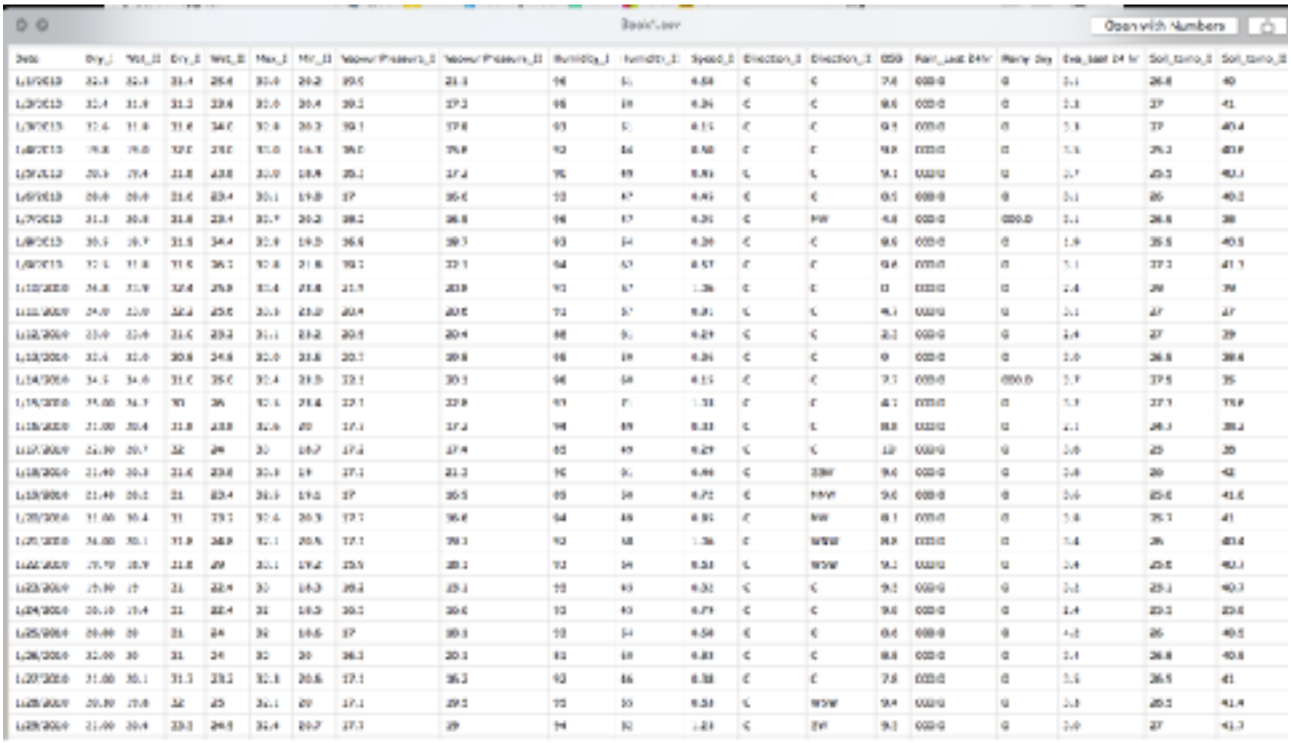
LAB EXAM: DATA ANALYSIS ON WEATHER DATA

Exploratory Data Analysis in R:

#Data



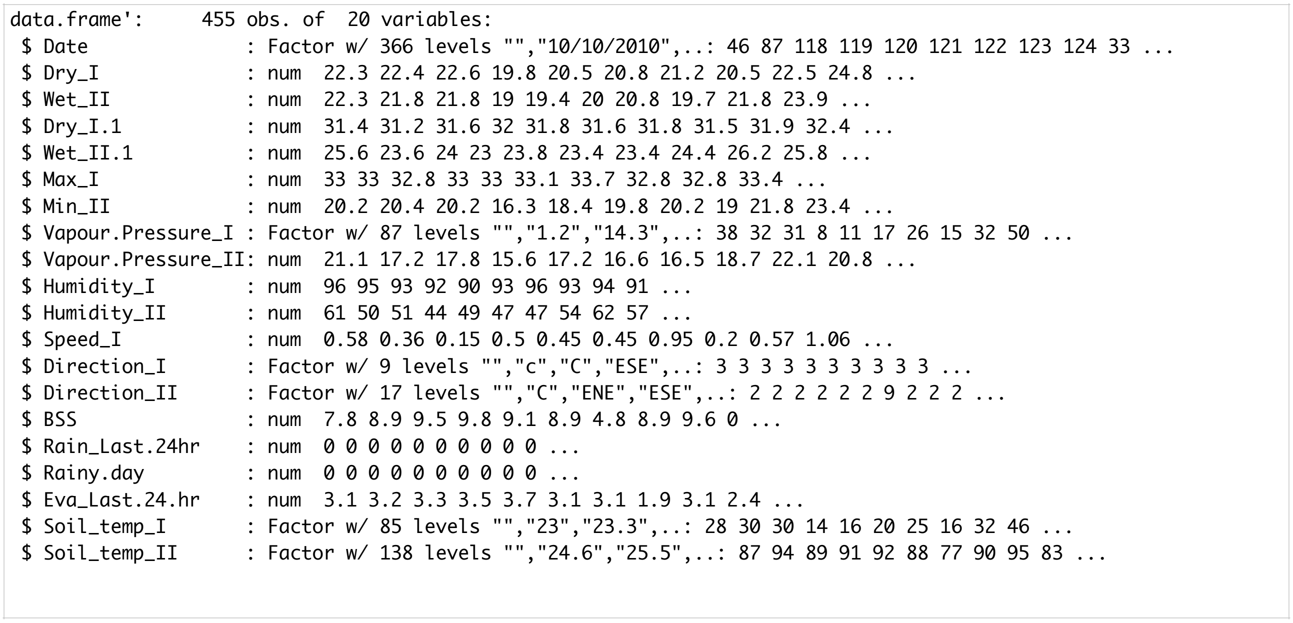
#Check dimensions (number of row & columns) in data set

```
>dim(data1)
```

```
[1] 455  20
```

