

# LAB EXAM: DATA ANALYSIS ON WEATHER DATA

## Exploratory Data Analysis in R:

#Data

Basic.csv																	Open with Numbers		
Date	Dry_1	Wet_1	Dry_2	Wet_2	Max_1	Min_1	Mean Pressure_1	Mean Pressure_2	Humidity_1	Humidity_2	Speed_1	Direction_1	Direction_2	OSD	Rain_Last 24hr	Rainy Day	Eva_Last 24 hr	Soil_temo_1	Soil_temo_2
1/1/2010	32.8	32.8	31.4	26.6	30.0	20.2	99.5	21.3	96	51	6.58	C	C	7.6	000-0	0	1.1	26.6	40
1/2/2010	32.4	31.8	31.2	23.6	30.0	20.4	99.2	27.2	96	18	6.36	C	C	9.6	000-0	0	1.3	27	41
1/3/2010	32.4	31.8	31.6	34.0	30.8	20.2	99.1	37.8	93	51	6.15	C	C	9.1	000-0	0	1.3	27	40.4
1/4/2010	29.8	29.0	32.0	23.0	30.0	16.3	96.0	36.8	92	64	6.50	C	C	9.9	000-0	0	1.5	26.2	40.8
1/5/2010	20.9	27.4	21.8	43.8	30.0	18.6	95.1	17.4	95	89	6.83	L	L	16.1	000-0	0	2.7	40.2	40.1
1/6/2010	20.6	20.6	21.6	23.4	30.1	19.0	97	35.6	92	47	6.45	C	C	9.5	000-0	0	2.1	26	40.2
1/7/2010	21.3	30.8	31.8	23.4	30.7	20.2	98.2	36.8	96	17	6.20	C	FW	4.8	000-0	000-0	1.1	26.8	38
1/8/2010	30.5	26.7	31.5	34.4	30.8	19.0	96.6	39.3	93	14	6.20	C	C	9.6	000-0	0	1.9	35.5	40.5
1/9/2010	32.1	31.8	31.6	36.2	30.8	21.8	99.1	27.1	94	53	6.57	C	C	9.8	000-0	0	1.1	27.2	41.1
1/10/2010	24.8	30.9	32.6	25.8	30.4	21.8	21.3	40.8	92	67	1.36	C	C	0	000-0	0	2.4	29	39
1/11/2010	24.0	22.0	32.4	45.0	30.5	23.0	40.4	40.0	72	57	6.81	L	L	4.1	000-0	0	2.1	27	27
1/12/2010	23.0	23.0	31.6	29.2	31.1	23.0	20.9	20.4	88	91	6.29	C	C	2.2	000-0	0	1.9	27	39
1/13/2010	32.6	32.0	30.8	24.8	30.0	23.8	20.2	39.8	98	18	6.36	C	C	9	000-0	0	2.0	26.8	38.6
1/14/2010	34.5	34.0	31.0	35.0	30.4	23.0	22.1	30.3	96	64	6.15	C	C	7.1	000-0	000-0	1.7	27.5	35
1/15/2010	33.00	34.7	30	36	30.1	21.4	22.1	22.8	89	71	1.31	C	C	4.1	000-0	0	1.3	27.1	35.8
1/16/2010	21.00	30.4	21.8	43.8	30.6	20	17.1	17.4	98	89	6.33	L	L	18.8	000-0	0	2.1	26.1	38.4
1/17/2010	22.00	20.7	32	29	30	18.7	17.4	17.4	82	89	6.29	C	C	12	000-0	0	2.0	25	20
1/18/2010	21.40	20.3	31.6	23.6	30.3	19	17.1	21.2	95	91	6.86	C	SSW	9.4	000-0	0	2.8	29	42
1/19/2010	21.40	20.2	21	23.4	30.0	19.0	17	30.5	89	18	6.72	C	FW	9.6	000-0	0	2.0	25.6	41.6
1/20/2010	31.00	30.4	31	33.2	30.4	20.3	17.1	36.6	94	48	6.91	C	SW	8.1	000-0	0	1.8	35.1	41
1/21/2010	34.00	30.1	31.8	26.8	30.1	20.6	17.1	39.2	92	68	1.36	C	WSW	18.8	000-0	0	1.4	26	40.4
1/22/2010	27.70	26.7	21.8	49	30.1	19.2	23.5	38.1	92	14	6.53	L	WSW	16.1	000-0	0	2.4	40.8	40.1
1/23/2010	27.90	27	21	22.4	30	18.3	19.4	43.1	92	43	6.30	C	C	9.2	000-0	0	2.0	29.1	40.3
1/24/2010	20.10	27.4	21	22.4	31	18.3	20.2	39.6	92	43	6.79	C	C	9.6	000-0	0	1.4	29.2	23.6
1/25/2010	20.40	20	21	24	30	18.6	17	30.1	92	14	6.56	C	C	9.4	000-0	0	1.0	26	40.2
1/26/2010	32.00	30	31	26	30	20	36.3	20.1	82	68	6.83	C	C	8.8	000-0	0	1.1	26.8	40.8
1/27/2010	31.00	30.1	31.3	23.2	30.8	20.6	17.1	36.2	92	66	6.38	C	C	7.8	000-0	0	1.5	26.5	41
1/28/2010	20.90	27.8	32	45	34.1	20	17.1	29.2	92	50	6.53	L	WSW	16.4	000-0	0	2.3	40.2	41.4
1/29/2010	21.00	20.4	33.0	24.0	34.4	20.7	17.1	29	94	92	1.23	C	FW	9.2	000-0	0	2.0	27	41.3

