Data Visualisation Assignment 2 (18230229)

Sai Krishna Lakshminarayanan

15 February 2019

Introduction

The given dataset deals with Daily air quality measurements in New York from May to July 1973. It consist of the attributes namely Ozone, Wind,Temp ,Month and Day. The data is obtained from the New York State Department of Conservation and the National Weather Service.

Part -1

The given dataset is loaded into the R Studio. It is seen that there are several NA values present. So , the NA values are then replaced with the mean values of the corresponding columns. This is because, it doesnt lead to any data loss and can give out a better result as the mean values are considered.

```
library(ggplot2)#importing the packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
airquality<-read.csv("airqualityv2.csv")#loading the dataset
airquality<-airquality %>% mutate_all(~ifelse(is.na(.x), mean(.x, na.rm = TRUE), .x))#replacing
 na with mean values of each column
head(airquality)
```

```
##      Ozone Wind Temp Month Day
## 1 41.00000  7.4   67     5   1
## 2 36.00000  8.0   72     5   2
## 3 12.00000 12.6   74     5   3
## 4 18.00000 11.5   62     5   4
## 5 39.60656 14.3   56     5   5
## 6 28.00000 14.9   66     5   6
```

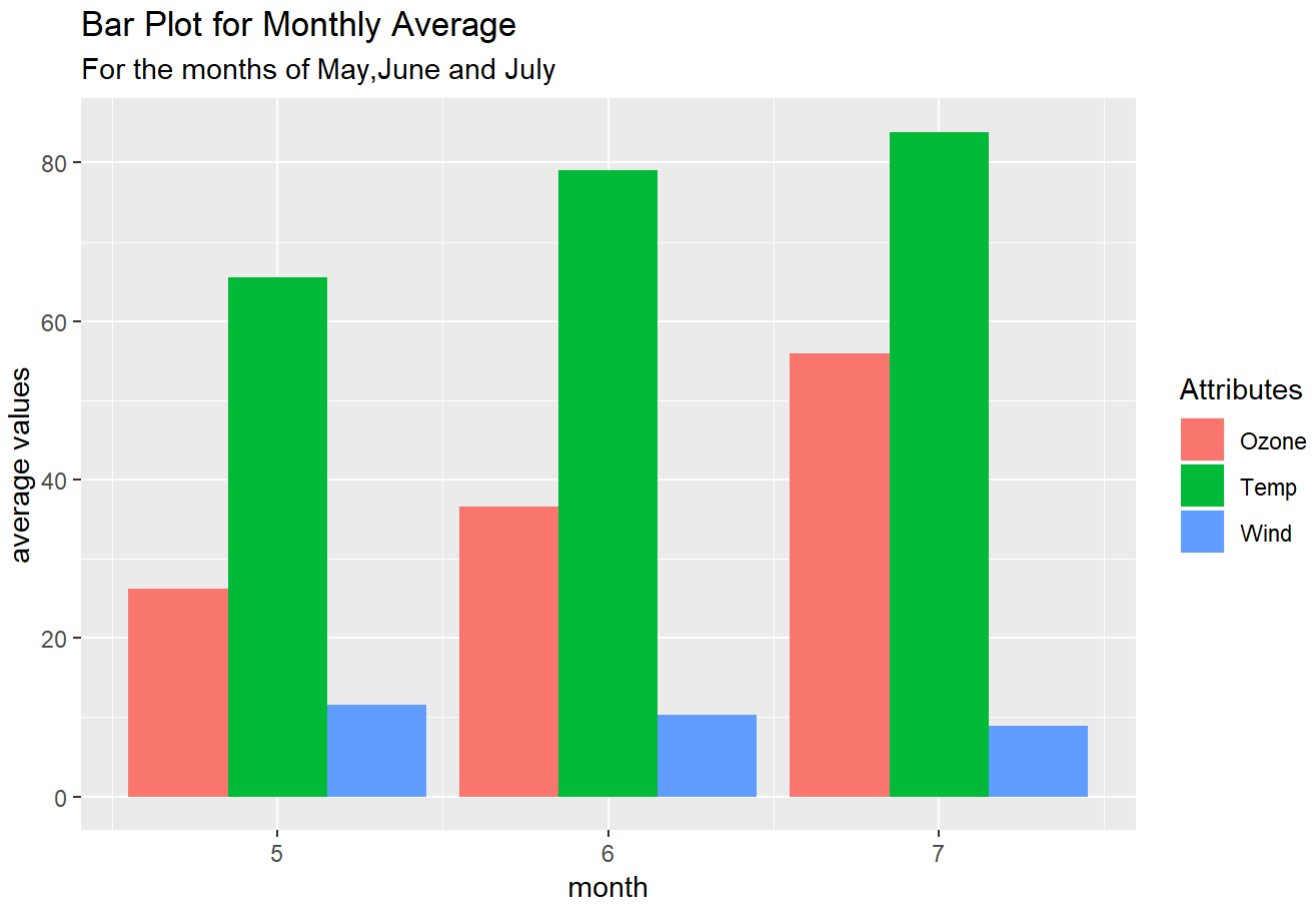
Now, the dataset is then grouped based on the month and arranged based on their values. For this, the functions of group_by() and summarise() are used. The average values of Ozone, temp and wind based on each month of May,June and July is obtained and binded into one dataframe.

```
a<-airquality %>%# getting the average ozone value per month. keeping column names for average a
nd attributes same for all 3 so that binding is proper in the end
  group_by(Month)%>%
 summarise(average=mean(Ozone),
           attributes="Ozone")
b<-airquality %>% # getting the average temperature value per month.
  group_by(Month)%>%
  summarise(average=mean(Temp),
            attributes="Temp")
c<-airquality %>% # getting the average wind value per month.
  group_by(Month)%>%
  summarise(average=mean(Wind),
            attributes="Wind")
d<-rbind(a,b,c)#binding all three
head(d)
```

```
## # A tibble: 6 x 3
##   Month average attributes
##   <int>   <dbl> <chr>
## 1     5    26.2 Ozone
## 2     6    36.6 Ozone
## 3     7    56.0 Ozone
## 4     5    65.5 Temp
## 5     6    79.1 Temp
## 6     7    83.9 Temp
```

Now the bar chart is designed for this average values for 3 months.It is seen that the values of Ozone and Temperature increased each month and the value of wind decreased across each month from May to July.

```
ggplot(data = d, aes(x=Month, y = average,fill=attributes)) +
geom_col(position="dodge")  +
labs(title="Bar Plot for Monthly Average", subtitle="For the months of May,June and July", y="av
erage values", x="month", caption="Plot 1",fill="Attributes")#providing the title and descriptio
n for the plot
```



Plot 1

Part-2

Now, in order to perform the divering bar charts for ozone and temperature, it is important to obtain the difference between the values and the mean. So, mutating to get the difference in ozone and temp from their values to their means and also getting two columns based on logical operators to check if the value is truly greater than the mean or not.

