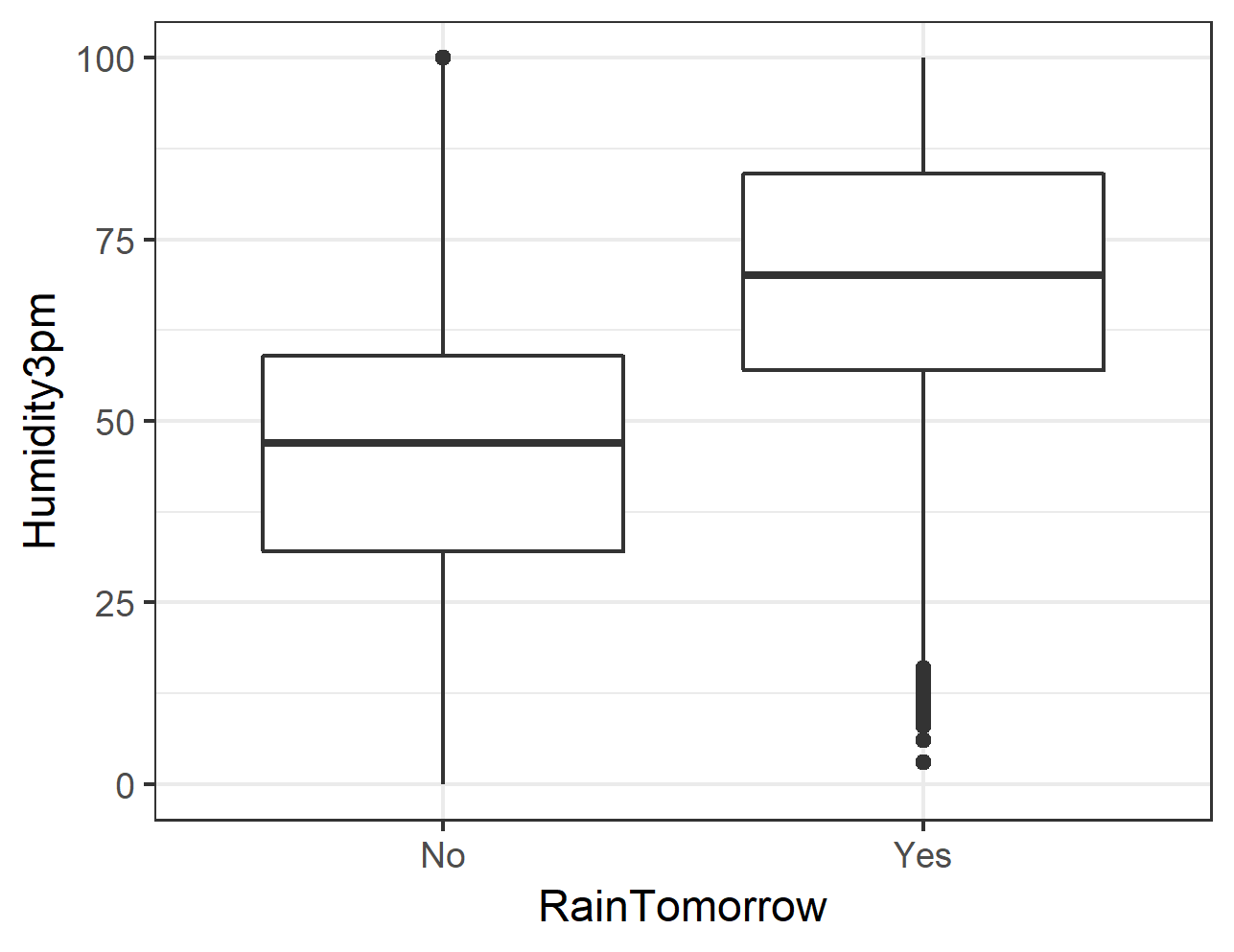
When observing the rain dataset we see it contains 28,003 observations with 20 variables. The variables are date, minimum temperature, maximum temperature, rainfall, wind gust direction and wind gusts speed (as well as the specific wind direction and speed for both 9 AM and 3 PM). The other variables are humidity, atmospheric pressure, fraction of the sky obscured by clouds and temperature for 9 AM and 3 PM as well as the precipitation level for that day and the prediction of whether it will rain tomorrow. Of these 20 variables 14 have missing values.