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

```
>dim(data1)
```

```
[1] 455 20
```

#check the variables and their types in data

```
>str(data1)
```

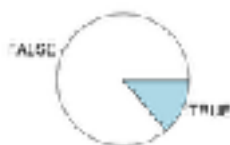
```
data.frame':    455 obs. of  20 variables:
 $ Date          : Factor w/ 366 levels "", "10/10/2010",...: 46 87 118 119 120 121 122 123 124 33 ...
 $ Dry_I         : num  22.3 22.4 22.6 19.8 20.5 20.8 21.2 20.5 22.5 24.8 ...
 $ Wet_II        : num  22.3 21.8 21.8 19 19.4 20 20.8 19.7 21.8 23.9 ...
 $ 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 ...
 $ Direction_II   : Factor w/ 17 levels "", "C", "ENE", "ESE",...: 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
```

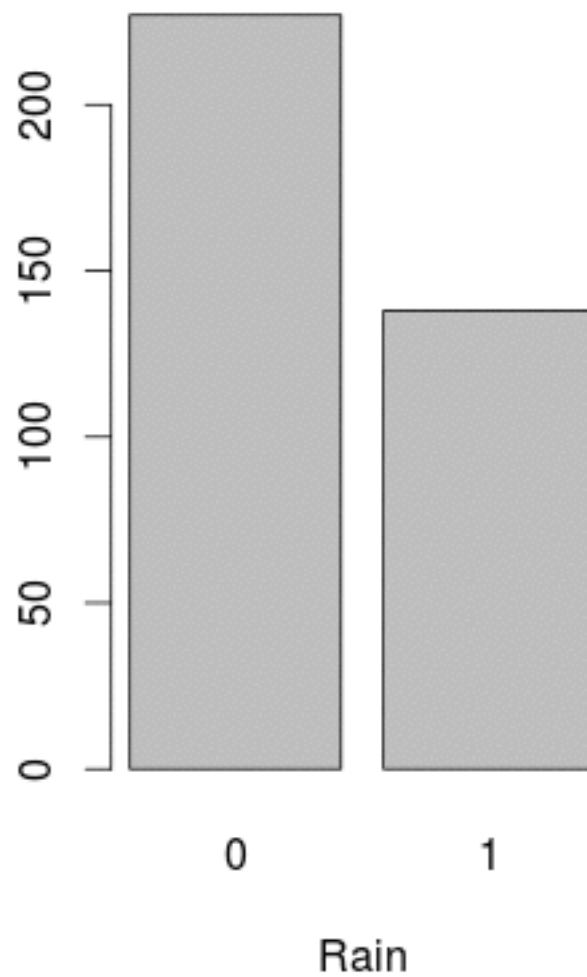
```
> length(unique(data1$Direction_II))
```

```
[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
```

