VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

FOUNDATION OF DATA ANALYTICS

FACULTY: PROF.TULASI PRASAD

PROJECT REPORT

ANALYSIS OF RAINFALL PREDICTION

TEAM MEMBERS:

V.SAI PREETHAM (19BCE1434)

B. SOUMITH RAM (19BPS1116)

TABLE OF CONTENTS :

INRODUCTION

Rainfall is the most crucial process of nature. All the living being rely on water and rainfall is a process that is responsible for the continual process of water cycle. Many farmers will depend upon the rainfall for cultivating many essential crops for us, government also looking ahead of prediction of rainfall which will be useful for the future purposes, many human activities like agriculture are dependent on rainfall, especially in a country like India. Thus, it is very important and necessary to predict the rainfall patterns to estimate the flooding and drowning events, hence we need to analyze the rainfall data to have such pattern and in this we must describe about the rainfall, diagnosis the rainfall patterns and predict the rainfall using the analytics of the rainfall patterns, to predict the rainfall Application of algorithms is the best way to forecast rainfall. These algorithms predict rainfall numerically. There are two kinds of approaches for it. They are empirical and dynamic methods. Empirical approach consists of evaluating historical data, identifying the pattern or relationship between the given atmospheric variables that determine rainfall

1.OVERVIEW

1.1 PROBLEM STATEMENT

Farmers have no idea of falling rains to cultivate the crops and they cannot predict when the rainfall will occur. To solve this problem, we will collect the data(information), datasets and will predict the rainfall percentage which will be useful to the farmers.

It is also important to understand the rainfall patterns across different years in the different states to predict the rainfall at that location

1.2. OBJECTIVES

Objective is to need to analyze the rainfall data to have such pattern and in this we must describe about the rainfall, diagnosis the rainfall patterns and predict the rainfall using the analytics of the rainfall patterns, and will visualize the graphs and have the patterns using visualizing tools, to predict the rainfall Application of algorithms is the best way to forecast rainfall. These algorithms predict rainfall numerically. There are two kinds of approaches for it. They are empirical and dynamic methods. Empirical approach consists of evaluating historical data, identifying the pattern or relationship between the given atmospheric variables that determine rainfall. It includes many kinds of clustering and classification techniques and the implementation of the rainfall prediction is very crucial in these days for the farmers and government hence there are many methods and many algorithms which will be very useful for the predicting the rainfall in these days, Dynamic approach consists of dynamically changing training samples whose results can be applied to other large samples of data. In this rainfall prediction by using the machine learning algorithms how rainfall fall can be predicted by machine learning algorithms good accuracy and using the machine learning algorithms it will be very good

for the finding of the rainfall prediction of the machine learning algorithms of dynamic approach like linear regression algorithms, SVM machine learning algorithm and will build UI website for the rainfall prediction

1.3 DATASOURCES

The data in this report was taken from data obtaining platforms such as Kaggle and bom.gov (Bureau of Meteorology) official weather forecasters of Australia.

1.4 TOOLS AND TECHNIQUES

In this work RStudio, Tableau are major tools used and we used all the data analytics techniques and used many visualizations to understand the patterns of the rainfall which will be detailed and elaborated in this paper work

1.5 LIMITATIONS

Some of the limitations of the project are:

- Data imputation was done using mean imputation, better imputation techniques can be used for better accuracy
- Only some part of the data set was used to prevent extensive computational time.
- This dataset is restricted only to Australian cities. Therefore, it's not possible to calculate to the result for the entire world
- The Algorithms which we use for the model building is the clean data which we use the imputation methods of mean and median so it is not possible to get some of the accurate predictions

2.DATA DESCRIPTION AND PREPARATION

2.1 DATA OF RAINFALL

Our dataset consists of 23 columns and 145460 rows in total.

$ Date: char

$ Location: char

$ MinTemp : num

$ MaxTemp : num

$ Rainfall: num

$ Evaporation: num

$ Sunshine: num

$ WindGustDir : char

$ WindGustSpeed: int

$ WindDir9am: char

$ WindDir3pm: char

$ WindSpeed9am: int

$ WindSpeed3pm: int

$ Humidity9am: int

$ Humidity3pm: int

$ Pressure9am: num

$ Pressure3pm: num

$ Cloud9am: int

$ Cloud3pm: int

$ Temp9am: num

$ Temp3pm: num

$ RainToday : char $Rain tomorrow:char

2.2 DATA EXPLORATION

In the Data Exploration phase dealing with the weather of the Australian data which helps to analyze and predict the rainfall patterns and which consists of columns as date, Location, MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, Wind Speed, Wind direction at different times, Cloud at different times, Humidity at different times, pressure at different times, Temperature at different times, Rain today, Rain tomorrow. These are the variables included in the dataset which will be useful for the Analyzing the rainfall patterns and analyzing the prediction of rainfall

2.3 DATA TABLE

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
1	2008-12-01	Albury	13.4	22.9	0.6	NA	NA	W
2	2008-12-02	Albury	7.4	25.1	0.0	NA	NA	WNW
3	2008-12-03	Albury	12.9	25.7	0.0	NA	NA	WSW
4	2008-12-04	Albury	9.2	28.0	0.0	NA	NA	NE
5	2008-12-05	Albury	17.5	32.3	1.0	NA	NA	W
6	2008-12-06	Albury	14.6	29.7	0.2	NA	NA	WNW
7	2008-12-07	Albury	14.3	25.0	0.0	NA	NA	W
8	2008-12-08	Albury	7.7	26.7	0.0	NA	NA	W
9	2008-12-09	Albury	9.7	31.9	0.0	NA	NA	NNW
10	2008-12-10	Albury	13.1	30.1	1.4	NA	NA	W
11	2008-12-11	Albury	13.4	30.4	0.0	NA	NA	N
12	2008-12-12	Albury	15.9	21.7	2.2	NA	NA	NNE
13	2008-12-13	Albury	15.9	18.6	15.6	NA	NA	W
14	2008-12-14	Albury	12.6	21.0	3.6	NA	NA	SW
15	2008-12-15	Albury	8.4	24.6	0.0	NA	NA	NA

2.4 DATA CLEANING AND PREPARATION

The Data Cleaning is done in the four phases accordance to the type of the variables present in the dataset. Hence it divided into 4 parts

1. Categorical
2. Discrete
3. Numerical
4. Continuous

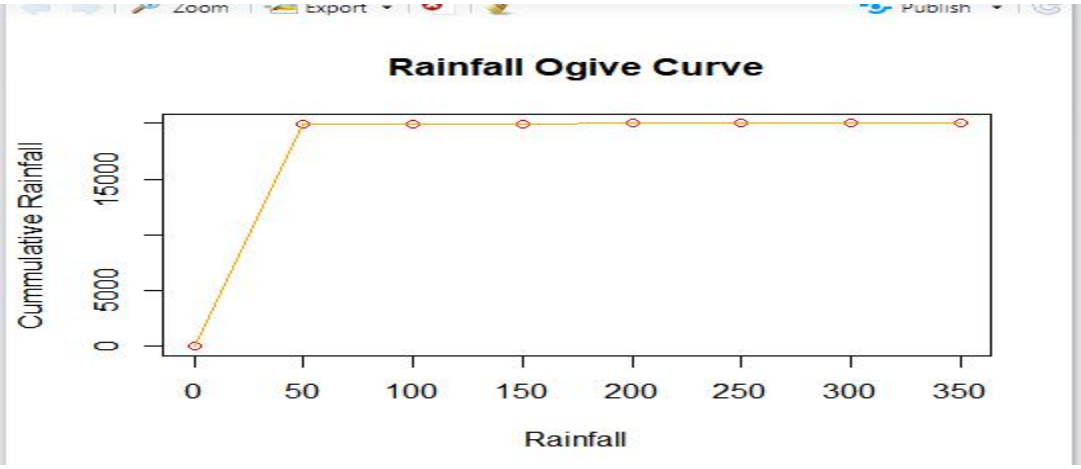
For the Discrete Data we had Imputed values by the method of Mode Imputation technique. For the categorical data imputed by the mean Imputation technique and for numeric values the median imputation has been imputed into the missing NA values and for the empty columns the Random Imputation technique has been done to get some relation between the Sunshine, Evaporation and the rainfall and for the windspeed direction has divided the directions based on the indexes of the numbers and made the count of the vales to the indexes and imputed with the following count of the wind speed direction and for the Rain tomorrow and rain today the imputation has done by proceeding with yes quantities to the variables. In the R-code the KNN technique and median technique is used for the accurate values of the missing values

