Rainfall Prediction Using Models

MIS 6356.504 – Business Analytics with R

Group 16
Anusha Sathravada – AXS220402
Nitisha Sree Venkatesan – NXV220065
Kovid Reddy Vayalpati – KXV220038
Rakshith Reddy Koturu – RRK230002

Contents:

1.ABSTRACT:

Rainfall prediction, being a challenging and uncertain task with profound societal consequences, underscores the importance of timely and precise forecasting in mitigating potential human and financial losses.

Within this project, our exploration focuses on employing supervised classification techniques, including K-Nearest Neighbors (KNN), Decision Trees, and Random Forest. The objective is to build a predictive model capable of determining whether it will rain on the following day. This prediction is based on various weather parameters, namely Temperature, Sunshine, Wind speed, Humidity, and Pressure, recorded daily in major cities across Australia. The extensive dataset utilized for this rainfall prediction project spans a decade and compiles daily weather observations from numerous weather stations located throughout Australia.

2.DATA SOURCE:

The dataset is open-source, and it is available to the public on the website Kaggle.

2.1. SOURCE LINK: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package.

3.DATA DESCRIPTION:

The dataset encompasses an extensive collection of 145,460 records, spanning the years 2008 to 2017, offering a comprehensive view of daily weather observations. The observational data originates from various weather stations across Australia, providing insights into weather patterns over the course of a decade.

Within this dataset, diverse features are captured, including quantitative parameters such as maximum and minimum temperature, evaporation, duration of sunshine, and wind speed. Complementing these quantitative aspects are categorical features like dates, locations, and wind direction, which add a nuanced layer to the dataset's richness.

Two crucial boolean features, RainToday and RainTomorrow, play a pivotal role in signaling the occurrence of rain, contributing to the dataset's predictive potential. In essence, this dataset serves as a valuable resource for in-depth exploration and analysis of Australian weather conditions over a significant temporal span.

Date	The date of observation.
Location	The common name of location of the weather station.
Min Temp	Minimum Temp in degree Celsius.
Max Temp	Maximum Temp in degree Celsius.
Rainfall	The amount of rainfall recorded for the day in mm.
Evaporation	The evaporation (mm) in the 24 hours to 9am
Sunshine	The number of hours of bright sunshine in the day.
WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Direction of the wind at 9am
WindDir3pm	Direction of the wind at 3pm
Windspeed9am	Wind speed (km/hr) averaged over 10 minutes prior to 9am
Windspeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Humidity (percent) at 9am
Humidity3pm	Humidity (percent) at 3pm
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	Fraction of sky obscured by cloud (in "oktas": eighths) at 9am.
Cloud3pm	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm.
Temp9am	Temperature (degrees C) at 9am
Temp3pm	Temperature (degrees C) at 3pm
RainToday	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RainTomorrow	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

4.DATA EXPLORATION:

Shape of the dataset: (145460, 23)

```
> dim(rain)
[1] 145460     23
```

Sample of data:

```
> head(rain)
        Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir WindGustSpeed
1 2008-12-01   Albury    13.4    22.9      0.6          NA       NA           W            44
2 2008-12-02   Albury     7.4    25.1      0.0          NA       NA         WNW            44
3 2008-12-03   Albury    12.9    25.7      0.0          NA       NA         WSW            46
4 2008-12-04   Albury     9.2    28.0      0.0          NA       NA          NE            24
5 2008-12-05   Albury    17.5    32.3      1.0          NA       NA           W            41
6 2008-12-06   Albury    14.6    29.7      0.2          NA       NA         WNW            56
  WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm
1          W        WNW           20           24          71          22      1007.7      1007.1
2        NNW        WSW            4           22          44          25      1010.6      1007.8
3          W        WSW           19           26          38          30      1007.6      1008.7
4         SE          E           11            9          45          16      1017.6      1012.8
5        ENE         NW            7           20          82          33      1010.8      1006.0
6          W          W           19           24          55          23      1009.2      1005.4
  Cloud9am Cloud3pm Temp9am Temp3pm RainToday RainTomorrow
1        8       NA    16.9    21.8        No           No
2       NA       NA    17.2    24.3        No           No
3       NA        2    21.0    23.2        No           No
4       NA       NA    18.1    26.5        No           No
5        7        8    17.8    29.7        No           No
6       NA       NA    20.6    28.9        No           No
>
```

Checking missing values in a dataset:

```
> # Check for missing values
> colSums(is.na(rain))
         Date      Location       MinTemp       MaxTemp      Rainfall   Evaporation      Sunshine   WindGustDir
            0             0          1485          1261          3261         62790         69835         10326
WindGustSpeed    WindDir9am    WindDir3pm  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm   Pressure9am
        10263         10566          4228          1767          3062          2654          4507         15065
  Pressure3pm      Cloud9am      Cloud3pm       Temp9am       Temp3pm     RainToday  RainTomorrow
        15028         55888         59358          1767          3609          3261          3267
```

Display the structure of the dataset:

```
> # Display the structure of the dataset
> str(rain)
'data.frame':   145460 obs. of  23 variables:
 $ Date         : chr  "2008-12-01" "2008-12-02" "2008-12-03" "2008-12-04" ...
 $ Location     : chr  "Albury" "Albury" "Albury" "Albury" ...
 $ MinTemp      : num  13.4 7.4 12.9 9.2 17.5 14.6 14.3 7.7 9.7 13.1 ...
 $ MaxTemp      : num  22.9 25.1 25.7 28 32.3 29.7 25 26.7 31.9 30.1 ...
 $ Rainfall     : num  0.6 0 0 0 1 0.2 0 0 0 1.4 ...
 $ Evaporation  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Sunshine     : num  NA NA NA NA NA NA NA NA NA NA ...
 $ WindGustDir  : chr  "W" "WNW" "WSW" "NE" ...
 $ WindGustSpeed: int  44 44 46 24 41 56 50 35 80 28 ...
 $ WindDir9am   : chr  "W" "NNW" "W" "SE" ...
 $ WindDir3pm   : chr  "WNW" "WSW" "WSW" "E" ...
 $ WindSpeed9am : int  20 4 19 11 7 19 20 6 7 15 ...
 $ WindSpeed3pm : int  24 22 26 9 20 24 24 17 28 11 ...
 $ Humidity9am  : int  71 44 38 45 82 55 49 48 42 58 ...
 $ Humidity3pm  : int  22 25 30 16 33 23 19 19 9 27 ...
 $ Pressure9am  : num  1008 1011 1008 1018 1011 ...
 $ Pressure3pm  : num  1007 1008 1009 1013 1006 ...
 $ Cloud9am     : int  8 NA NA NA 7 NA 1 NA NA NA ...
 $ Cloud3pm     : int  NA NA 2 NA 8 NA NA NA NA NA ...
 $ Temp9am      : num  16.9 17.2 21 18.1 17.8 20.6 18.1 16.3 18.3 20.1 ...
 $ Temp3pm      : num  21.8 24.3 23.2 26.5 29.7 28.9 24.6 25.5 30.2 28.2 ...
 $ RainToday    : chr  "No" "No" "No" "No" ...
 $ RainTomorrow : chr  "No" "No" "No" "No" ...
```

