PGDM – Analytics
Aparajito Sengupta
IISWBM **Date 7th January 2019**

Dataset: **Air Quality** (R Dataset Package)

Examining the missing values & outliers in the dataset

```
> airquality <- datasets::airquality ##Calling the dataset from R Dataset and
  save it as airquality
```

Data Structure :

```
class(airquality)
[1] "data.frame"  #Data structure
```

Dimension :

```
dim(airquality)
[1] 153   6       #153 observations with 6 variables
```

```
> summary(airquality) #To check the missing value at a glance I prefer to use
  the "Summary" Function which yields a workable result.
```

```
     Ozone          Solar.R          Wind            Temp           Month
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.
000
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.
000
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.
000
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.
993
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.
000
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.
000
 NA's   :37       NA's   :7
      Day
 Min.   : 1.0
 1st Qu.: 8.0
 Median :16.0
 Mean   :15.8
 3rd Qu.:23.0
 Max.   :31.0
```

We can see in the said dataset there are 37+ 7 = 44 missing values .

##Ozone consists of **37** missing values and Solar.R consists of 7 missing values.

Wind, Temp, Month and Day have no missing values.

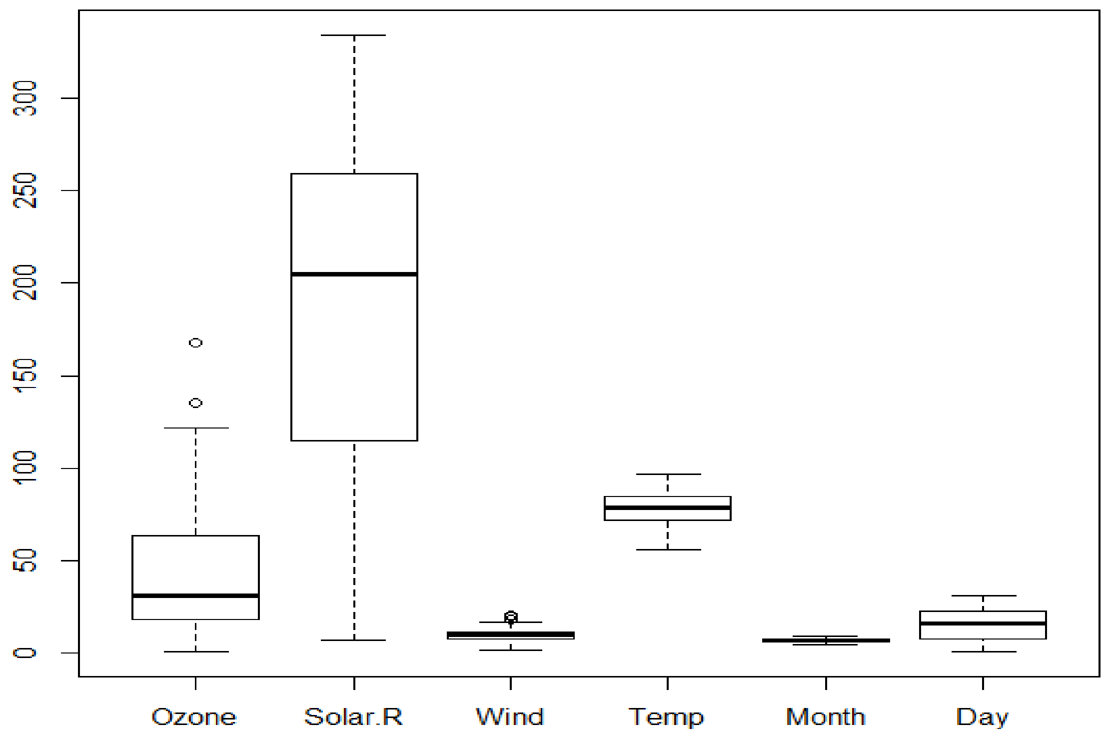
Alternatively , for a comprehensive sum total of missing values (variable wise) we can use “sapply” and “is.na” functions in the following way to get missing values,

```
sapply(airquality,function(x)sum(is.na(x)))
  Ozone Solar.R    Wind    Temp   Month     Day
     37       7       0       0       0       0
```

Outlier detection

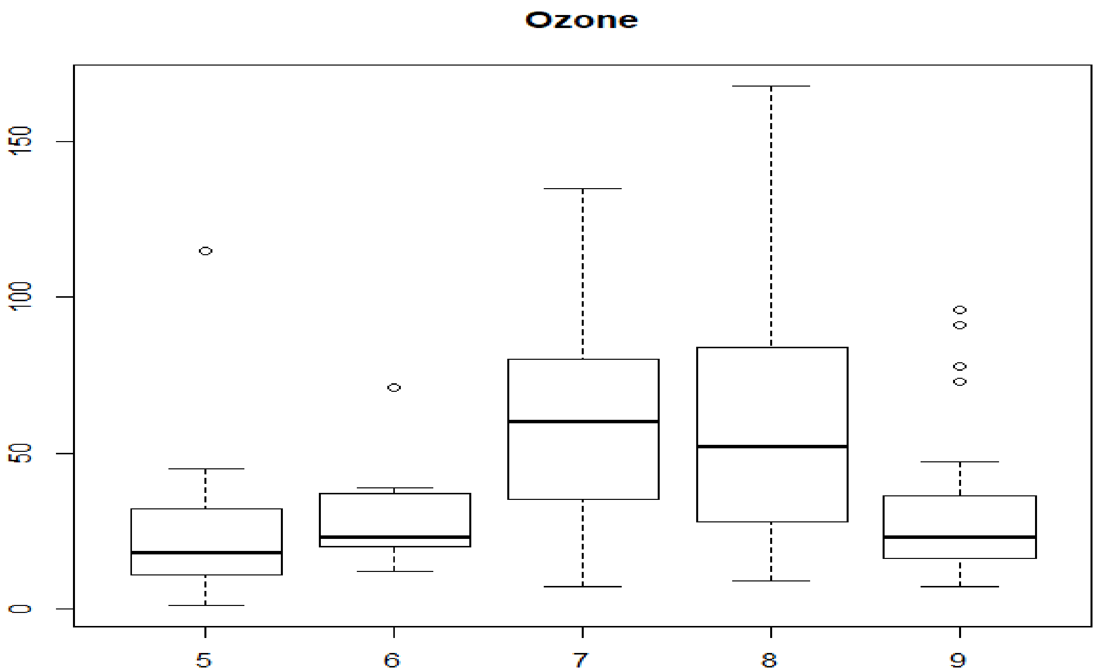
For a given continuous variable, outliers are those observations that lie out side 1.5*IQR (“Inter Quartile Range”) and “Inter Quartile Range” which is the difference between 75th & 25th quartiles.

```
OutVals = boxplot(airquality)$out
```