#check the variables and their types in data

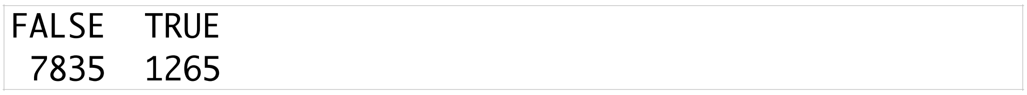
```
>str(data1)
```



```
data.frame':	455 obs. of  20 variables:
 $ Date              : Factor w/ 366 levels "","10/10/2010",..: 46 87 118 119 120 121 122 123 124 33 ...
 $ Dry_I             : num  22.3 22.4 22.6 19.8 20.5 20.8 21.2 20.5 22.5 24.8 ...
 $ Wet_II            : num  22.3 21.8 21.8 19 19.4 20 20.8 19.7 21.8 23.9 ...
 $ Dry_I.1           : num  31.4 31.2 31.6 32 31.8 31.6 31.8 31.5 31.9 32.4 ...
 $ Wet_II.1          : num  25.6 23.6 24 23 23.8 23.4 23.4 24.4 26.2 25.8 ...
 $ Max_I             : num  33 33 32.8 33 33 33.1 33.7 32.8 32.8 33.4 ...
 $ Min_II            : num  20.2 20.4 20.2 16.3 18.4 19.8 20.2 19 21.8 23.4 ...
 $ Vapour.Pressure_I : Factor w/ 87 levels "","1.2","14.3",..: 38 32 31 8 11 17 26 15 32 50 ...
 $ Vapour.Pressure_II: num  21.1 17.2 17.8 15.6 17.2 16.6 16.5 18.7 22.1 20.8 ...
 $ Humidity_I        : num  96 95 93 92 90 93 96 93 94 91 ...
 $ Humidity_II       : num  61 50 51 44 49 47 47 54 62 57 ...
 $ Speed_I           : num  0.58 0.36 0.15 0.5 0.45 0.45 0.95 0.2 0.57 1.06 ...
 $ Direction_I       : Factor w/ 9 levels "","c","C","ESE",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Direction_II      : Factor w/ 17 levels "","C","ENE","ESE",..: 2 2 2 2 2 2 9 2 2 2 ...
 $ BSS               : num  7.8 8.9 9.5 9.8 9.1 8.9 4.8 8.9 9.6 0 ...
 $ Rain_Last.24hr    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Rainy.day         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Eva_Last.24.hr    : num  3.1 3.2 3.3 3.5 3.7 3.1 3.1 1.9 3.1 2.4 ...
 $ Soil_temp_I       : Factor w/ 85 levels "","23","23.3",..: 28 30 30 14 16 20 25 16 32 46 ...
 $ Soil_temp_II      : Factor w/ 138 levels "","24.6","25.5",..: 87 94 89 91 92 88 77 90 95 83 ...
```

#check if this data has missing values

```
>table(is.na(data1))
```



```
FALSE  TRUE
 7835  1265
```

```
> pie(colSums(is.na(data1)))
```

#Column wise finding out the missing value

```
>colSums(is.na(data1))

           Date             Dry_I
             0               90

                 Wet_II          Dry_I.1
             90               90
        Wet_II.1            Max_I
             90               90
          Min_II  Vapour.Pressure_I
             90                0
Vapour.Pressure_II        Humidity_I
             90               90
      Humidity_II           Speed_I
             90               93
      Direction_I       Direction_II
             0                0
            BSS      Rain_Last.24hr
             91               90
        Rainy.day      Eva_Last.24.hr
             90               91
      Soil_temp_I       Soil_temp_II
             0                0
```

