Airquality

Sowmiya S

2023-02-19

```
data()
```

```
library(MASS)
df=airquality
View(df)

#STRUCTURE OF THE DATA
str(df)
```

```
## 'data.frame':    153 obs. of  6 variables:
```

```
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
summary(df)
##      Ozone           Solar.R         Wind             Temp
Month           Day
```

```
##  Min.   :  1.0   Min.   :  7   Min.   : 1.70   Min.   :56.0   Min.
:5.00   Min.   : 1.0
##  1st Qu.: 18.0   1st Qu.:116   1st Qu.: 7.40   1st Qu.:72.0   1st
Qu.:6.00   1st Qu.: 8.0
##  Median : 31.5   Median :205   Median : 9.70   Median :79.0   Median
:7.00   Median :16.0
##  Mean   : 42.1   Mean   :186   Mean   : 9.96   Mean   :77.9   Mean
:6.99   Mean   :15.8
##  3rd Qu.: 63.2   3rd Qu.:259   3rd Qu.:11.50   3rd Qu.:85.0   3rd
Qu.:8.00   3rd Qu.:23.0
##  Max.   :168.0   Max.   :334   Max.   :20.70   Max.   :97.0   Max.
:9.00   Max.   :31.0
##  NA's   :37      NA's   :7
```

```
#UNDERSTANDING THE DATA
```

```
head(df)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
tail(df)
##     Ozone Solar.R Wind Temp Month Day
## 148    14      20 16.6   63     9  25
## 149    30     193  6.9   70     9  26
## 150    NA     145 13.2   77     9  27
## 151    14     191 14.3   75     9  28
## 152    18     131  8.0   76     9  29
## 153    20     223 11.5   68     9  30
```

```
dim(df)
## [1] 153   6
colnames(df)
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
colSums(is.na(df))
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

```
#SUBSETTING THE DATASET
```

```
library(dplyr)
#SELECT FUNCTION
df1=select(df,Ozone,Day,Month)
head(df1)
```

```
##   Ozone Day Month
## 1    41   1     5
## 2    36   2     5
## 3    12   3     5
## 4    18   4     5
## 5    NA   5     5
## 6    28   6     5
```

```
df2=select(df,Ozone:Wind)
```

```
head(df2)
```

```
##   Ozone Solar.R Wind
```

```
## 1    41     190  7.4
## 2    36     118  8.0
## 3    12     149 12.6
## 4    18     313 11.5
## 5    NA      NA 14.3
## 6    28      NA 14.9
```

```
df3=select(df,-Solar.R)
```

head(df3)

```
##   Ozone Wind Temp Month Day
```

```
## 1    41  7.4   67     5   1
## 2    36  8.0   72     5   2
## 3    12 12.6   74     5   3
## 4    18 11.5   62     5   4
## 5    NA 14.3   56     5   5
## 6    28 14.9   66     5   6
```

```
head(select(df,-(Temp:Day)),3)
##   Ozone Solar.R Wind
```

```
## 1    41     190  7.4
## 2    36     118  8.0
## 3    12     149 12.6
```

```
df4=select(df,contains("O"))
```

head(df4)

```
##   Ozone Solar.R Month
```

```
## 1    41     190     5
## 2    36     118     5
## 3    12     149     5
## 4    18     313     5
## 5    NA      NA     5
## 6    28      NA     5
```

```
#FILTER FUNCTION
```

filter(df,Month==9,Temp>90)

```
##   Ozone Solar.R Wind Temp Month Day
```

```
## 1    96     167  6.9   91     9   1
## 2    78     197  5.1   92     9   2
## 3    73     183  2.8   93     9   3
## 4    91     189  4.6   93     9   4
```

```
filter(df,Day<5&Solar.R>=200)
##   Ozone Solar.R Wind Temp Month Day
```

```
## 1    18     313 11.5   62     5   4
## 2    NA     286  8.6   78     6   1
## 3    NA     287  9.7   74     6   2
## 4    NA     242 16.1   67     6   3
## 5   135     269  4.1   84     7   1
## 6    49     248  9.2   85     7   2
## 7    32     236  9.2   81     7   3
```

```
head(filter(df,Month==8|Wind<5),5)
##   Ozone Solar.R Wind Temp Month Day
```

```
## 1    NA      59  1.7   76     6  22
## 2    NA      91  4.6   76     6  23
## 3   135     269  4.1   84     7   1
## 4    64     175  4.6   83     7   5
## 5    39      83  6.9   81     8   1
```