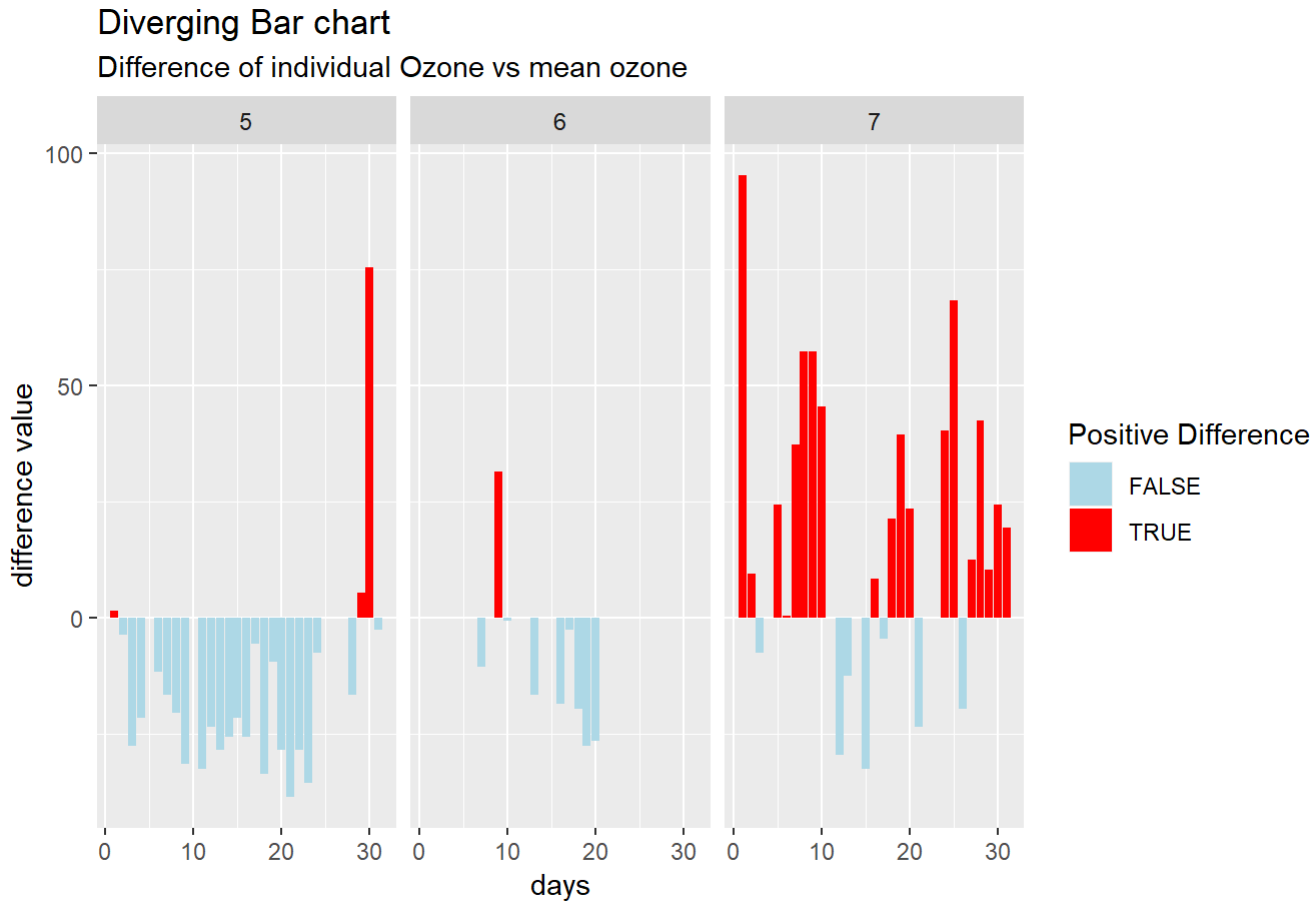
```
airquality <- airquality %>% #obtaining the difference between ozone and temp with their means r
espectively and storing them
  mutate(pos_ozone = Ozone >=mean(Ozone),
         pos_temp= Temp >= mean(Temp),
         Difference_ozone= Ozone -mean(Ozone),
         Difference_Temp=Temp-mean(Temp))
head(airquality)
```

```
##      Ozone Wind Temp Month Day pos_ozone pos_temp Difference_ozone
## 1 41.00000  7.4   67     5   1      TRUE    FALSE         1.393443
## 2 36.00000  8.0   72     5   2     FALSE    FALSE        -3.606557
## 3 12.00000 12.6   74     5   3     FALSE    FALSE       -27.606557
## 4 18.00000 11.5   62     5   4     FALSE    FALSE       -21.606557
## 5 39.60656 14.3   56     5   5      TRUE    FALSE         0.000000
## 6 28.00000 14.9   66     5   6     FALSE    FALSE       -11.606557
##   Difference_Temp
## 1       -9.152174
## 2       -4.152174
## 3       -2.152174
## 4      -14.152174
## 5      -20.152174
## 6      -10.152174
```

Diverging bar for Ozone

In the diverging bar, the default colour for True positive difference is blue and False positive difference is red. The colour combination seems to be contradicting because negative difference means there is lower ozone level(good conditions) than the mean and positive difference means there is higher level (bad conditions) than the mean. So, swapping the colours for better representation.

```
ggplot(airquality, aes(x = Day, y = Difference_ozone, fill = pos_ozone)) +
  geom_col(position = "identity")+#providing the values for the diverging bar
  labs(title="Diverging Bar chart", subtitle="Difference of individual Ozone vs mean ozone", y=
"difference value", x="days", caption="Plot 2",fill="Positive Difference")+
  scale_fill_manual(values = c("lightblue","red"))+
  facet_wrap(~Month)
```



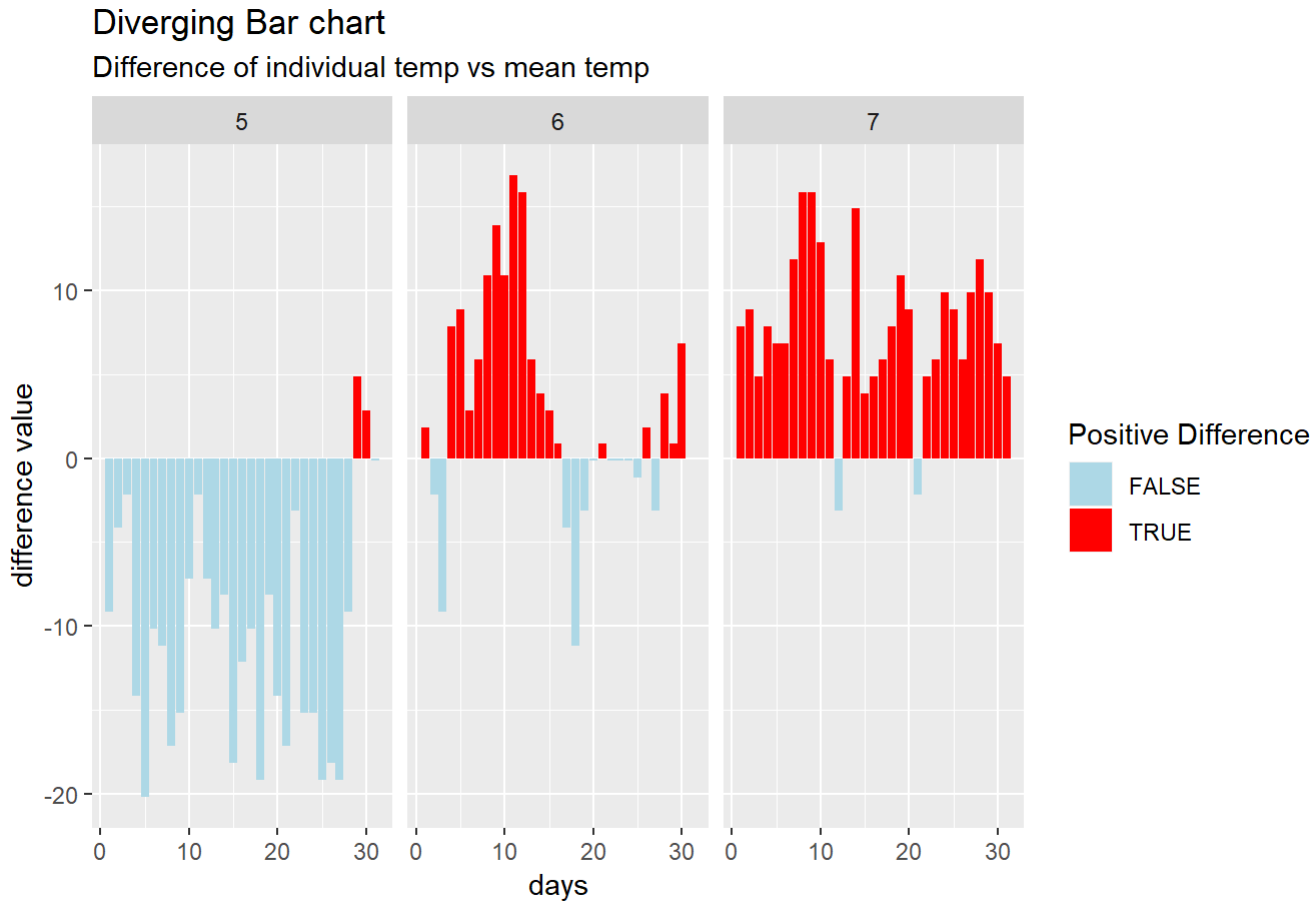
Plot 2

It is seen that the ozone values were predominantly lower to the mean in May and nearer to the mean in June and higher than to the mean in July thereby showing us that its value is increasing predominantly from May to July.

