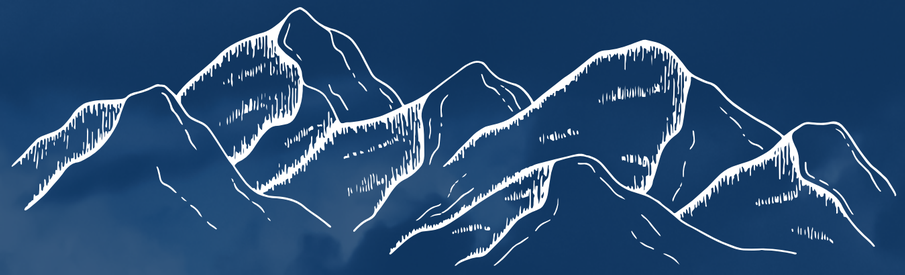
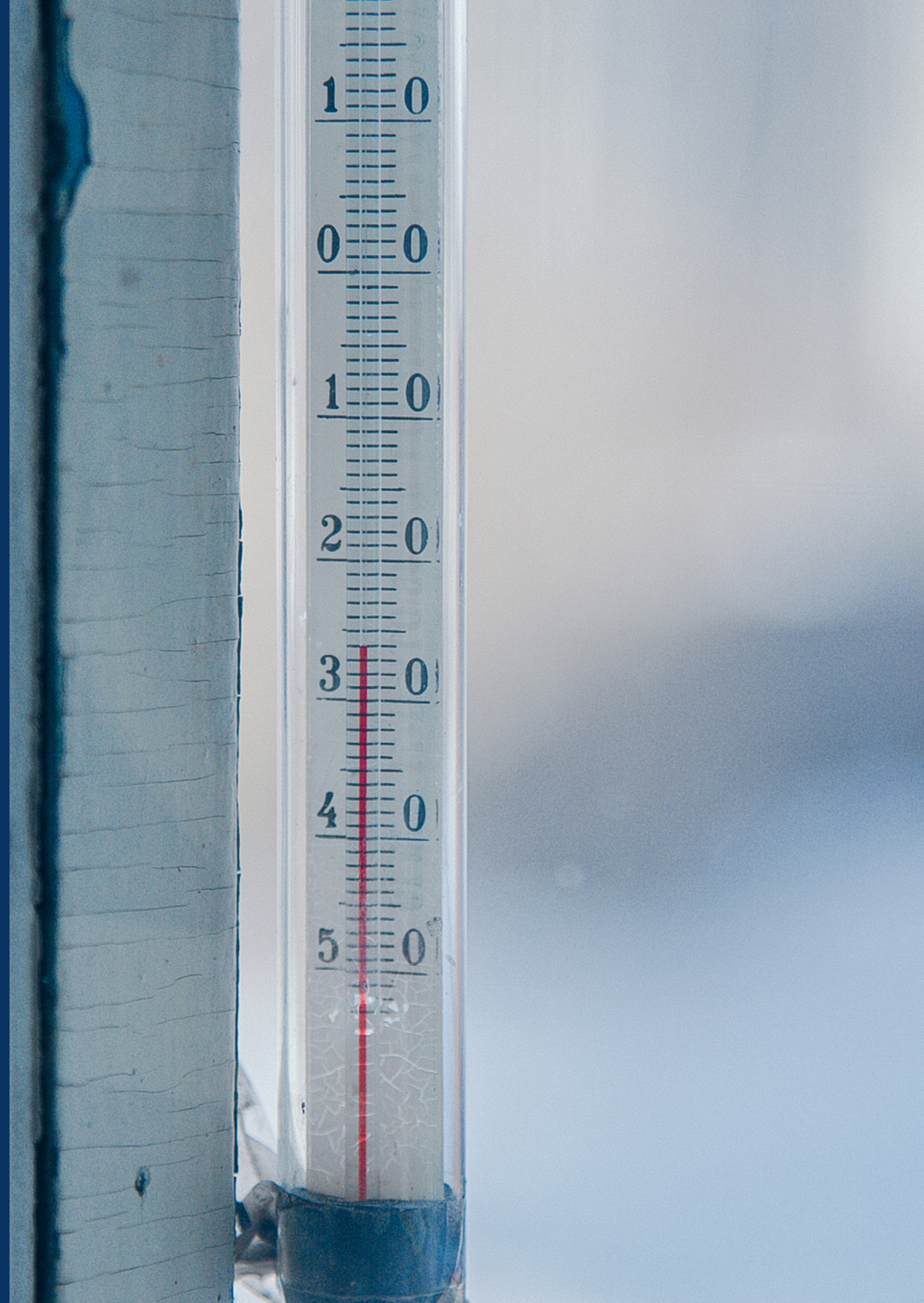


RAINFALL PREDICTION USING MODELS

GROUP 16
ANUSHA SATHRAVADA – AXS220402
NITISHA SREE VENKATESAN – NXV220065
KOVID REDDY VAYALPATI – KXV220038
RAKSHITH REDDY KOTURU – RRK230002

PROJECT CONTEXT

- Rainfall prediction is a complex task with considerable societal impact.
- Timely and accurate forecasting is crucial to proactively mitigate human and financial losses.
- Overall Goal: Develop a reliable model for informed decision-making regarding rain predictions.



ABOUT DATASET

- Dataset comprises daily weather records over a decade in various Australian locations.
- Objective: Develop a model predicting next-day rain occurrence based on daily weather data (Temperature, Sunshine, Wind speed, Humidity, Pressure) in major Australian cities.
- Key Variables: Temperature, Sunshine, Wind speed, Humidity, and Pressure play vital roles in the prediction model.
- Geographical Focus: Major cities in Australia serve as locations for data collection.

DATASET DESCRIPTION

- **Dataset:** Captures 10 years of daily weather data across multiple Australian stations.
- **Structure:** Comprises 23 columns and 145,460 records spanning 2008 to 2017.
- **Features:** Quantitative data includes max/min temperature, evaporation, sunshine duration, and wind speed.
- **Categories:** Categorical features encompass dates, locations, and wind direction.
- **Boolean Flags:** Indicators include RainToday and RainTomorrow, signaling rain occurrences.

PROJECT OVERVIEW

- Extensive exploration of the dataset is conducted, including summary statistics, visualization of relationships, and handling missing values.
- Three distinct models—Random Forest, Decision Tree, and k-Nearest Neighbors are implemented for rainfall prediction.
- A model that can predict whether it will rain tomorrow or not based on the weather data (Temperature, Sunshine, Wind speed, Humidity, Pressure) for that day in major cities in Australia.

USED LIBRARIES

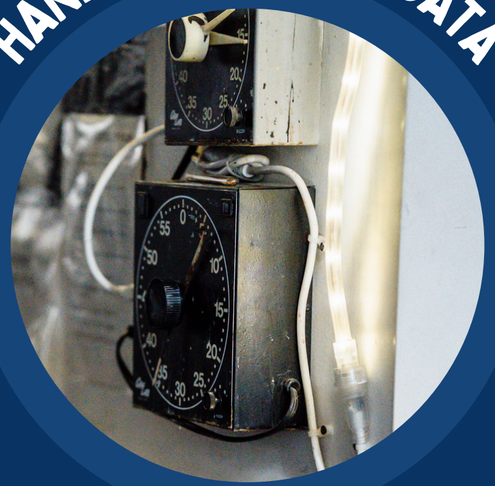
- **data.table** – for tabular data.
- **Tidyverse** – for data analysis.
- **gridExtra** – to combine multiple plots.
- **Plotly** – for creating interactive web-based plots.
- **Rpart** – for building the classification and regression trees.
- **rpart.plot** – for plotting the classification and regression trees.
- **randomForest** – for building ensemble of decision trees.
- **Dplyr** – for manipulating tabular data.
- **Ggplot2** – for data visualization.
- **Scales** - for percentage scales.
- **Neuralnet** – for neural networks.
- **Caret** – for data preparation model building and evaluation.