```
head(filter(df,!is.na(Ozone)),5)
##   Ozone Solar.R Wind Temp Month Day
```

```
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    28      NA 14.9   66     5   6
```

```
#ARRANGE FUNCTION
```

```
df=arrange(df,Day)
head(df)
```

```
##   Ozone Solar.R Wind Temp Month Day
```

```
## 1    41     190  7.4   67     5   1
## 2    NA     286  8.6   78     6   1
## 3   135     269  4.1   84     7   1
## 4    39      83  6.9   81     8   1
## 5    96     167  6.9   91     9   1
## 6    36     118  8.0   72     5   2
```

```
df=arrange(df,desc(Temp))
```

head(df)

```
##   Ozone Solar.R Wind Temp Month Day
## 1    76     203  9.7   97     8  28
## 2    84     237  6.3   96     8  30
## 3   118     225  2.3   94     8  29
## 4    85     188  6.3   94     8  31
## 5    73     183  2.8   93     9   3
## 6    91     189  4.6   93     9   4
```

```
df=arrange(df,Day,desc(Month))
```

head(df)

```
##   Ozone Solar.R Wind Temp Month Day
## 1    96     167  6.9   91     9   1
## 2    39      83  6.9   81     8   1
## 3   135     269  4.1   84     7   1
## 4    NA     286  8.6   78     6   1
## 5    41     190  7.4   67     5   1
## 6    78     197  5.1   92     9   2
```

```
#MUTATE FUNCTION
```

df=mutate(df,temp_celsius=(Temp-32)*5/9)
head(df)

```
##   Ozone Solar.R Wind Temp Month Day temp_celsius
## 1    96     167  6.9   91     9   1         32.8
## 2    39      83  6.9   81     8   1         27.2
## 3   135     269  4.1   84     7   1         28.9
## 4    NA     286  8.6   78     6   1         25.6
## 5    41     190  7.4   67     5   1         19.4
## 6    78     197  5.1   92     9   2         33.3
```

```
df=mutate(df,TempCat=factor((Temp>80),labels=c("cold","hot")))
```

head(df)

```
##   Ozone Solar.R Wind Temp Month Day temp_celsius TempCat
## 1    96     167  6.9   91     9   1         32.8     hot
## 2    39      83  6.9   81     8   1         27.2     hot
```

```
## 3   135     269  4.1   84     7   1         28.9     hot
## 4    NA     286  8.6   78     6   1         25.6    cold
## 5    41     190  7.4   67     5   1         19.4    cold
## 6    78     197  5.1   92     9   2         33.3     hot
```

```
#SUMMARISE FUNCTION
```

summarise(df,median_Oz=median(Ozone,na.rm=TRUE))

```
##   median_Oz
```

```
## 1      31.5
```

```
summarise(df,max_temp=max(Temp),min_temp=min(Temp))
##   max_temp min_temp
```

```
## 1       97       56
```

```
summarise(df,Ozone=mean(Ozone,na.rm=TRUE))
##   Ozone
```

```
## 1  42.1
```

```
#DATA TRANSFORMATION
```

#HANDLING MISSING VALUES
NROW(df$Ozone)