After observing the graph of missingness we see that cloud 9 AM and cloud 3 PM are missing almost 40% of their data. Since a large portion of these variables are missing, if we replace the missing values through imputation, we risk affecting how these variables will interact with predicting whether there will be rain tomorrow or not. Therefore, for both cloud 9 AM and cloud 3 PM I am opting for column wise deletion. We can then observe the missing values for the categorical variables. Noting that these missing values are less than 15% of the data I opted for row-wise deletion. The reason I opt for deletion is that imputing the missing variables for wind gust direction, wind direction 9am, wind direction 3 pm and rain today with the most commonly occurring value could affect the data in the prediction for whether it will rain tomorrow. The variables with the next highest proportion of missing data are pressure 9 AM and pressure 3 PM. These are followed by humidity 9 AM and humidity 3pm. Each of these variables are missing 10% of their data or less. Even though these variables are missing a small proportion of their data removing them would delete potentially valuable information therefore I have opted for imputation with the mean value.

When observing the density plots for the variables with missing values we see that some of the plots had more realistic results for their variable than others. For example, the density plots for pressure 9 AM, pressure 3 PM and humidity 3 PM appear to match the actual distribution of the original data for that variable and have an extremely similar shape and pattern indicating realistic values for those variables. The density plots for the minimum temperature, maximum temperature and humidity 9 AM have shapes that are overall similar to the original data, however, the imputed values are slightly less realistic. The density plots for temperature 9 AM and temperature 3 PM do not have a similar shape to the original data indicating that the imputed values for these variables are not realistic based on our original values. Since these density plots did not do a good job representing the distribution of the actual dataset, I decided to opt for row wise deletion for temperature 9 AM and temperature 3 PM to help prevent any potential effects on the data and subsequent analysis.

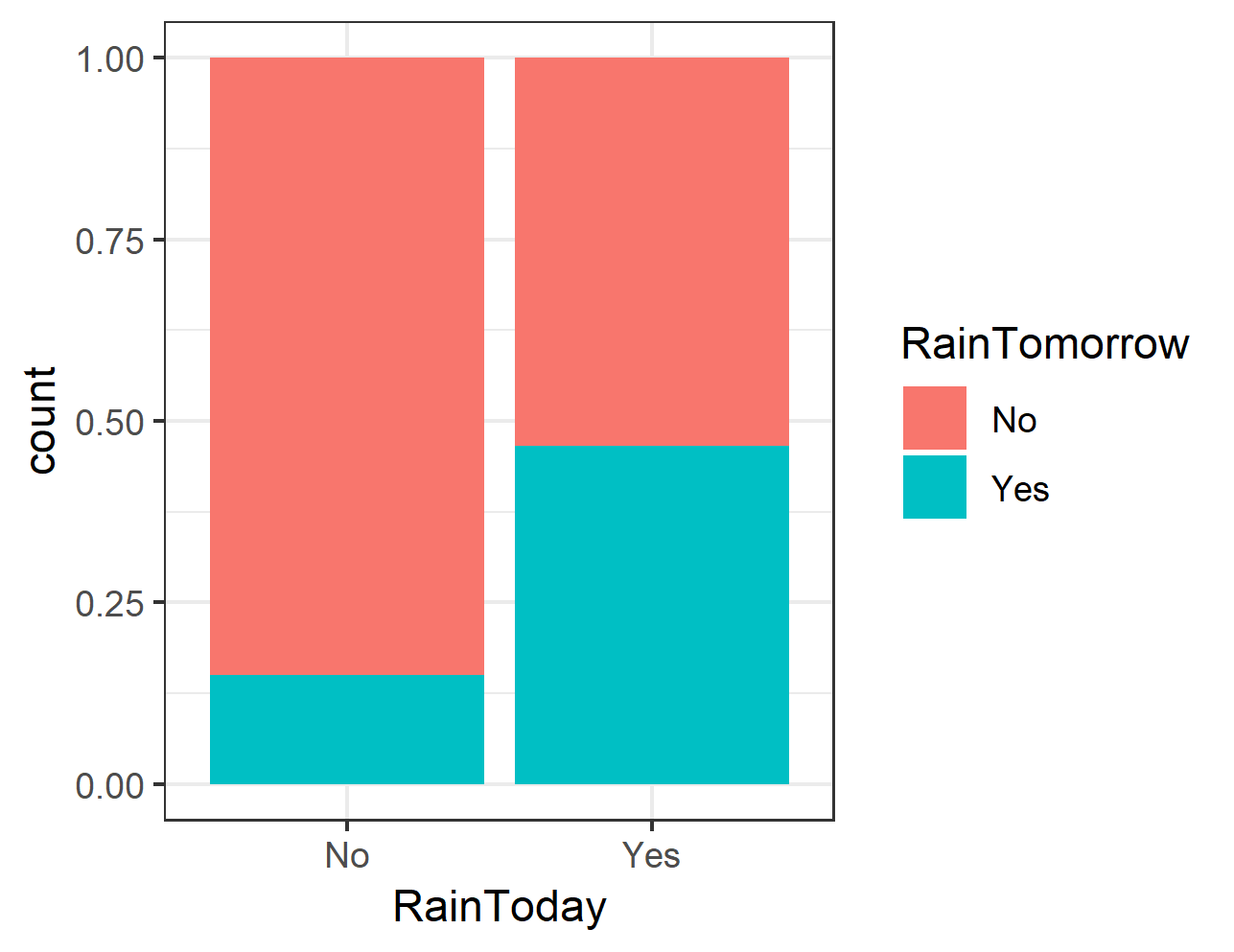
I then created visualizations and tables to identify which variables appear to be the most predictive of the variable rain tomorrow. When observing the box plot for minimum temperature versus rain tomorrow we see that on days where it did rain the next day the minimum temperature was slightly higher than on days where it did not rain the next day. This indicates that minimum temperature slightly affects and might be able to predict whether or not it will rain the next day. The box plot for maximum temperature versus rain tomorrow illustrates that when it does not rain the next day the maximum temperature is higher and when it does rain the next day, the maximum temperature is lower. This indicates that maximum temperature does affect and might be able to predict whether or not it will rain tomorrow. After observing the box plot for rainfall versus rain tomorrow we see that on days where it did rain the next day, the rainfall was slightly higher than on days where it did not rain the next day. This indicates that rainfall may affect and might be able to predict if it will rain tomorrow.