#Full summary of each column in the dataset

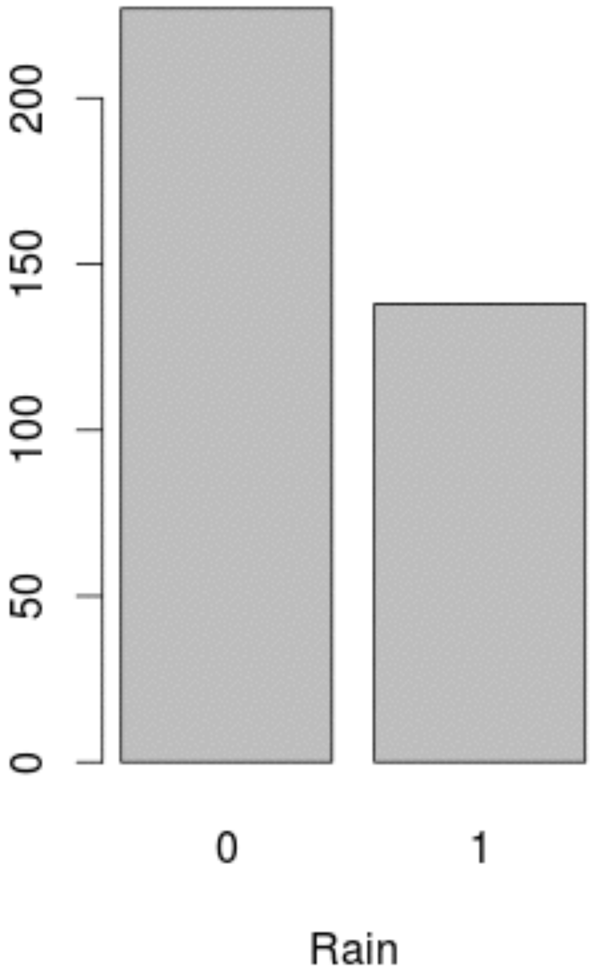
```
> summary(train)

       Date          Dry_I           Wet_II
          : 90   Min.   :19.50   Min.   :18.00
 10/10/2010:  1   1st Qu.:23.50   1st Qu.:23.00
 10/11/2010:  1   Median :24.30   Median :23.80
 10/1/2010 :  1   Mean   :24.59   Mean   :23.54
 10/12/2010:  1   3rd Qu.:26.00   3rd Qu.:24.50
 10/13/2010:  1   Max.   :30.20   Max.   :28.00
 (Other)   :360   NA's   :90      NA's   :90


    Dry_I.1         Wet_II.1          Max_I
 Min.   :23.00   Min.   :22.40   Min.   :25.20
 1st Qu.:28.90   1st Qu.:24.80   1st Qu.:30.20
 Median :30.50   Median :25.50   Median :31.80
 Mean   :30.24   Mean   :25.69   Mean   :31.71
 3rd Qu.:32.00   3rd Qu.:26.50   3rd Qu.:33.50
 Max.   :35.00   Max.   :35.00   Max.   :36.20  
 NA's   :90      NA's   :90      NA's   :90     


     Min_II      Vapour.Pressure_I
 Min.   :16.30          : 90      
 1st Qu.:22.50   22.4   : 18      
 Median :23.30   22     : 17      
 Mean   :23.17   21.9   : 15      
 3rd Qu.:24.10   21.7   : 14      
 Max.   :29.20   22.1   : 14      
 NA's   :90      (Other):287      


 Vapour.Pressure_II   Humidity_I    
 Min.   :15.10      Min.   : 56.00  
 1st Qu.:20.60      1st Qu.: 87.00  
 Median :22.10      Median : 93.00  
 Mean   :21.93      Mean   : 91.43  
 3rd Qu.:23.30      3rd Qu.: 96.00  
 Max.   :29.50      Max.   :100.00  
 NA's   :90         NA's   :90      




  Humidity_II        Speed_I       Direction_I
 Min.   : 44.00   Min.   :0.010   C      :353  
 1st Qu.: 60.00   1st Qu.:0.740          : 92  
 Median : 67.00   Median :1.090   c      :  2  
 Mean   : 70.54   Mean   :1.275   ESE    :  2  
 3rd Qu.: 76.00   3rd Qu.:1.800   SSW    :  2  
 Max.   :663.00   Max.   :3.800   NNE    :  1  
 NA's   :90       NA's   :93      (Other):  3  


  Direction_II      BSS          Rain_Last.24hr  
 C      :164   Min.   : 0.000   Min.   :  0.00  
        : 95   1st Qu.: 1.975   1st Qu.:  0.00  
 SW     : 47   Median : 6.050   Median :  0.00  
 NW     : 42   Mean   : 5.311   Mean   : 11.57  
 WSW    : 29   3rd Qu.: 8.600   3rd Qu.:  9.60  
 SSW    : 27   Max.   :11.300   Max.   :173.00  
 (Other): 51   NA's   :91       NA's   :90      


   Rainy.day       Eva_Last.24.hr  Soil_temp_I
 Min.   :0.0000   Min.   :0.200          : 90  
 1st Qu.:0.0000   1st Qu.:2.400   27     : 32  
 Median :0.0000   Median :3.100   26     : 28  
 Mean   :0.3781   Mean   :3.287   25     : 23  
 3rd Qu.:1.0000   3rd Qu.:4.200   26.5   : 23  
 Max.   :1.0000   Max.   :9.300   25.5   : 22  
 NA's   :90       NA's   :91      (Other):237  
 Soil_temp_II
       : 90
35     : 13
47     : 12
34     : 11
36.5   : 11
36     :  9
(Other):309
```

```
#Dropping missing values

>data1=na.omit(data1)


#Summary after removing NA values.

>summary(data1)
```

```
       Date         Dry_I           Wet_II
10/10/2010:  1   Min.   :19.50   Min.   :18.00
10/11/2010:  1   1st Qu.:23.50   1st Qu.:22.98
10/1/2010 :  1   Median :24.30   Median :23.80
10/12/2010:  1   Mean   :24.59   Mean   :23.54
10/13/2010:  1   3rd Qu.:26.00   3rd Qu.:24.50
10/14/2010:  1   Max.   :30.20   Max.   :28.00
(Other)   :354
   Dry_I.1         Wet_II.1         Max_I
Min.   :23.00   Min.   :22.40   Min.   :25.2
1st Qu.:28.90   1st Qu.:24.80   1st Qu.:30.1
Median :30.50   Median :25.50   Median :31.8
Mean   :30.24   Mean   :25.69   Mean   :31.7
3rd Qu.:32.00   3rd Qu.:26.50   3rd Qu.:33.5
Max.   :35.00   Max.   :35.00   Max.   :36.2

    Min_II      Vapour.Pressure_I
Min.   :16.30   22     : 17
1st Qu.:22.50   22.4   : 17
Median :23.30   21.9   : 15
Mean   :23.16   21.7   : 14
3rd Qu.:24.12   22.1   : 14
Max.   :29.20   20.8   : 13
                (Other):270
Vapour.Pressure_II   Humidity_I
Min.   :15.10      Min.   : 56.00
1st Qu.:20.60      1st Qu.: 87.00
Median :22.10      Median : 93.00
Mean   :21.93      Mean   : 91.43
3rd Qu.:23.50      3rd Qu.: 96.00
Max.   :29.50      Max.   :100.00

 Humidity_II       Speed_I         Direction_I
Min.   : 44.00   Min.   :0.0100   C      :350
1st Qu.: 60.00   1st Qu.:0.7375   c      :  2
Median : 67.00   Median :1.0750   ESE    :  2
```

```
 Mean   : 70.57   Mean   :1.2716   SSW     :  2  
 3rd Qu.: 76.00   3rd Qu.:1.8000   NNE     :  1  
 Max.   :663.00   Max.   :3.8000   SE      :  1  
                                   (Other):  2  
  Direction_II      BSS          Rain_Last.24hr  
 C      :163   Min.   : 0.000   Min.   :  0.00  
 SW     : 46   1st Qu.: 1.875   1st Qu.:  0.00  
 NW     : 41   Median : 6.000   Median :  0.00  
 WSW    : 29   Mean   : 5.294   Mean   : 11.66  
 SSW    : 27   3rd Qu.: 8.600   3rd Qu.:  9.95  
 WNW    : 18   Max.   :11.300   Max.   :173.00  
 (Other): 36                                    
   Rainy.day      Eva_Last.24.hr   Soil_temp_I 
 Min.   :0.0000   Min.   :0.200   27     : 31  
 1st Qu.:0.0000   1st Qu.:2.400   26     : 28  
 Median :0.0000   Median :3.100   25     : 23  
 Mean   :0.3778   Mean   :3.283   26.5   : 23  
 3rd Qu.:1.0000   3rd Qu.:4.200   25.5   : 22  
 Max.   :1.0000   Max.   :9.300   28     : 10  
                                  (Other):223  
  Soil_temp_II
 35     : 13  
 47     : 12  
 34     : 11  
 36.5   : 11  
 36     :  9  
 37     :  9  
 (Other):295  
```

```
#Checking if the feature is categorical or not

>length(unique(data1$Rainy.day))

[1] 3

> length(unique(data1$Direction_I))
[1] 8

> length(unique(data1$Direction_II))
[1] 17


#Plotting the categorical variable

>counts<- table(data1$Rainy.day)
>barplot(counts, xlab = 'Rain')
```



