# Weather Report of Australia

25/03/2021

## Dataset: Australian weather report from 10years.

```
The task begins with loading the packages and the weatherAUS dataset.

The dataset consists of 145460 rows and 23 columns.
```

## Dataset loaded and NA values removed – currently 56420 rows.

```
there are total of 3,00,000+ NA values all over the dataset.
```

```
data <- data[complete.cases(data),]
```

```
this is the syntax for removing the NA values.
```

## The following are the columns of the dataset:

```
names(data)

## [1] "Date"          "Location"      "MinTemp"       "MaxTemp"
## [5] "Rainfall"      "Evaporation"   "Sunshine"      "WindGustDir"
## [9] "WindGustSpeed" "WindDir9am"    "WindDir3pm"    "WindSpeed9am"
## [13] "WindSpeed3pm" "Humidity9am"   "Humidity3pm"   "Pressure9am"
## [17] "Pressure3pm"   "Cloud9am"      "Cloud3pm"      "Temp9am"
## [21] "Temp3pm"       "RainToday"     "RainTomorrow"
```

**Basic statistics** of the dataset includes minimum value, maximum value , median , 1st quadrant, 3rd quadrant, length etc;

```
summary(data)

##      Date             Location           MinTemp         MaxTemp
##  Length:56420       Length:56420       Min.   :-6.70   Min.   : 4.10
##  Class :character   Class :character   1st Qu.: 8.60   1st Qu.:18.70
##  Mode  :character   Mode  :character   Median :13.20   Median :23.90
##                                        Mean   :13.46   Mean   :24.22
##                                        3rd Qu.:18.40   3rd Qu.:29.70
##                                        Max.   :31.40   Max.   :48.10
##     Rainfall        Evaporation         Sunshine       WindGustDir
##  Min.   :  0.00   Min.   : 0.000   Min.   : 0.000   Length:56420
##  1st Qu.:  0.00   1st Qu.: 2.800   1st Qu.: 5.000   Class :character
##  Median :  0.00   Median : 5.000   Median : 8.600   Mode  :character
```

```
##  Mean   :  2.13   Mean   : 5.503   Mean   : 7.736  
##  3rd Qu.:  0.60   3rd Qu.: 7.400   3rd Qu.:10.700  
##  Max.   :206.20   Max.   :81.200   Max.   :14.500  
##  WindGustSpeed     WindDir9am         WindDir3pm         WindSpeed9am  
##  Min.   :  9.00   Length:56420       Length:56420       Min.   : 2.00  
##  1st Qu.: 31.00   Class :character   Class :character   1st Qu.: 9.00  
##  Median : 39.00   Mode  :character   Mode  :character   Median :15.00  
##  Mean   : 40.88                                         Mean   :15.67  
##  3rd Qu.: 48.00                                         3rd Qu.:20.00  
##  Max.   :124.00                                         Max.   :67.00  
##   WindSpeed3pm    Humidity9am      Humidity3pm     Pressure9am    
##  Min.   : 2.00   Min.   :  0.00   Min.   :  0.0   Min.   : 980.5  
##  1st Qu.:13.00   1st Qu.: 55.00   1st Qu.: 35.0   1st Qu.:1012.7  
##  Median :19.00   Median : 67.00   Median : 50.0   Median :1017.2  
##  Mean   :19.79   Mean   : 65.87   Mean   : 49.6   Mean   :1017.2  
##  3rd Qu.:26.00   3rd Qu.: 79.00   3rd Qu.: 63.0   3rd Qu.:1021.8  
##  Max.   :76.00   Max.   :100.00   Max.   :100.0   Max.   :1040.4  
##   Pressure3pm        Cloud9am        Cloud3pm        Temp9am     
##  Min.   : 977.1   Min.   :0.000   Min.   :0.000   Min.   :-0.7  
##  1st Qu.:1010.1   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:13.1  
##  Median :1014.7   Median :5.000   Median :5.000   Median :17.8  
##  Mean   :1014.8   Mean   :4.242   Mean   :4.327   Mean   :18.2  
##  3rd Qu.:1019.4   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:23.3  
##  Max.   :1038.9   Max.   :8.000   Max.   :9.000   Max.   :39.4  
##     Temp3pm       RainToday         RainTomorrow      
##  Min.   : 3.70   Length:56420       Length:56420      
##  1st Qu.:17.40   Class :character   Class :character  
##  Median :22.40   Mode  :character   Mode  :character  
##  Mean   :22.71                                        
##  3rd Qu.:27.90                                        
##  Max.   :46.10
```

```
str(data)
```

str() function describes the structure of the dataset.

Below given is the output of the str() function showing the datatypes of the features/columns.

It is a dataframe of 56420 rows and 23 columns.

```
## 'data.frame':    56420 obs. of  23 variables:
```

Date is the char type of variable that needs to be modified into the

**Date() type.**

```
##  $ Date         : chr  "01-01-2009" "02-01-2009" "04-01-2009" "05-01-2009" ...
##  $ 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 ...
```

```
These are the very important variables for the project with chr type and to b
e modified into integer data type.
```

```
##  $ RainToday    : chr  "No" "No" "No" "No" ...
##  $ RainTomorrow : chr  "No" "No" "No" "No" ...
```

```
data[data$RainToday == "No",]$RainToday <- 0
data[data$RainToday == "Yes",]$RainToday <- 1
data[data$RainTomorrow == "Yes",]$RainTomorrow <- 1
data[data$RainTomorrow == "No",]$RainTomorrow <- 0
```

```
data$RainToday <- as.integer(data$RainToday)
data$RainTomorrow <- as.integer(data$RainTomorrow)
```

```
datatypes changed!!!
```

```
str(data)
```

```
##  $ RainToday    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RainTomorrow : int  0 0 0 0 0 0 0 0 0 0 ...
```

# EXPLORATORY DATA ANALYSIS

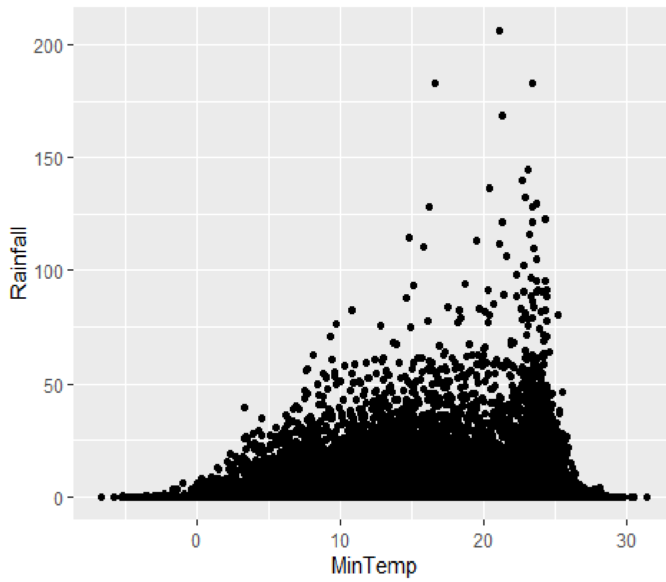
The three ways of looking into the data are

**Variables, Variables with respect to location , Variables with respect to time.**

These plots are plotted only to what kind of relationship they have with the rainfall parameter…

These doesn't matter when it comes to Raintomorrow (our dependent variable).

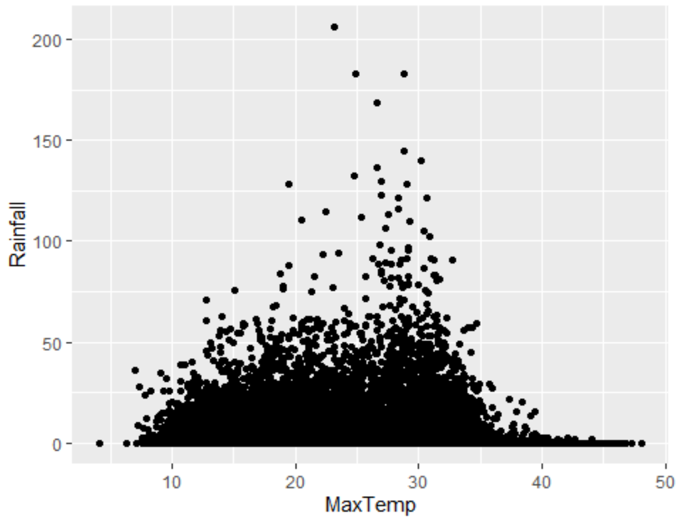
## Mintemp vs Rainfall:




**Inference from the plot:**

The above graph is the overall relationship between the Mintemp variable and the Rainfall variable. From the graph it can be said that there is no positive or negative correlation between the variables . but it seems like the rainfall is more common or there Rainfall generally exists between 0 to 50pts and that increases between 20 to 30 temperatures.

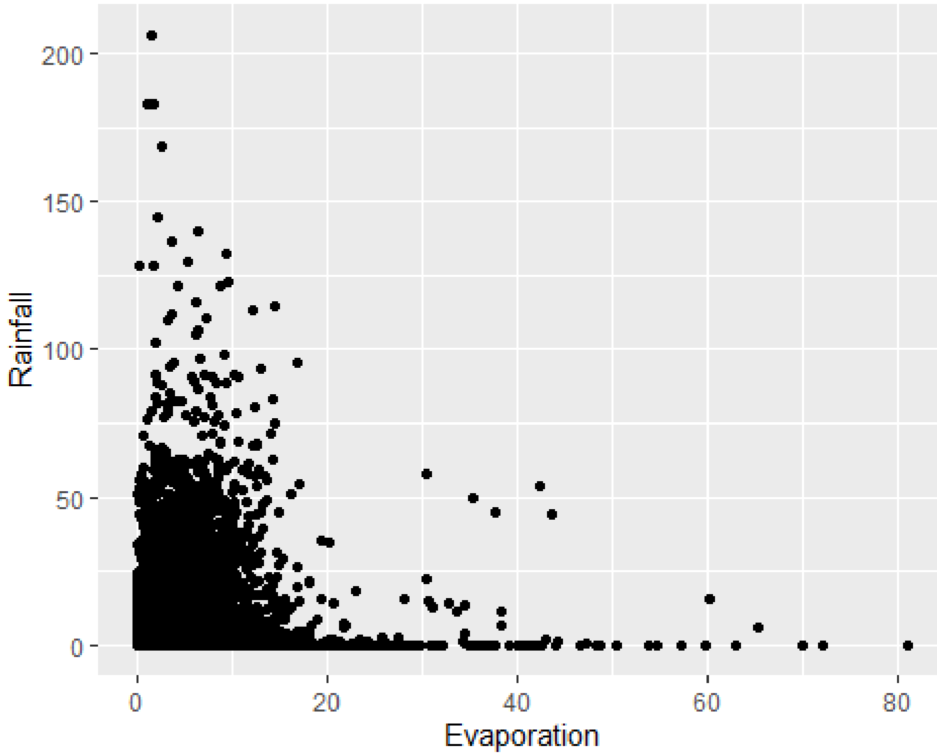
## Maxtemp vs Rainfall:




**Inference from the plot:**

The above graph is the overall relationship between the Maxtemp variable and the Rainfall variable. From the graph it can be said that there is no positive or negative correlation between the variables . the outliers among the data variables seem to have more effect on these plots.

