Kellie McLiverty

Dr. Hill

BAN 502

Course Project Part 1

Data Exploration, Preparation, and Visualization

For this portion of the Course Project, I was required to gain an understanding of the Rain dataset and of the relationships the dataset variables have with the response variable (RainTomorrow). In this write up, I will explain the methodologies I employed to make sense of the data and perform the necessary lines of R code to visualize the variable relationships. Please note, for the sake of computing time and space, the is.na(rain) part of the section titled ***Testing for Missing Data*** in the R Markdown document has been commented out during the knitting process as it added over 100 extra pages of notes.

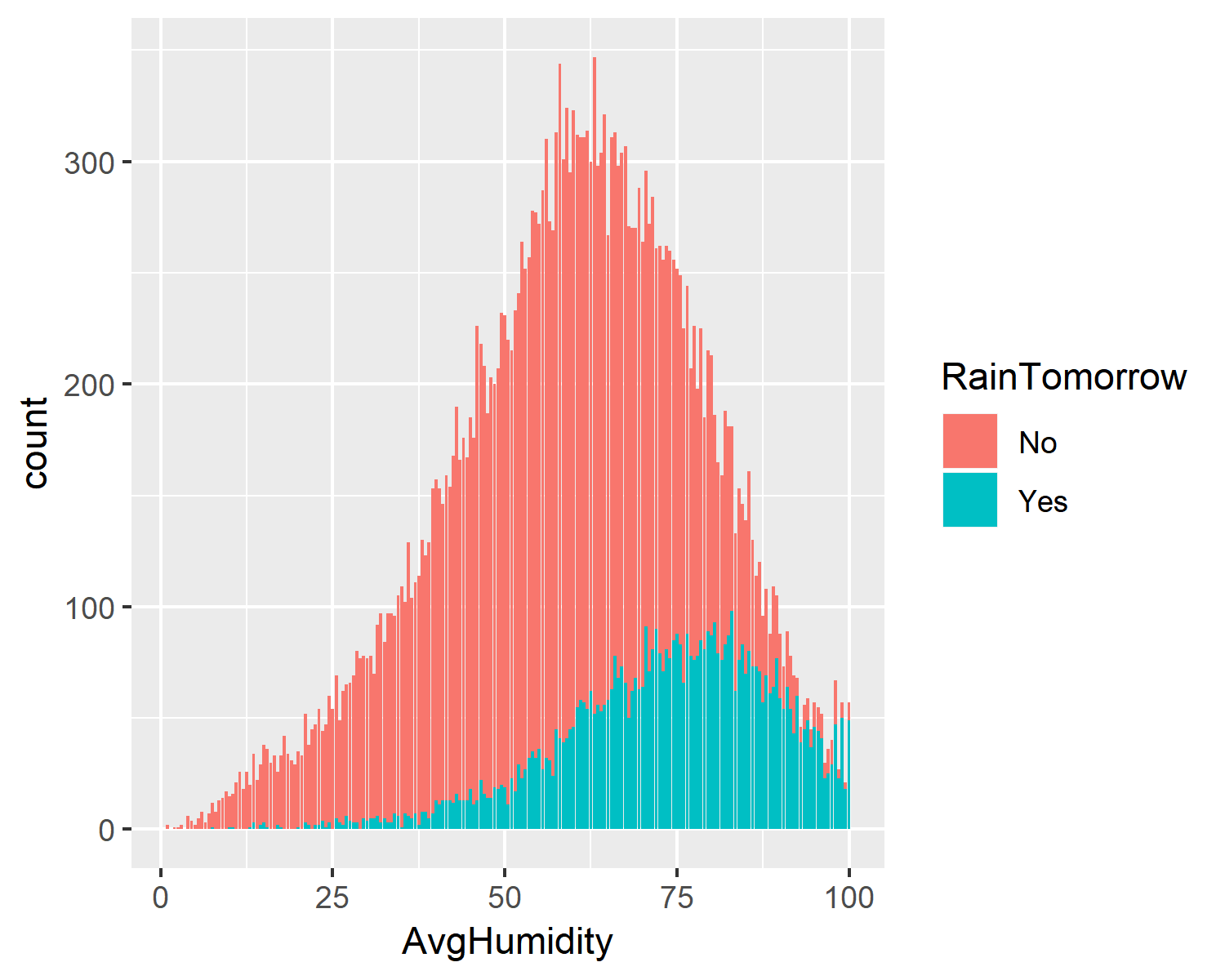
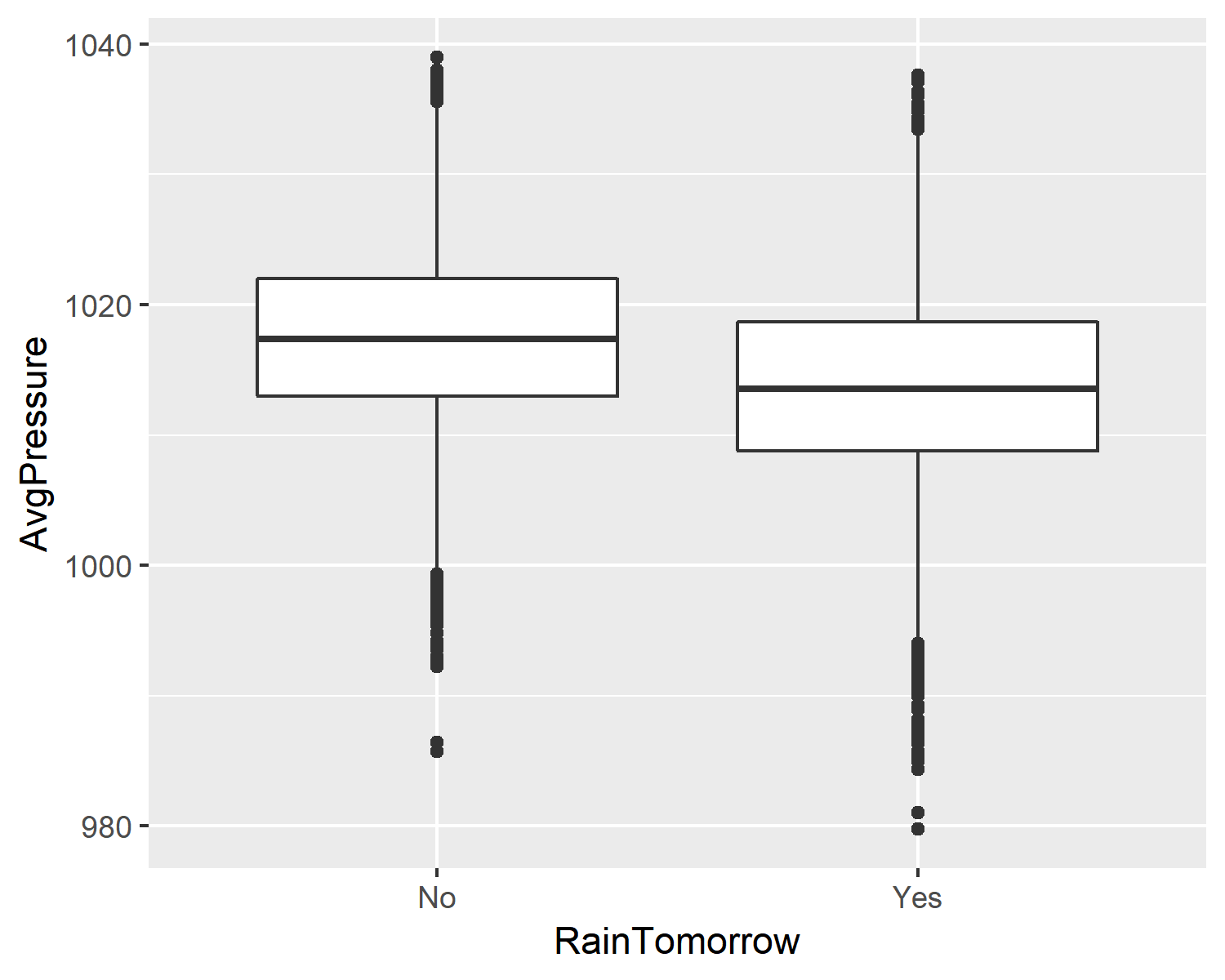
To begin this project, I read in the Rain dataset and reviewed its structure, looking over each of the variables to find which ones needed recoding. For this dataset the Date, RainToday, RainTomorrow, WindGustDir, WindDir9am, and WindDir3pm variables needed to be recoded into factors. This was done by using the mutate function of the dplyr package. Then again, I reviewed the structure of the data to insure conversion took place. From here, I moved on to tackling the missingness of the data.

