Airquality-dataset.R

Yashwanth Rajan P. R

2023-03-21

```
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
```

```
#RENAME FUNCTION
```

rename(df,Temperature=Temp)

```
##     Ozone Solar.R Wind Temperature 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
## 7       9      24 13.8          81     8   2         27.2     hot
## 8      49     248  9.2          85     7   2         29.4     hot
## 9      NA     287  9.7          74     6   2         23.3    cold
## 10     36     118  8.0          72     5   2         22.2    cold
## 11     73     183  2.8          93     9   3         33.9     hot
## 12     16      77  7.4          82     8   3         27.8     hot
## 13     32     236  9.2          81     7   3         27.2     hot
## 14     NA     242 16.1          67     6   3         19.4    cold
## 15     12     149 12.6          74     5   3         23.3    cold
## 16     91     189  4.6          93     9   4         33.9     hot
## 17     78      NA  6.9          86     8   4         30.0     hot
## 18     NA     101 10.9          84     7   4         28.9     hot
## 19     NA     186  9.2          84     6   4         28.9     hot
```

```
## 20     18     313 11.5          62     5   4          16.7    cold
## 21     47      95  7.4          87     9   5          30.6     hot
## 22     35      NA  7.4          85     8   5          29.4     hot
## 23     64     175  4.6          83     7   5          28.3     hot
## 24     NA     220  8.6          85     6   5          29.4     hot
## 25     NA      NA 14.3          56     5   5          13.3    cold
## 26     32      92 15.5          84     9   6          28.9     hot
## 27     66      NA  4.6          87     8   6          30.6     hot
## 28     40     314 10.9          83     7   6          28.3     hot
## 29     NA     264 