Diverging bar for temperature

In the diverging bar, the default colour for True positive difference is blue and False positive difference is red. The colour combination seems to be contradicting because negative difference means there is lower temperature than the mean and positive difference means there is higher temperature than the mean. So, swapping the colours for better representation.

```
ggplot(airquality, aes(x = Day, y = Difference_Temp, fill = pos_temp)) +
  geom_col(position = "identity")+
  labs(title="Diverging Bar chart", subtitle="Difference of individual temp vs mean temp", y="di
fference value", x="days", caption="Plot 3",fill="Positive Difference")+
  scale_fill_manual(values = c("lightblue","red"))+
  facet_wrap(~Month)
```



Plot 3

It is seen that temperature is also following the similar pattern with the most of thelowest temperatures are in May and most of the highest temperatures are in June and July.

Part-3

Slope chart for average values across 3 months

In order to perform slope chart, data transformation has to be done. So, considering a data frame which is having the months as columns and the average values as row for the best visual output.

```
library(scales)
f<-airquality %>%
  group_by(Month)%>% #grouping by month and getting the average
  summarise(avg_ozone=mean(Ozone),
            avg_wind=mean(Wind),
            avg_temp=mean(Temp))
f<-as.data.frame(t(f))#transposing the dataframe
colnames(f)<-c("May","June","July")#giving the column names
f<-f[-1,]#removing redundant row
head(f)
```

```
##                May     June      July
## avg_ozone 26.19461 36.55792 55.968800
## avg_wind  11.62258 10.26667  8.941935
## avg_temp  65.54839 79.10000 83.903226
```

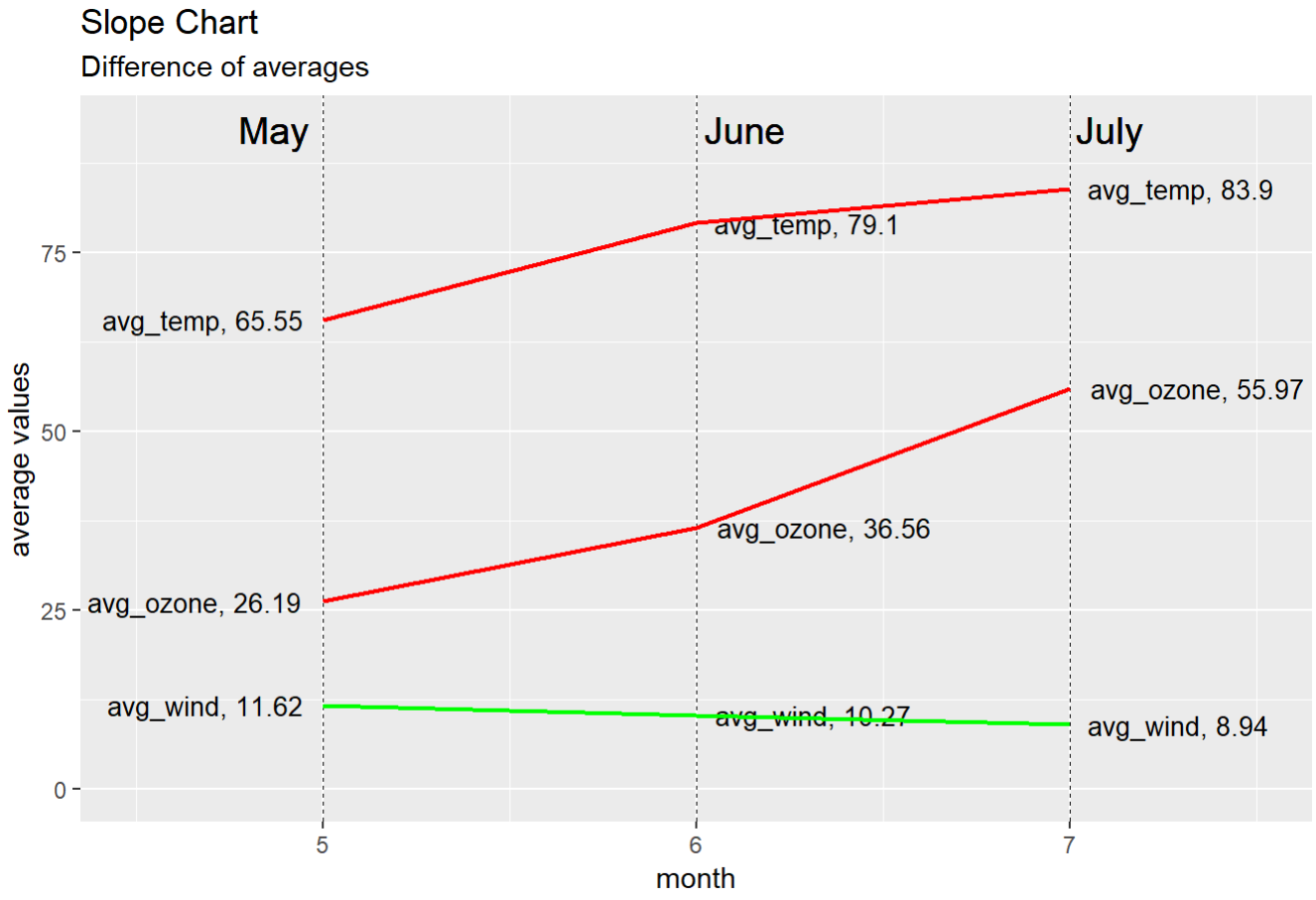
Now, generating the labels for the 3 lines and the class colour to show the difference.If the difference is higher than the mean, then it is indicated by red colour as it represents higher temperature and ozone level. Similarly, lower winds is represented in green colour.

```
left_label <- paste(rownames(f), round(f$`May`,digits = 2),sep=", ")#giving the label names
middle_label <- paste(rownames(f), round(f$`June`,digits = 2),sep=", ")
right_label <- paste(rownames(f), round(f$`July`,digits = 2),sep=", ")
f$class <- ifelse((f$July - f$May) < 0, "red", "green")#condition for slope colour
```

```
p <- ggplot(f) +
  geom_vline(xintercept=5, linetype="dashed", size=.1) + #generating the slope
  geom_vline(xintercept=6, linetype="dashed", size=.1) +
  geom_vline(xintercept=7, linetype="dashed", size=.1) +

  labs(x="Months", y="Mean of Ozone Temp and Wind") +
  xlim(4.5, 7.5) +
  ylim(0,(1.1*(max(f$May, f$June,f$July))))  # X and Y axis limits
p <- p + geom_text(label=left_label, y=f$May,x=rep(5, NROW(f)), hjust=1.1, size=3.5)
p <- p + geom_text(label=middle_label, y=f$June, x=rep(6, NROW(f)), hjust=-0.1, size=3.5)
p <- p + geom_text(label=right_label, y=f$July, x=rep(7, NROW(f)), hjust=-0.1, size=3.5)
p <- p + geom_text(label=colnames(f)[1], x=5, y=1.1*(max(f$May, f$June,f$July)), hjust=1.2, size
=5)  # title of left line
p <- p + geom_text(label=colnames(f)[2], x=6, y=1.1*(max(f$May, f$June,f$July)), hjust=-0.1, siz
e=5) #title of middle line
p <- p + geom_text(label=colnames(f)[3], x=7, y=1.1*(max(f$May, f$June,f$July)), hjust=-0.1, siz
e=5) # title of right line
```

```
p <- p + geom_segment(aes(x=5, xend=6, y=f$May, yend=f$June, col=f$class), size=.75, show.legend
=F) +
  geom_segment(aes(x=6, xend=7, y=f$June, yend=f$July, col=f$class), size=.75, show.legend=F) +
  scale_color_manual(labels = c("Up", "Down"),
                     values = c("red","green"))+
  labs(title="Slope Chart", subtitle="Difference of averages", y="average values", x="month", ca
ption="Plot 4")
p
```