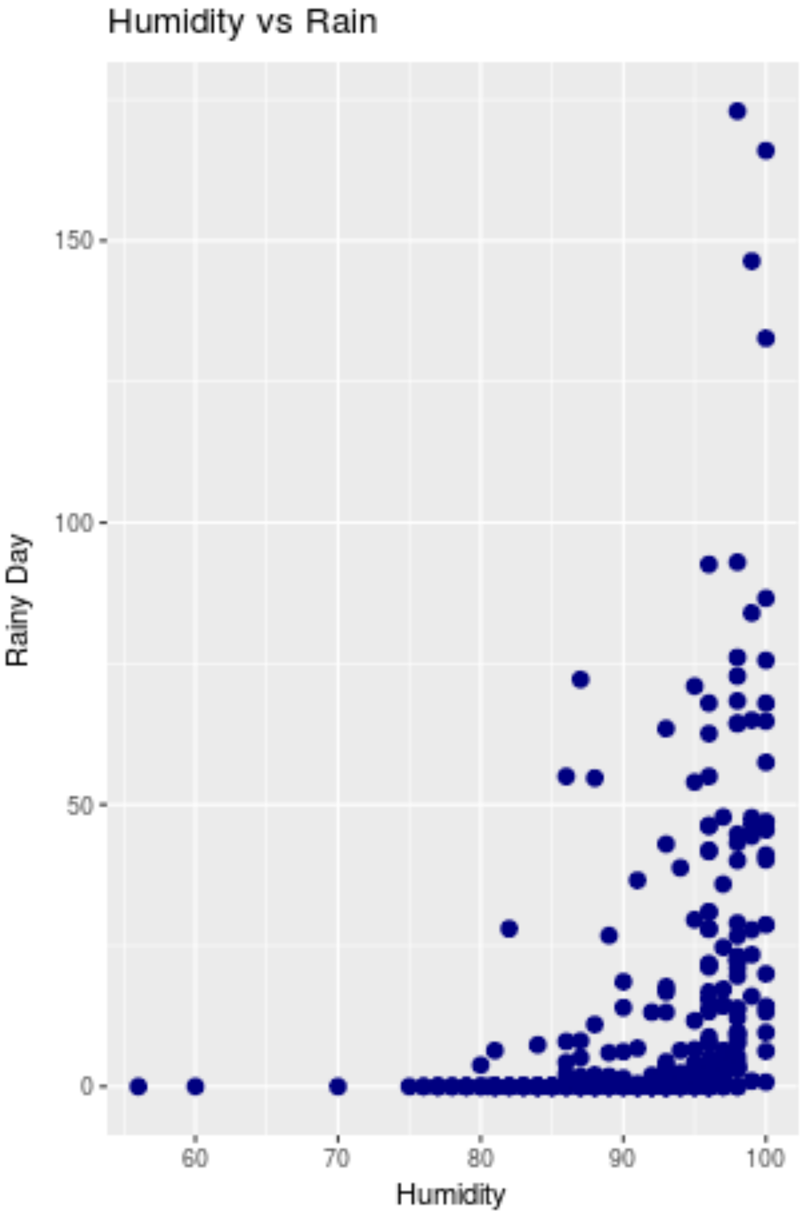
```
#Finding Correlation between feature columns

> cor(data1$Rain_Last.24hr,data1$BSS)
[1] -0.4707937
> cor(data1$Rain_Last.24hr,data1$Humidity_I)
[1] 0.3872418
> cor(data1$Rain_Last.24hr,data1$Humidity_II)
[1] 0.193109
```