```
## [1] 153
#REMOVING MISSING VALUES
```

x=na.omit(df$Ozone)
NROW(x)

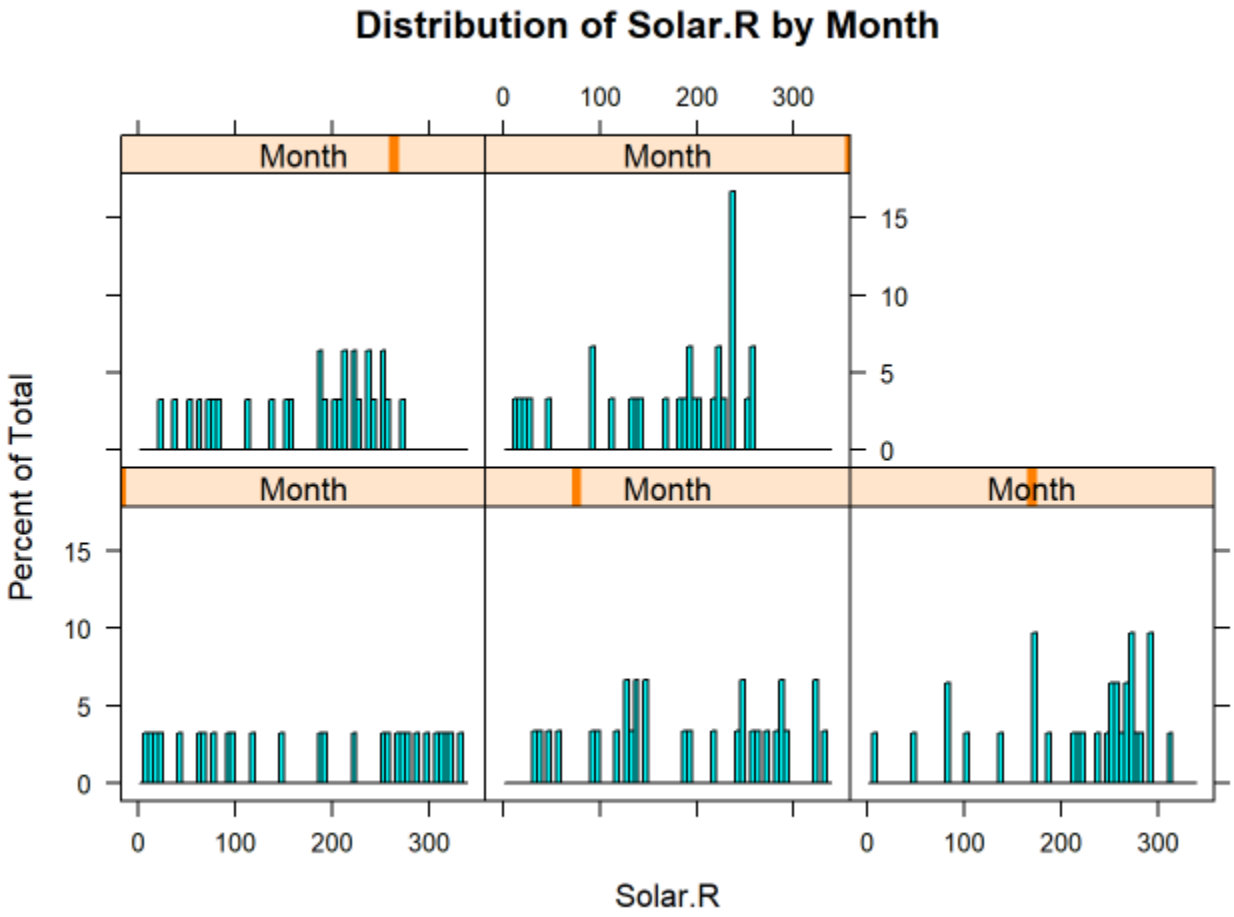
```
## [1] 116
Q1=quantile(df$Wind,0.25)
```

Q3=quantile(df$Wind,0.75)
IQR=IQR(df$Wind)
no_outliers=subset(df,df$Wind>(Q1-1.5*IQR)&df$Wind<(Q3+1.5*IQR))
NROW(no_outliers)