14.3          79     6   6          26.1    cold
## 30     28      NA 14.9          66     5   6          18.9    cold
## 31     20     252 10.9          80     9   7          26.7    cold
## 32    122     255  4.0          89     8   7          31.7     hot
## 33     77     276  5.1          88     7   7          31.1     hot
## 34     29     127  9.7          82     6   7          27.8     hot
## 35     23     299  8.6          65     5   7          18.3    cold
## 36     23     220 10.3          78     9   8          25.6    cold
## 37     89     229 10.3          90     8   8          32.2     hot
## 38     97     267  6.3          92     7   8          33.3     hot
## 39     NA     273  6.9          87     6   8          30.6     hot
## 40     19      99 13.8          59     5   8          15.0    cold
## 41     21     230 10.9          75     9   9          23.9    cold
## 42    110     207  8.0          90     8   9          32.2     hot
## 43     97     272  5.7          92     7   9          33.3     hot
## 44     71     291 13.8          90     6   9          32.2     hot
## 45      8      19 20.1          61     5   9          16.1    cold
## 46     24     259  9.7          73     9  10          22.8    cold
## 47     NA     222  8.6          92     8  10          33.3     hot
## 48     85     175  7.4          89     7  10          31.7     hot
## 49     39     323 11.5          87     6  10          30.6     hot
## 50     NA     194  8.6          69     5  10          20.6    cold
## 51     44     236 14.9          81     9  11          27.2     hot
## 52     NA     137 11.5          86     8  11          30.0     hot
## 53     NA     139  8.6          82     7  11          27.8     hot
## 54     NA     259 10.9          93     6  11          33.9     hot
## 55      7      NA  