CLEANING AND MERGING DATASETS

MODIFICATIONS

Removed Date column as the dataset isn't time-series

HANDLING MISSING DATA



Dropped columns with NA rate below 1%, considering insignificance

QUANTITATIVE COLUMNS

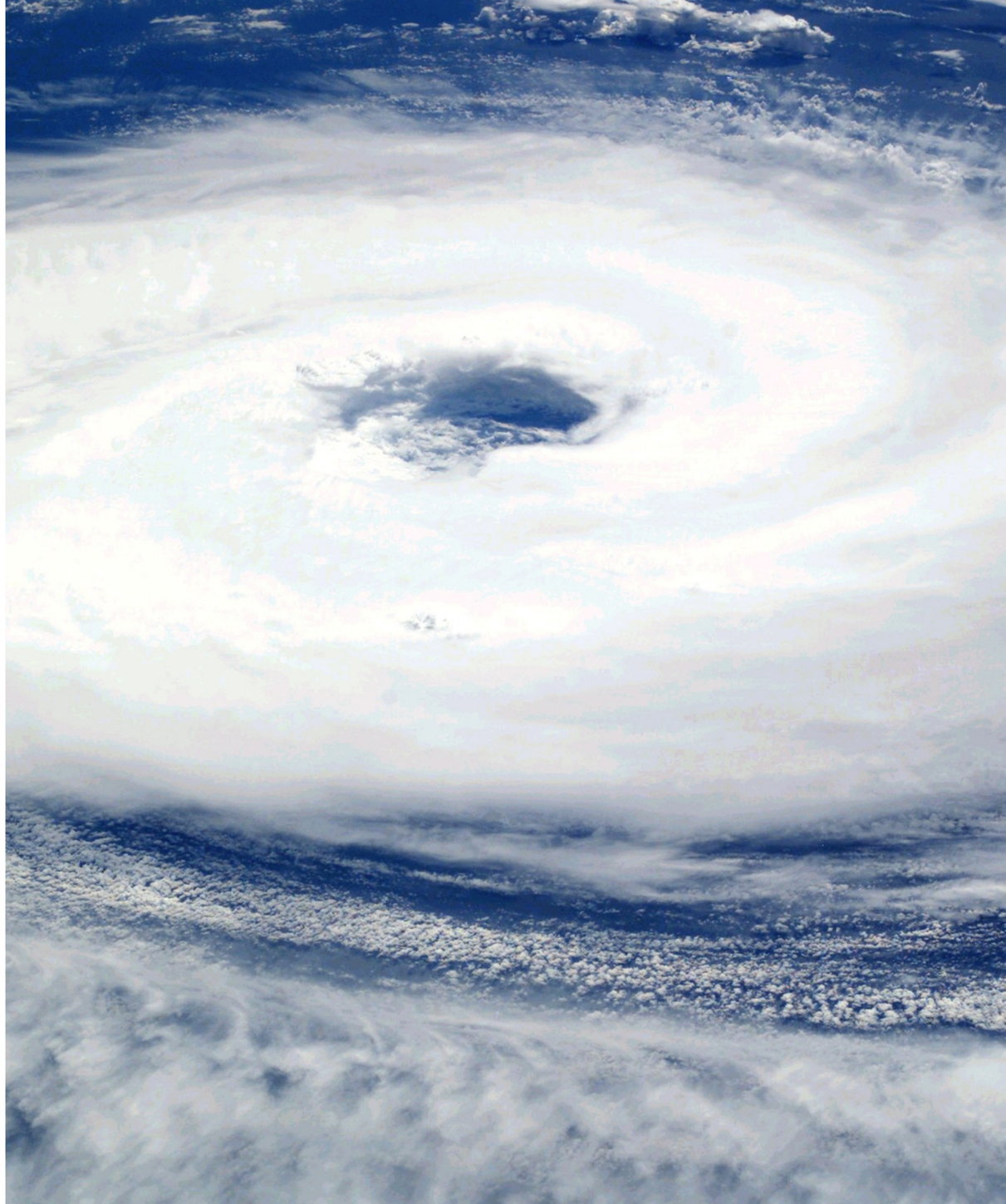
Used mean to fill NAs with 5%, 10%, or higher missing values.

CATEGORICAL COLUMNS

Replaced NA values with meaningful data, e.g., "No Wind" for WindDir9am.

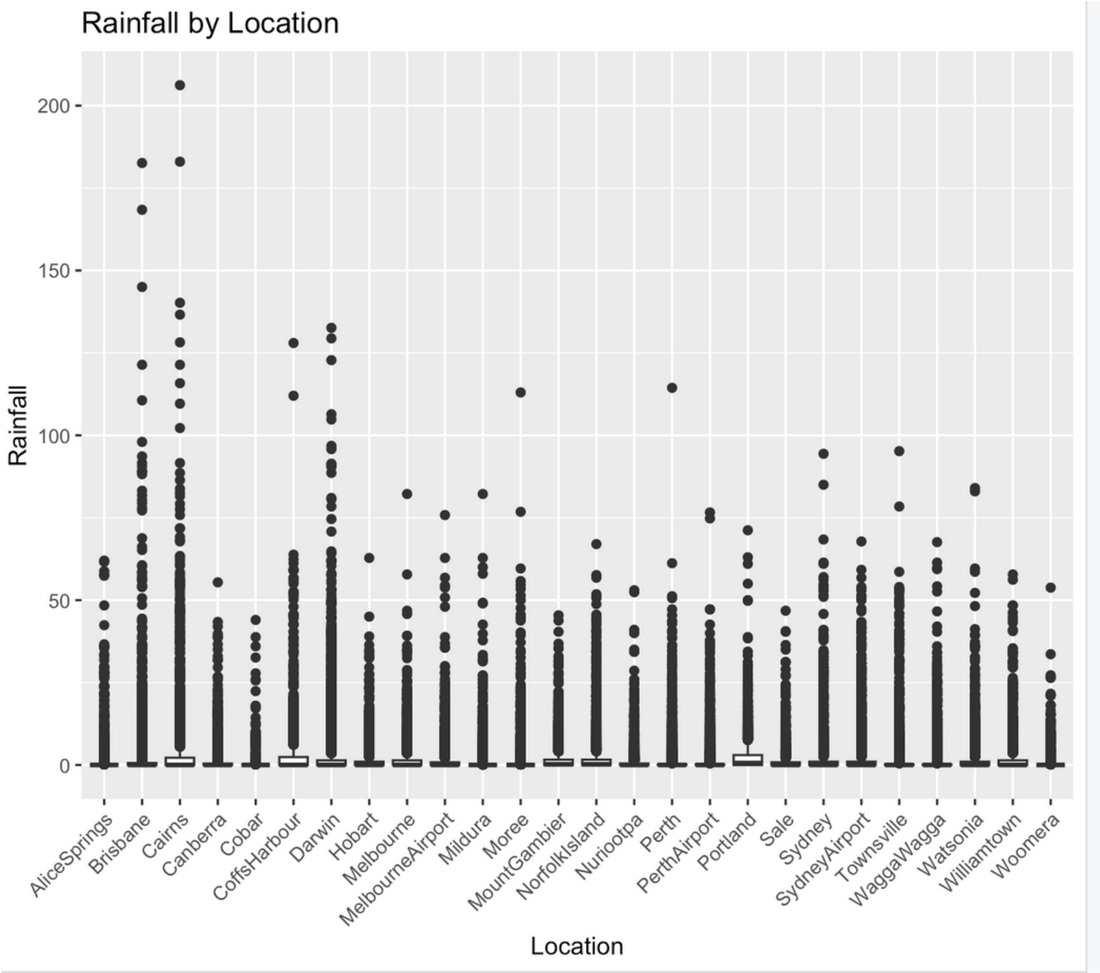
PROCESS FLOW

- Import libraries.
- Data importing.
- Data cleaning and merging.
- Data insights.
- Analysing the weather for rainfall prediction.
- Model Building.
- Forecasting weather for the next day.