## Evaporation vs Rainfall



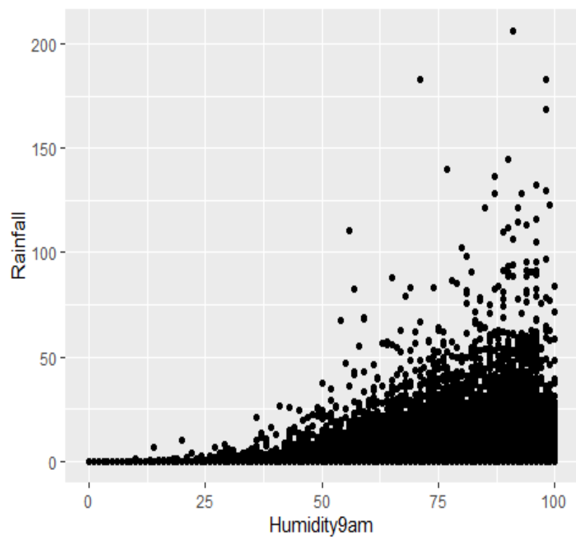

**Inference from the plot:**

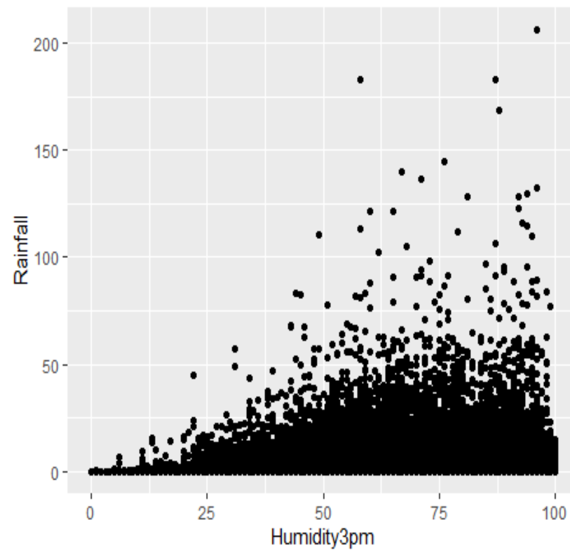
The above graph is the overall relationship between the Evaporation variable and the Rainfall variable. It is seen that the points ranging 0 to 20 of evaporation has more Rainfall rates in Australia.

This dataset is collected all over the Australia and the continent is surrounded by oceans and seas.

That being the case with the lower evaporation rates there is high chance of Rainfall.

## Humidity9am and humdity3pm vs Rainfall

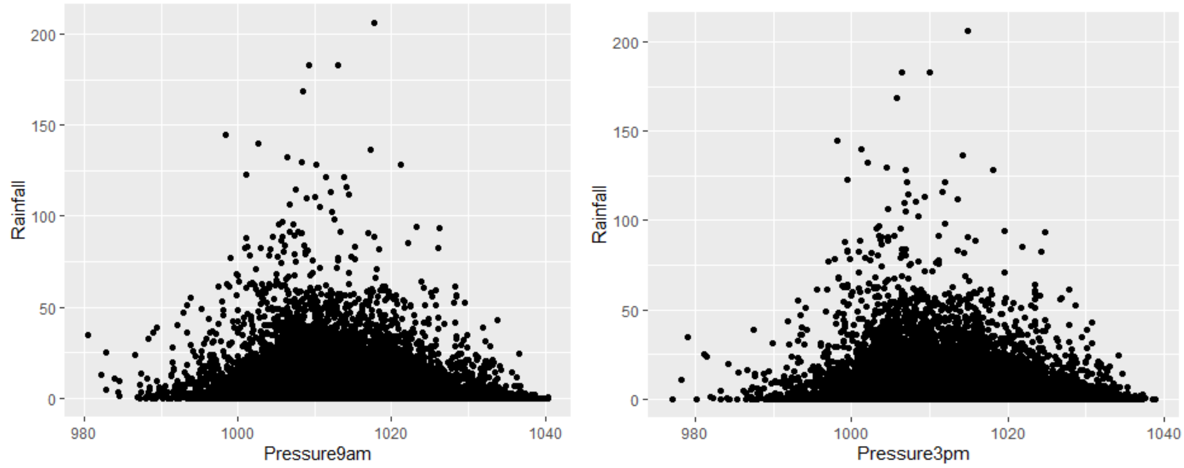



**Inference from the plot:**

The above graph is the overall relationship between the Mintemp variable and the Rainfall variable. From the graph it can be said that there is no positive or negative correlation between the variables . but it seems like the rainfall is more common or there Rainfall generally exists between 0 to 50pts and that increases between 20 to 30 temperatures.

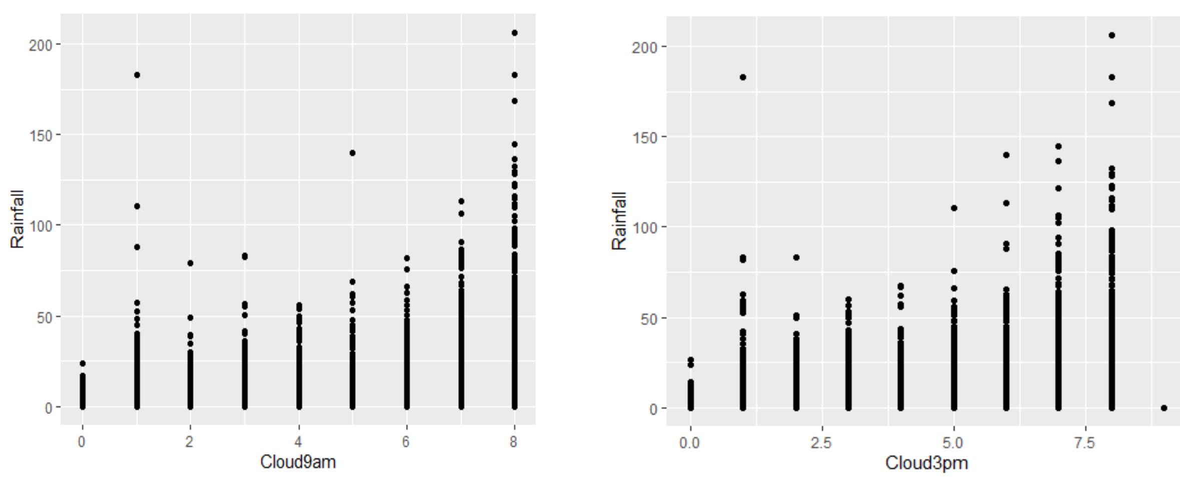
## Pressure9am Pressure3pm vs Rainfall




**Inference from the plot:**

The above graph is the overall relationship between the Pressure variable and the Rainfall variable. Pressure is the very important factor for determing the Rainfall. This graph shows that the pressure and Rainfall relation is constant throughout the day irrespective of timing

## cloud9am and cloud3pm vs Rainfall

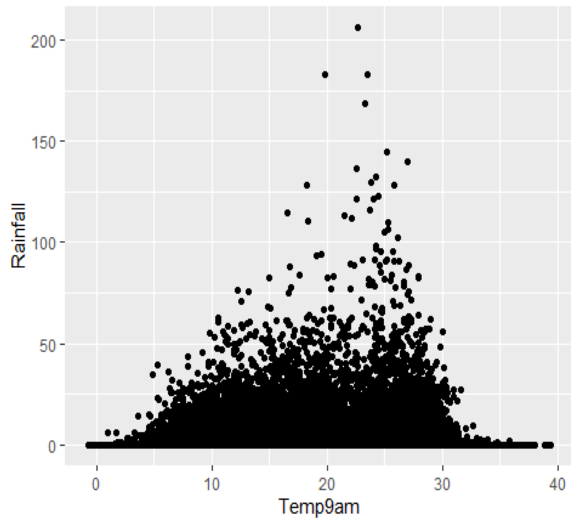


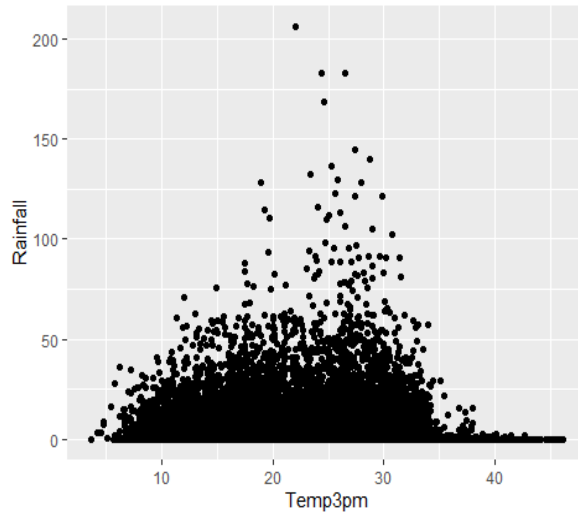

**Inference from the plot:**

The above graph is the overall relationship between the Cloud variables and the Rainfall variable. From the graphs it can be said that cloud variable with respect to Rainfall variable is a factor variable. It is also seen that clouds at 7.5 and 8 values have much Rainfall rates.