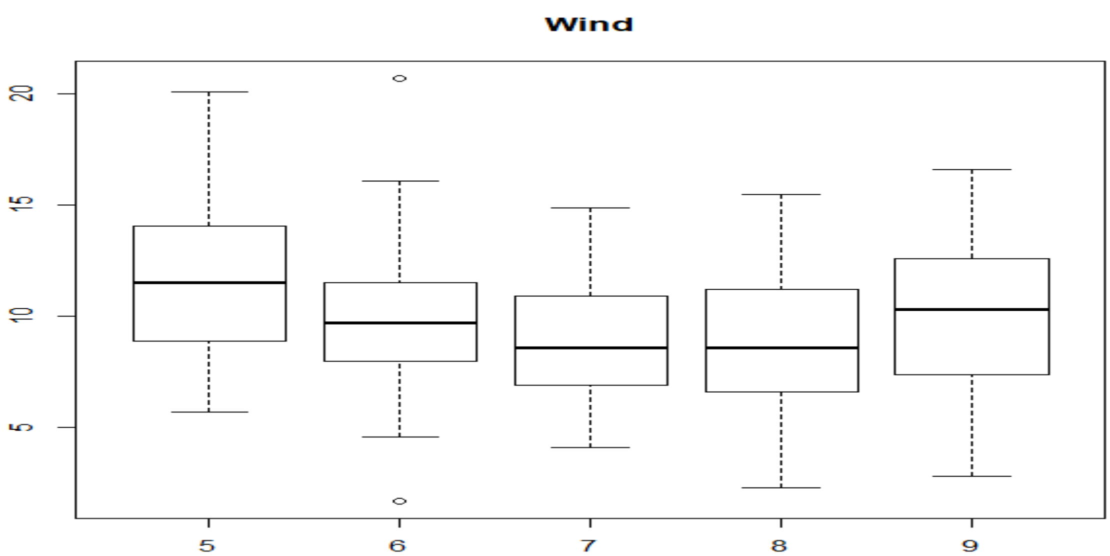


##Clearly visible that Ozone & Wind have outliers. Let us have a close look of these two variables by plotting them individually with respect to months.

```
windows(10,10)
> boxplot(airquality$Ozone ~ airquality$Month , main = "Ozone")
```




```
windows(10,10)
> boxplot(airquality$Wind ~ airquality$Month, main = "Wind")
```



#In the 5th, 6th & 9th month “Ozone” has outliers and in the 6th month “Wind” has outliers.

Finding Pattern in the missing values with respect to Date & Month

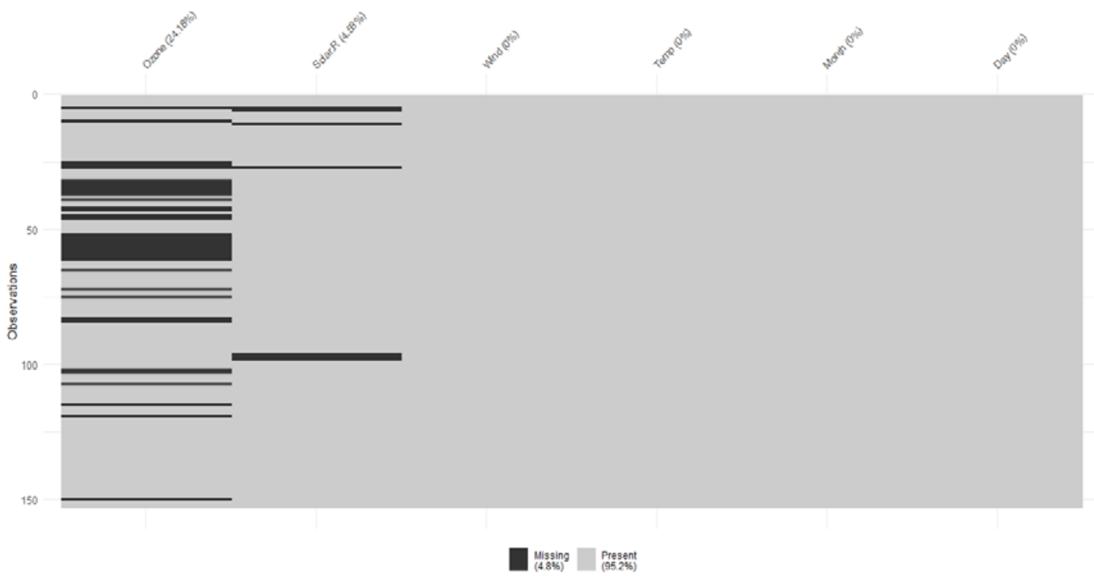
#As early detected Ozone has 37 & Solar.R has 7 missing values

```
> colnames(airquality)[colSums(is.na(airquality)) > 0]
[1] "Ozone"   "Solar.R"
```

#Percentage of Columns & Rows with missing values from these two variables can be identified easily and visually by using ,

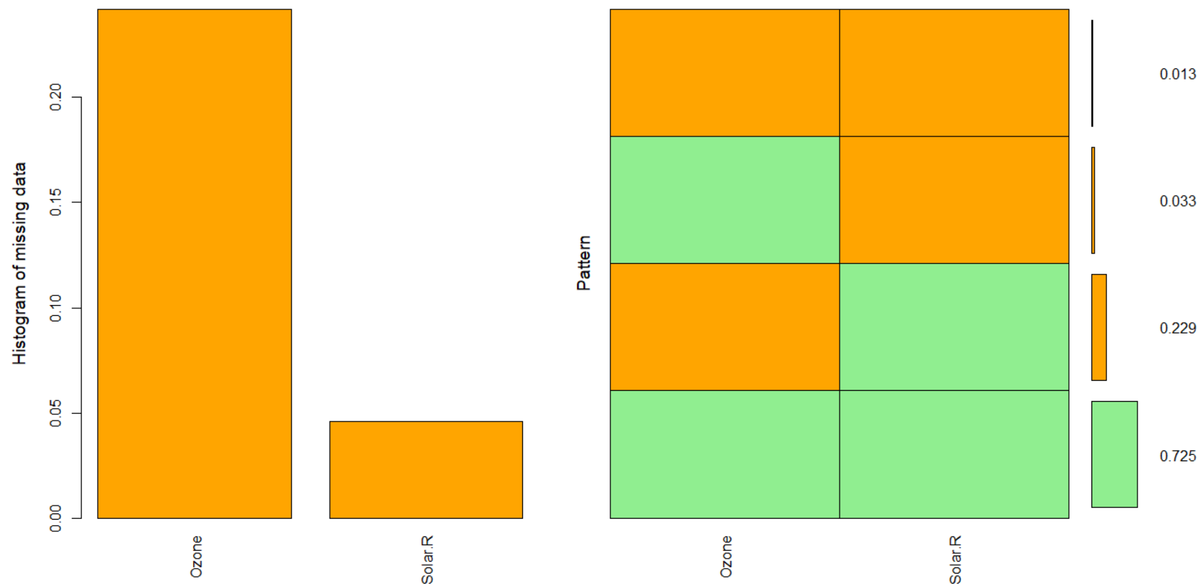
```
>
```

#Which yields the following graph stating 24.18% Ozone data & 4.58% of the Solar.R are missing values :



Another way to check it with a comprehensive plot with the package VIM.

```
> library('VIM')
> windows(10,10)
> aggr_plot <- aggr(airquality[c(1,2)],col =c('light green','orange'),numbers
= TRUE,sortVars = TRUE, ylab=c("Histogram of missing data","Pattern"))

 Variables sorted by number of missings:
 Variable      Count
    Ozone 0.24183007
  Solar.R 0.04575163
```