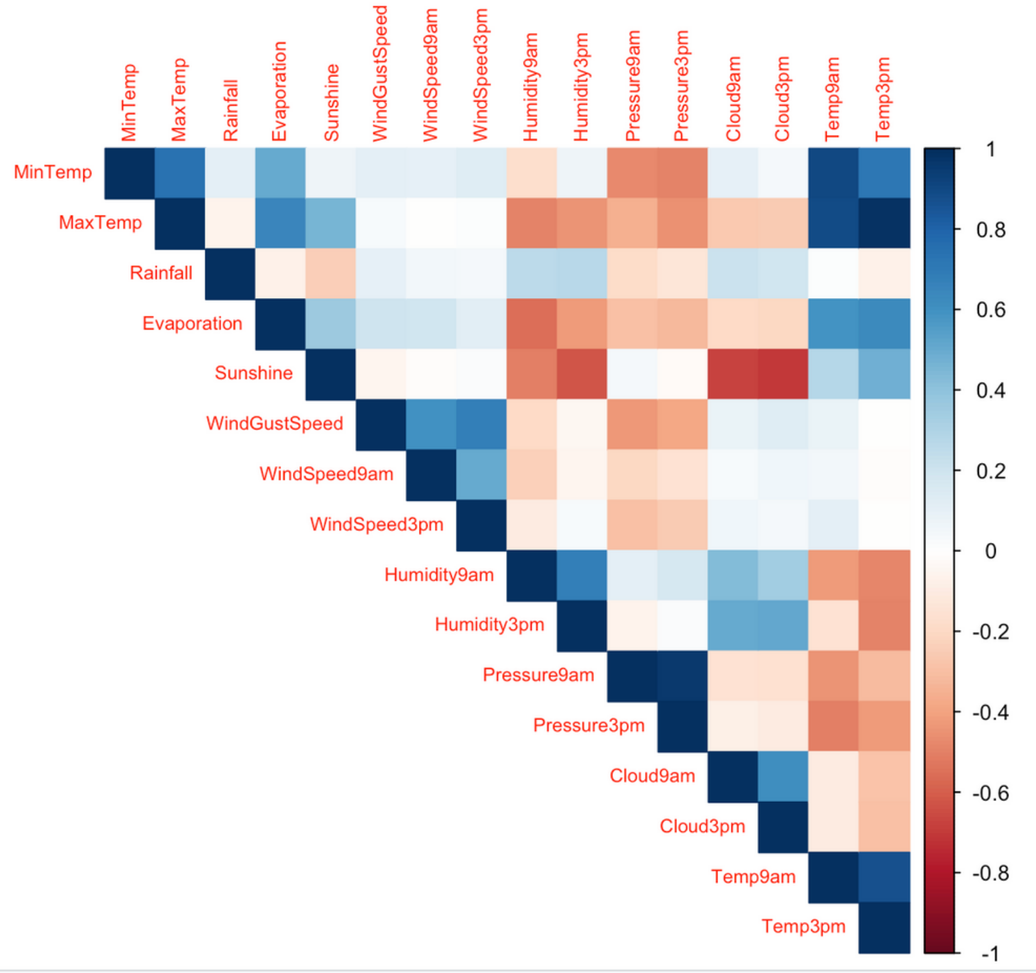


DATA INSIGHTS

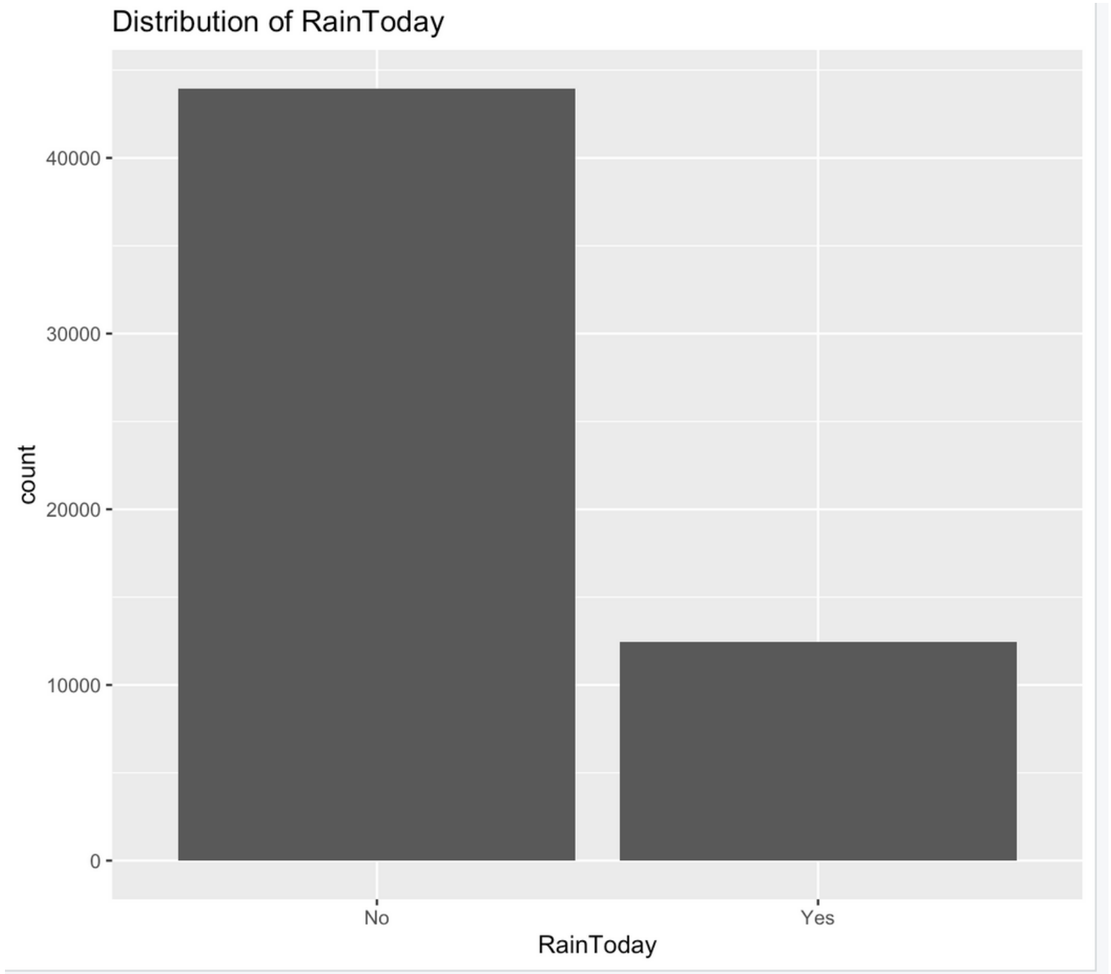
Rainfall By Location

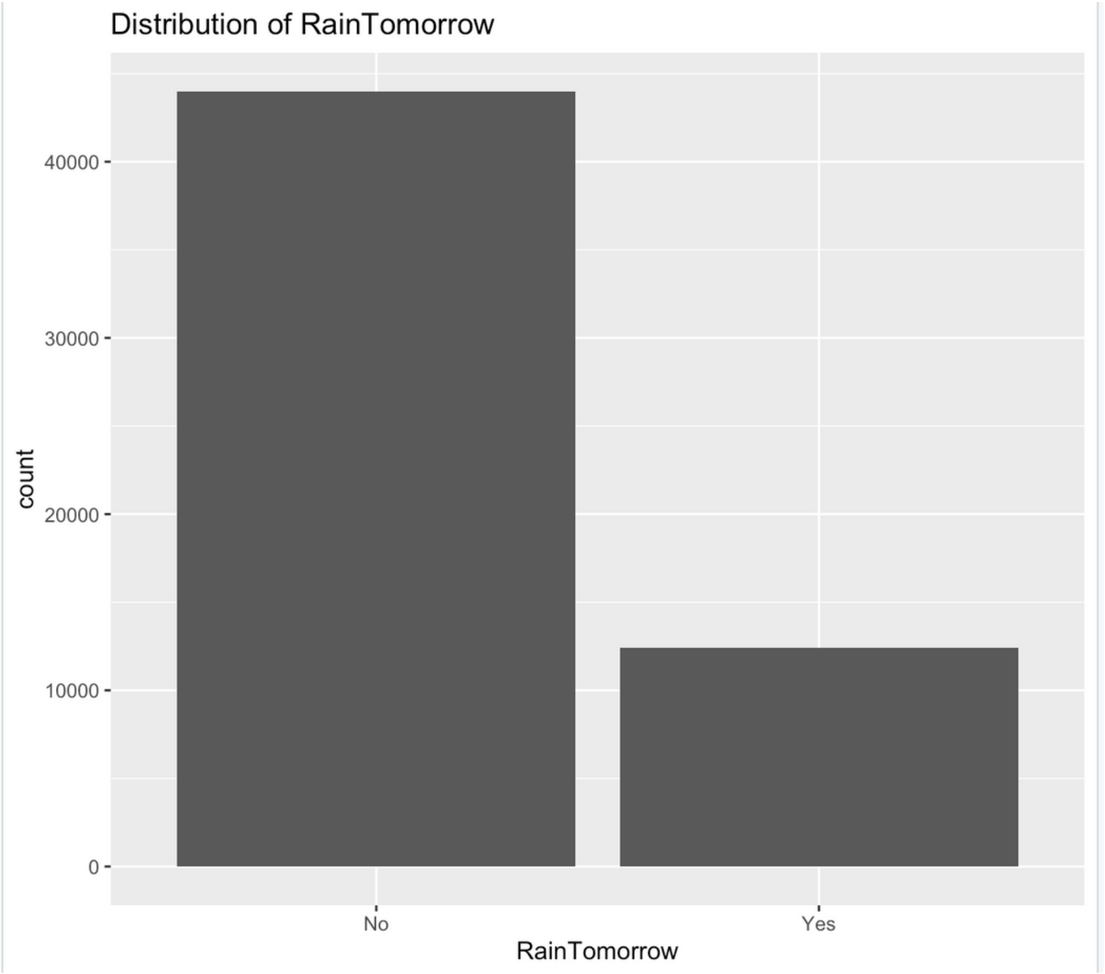


Correlation Matrix