Summary statistics for numerical variables:

```
> # Summary statistics for numerical variables
> summary(rain[, c("MinTemp", "MaxTemp", "Rainfall", "Evaporation", "Sunshine",
+                  "WindGustSpeed", "WindSpeed9am", "WindSpeed3pm",
+                  "Humidity9am", "Humidity3pm", "Pressure9am", "Pressure3pm",
+                  "Cloud9am", "Cloud3pm", "Temp9am", "Temp3pm")])
    MinTemp          MaxTemp         Rainfall       Evaporation        Sunshine     WindGustSpeed
 Min.   :-8.50   Min.   :-4.80   Min.   :  0.000   Min.   :  0.00   Min.   : 0.00   Min.   :  6.00
 1st Qu.: 7.60   1st Qu.:17.90   1st Qu.:  0.000   1st Qu.:  2.60   1st Qu.: 4.80   1st Qu.: 31.00
 Median :12.00   Median :22.60   Median :  0.000   Median :  4.80   Median : 8.40   Median : 39.00
 Mean   :12.19   Mean   :23.22   Mean   :  2.361   Mean   :  5.47   Mean   : 7.61   Mean   : 40.03
 3rd Qu.:16.90   3rd Qu.:28.20   3rd Qu.:  0.800   3rd Qu.:  7.40   3rd Qu.:10.60   3rd Qu.: 48.00
 Max.   :33.90   Max.   :48.10   Max.   :371.000   Max.   :145.00   Max.   :14.50   Max.   :135.00
 NA's   :1485    NA's   :1261    NA's   :3261      NA's   :62790    NA's   :69835   NA's   :10263
  WindSpeed9am     WindSpeed3pm    Humidity9am      Humidity3pm      Pressure9am      Pressure3pm
 Min.   :  0.00   Min.   : 0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 980.5   Min.   : 977.1
 1st Qu.:  7.00   1st Qu.:13.00   1st Qu.: 57.00   1st Qu.: 37.00   1st Qu.:1012.9   1st Qu.:1010.4
 Median : 13.00   Median :19.00   Median : 70.00   Median : 52.00   Median :1017.6   Median :1015.2
 Mean   : 14.04   Mean   :18.66   Mean   : 68.88   Mean   : 51.54   Mean   :1017.6   Mean   :1015.3
 3rd Qu.: 19.00   3rd Qu.:24.00   3rd Qu.: 83.00   3rd Qu.: 66.00   3rd Qu.:1022.4   3rd Qu.:1020.0
 Max.   :130.00   Max.   :87.00   Max.   :100.00   Max.   :100.00   Max.   :1041.0   Max.   :1039.6
 NA's   :1767     NA's   :3062    NA's   :2654     NA's   :4507     NA's   :15065    NA's   :15028
    Cloud9am        Cloud3pm        Temp9am         Temp3pm
 Min.   :0.00    Min.   :0.00    Min.   :-7.20   Min.   :-5.40
 1st Qu.:1.00    1st Qu.:2.00    1st Qu.:12.30   1st Qu.:16.60
 Median :5.00    Median :5.00    Median :16.70   Median :21.10
 Mean   :4.45    Mean   :4.51    Mean   :16.99   Mean   :21.68
 3rd Qu.:7.00    3rd Qu.:7.00    3rd Qu.:21.60   3rd Qu.:26.40
 Max.   :9.00    Max.   :9.00    Max.   :40.20   Max.   :46.70
 NA's   :55888   NA's   :59358   NA's   :1767    NA's   :3609
```

5.DATASET CLEANING:

(i) This section aims to understand the distribution of categorical variables in the dataset. The table () function is used to display the unique values and their frequency for each categorical variable, including Location, WindGustDir, WindDir9am, WindDir3pm, RainToday, and RainTomorrow.

```
> # Explore unique values and frequency for categorical variables
> table(rain$Location)

    AliceSprings         Brisbane           Cairns         Canberra            Cobar     CoffsHarbour
            2223             2953             2444             1078              534             1380
          Darwin          Hobart        Melbourne MelbourneAirport          Mildura           Moree
            3062             1939             1898             2929             2594             1913
    MountGambier    NorfolkIsland        Nuriootpa            Perth     PerthAirport        Portland
            2465             2464             2008             3025             2913             1863
            Sale           Sydney    SydneyAirport       Townsville       WaggaWagga        Watsonia
            1678             1690             2870             2419             2416             2730
     Williamtown          Woomera
            1198             1734
> table(rain$WindGustDir)

   E  ENE  ESE    N   NE  NNE  NNW   NW    S   SE  SSE  SSW   SW    W  WNW  WSW
4516 4028 3312 4210 3185 2516 2289 2612 3636 3930 3295 3898 4052 4161 2989 3791
> table(rain$WindDir9am)

   E  ENE  ESE    N   NE  NNE  NNW   NW    S   SE  SSE  SSW   SW    W  WNW  WSW
4456 3932 3400 4967 3390 3267 3016 2854 3421 3880 3893 2926 3356 3707 2918 3037
> table(rain$WindDir3pm)

   E  ENE  ESE    N   NE  NNE  NNW   NW    S   SE  SSE  SSW   SW    W  WNW  WSW
3753 3946 3703 3626 3390 2440 2766 2727 4109 4153 3332 3485 4012 3922 3200 3856
> table(rain$RainToday)

   No   Yes
43958 12462
> table(rain$RainTomorrow)

   No   Yes
43993 12427
```

(ii) The na.omit() function is applied to remove any rows with missing values in the dataset. This step helps in ensuring the dataset is free from incomplete observations. The as.Date() function is utilized to convert the Date column to a date object. This ensures that the date information is represented in a standardized format. The as.factor() function is employed to convert the RainToday and RainTomorrow columns into factors, which is a categorical data type. This is essential for classification tasks.

```
# Data Cleaning

# Remove rows with missing values
rain <- na.omit(rain)

# Convert Date to a date object
rain$Date <- as.Date(rain$Date)

# Convert RainToday and RainTomorrow to factors
rain$RainToday <- as.factor(rain$RainToday)
rain$RainTomorrow <- as.factor(rain$RainTomorrow)
```

