R Notebook

Title: “IST687 – Air quality Analysis”
Name: Sathish Kumar Rajendiran
Week: 6
Date: 05/12/2020

Exercise: Air quality Analysis

Install necessary packages

```
install.packages( pkgs=c("ggplot2","reshape2","ggeasy","viridis"),repos = "http://cran.us.r-project.org
```

```
##
## The downloaded binary packages are in
##  /var/folders/_z/ltmjkt4156b37rsk7cgvj7l80000gn/T//Rtmpn4zykq/downloaded_packages
```

```
install.packages("ggplot2", repos ="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/_z/ltmjkt4156b37rsk7cgvj7l80000gn/T//Rtmpn4zykq/downloaded_packages
```

```
# install.packages("ggplot2")
# install.packages("reshape2")
# install.packages("ggeasy")
# install.packages("viridis")


library(ggplot2)
library(ggcorrplot)
library(reshape2)
library(ggeasy)
library(viridis)
```

```
## Loading required package: viridisLite
```

Step 1: Load the data

Step 2: Clean the data

```
#   Step 1: Load the data

    ?airquality

    myairquality <- data.frame(airquality)
    str(myairquality)
```

```
## '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 ...
```

```
    #look for columns having NAs
    clnames <- colnames(myairquality)[colSums(is.na(myairquality)) > 0]
    clnames
```

```
## [1] "Ozone"   "Solar.R"
```

```
    #create subset of dataframe rows having NAs
    na_myairquality <- myairquality[rowSums(is.na(myairquality)) > 0,]
    # na_myairquality

    #review the columns with distinct values and look for NAs
    sort(unique(myairquality$Ozone),decreasing = FALSE,na.last = FALSE)
```

```
##  [1]  NA   1   4   6   7   8   9  10  11  12  13  14  16  18  19  20  21  22  23
## [20]  24  27  28  29  30  31  32  34  35  36  37  39  40  41  44  45  46  47  48
## [39]  49  50  52  59  61  63  64  65  66  71  73  76  77  78  79  80  82  84  85
## [58]  89  91  96  97 108 110 115 118 122 135 168
```

```
    sort(unique(myairquality$Solar.R),decreasing = FALSE,na.last = FALSE)
```

```
##   [1]  NA   7   8  13  14  19  20  24  25  27  31  36  37  44  47  48  49  51
##  [19]  59  64  65  66  71  77  78  81  82  83  91  92  95  98  99 101 112 115
##  [37] 118 120 127 131 135 137 138 139 145 148 149 150 153 157 167 175 183 186
##  [55] 187 188 189 190 191 192 193 194 197 201 203 207 212 213 215 220 222 223
##  [73] 224 225 229 230 236 237 238 242 244 248 250 252 253 254 255 256 258 259
##  [91] 260 264 266 267 269 272 273 274 275 276 279 284 285 286 287 290 291 294
## [109] 295 299 307 313 314 320 322 323 332 334
```

```
    sort(unique(myairquality$Wind),decreasing = FALSE,na.last = FALSE)
```

```
##  [1]  1.7  2.3  2.8  3.4  4.0  4.1  4.6  5.1  5.7  6.3  6.9  7.4  8.0  8.6  9.2
## [16]  9.7 10.3 10.9 11.5 12.0 12.6 13.2 13.8 14.3 14.9 15.5 16.1 16.6 18.4 20.1
## [31] 20.7
```

```
    sort(unique(myairquality$Temp),decreasing = FALSE,na.last = FALSE)
```

```
##  [1] 56 57 58 59 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
## [26] 82 83 84 85 86 87 88 89 90 91 92 93 94 96 97
```

```
    sort(unique(myairquality$Month),decreasing = FALSE,na.last = FALSE)
```

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

```
    sort(unique(myairquality$Day),decreasing = FALSE,na.last = FALSE)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## [26] 26 27 28 29 30 31
```

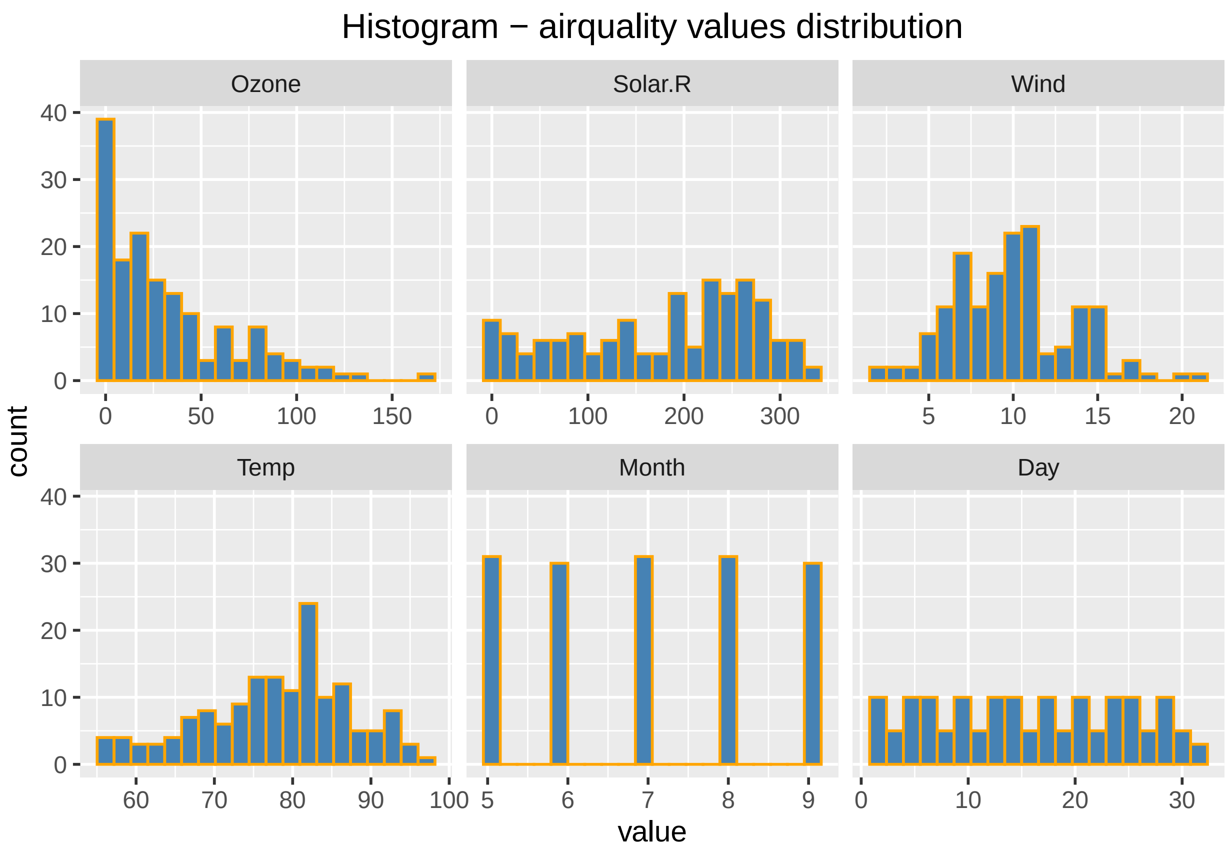
```
    #replace NAs with 0
    myairquality[is.na(myairquality)] <- 0
```