```
## [1] 150
# VISUALIZING THE DATASET
```

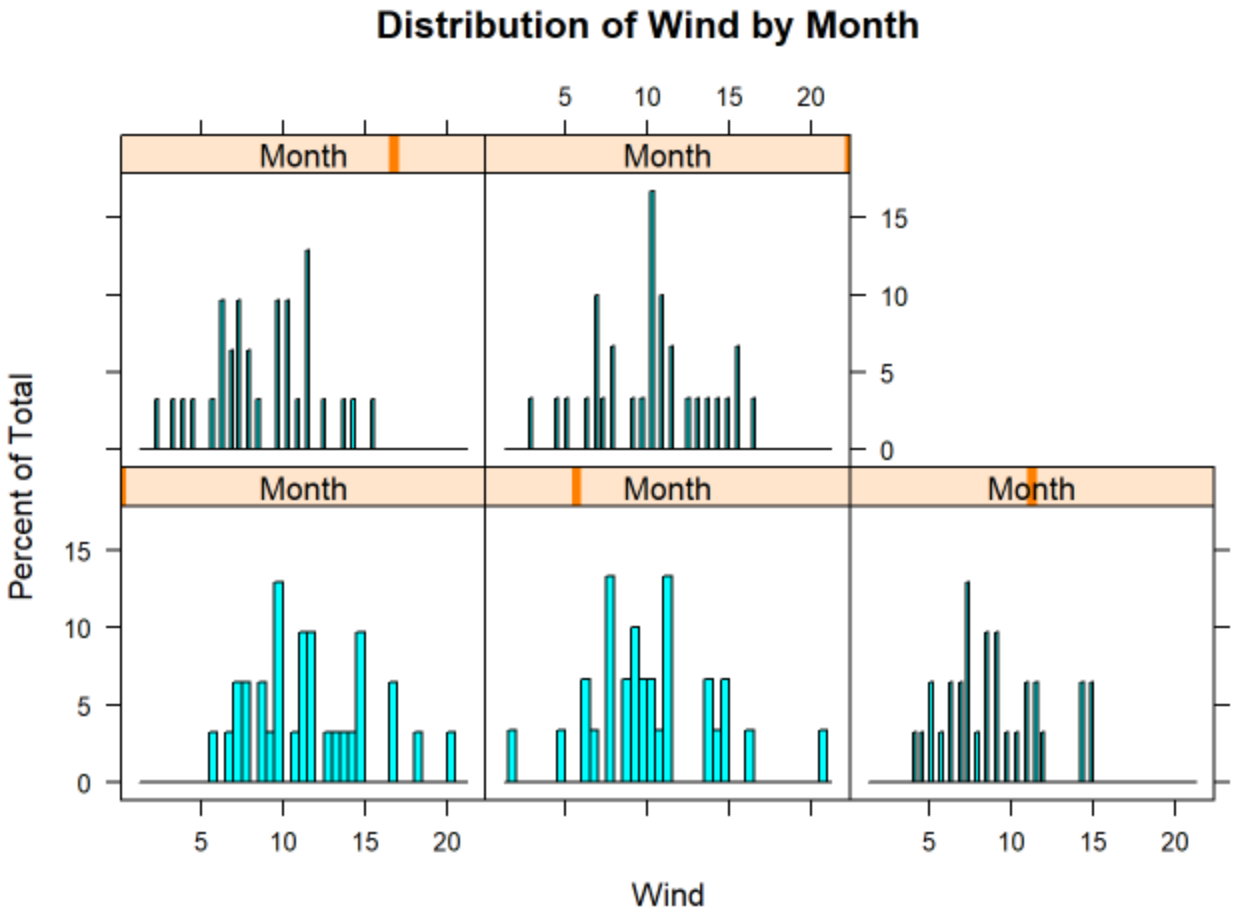
#1: Which month got the most Solar radiation?
#Using histogram to find out the the maximum solar radiation in Month wise analysis

```
library(lattice)
histogram(~Solar.R|Month,data=df,breaks=50,main="Distribution of Solar.R by Month")
```



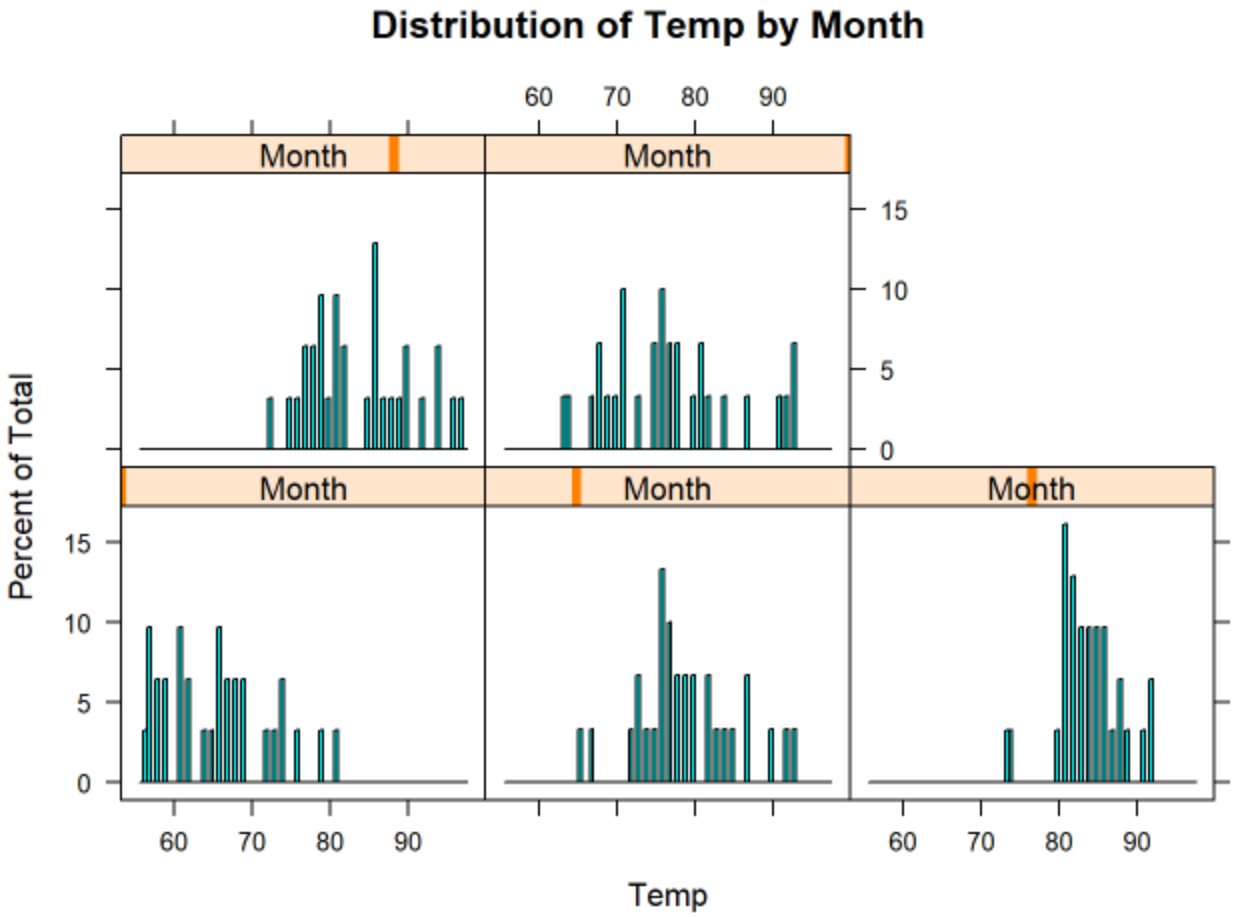