(iii) The str() function is used to display the structure of the dataset, providing information about data types and the first few observations. The summary() function gives a statistical summary, offering insights into the central tendency and distribution of numerical variables.

```
> str(rain)
'data.frame':   56420 obs. of  23 variables:
 $ Date         : Date, format: "2009-01-01" "2009-01-02" "2009-01-04" ...
 $ Location     : chr  "Cobar" "Cobar" "Cobar" "Cobar" ...
 $ MinTemp      : num  17.9 18.4 19.4 21.9 24.2 27.1 23.3 16.1 19 19.7 ...
 $ MaxTemp      : num  35.2 28.9 37.6 38.4 41 36.1 34 34.2 35.5 35.5 ...
 $ Rainfall     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Evaporation  : num  12 14.8 10.8 11.4 11.2 13 9.8 14.6 12 11 ...
 $ Sunshine     : num  12.3 13 10.6 12.2 8.4 0 12.6 13.2 12.3 12.7 ...
 $ WindGustDir  : chr  "SSW" "S" "NNE" "WNW" ...
 $ WindGustSpeed: int  48 37 46 31 35 43 41 37 48 41 ...
 $ WindDir9am   : chr  "ENE" "SSE" "NNE" "WNW" ...
 $ WindDir3pm   : chr  "SW" "SSE" "NNW" "WSW" ...
 $ WindSpeed9am : int  6 19 30 6 17 7 17 15 30 15 ...
 $ WindSpeed3pm : int  20 19 15 6 13 20 19 6 9 17 ...
 $ Humidity9am  : int  20 30 42 37 19 26 33 25 46 61 ...
 $ Humidity3pm  : int  13 8 22 22 15 19 15 9 28 14 ...
 $ Pressure9am  : num  1006 1013 1012 1013 1011 ...
 $ Pressure3pm  : num  1004 1012 1009 1009 1007 ...
 $ Cloud9am     : int  2 1 1 1 1 8 3 1 1 1 ...
 $ Cloud3pm     : int  5 1 6 5 6 8 1 1 5 5 ...
 $ Temp9am      : num  26.6 20.3 28.7 29.1 33.6 30.7 25 20.7 23.4 24 ...
 $ Temp3pm      : num  33.4 27 34.9 35.6 37.6 34.3 31.5 32.8 33.3 33.6 ...
 $ RainToday    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ RainTomorrow : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:89040] 1 2 3 4 5 6 7 8 9 10 ...
  ..- attr(*, "names")= chr [1:89040] "1" "2" "3" "4" ...
> summary(rain)
      Date               Location           MinTemp          MaxTemp          Rainfall      
 Min.   :2007-11-01   Length:56420       Min.   :-6.70   Min.   : 4.10   Min.   :  0.00  
 1st Qu.:2010-07-19   Class :character   1st Qu.: 8.60   1st Qu.:18.70   1st Qu.:  0.00  
 Median :2012-07-28   Mode  :character   Median :13.20   Median :23.90   Median :  0.00  
 Mean   :2012-09-17                      Mean   :13.46   Mean   :24.22   Mean   :  2.13  
 3rd Qu.:2014-10-10                      3rd Qu.:18.40   3rd Qu.:29.70   3rd Qu.:  0.60  
 Max.   :2017-06-25                      Max.   :31.40   Max.   :48.10   Max.   :206.20  
  Evaporation        Sunshine      WindGustDir        WindGustSpeed     WindDir9am       
 Min.   : 0.000   Min.   : 0.000   Length:56420       Min.   :  9.00   Length:56420      
 1st Qu.: 2.800   1st Qu.: 5.000   Class :character   1st Qu.: 31.00   Class :character  
 Median : 5.000   Median : 8.600   Mode  :character   Median : 39.00   Mode  :character  
 Mean   : 5.503   Mean   : 7.736                      Mean   : 40.88                     
 3rd Qu.: 7.400   3rd Qu.:10.700                      3rd Qu.: 48.00                     
 Max.   :81.200   Max.   :14.500                      Max.   :124.00                     
  WindDir3pm         WindSpeed9am    WindSpeed3pm    Humidity9am      Humidity3pm     Pressure9am   
 Length:56420       Min.   : 2.00   Min.   : 2.00   Min.   :  0.00   Min.   :  0.0   Min.   : 980.5  
 Class :character   1st Qu.: 9.00   1st Qu.:13.00   1st Qu.: 55.00   1st Qu.: 35.0   1st Qu.:1012.7  
 Mode  :character   Median :15.00   Median :19.00   Median : 67.00   Median : 50.0   Median :1017.2  
                    Mean   :15.67   Mean   :19.79   Mean   : 65.87   Mean   : 49.6   Mean   :1017.2  
                    3rd Qu.:20.00   3rd Qu.:26.00   3rd Qu.: 79.00   3rd Qu.: 63.0   3rd Qu.:1021.8  
                    Max.   :67.00   Max.   :76.00   Max.   :100.00   Max.   :100.0   Max.   :1040.4  
  Pressure3pm        Cloud9am        Cloud3pm        Temp9am         Temp3pm       RainToday     
 Min.   : 977.1   Min.   :0.000   Min.   :0.000   Min.   :-0.7   Min.   : 3.70   No :43958  
 1st Qu.:1010.1   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:13.1   1st Qu.:17.40   Yes:12462  
 Median :1014.7   Median :5.000   Median :5.000   Median :17.8   Median :22.40              
 Mean   :1014.8   Mean   :4.242   Mean   :4.327   Mean   :18.2   Mean   :22.71              
 3rd Qu.:1019.4   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:23.3   3rd Qu.:27.90              
 Max.   :1038.9   Max.   :8.000   Max.   :9.000   Max.   :39.4   Max.   :46.10              
 RainTomorrow
 No :43993
 Yes:12427
```

6.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.

7.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.

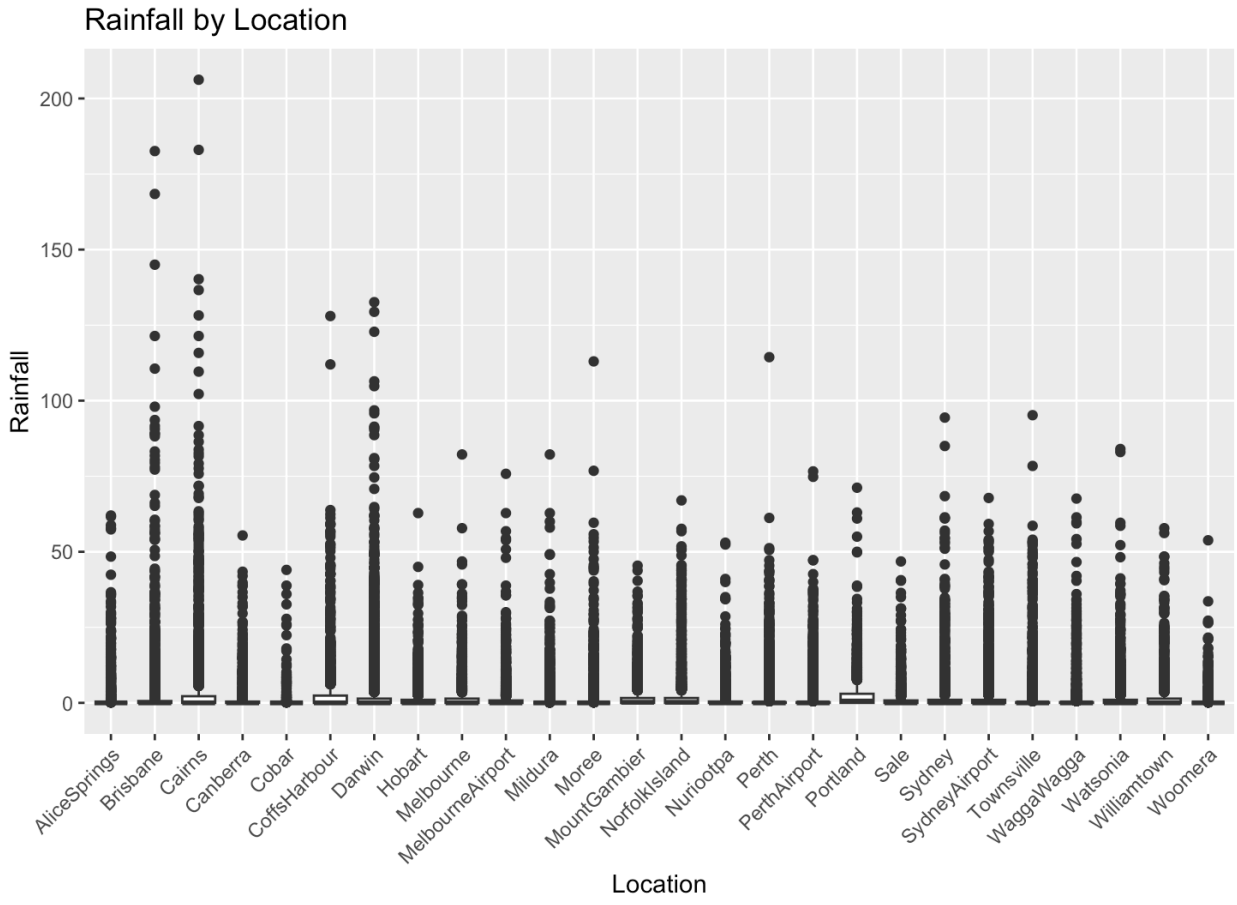
8.DATA VISUALIZATION:

Insights:

Visualization of Rainfall over the years by Location:

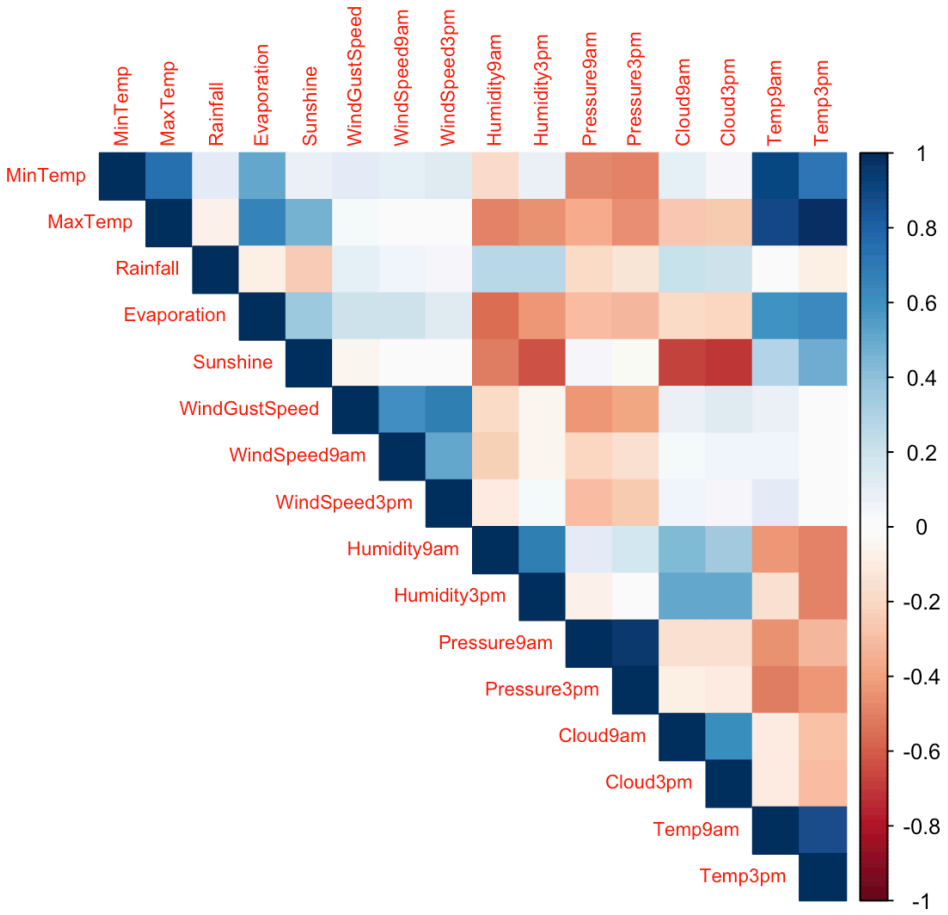
This code segment serves as part of Exploratory Data Analysis (EDA), specifically focusing on visualizing the relationship between rainfall and location. The boxplot is an effective graphical representation that allows for the comparison of rainfall distributions across different locations. Each boxplot summarizes the central tendency and spread of rainfall values, offering insights into

potential variations and outliers. The chosen customization in the theme improves the presentation of the x-axis labels, making them more readable when rotated. The overall purpose is to gain a visual understanding of how rainfall varies among different locations, aiding in the identification of patterns or trends in the dataset.



Correlation matrix and plot:

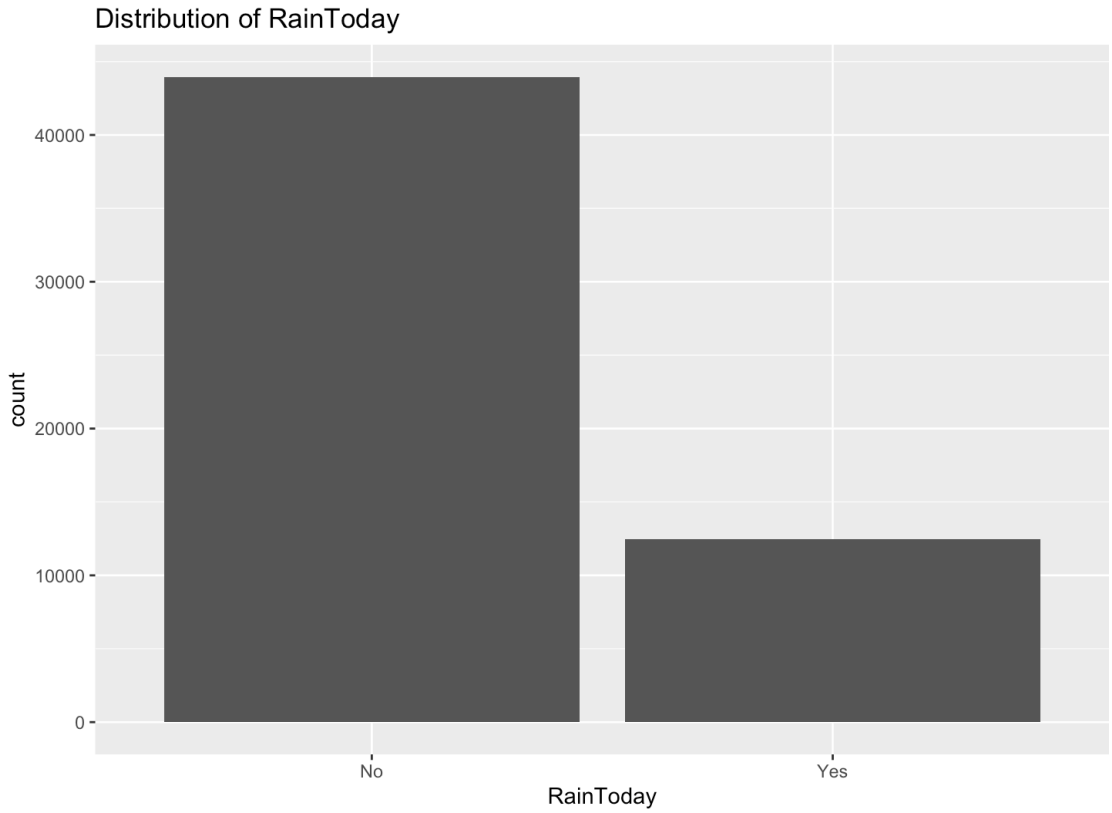
This correlation plot is valuable for identifying patterns and associations between different weather-related variables. Strong correlations (either positive or negative) suggest potential relationships, which can be crucial for feature selection or gaining insights into the underlying dynamics of the dataset. The cor() function computes the pairwise correlations between the specified numerical variables, creating a matrix (cor_matrix) of correlation coefficients.