Step 3: Understand the data distribution

```
# Create Histograms for each of the variables

#histogram for all variable

  hcolor <- c("orange")
  hfill <- c("steelblue")
  htitle <- c("Histogram - airquality values distribution")
  theme <-theme(plot.title = element_text(hjust = 0.5),axis.title = element_text())

  gghist <- ggplot(data=melt(myairquality),mapping = aes(x= value))
```

```
## No id variables; using all as measure variables
```

```
  gghist+geom_histogram(bins = 20,color=hcolor,fill=hfill)+facet_wrap(~variable,scales = "free_x")+ ggt
```



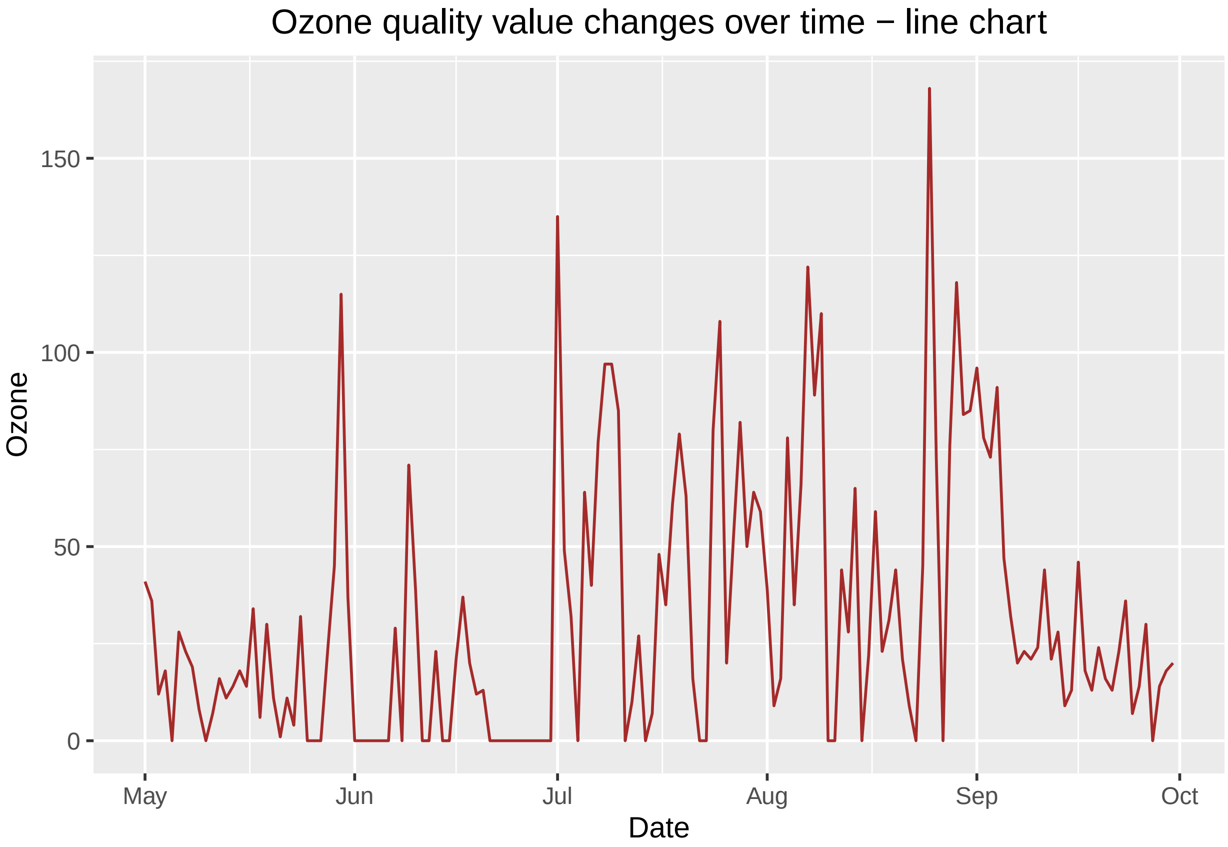
```
myairquality[1,]
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
```