6.9          74     5  11          23.3    cold
## 56     21     259 15.5          76     9  12          24.4    cold
## 57     44     192 11.5          86     8  12          30.0     hot
## 58     10     264 14.3          73     7  12          22.8    cold
## 59     NA     250  9.2          92     6  12          33.3     hot
## 60     16     256  9.7          69     5  12          20.6    cold
## 61     28     238  6.3          77     9  13          25.0    cold
## 62     28     273 11.5          82     8  13          27.8     hot
## 63     27     175 14.9          81     7  13          27.2     hot
## 64     23     148  8.0          82     6  13          27.8     hot
## 65     11     290  9.2          66     5  13          18.9    cold
```

```
##  66      9      24 10.9         71     9  14         21.7    cold
##  67     65     157  9.7         80     8  14         26.7    cold
##  68     NA     291 14.9         91     7  14         32.8     hot
##  69     NA     332 13.8         80     6  14         26.7    cold
##  70     14     274 10.9         68     5  14         20.0    cold
##  71     13     112 11.5         71     9  15         21.7    cold
##  72     NA      64 11.5         79     8  15         26.1    cold
##  73      7      48 14.3         80     7  15         26.7    cold
##  74     NA     322 11.5         79     6  15         26.1    cold
##  75     18      65 13.2         58     5  15         14.4    cold
##  76     46     237  6.9         78     9  16         25.6    cold
##  77     22      71 10.3         77     8  16         25.0    cold
##  78     48     260  6.9         81     7  16         27.2     hot
##  79     21     191 14.9         77     6  16         25.0    cold
##  80     14     334 11.5         64     5  16         17.8    cold
##  81     18     224 13.8         