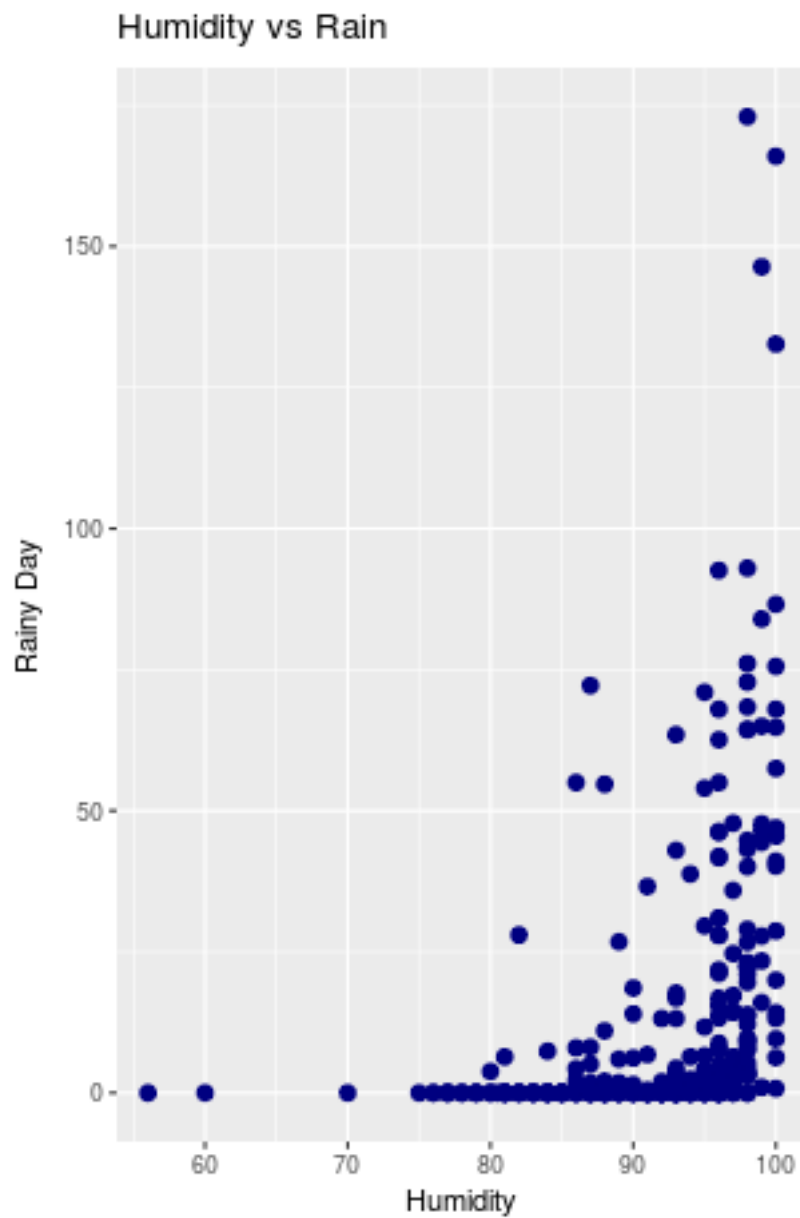
Some of the inferences drawn from variables in the data set:

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")
```

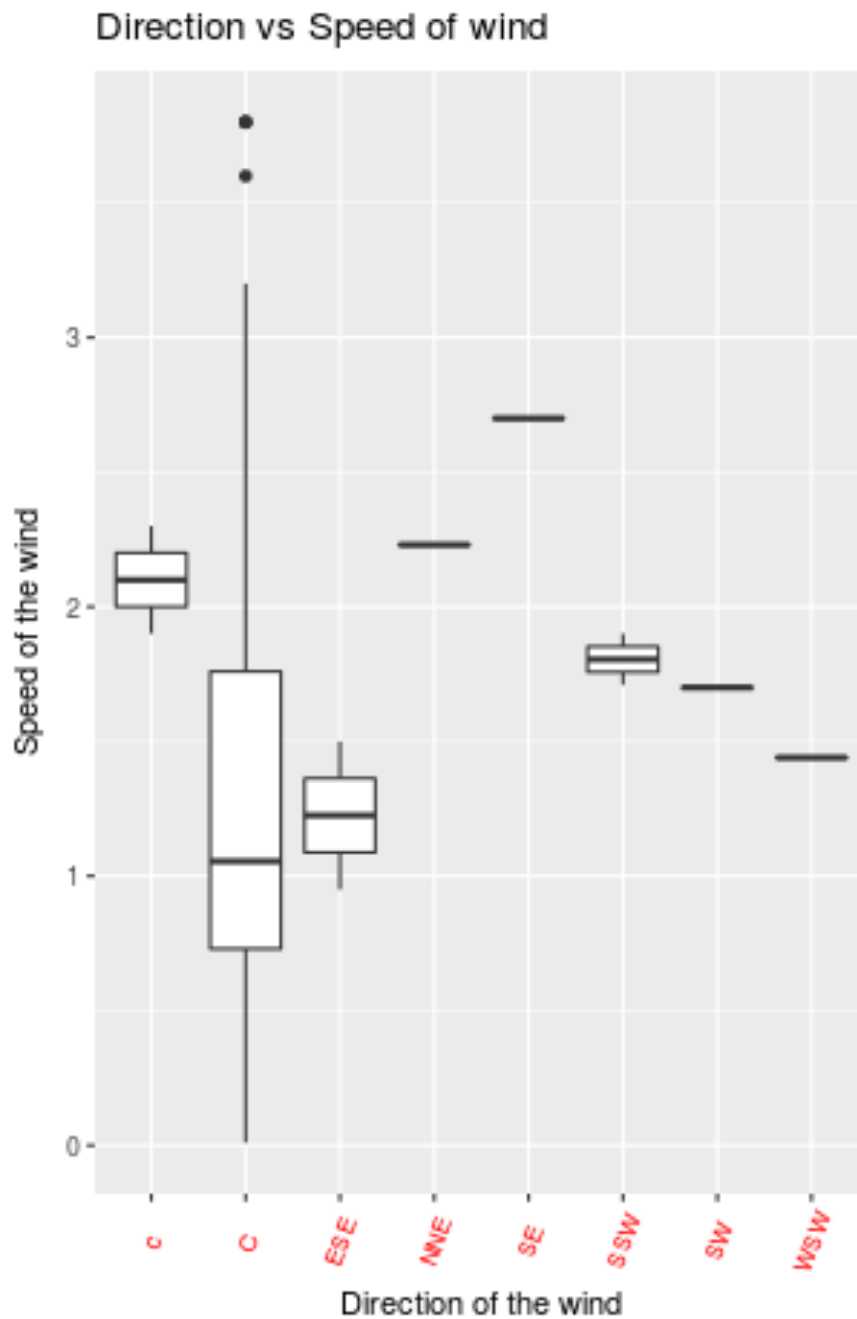


Direction vs Speed of wind

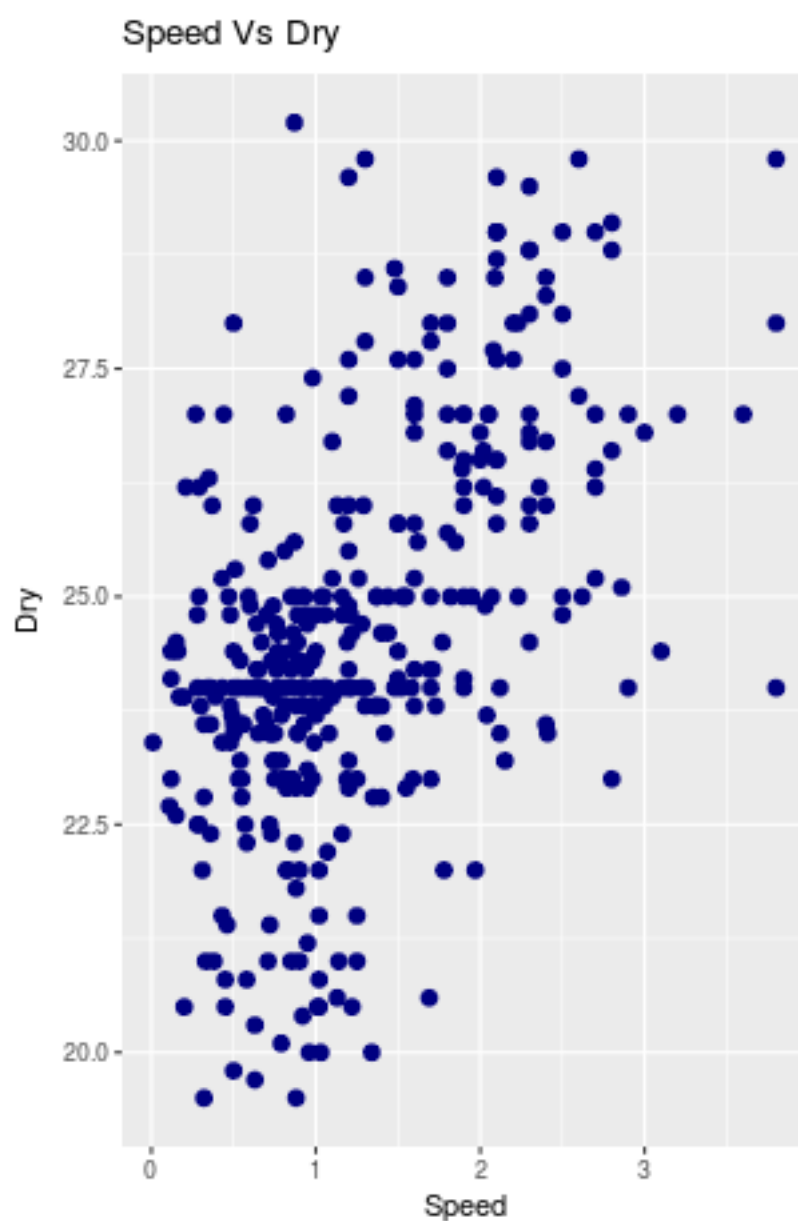
Speed of the wind

Direction of the wind

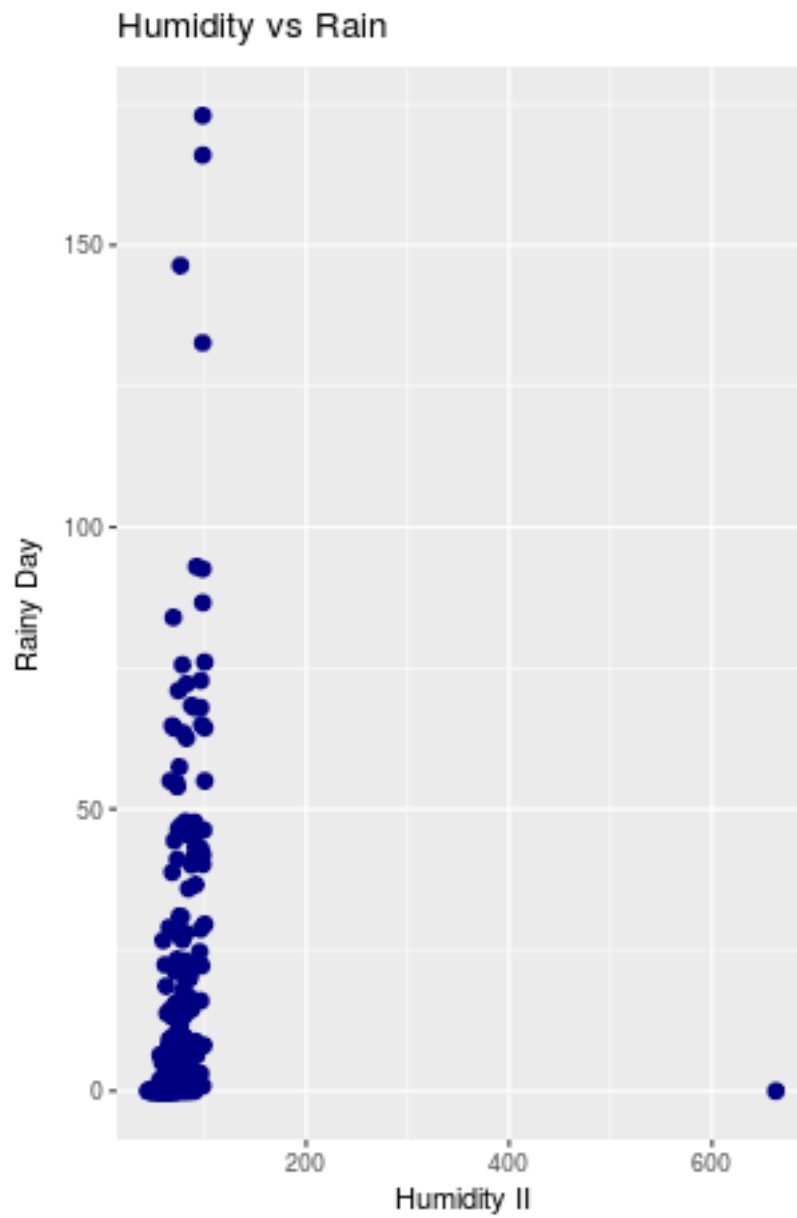
Direction of the wind	Min	Q1	Median	Q3	Max	Outliers
c	1.9	2.0	2.1	2.2	2.3	