```
#2:Find out Which month got the maximum wind speed?
```

```
histogram(~Wind|Month,data=df,breaks=50,main="Distribution of Wind by Month")
```



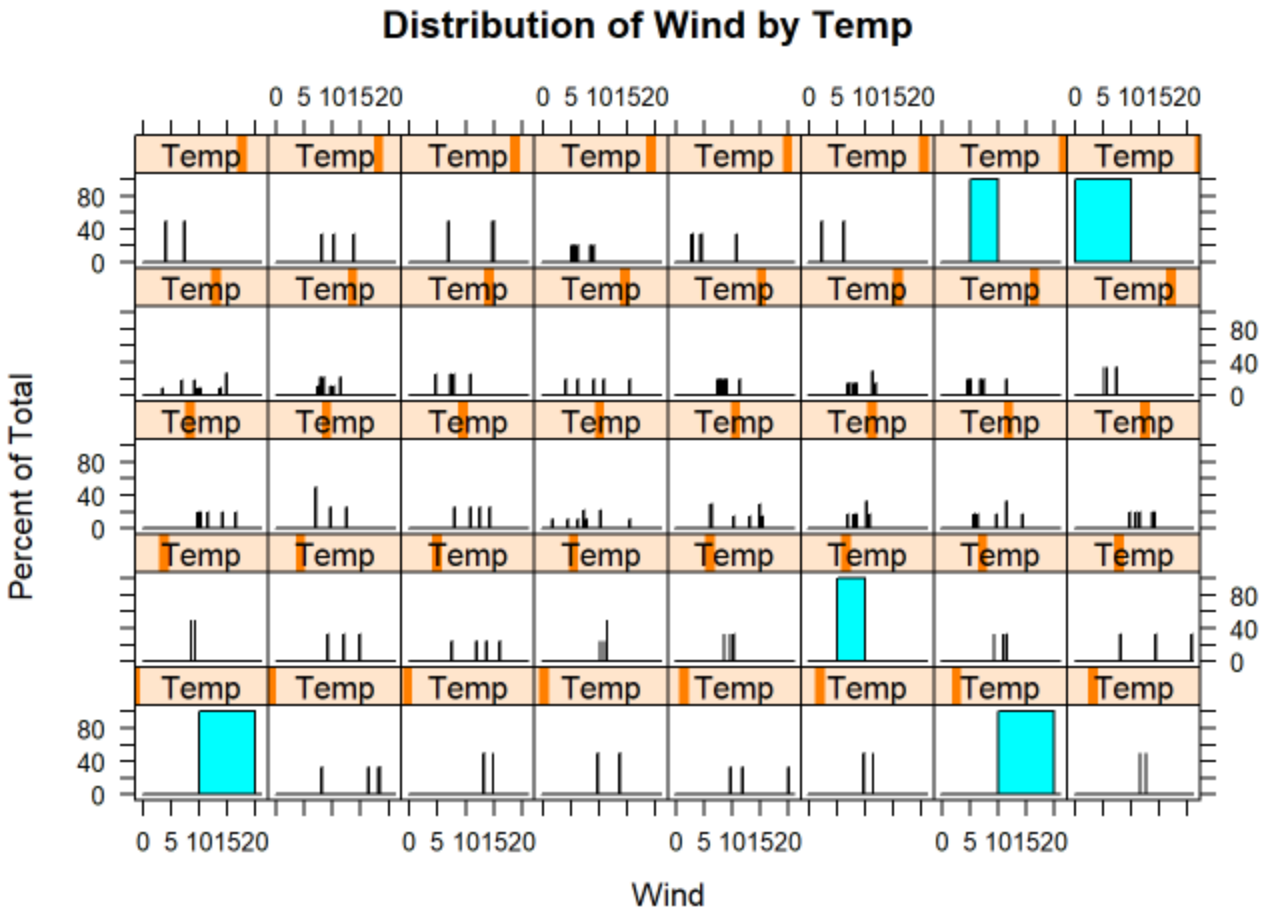