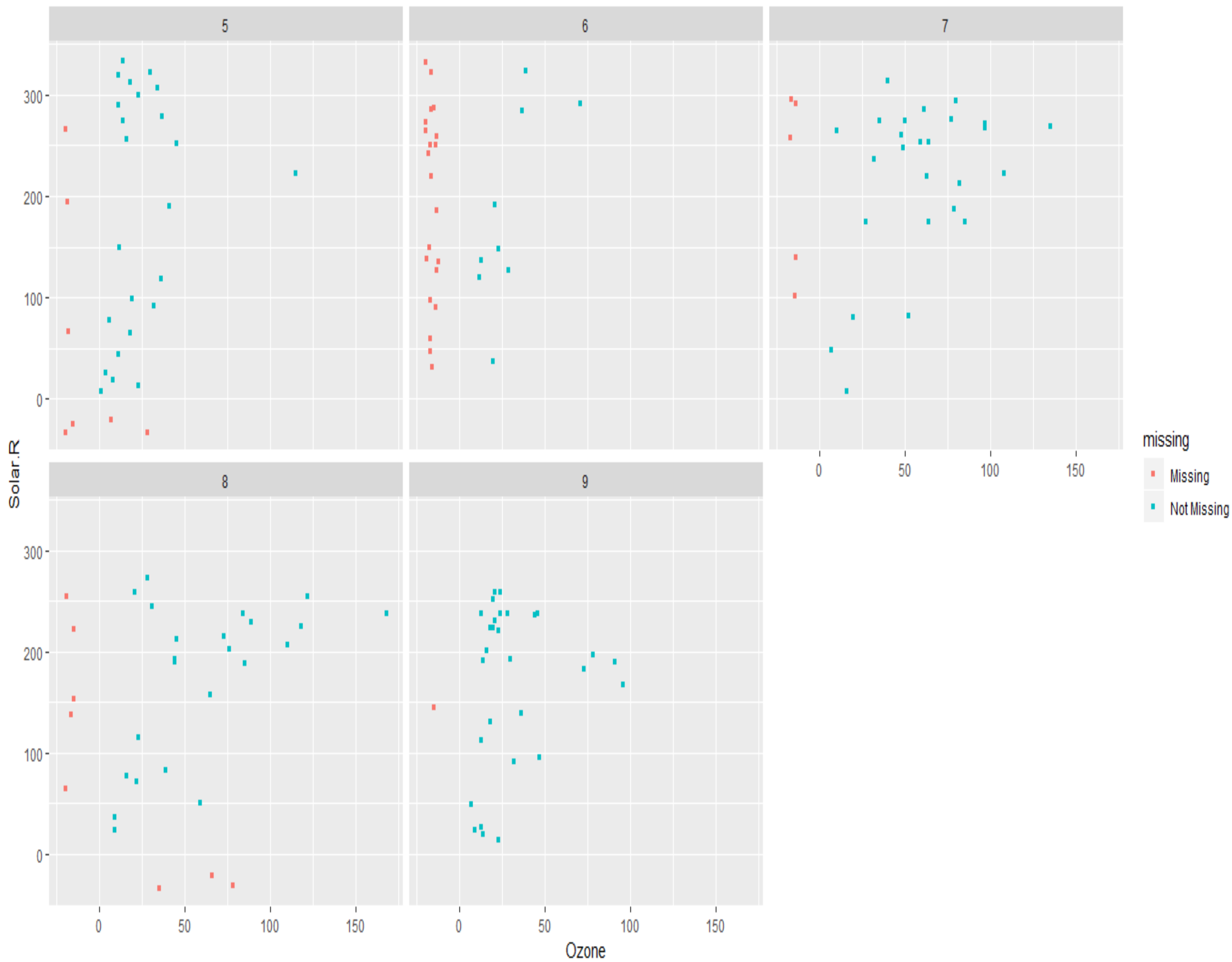


#This graph clearly makes us know the missing percentage of data . It says only 1% of data for Ozone & Solar. R is jointly missing. 3.3% of Solar.R data is missing when the same is available for Ozone 22.9% data is missing for Ozone when the same is available for Solar.R. And 72.5% data is available for both of these variables.

#and rest of the variables have no missing value.Now, I would like to see the pattern of missing values of these two variables together with respect to date & month.

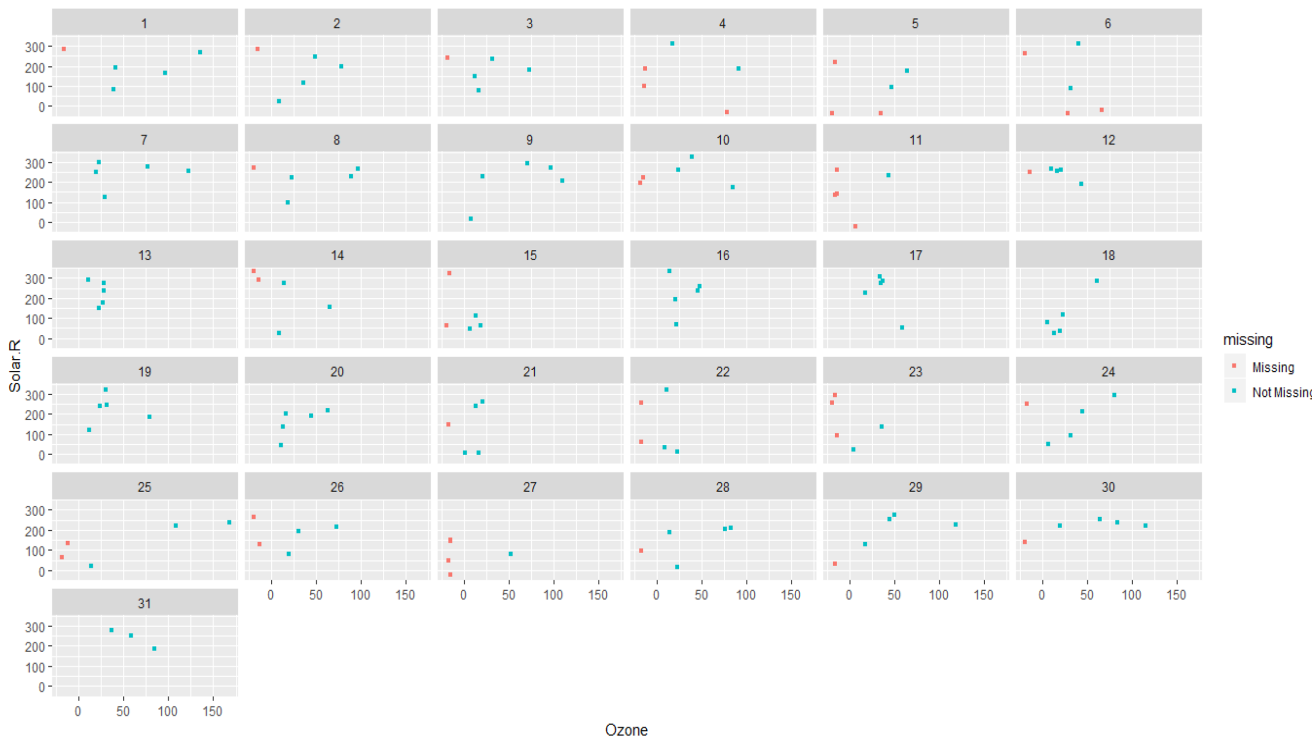
#Let us see how the graph looks like if we plot Solar.R in Y - axis and Ozone in X – axis wrt. Month using ggplot,

```
> windows(10,10)
> ggplot(airquality,aes(x = Ozone, y = Solar.R))+naniar::geom_miss_point()+fa
cet_wrap(~ Month)

#Sir, lots of hard work put-in to get this graph. I learned about lots of pac
kages to plot missing values, like “naniar”, “mice”,”VIM” , geo_miss_plot onl
y to do this exercise]

#It is visible from the plot that in the month of “June” the proportion of mi
ssing value is highest and in September it is least.
```



#To check missing values with respect to days I will use the same function changing the facet_wrap to (~ Day),

```
> windows(10,10)
> ggplot(airquality,aes(x = Ozone, y = Solar.R))+naniar::geom_miss_point()+fa
  cet_wrap(~ Month) + facet_wrap(~ Day)#The resulting plot is like,
```



#Though there is no specific patterns found in day wise missing values but still 4, 5,6 ,11,23,27 are the dates where we can find the little surge in missing values. June is

1.C) Justify your decision with respect to the treatment to missing values strategy.

We can see that the data missing completely at random in the given dataset. But too much missing data is a problem for further analysis. Considering 5% missing data as threshold let us check Ozone & Solar.R stand where.

From my early computation :

```
Variables sorted by number of missings:
 Variable       Count
    Ozone 0.24183007 = 24%
    Solar.R 0.04575163 = 4.6%
```

But if we drop "ozone" then it is a huge data loss and we have not yet verified Ozone's impact on Temperature change & vice versa , Solar.R 's effect on Ozone and Ozone's effect on wind.

It won't be a prudent decision to remove Ozone's missing values.