Cloud variable ranges from 0 to 8 in the given dataset.

## Temp9am Temp3pm vs Rainfall




**Inference from the plot:**

The above graph is the overall relationship between the Temperature variables and the Rainfall variable. From the graph it can be said temperature and Rainfall are consistent through the day. On any point of time be it morning or evening the occurance of Rainfall can be same with respect to Temperature. Outliers are still in our consideration.

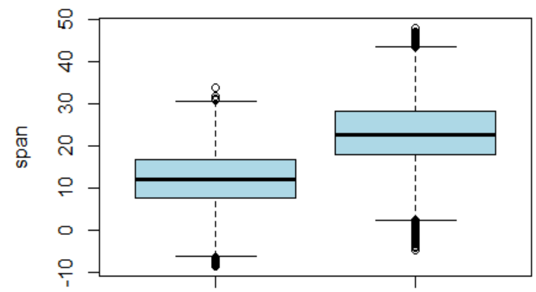
# Outliers: Boxplot

**boxplots** generally helps in extracting the information regarding the span of the data.

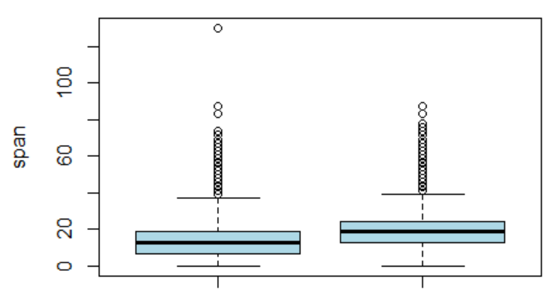
How the data is spanned. Is it skewed or normal etc etc;

In the working of the project the outliers are removed.

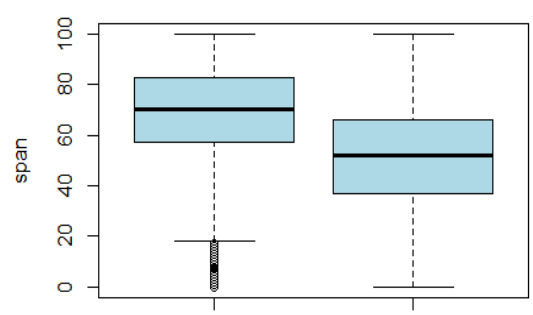
# Box plots



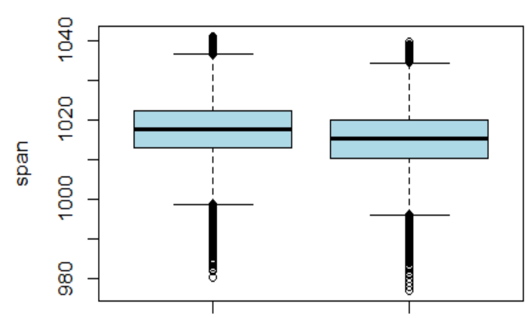

MinTemp and MaxTemp



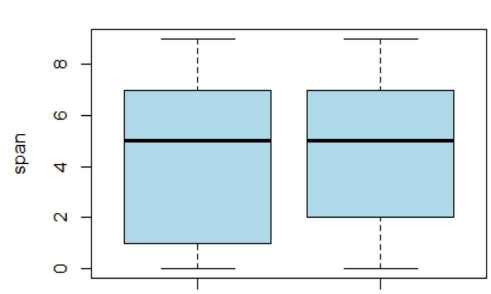

WindSpeed9am and WindSpeed3pm



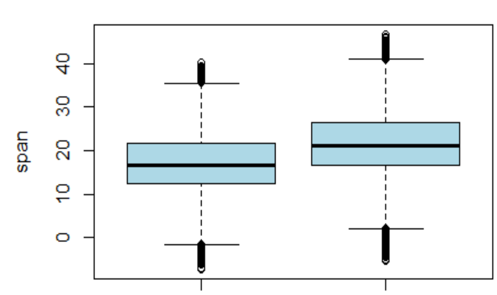

Humidity9am and Humdity3pm




Pressure9am and pressure3pm




Cloud9am and Cloud3pm




temp9am and temp3pm

## **PREPROCESSING THE DATA…** with respect to Locations!

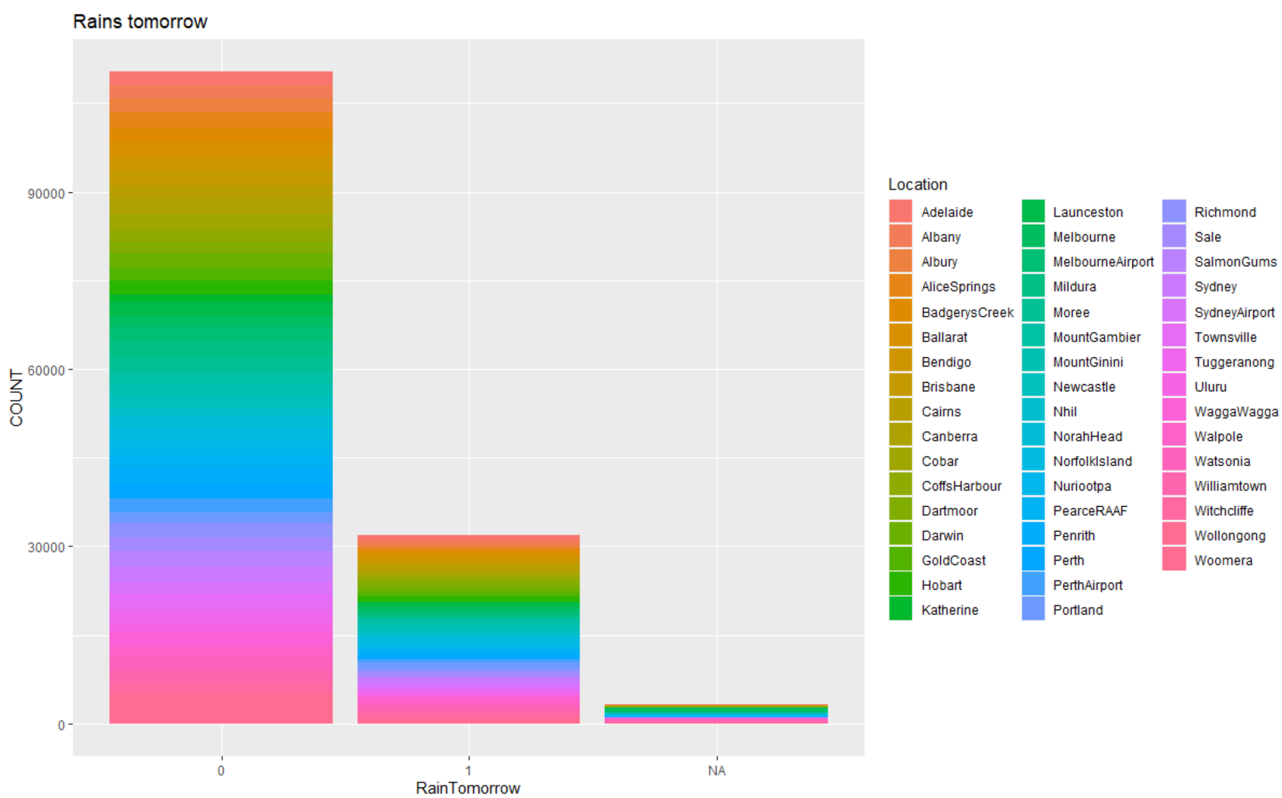
## column RainTomorrow and RainToday is in yes/no format

## Converting it to 0 or 1

```
data$RainToday<-str_replace_all(data$RainToday,"No","0")
data$RainToday<-str_replace_all(data$RainToday,"Yes","1")
data$RainTomorrow<-str_replace_all(data$RainTomorrow,"No","0")
data$RainTomorrow<-str_replace_all(data$RainTomorrow,"Yes","1")
```

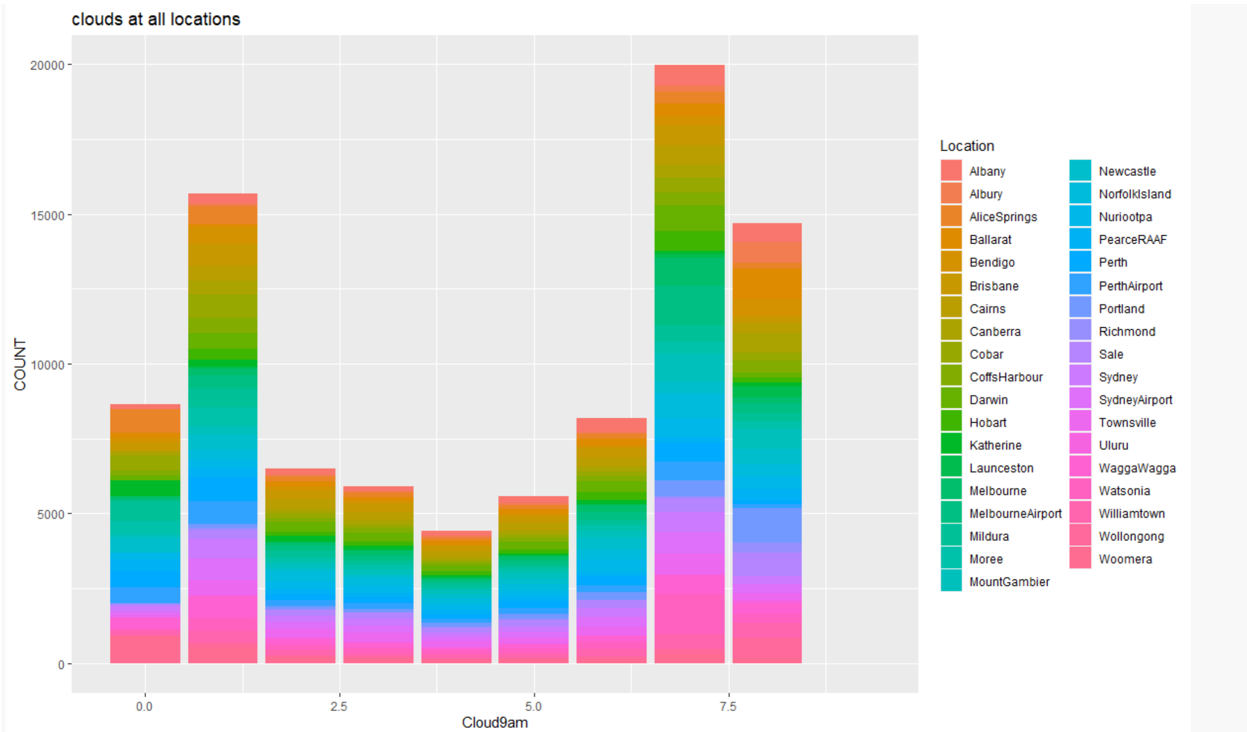
### Time for some plots

Raintomorrow values over various locations.