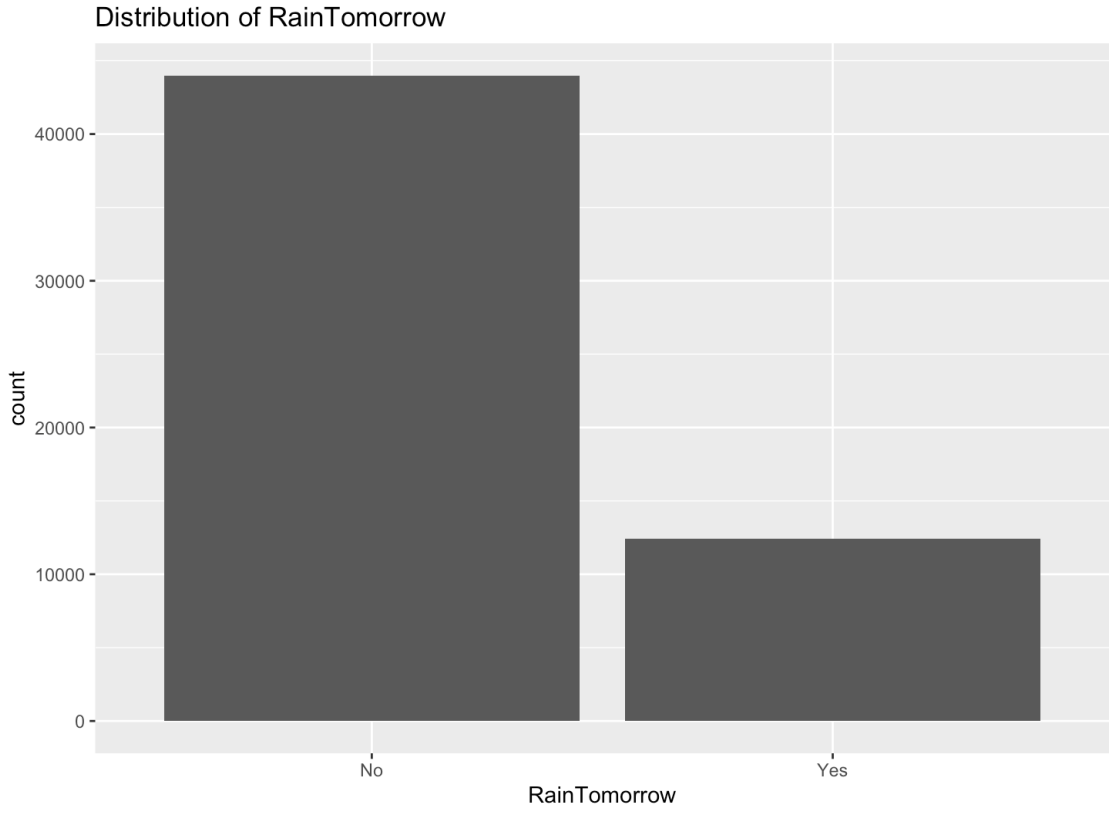


The corrplot() function generates a correlation plot based on the correlation matrix. The plot uses color to indicate the strength and direction of correlations, with positive correlations in one color spectrum and negative correlations in another. The upper triangle of the plot is chosen for visualization to avoid redundancy, as the lower triangle is a mirror reflection.

Bar plot whether it's raining today and tomorrow:

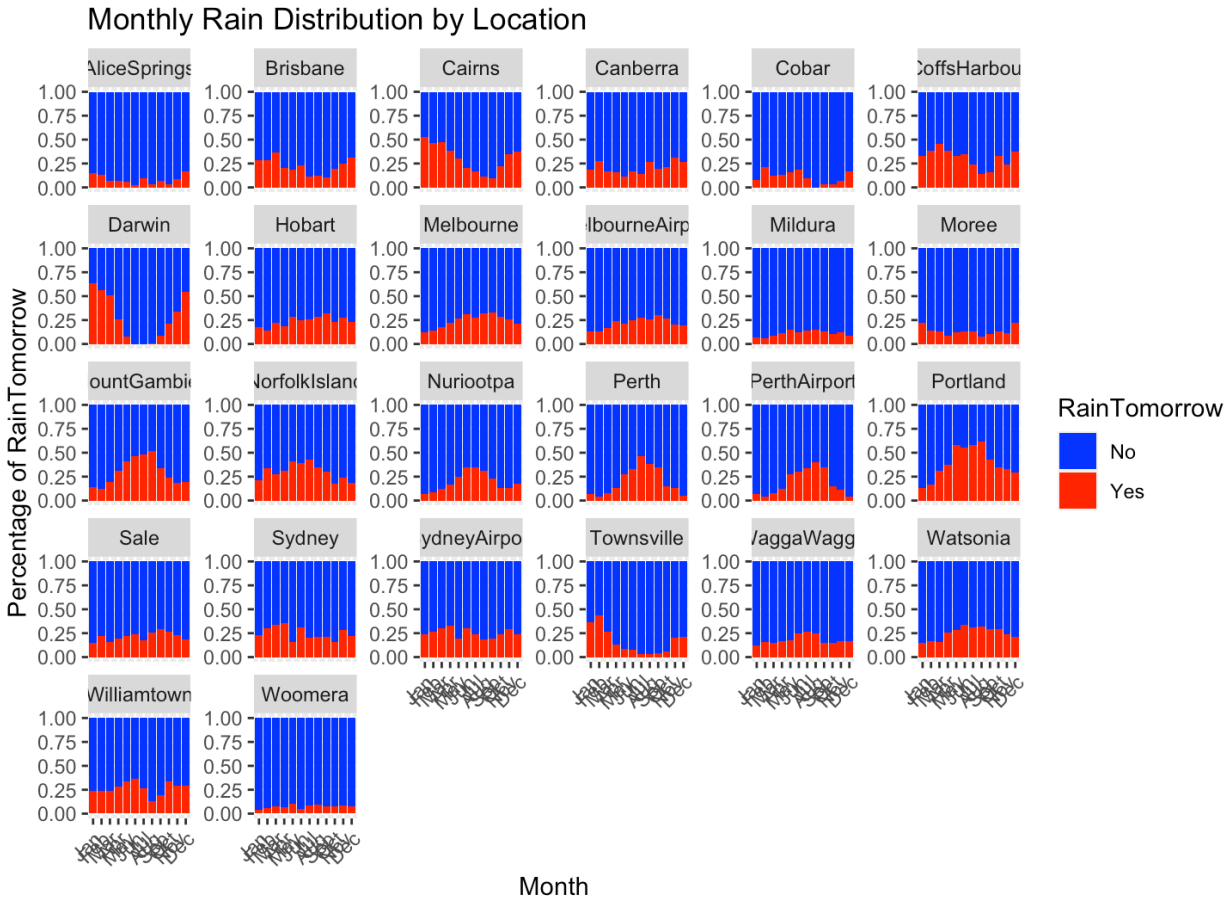
This segment is designed to create a bar plot illustrating the distribution of the binary variable RainTomorrow and RainToday. A bar plot is a simple and effective way to visualize the frequency or count of different categories in a categorical variable. In this case, it helps understand the balance or imbalance in the dataset regarding the occurrence of rain (RainTomorrow, RainToday being either "Yes" or "No").