67     9  17         19.4    cold
##  82     59      51  6.3         79     8  17         26.1    cold
##  83     35     274 10.3         82     7  17         27.8     hot
##  84     37     284 20.7         72     6  17         22.2    cold
##  85     34     307 12.0         66     5  17         18.9    cold
##  86     13      27 10.3         76     9  18         24.4    cold
##  87     23     115  7.4         76     8  18         24.4    cold
##  88     61     285  6.3         84     7  18         28.9     hot
##  89     20      37  9.2         65     6  18         18.3    cold
##  90      6      78 18.4         57     5  18         13.9    cold
##  91     24     238 10.3         68     9  19         20.0    cold
##  92     31     244 10.9         78     8  19         25.6    cold
##  93     79     187  5.1         87     7  19         30.6     hot
##  94     12     120 11.5         73     6  19         22.8    cold
##  95     30     322 11.5         68     5  19         20.0    cold
##  96     16     201  8.0         82     9  20         27.8     hot
##  97     44     190 10.3         78     8  20         25.6    cold
##  98     63     220 11.5         85     7  20         29.4     hot
##  99     13     137 10.3         76     6  20         24.4    cold
## 100     11      44  9.7         62     5  20         16.7    cold
## 101     13     238 12.6         64     9  21         17.8    cold
## 102     21     259 15.5         77     8  21         25.0    cold
## 103     16       7  6.9         74     7  21         23.3    cold
## 104     NA     150  6.3         77     6  21         25.0    cold
## 105      1       8  9.7         59     5  21         15.0    cold
## 106     23      14  9.2         71     9  22         21.7    cold
## 107      9      36 14.3         72     8  22         22.2    cold