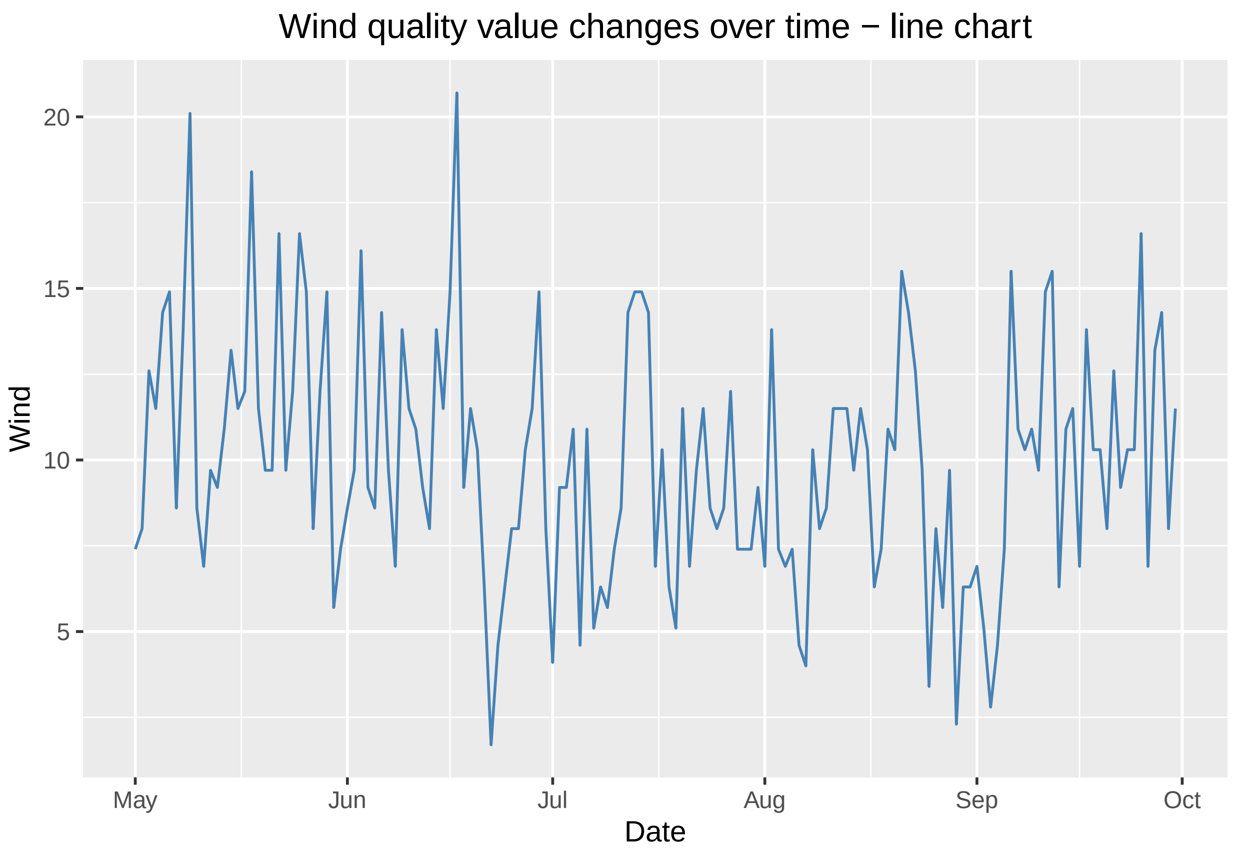
```
myairquality$Date <- as.Date(paste("1973",myairquality$Month,myairquality$Day,sep = "-"))
```

Step 3: Explore how the data changes over time

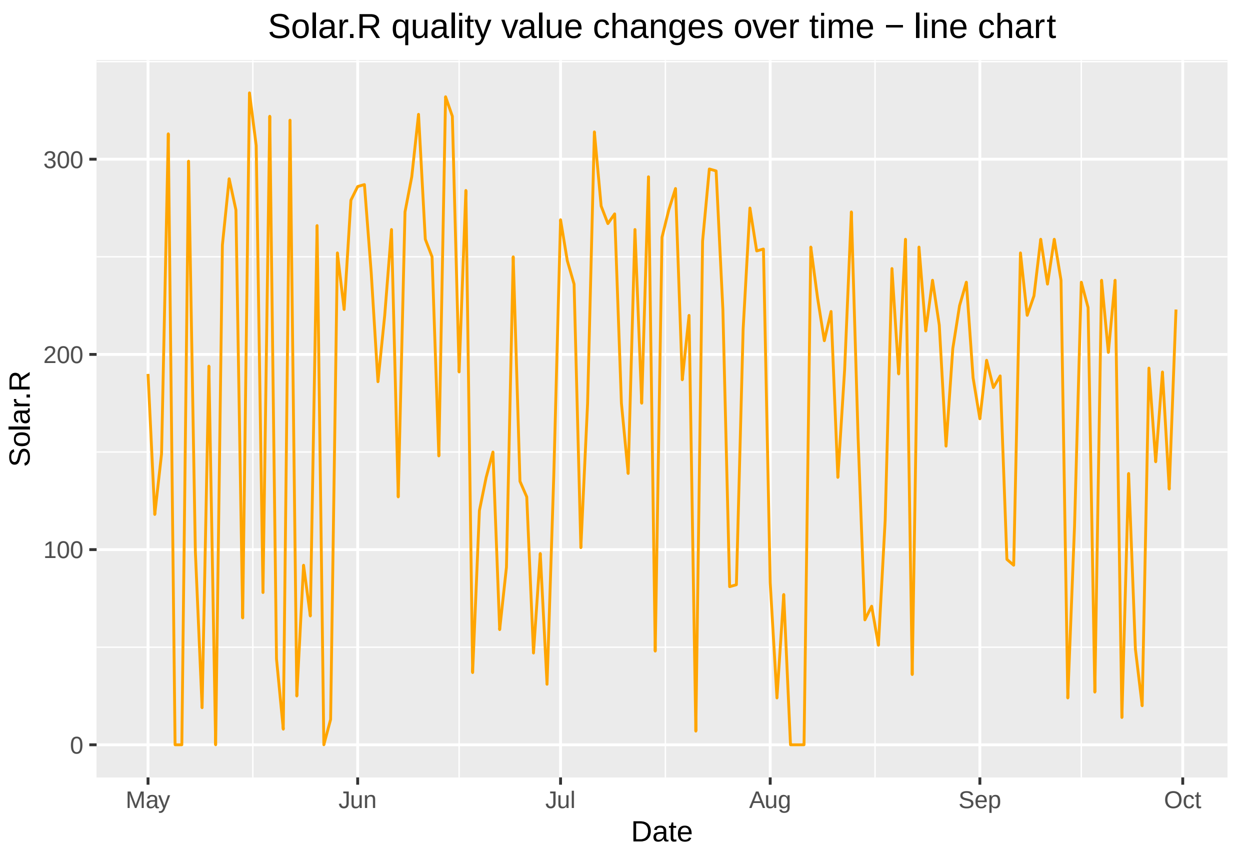
```
# Create line charts for ozone, temp, wind and solar

  glChart <- function(d,xcol,ycol,c,ctitle)
    {
      x <- d[,which(colnames(d)==xcol)]
      y <- d[,which(colnames(d)==ycol)]
      t <- paste(ycol,ctitle)
      ggchart <- ggplot(d,aes(x,y)) + geom_line(color=c)+ ggtitle(t) + xlab(xcol)+ ylab(ycol) + theme
      return(ggchart)
    }

  glChart(myairquality,"Date","Ozone","brown","quality value changes over time - line chart")
```