3 EXPLORATORY DATA ANALYTICS

3.1 Statistical measures

```
> rainrange=range(rainfall1)
> rainbreaks=seq(0,371,by=50)
> rainbreaks
[1]   0  50 100 150 200 250 300 350
> rainbreaks_cut=cut(rainfall1,rainbreaks,right=FALSE,left=FALSE)
> rainfreq=table(rainbreaks_cut)
> rainfreq
rainbreaks_cut
   [0,50)  [50,100) [100,150) [150,200) [200,250) [250,300) [300,350)
    19905        77        11         3         2         1         0
> raincumfreq=cumsum(rainfreq)
> raincumfreq
   [0,50)  [50,100) [100,150) [150,200) [200,250) [250,300) [300,350)
    19905     19982     19993     19996     19998     19999     19999
> raincumfreq0=c(0,cumsum(rainfreq))
> plot(rainbreaks,raincumfreq0,
+      main="Rainfall Ogive Curve",
+      xlab="Rainfall",
+      ylab="Cummulative Rainfall",
+      col="red")
> lines(rainbreaks,raincumfreq0,col="Orange")
> rainrange
[1]   0 371
```

The Ogive Curve for the Rainfall is to be: -



The Statistical measures are found to be: -

```
> #Statistical Measures
> mean(data1$Rainfall)
[1] 2.184417
> median(data1$Rainfall)
[1] 0
> maxRainfall=max(data1$Rainfall)
> maxRainfall
[1] 371
> #highest rainfall occured
> highestRainfall=data1[order(-data1$Rainfall),c(1,2)]
> highestRainfall[1,1:2]
           Date     Location
9369 2009-11-07 CoffsHarbour
> quantile(data1$Rainfall)
   0%   25%   50%   75%  100%
  0.0   0.0   0.0   0.8 371.0
> var(data1$Rainfall)
[1] 71.07832
> sqrt(var(data1$Rainfall))
[1] 8.430796
> IQR(data1$Rainfall)
[1] 0.8
> library(ggplot2)
> library(moments)
> sk=skewness(data1$Rainfall)
> sk
[1] 13.84239
> if (sk <0) {
+         print('Negitive Skewness')
+ } else if (sk>0 ) {
+         print('Positive Skewness')
+ } else {
+         print('No Skewness')
+ }
[1] "Positive Skewness"
```

The Data is found to be positive Skewness data and the IQR is found to be 0.8 and the place in which maximum rainfall occurred is Cardsharper which is on the date of 07-11-2009

```
> ku=kurtosis(data1$Rainfall)
> ku
[1] 366.025
> if (ku<3){
+         print('PlatyKurtic')
+ }else if(ku>3){
+         print('Leptokurtic')
+ }else{
+         print('Mesokurtic')
+ }
[1] "Leptokurtic"
```

The Kurtosis is found to be 366.02 which is extremely high belongs to the Leptokurtic range

3.3

The Correlation for the variables is given as: -

```
> cor.test(data1$Rainfall,data1$MaxTemp)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$MaxTemp
t = -10.109, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08507987 -0.05750239
sample estimates:
        cor
-0.07130476

> cor.test(data1$Rainfall,data1$WindSpeed3pm)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$WindSpeed3pm
t = 5.7257, df = 19998, p-value = 1.045e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02661156 0.05428458
sample estimates:
       cor
0.04045583
```

```
> cor.test(data1$Rainfall,data1$WindSpeed9am)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$WindSpeed9am
t = 10.809, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06241783 0.08997525
sample estimates:
      cor
0.0762111

> cor.test(data1$Rainfall,data1$Humidity9am)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$Humidity9am
t = 31.611, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2049093 0.2313089
sample estimates:
     cor
0.218149
```

```
> cor.test(data1$Rainfall,data1$Humidity3pm)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$Humidity3pm
t = 34.656, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2249062 0.2510544
sample estimates:
      cor
0.2380234

> cor.test(data1$Rainfall,data1$Pressure9am)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$Pressure9am
t = -19.539, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1504414 -0.1232421
sample estimates:
       cor
-0.1368675
```

```
> cor.test(data1$Rainfall,data1$Temp9am)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$Temp9am
t = 0.90889, df = 19998, p-value = 0.3634
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.007432845  0.020284391
sample estimates:
        cor
0.006427007

> cor.test(data1$Rainfall,data1$Temp3pm)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$Temp3pm
t = -9.982, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08418849 -0.05660750
sample estimates:
        cor
-0.07041145
```

```
> cor.test(data1$Rainfall,data1$MinTemp)

        Pearson's product-moment correlation

data:  data1$Rainfall and data1$MinTemp
t = 13.303, df = 19998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07990289 0.10737818
sample estimates:
       cor
0.09365837
```

3.3 GRAPHS AND VISUALIZATIONS

Implementation is done in two phases using R-studio and using Python

R-studio Visualization:

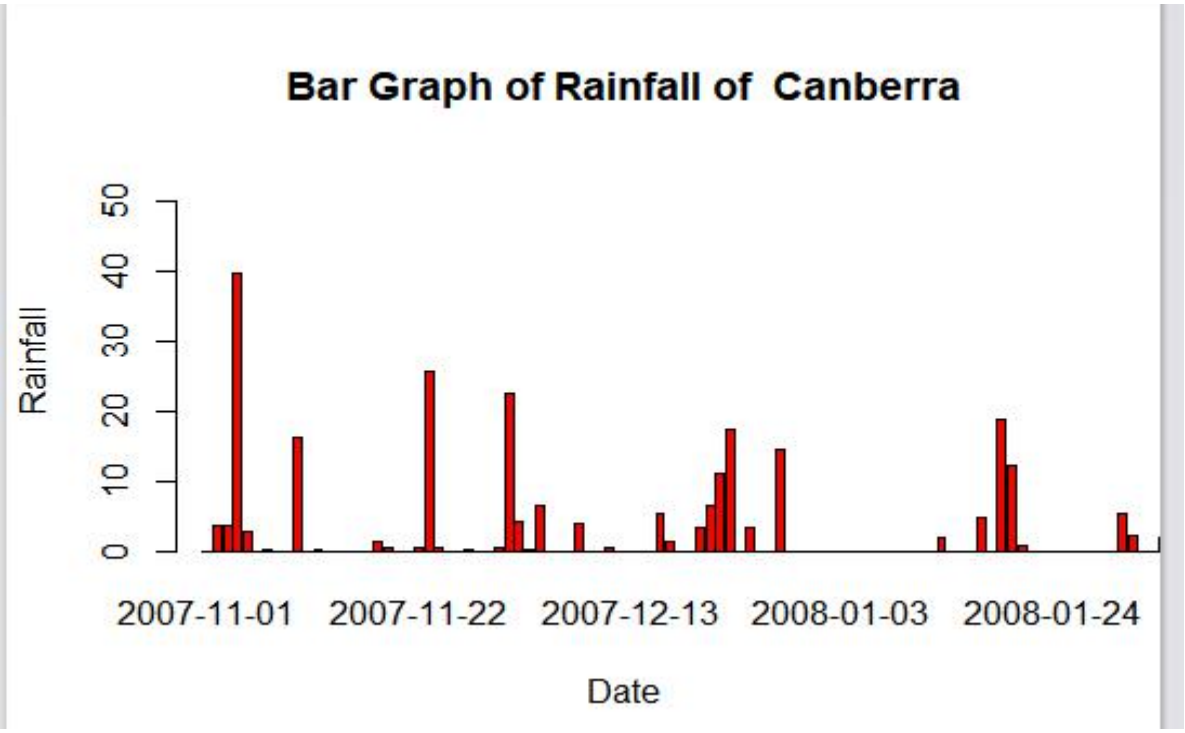
In the R-studio with the help of user input we will enter the name of the location in Australia in which the visualizations are identified

```
Please Enter the Location :
```

```
Please Enter the Location : Canberra
> var
[1] "Canberra"
>
```

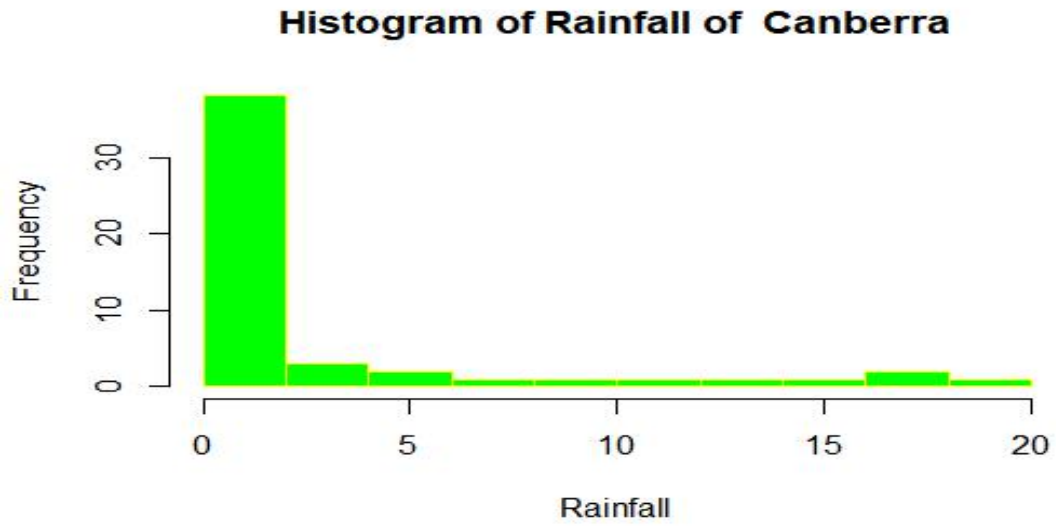
After the Entry of the location from the user the visualizations of the location to be are: -