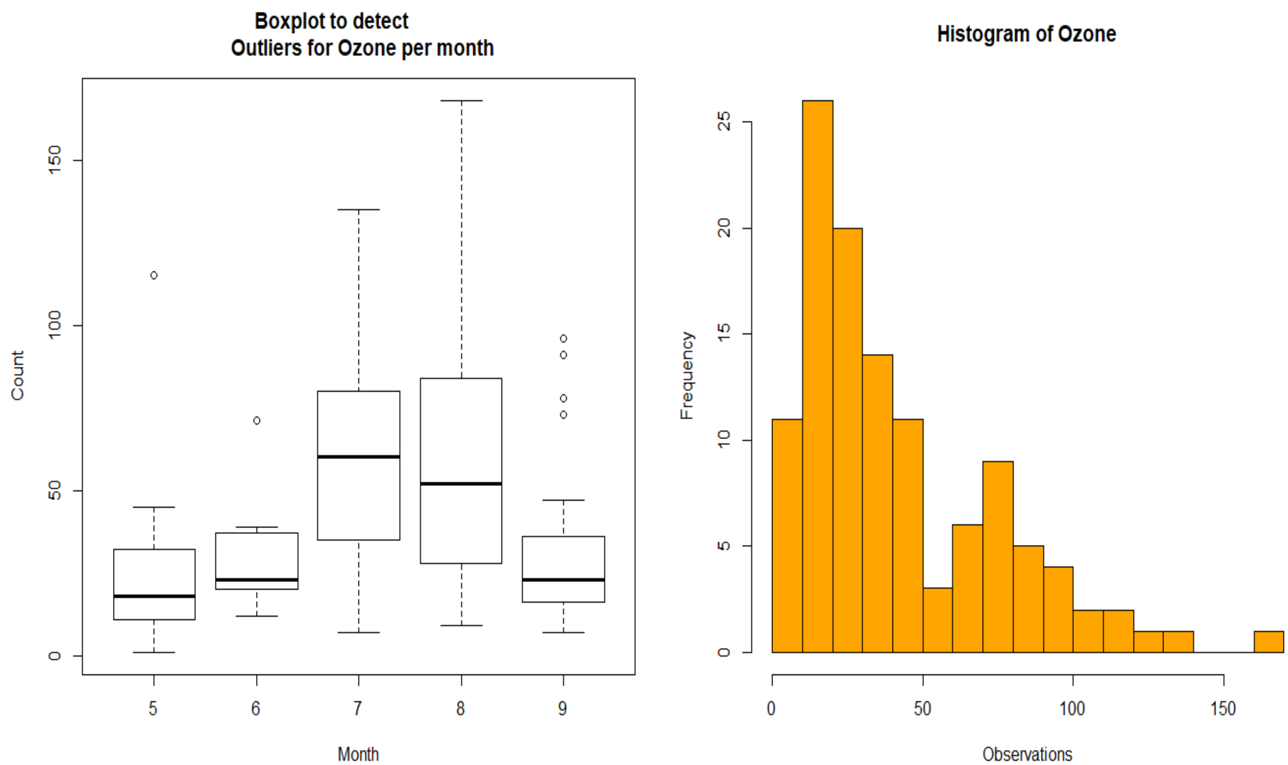
I would rather go for a mean replacement for Ozone & Solar.R.

```
#Replacing the missing values in Ozone & Solar.R columns by their respective
Mean values,
```

```
> airquality$Ozone <- ifelse(is.na(airquality$Ozone),mean(airquality$Ozone,na
.rm=TRUE),airquality$Ozone)

> airquality$Solar.R <- ifelse(is.na(airquality$Solar.R),mean(airquality$Sola
r.R, na.rm = TRUE),airquality$Solar.R)

> summary(airquality)
     Ozone            Solar.R           Wind              Temp            Month
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.
000
 1st Qu.: 21.00   1st Qu.:120.0   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.
000
 Median : 42.13   Median :194.0   Median : 9.700   Median :79.00   Median :7.
000
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.
993
 3rd Qu.: 46.00   3rd Qu.:256.0   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.
000
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.
000
      Day
 Min.   : 1.0
 1st Qu.: 8.0
 Median :16.0
 Mean   :15.8
 3rd Qu.:23.0
 Max.   :31.0
```

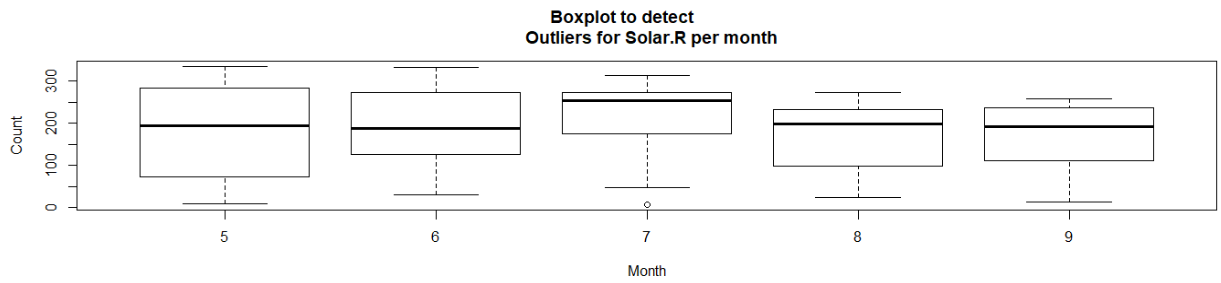
The presence of outliers for each variables

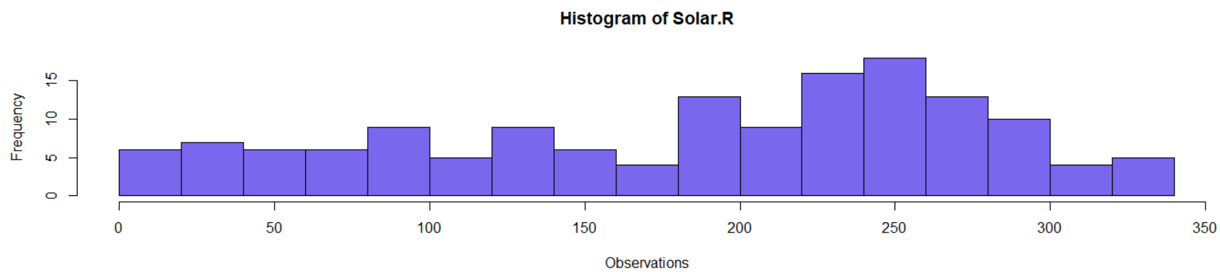
```
#For OZONE OUTLIERS PER MONTH :
> windows(8,8)
> par(mfrow = c(1,2))
> boxplot(airquality$Ozone ~ airquality$Month, main = "Boxplot to detect
+   Outliers for Ozone per month",xlab = "Month",ylab = "Count")
> hist(airquality$Ozone,col = 'orange',main = "Histogram of Ozone",
xlab = "Observations",breaks = 20)
```



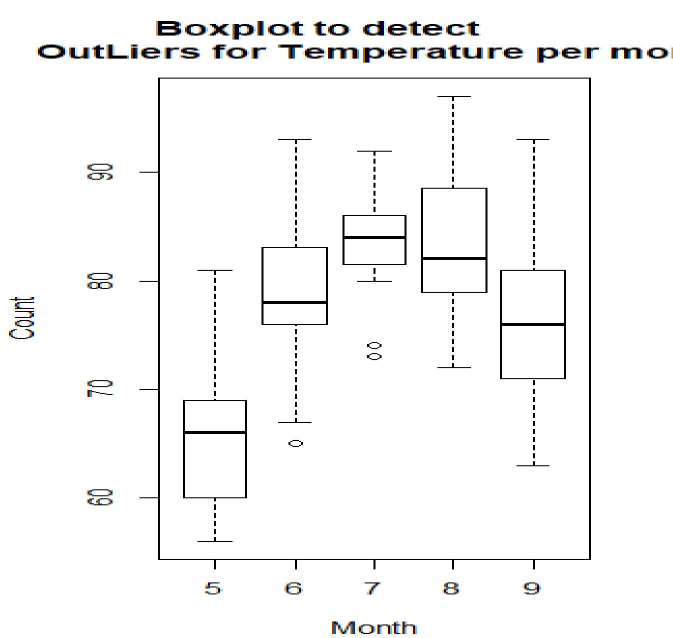

```
#There is 0 observation below the 1st Quantile.
#There are 2 observations above 3rd Quantile.
#Month wise outliers above 3rd Quantile are as follows May(5) = 1,June(6) = 1,
Sept(9) = 4,July & Aug no outliers.
```

