Course Project: Part 1 – Data Exploration, Preparation & Visualization

Attn: Dr. Stephen Hill

BAN 502: Introduction to Predictive Analytics

Stephen AJ Goode

There’s a good amount of data missing from our dataset. Visualizations were created to analyze missingness in our dataset before and after omitting/imputations. The majority of the missing data is from the Clouds variable, “Cloud9am” and “Cloud3pm”, which is missing around 38% and 40% of the data respectively. Besides the “Date” variable, the Temperature variables, “MaxTemp” & “MinTemp”, are missing the least amount of data with less than 1%. In total, 6.7% of our data is missing out of a total of 28,003 observations.

I decided to omit the missing values but the dataset was cut in half, shrinking from over 28,000 observations to 13,887. This is an extraordinary loss in data. For fear of losing so many observations, I chose to create a new dataset with imputed values titled “rain\_complete”.

I also changed the Date variable and separated it into three separate columns – Year, Month and Day – to create a new variable called Month (I chose to exclude Day & Year). By looking at the Month variable, we can look more closely at seasonal rainfall and weather patterns.

Utilizing ggplot and its visualization functions, there appears to be several strong predictors influencing (or influenced by) our response variable. The most visually obvious predictors appear to be cloud coverage and humidity. There also appears to be a negative relationship between our response variable and both Temperature at 3pm and Atmospheric Pressure – lower temperature (at 3pm) and lower pressure (morning & afternoon) seems to increase the likelihood of rain the next day. It makes sense that the temperature is lower the day before a rain shower because there may be more cloud coverage, and other storm effects – such as lower pressure, higher humidity, and wind.

