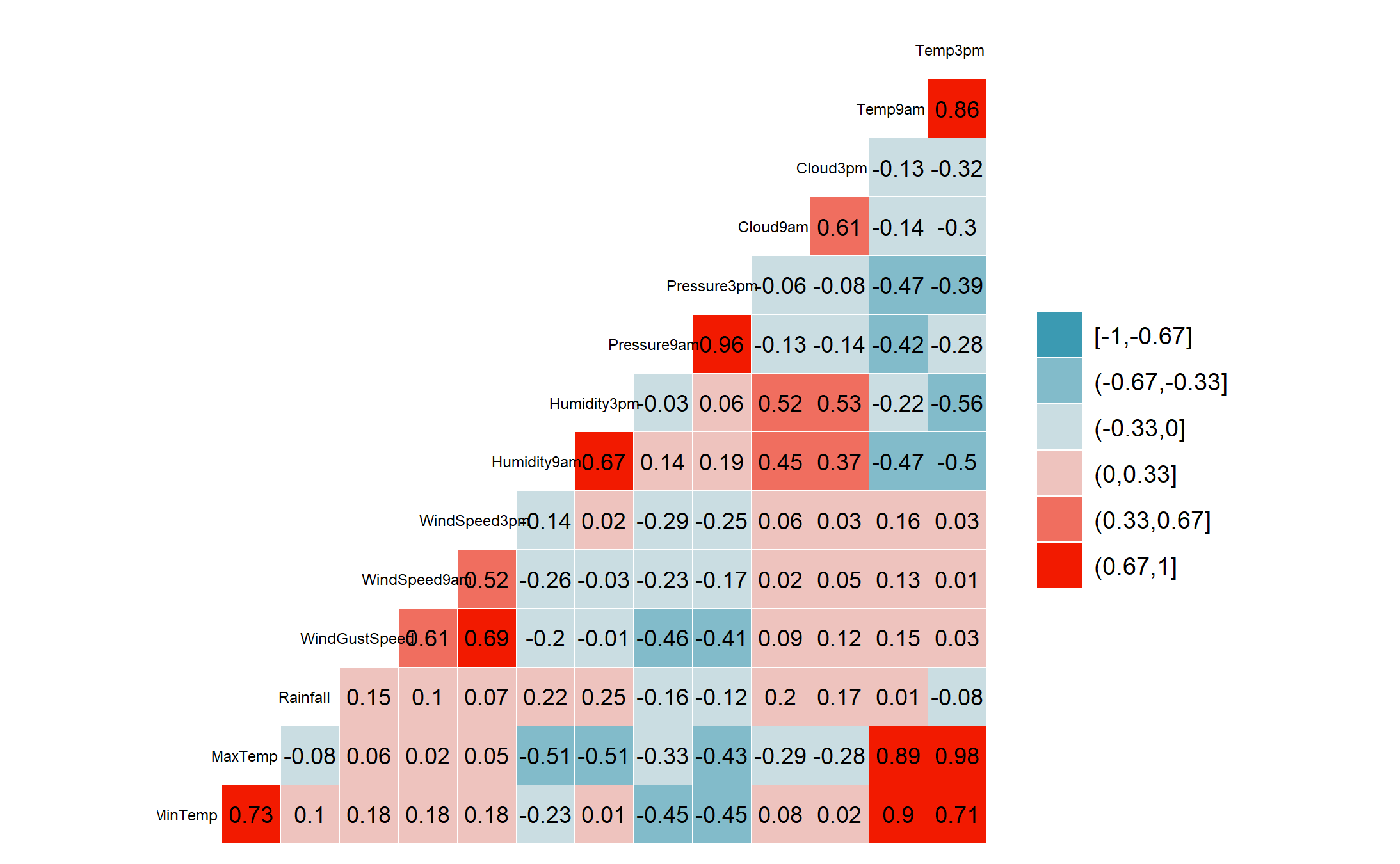
Our RAIN dataset started off with 28,003 rows of weather-related measurements that ranged from 2007 to 2017. Unfortunately for the class, of the 20 included columns, only “Date” and “RainTomorrow” had zero missing values (Note to readers: Throughout this writeup missing data and “NA” values will be used interchangeably). Most of these “NA” values fluctuated in the hundreds, but there were several fields that climbed up into the thousands. Some even having as much as 40% of its values missing (e.g. Cloud9am and Cloud3pm). I attempted to use the VIM *aggr()* function to help visualize any rows with several columns of missing data to look for potential patterns, but there were too many observations in the source file for this command to properly show how the missing fields lined up with each other. I decided to move on and try *aggr()* at a later time.

And this leads us to our first, and likely, our most important decision that had to be made. Do we completely remove the columns with the large counts of missing data or do we strategically remove rows with “NA” values in the hopes that the removed rows will also line up with missing data in other fields? Or do we impute the data using the VIM package’s *mice()* function? I can tell you now that I used all three methods, but I could not make the necessary choices until I figured out how all the columns related to each other and how they related to the response variable “RainTomorrow". Before doing this, I decide to check for one more type of missing data before going on to the next step. This involved confirming that there were no “ “ values (i.e. blank) or whitespace having values in our RAIN dataset. I can report that none were found.