Plot 4

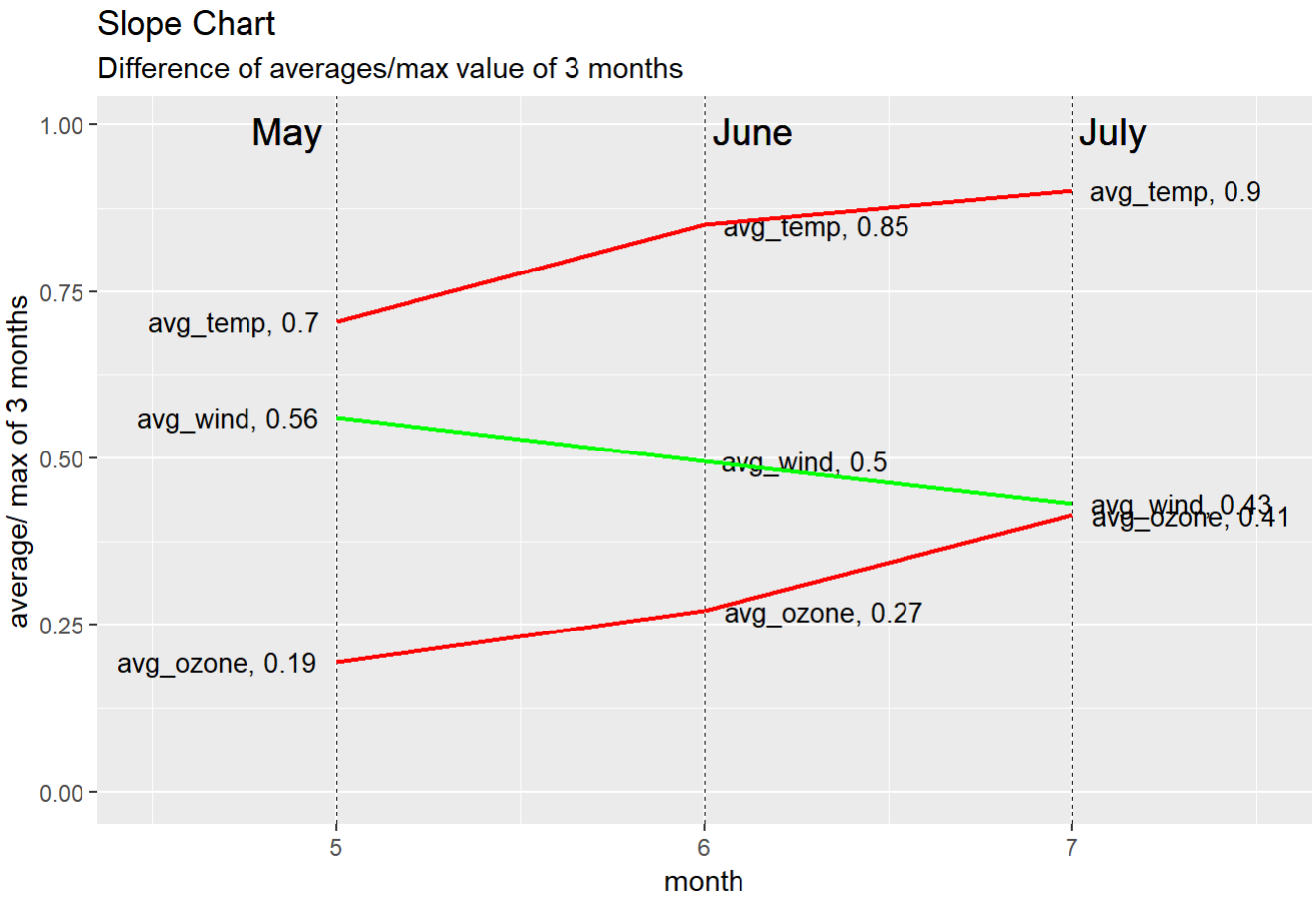
From the slope chart, it is seen that the temperature and ozone average values are increasing at an alarming rate from may to july. Only the winds average value reduced slightly from May to July.

Slope chart for average value divided by 3 month maximum value

Now, data preprocessing has to be done to divide the value of the average to the maximum value. Then, similar process and colouring is carried out as increase in temperature and ozone means more heat and hence red colour and decrease means green colour to get the slope chart as above.

```
f<-f[,-4]
f[1,]<-f[1,]/max(airquality$Ozone)#Obtaining the average/max of 3 month value
f[2,]<-f[2,]/max(airquality$Wind)
f[3,]<-f[3,]/max(airquality$Temp)
f$class <- ifelse((f$July - f$May) < 0, "red", "green")
left_label <- paste(rownames(f), round(f$`May`,digits = 2),sep=", ")#providing the labels
middle_label <- paste(rownames(f), round(f$`June`,digits = 2),sep=", ")
right_label <- paste(rownames(f), round(f$`July`,digits = 2),sep=", ")
```

```
p <- ggplot(f) +
  geom_vline(xintercept=5, linetype="dashed", size=.1) +
  geom_vline(xintercept=6, linetype="dashed", size=.1) +
  geom_vline(xintercept=7, linetype="dashed", size=.1) +
  labs(x="Months", y="Mean of Ozone Temp and Wind") +  # Axis labels
  xlim(4.5, 7.5) +
  ylim(0,(1.1*(max(f$May, f$June,f$July))))  # X and Y axis limits
p <- p + geom_text(label=left_label, y=f$May,x=rep(5, NROW(f)), hjust=1.1, size=3.5)
p <- p + geom_text(label=middle_label, y=f$June, x=rep(6, NROW(f)), hjust=-0.1, size=3.5)
p <- p + geom_text(label=right_label, y=f$July, x=rep(7, NROW(f)), hjust=-0.1, size=3.5)
p <- p + geom_text(label=colnames(f)[1], x=5, y=1.1*(max(f$May, f$June,f$July)), hjust=1.2, size
=5)  # title of left line
p <- p + geom_text(label=colnames(f)[2], x=6, y=1.1*(max(f$May, f$June,f$July)), hjust=-0.1, siz
e=5)#title of middle line
p <- p + geom_text(label=colnames(f)[3], x=7, y=1.1*(max(f$May, f$June,f$July)), hjust=-0.1, siz
e=5) # title of reight line
p <- p + geom_segment(aes(x=5, xend=6, y=f$May, yend=f$June, col=f$class), size=.75, show.legend
=F) +
  geom_segment(aes(x=6, xend=7, y=f$June, yend=f$July, col=f$class), size=.75, show.legend=F) +
  scale_color_manual(labels = c("Up", "Down"),
                     values = c("red", "green"))+   # color of lines
  labs(title="Slope Chart", subtitle="Difference of averages/max value of 3 months", y="average/
max of 3 months", x="month", caption="Plot 5")
p
```



Plot 5

From this, it is seen that the trend is similar, with the value of average/max of 3 months of ozone and temperature increasing from may to july constantly and decreasing for winds.