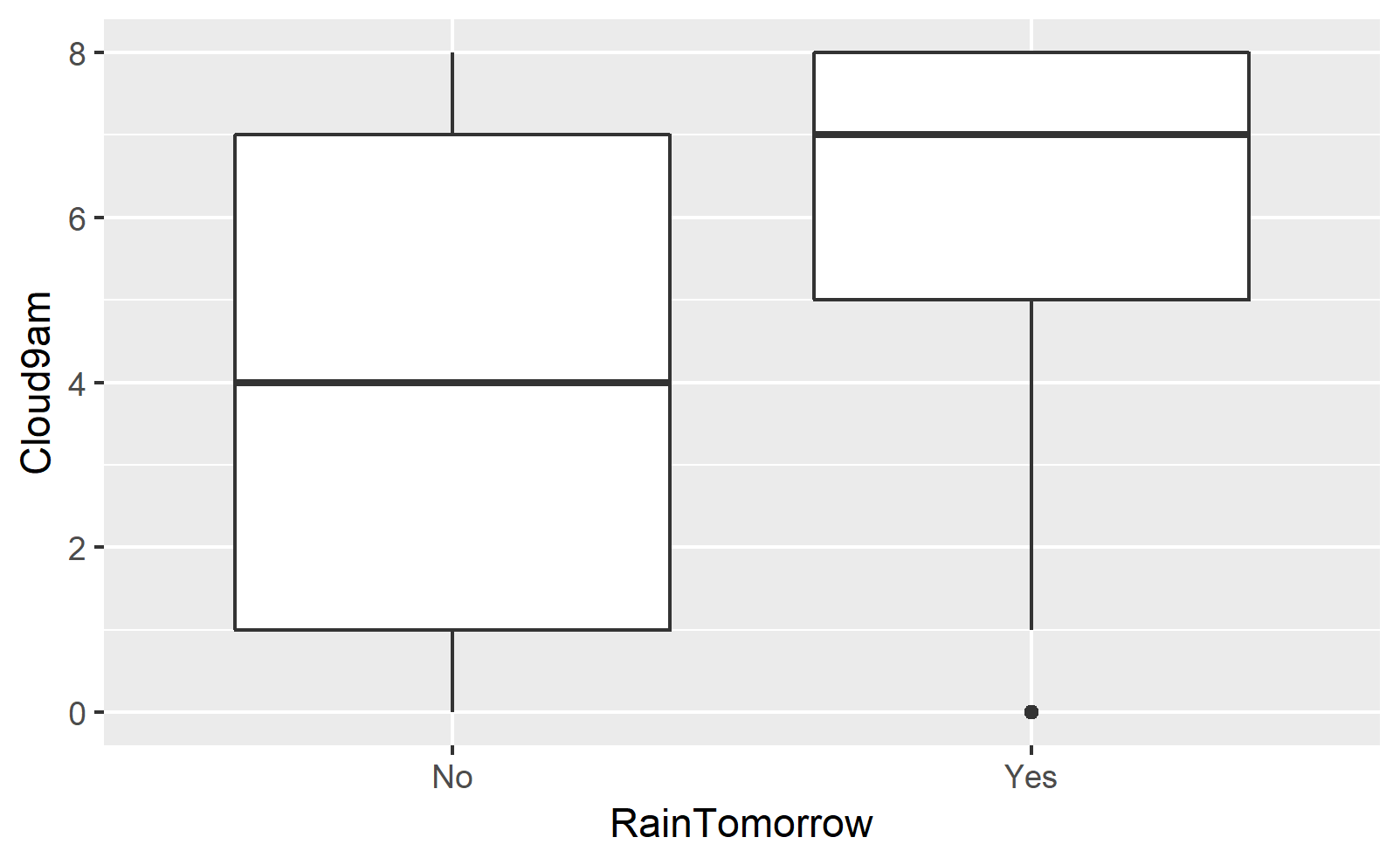
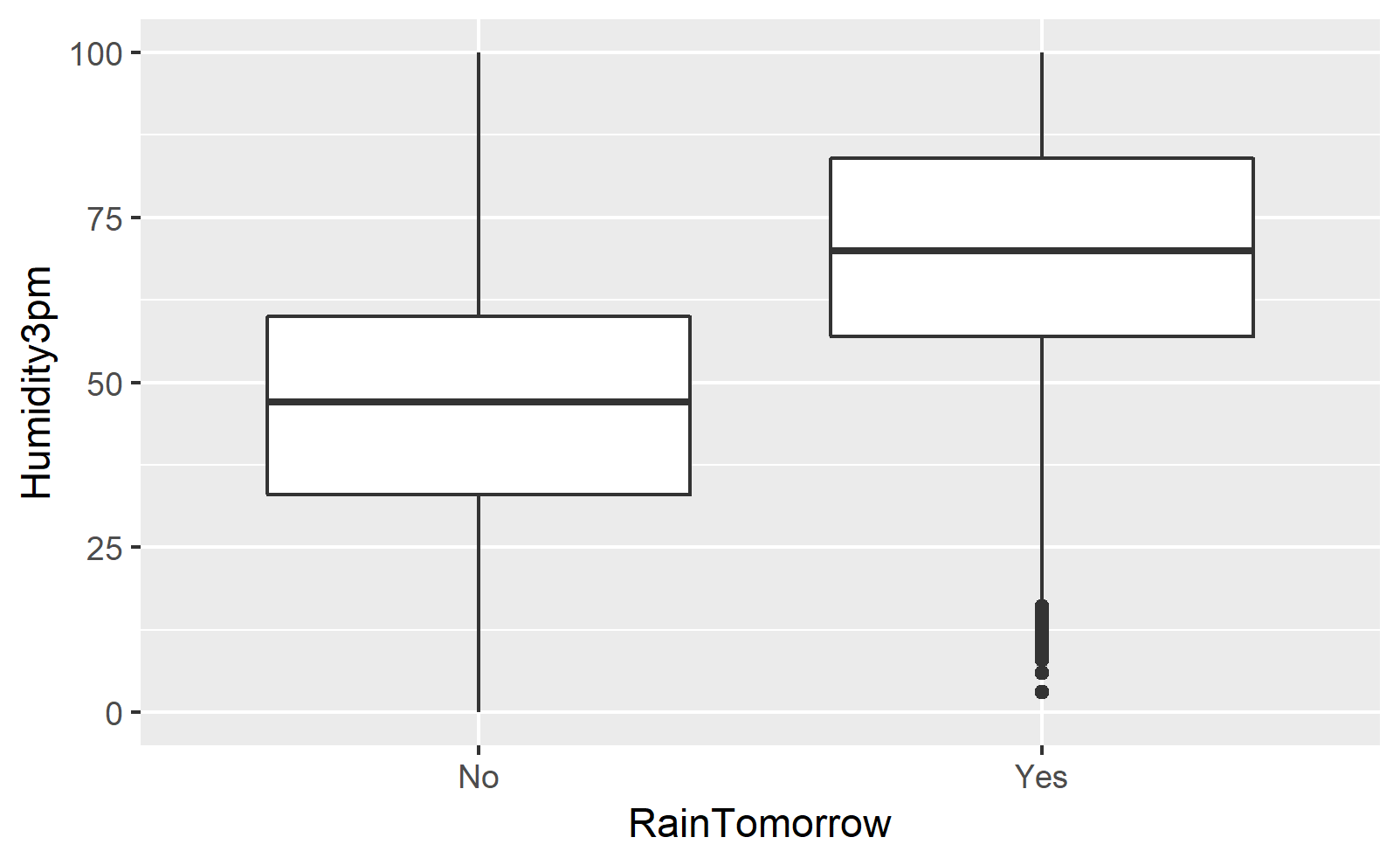


Figure 1: Humidity3pm x RainTomorrow Figure : Cloud9am x RainTomorrow

Several bar charts (with aesthetic fill) are also included in addition to the boxplots. The number of rain showers seems to vary with the months. There’s an uptick in rain showers in the summer months (May – July) with a slight and gradual decline moving into the Fall and Winter. Bar charts were also created to visualize Wind Gust Speed and Wind Direction in the morning and afternoon. At first glance, nothing really stands out or pops at the viewer, but if you look closely at Wind Direction at 9am, there’s a high volume of wind coming from the North (NNE, NE & NNW) which may provide insight into the geographical location of our datapoints by analyzing incoming storm directions against jet streams and the Coriolis effect.