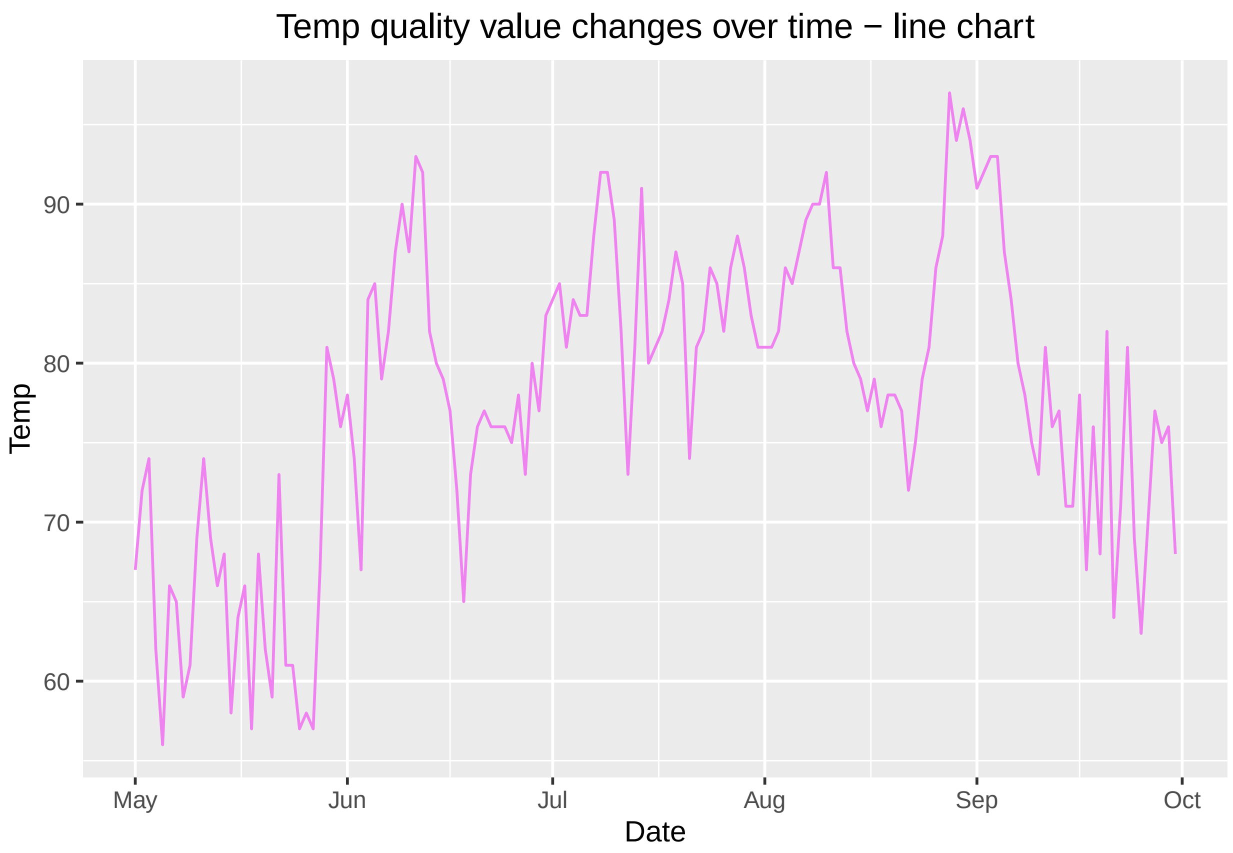
```
glChart(myairquality,"Date","Wind","steelblue","quality value changes over time - line chart")
```