Part-4

Inorder to find the distribution of all the 3 ozone temperature and wind together across 3 months, their values have to be brought together in a single column. So, data preprocessing is done to obtain the values,variable name,date,day and month into each column.

```
library(reshape2)
g<-melt(airquality[1:5])#rearraning the original dataframe
```

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

```
g<-g[1:276,1:2]#taking only the needed values
g$Date<-seq(as.Date("1973/05/01"),by="day",length.out = NROW(airquality))#obtaining the date for
mat
g$Month<-substr(g$Date,7,7)#extracting month from date
g$Day<-substr(g$Date,9,10)#extracting day from date
head(g)
```

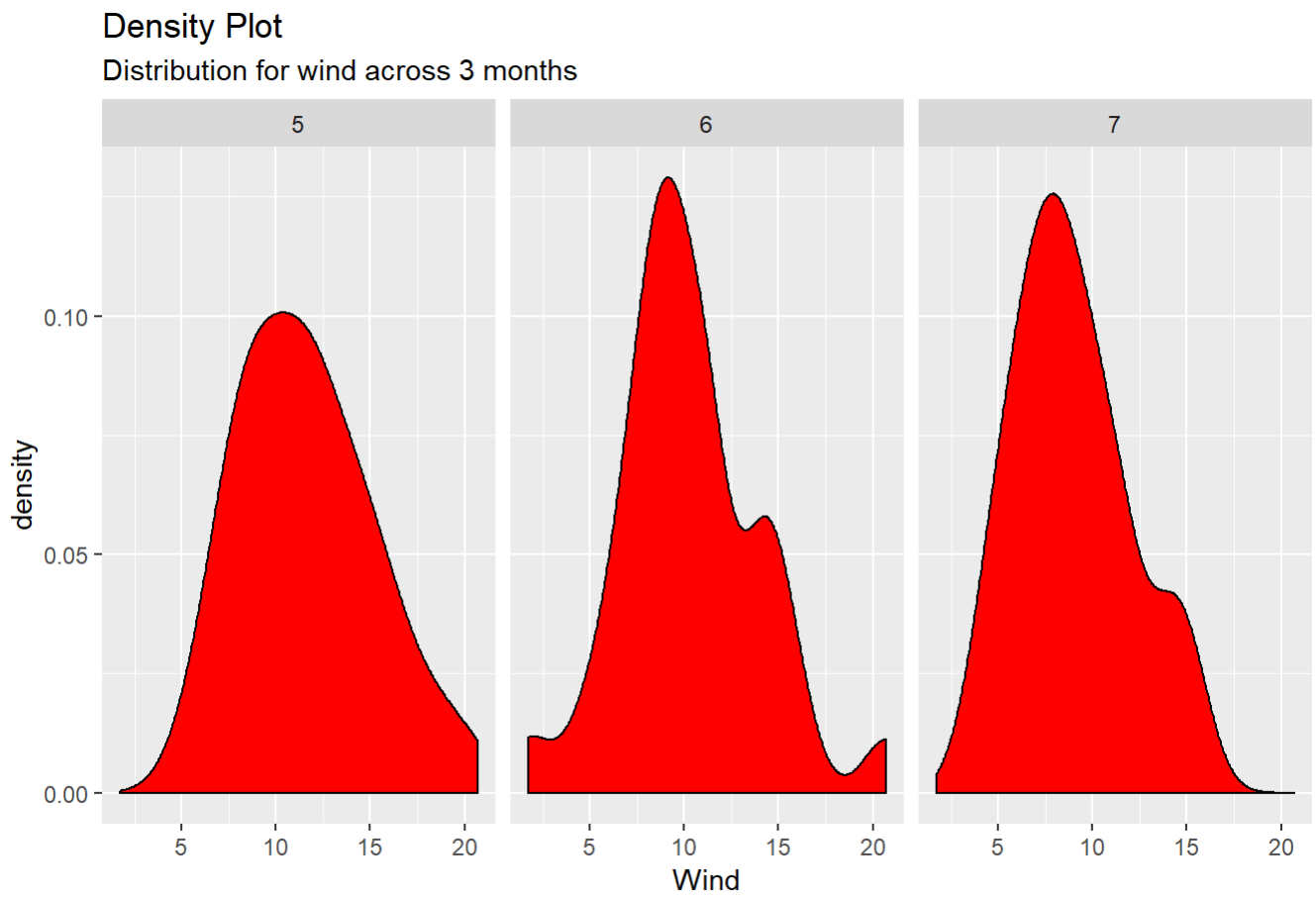
```
##   variable    value       Date Month Day
## 1    Ozone 41.00000 1973-05-01     5  01
## 2    Ozone 36.00000 1973-05-02     5  02
## 3    Ozone 12.00000 1973-05-03     5  03
## 4    Ozone 18.00000 1973-05-04     5  04
## 5    Ozone 39.60656 1973-05-05     5  05
## 6    Ozone 28.00000 1973-05-06     5  06
```

Visualisation 1- Density Chart

For Wind across 3 months

The density plot is obtained using the geom_density(). It is seen that the distribution of winds for May is consistent and nearly normally distributed. For June, it is seen that it is slightly heavier towards its left, indicating that its mean and median are different. For July, it is slightly right skewed due to the presence of a tail in right side.

```
ggplot(airquality)+
  geom_density(aes(x=Wind),fill="Red")+#density plot for wind
  facet_wrap(~Month)+
labs(title="Density Plot", subtitle="Distribution for wind across 3 months", caption="Plot 6")
```