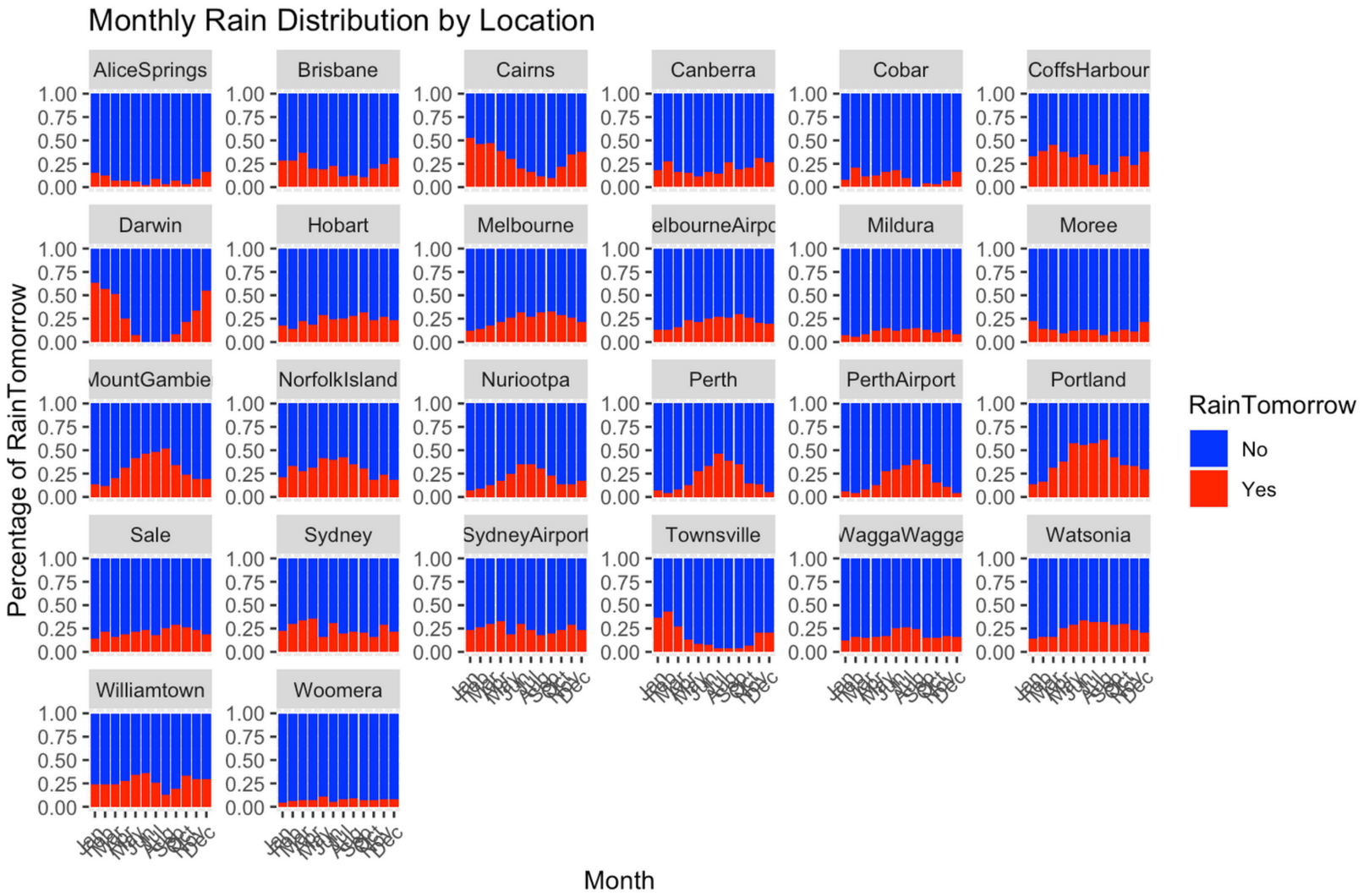


Distribution of Raintoday and Raintomorrow

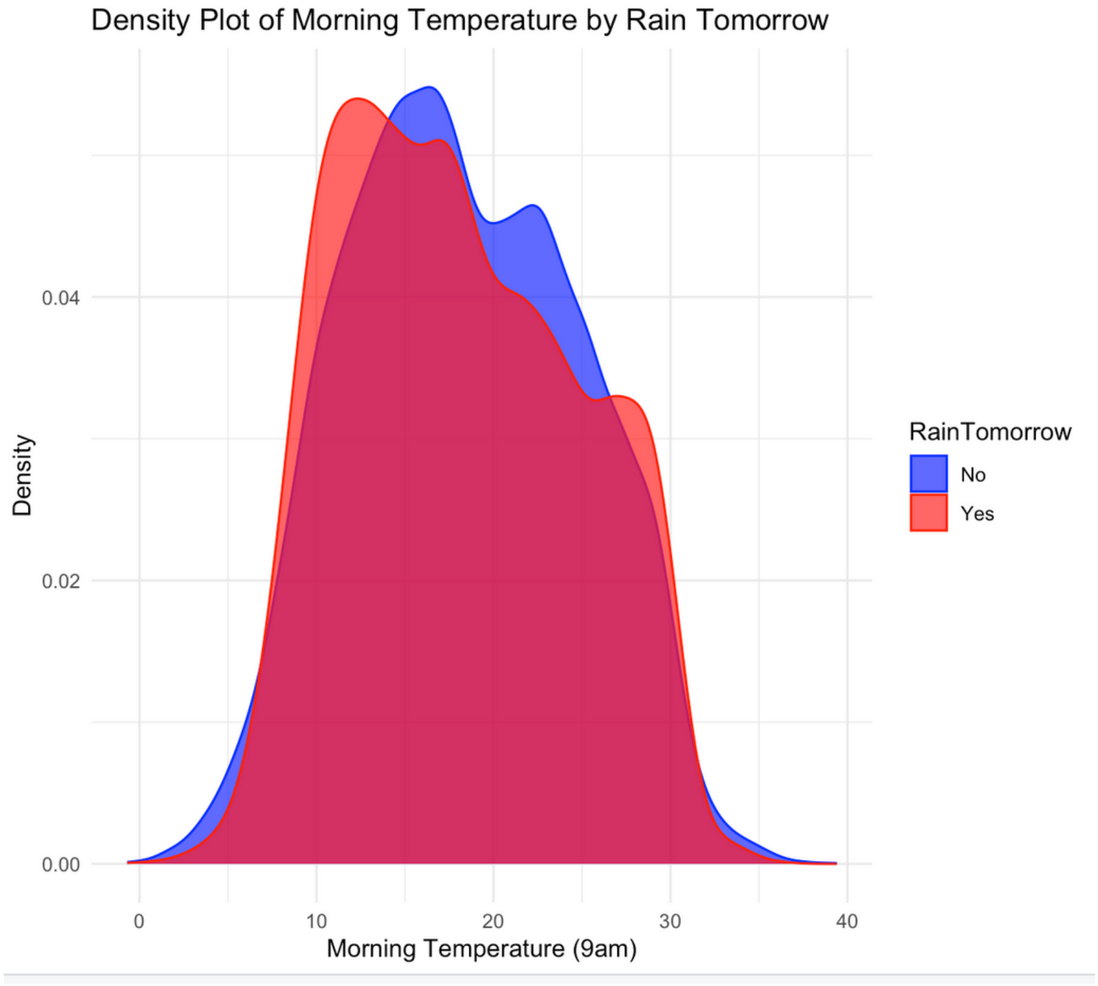


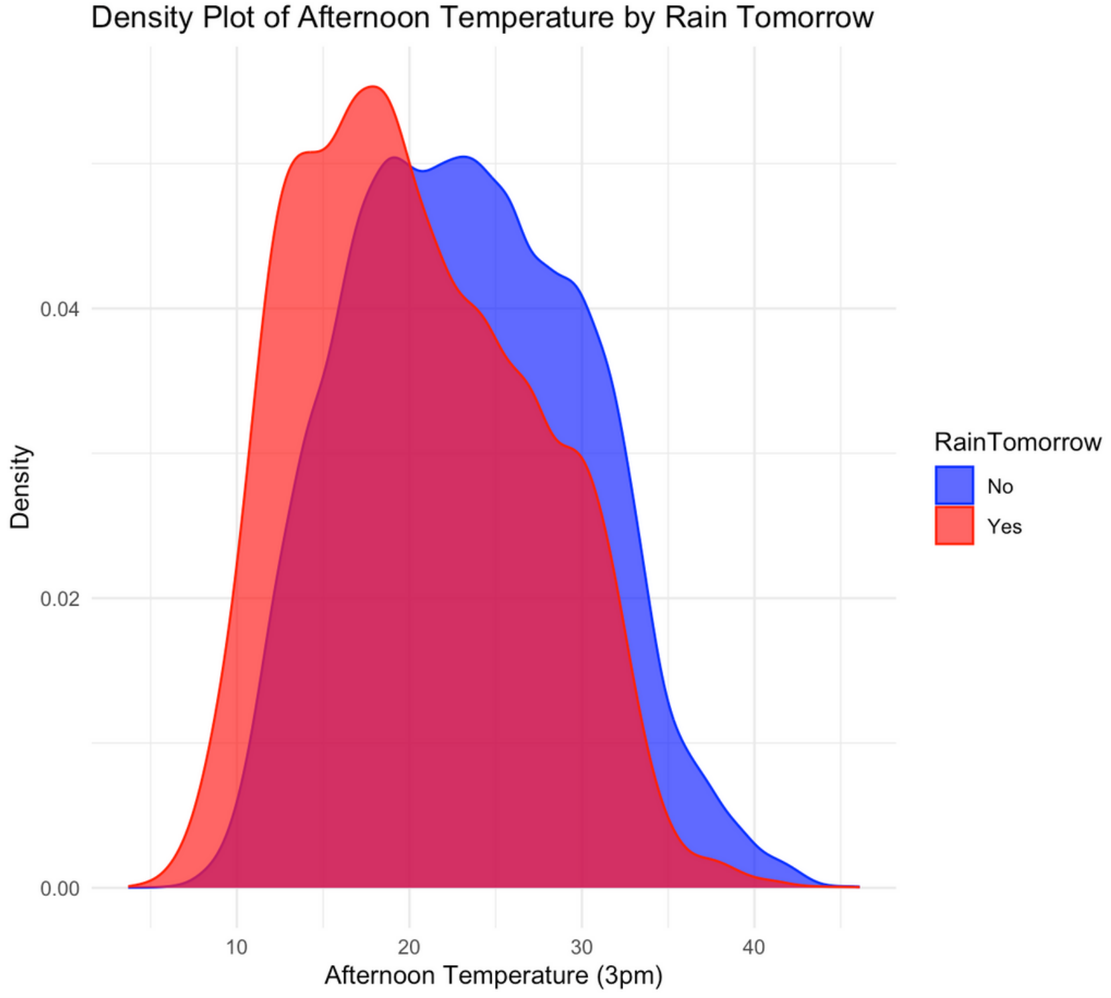