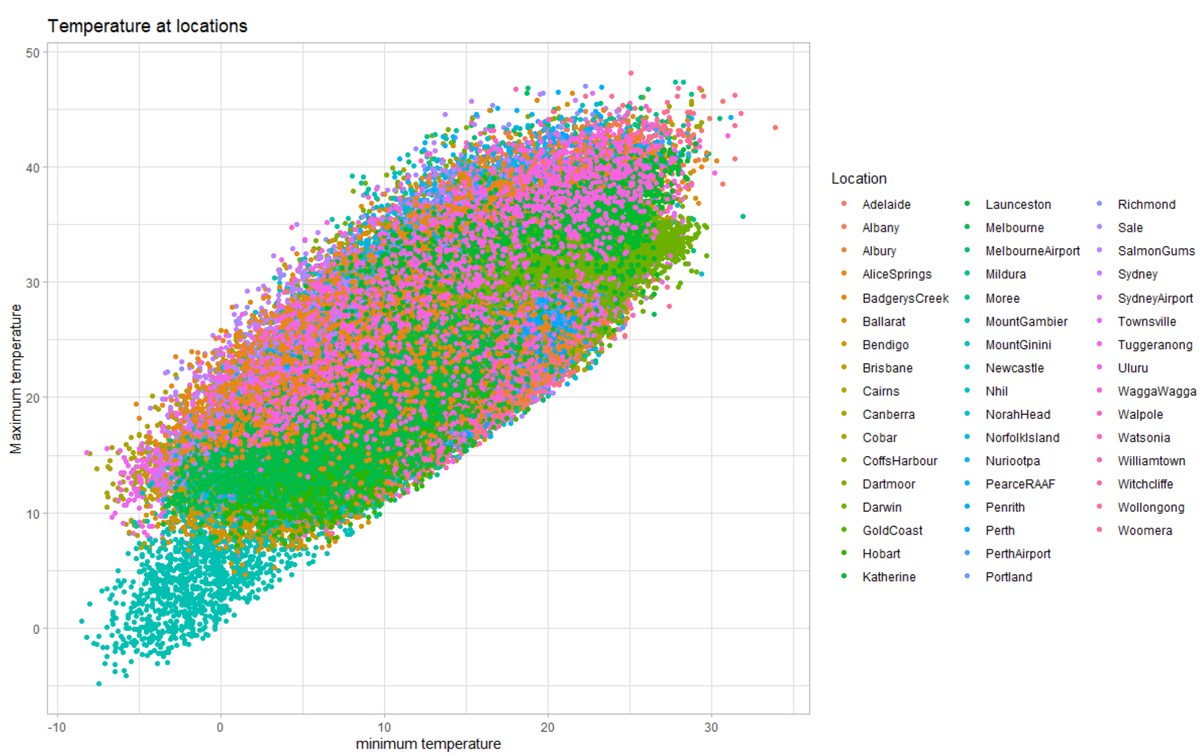



```
The feature has more number of  no-rains (0) than the yes_rains(1).
```

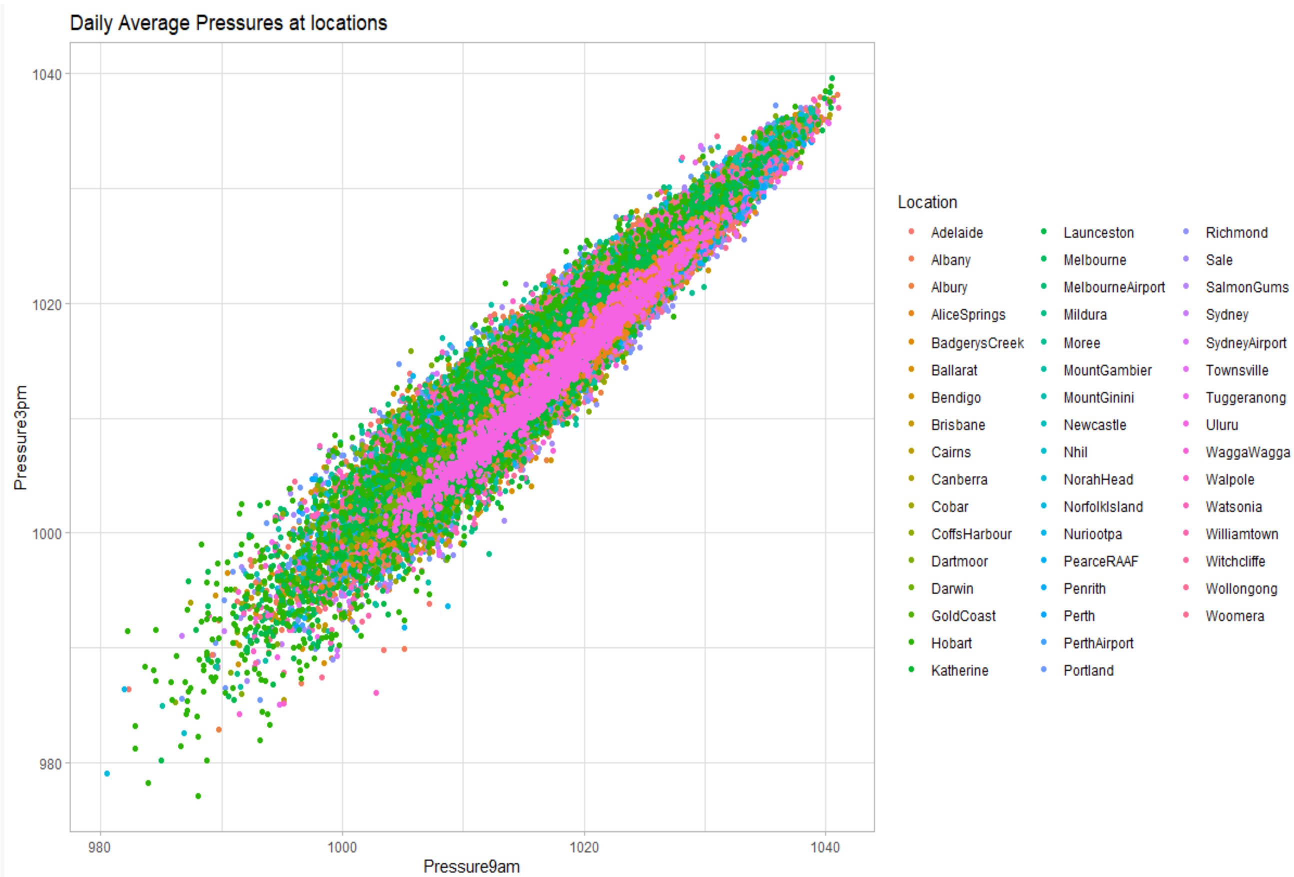
# Clouds at various Locations.




## Temperatures at different locations



## Pressure at different locations

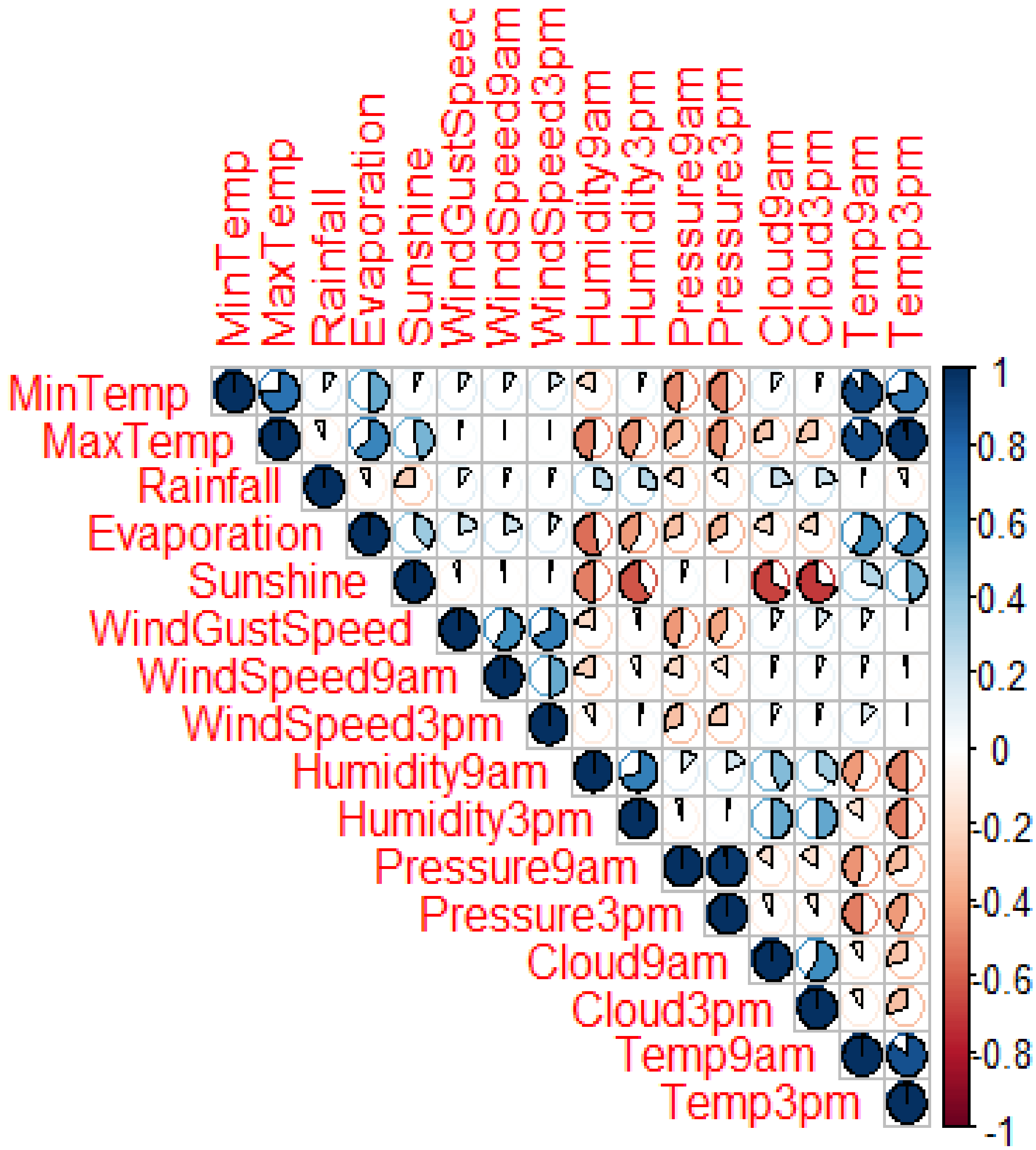



```
data$MinTemp<-as.numeric(data$MinTemp)
data$RainToday<-as.numeric(data$RainToday)
data$RainTomorrow<-as.numeric(data$RainTomorrow)

the most important variables/features for this project of predicting the

Raintomorrow variable are to be converted into numeric datatype…
```