1.Bar graph



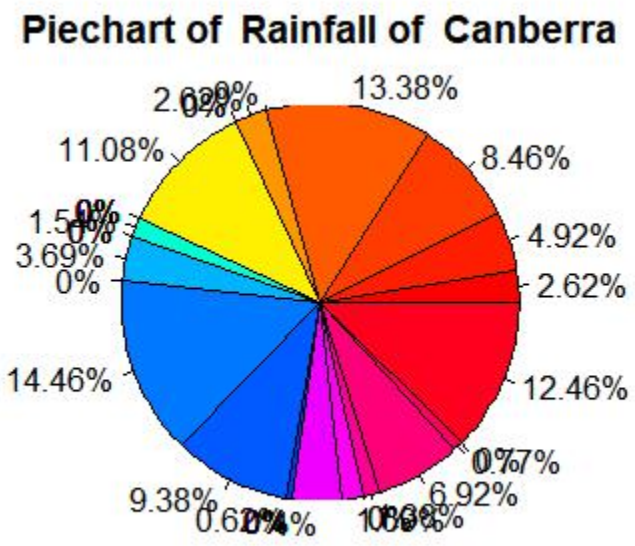
It represents the bar graph of Canberra city rainfall with the accordance of the date

2.Histogram



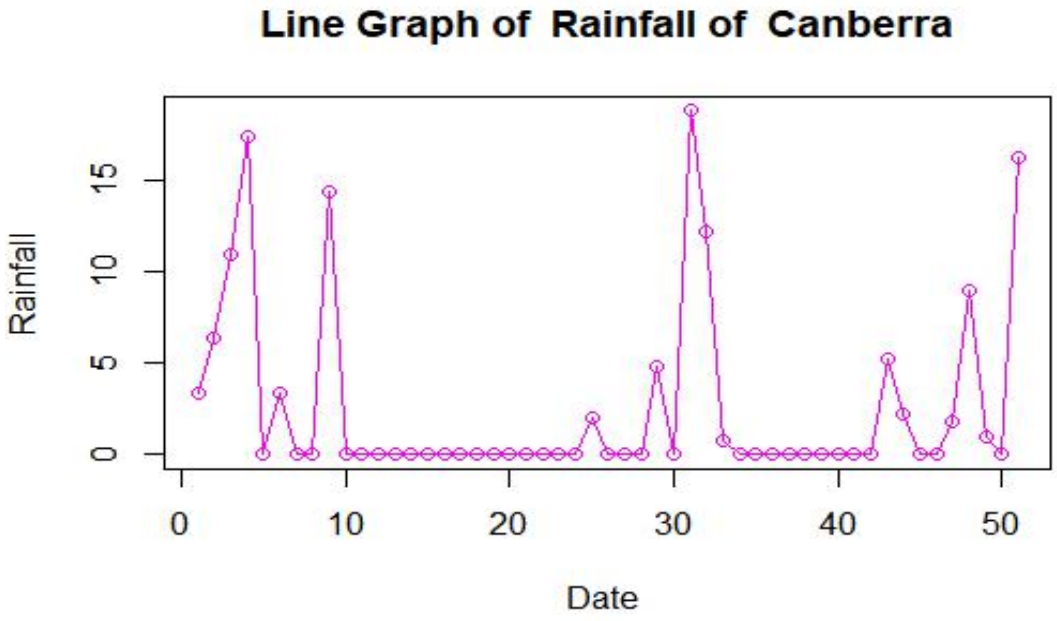
This graph gives the number of times of rainfall fall in the location of the Canberra and frequency of the value of the rainfall had fall

3.Pie chart



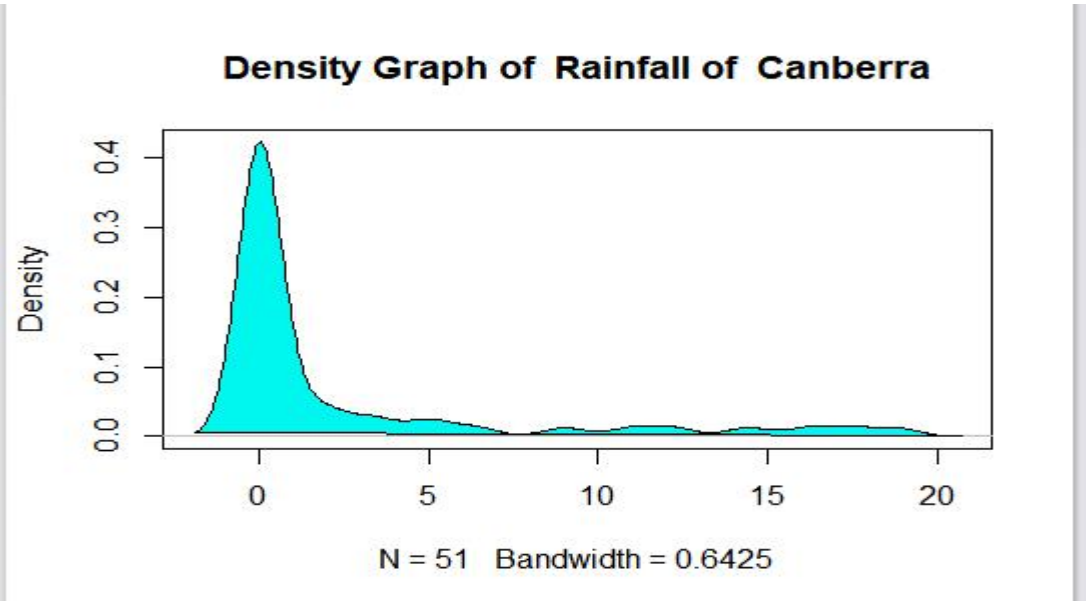
Pie chart gives the value of different values of rainfall in the Canberra location of with the percentage in the months

4.Line Graph



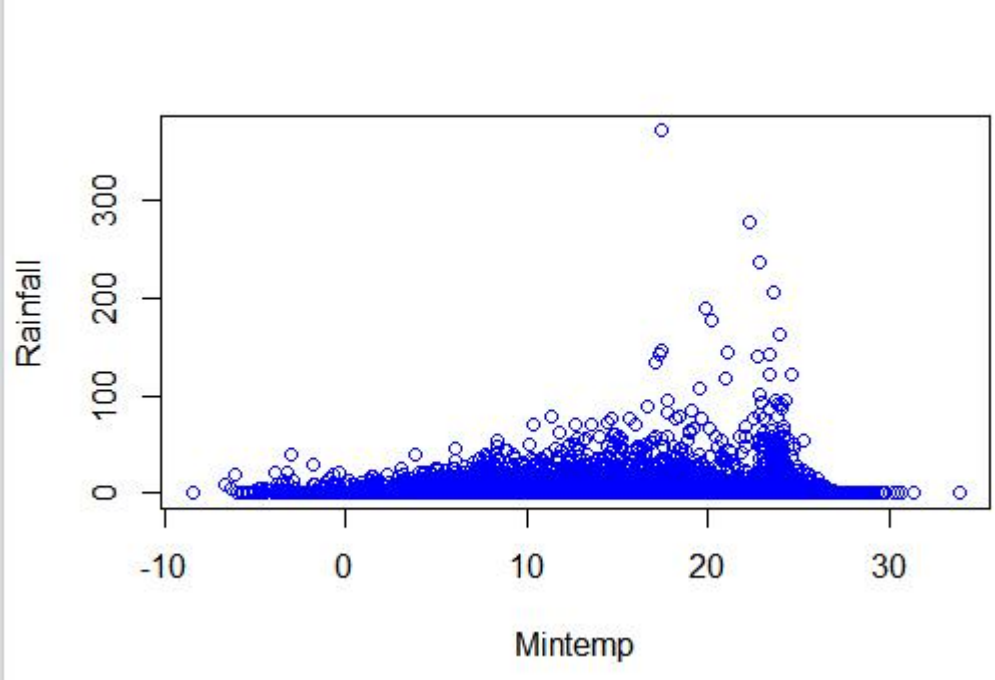
Line graph helps to find the max and lowest value of the rainfall at date in the location of the Canberra