Monthly distribution of rain by Location

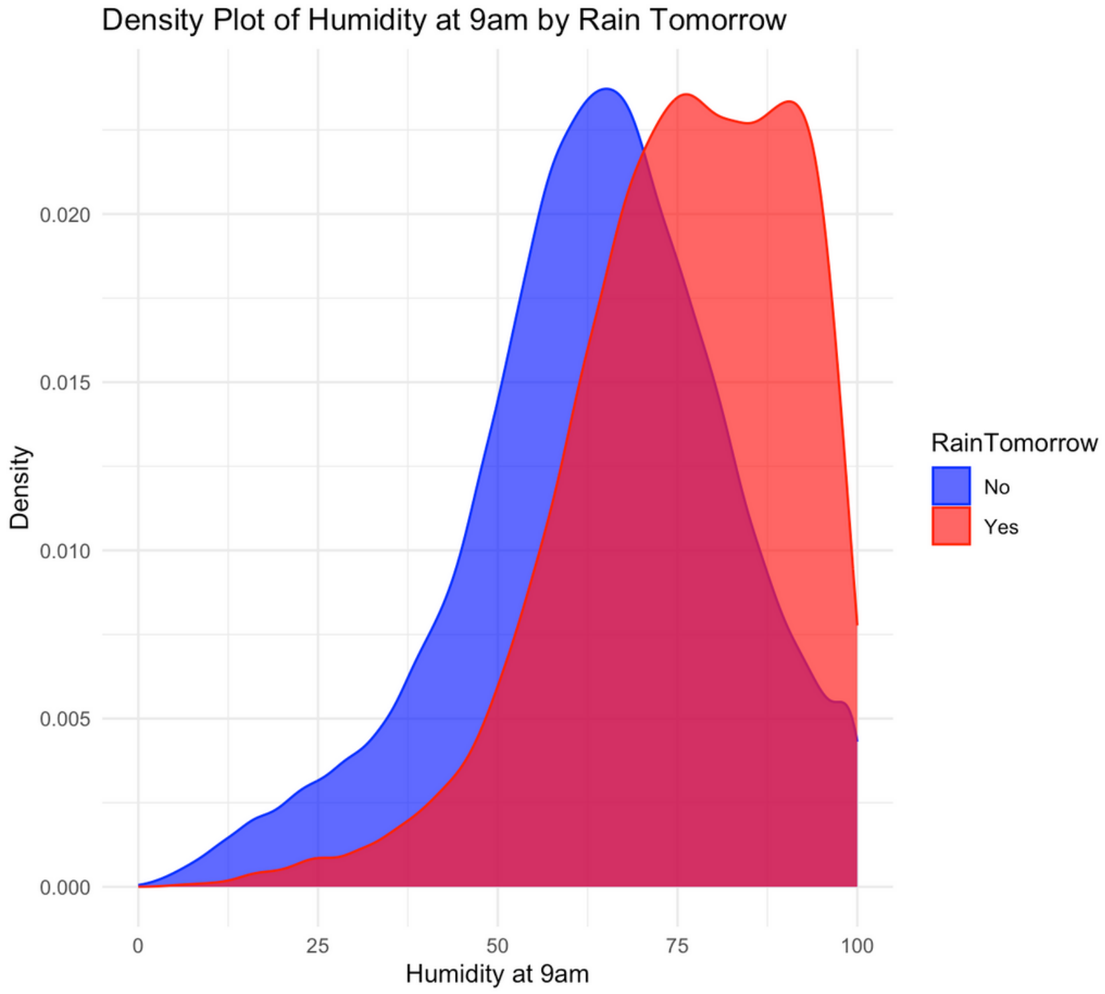


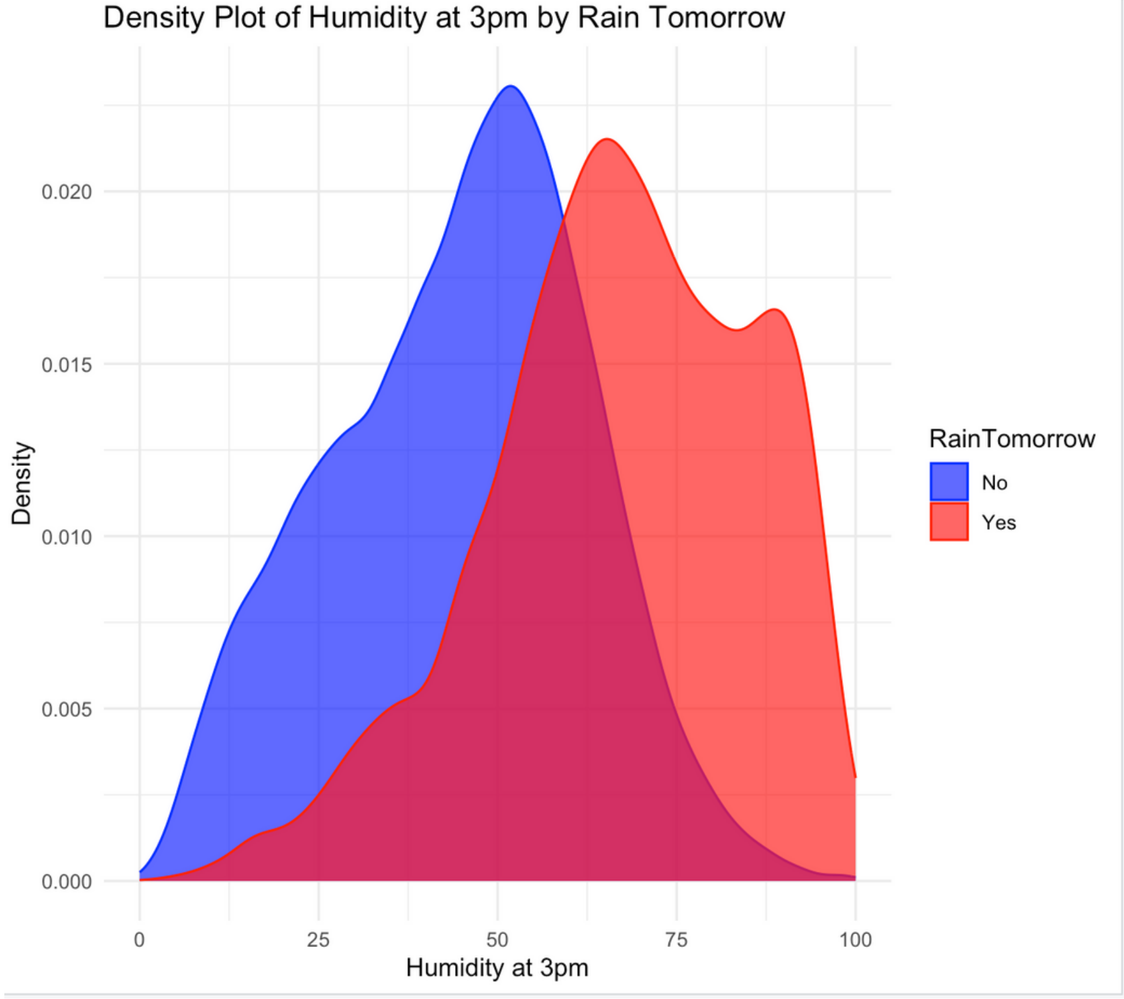
Density plot of Morning and Afternoon Temperature by RainTomorrow