C	0.0	0.7	1.1	1.8	3.2	3.5, 3.7
ESE	0.9	1.1	1.2	1.4	1.5	
NNE	2.2	2.2	2.2	2.2	2.2	
SE	2.7	2.7	2.7	2.7	2.7	
SSW	1.7	1.8	1.8	1.9	1.9	
SW	1.7	1.7	1.7	1.7	1.7	
WSW	1.4	1.4	1.4	1.4	1.4	



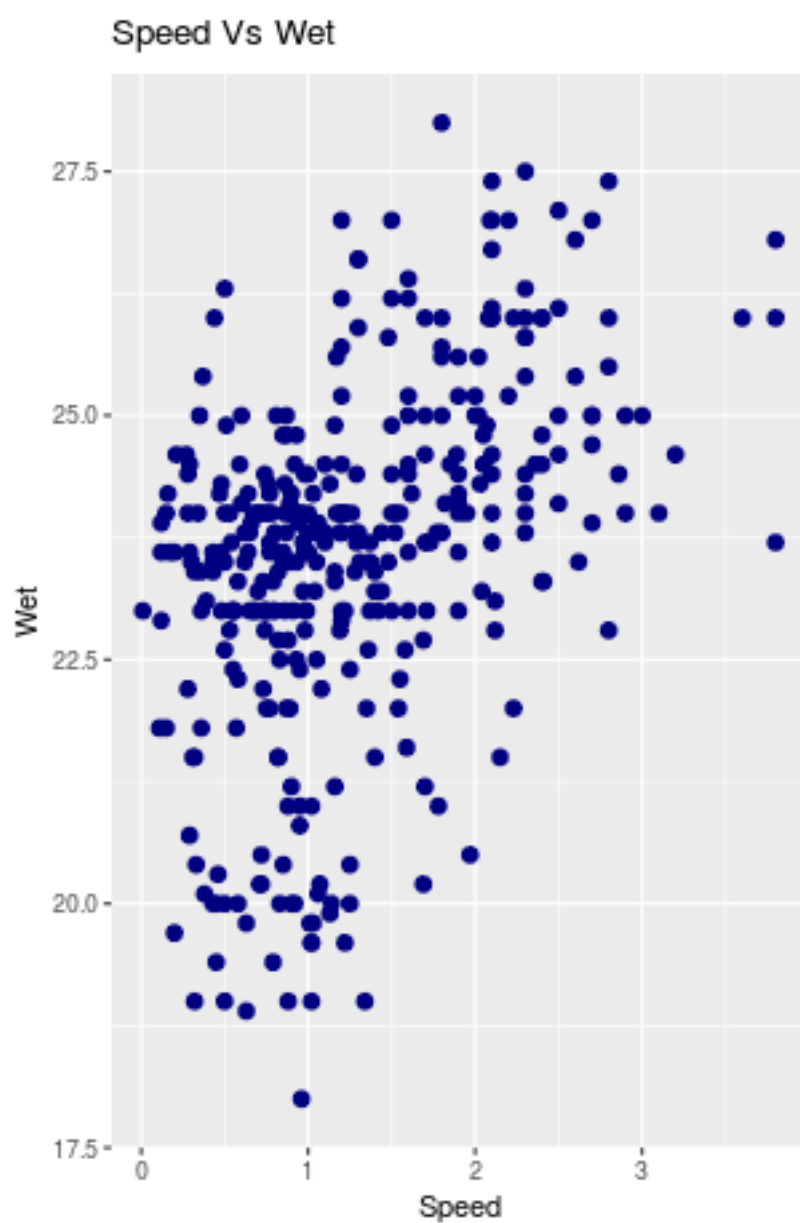
```
>ggplot(data1, aes(x= data1$Speed_I, y = data1$Dry_I)) + geom_point(size = 2.5,  
color="navy") + xlab("Speed") + ylab("Dry") + ggtitle("Speed Vs Dry")
```



```
> 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  $75 < \text{Humidity} < 100$  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

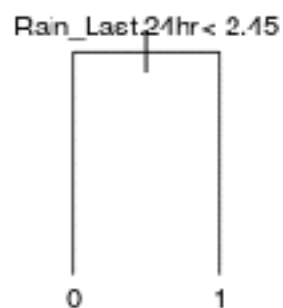
Kappa : 0.9438  
McNemar's Test P-Value : 0.4795

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) *
```