When observing the proportion bar plot for wind gust direction to predict if it will rain tomorrow or not, it appears that the wind direction does seem to slightly affect whether or not it will rain the next day. It appears that when the wind direction is North, north north west, north west, west and west north west it is more likely to rain the next day then if the wind is coming from any other direction. After creating bar plots and tables for the variables wind direction 9 AM and wind direction 3 PM we see a similar trend. Based on the bar plots and the tables it appears that wind gust direction, wind direction 9 AM and wind direction 3 PM may help predict whether or not it will rain tomorrow.

The box plot for wind gust speed versus rain tomorrow indicates that wind gust speed does predict if it will rain tomorrow. The box plot illustrates that when it does rain the next day the wind speed is higher versus when it does not rain the next day the wind speed is lower. After creating a boxplot for the variables wind speed 9 AM and wind speed 3 PM we note a similar trend. This indicates that wind gust speed, wind speed 9 AM and wind speed 3 PM might be able to predict whether or not it will rain the next day.

The box plots for humidity 9 AM and humidity 3 PM versus rain tomorrow indicate that the humidity level does predict whether or not it will rain tomorrow. The box plot illustrates that when it does rain the next day the humidity is significantly higher and when it does not rain the next day, the humidity is lower. This indicates that humidity 9 AM and humidity 3 PM might be able to predict if it will rain tomorrow.

The box plots for pressure 9 AM and pressure 3 PM versus rain tomorrow indicate that the pressure level does predict whether or not it will rain tomorrow. The box plot illustrates that when it does not rain the next day the pressure level is significantly higher and when it does rain the next day the pressure level is lower. This indicates that pressure 9 AM and pressure 3 PM might be able to predict whether or not it will rain the next day.

When observing the box plot for temperature 9 AM versus rain tomorrow we see that the temperature at 9 AM on its own does not seem to predict if it will rain tomorrow. On the other hand the box plot for temperature 3 PM versus rain tomorrow indicates that the temperature at 3pm does predict whether or not it will rain tomorrow. This indicates that temperature may be able to predict whether or not it will rain tomorrow.

After observing the proportion bar plot for rain today to predict if it will rain tomorrow it appears that whether it rained today does affect if it will rain tomorrow. It appears that if it rained today it is more likely to rain tomorrow than if it did not rain today. After creating a table to see the proportions of the data we see a similar trend. About 46% of the time when it rained today it will rain tomorrow. Based on the bar plot and the table it appears that rain today can help predict whether or not it will rain the next day. Overall, based on the visualizations the variables that seem to be able to predict whether or not it will rain tomorrow, the best are maximum temperature, wind gust speed, humidity, pressure and rain today.