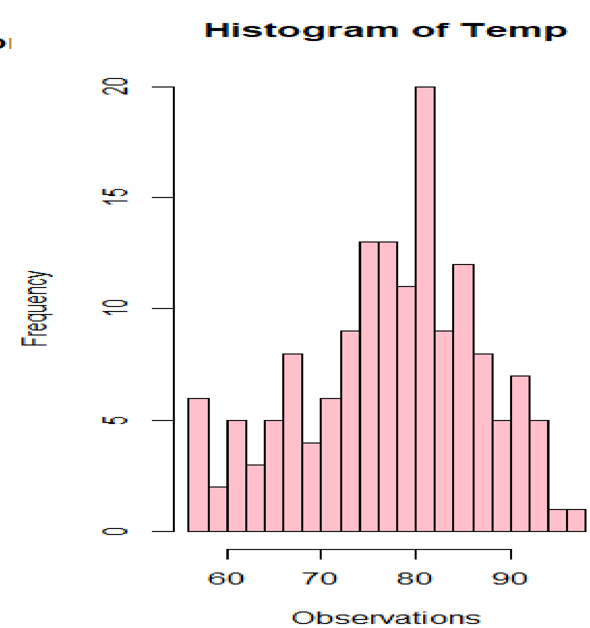
```
#FOR SOLAR.R
windows(8,8)
> boxplot(airquality$Solar.R~airquality$Month, main = "Boxplot to detect
+   OutLiers for Solar.R per month",xlab = "Month",ylab = "Count")
```



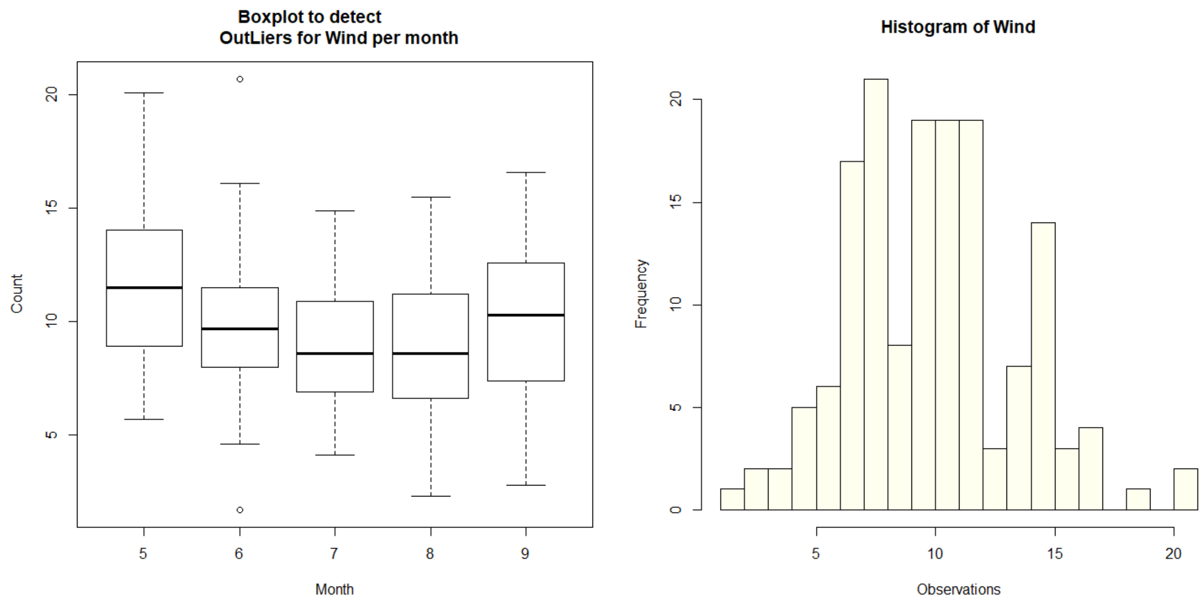



```
#There is 0 observation below 1st Quantile & 0 observation above 3rd Quntile.

#FOR TEMPERATURE:
windows(8,8)
par(mfrow = c(1,2))
> boxplot(airquality$Temp ~ airquality$Month, main = "Boxplot to detect
+ OutLiers for Temperature per month",xlab = "Month",ylab = "Count")
> hist(airquality$Temp,col = "pink",main = "Histogram of Temp",
xlab = "Observations",breaks = 15)
```





```
#Though in June & July there are 1 & 2 outliers existing below 1st quantile bu
t overall there is no outlier below the 1st quantile or above the 3rd quantile.

#FOR WInd:

windows(8,8)
> par(mfrow = c(1,2))
> boxplot(airquality$Wind ~ airquality$Month, main = "Boxplot to detect
+   OutLiers for Wind per month",xlab = "Month",ylab = "Count")
> hist(airquality$Wind,col = "ivory",main = "Histogram of Wind", xlab = "Obse
rvations",breaks = 15)
```




```
# There are two outliers visible one is below the 1st quantile and one is
above the 3rd quantile in the month of June and overall.
```

Patterns in the outliers

As mentioned above these are the following observations,

For Ozone,

```
#There is 0 observation below the 1st Quantile.
#There are 2 observations above 3rd Quantile.
#Month wise outliers above 3rd Quantile are as follows May(5) = 1,June(6) = 1,
Sept(9) = 4,July & Aug no outliers.
```

For Solar R

```
#There is 0 observation below 1st Quantile & 0 observation above 3rd Quntile.
```

For Temperature

#Though in June & July there are 1 & 2 outliers existing below 1st quantile but overall there is no outlier below the 1st quantile or above the 3rd quantile.

For Wind

There are two outliers visible, one is below the 1st quantile and one is above the 3rd quantile in the month of June and overall.

The ouliers of Ozone and Wind are quite random in nature. Where out value is attributed by June's observation to Wind where as for Ozone it is mainly due to September.

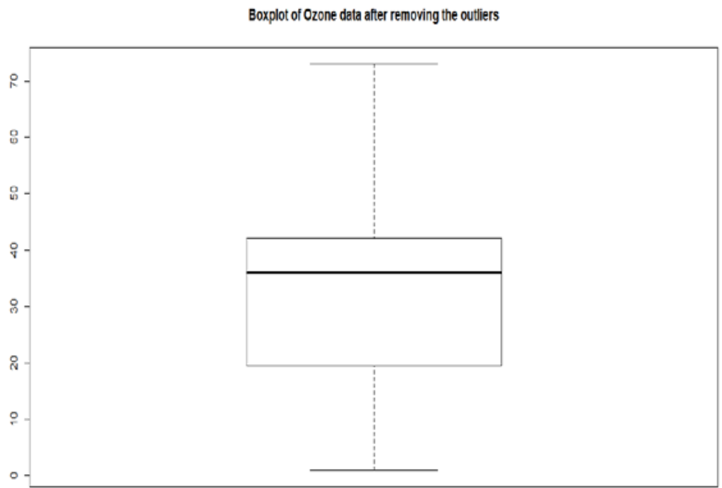
Justifying my decision with respect to the treatment of outliers

#Since Ozone & Wind variables are having outliers I would like to remove them from the dataset.