```
#3:Find out Which month got the maximum daily temperature?
```

histogram(~Temp|Month,data=df,breaks=50,main="Distribution of Temp by Month")



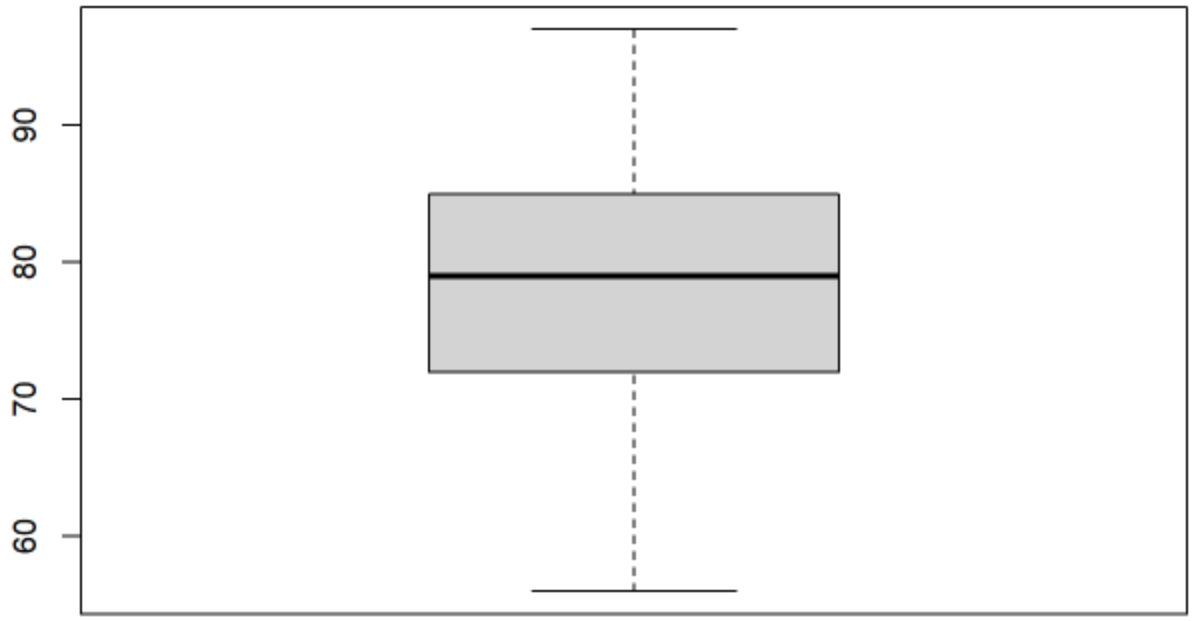