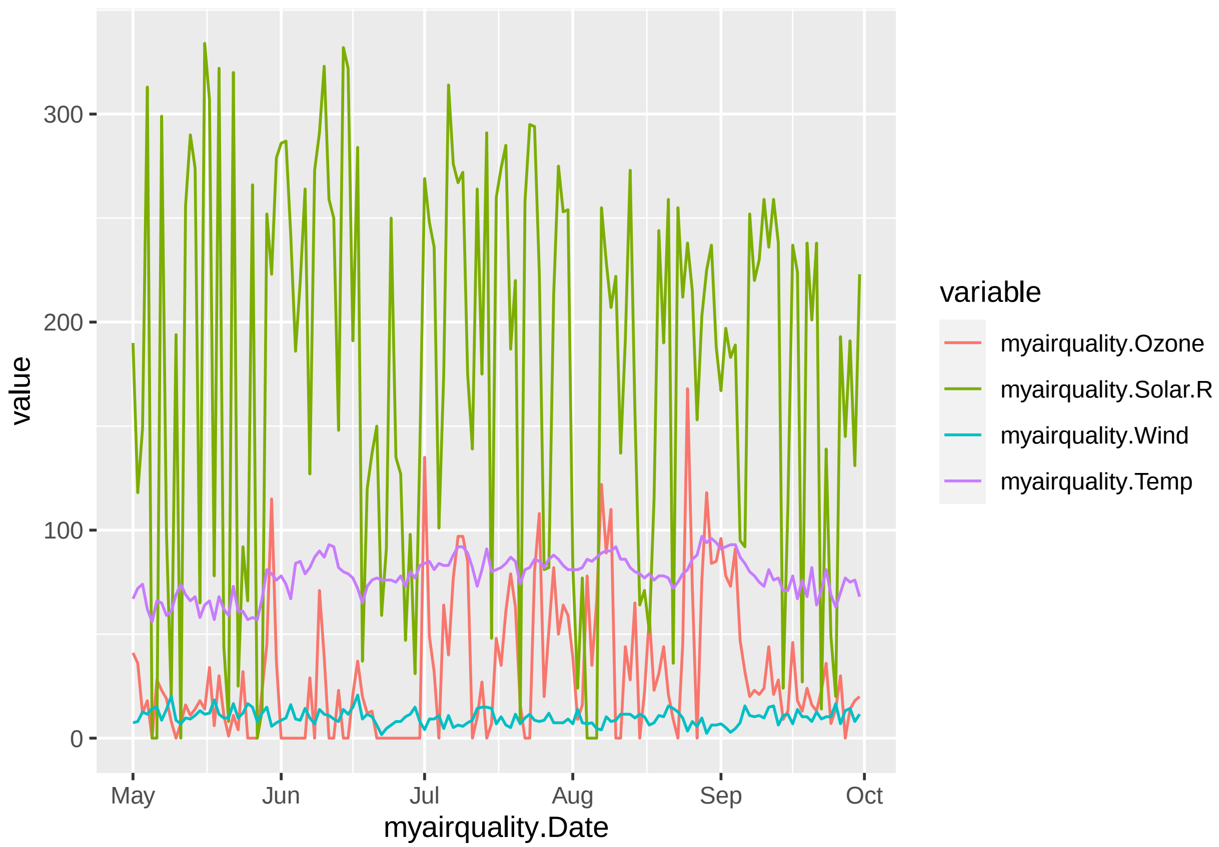
```
glChart(myairquality,"Date","Solar.R","orange","quality value changes over time - line chart")
```



```r
glChart(myairquality,"Date","Temp","violet","quality value changes over time - line chart")
```



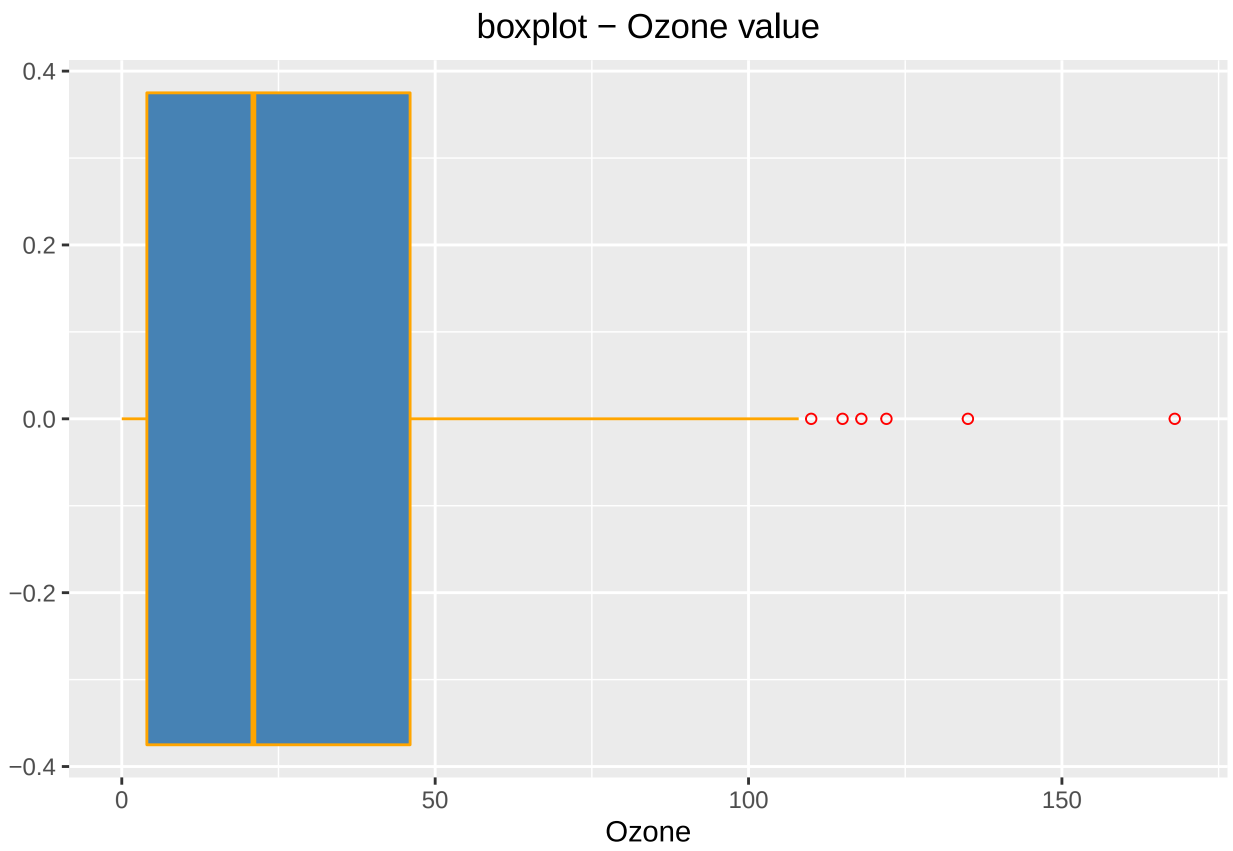
```
# All values in one chart

df <-data.frame(myairquality$Ozone,myairquality$Solar.R,myairquality$Wind,myairquality$Temp,myairqual

df <- melt(df,id=c("myairquality.Date"))
gghist <- ggplot(df,aes(x= myairquality.Date,y=value,color=variable))

gghist+geom_line()
```



```
# Create Boxplot for Ozone

    myairquality[1,]
```

```
##   Ozone Solar.R Wind Temp Month Day       Date
## 1    41     190  7.4   67     5   1 1973-05-01
```