#For Ozone

```
> min(boxplot.stats(airquality$Ozone)$out)

  [1] 84
> summary(airquality$Ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   21.00   42.13   42.13   46.00  168.00
> summary(airquality$Ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   21.00   42.13   42.13   46.00  168.00
> #and IQR of Ozone is,
> IQR(airquality$Ozone)
[1] 25
> #Thus Limiting Value of Outliers to be
> Ozone_Limiting_Value <- 46 + 1.5*25
> Ozone_Limiting_Value
[1] 83.5

> airquality <- airquality[airquality$Ozone <= 83.5,]
> boxplot(airquality$Ozone,horaizonta = T)
> boxplot(airquality$Ozone)
#Now let us see what is the summary of Ozone now,
> summary(airquality$Ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   20.00   39.00   35.11   42.13   82.00
> #But there are still the outvalues.Let us repeat the process once again,
> IQR(airquality$Ozone)
[1] 22.12931
> Ozone_Limiting_Value2 <- 42.13 +1.5* 22.13
> Ozone_Limiting_Value2
[1] 75.325
> #Let us resrict the limiting value to 75 and repeat the process
> airquality <- airquality[airquality$Ozone <= 75,]
> boxplot(airquality$Ozone)
> windows(10,10)
```

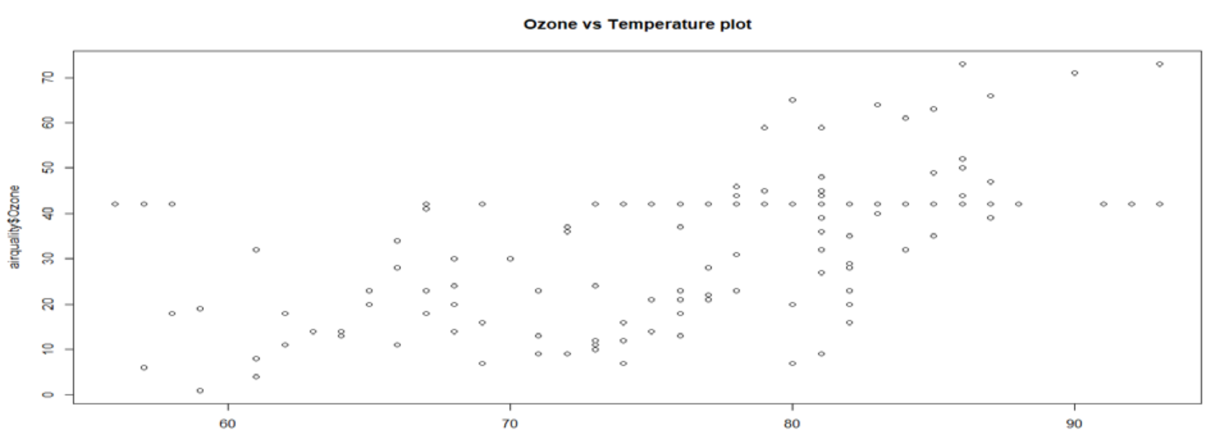
```
> boxplot(airquality$Ozone, main = "Boxplot of Ozone data after removing the
outliers")
> summary(airquality$Ozone)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   19.50   36.00   32.79   42.13   73.00
```




```
#Let us also remove the outliers of wind variable,

> summary(airquality$Wind)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.70    8.00   10.30   10.58   12.60   20.70
> IQR(airquality$Wind)
[1] 4.6
> wind_limiting_value <- 12.6+ 1.5*IQR(airq$Wind)
> wind_limiting_value
[1] 19.5
> airquality <- airqualty[airquality$Wind < 19.5,]

> boxplot(airquality$Wind) # There are some outliers visible let us repeat
the process again.

> summary(airq$Wind)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.70    8.00   10.30   10.43   12.00   18.40

> wind_limiting_value2 <- 12+ 1.5*IQR(airquality$Wind)
> wind_limiting_value2
[1] 18
> wind_limiting_value3 <- 12- 1.5*IQR(airquality$Wind)
> win_lim2
[1] 6
> airquality <- airquality[airquality$Wind < 18,]

> airquality <- airquality[airquality$Wind > 6,]

> boxplot(airquality$Wind) #There are no outliers anymore
```

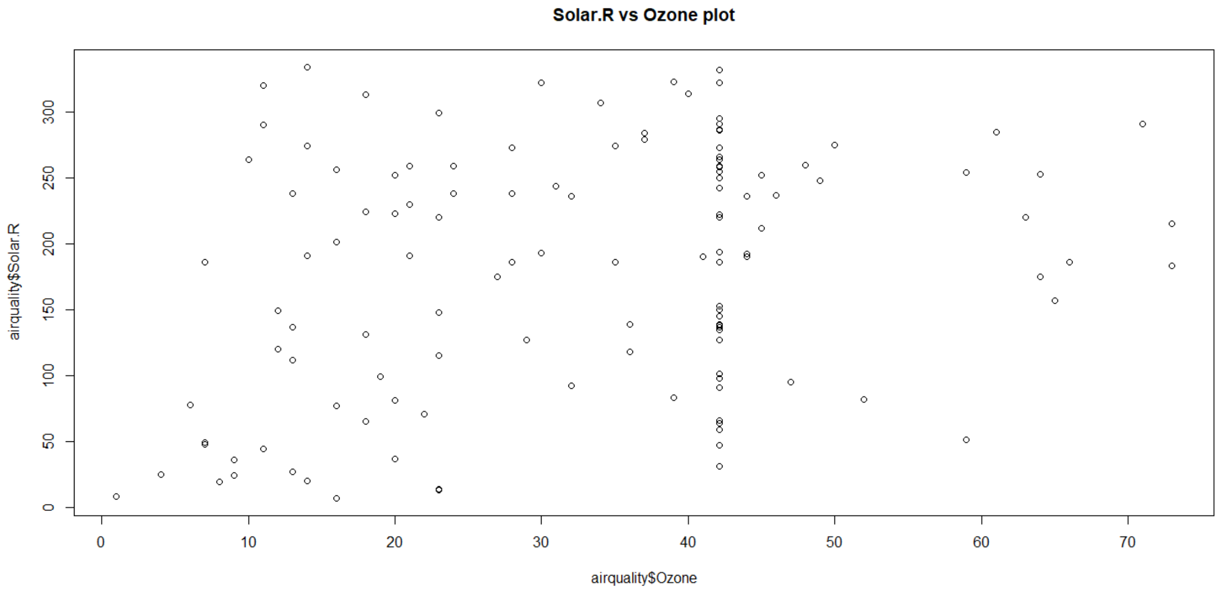
Examining the correlation among temperature, ozone, wind & solar radiation

```
> library(lattice)
> library(survival)
> library(Formula)
> library(ggplot2)
> install.packages("Hmisc")

> airquality_cor <- rcorr(as.matrix(airquality[,c(1,2,3,4)]),type =
"pearson")
> airquality_cor

##Correlation Coefficients among variables,

        Ozone Solar.R  Wind  Temp
Ozone    1.00    0.29 -0.33  0.54
Solar.R  0.29    1.00  0.03  0.22
Wind    -0.33    0.03  1.00 -0.34
Temp     0.54    0.22 -0.34  1.00

n= 131

##P Values among variables,

P
        Ozone  Solar.R Wind   Temp
Ozone          0.0006  0.0001 0.0000
Solar.R 0.0006         0.7309 0.0119
Wind    0.0001 0.7309         0.0000
Temp    0.0000 0.0119  0.0000


### Here we can see that there is a positive correlation between,
a)Ozone and temperature (Ozone increases Temperature increases , strong
association)
> windows(10,10)
>plot(airquality$Ozone ~ airquality$Temp, main = "Ozone vs Temperature plot")
```