Distribution of RainToday
40000
30000
20000
10000
0
count
No
Yes
RainToday


Distribution of RainTomorrow
40000
30000
20000
10000
0
count
No
Yes
RainTomorrow

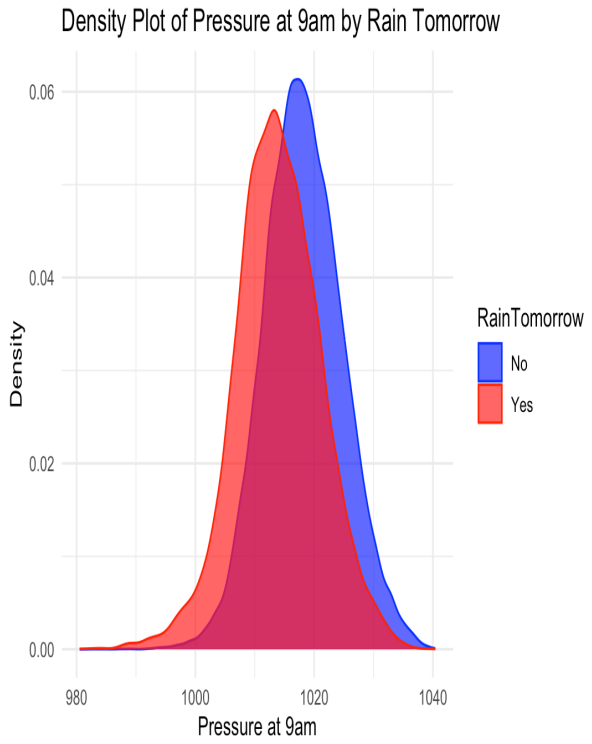
Exploring Monthly Rain distribution by location:

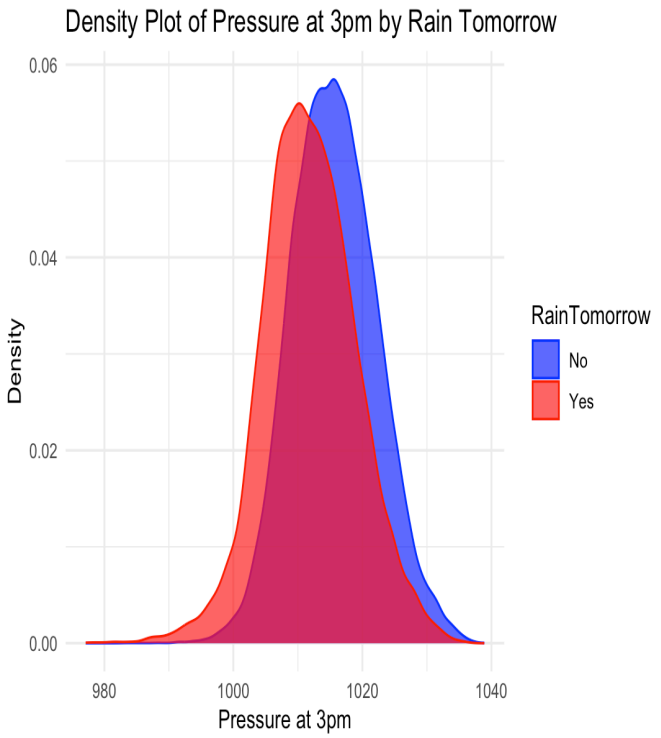
The rainfall in different months among different locations is visualized through the graph. The rainfall differs from month to month and across different locations. Very few regions have less variations of rainfall through the months and majority of regions show variation.



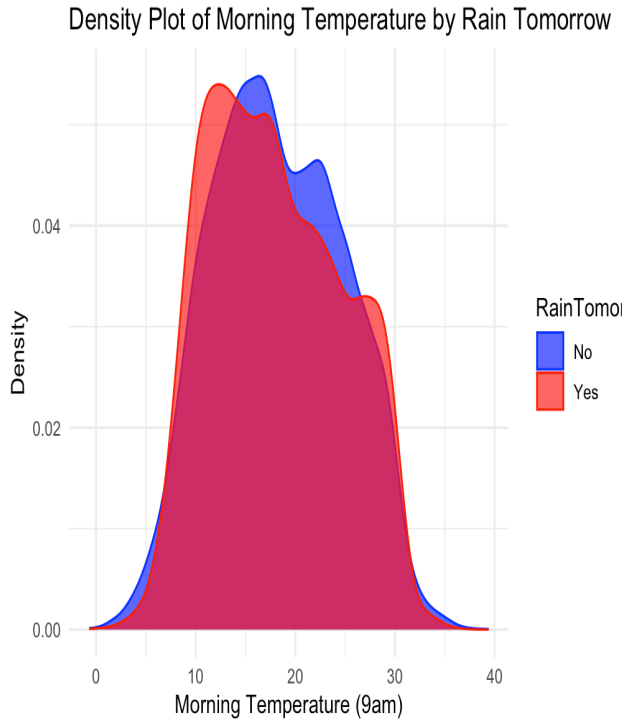
Density plot for pressure at Morning and Afternoon by RainTomorrow:

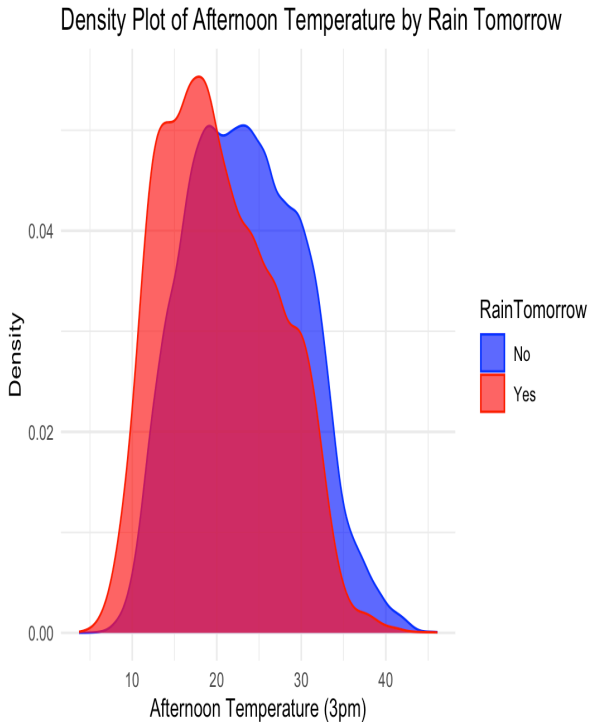
These density plots are useful for visualizing the probability distribution of atmospheric pressure in the morning and afternoon, respectively, with differentiating colors based on whether rain occurs the next day (RainTomorrow). The transparency (alpha) allows for better visualization of overlapping distributions, and the color customization enhances the distinction between categories. These visualizations can aid in identifying potential patterns and differences in the weather conditions concerning the occurrence of rain.