Some of the inferences drawn from variables in the data set:

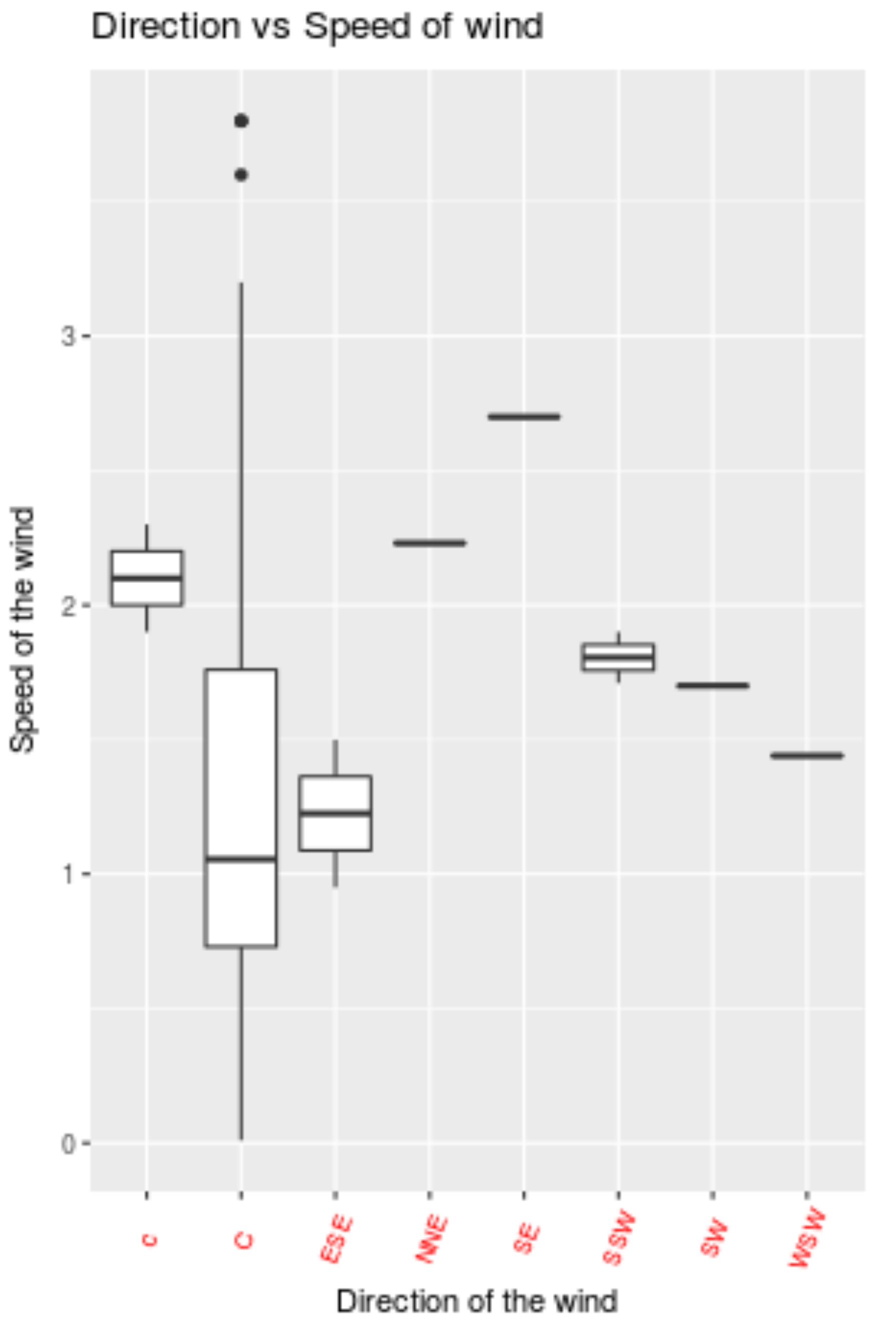
1. There are total 90 days when the observations weren’t taken for any factor, this could be due to failure of instruments/shutdown.
2. BSS Column has the value regarding best sunshine, the min value is 0 which is highly impossible(there goes no day without sunshine), hence the 0 value’s are the missing data or the instrument is less accurate.
3. Rainy.day column is a categorical variable, where 0 means no rain and 1 means rain. Also it rained approximately 30% of the year.
4. The occurrence of rain and humidity are highly correlated.