## 108     NA     258  9.7         81     7  22         27.2     hot
## 109     NA      59  1.7         76     6  22         24.4    cold
## 110     11     320 16.6         73     5  22         22.8    cold
## 111     36     139 10.3         81     9  23         27.2     hot
```

```
## 112    NA     255 12.6          75     8  23         23.9    cold
## 113    NA     295 11.5          82     7  23         27.8     hot
## 114    NA      91  4.6          76     6  23         24.4    cold
## 115     4      25  9.7          61     5  23         16.1    cold
## 116     7      49 10.3          69     9  24         20.6    cold
## 117    45     212  9.7          79     8  24         26.1    cold
## 118    80     294  8.6          86     7  24         30.0     hot
## 119    NA     250  6.3          76     6  24         24.4    cold
## 120    32      92 12.0          61     5  24         16.1    cold
## 121    14      20 16.6          63     9  25         17.2    cold
## 122   168     238  3.4          81     8  25         27.2     hot
## 123   108     223  8.0          85     7  25         29.4     hot
## 124    NA     135  8.0          75     6  25         23.9    cold
## 125    NA      66 16.6          57     5  25         13.9    cold
##  [ reached 'max' / getOption("max.print") -- omitted 28 rows ]
```

```
#DATA TRANSFORMATION
```

#HANDLING MISSING VALUES
NROW(df$Ozone)

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

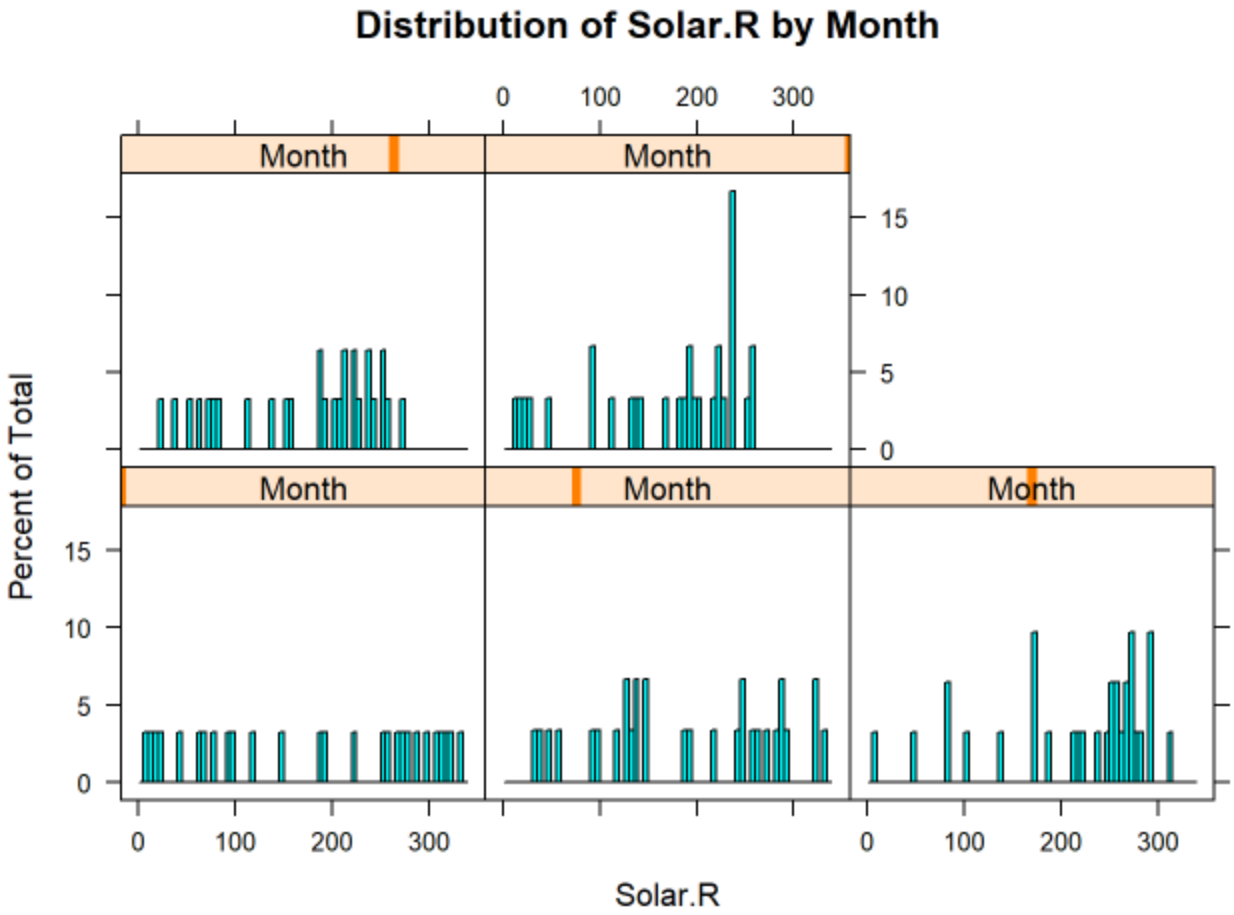
x=na.omit(df$Ozone)
NROW(x)

```
## [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)

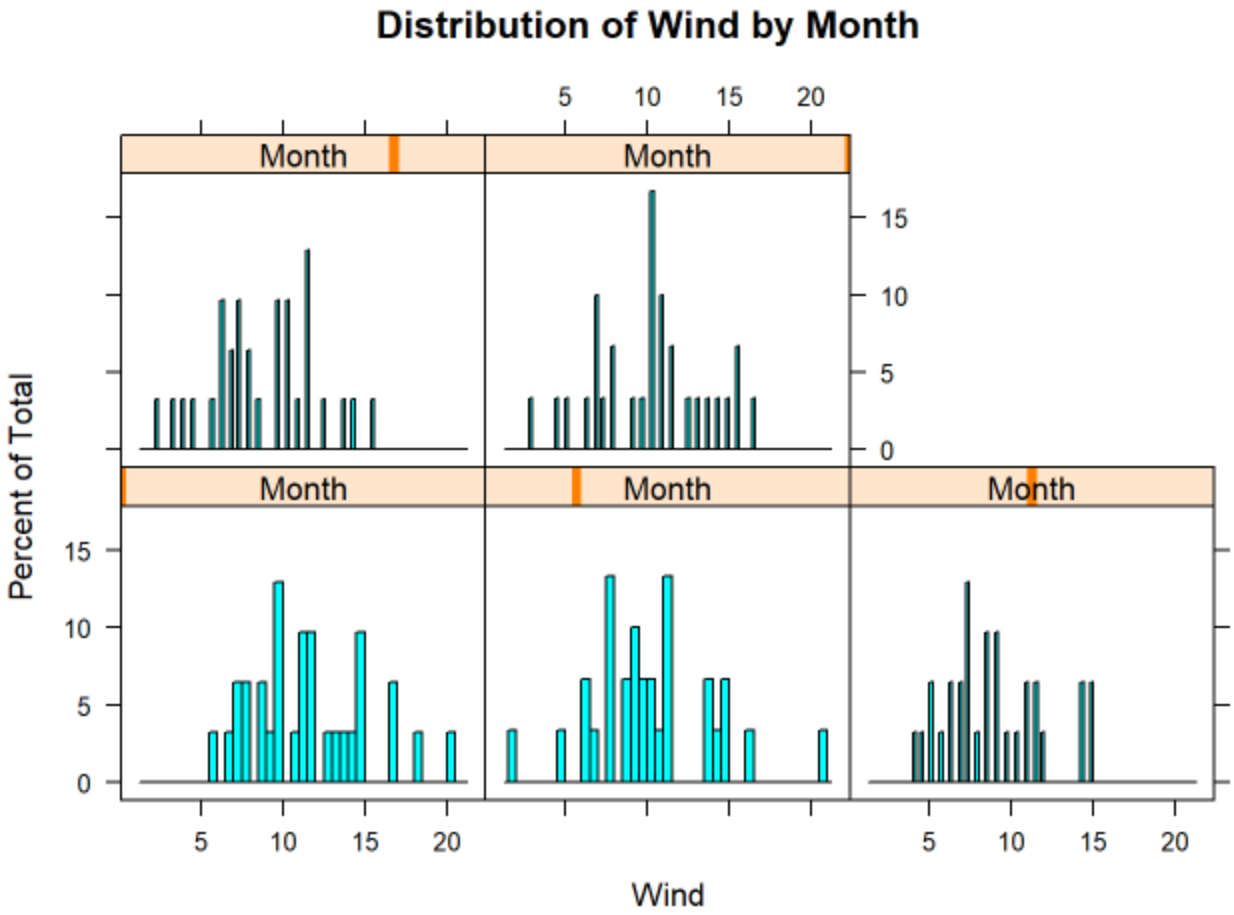
```
## [1] 150
# VISUALIZING THE DATASET
```

#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")



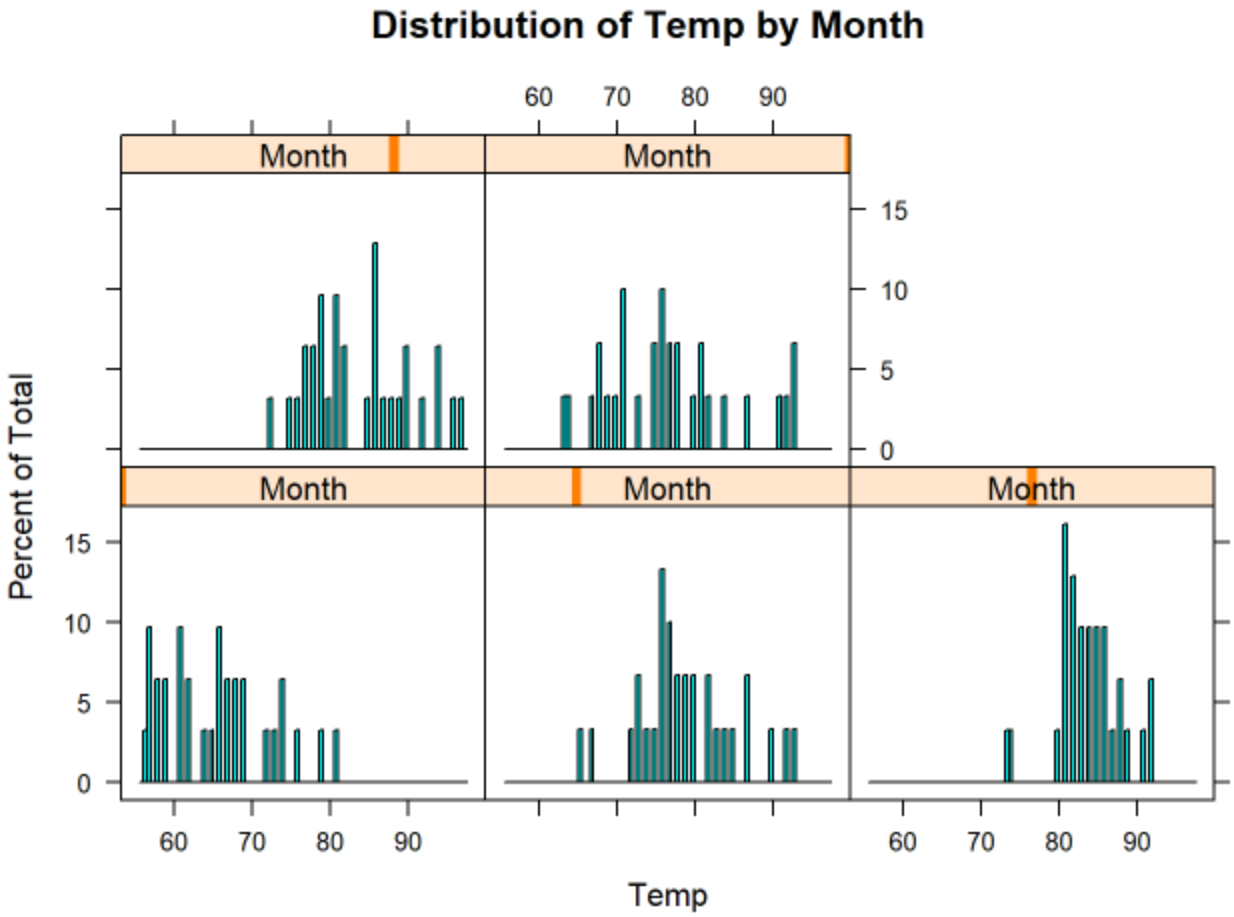

```
#2:Find out Which month got the maximum wind speed?