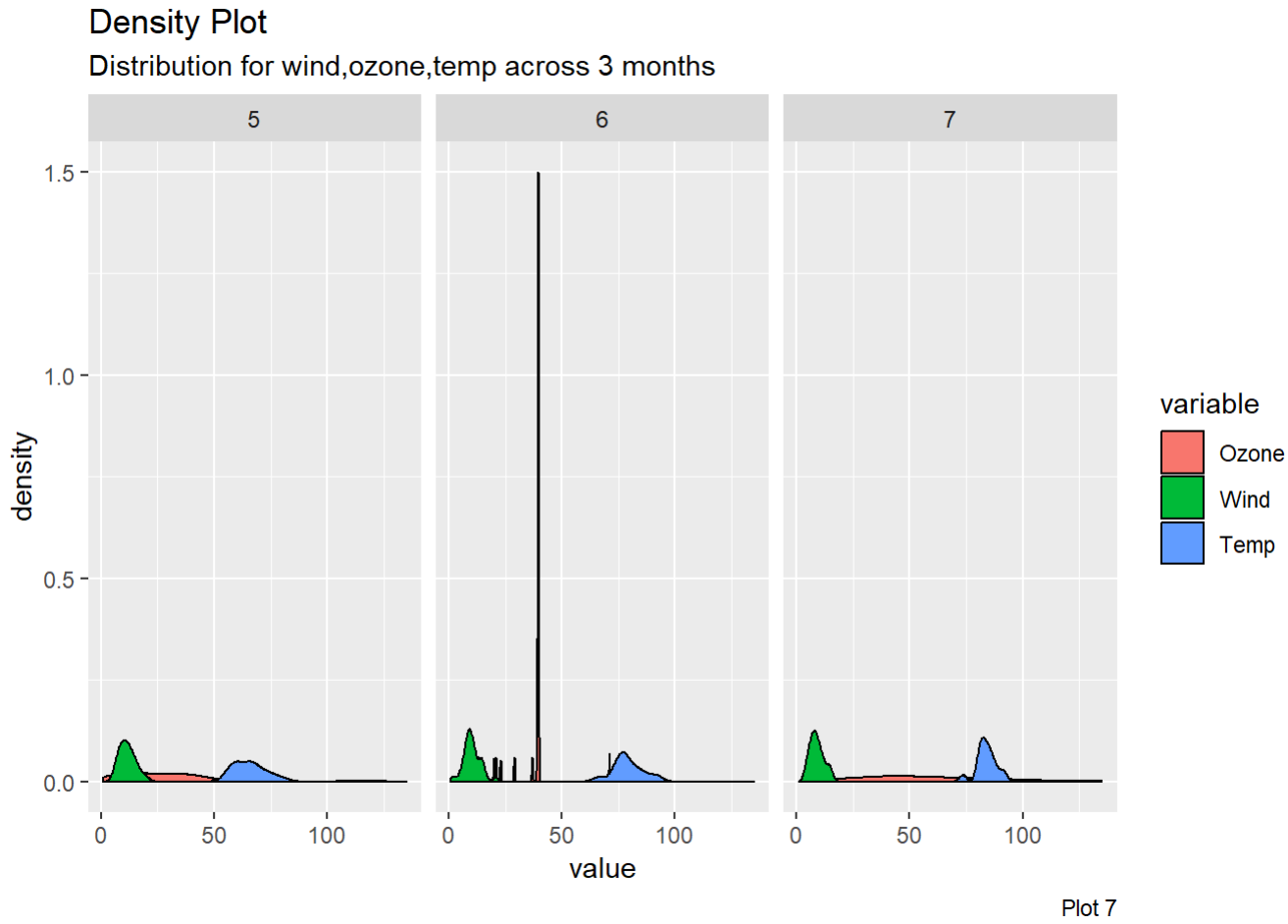


Plot 6

For all 3 variables

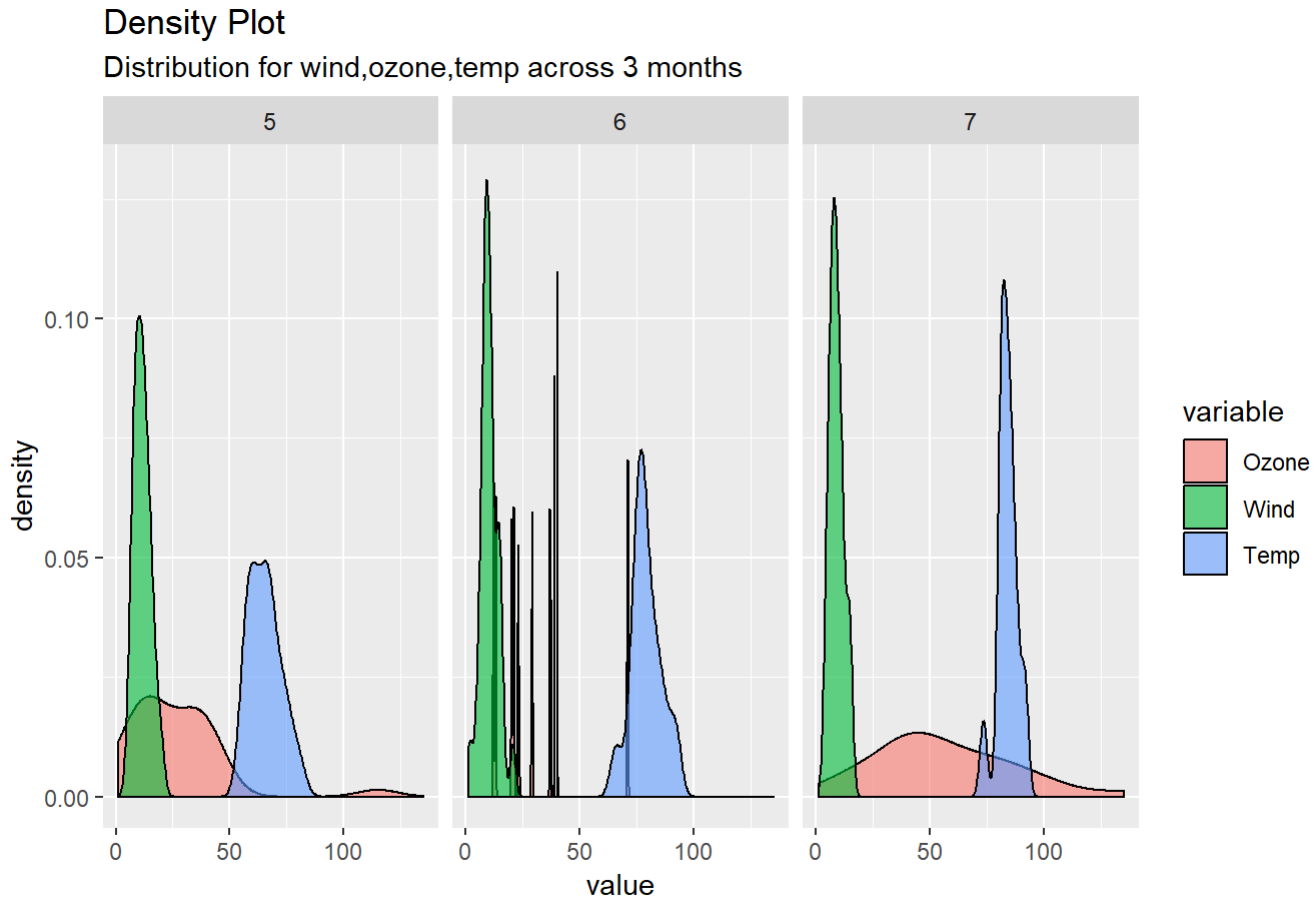
Now, developing a density plot for the distrbution of ozone,temperature and winds across 3 months. It is seen that one of the density of ozone is very high and hence the plot is too small to figure out.

```
ggplot(g)+
  geom_density(aes(x=value,fill=variable))+
  facet_wrap(~Month)+
labs(title="Density Plot", subtitle="Distribution for wind,ozone,temp across 3 months", caption=
"Plot 7")
```



Therefore, y limits are provided inorder to give out a better meaning density plot. It is seen that ozone distrbution is right skewed in months of may and kuly and scattered for june . Distribution for temperature seems to be right skewed in May and normally distributed for the other two months. The default colours seemed to be differentiating well and the classic theme kept the things simple. So those are retained as given.

```
ggplot(g)+
  geom_density(aes(x=value,fill=variable),alpha=0.6)+
  ylim(0,0.13)+
  facet_wrap(Month ~.)+
  labs(title="Density Plot", subtitle="Distribution for wind,ozone,temp across 3 months", captio
n="Plot 8")
```