```
#4:Find out Which temperature got the maximum Wind ?
```

histogram(~Wind|Temp,data=df,breaks=50,main="Distribution of Wind by Temp")



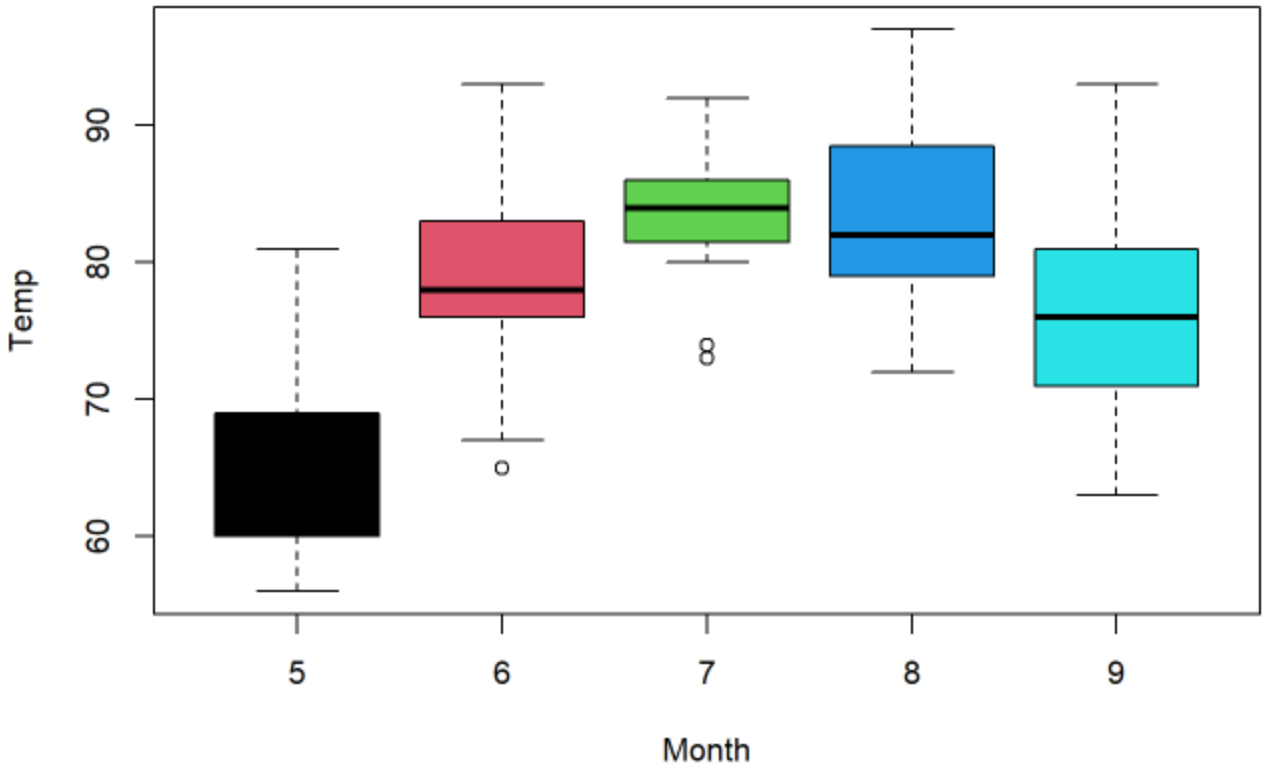

```
#BOXPLOT
```

#1
with(df,boxplot(Temp))