```

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



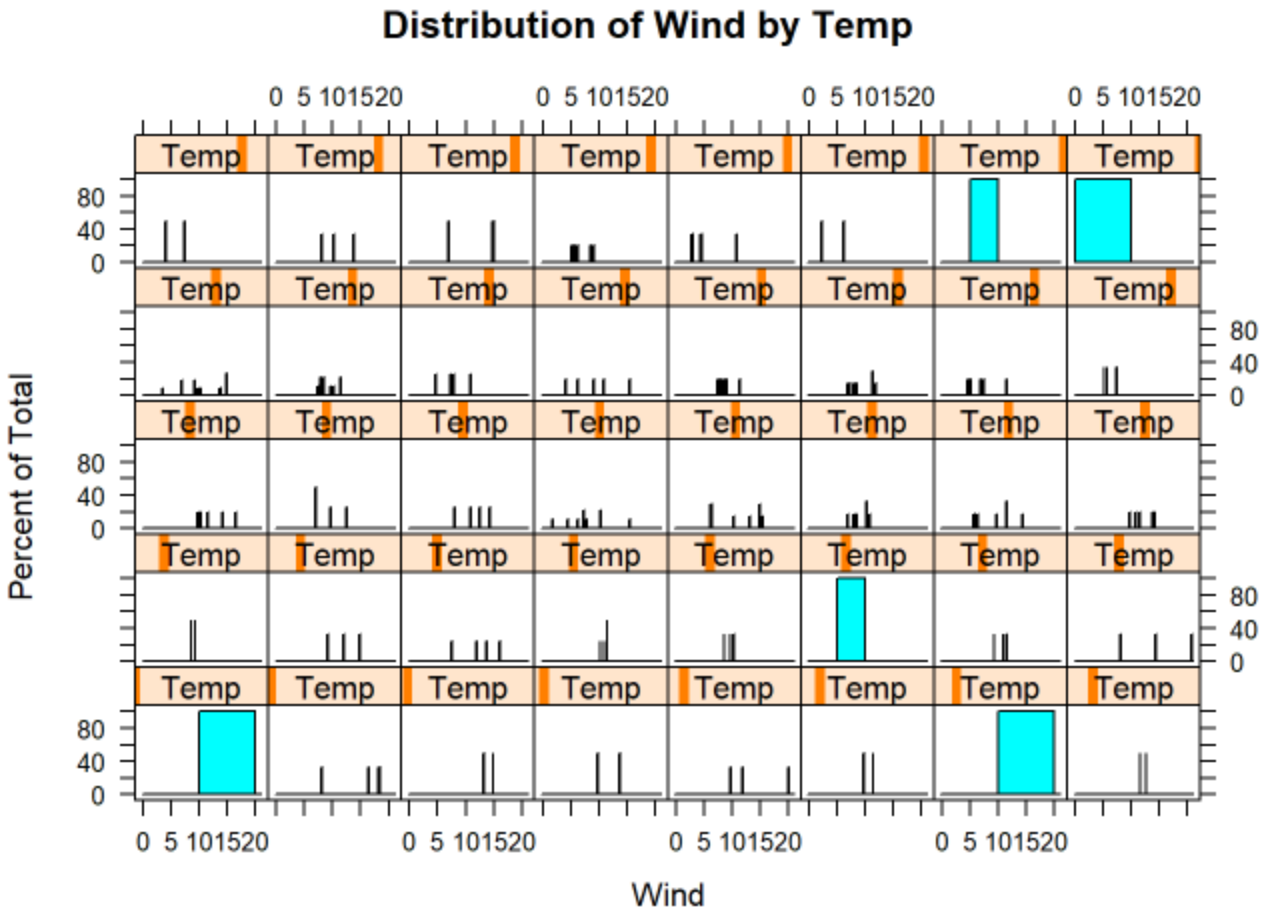

```
#3:Find out Which month got the maximum daily temperature?