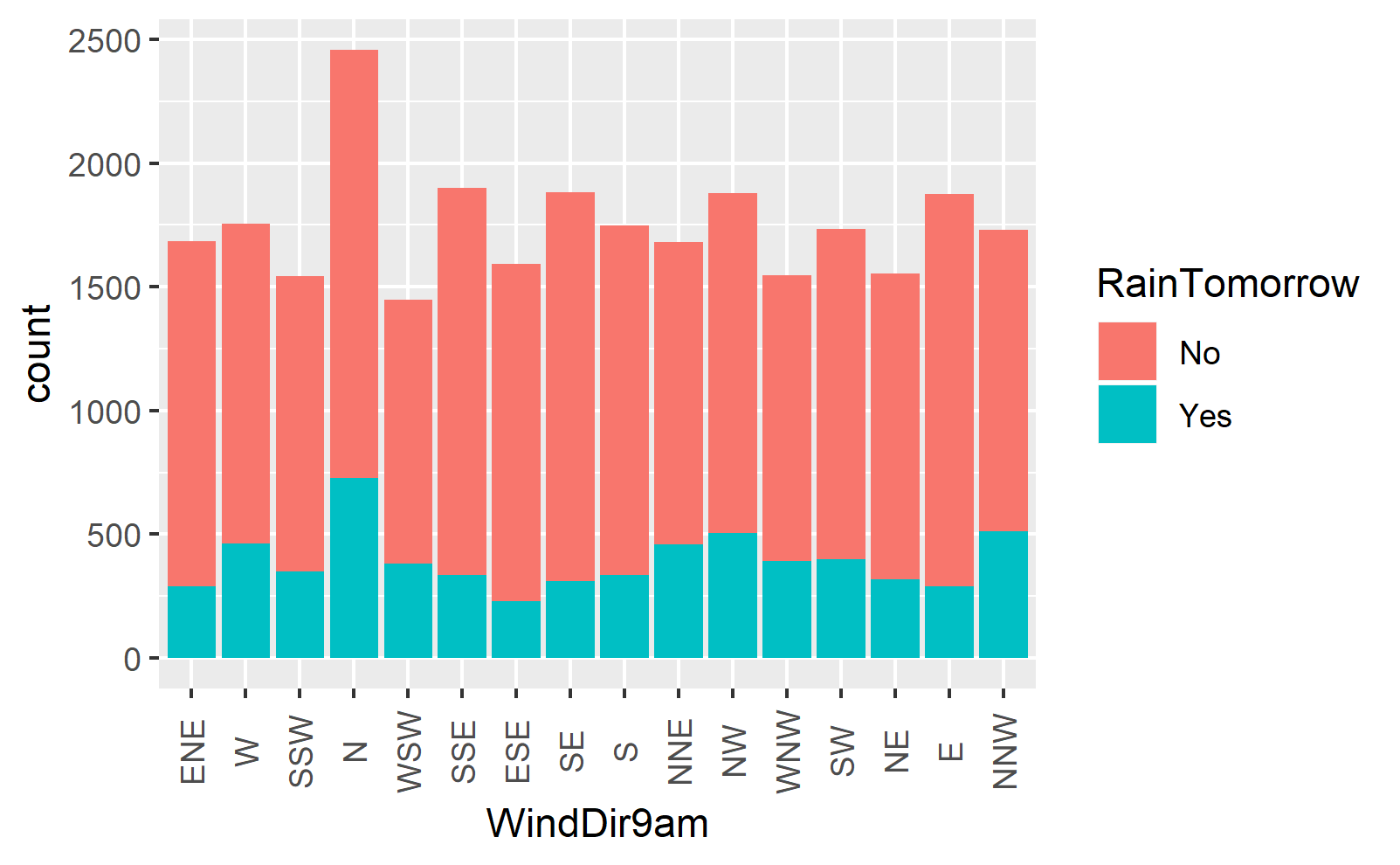


Figure 3: Wind Direction at 9am

To analyze our other variables and gain further insight into the data, I used stepwise regression. Significant variables include: *Months (April – Oct), MinTemp, MaxTemp, RainFall, WindGustSpeed, WindDir9am (N & NNE), WindDir3pm (E, ESE, ENE, NE, SSE, N), WindSpeed9am, Windspeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am and RainTodayYes.*

That’s nearly all of our variables except for Temp3pm, WindGustDir and Location (which was left out of our data). Days and Years were purposely excluded. The model makes sense for the most part. I’m a little surprised that Temp3pm was left out of the optimal model. I thought there would be a stronger correlation after looking at the boxplots. I’m also a little wary of the RainTodayYes variable because an overnight storm system can trigger both variables.

Stepwise regression verified earlier predictions with respect to wind direction from the north at 9am. The North & North-Northeast factors from our WindDir9am variable are apparent in our visualizations as well as our regressions.