#Visualising Data

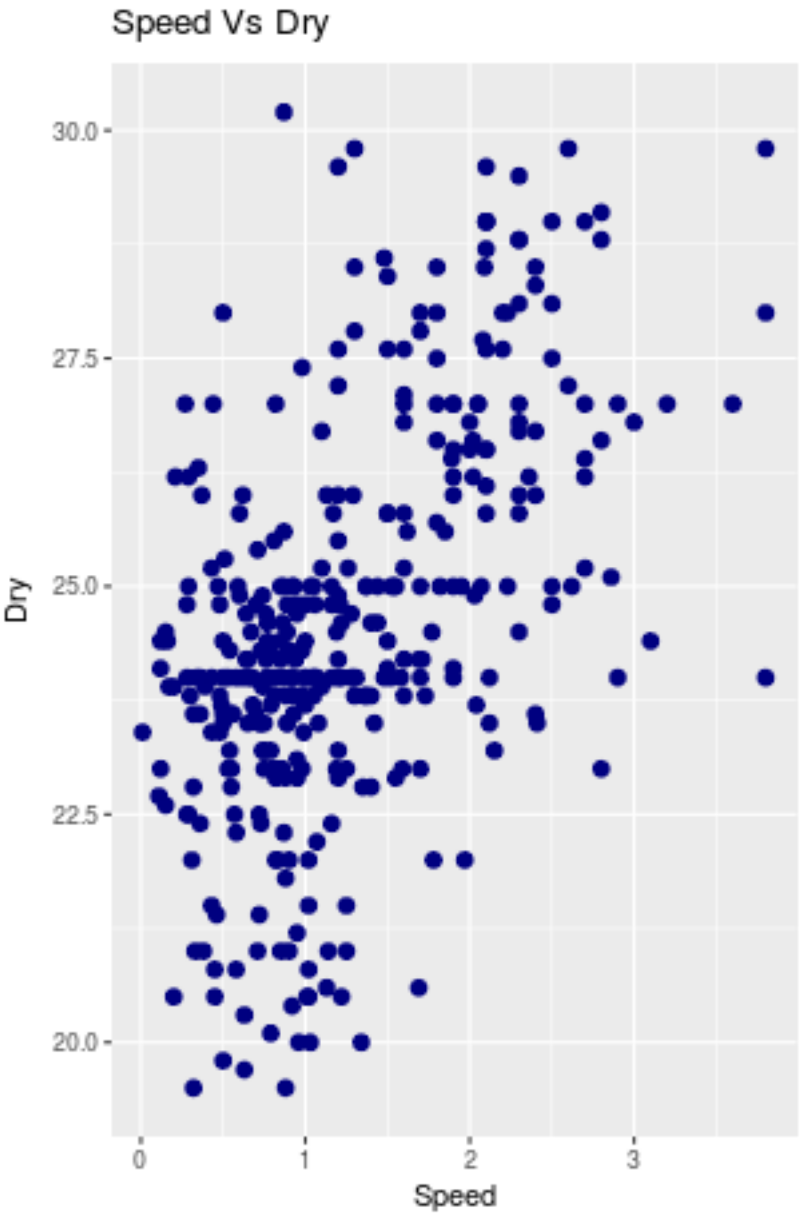
```
> ggplot(data1, aes(x= data1$Humidity_I, y = data1$Rain_Last.24hr)) + geom_point(size
= 2.5, color="navy") + xlab("Humidity I") + ylab("Rainy Day") + ggtitle("Humidity vs Rain")
```