5.Density Graph

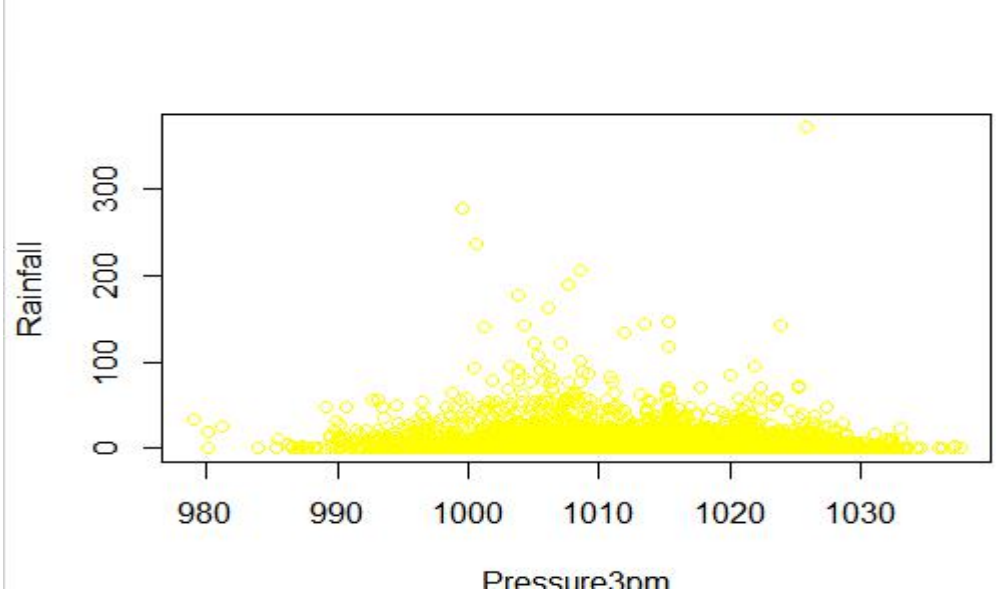


Density graph helps to find the most area in which the denser rainfall had fall in the city of the Canberra and helps to identify for the future analytics of the data

6.Scatter Plot

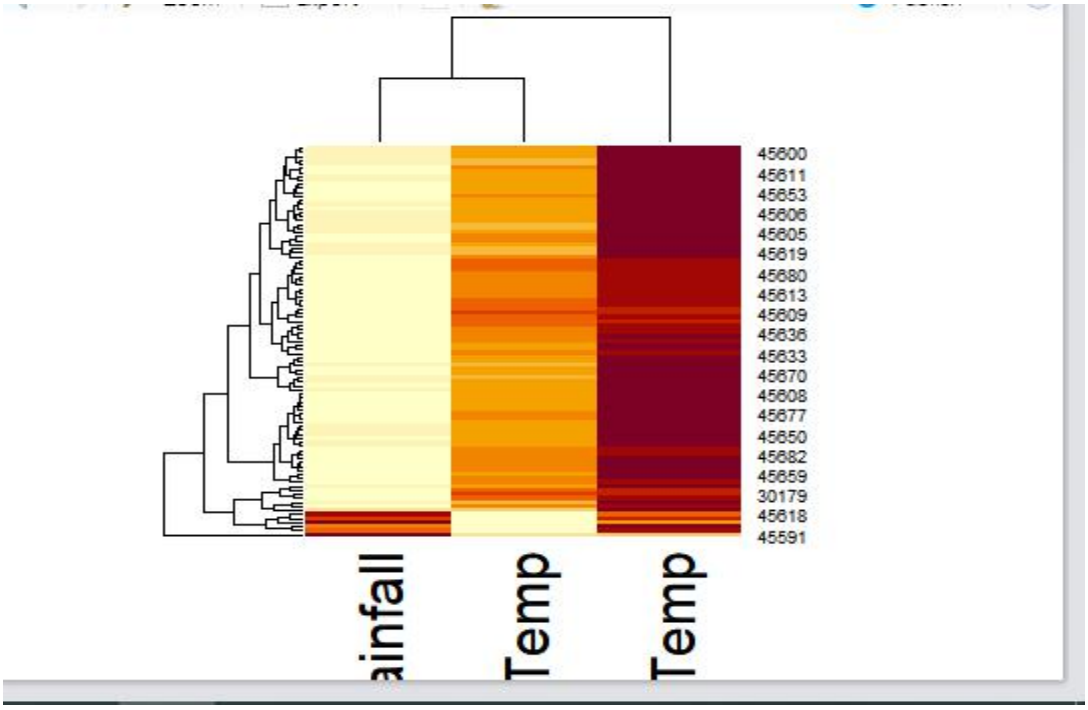


Scatter plot is for the entire data in which the rainfall vs mintemp scatter plot is plotted so we can conclude that if the mintemp is about 20-30 then the possibility of rain is higher than remaining



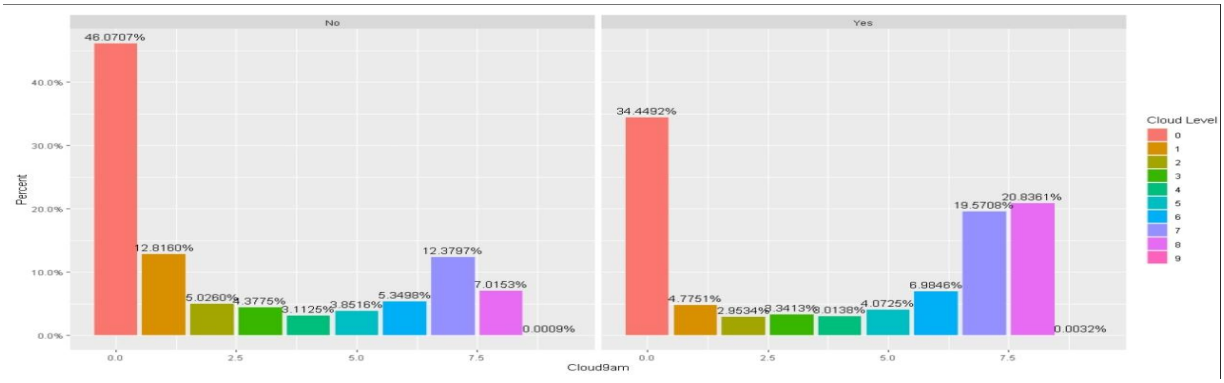
This is the Scatter plot of rainfall vs Pressure so we can conclude that at the pressure of 990-1020 the rainfall is slightly to be higher range compared to remaining

7.HeatMap



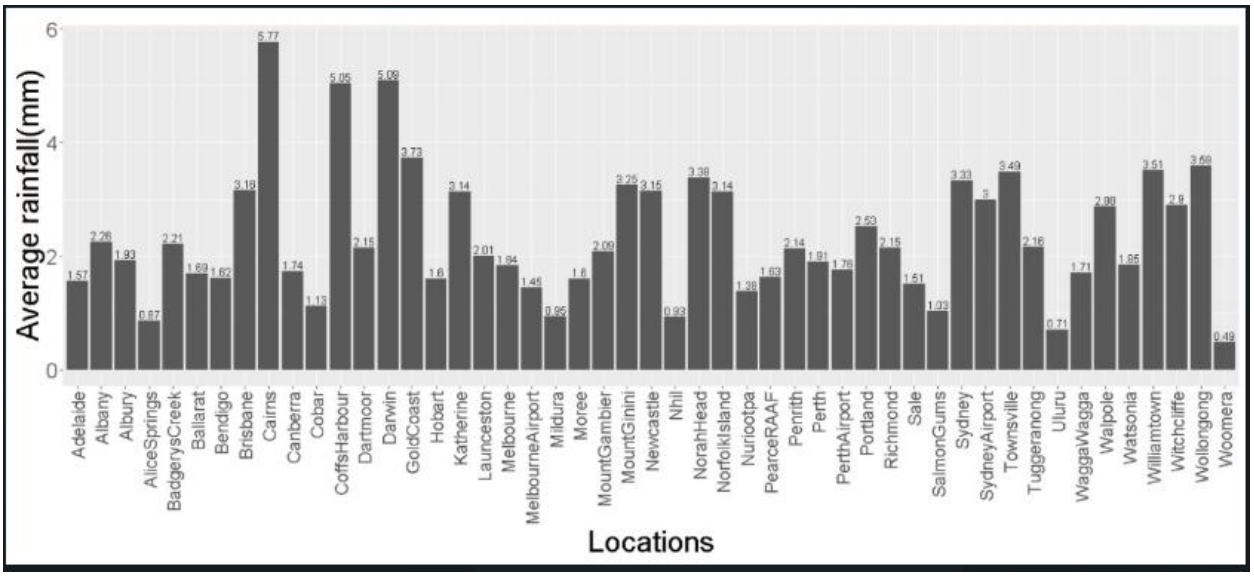
In this Heat map we can say that if the min temp is lower and max temp is lower than the possibility of rainfall is very high and when the min temp and max temp is above 45659 there is no possibility of rainfall that means the min and max temp Is very high compared to all the remaining temperatures