**Data manipulation**

Before looking for relationships between the columns and our response variable, I needed to stop and adjust some of the presented data by both aggressive manipulation and by data type. I first started with the “Date” column. I just did not like this field. It did not have any “NA” values, but it was a date value as CHARACTER data type in a MM/DD/YYYY format. To me, this was worthless data even if it was converted to a DATE data type. What correlation or significance could a single day in a random year have on our given response variable. I was going to drop the column from the dataset completely when I decided to give field one chance to show me worth to our future prediction models. I decide to mutate the “Date” field into three separate date measurements (e.g. month, day, and year) and then to convert these new fields into factors to see if there was any correlation to “RainTomorrow”. My theory was that “year” and “day” would still be next to worthless to us, but I did have high hopes for “month”. I could easily imagine a higher chance of rain in the spring months compared to winter months. And then I also mutated the rest of the character fields in the RAIN data set to factors. This included the two cloud coverage variables, the wind variables, and then the two Rain = Yes variables. I first did this to make sure “Cloud9am” and “Cloud3pm” were treated as categorical fields instead of numbers, but then I decided that I also wanted the wind directions treaded the same way and I was thinking this could help later on if/and when it came time to impute the data. If, for example, a wind direction of “NW” had a factor level of 1, then it may be easier to educationally “fake” its missing values using the *mice()* function. I am not sure if this step mattered in the long run, but I don’t think it hurt my results.  
 **Visualization, Correlation and/or Possible Significance**

The first visualization that I wanted to see was my newly split “Date” fields. For this I used “RainTomorrow” as the X-axis in three bar plot graphs while using “month”, “day”, and “year” as the fill criteria. The results for this experiment were interesting, but nothing earth shattering in terms of a relationship with the response variable. As expected, there were a few YES peaks in the April through August months (rainy seasons?), but these were minimal at best. As for the other two created columns, there were too many factors in the “day” bar plot to support any relevance and, while there were a few years that looked like it produced more RainTomorrow = YES values, I don’t think the “year” date point will be useful in future modeling.

Next came a correlation matrix plot using the *ggcorr()* function to calculate any relationships between our quantitative data points (see below). There were some very interesting numbers within this table, but the first thing that I noticed was that there were almost no relevant positive correlation values above a .3 rating. I say relevant because of the obvious relationship between the two temperature fields and the MIN/MAX temperature fields. The same can be said of the “WindGustSpeed” and the two wind speed measurements. There were some interesting relationships to keep in mind for later modeling decisions. For example, there is a semi-decent connection between humidity and temperature, pressure and top wind gust speed, and pressure and temperature. The last thing to note below is that “Rainfall” shows almost no correlation with the other data points.   
  
The rest of the visualizations were made using box plots for the quantitative fields and bar plots for the categorical fields while using “RainTomorrow” as the X variable. The high points for this process are as follows:

* The MIN/MAX and other temperature values appear to be meaningless to the response variable
* The “Rainfall” field’s measurement values are too small to really notice a meaningful relationship to “RainTomorrow” when graphed: scatter plot and box plot.
  + There also appears to be several huge outliers that may throw off any future imputing methods
* The wind direction fields have some interesting peaks when plotted with our response variable.
  + For YES - winds from the North and West
  + For NO - winds from the South and East
* The wind speed plots show nothing spectacular in terms of a connection to “RainTomorrow”
* The two humidity plots show a strong relationship and will be useful in future models
* Pressure also shows some possible influence and will be useful in Phase 2
* The cloud coverage fields showed the expected results (especially the 3pm measurement).
  + FACT: There must be clouds in the sky for it to rain.
* “RainToday” did show that if it rained today, it had a better chance to rain the next day
  + However, the low number of RainTomorrow = YES values in the no column almost nullifies this tendency’s worth.

**Conclusion**

I may be in the minority but when the time comes to build predictive models, using the data within the rain.csv file in Phase 2, I will be forfeiting a large chunk of this 28k row dataset. This will mostly be because I have come to the conclusion that Cloud9am and Cloud3pm are just too important to our “RainTomorrow” response variable to remove completely. Maybe not by themselves, but I think we will discover that this data point will be very useful when used in conjunction with fields like Humidity and Pressure as both show a strong relationship to “RainTomorrow” (see graph below). Because of this opinion, I will be removing all rows with NA values in all six of these columns and then I will impute the missing data in the other fields. However, I will be completely removing the “Date” and “Rainfall” columns first. The date information does not appear to be useful and the total rain fall in a given day seems to matter very little to our response variable and the low average, plus several large outliers, may skew any imputed data in the other columns. And lastly, we come to the temperature measurements. While the temperature measurements do not appear to show much correlation to “RainTomorrow”, temperature certainly has a noticeable relationship to pressure and humidity and these temperature values may show a daisy chain relationship when put through our modeling techniques.

A screenshot of a cell phone

Description generated with high confidence