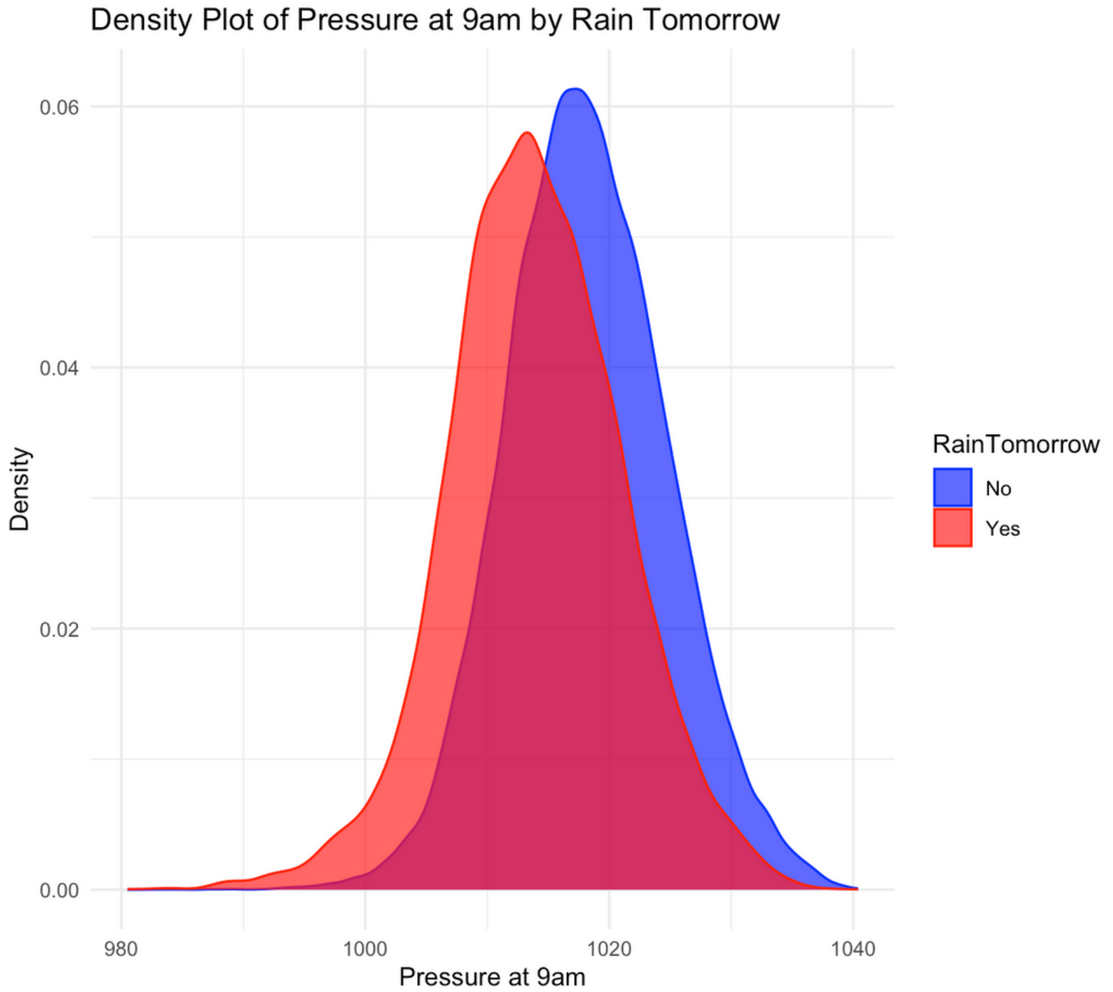


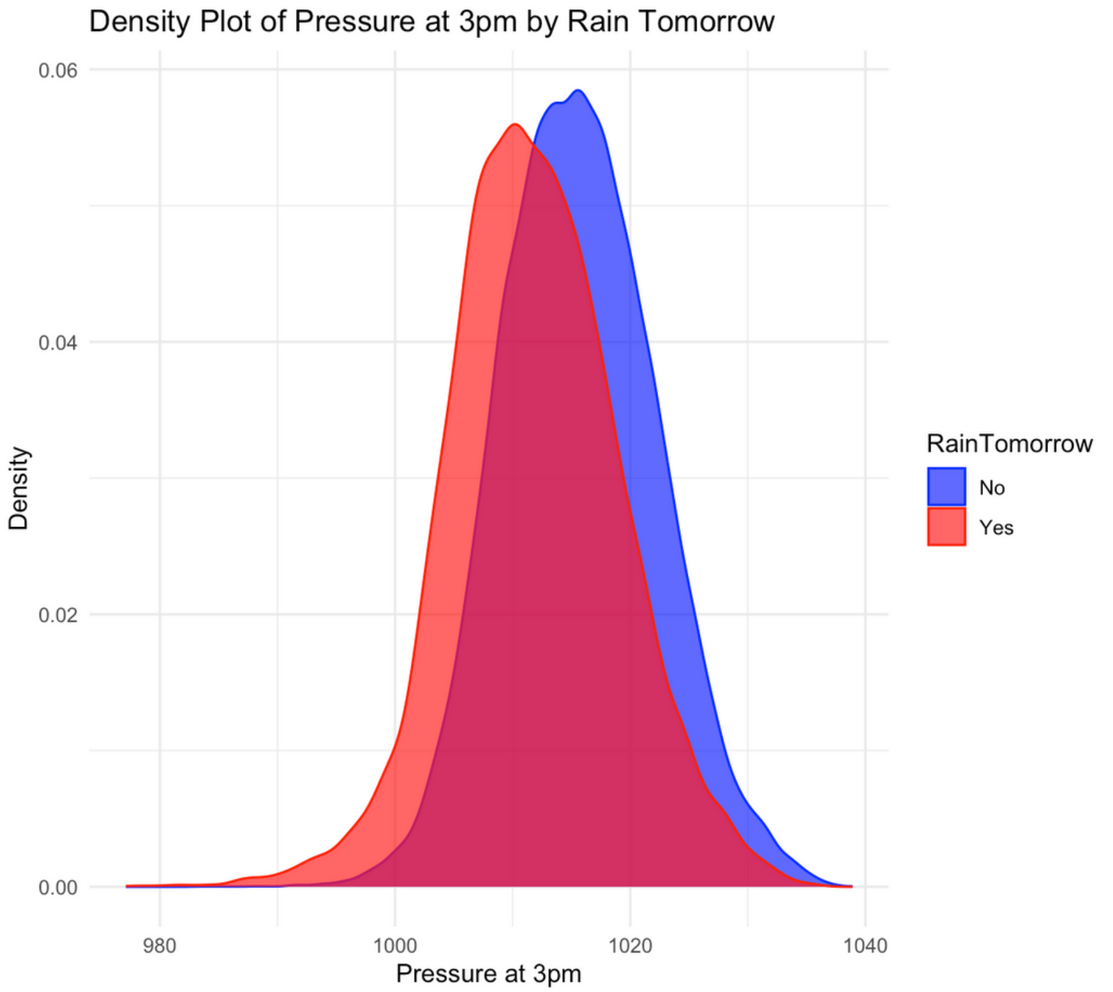
Density plot of Humidity at Morning and Afternoon by RainTomorrow