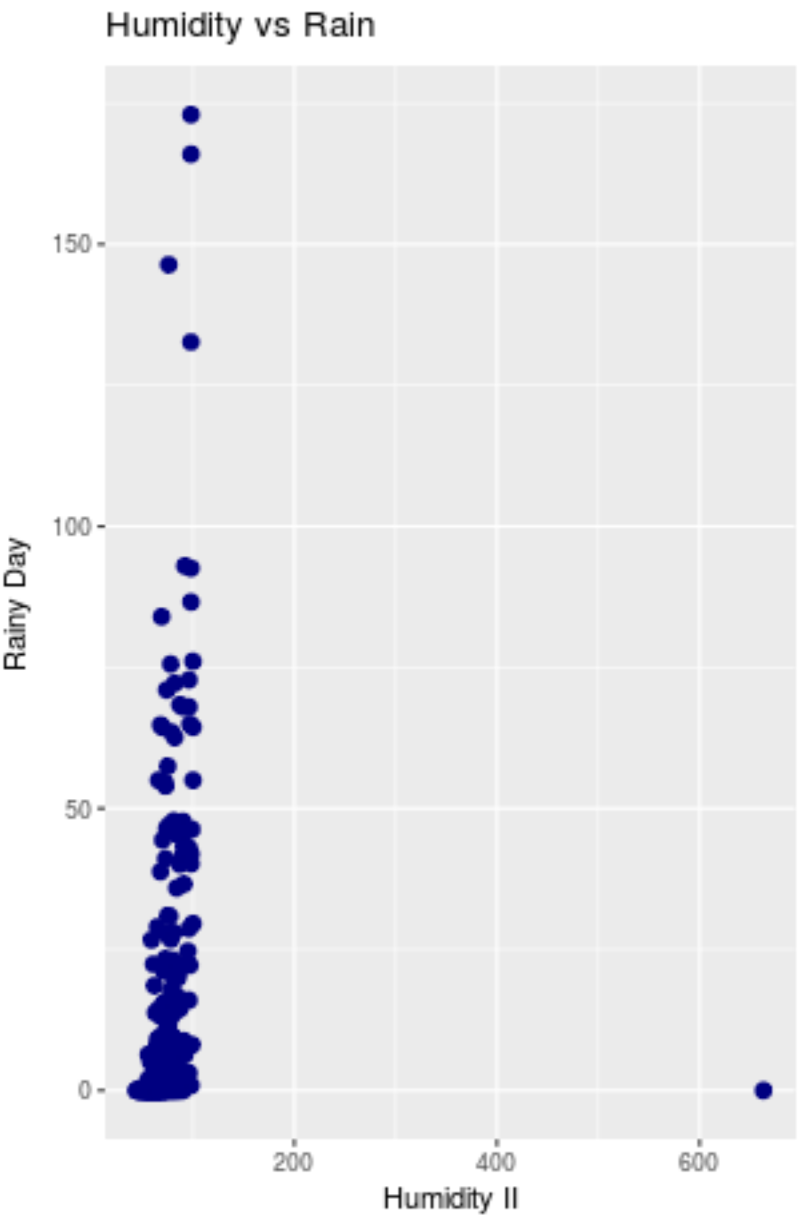
```
> ggplot(data1, aes(data1$Direction_I, data1$Speed_I))
+geom_boxplot() +ggtitle("Box Plot") + theme(axis.text.x =
element_text(angle = 70, vjust = 0.5, color = "red")) +
xlab("Direction of the wind") + ylab("Speed of the wind") +
ggtitle("Direction vs Speed of wind")
```