# Correlation Plot:

For the correlation plot the unnecessary and the categorical variables are removed.

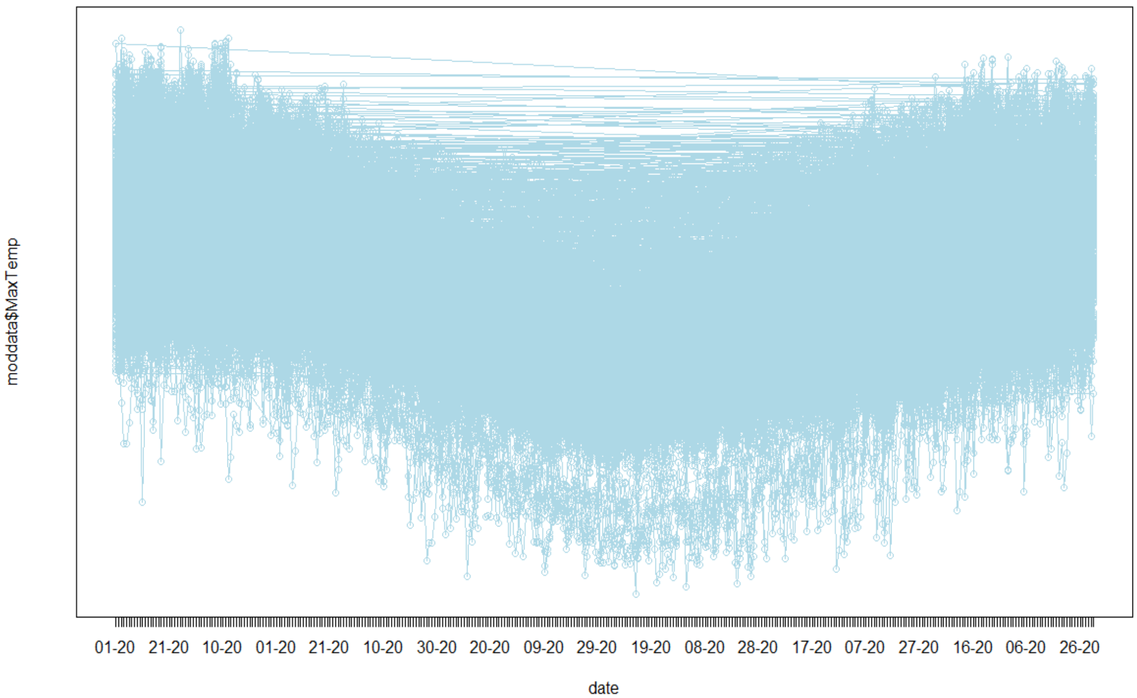
```
corrplot(cor(corrdata),
  method = "pie",
  type = "upper" # show only upper side)
```



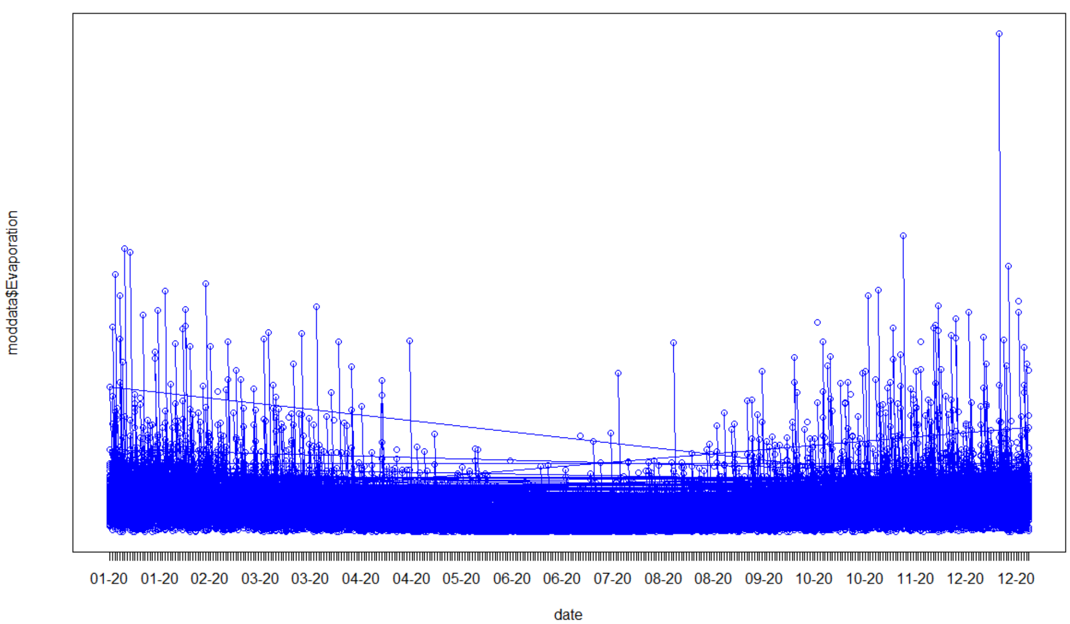

This is the correlation plot that shows the effect of variables on each other by calculating the correlation matrix. Here the unnecessary and categorical variables (date, location,RainToday, RainTomorrow) are ignored!!!

## Plots of variables over the years…

**Maximum temperatures over the years.**




**Evaporation over the time**




Here we end the exploratory data analysis part and start doing the modelling!!!

## Linear Modelling

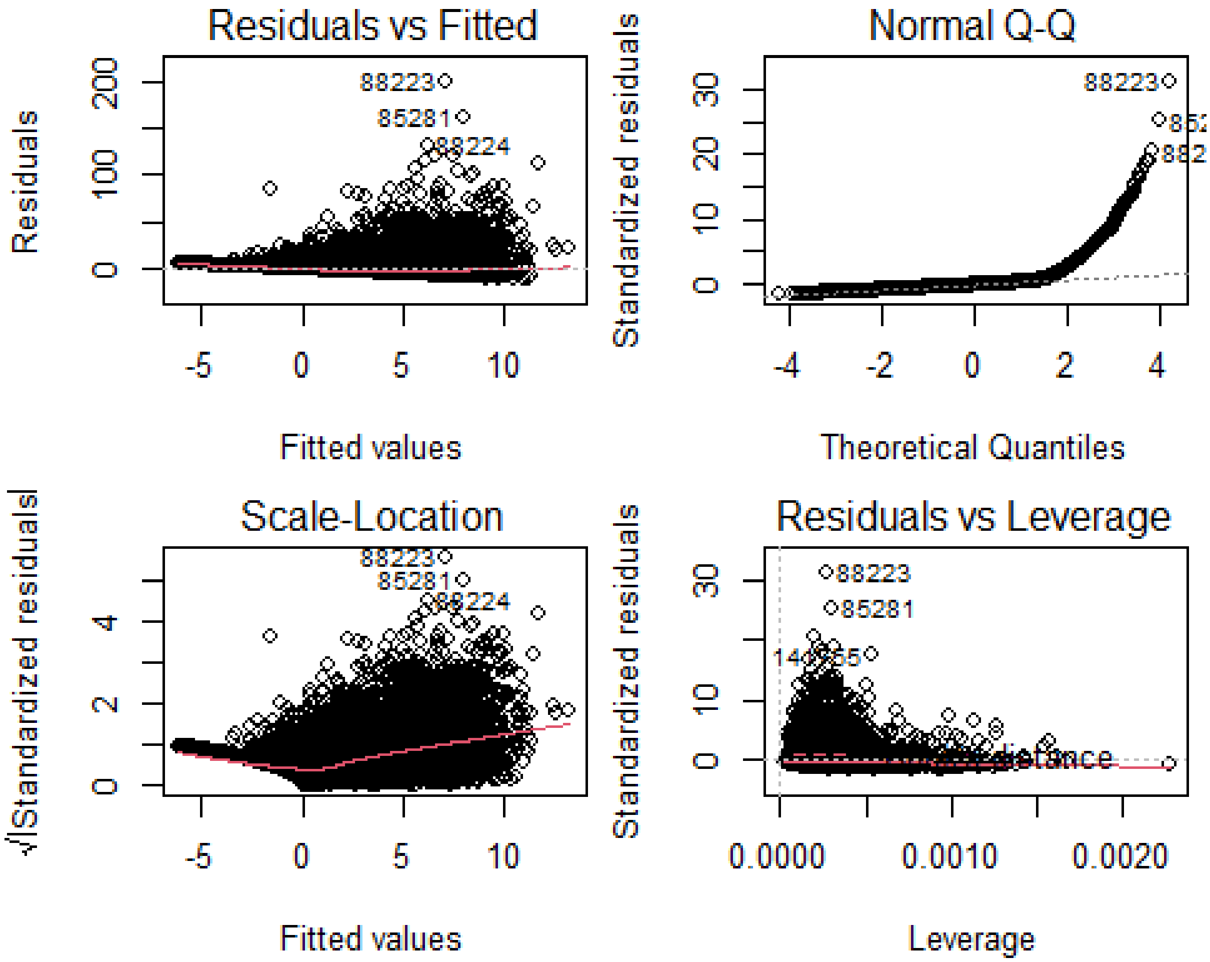
## Linear Regression…

The data now which is being used is not normalized and not scaled data from the weatherAUS dataset.