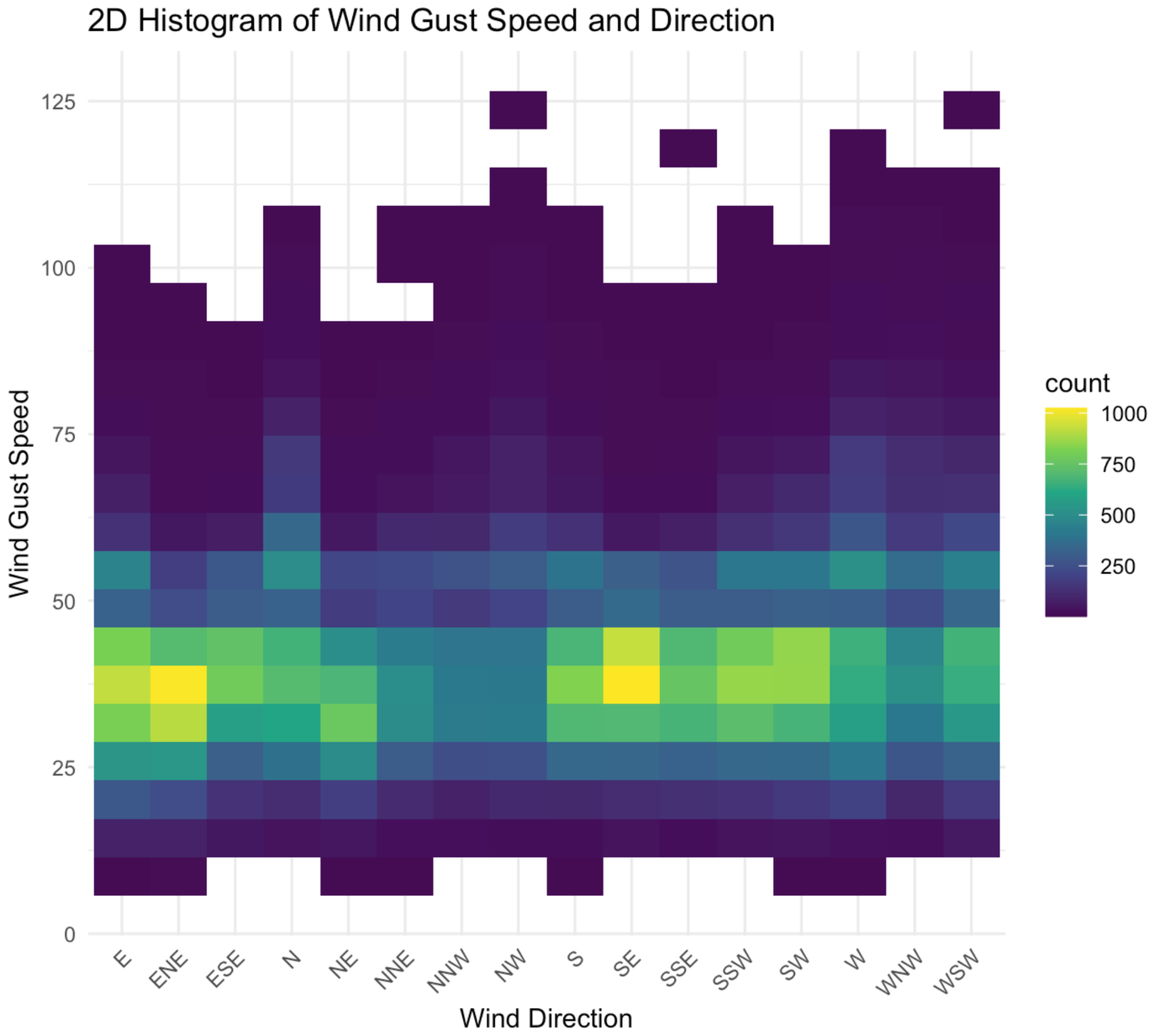


Density plot of Pressure at Morning and Afternoon by RainTomorrow

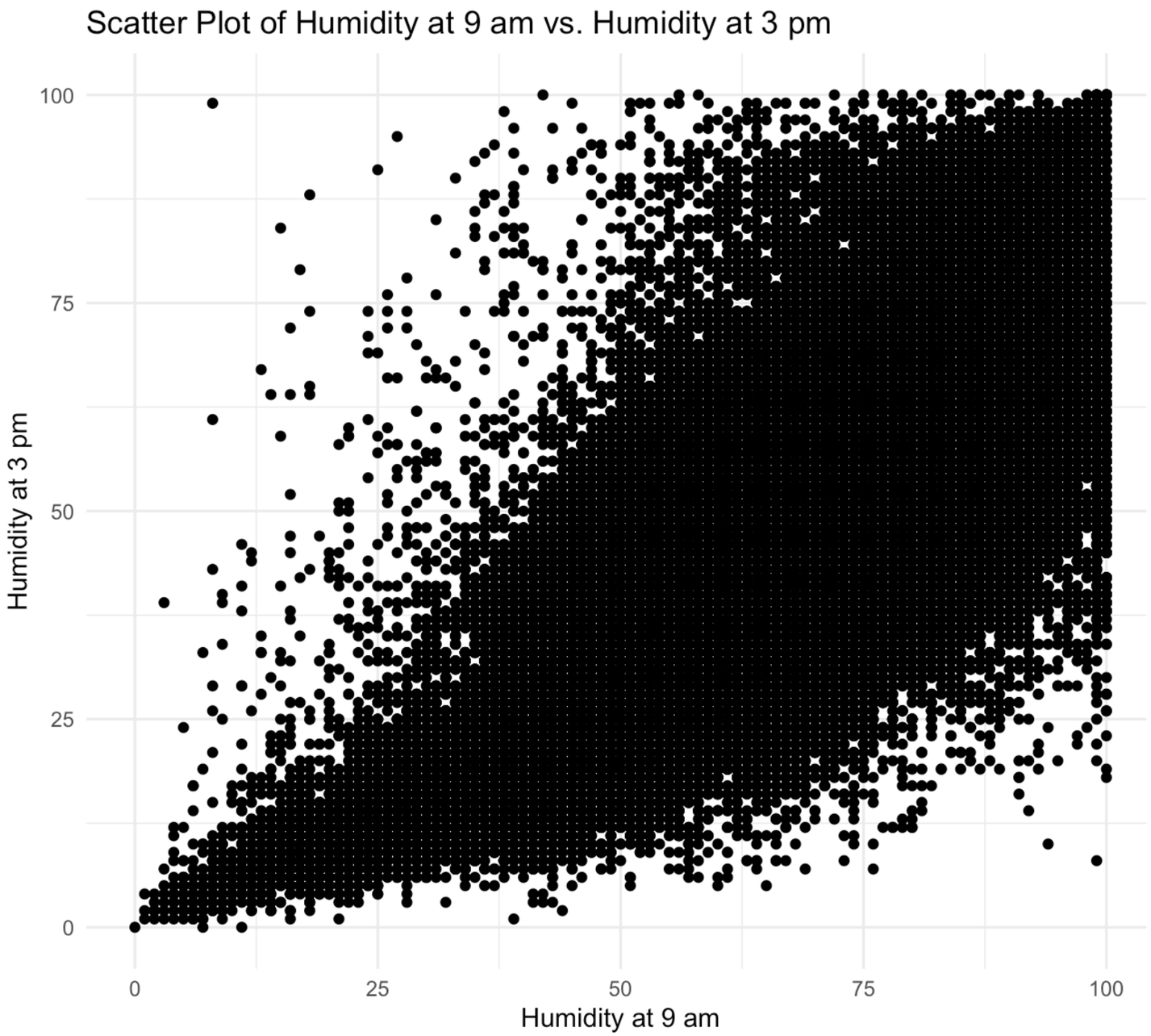




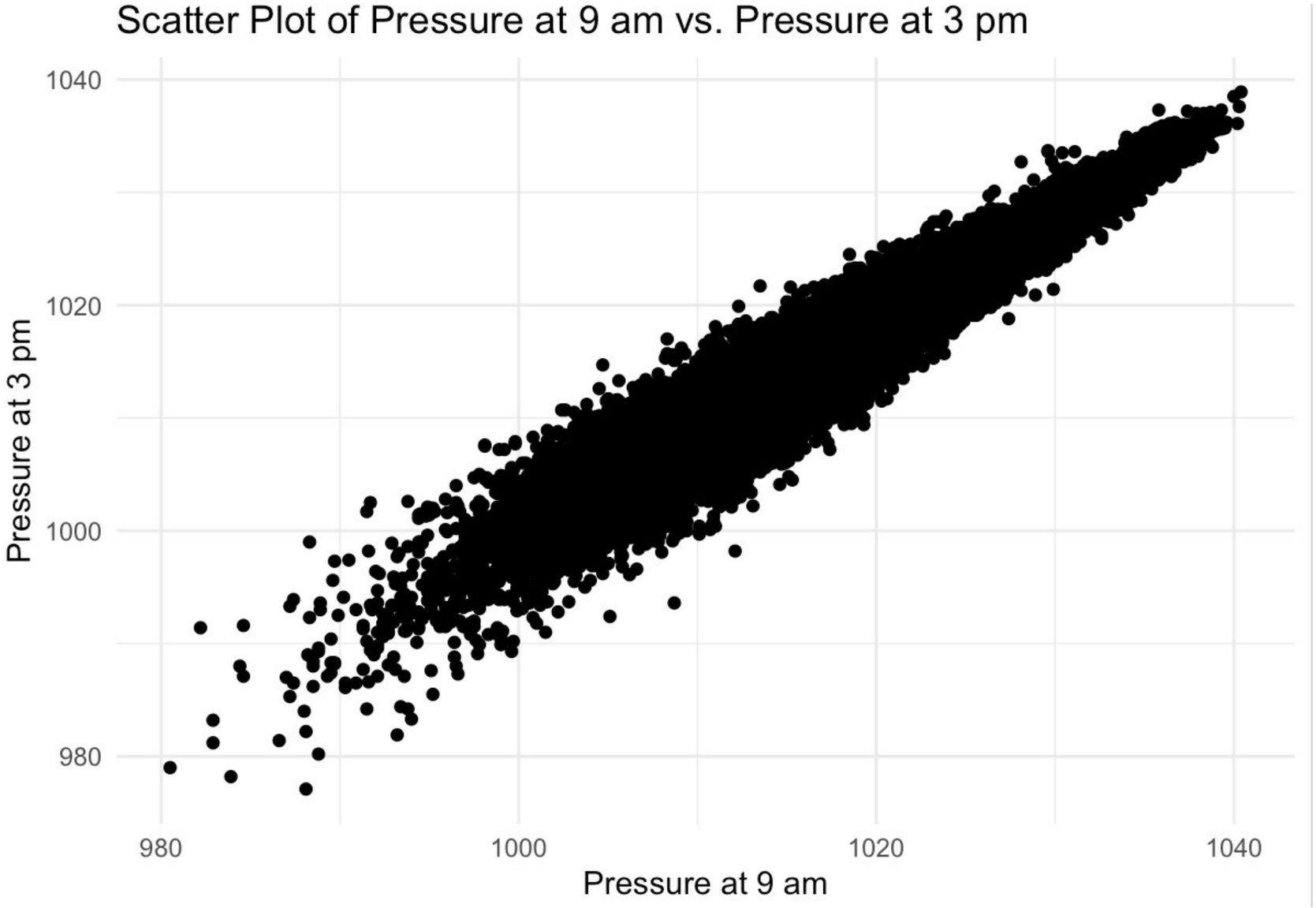
Histogram of Wind gust speed and direction



Scatterplot of Humidity at Morning and Afternoon



Scatterplot of Pressure at Morning and Afternoon

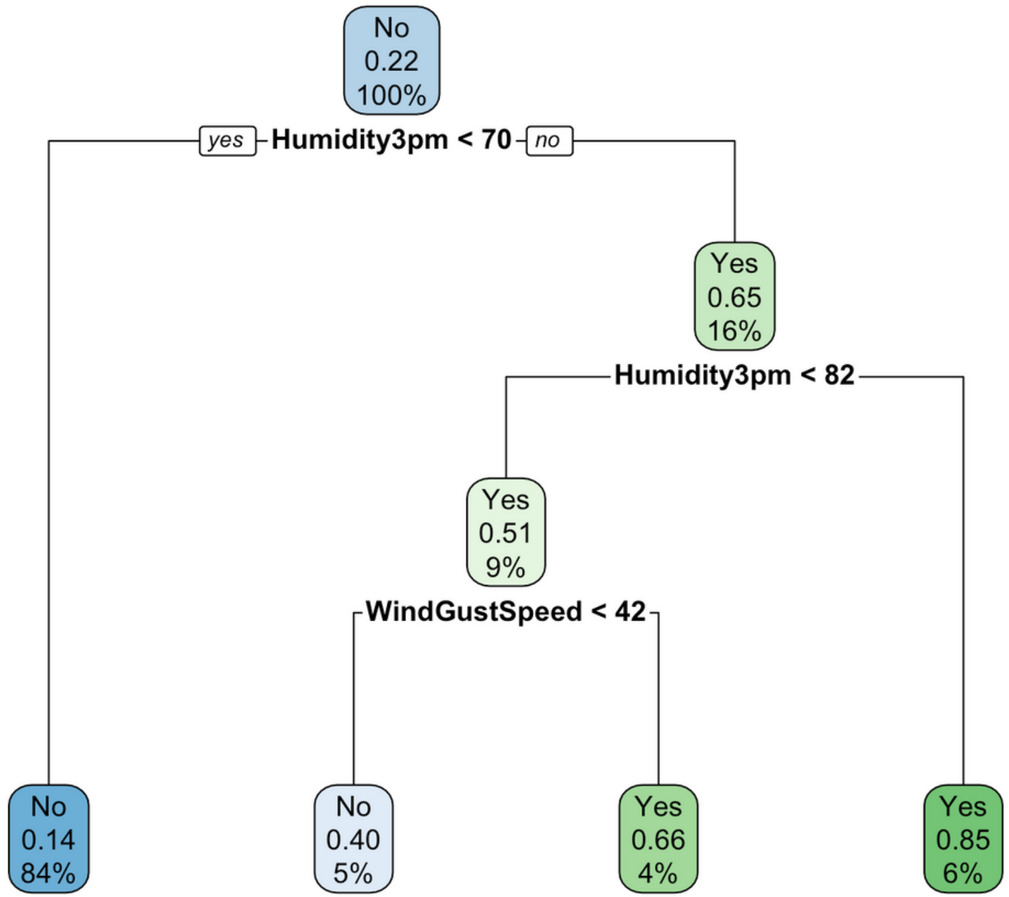


CONFUSION MATRIX FOR RANDOM FOREST

```
          Reference
Prediction    No   Yes
       No  12446  1515
       Yes   751  2213

               Accuracy : 0.8661
                 95% CI : (0.8609, 0.8712)
    No Information Rate : 0.7797
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5793

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9431
            Specificity : 0.5936
         Pos Pred Value : 0.8915
         Neg Pred Value : 0.7466
             Prevalence : 0.7797
         Detection Rate : 0.7354
   Detection Prevalence : 0.8249
      Balanced Accuracy : 0.7684

       'Positive' Class : No
```

CONFUSION MATRIX FOR DECISION TREE

```
          Reference
Prediction   No  Yes
       No  8486 1616
       Yes  278  904

               Accuracy : 0.8322
                 95% CI : (0.8251, 0.839)
    No Information Rate : 0.7767
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4033

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9683
            Specificity : 0.3587
         Pos Pred Value : 0.8400
         Neg Pred Value : 0.7648
             Prevalence : 0.7767
         Detection Rate : 0.7520
   Detection Prevalence : 0.8952
      Balanced Accuracy : 0.6635

       'Positive' Class : No
```



CONFUSION MATRIX FOR KNN MODEL

```
          Reference
Prediction    No   Yes
       No  12314  1816
       Yes   883  1912

               Accuracy : 0.8405
                 95% CI : (0.8349, 0.846)
    No Information Rate : 0.7797
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.49

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9331
            Specificity : 0.5129
         Pos Pred Value : 0.8715
         Neg Pred Value : 0.6841
             Prevalence : 0.7797
         Detection Rate : 0.7276
   Detection Prevalence : 0.8349
      Balanced Accuracy : 0.7230

       'Positive' Class : No
```

THANK YOU!