7.Cloud level Chart



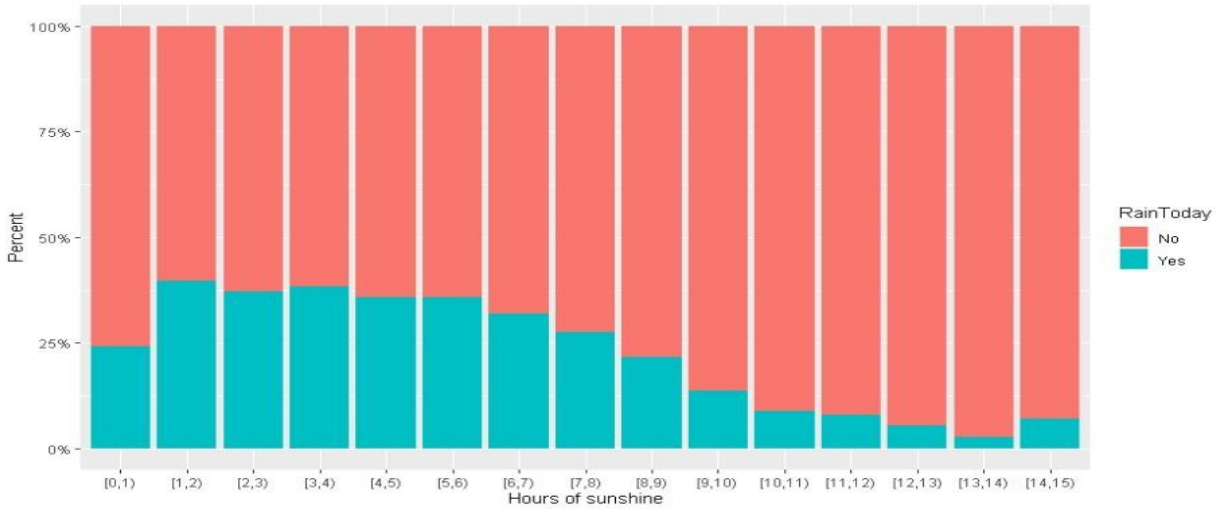
In this plot we can observe the variation in cloud level at 9am depending upon the rainfall. Therefore, we can reach to conclusion that highly intensive clouds are observed during the days with guaranteed rainfall

7.Average Rainfall Chart



We can see the variation of average rainfall across different locations in Australia

8.Sunshine chart



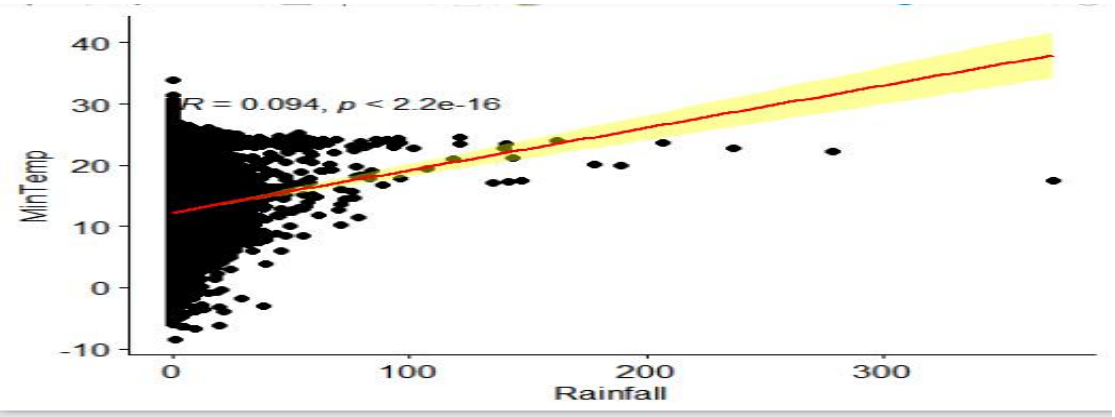
We can observe the variation of sunshine during the days of rainfall and during the days without rainfall

4 MODEL BUILDING

4.1) Linear Regression model

Linear regression model is the supervised machine learning algorithm which helps to predict the datapoints using the linear equation which helps to find the relation between the two or more variables (independent variables) and helps to predict the other variable which is dependent variable

The relation between the rainfall and mintemp is to be

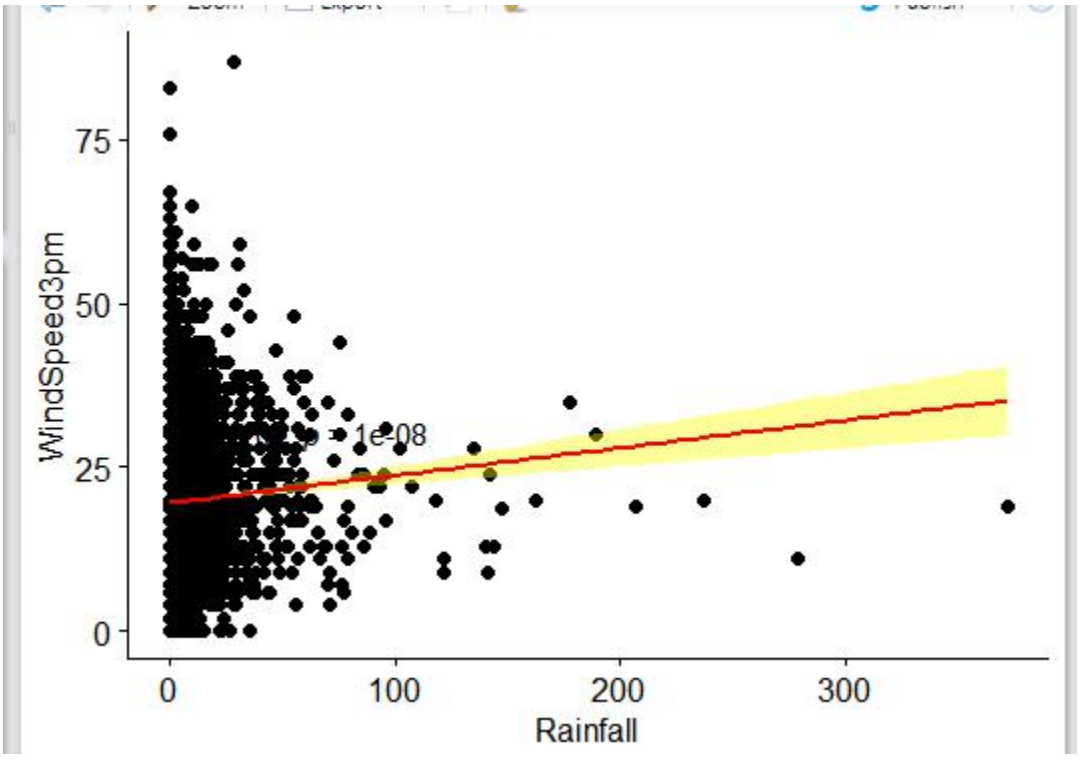


The Summary of the model which is for rainfall vs mintemp is to be: -

```
Call:
lm(formula = Rainfall ~ mintemp, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
 -4.91  -2.52  -1.79  -0.78 368.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.611527   0.132297   4.622 3.82e-06 ***
mintemp     0.126701   0.009524  13.303  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.394 on 19998 degrees of freedom
Multiple R-squared:  0.008772,  Adjusted R-squared:  0.008722
F-statistic:   177 on 1 and 19998 DF,  p-value: < 2.2e-16
```

In this model the residual standard error is very huge that is 8.394 and multiple r-squared is 0.008 in which the linear regression is not at all suitable for these two variables

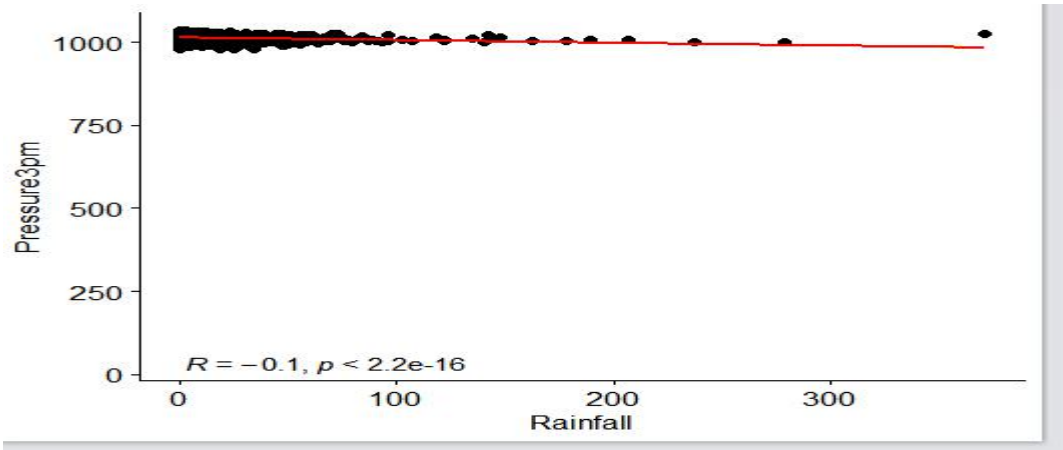
The relation between the rainfall and the windspeed is found to be