Here the dataset is splitted into two subsets as traindata and testdata with the ratios of 70% and 30% respectively.

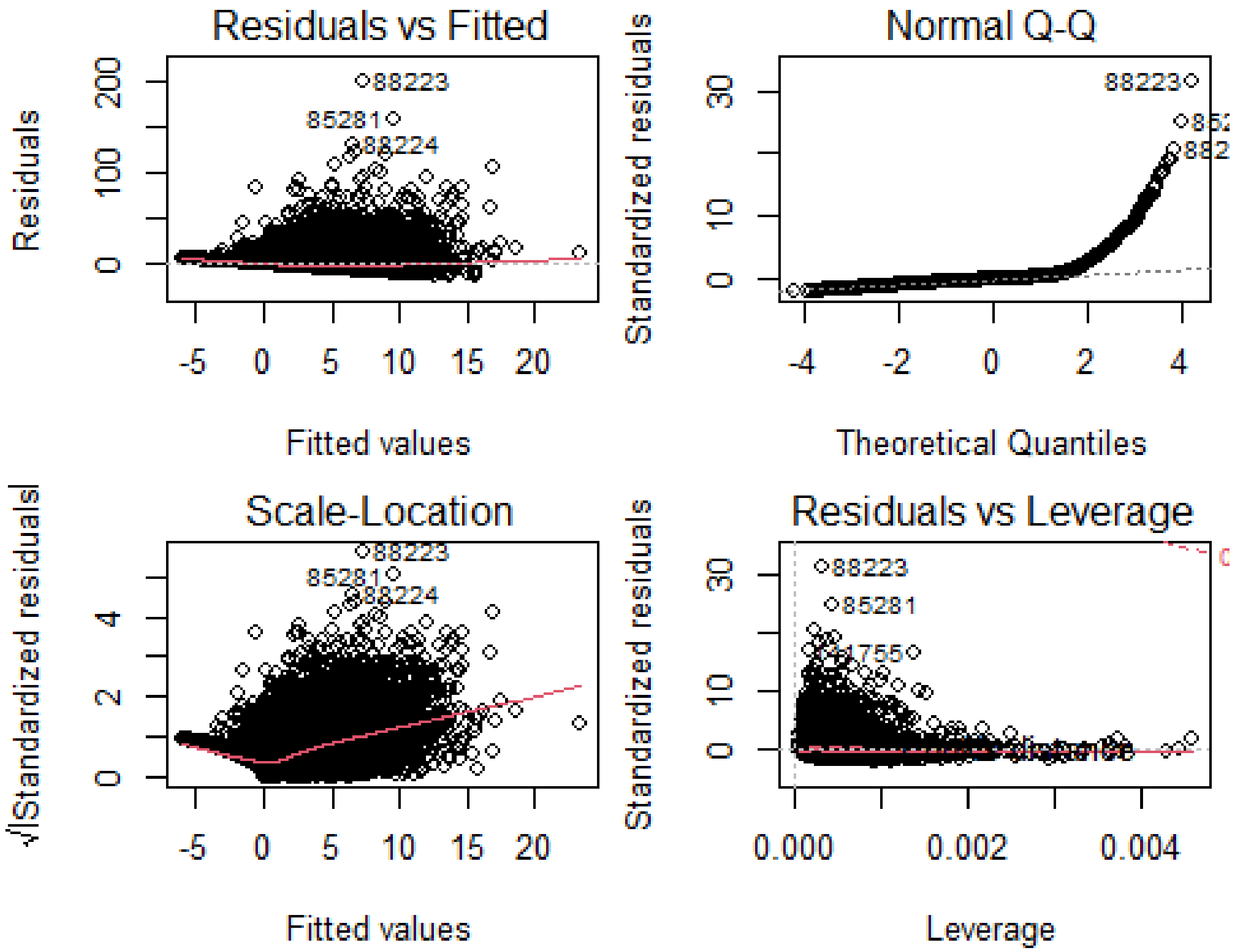
```
smp_size <- floor(0.70 * nrow(data))


traindata <-data[train_index, ]
testdata <- data[-train_index, ]
```

```
Model1: trained using the traindata.

Rainfall ~ MaxTemp+Sunshine+WindSpeed9am+Humidity9am+Humidity3pm+Pressure9am+
Pressure3pm


     adj.r.squared       sigma     AIC              BIC
          <dbl>          <dbl>   <dbl>             <dbl>
          0.142            6.    4125883          258913.

summary(model1)


## Residuals:
##      Min       1Q  Median      3Q     Max
## -10.469  -2.659  -1.000   0.772 199.073

## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    139.065905   5.774541   24.08   <2e-16 ***
## MaxTemp          0.100559   0.006604   15.23   <2e-16 ***
## Sunshine        -0.205757   0.011577  -17.77   <2e-16 ***
## WindSpeed9am     0.058312   0.004215   13.84   <2e-16 ***
## Humidity9am      0.086976   0.002614   33.27   <2e-16 ***
## Humidity3pm      0.024303   0.002537    9.58   <2e-16 ***
## Pressure9am     -0.366662   0.018780  -19.52   <2e-16 ***
## Pressure3pm      0.224044   0.019453   11.52   <2e-16 ***
## Multiple R-squared:  0.1425, Adjusted R-squared:  0.1423
## F-statistic: 937.2 on 7 and 39486 DF,  p-value: < 2.2e-16
```




Working on the data that is not normal…

```
model2 <- lm(Rainfall ~ (MaxTemp*Pressure9am)+Sunshine+WindGustSpeed+Humidity
9am+(Humidity3pm*Pressure3pm)+Temp3pm,traindata)


##   adj.r.squared sigma     AIC     BIC
##           <dbl> <dbl>   <dbl>   <dbl>
##           0.164  6.33 257846. 257949.


## Residual standard error: 6.33 on 39483 degrees of freedom
## Multiple R-squared:  0.1638, Adjusted R-squared:  0.1636
## F-statistic: 773.5 on 10 and 39483 DF,  p-value: < 2.2e-16
```



Residuals vs Fitted
Residuals
Fitted values
88223
85281
88224
Normal Q-Q
Standardized residuals
Theoretical Quantiles
88223
Scale-Location
√|Standardized residuals|
Fitted values
88223
85281
88224
Residuals vs Leverage
Standardized residuals
Leverage
88223
85281
41755

## Logistic Regression…

```
logisticmod <- glm(RainTomorrow ~ Rainfall+RainToday,data=traindata,family = 
"binomial")
summary(logisticmod)

## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.8338  -0.5733  -0.5733  -0.5733   1.9427  

## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.722640   0.015884 -108.45   <2e-16 ***
## Rainfall     0.036910   0.002253   16.39   <2e-16 ***
## RainToday    1.238666   0.033144   37.37   <2e-16 ***

##     Null deviance: 41686  on 39493  degrees of freedom
## Residual deviance: 37931  on 39491  degrees of freedom


AIC: 37937


logisticmod2 <- glm(RainTomorrow ~ Rainfall+RainToday+MinTemp+MaxTemp+Sunshin
e+WindGustSpeed+WindSpeed9am+WindSpeed3pm+Humidity9am+Humidity3pm+Pressure9am
+Pressure3pm+Cloud3pm+Temp9am+Temp3pm,data=traindata,family = "binomial")
summary(logisticmod2)


## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.2087  -0.5120  -0.2831  -0.1268   3.2131  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   57.5008364  2.9303688  19.622  < 2e-16 ***
## Rainfall       0.0115716  0.0025806   4.484 7.33e-06 ***
## RainToday      0.5000591  0.0430650  11.612  < 2e-16 ***
## MinTemp       -0.0520319  0.0085910  -6.057 1.39e-09 ***
## MaxTemp        0.0060878  0.0140848   0.432  0.66558    
## Sunshine      -0.1445786  0.0066074 -21.881  < 2e-16 ***
## WindGustSpeed  0.0577428  0.0018792  30.728  < 2e-16 ***
## WindSpeed9am  -0.0145281  0.0025222  -5.760 8.41e-09 ***
## WindSpeed3pm  -0.0260846  0.0026048 -10.014  < 2e-16 ***
## Humidity9am   -0.0003748  0.0018320  -0.205  0.83788    
## Humidity3pm    0.0570280  0.0019932  28.611  < 2e-16 ***
## Pressure9am    0.1365539  0.0095445  14.307  < 2e-16 ***
## Pressure3pm   -0.1996735  0.0096206 -20.755  < 2e-16 ***
## Cloud3pm       0.1138662  0.0095843  11.881  < 2e-16 ***
## Temp9am        0.0358274  0.0130025   2.755  0.00586 ** 
```

```
## Temp3pm        0.0111949  0.0160324   0.698  0.48501

##     Null deviance: 41686  on 39493  degrees of freedom
## Residual deviance: 26460  on 39478  degrees of freedom
## AIC: 26492
```

```r
logisticmod3 <- glm(RainTomorrow ~ Rainfall+RainToday+MinTemp+Sunshine+WindGustSpeed+WindSpeed9am+WindSpeed3pm+Humidity3pm+Pressure9am+Pressure3pm+Cloud3pm+Temp9am,data=traindata,family = "binomial")
summary(logisticmod3)
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2192  -0.5138  -0.2826  -0.1254   3.2165
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   57.985091   2.907793  19.941  < 2e-16 ***
## Rainfall       0.011852   0.002566   4.620 3.85e-06 ***
## RainToday      0.498220   0.042139  11.823  < 2e-16 ***
## MinTemp       -0.049943   0.007811  -6.394 1.62e-10 ***
## Sunshine      -0.142947   0.006529 -21.893  < 2e-16 ***
## WindGustSpeed  0.057694   0.001831  31.510  < 2e-16 ***
## WindSpeed9am  -0.014933   0.002414  -6.187 6.14e-10 ***
## WindSpeed3pm  -0.026620   0.002553 -10.427  < 2e-16 ***
## Humidity3pm    0.055124   0.001191  46.299  < 2e-16 ***
## Pressure9am    0.143024   0.008667  16.503  < 2e-16 ***
## Pressure3pm   -0.206431   0.008725 -23.659  < 2e-16 ***
## Cloud3pm       0.114420   0.009510  12.031  < 2e-16 ***
## Temp9am        0.048801   0.007874   6.198 5.73e-10 ***

##     Null deviance: 41686  on 39493  degrees of freedom
## Residual deviance: 26463  on 39481  degrees of freedom
## AIC: 26489
```