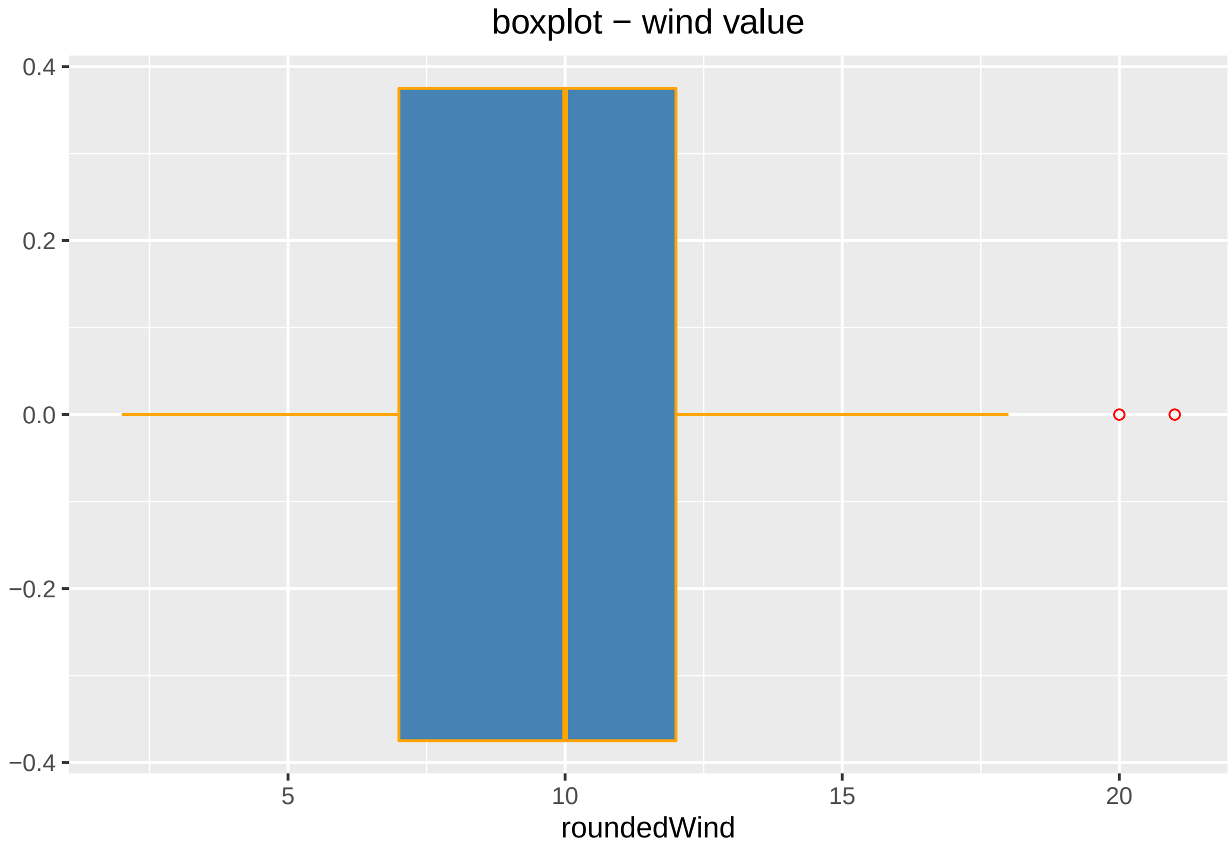
```
    # unique(myairquality$Ozone)

    ggOzoneboxplot <- ggplot(myairquality,aes(Ozone)) +geom_boxplot(fill = "steelblue", colour = "orang

    ggOzoneboxplot+theme
```



```
# Create Boxplot for wind values (round the wind to get a good number of "buckets")

    myairquality[1,]
```