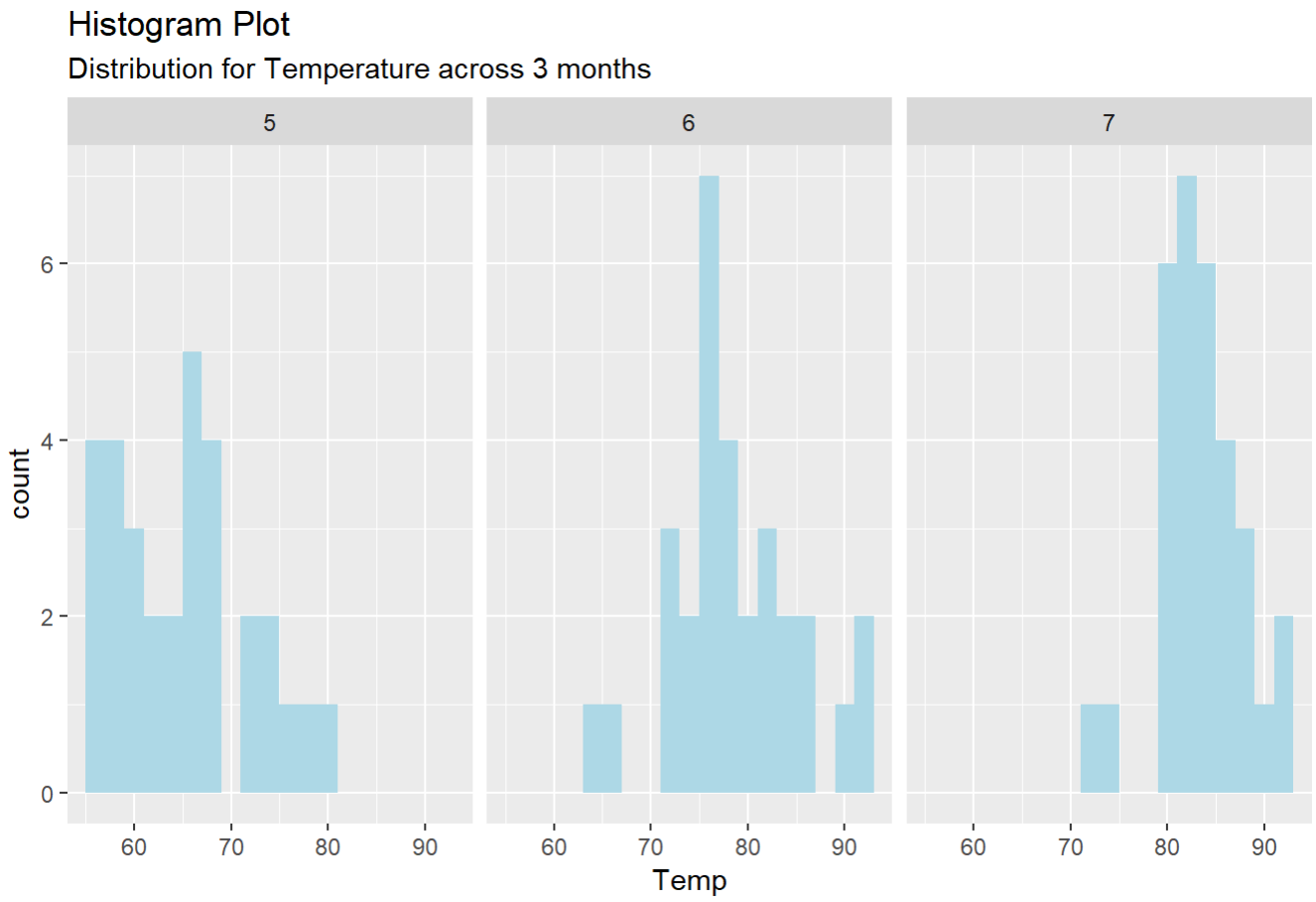
Plot 8

Visualisation 2- Histogram

For Temperature

From the historgram of distribution of temperature in 3 months, it is seen that the count is normally distributed in june and the lower temperature counts are greater in may and higher temperature counts are greater in july.

```
ggplot(airquality)+
  geom_histogram(aes(x=Temp),binwidth=2,fill="lightblue")+#histogram plot
  facet_wrap(~Month)+
  labs(title="Histogram Plot", subtitle="Distribution for Temperature across 3 months", caption=
"Plot 9")
```