```
#2
```

```
with(df,boxplot(Temp~Month,col=c(1,2,3,4,5)))
```



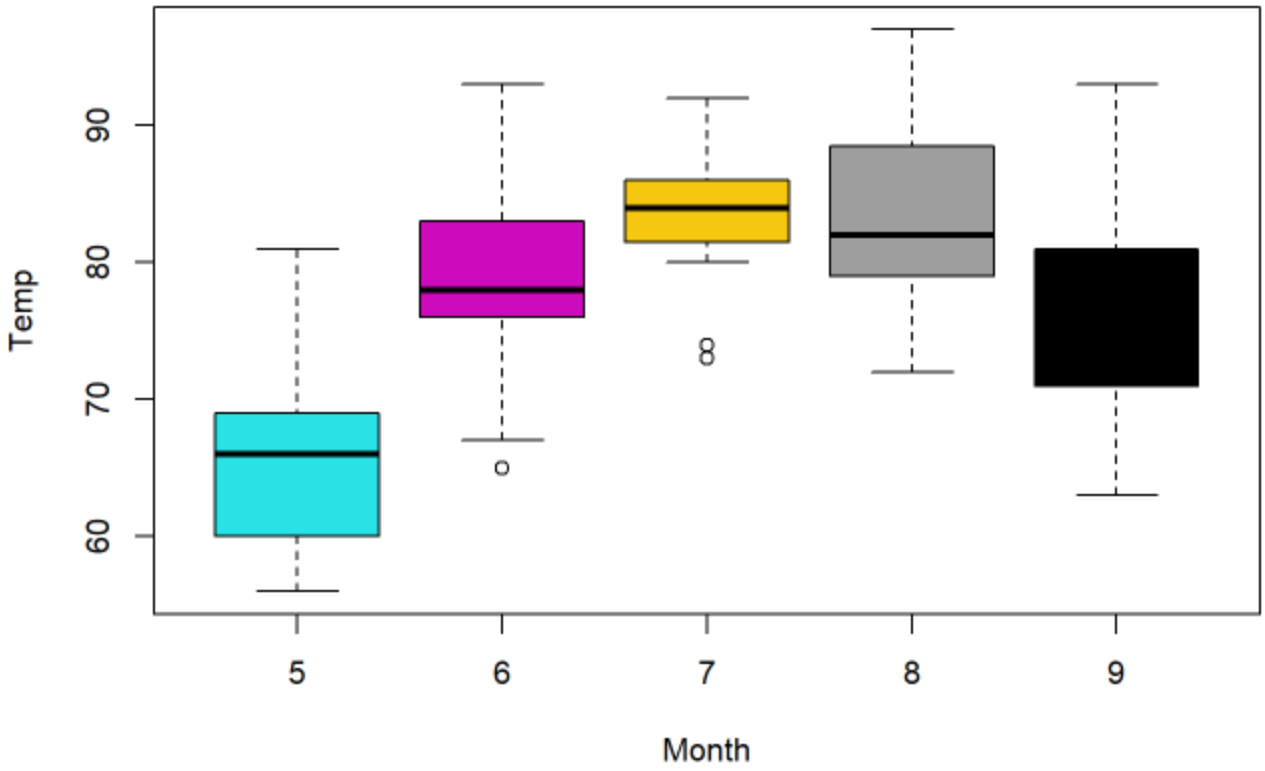

```
#3
```

with(df,as.factor(Month))

```
##   [1] 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9
8 7 6 5 9 8 7 6 5 9 8 7 6
```

```
##  [45] 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5
9 8 7 6 5 9 8 7 6 5 9 8 7
##  [89] 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6
5 9 8 7 6 5 9 8 7 6 5 9 8
## [133] 7 6 5 9 8 7 6 5 9 8 7 6 5 9 8 7 6 5 8 7 5
## Levels: 5 6 7 8 9
```

```
levels(with(df,as.factor(Month)))
## [1] "5" "6" "7" "8" "9"
with(df,boxplot(Temp~Month,col=levels(with(df,as.factor(Month)))))
```




```
#SCATTERPLOT
```

#1 Which month has the maximum temperature?

```
library(plotly)
fig=plot_ly(data=df,x=~Month,y=~Temp,type="scatter")%>%layout(title="Scatterplot between Month and Temp")
fig
```

```
## No scatter mode specifed:
```

```
##   Setting the mode to markers
##   Read more about this attribute ->
https://plotly.com/r/reference/#scatter-mode
```

Data Description: The dataset contains the details of daily air quality measurements in New York from May to September 1973 over a period of 5 months. It is classified by Ozone, Solar.R, Wind, Temp, Month and Day.

ASSUMPTIONS:

The dataset contains air quality measurements of 1973 for five months from may to September recorded daily. I assumed to subset each of the attribute and try to conclude about the data. I try to find out in which month we get maximum solar radiation, wind speed and temperature. Depends upon these factors the variations may occur.

INFERENCE:

1. Averagely 185 mph are wind speed. Speed which are above 7 mph and below 11 mph has completed the wind speed.

2. Averagely 77 degrees in F has completed the maximum temperature.

3. Averagely 6 days to complete in a month.

4. Average 42 parts per billion complete the Ozone readings.

5. DISTRIBUTION OF SOLAR.R BY MONTH: These distributions are multi model distributions and may be variations in these data.

6. DISTRIBUTION OF WIND BY TEMP: These distributions between 10-15 mph speed occurs. So many variations in these speed levels.

INSIGHTS:

1. The 200-300 langleys are most completed their Solar Radiation.

2. Maximum wind speed is in the month of 10.

3.Temperature between 80-85 degrees in F are most completed in a month.

4.The wind speed between 10-20 are maximum speed in a month.

5. Comparing with the month and temperature , above the temperature of 70 degrees in F at a month of 8 are highly corresponded.

6. In the month of 10 we get the maximum temperature and wind speed. So by comparatively it is higher prioritize than other attributes.

7. Month is mostly preferrable attribute in a dataset. It helps to predict that in a year which month gives most needable aspect to analyze the daily airquality of Newyork.