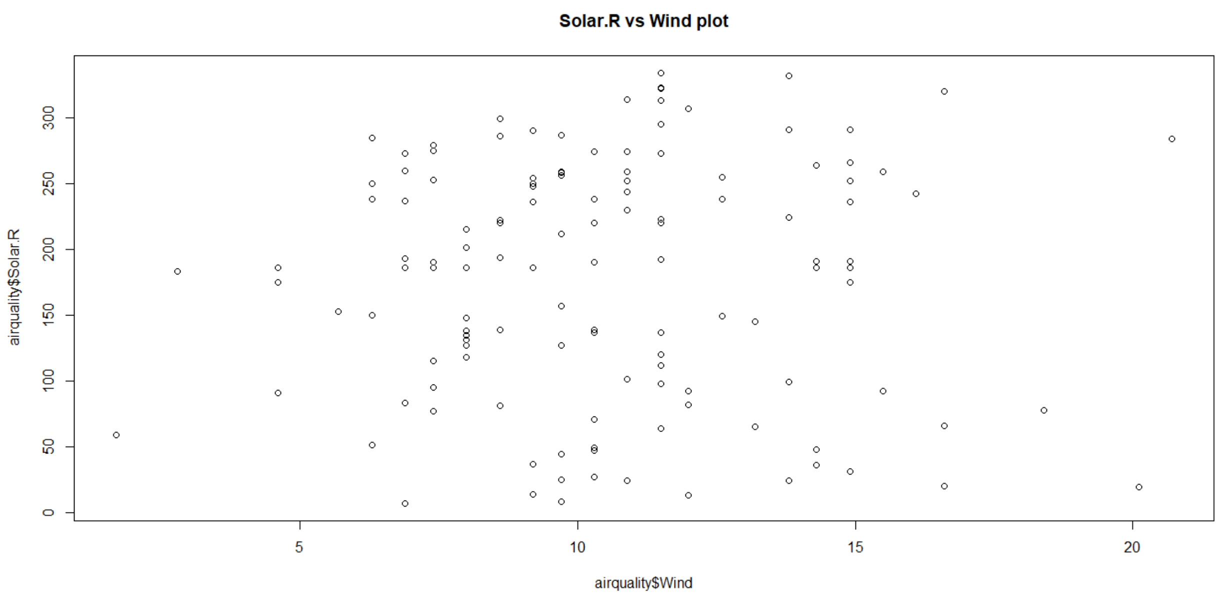



```
b)Solar.R & Ozone (Solar Radiation increases and Ozone increases moderately)
```

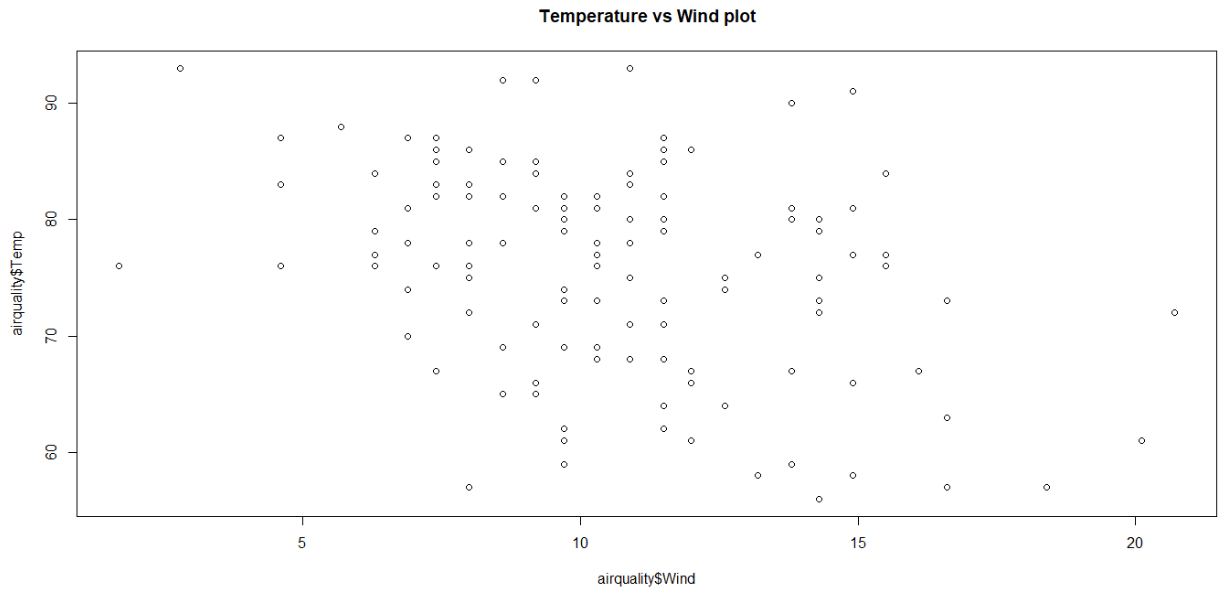
```
>plot(airquality$Solar.R ~ airquality$Ozone, main = "Solar.R vs Ozone plot")
```



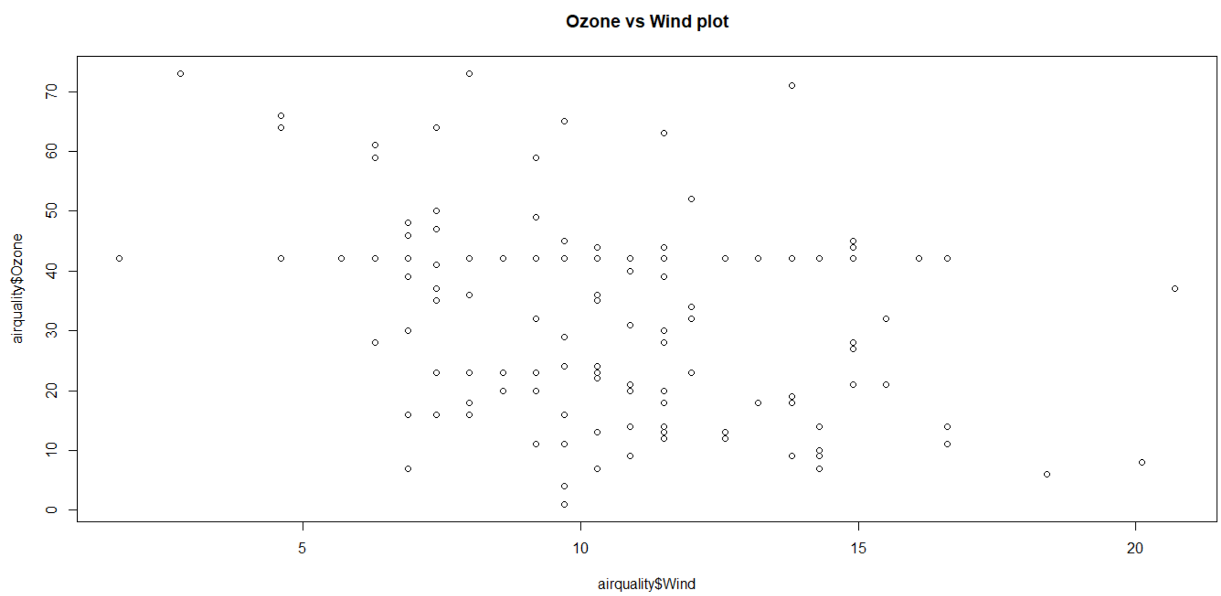
c)Solar.R & Wind (Solar Radiation & wind have very weak correlation)



And negative Correlation between,
d)Temp and wind (Temp increases & wind decreases moderately)



e)Ozone & wind (Ozone increases Wind decreases)



##The other Correlation coefficients among variables are very small and hence can be considered negligible.