Using the is.na(rain) command in the ***Testing for Missing Data*** section of the R Markdown, I discovered there was a lot of missing data in the rain dataset. To have an accurate dataset, I needed to remove this missing data. I executed a vim\_plot to find the columns for Cloud Coverage at 9 am and 3pm were missing loads of data to the point where using Row-wise deletion to solve it would deplete the dataset. As such, I used Column-wise deletion to remove these two columns and the Date column, which is a relatively useless variable for the analysis.

After these columns were deleted, I created another vim\_plot to look for more missing data. As there was still missing data present, I decided the next step would be to use the “mice” package to do imputation on the rest of the missing data. I chose this method instead of the brute force method of using na.omit to preserve the integrity of the dataset, as it would remove a vast number of rows from the dataset. To start the imputation process, I went in and used mutate to aggregate the Temp, Humidity, Pressure, and Windspeed variables into new variables that were an average of the 9am and 3pm observations. As I want to work with as few variables as I can for imputation, this process made the most sense to do, since separately these variables would be extremely time consuming to work with, but together they may provide useful data.

I created a new data frame called rain3 that only included the RainTomorrow, RainToday, Rainfall, MaxTemp, MinTemp, WindGustDir, WindGustSpeed, AvgHumidity, AvgPressure, AvgWindSpeed, and AvgTemp variables. I used this data frame to go through the imputation process of the data. Finally, I merged the imputed values into a new data frame called rain\_complete. To see if my work was correct, I created one last vim\_plot to check for any missing data. The Dataset was now clean of missingness and prepared to be used for visualizations.

Continuing, I began to look for important variables that would have significant relationships to my response variable. First, I ran ggpairs to visualize the correlating relationships in the data, please see ***Looking for important variables*** section of R Markdown. After reviewing the output, I used ggplot to examine individual variable relationships to the response variable. The initial analysis of the graphs indicated to me that RainToday, AvgHumidity, AvgPressure, MinTemp, MaxTemp, WindGustSpeed, WindGustDir, and Rainfall could be strong predictors of RainTomorrow. Below is the bar chart for AvgHumidity and box plot for AvgPressure.



Since there is the potential that I misread a graph or overlooked a variable, I decided to examine the relationships further to be sure of my initial findings. To do this I created a Backward Stepwise Model to test for significant relationships between the variables and the response variable. Examining the coefficients, I concluded I did indeed misread some of my ggplots. The model showed that the variables RainToday, Rainfall, MaxTemp, MinTemp, WindGustSpeed, AvgHumidity, AvgPressure, AvgWindSpeed, and AvgTemp are significant indicators of if it will RainTomorrow. However, though it did show up in the model, WindGustDirection is conditionally significant to the response variable, in that only three directions [NNW, N, SSW] showed to be significant predictors of the response variable, with the N Wind having the most significant impact.

To conclude this portion of the project, I prepared my data for the second part of the analysis by creating my training and testing split using a set.seed of 1234. In the next deliverable, I will be creating prediction models for the response variable RainTomorrow using the variables I discovered here.