*the Accuracy:*

```
## [1] 0.07313604
```

## Confusion Matrix

```
## Confusion Matrix and Statistics

##          actual
## Predicted     0     1
##         0 13177   258
##         1 17602  8457


##                Accuracy : 0.5478
##                  95% CI : (0.5429, 0.5527)


##             Sensitivity : 0.9704
##             Specificity : 0.4281
##          Pos Pred Value : 0.3245
##          Neg Pred Value : 0.9808
##              Prevalence : 0.2207
##          Detection Rate : 0.2141
##    Detection Prevalence : 0.6598
##       Balanced Accuracy : 0.6993
##
##        'Positive' Class : 1
```

# LOGISTIC REGRESSION MODEL

## Importing the standardized data

After the dataset is normalized it is again written into another csv file named as normaldata which consists of 18 variables.

The dataset is loaded into a dataframe called Rain.

The contents in Rain data frame are splitted into the ratios of 75% and 25% as traindata and testdata respectively.

```
Rain = read.csv("normaldata.csv")
```

## Fitting Logistic Regression to the Training set:

```
RainTomorrow ~ WindGustSpeed + Humidity3pm + Pressure3pm

the formula used for the fitting the dataset.

## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5182  -0.5696  -0.3713  -0.2191   3.0641

The deviance residuals are good in this case as they are centered toward 0

(approx).

And also approximately symmetric as Min is -2.5 from 0

                                    Max is +3.0 from 0

##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.03003    0.01337 -151.87   <2e-16 ***
## WindGustSpeed  0.39822    0.01141   34.90   <2e-16 ***
## Humidity3pm    1.10273    0.01198   92.06   <2e-16 ***
## Pressure3pm   -0.47333    0.01181  -40.09   <2e-16 ***
```

all the pvalues are well below the significance level  0.05 thus the both log(odds) and log(oddratios) are both statistically significant with decent effect sizes.

```
## Null deviance: 67652  on 73899  degrees of freedom
## Residual deviance: 53351  on 73896  degrees of freedom
```

The null deviance measures the deviance using the intercept and the residual deviance measures the deviance using the independent variables.

In this case, smaller these values better the model.

```
## AIC: 53359
```

here too, smaller the AIC value better the logistic model.

## Test data Prediction using the predict function.

The predict function is used to predict the values of testdataset using the model that is trained using the traindata.

The model's true-false matrix.

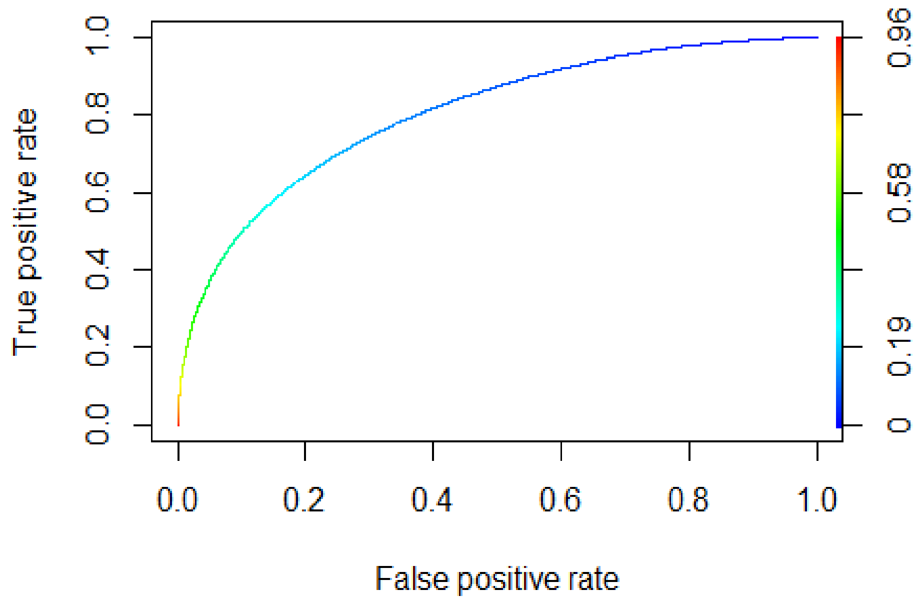
```
##     FALSE  TRUE
##   0 20310   107
##   1  3709   507
```

True positive and False positive Graph:

True positive and False positive Graph:

True positive and False positive Graph:

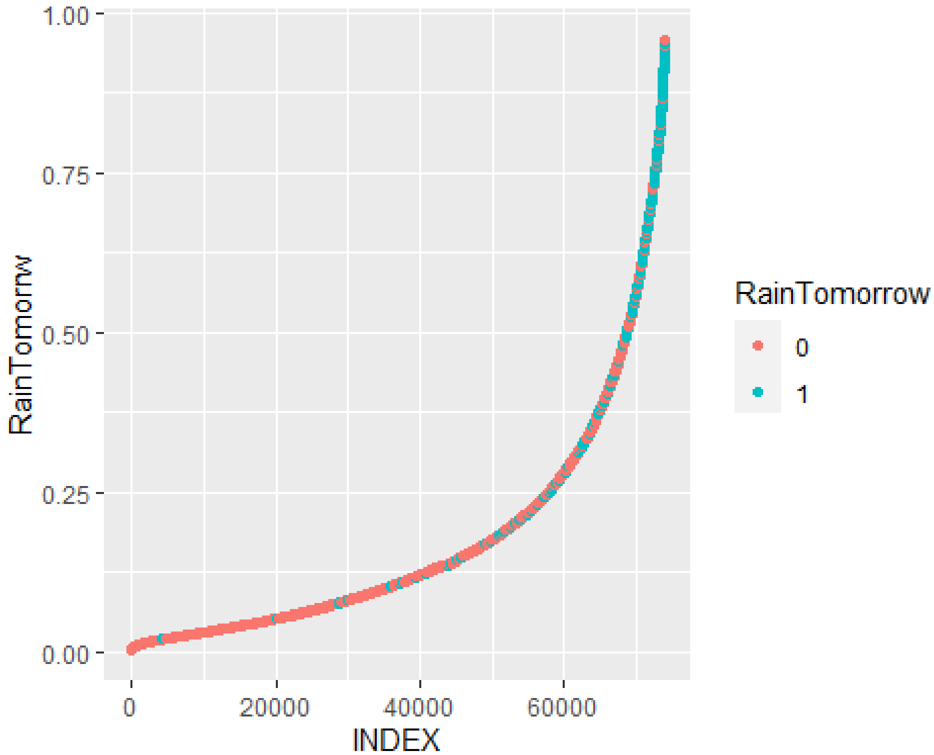
## True positive and False positive Graph:




```
Performance of the model:
## [1] 0.8030272

anova(model, test = "Chisq")

## Response variable : RainTomorrow

## Analysis of Deviance Table

## WindGustSpeed  1   1911.3     73898      65741 < 2.2e-16 ***
## Humidity3pm    1  10690.5     73897      55050 < 2.2e-16 ***
## Pressure3pm    1   1698.8     73896      53351 < 2.2e-16 ***

## Model: binomial
```

```
Predict function

predictdata<-data.frame(ProbRaintomorrow=model$fitted.values,RainTomorrow=tra
indata$RainTomorrow)
```

The graph showing the predictions done by the model for the response variable Raintomor row.

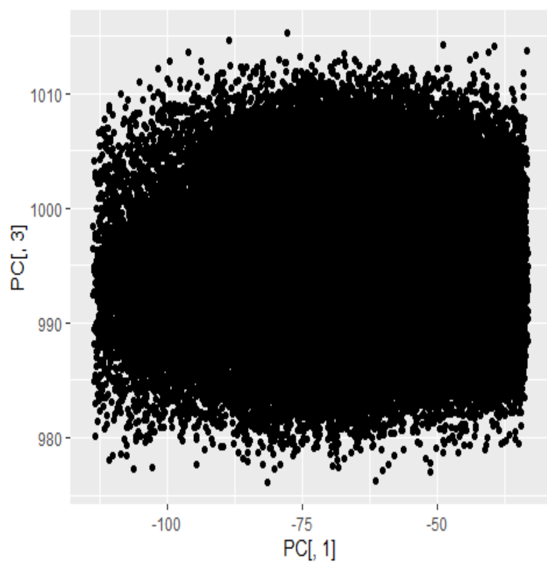


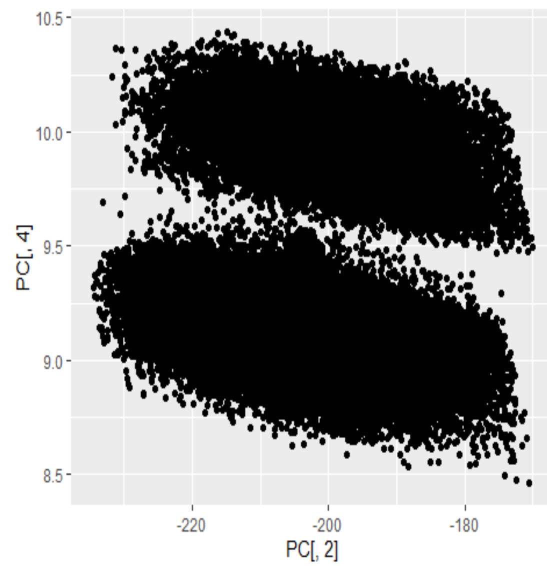

Raintomorrow is the target variable to be predicted as either 0 or 1.

```
sum(diag(tab))/sum(tab)

## [1] "accuracytest: 0.844614343707713"
```