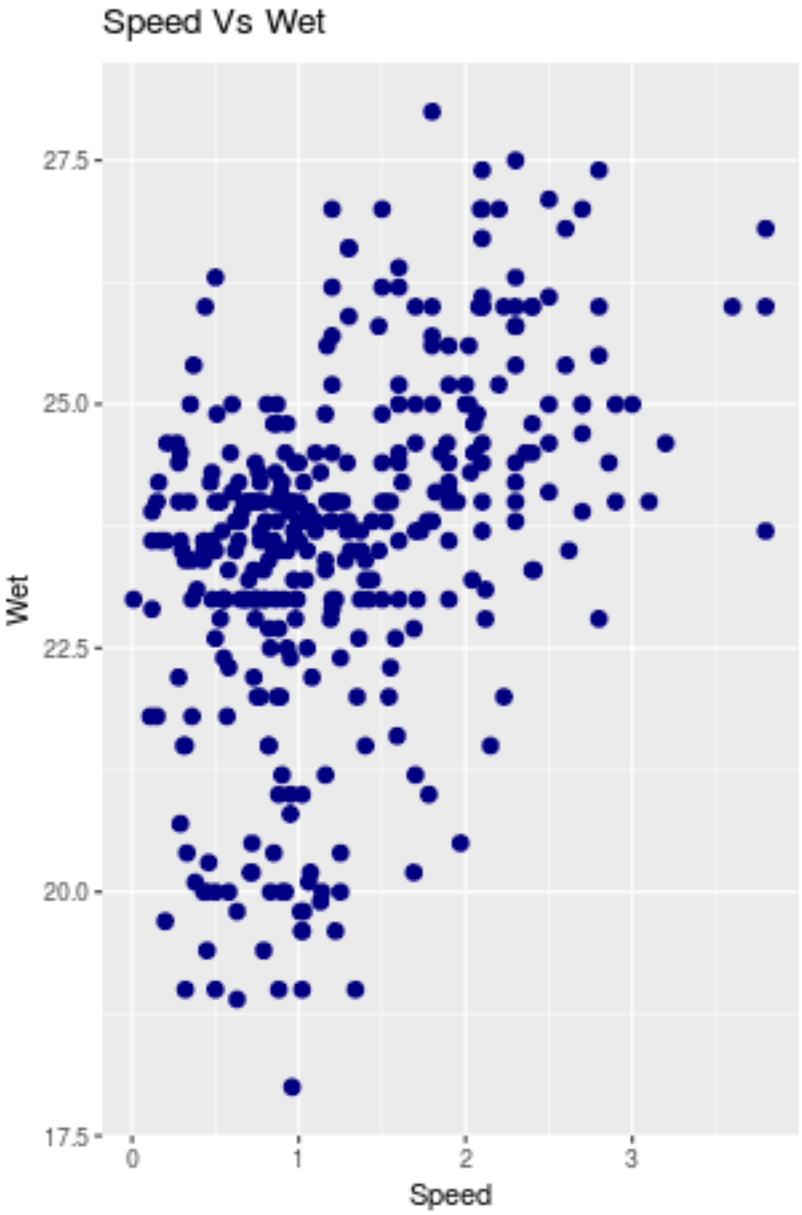
```
>ggplot(data1, aes(x= data1$Speed_I, y = data1$Dry_I)) + geom_point(size = 2.5,
color="navy") + xlab("Speed") + ylab("Dry") + ggtitle("Speed Vs Dry")
```



```
> ggplot(data1, aes(x= data1$Humidity_II, y = data1$Rain_Last.24hr)) + geom_point(size
= 2.5, color="navy") + xlab("Humidity II") + ylab("Rainy Day") + ggtitle("Humidity vs Rain”)
```



```
>ggplot(data1, aes(x= data1$Speed_I, y = data1$Wet_II)) + geom_point(size = 2.5,
color="navy") + xlab("Speed") + ylab("Wet") + ggtitle("Speed Vs Wet")
```



Some of the inferences drawn from the visualisation:

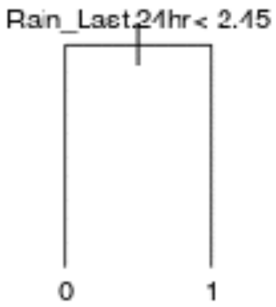
1. From first 2 plots, it is inferred, If 75<Humidity<100 in last 24 hours then it rains.
2. From Box plot, it is inferred, Wind has variable speed in C direction, constant speed in NNE, SE, SW, WSW direction.
3. From last two plots, it is inferred that there is not much dependency between type of wind(dry or wet) and speed of the wind.

PREDICTIVE ANALYSIS IN R:

```
#Using Decision tree classifier to classify if It will rain or not on basis of all other weather factors.

library(rpart)
library(caret)
data1=read.csv(file.choose(),header = T)
data1=na.omit(data1)

#split data into test train
data<-data1
dt<-sort(sample(nrow(data),nrow(data)*0.8))
train<-data[dt,]
test<-data[-dt,]
trainx<-subset(train, select = -data$Rainy.day)
trainy<- train$Rainy.day
testx<- subset(test, select = -data$Rainy.day)
testy<- test$Rainy.day


#fit Decision Tree
dtreeClass<-rpart(trainy~.,data=trainx,method="class")


#predict using fit model
pred<-predict(dtreeClass,testx,type = "class")


#show required parameters
xtab<-table(pred,testy)

#confusionMatrix(xtab)
confusionMatrix(xtab)
```

```
Confusion Matrix and Statistics

    testy
pred  0  1
   0 39  2
   1  0 31

               Accuracy : 0.9722
```

```
                 95% CI : (0.9032, 0.9966)
    No Information Rate : 0.5417
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9438
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9394
         Pos Pred Value : 0.9512
         Neg Pred Value : 1.0000
             Prevalence : 0.5417
         Detection Rate : 0.5417
   Detection Prevalence : 0.5694
      Balanced Accuracy : 0.9697

       'Positive' Class : 0

>plot(dtreeClass,branch = 1,uniform = true(),margin = 1)

> text(dtreeClass,cex=.7)
```



```
> print(dtreeClass)
n= 288 

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 288 103 0 (0.6423611 0.3576389)  
  2) Rain_Last.24hr< 2.55 185   0 0 (1.0000000 0.0000000) *
  3) Rain_Last.24hr>=2.55 103   0 1 (0.0000000 1.0000000) *
```