The model summary is to be which will be not suitable for the linear regression due to Heredia city condition

```
Call:
lm(formula = Rainfall ~ WindSpeed9am, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
 -7.11  -2.37  -1.79  -1.14 368.92

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.144356   0.113104   10.12   <2e-16 ***
WindSpeed9am 0.071929   0.006655   10.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.406 on 19998 degrees of freedom
Multiple R-squared:  0.005808,  Adjusted R-squared:  0.005758
F-statistic: 116.8 on 1 and 19998 DF,  p-value: < 2.2e-16
```

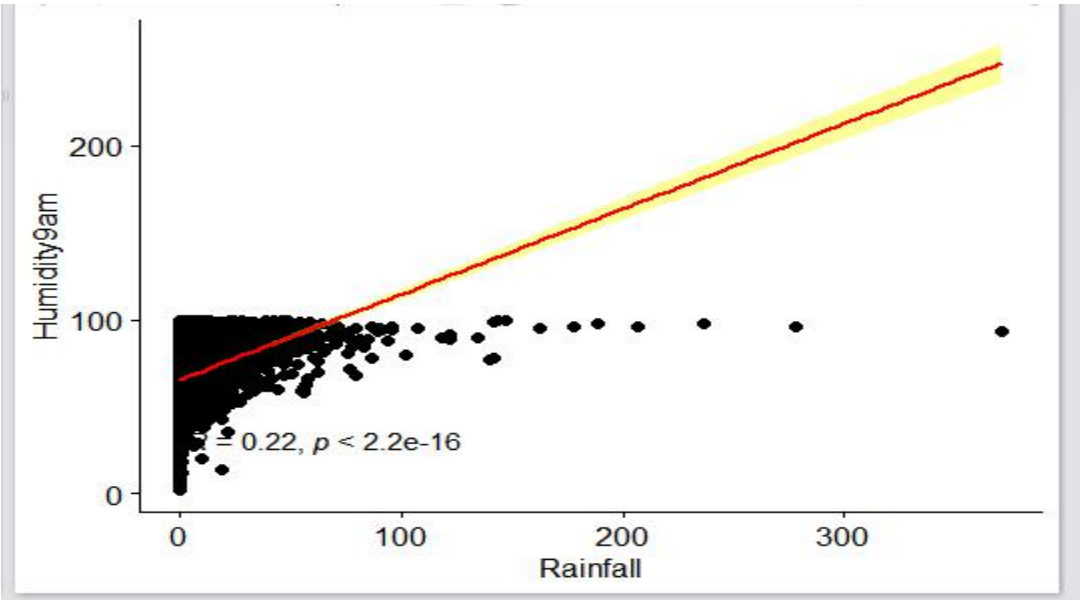
The Relation between Rainfall and pressure is to be: -



The model summary in which is suitable for the linear regression

```
Call:
lm(formula = Rainfall ~ pressure3, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
 -6.73  -2.47  -1.83  -0.73 370.36

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.331708   9.105770   15.08   <2e-16 ***
pressure3    -0.133239   0.008977  -14.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.385 on 19998 degrees of freedom
Multiple R-squared:  0.0109,    Adjusted R-squared:  0.01085
F-statistic: 220.3 on 1 and 19998 DF,  p-value: < 2.2e-16
```

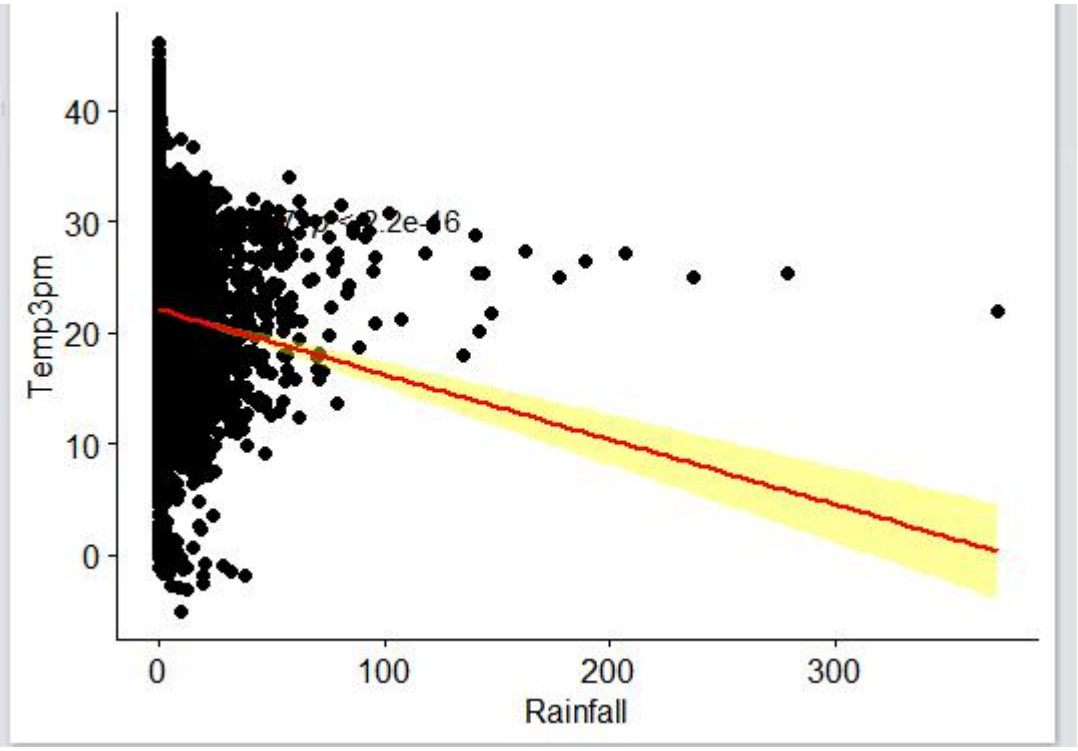
The Relation between the rainfall and humidity is to be: -



The model summary of rainfall vs humidity is not fit for the linear regression

```
Call:
lm(formula = Rainfall ~ Humidity3pm, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
 -7.04  -2.78  -1.33   0.46 365.80

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.631854   0.150554  -17.48   <2e-16 ***
Humidity3pm  0.096712   0.002791   34.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.189 on 19998 degrees of freedom
Multiple R-squared:  0.05666,   Adjusted R-squared:  0.05661
F-statistic:  1201 on 1 and 19998 DF,  p-value: < 2.2e-16
```

The Relation between the temperature and rainfall is followed to be: -



The model summary of rainfall vs temperature is not fit for the linear regression

```
Call:
lm(formula = Rainfall ~ Temp9am, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
 -2.36  -2.21  -2.15  -1.44 368.80

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.040214   0.169488  12.038   <2e-16 ***
Temp9am     0.008389   0.009230   0.909    0.363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.431 on 19998 degrees of freedom
Multiple R-squared:  4.131e-05, Adjusted R-squared:  -8.697e-06
F-statistic: 0.8261 on 1 and 19998 DF,  p-value: 0.3634
```

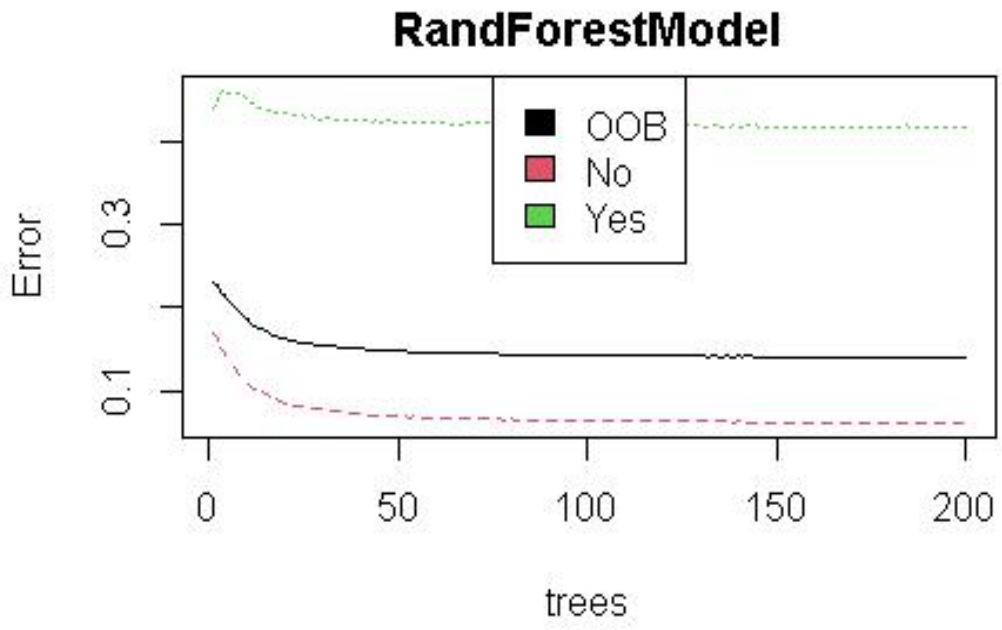
4.2) Random Forest Classifier Model: -

Random Forest classifier is a supervised learning algo which is useful when there are many decisions which creates the multitude of decision trees at the training time and them taking a classification of the output helps to produce more effective results when this data is passed to the model, we got the accuracy of 84 percentage which is likely better result has more recall, f1-score and has high support for the dataset