```

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



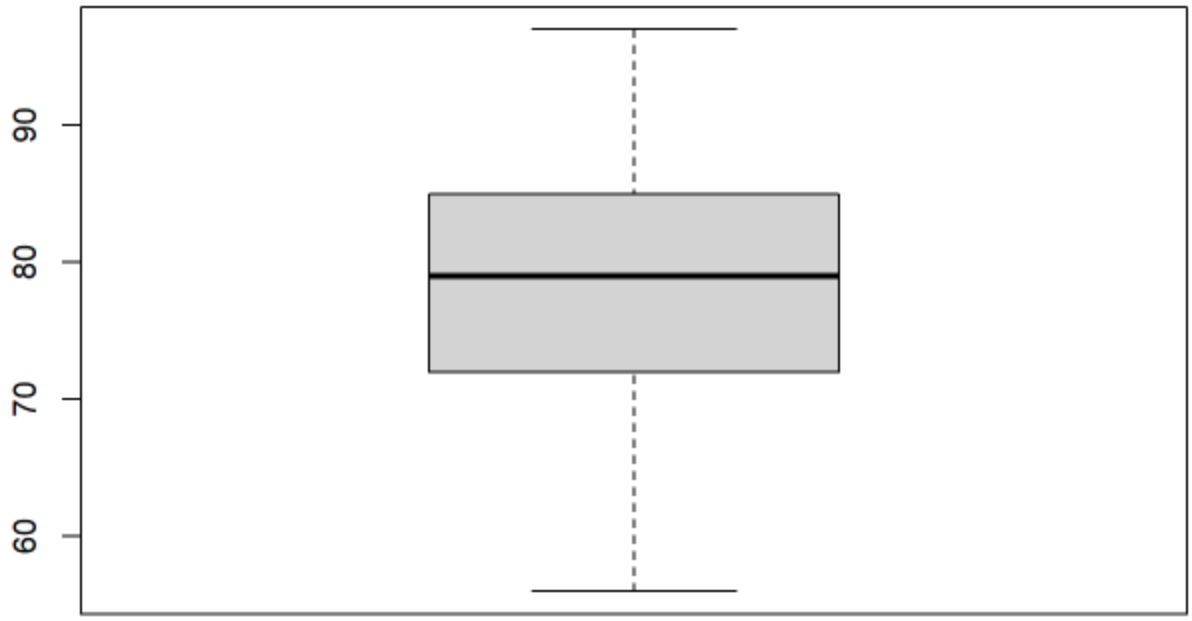

```
#4:Find out Which temperature got the maximum Wind ?
```

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



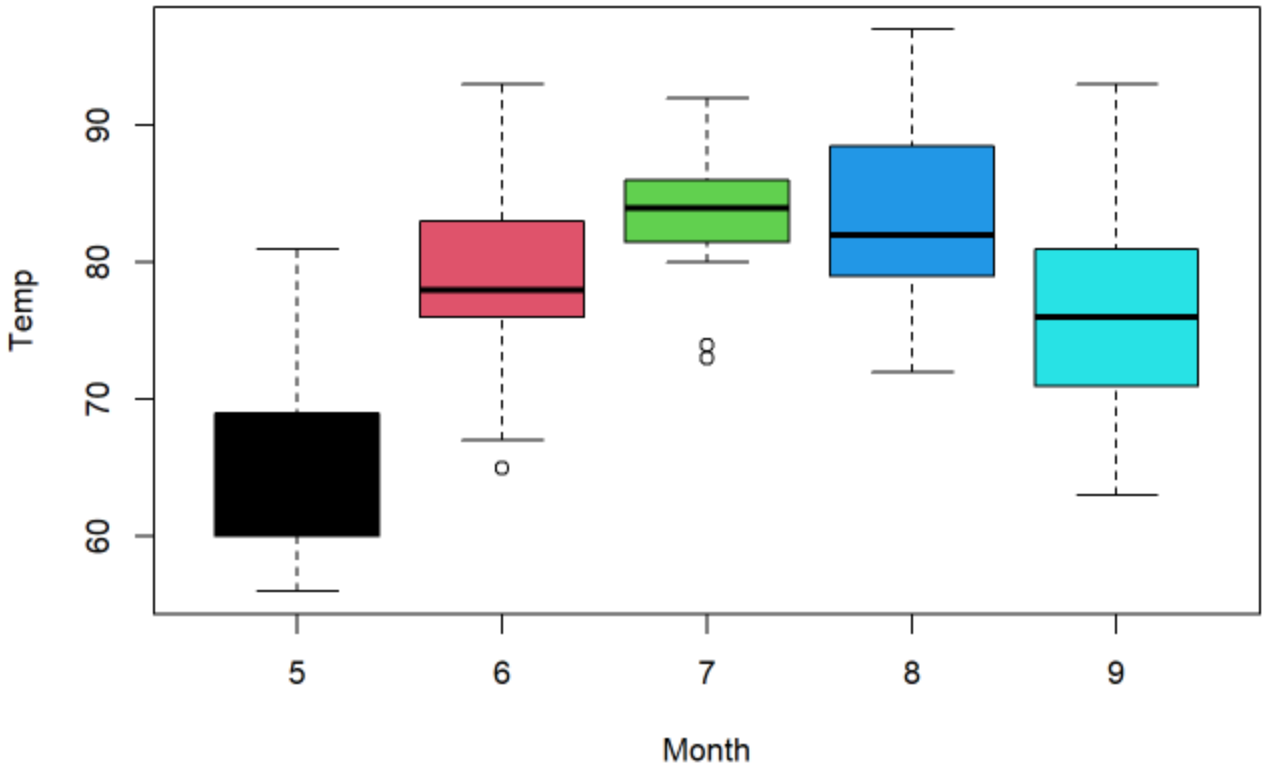