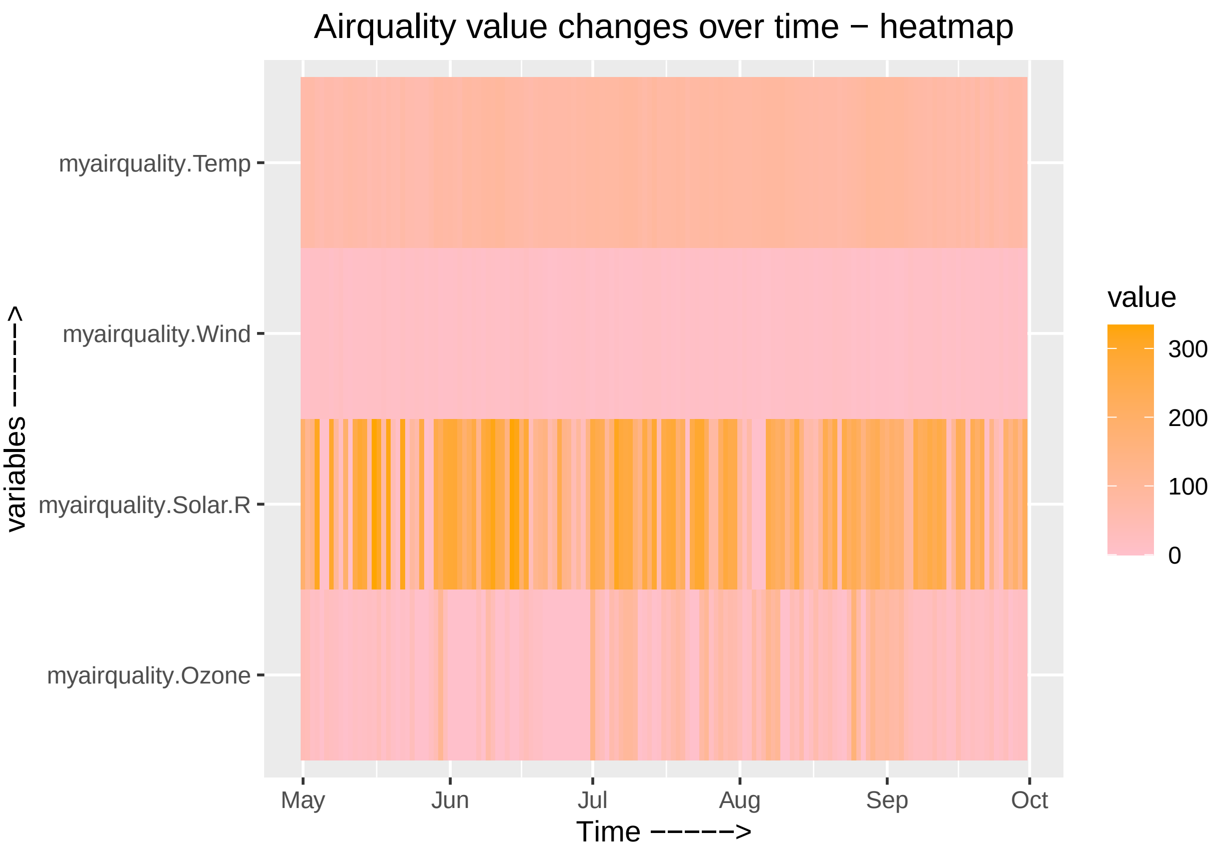
```
##   Ozone Solar.R Wind Temp Month Day       Date
## 1    41     190  7.4   67     5   1 1973-05-01
```

```
    # unique(myairquality$Wind)

    roundedWind <- round(myairquality$Wind)


    ggWindboxplot <- ggplot(myairquality,aes(roundedWind)) +geom_boxplot(fill = "steelblue", colour = "

    ggWindboxplot +theme
```



Step 4: Look at all the data via a Heatmap

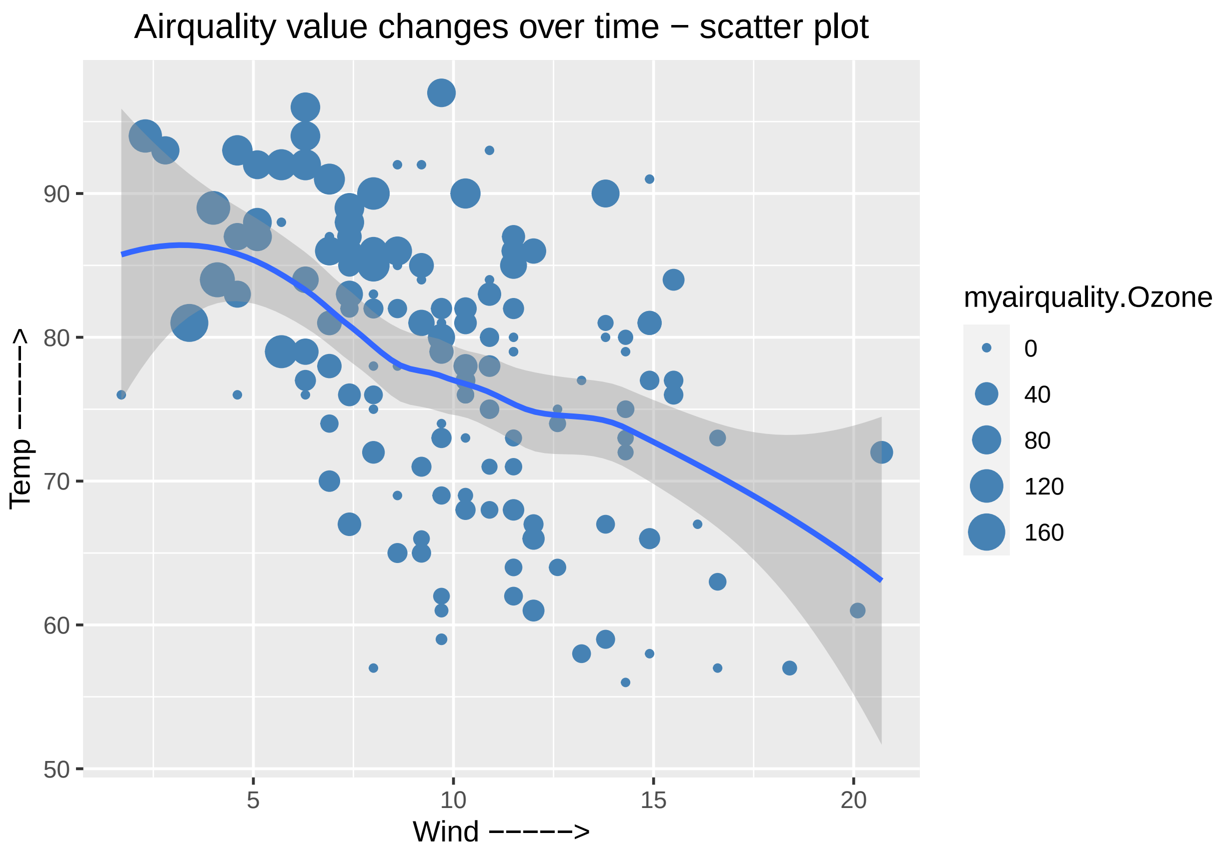
```
dfheat <-data.frame(myairquality$Ozone,myairquality$Solar.R,myairquality$Wind,myairquality$Temp,myair
# df
melted_dfheat <- melt(dfheat,id=c("myairquality.Date"))
# melted_df

gghist <- ggplot(melted_dfheat,aes(x= myairquality.Date,y=variable))

gghist+geom_tile(aes(fill=value))+scale_fill_gradient(low="pink", high="orange")+ggtitle("Airquality 
```