Computing whether mean temperature across the months are significantly different from each other

To compute the mean temperature in each month I am using taaply function:

```
> tapply(airquality$Temp,airquality$Month, mean)
       5        6        7        8        9
```

```
  65.10000 79.10000 82.27273 81.18182 75.22222
```

#Mean Temperature Across months are 65.1(May), 79.1(June), 82.27(July), 81.18 (Aug), 75.22(Sept)

To check whether the mean temperatures per month are significantly different or not from each other lets perform One Way ANOVA (Analysis of variance)

#Let us make the hypothesis as

H_0 (Null Hypothesis) : "The mean temperature from the different months are same (or not significantly different from each other)".

H_1 (Alternative Hypothesis) :"The mean temperatures are significantly different from each other with respective to the respective months(Or atleast one of them is significantly different)".

##Subsetting the data set with respect to Temperature and Month

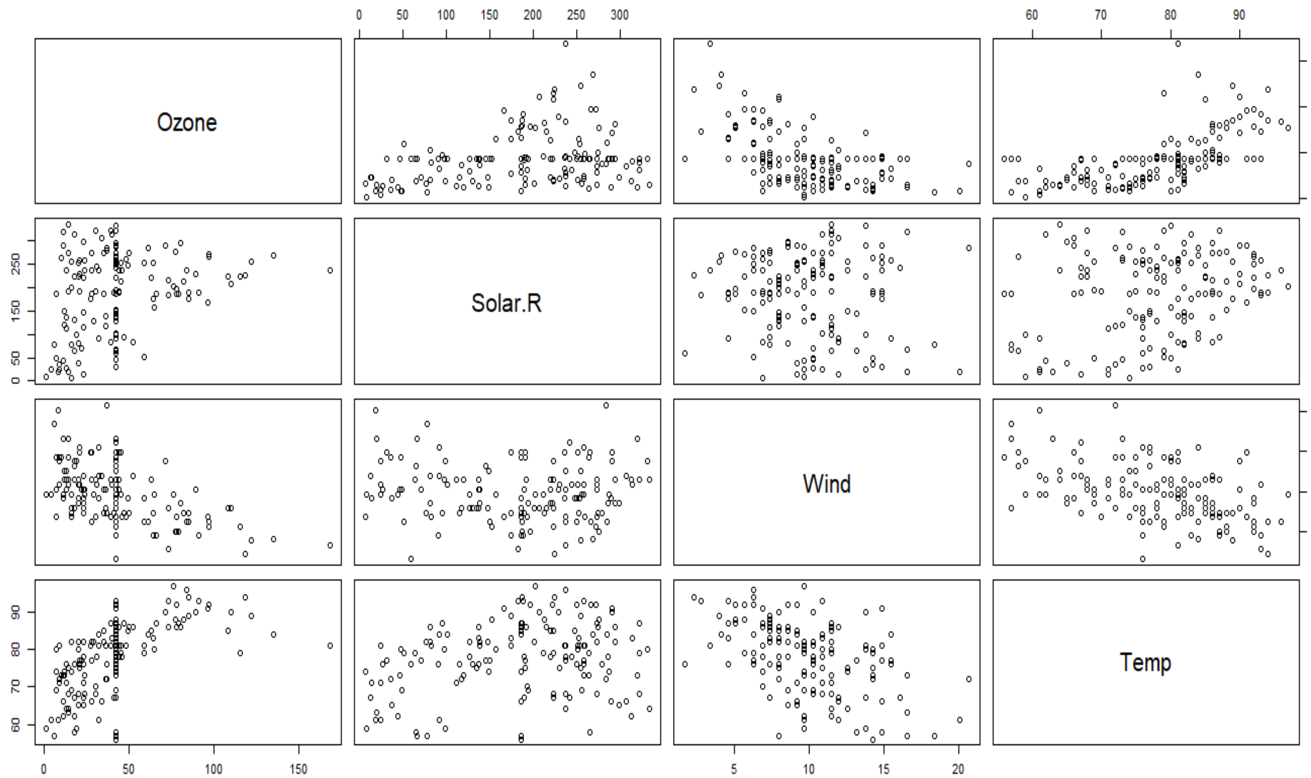
```
> airquality_temp_mnth <- airquality[,c(4,5)]


> Anova_Result <- aov(airquality_temp_mnth$Temp ~ airquality_temp_mnth$Month,
data = airquality_temp_mnth)

> summary(Anova_Result)
                             Df Sum Sq Mean Sq F value   Pr(>F)    
airquality_temp_mnth$Month   1   1437  1437.3   21.98 6.89e-06 ***
Residuals                  129   8434    65.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since we are getting a pretty lower value than the 5% significance level we are rejecting the null hypothesis and accepting the alternative hypothesis which says "
"The mean temperatures are significantly different from each other with respective to the respective months(Or at least one of them is significantly different)".

Examining whether ozone, wind and solar radiation plays any significant role in explaining the temperature variation across months

#Let us check the same graphically with the pair function

```
>airquality <- airquality[,c(1:4)]
> windows(10,10)
> pairs(airquality)
```



#And checking our previously calculated Correlation Cofficient data we can conclude Ozone has a significant impact on Temperature and Solar.R is also having a small positive impact on Temperature but Temperature is having negative correlation with Wind

```
>airquality_cor <- rcorr(as.matrix(airquality[,c(1,2,3,4)]),type = "pearson")
> airquality_cor

##Correlation Coefficients among variables,

        Ozone Solar.R  Wind  Temp
Ozone    1.00    0.29 -0.33  0.54
Solar.R  0.29    1.00  0.03  0.22
Wind    -0.33    0.03  1.00 -0.34
Temp     0.54    0.22 -0.34  1.00

n= 131

##P Values among variables,

P
        Ozone  Solar.R Wind   Temp
Ozone          0.0006  0.0001 0.0000
Solar.R 0.0006         0.7309 0.0119
Wind    0.0001 0.7309         0.0000
Temp    0.0000 0.0119  0.0000
```

#But still considering , Solar.R , Ozone and Wind as independent variables to Temperature let us do a regression analysis to see the impact of each of the variables on Temperature.

```
> lm(x$Temp ~ x$Ozone+x$Solar.R+x$Wind) -> m
> summary(m)

Call:
lm(formula = x$Temp ~ x$Ozone + x$Solar.R + x$Wind)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.954  -4.642   1.019   4.612  14.771 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 74.645981   2.846478  26.224  < 2e-16 ***
x$Ozone      0.154142   0.025961   5.937 1.96e-08 ***
x$Solar.R    0.011809   0.007187   1.643  0.10249    
x$Wind      -0.547642   0.201812  -2.714  0.00744 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.364 on 149 degrees of freedom
Multiple R-squared:  0.4066,	Adjusted R-squared:  0.3947 
F-statistic: 34.03 on 3 and 149 DF,  p-value: < 2.2e-16
```

Conclusion :

a) We can see Ozone has a significant positive relationship with Temperature followed by Wind in some significantly negative way.
b) Solar.R has no direct significance to the temperature but since Solar.R mildly impacts Ozone so it may indirectly impact Temperature as well.
c) The respective plots are suggestive but we don’t have sufficient data to establish a strong model connecting these 4 variables but roughly our regr ession equation would be,

Temperature = 74.65 + 0.15Ozone + .012Solar.R – 0.55 Wind

#Test

```
> predict(m,"x$Ozone" = 42.13,"x$Solar.R" = 185.9, "x$Wind" = 9.958)
```

Would yield a result for Temperature for all 153 observations.

[N.B : The last part was exploratory and has nothing to do with the last que stion. But it gave me an insight . I do not quite know for now how to get a single Temperature prediction value for single values of the input variables But I will surely find out]

```
ggplot(train, aes(Item_MRP)) + geom_histogram(binwidth = 2)+
```

```
scale_x_continuous("Item MRP", breaks = seq(0,270,by = 30))+
scale_y_continuous("Count", breaks = seq(0,200,by = 20))+
labs(title = "Histogram")
```