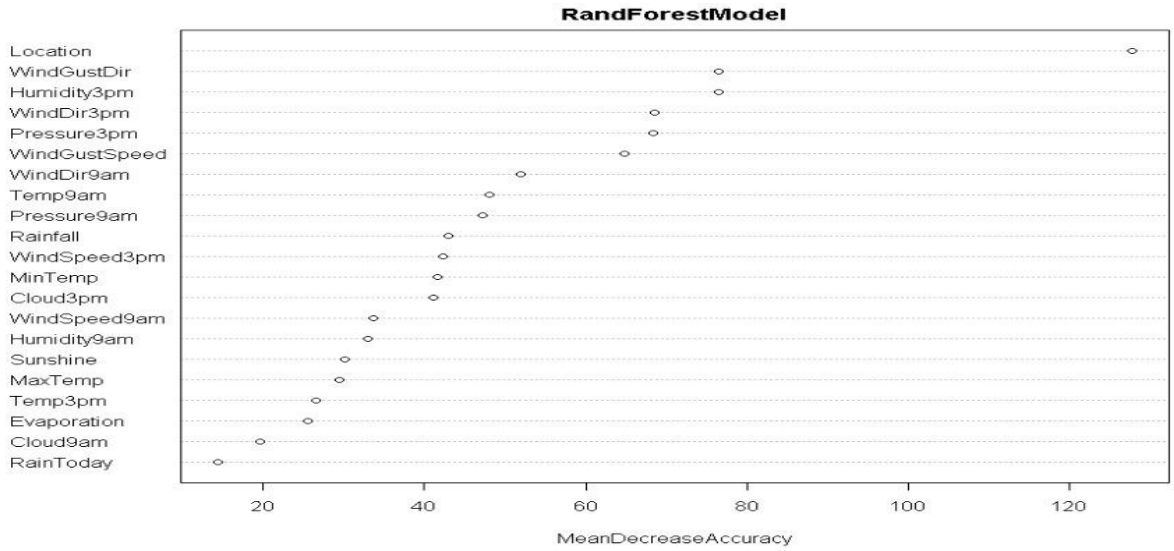
```
RF_misclassification_rate = error/sizeTestSet
RF_misclassification_rate
```

```
[1] 0.1419845
```

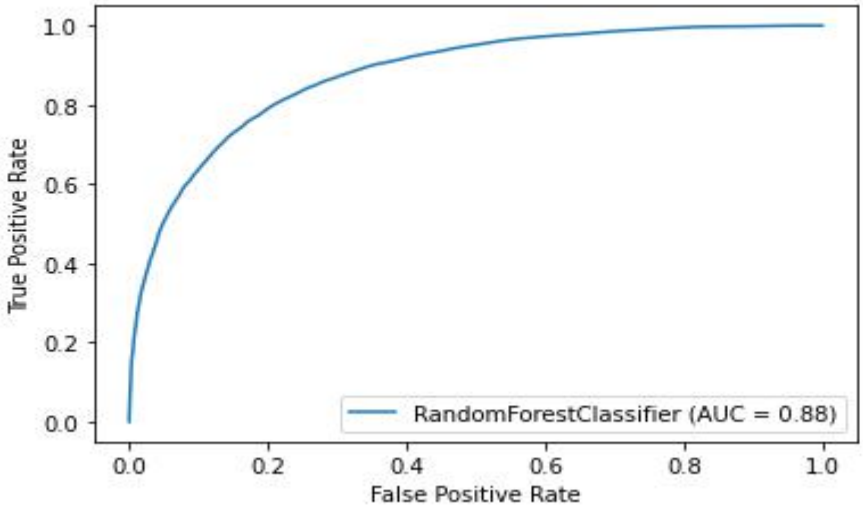
So, the misclassification rate of the random forest we got as 14 percentage and the accuracy value we got as86 percentage



Mean decease Accuracy Graph or Var Imp graph of random forest: -



The metrics of the Random Forest is: -



4.3) Logistic Regression Model Building: -

When we train our data in the logistic regression model the accuracy is found to be 78 Percentage since the logistic is the supervised classification algorithm the model predicts the probability of the data based on the input of the data and classifies whether the rainfall occurs tomorrow or not represented by 1 or 0

The model is trained with logistic regression the precision is likely to be higher that to be not rain tomorrow

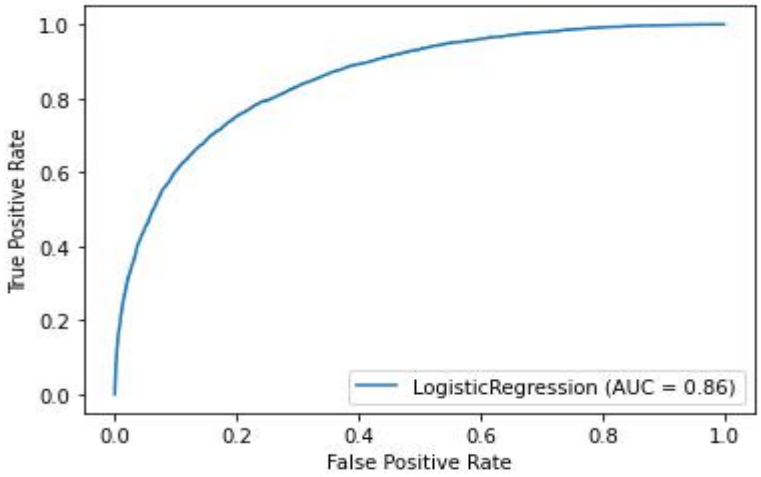
Since the Rain tomorrow and rain today variables are in the output of the code in which the logistic regression would be very trained and tested to give the much effective accuracy and classification rate is higher in the logistic regression

```
    print("accuracy:")
    print(LR_accuracy_rate) #


```

```
[1] "accuracy:"
[1] 0.849208
```

Using Logistic regression, the accuracy percentage that got in this is equals to 85 percentage

The Metrics for the logistic regression is: -



4.4) KNN Model Building: -

Division of data takes place in two ways that will be training and testing we train our data with the passing of the 80 percent of the data to the model and remaining 20 percentage is provided to be testing for the model. In this type the data is trained using the KNN machine learning algorithm in which the K-nearest

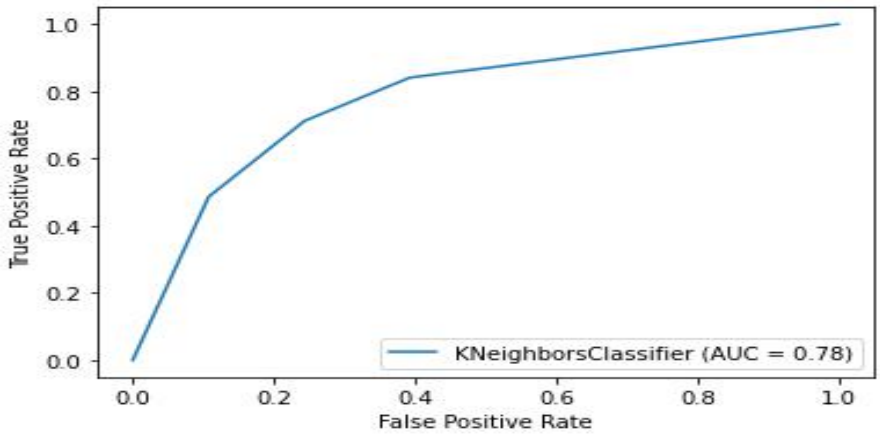
Neighbours could be found and can be trained as model the results of KNN as follows:

```
[1] 3
[1] 0.1800909
[1] 4
[1] 0.181689
[1] 5
[1] 0.1700405
[1] 6
[1] 0.1722423
[1] 7
[1] 0.1667732
```

KNN for the Neighbours of 3 is 0.18 and 4 is 0.18 like this so on the least value of the misclassification is 7 Neighbour so we consider 7 Neighbour

When we fit the data into KNN the accuracy percentage is likely to be 84 percentage