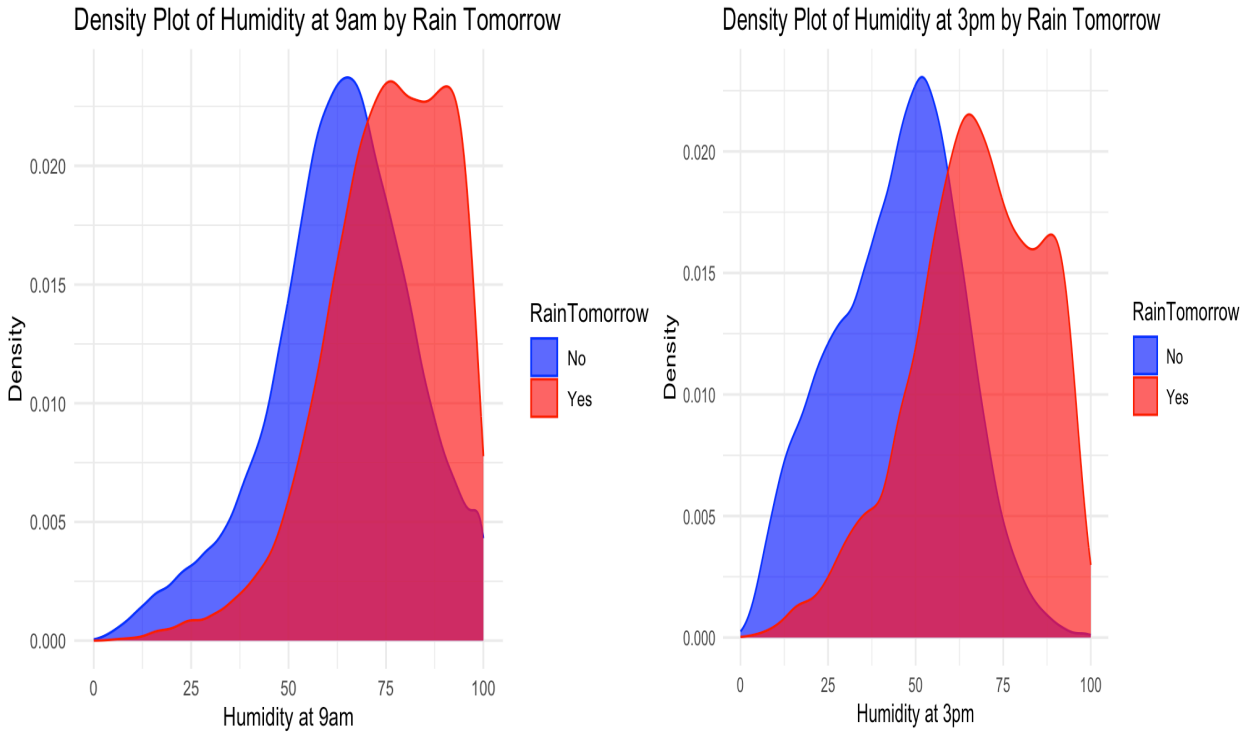
Density plot for Morning and Afternoon temperature by RainTomorrow: The plotted graph suggests that lower temperatures, especially in the afternoon (Temp3pm), correlate with an increased chance of rainfall the next day (RainTomorrow). This highlights the significant impact of temperature on rainfall likelihood.





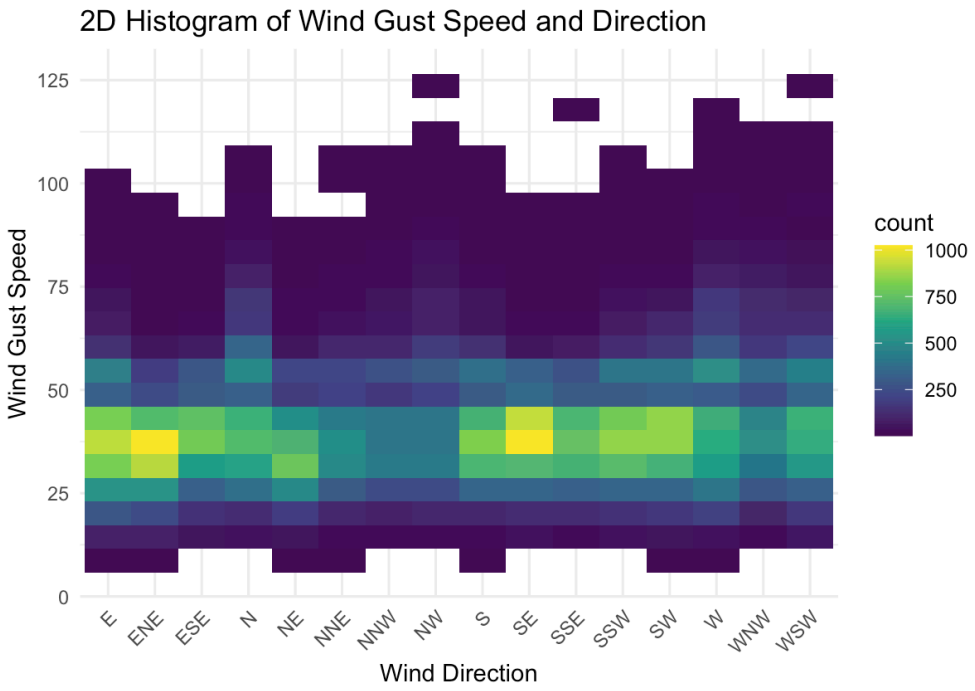
Density plot for humidity at Morning and Afternoon by RainTomorrow:

The plotted graph indicates a higher likelihood of rainfall when humidity levels are elevated, specifically in the afternoon (Humidity3pm). This observation underscores the considerable influence of humidity on rainfall occurrences, emphasizing the correlation between increased afternoon humidity and a subsequent rise in rainfall likelihood.



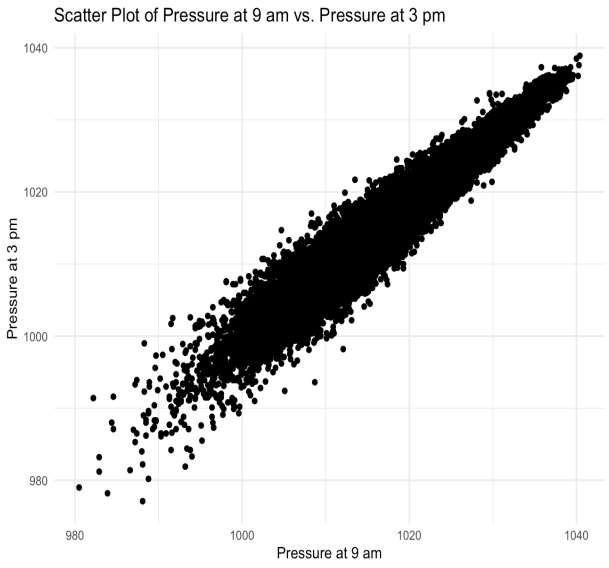
2D - Histogram (square plot) for WindGustSpeed and WindGustDir:

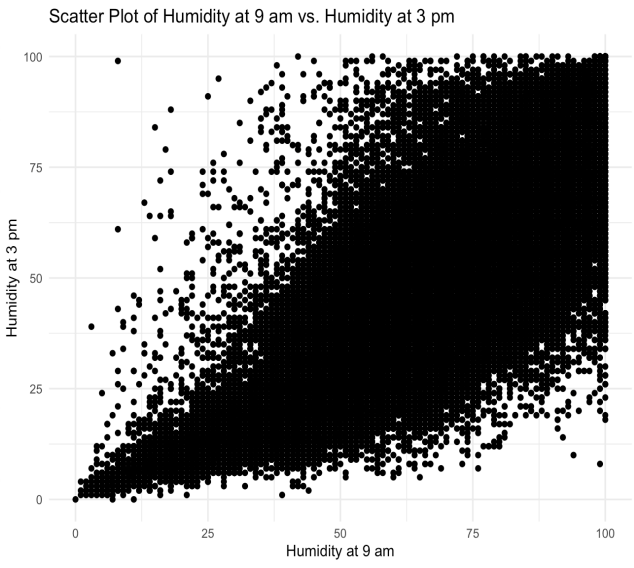
This code generates a 2D histogram to visualize the joint distribution of wind gust speed and wind direction. Each square in the plot represents a bin, and the color intensity reflects the frequency of observations within each bin. The use of the Viridis color palette enhances the visibility of patterns in the data. The resulting plot provides insights into the relationship between wind gust speed and wind direction, allowing for a better understanding of their joint distribution.