```
#BOXPLOT
```

#1
with(df,boxplot(Temp))




```
#2
```

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



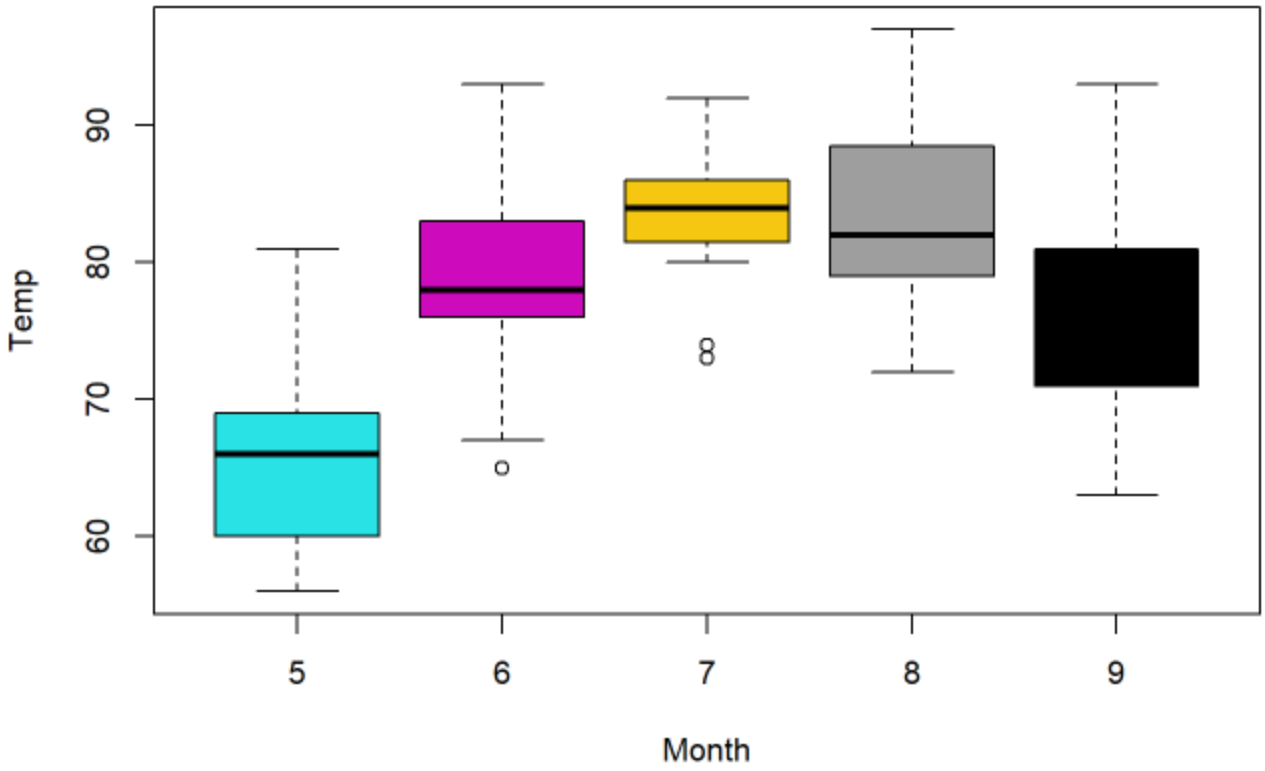

```
#3
```

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
```