# Principal Component Analysis

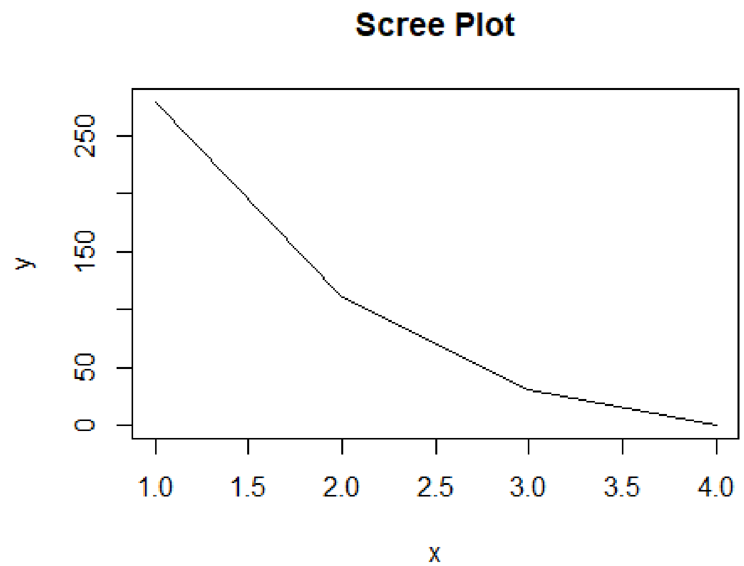
```
eigenvalues: 278.9695328 110.0631689  30.1242673   0.1124081
eigen vectors[,1]
-0.008033275  0.049846573 -0.998621870 -0.014322900
eig_val$vectors[,2]
 0.007368061  0.969554651  0.051767556 -0.239227109
eig_val$vectors[,3]
-0.008157726  0.239700164 -0.001893671  0.970810845
eig_val$vectors[,4]
 0.999907311 -0.004788342 -0.008419859  0.009568076
```



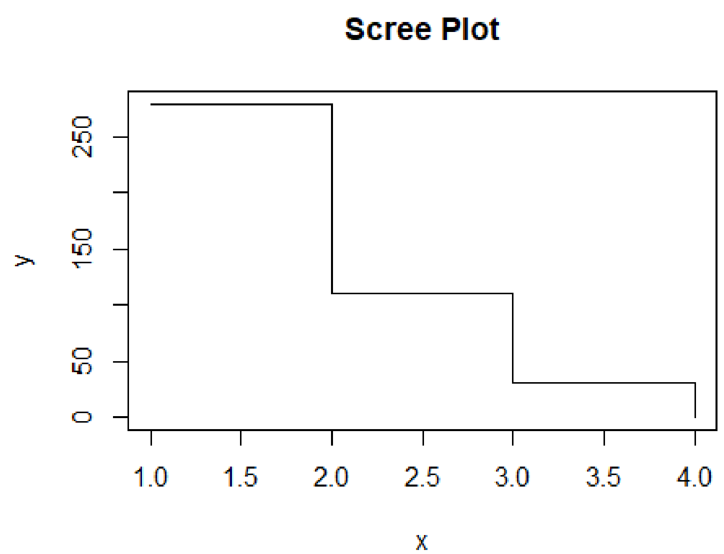

**Relation between the principal components:**

These are the scatter plots of pc1 with pc3 and Pc2 with pc4. The latter one seems to be negatively correlated. But the 1st seems to be not related to each other.




*From Principal component analysis it is Derived that the first three principal components are sufficient to estimate the data.*

<ins>Screeplot with line</ins>




<ins>Screeplot with stepfunction.</ins>

PC1 = (-0.008)RainTomorrow + (0.049)WindGustSpeed + (-0.998) Humidity3pm + (-0.014)Pressure3pm

PC2 = (0.007)RainTomorrow + (0.969)WindGustSpeed + (0.051) Humidity3pm + (-0.239)Pressure3pm

PC3 = (-0.008)RainTomorrow + (0.239)WindGustSpeed + (-0.002) Humidity3pm + (0.970)Pressure3pm

PC4 = (0.999)RainTomorrow + (-0.004)WindGustSpeed + (-0.008) Humidity3pm + (0.009)Pressure3pm

# KNN Algorithm

Knn is a machine learning algorithm used for nonparametric and supervised learning models.

Here we chose only 3 variables for modelling and training to predict the target variable

RainTomorrow.

```
subset <- dataknn[c('RainTomorrow','WindGustSpeed','Humidity3pm','Pressure3pm')]
```

later the data is normalized using the min-max function.

```
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x))) }
```

then the dataset is splitted in the ratios of 70% for training and the 30% for testing purposes.

```
nrow(train)
nrow(test)
```

70751
30322

Using the method “ Repeatedcv ” with the function traincontrol() having the parameters Number=10 and repeats=3

Train() for training the traindata set with the method KNN.

The KNN is instance based , non parametric algorithm so it will not have any explicit functionalities such as y = f(x) etc etc.

**k-fold cross Validation:**

Cross-validation can be used to estimate the test error associated with a learning method in order to evaluate its performance, or to select the appropriate level of flexibility.

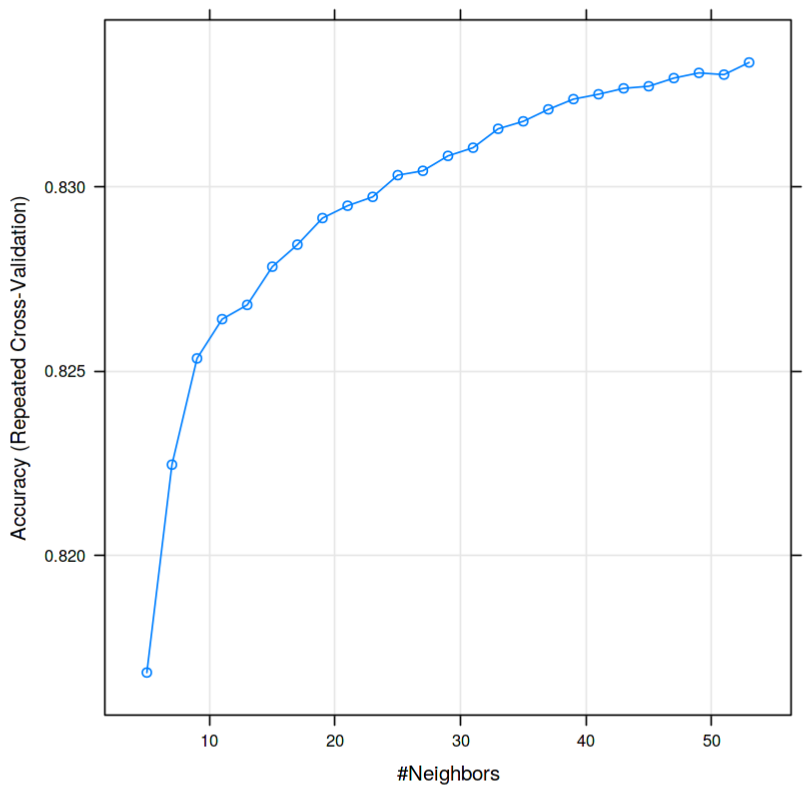
Here the k value is selected as a values which is optimal model with the large values

Here the k value is : 53

Finally the accuracy is calculated!!!

```
  k   Accuracy   Kappa
   5  0.8168270  0.3036266
   7  0.8224666  0.3112242
   9  0.8253499  0.3170634
  11  0.8264147  0.3166481
  13  0.8268010  0.3131528
  15  0.8278375  0.3148223
  17  0.8284358  0.3139846
  19  0.8291567  0.3147686
  21  0.8294912  0.3134858
  23  0.8297315  0.3128237
  25  0.8303204  0.3128565
  27  0.8304382  0.3111121
  29  0.8308433  0.3107070
  31  0.8310648  0.3093233
  33  0.8315783  0.3083491
  35  0.8317809  0.3066045
  37  0.8321060  0.3066131
  39  0.8323840  0.3056973
  41  0.8325159  0.3050143
  43  0.8326761  0.3051338
  45  0.8327326  0.3046992
  47  0.8329541  0.3045153
  49  0.8330954  0.3042392
  51  0.8330483  0.3042739
  53  0.8333781  0.3055463

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 53.
```

The Accuracy of the KNN model in predicting the target variable from k= 5 - 53.




**Confusion Matrix for the 4 classes of target variable:**

```
Confusion Matrix and Statistics

                   Reference
Prediction              0 0.2241812 0.224181218484736     1
  0                 23938       640                 7  3581
  0.2241812            18        18                 0     1
  0.224181218484736     0         0                 0     0
  1                   639        55                 0  1425
```

**Accuracy :** 0.837

**95% CI :** (0.8328, 0.8412)

**Overall statistics of the KNN model:**

| | **Class:** 0 | **Class:** 0.2241812 | **Class**: 0.224181218484736 | **Class:** 1 |
|---|---|---|---|---|
| **Sensitivity** | 0.9733 | 0.0252454 | 0.0000000 | 0.28460 |
| **Specificity** | 0.2617 | 0.9993583 | 1.0000000 | 0.97259 |
| **Pos Pred Value** | 0.8499 | 0.4864865 | NaN | 0.67249 |
| **Neg Pred Value** | 0.6953 | 0.9770513 | 0.9997691 | 0.87299 |
| **Prevalence** | 0.8111 | 0.0235143 | 0.0002309 | 0.16513 |
| **Detection Rate** | 0.7895 | 0.0005936 | 0.0000000 | 0.04700 |
| **Detection Prevalence** | 0.9289 | 0.0012202 | 0.0000000 | 0.06988 |
| **Balanced Accuracy** | 0.6175 | 0.5123019 | 0.5000000 | 0.62859 |