The purpose of this project is to study weather data and predict rainfall based on historical trends. The data we’re provided involves dates from November 2007 to late-June 2017. This dataset includes 28,003 observations with 20 variables describing various factors. Variables include information such as wind speed, humidity, temperature, and cloud cover. Among these observations are numerous missing values for some variables. Phase one of this project will be preparing the data for analysis by eliminating or filling those missing values, then identifying potentially strong predictors of the “RainTomorrow” variable.

With over 28,000 observations, it's not too surprising to see most variables are each missing at least a few values. the total missing can range widely from 64 “MaxTemp” values to over 11,341 missing values on cloud cover. Cloud9am and Cloud3pm suffer the most from missing data, with almost 40% of values missing. If we were to remove all observations with a missing value, our dataset size would reduce by half to 13,887 observations.

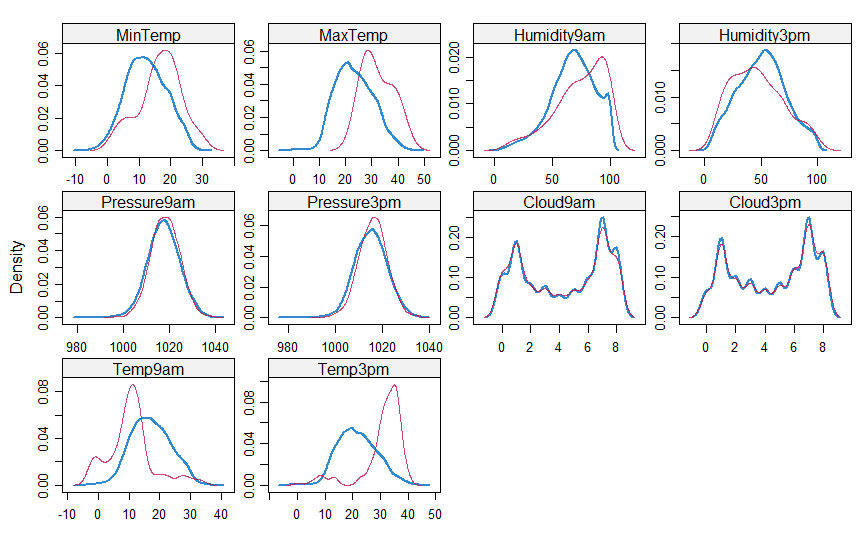
The dataset contains both categorical and quantitative variables, both of which have missing values. For simplicity, we’ll opt to eliminate any observations with missing categorical values. The following code checks for any missing, or “NA”, values in our categorical variables and eliminates the entire observation. Doing so reduces our dataset to 24,351 observations.

*rain = rain %>% drop\_na(WindGustDir, WindDir9am, WindDir3pm, RainToday, RainTomorrow)*

The next section of code imputes missing values using predictive mean matching, which essentially uses a predicted regression model from known data to estimate a missing value. With this code executed, all missing values in the quantitative variables have been replaced with an estimation based on the present data.

*rain\_imp = mice(rain, m=1, method = "pmm", seed = 12345)*

As we can see below, the imputed data (red) fits well with the observed values (blue). We’ll merge the imputed values with the original data to create a completed dataset.



With missingness in the data accounted for, we can now try to pick out variables which might predict the RainTomorrow variable. Common sense tells us Date will probably not be a predictor. The remaining variables are divided into 3 matrices of plots to show any correlation at a glance. We can see variables related to humidity, cloud cover, and temperature may be strong predictors of RainTomorrow. The remaining variables appear to have little to no correlation with the response variable individually.

We also created 4 new variables in the dataset to study the effect of differences in humidity, temperature, pressure, and wind speed on the response variable. It appears that changes in temperature and humidity may be predictors of RainTomorrow, while changes in either pressure or wind speed appear to not have as much of an effect.