The metrics plot for KNN classifier is: -



4.5) Decision tree Model

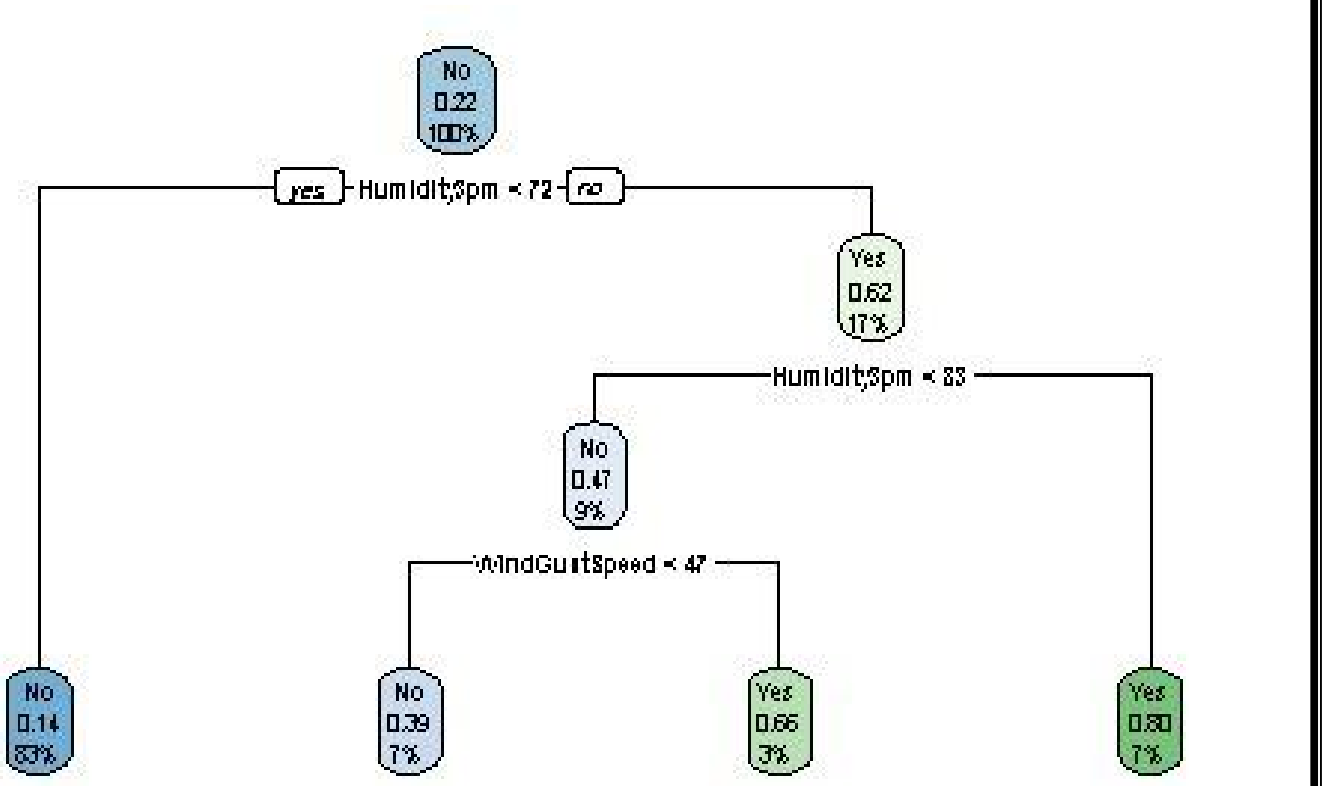
Division of data takes place in two ways that will be training and testing we train our data with the passing of the 80 percent of the data to the model and remaining 20 percentage is provided to be testing for the model. In this type the

data is trained using the Decision tree model. The accuracy of this model that got using this dataset is 82 percentage

```
# Calculate the accuracy rate of decison tree
accuarcy_rate = 1-error/sizeTestSet
accuarcy_rate
```

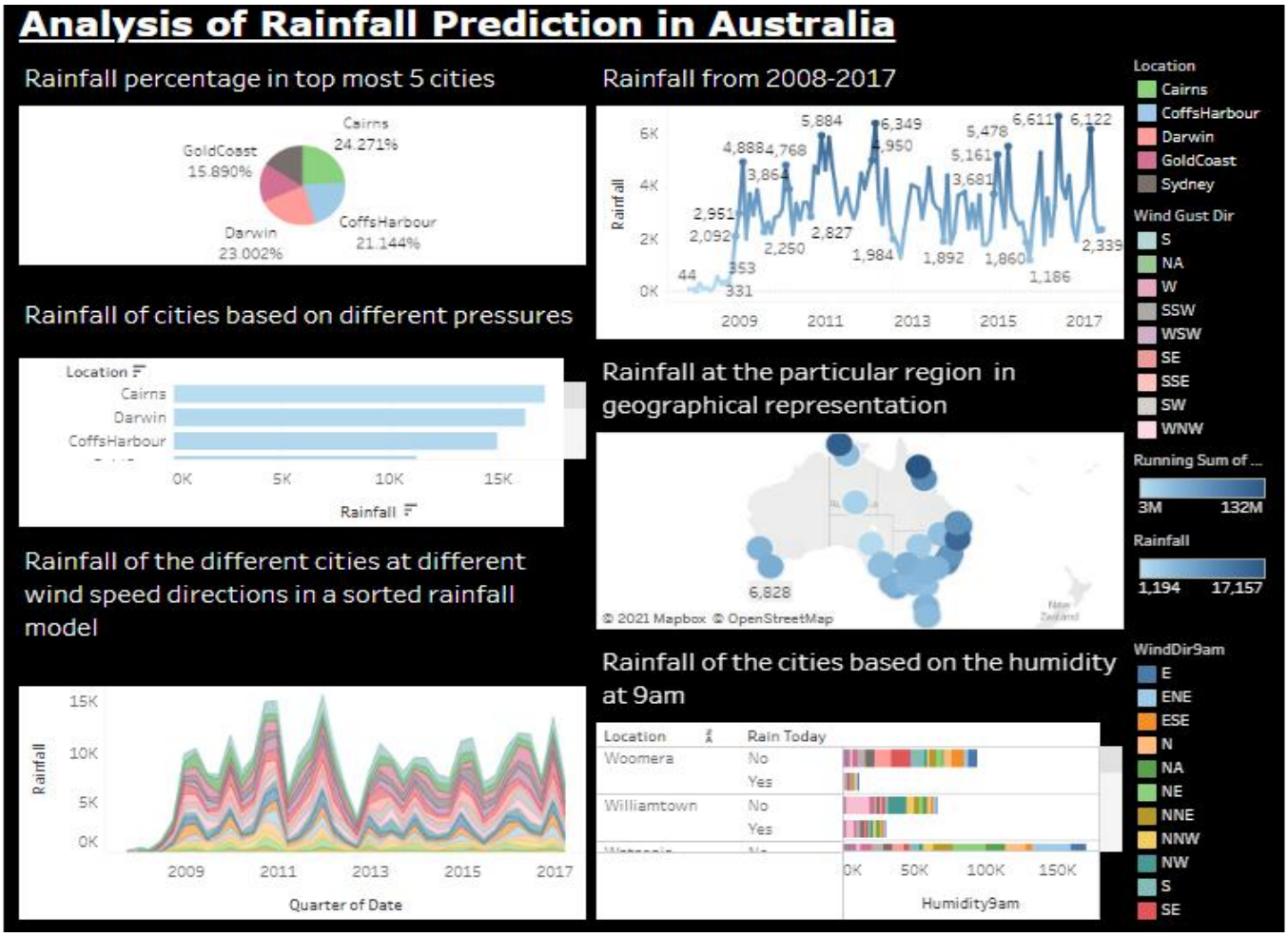
```
[1] 0.8283969
```

The Decision tree model is given by: -



So, we can say that the humidity is the root node which has more correlation to the rainfall which pass from the level wise of the variables

5.DASHBOARD VISUALIZATION



6.CONCLUSIONS:

The estimation of rainfall is important in terms of resource management and farming. It can be met with an incorrect or incomplete estimation as rainfall changes from region to region. In this paper, the weather Australia dataset is taken where it features fields like humidity, temperature, pressure, sunlight, evaporation. Five machine learning models have been developed: linear regression, logistic regression, random forest, KNN. The models are trained with 80% of the data and the rest of the data is used for testing, amongst five models Random Forest and logistic regression 83 and 85% respectively, because of the non-linear relation between the fields linear regression had very less accuracy. The daily wise data-set makes the model more reliable and accurate.

7.REFERENCES:

1. A rainfall prediction model using artificial neural network

L Shaikh, K Sawlani - International Journal of Technical Research and …, 2017 - ijsrnsc.org

2.Development of advanced artificial intelligence models for daily rainfall prediction

BT Pham, LM Le, TT Le, KTT Bui, VM Le, HB Ly… - Atmospheric …, 2020 - Elsevier

3. Rainfall Prediction using Machine Learning & Deep Learning Techniques

CZ Basha, N Bhavana, P Bhavya… - … on Electronics and …, 2020 - ieeexplore.ieee.org