Scatter plot for humidity and pressure Morning and Afternoon by RainTomorrow:

Both sections aim to explore and visualize the relationships between meteorological variables at different times of the day. Scatter plots are effective for identifying patterns, trends, and potential correlations between the specified pairs of variables. The removal of missing values ensures a comprehensive and accurate representation of the relationships within the datasets.





9.RAINFALL PREDICTION USING MODELS:

1. Random Forest

2. Decision Tree Model

3. KNN Classification

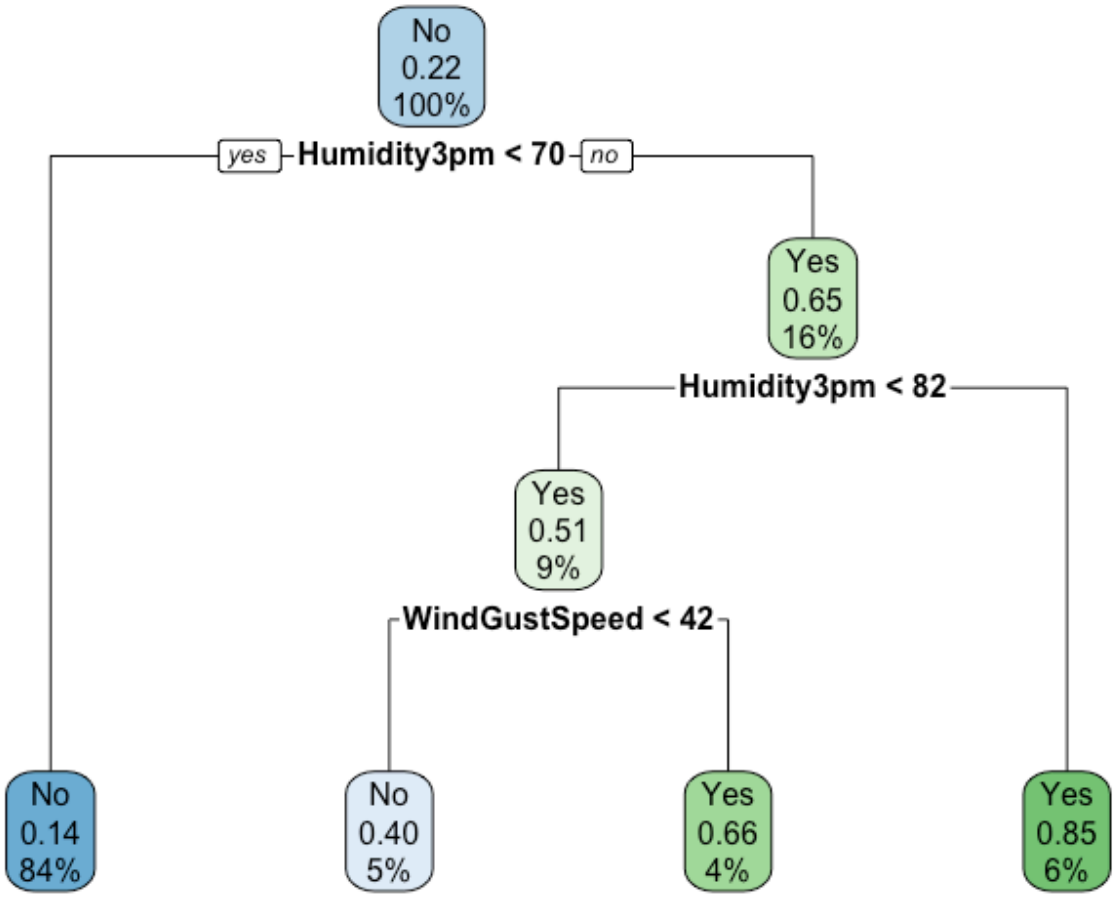
9.1 Random Forest:

Random Forest is a versatile ensemble learning technique widely employed in machine learning and predictive modeling. Comprising an ensemble of decision trees, Random Forest improves predictive accuracy and reduces overfitting by aggregating the outputs of individual trees. Each tree is trained on a random subset of the data and features, contributing to a diverse set of learners. Through a majority voting mechanism, Random Forest generates robust predictions, making it resilient to noise and capable of handling complex relationships within the data. Its adaptability, scalability, and ability to handle both classification and regression tasks make Random Forest a valuable tool in various domains, including finance, healthcare, and remote sensing.

```
Confusion Matrix and Statistics

          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
```

9.2 Decision Tree Model:

A decision tree is a powerful and interpretable machine learning algorithm commonly used for classification and regression tasks. It operates by recursively partitioning the dataset based on the most informative features, creating a tree-like structure of decision nodes. Each node represents a specific feature and a corresponding decision rule, leading to subsequent nodes or terminal leaves with predicted outcomes. Decision trees excel in capturing complex decision-making processes and are adept at handling both categorical and numerical data. Their visual clarity aids in understanding the model's decision logic, making decision trees valuable for tasks where interpretability is crucial. Despite their susceptibility to overfitting, techniques like pruning and ensemble methods, such as Random Forests, enhance their robustness and generalization capabilities.



```
Confusion Matrix and Statistics

          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
```

9.3 KNN Classification:

K-Nearest Neighbors (KNN) is a straightforward, yet powerful supervised machine learning algorithm used for both classification and regression tasks. KNN classifies or predicts the target variable by considering the majority class or average of its k nearest neighbors in the feature space. The choice of 'k' influences the model's sensitivity to local variations, providing flexibility in handling different data patterns. KNN is non-parametric, meaning it doesn't assume a specific underlying data distribution, making it applicable across diverse datasets. While computationally efficient for small to moderately sized datasets, KNN's performance may be impacted by the curse of dimensionality in high-dimensional spaces. Overall, KNN stands out for its simplicity and effectiveness in scenarios where local patterns play a crucial role.

```
Confusion Matrix and Statistics

          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
```

10. CONCLUSION:

Upon comparing the performance of KNN, Decision Tree, and Random Forest models it is evident that Random Forest stands out as the most accurate and computationally efficient classification model. The accuracy metrics for each model are as follows:

- KNN: 84.05%
- Decision Tree: 83.22%
- Random Forest: 86.66%

Key Factors for Rain Prediction: Humidity, pressure, and temperature emerge as pivotal factors influencing rain prediction. These meteorological variables play a crucial role in determining the likelihood of rainfall. Temporal Influence on Rain Prediction: Notably, data recorded in the evening (3 pm) holds greater significance in predicting rain for the following day. This temporal aspect underscores the importance of considering specific time frames for more accurate rain forecasts. In summary, Random Forest proves to be the superior choice for rain prediction, offering high accuracy and computational efficiency. The emphasis on key meteorological variables, coupled with the recognition of temporal patterns, enhances the overall predictive capabilities of the model.