Step 5: Look at all the data via a scatter chart

```
dfscatter <-data.frame(myairquality$Ozone,myairquality$Solar.R,myairquality$Wind,myairquality$Temp,myai

# df

 gghist <- ggplot(dfscatter,aes(x= myairquality.Wind,y=myairquality.Temp))
 gghist+geom_point(color="steel blue",aes(size=myairquality.Ozone,color=myairquality.Solar.R)) +ggtitle
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Step 6: Final Analysis

- Do you see any patterns after exploring the data? • What was the most useful visualization?

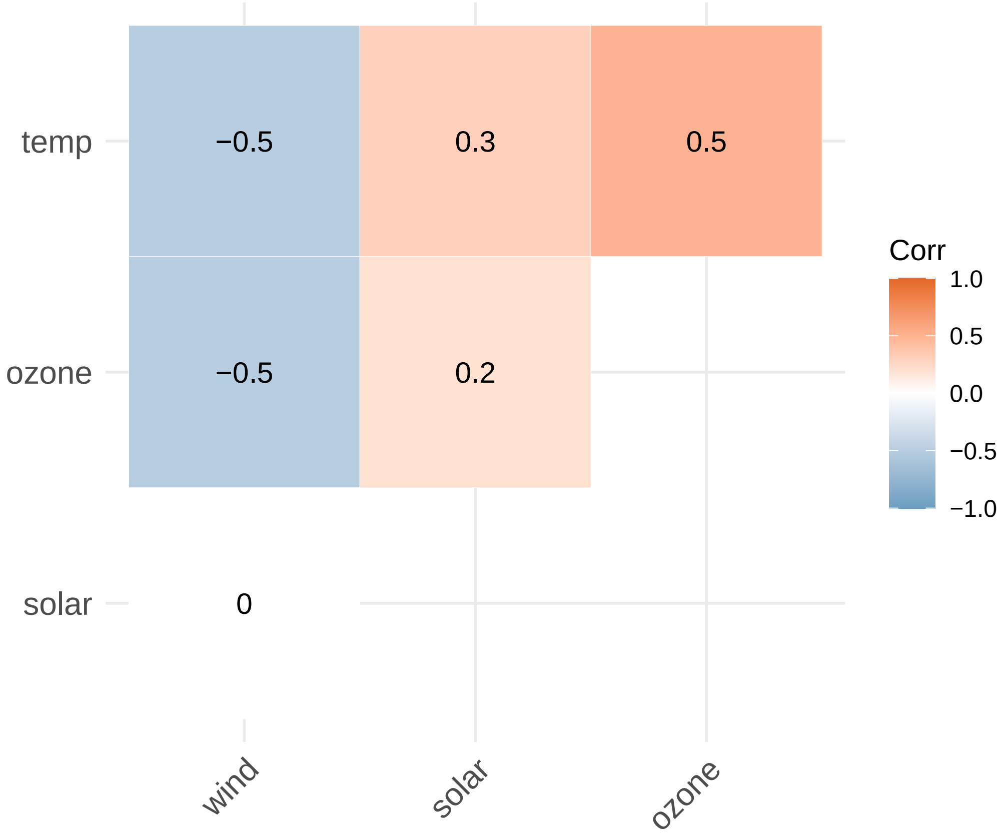
```
dfscorr <-data.frame(myairquality$Ozone,myairquality$Solar.R,myairquality$Wind,myairquality$Temp)

colnames(dfscorr)
```

```
## [1] "myairquality.Ozone"   "myairquality.Solar.R" "myairquality.Wind"
## [4] "myairquality.Temp"
```

```
colnames(dfscorr) <- c("ozone","solar","wind","temp")

corr <- round(cor(dfscorr),1)

head(corr[,1:4])
```

```
##       ozone solar wind temp
## ozone   1.0   0.2 -0.5  0.5
## solar   0.2   1.0  0.0  0.3
## wind   -0.5   0.0  1.0 -0.5
## temp    0.5   0.3 -0.5  1.0
```

```
ggcorrplot(corr, hc.order = TRUE, lab = TRUE, outline.col = "white",type = "upper", colors = c("#6D9EC1
```



```
# • Do you see any patterns after exploring the data?
    # As we can see from the correlation chart below, Temp and Ozone are highly correlated.

# • What was the most useful visualization?

    # I liked almost all the charts; However, BoxPlot, ScatterPlot and Correlation charts are really in
```