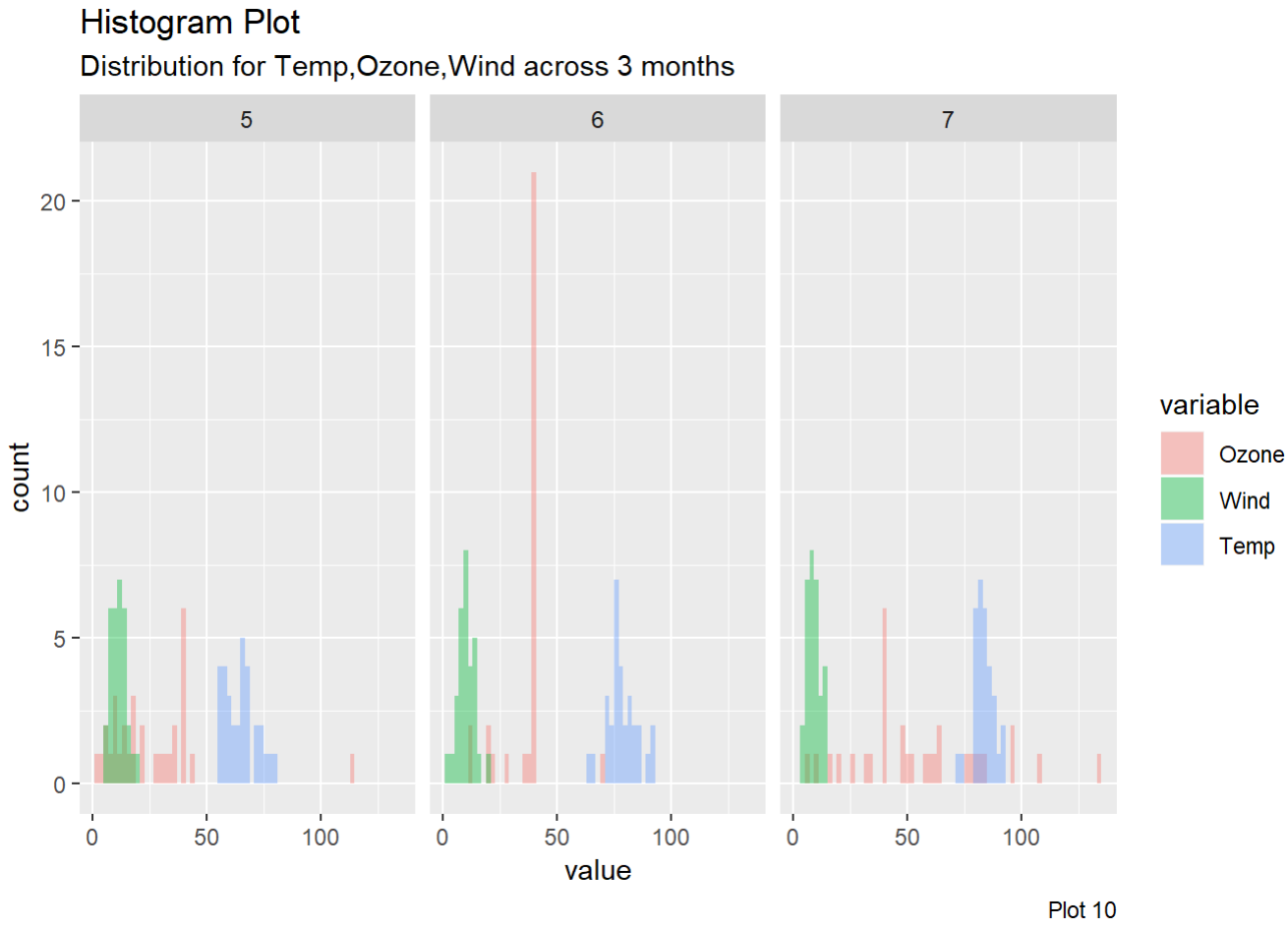


Plot 9

For all 3 variables

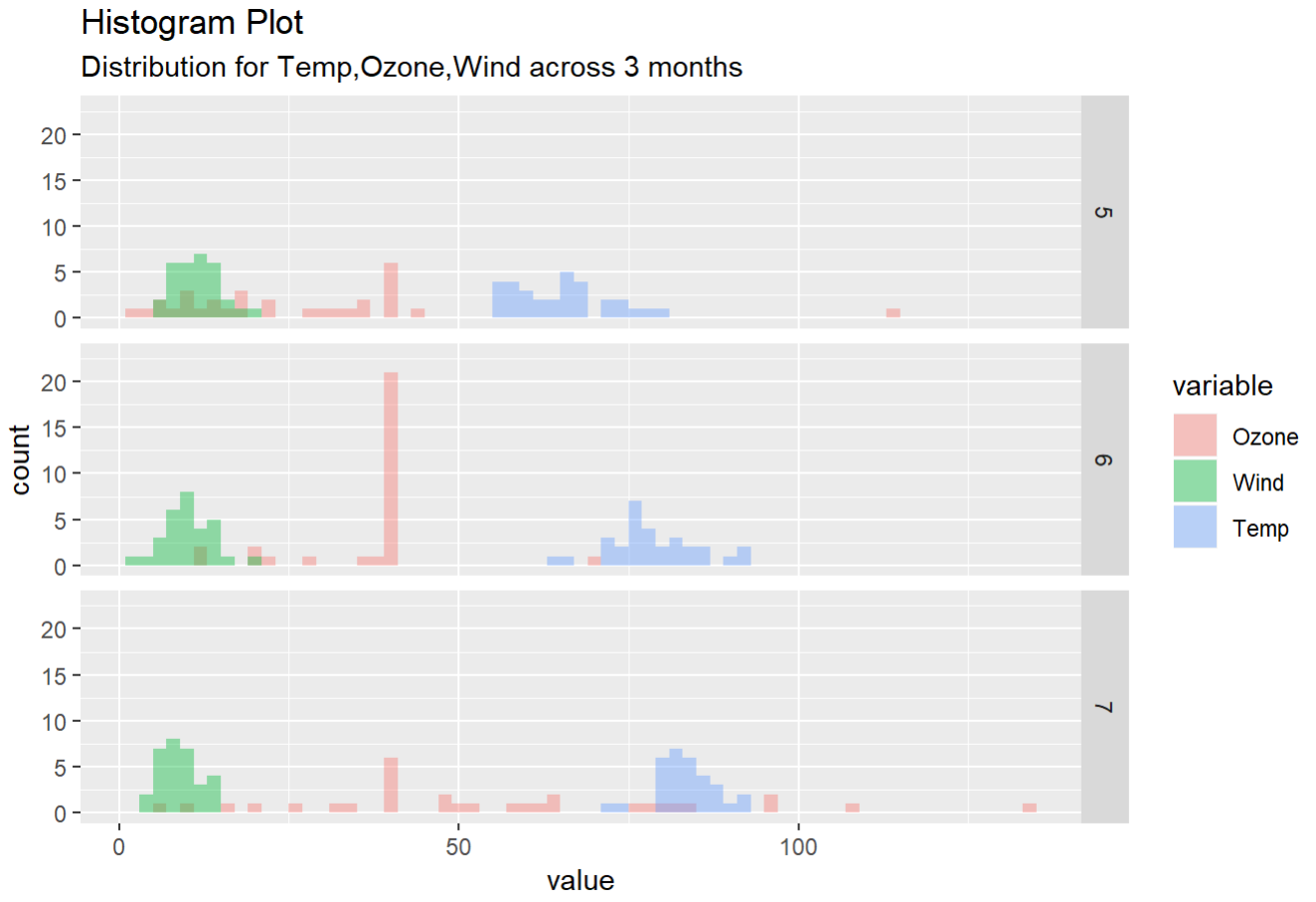
In this, different colour is given to each of ozone,wind and temperature. There are overlapping and so alpha is used to increase transparency.

```
ggplot(g)+
  geom_histogram(aes(x=value,fill=variable),binwidth = 2,alpha=0.4,position = "identity")+
  facet_wrap(~Month)+
  labs(title="Histogram Plot", subtitle="Distribution for Temp,Ozone,Wind across 3 months", capt
ion="Plot 10")
```



From the above, it is seen that one value of ozone has alot of count which reduces the visibility of other counts. Therefore, using facet_grid() in place of facet_wrap() for better readability.

```
ggplot(g)+
  geom_histogram(aes(x=value,fill=variable),binwidth = 2,alpha=0.4,position = "identity")+
  ylim(0,23)+
  facet_grid(Month~.)+
labs(title="Histogram Plot", subtitle="Distribution for Temp,Ozone,Wind across 3 months", captio
n="Plot 11")
```



Plot 11

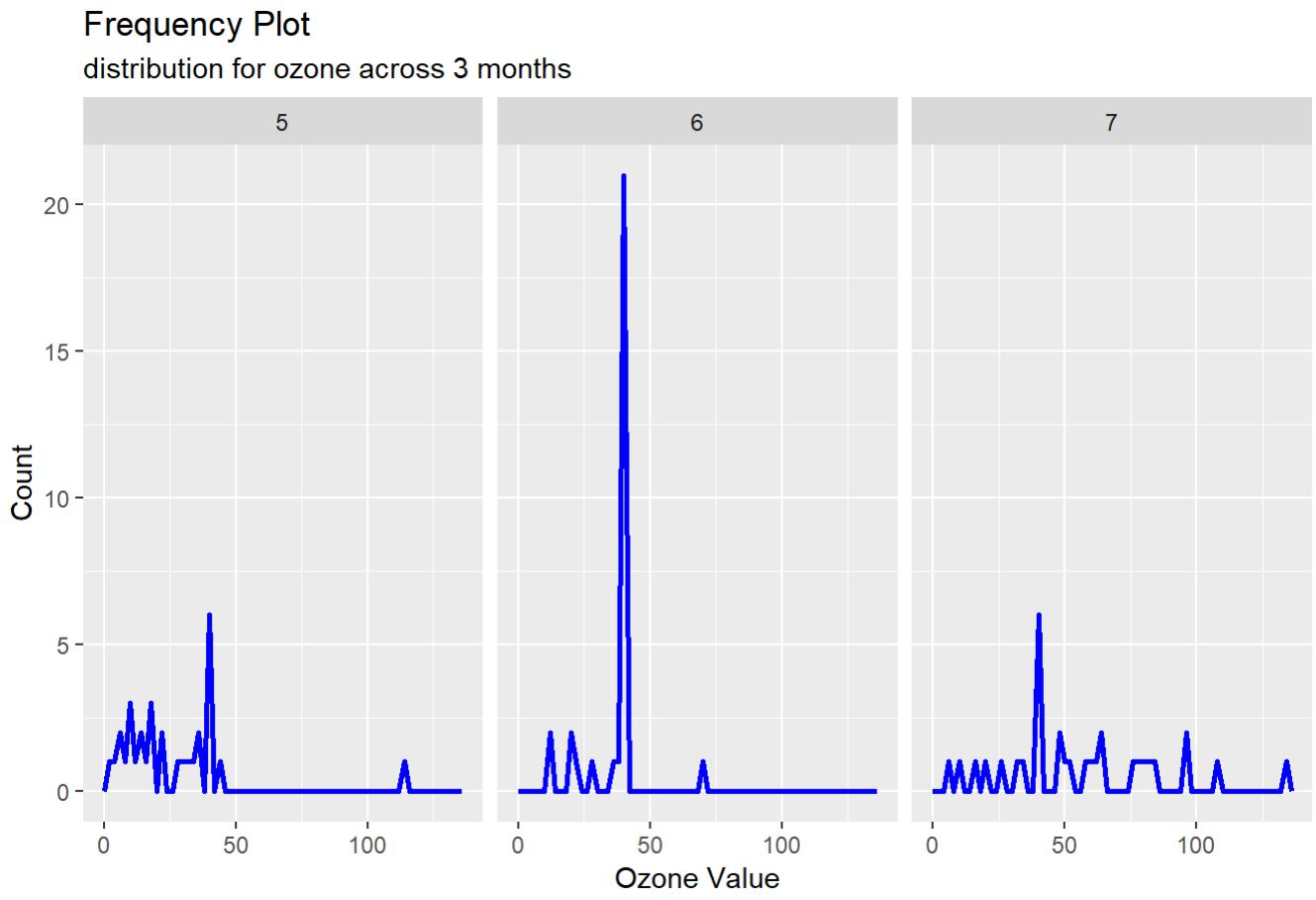
Visualisation-3 Frequency Polygon

For Ozone

The frequency polygon for ozone is obtained by using the geom_freqpoly().It gives the count of the ozone values across three months.

```
ggplot(airquality, aes(x=Ozone)) +
  geom_freqpoly(binwidth=2,color="blue",size = 1)+#frequency polygon chart

  facet_wrap(~Month)+
  labs(title="Frequency Plot", subtitle="distribution for ozone across 3 months", y="Count", x=
"Ozone Value",caption="Plot 12")
```

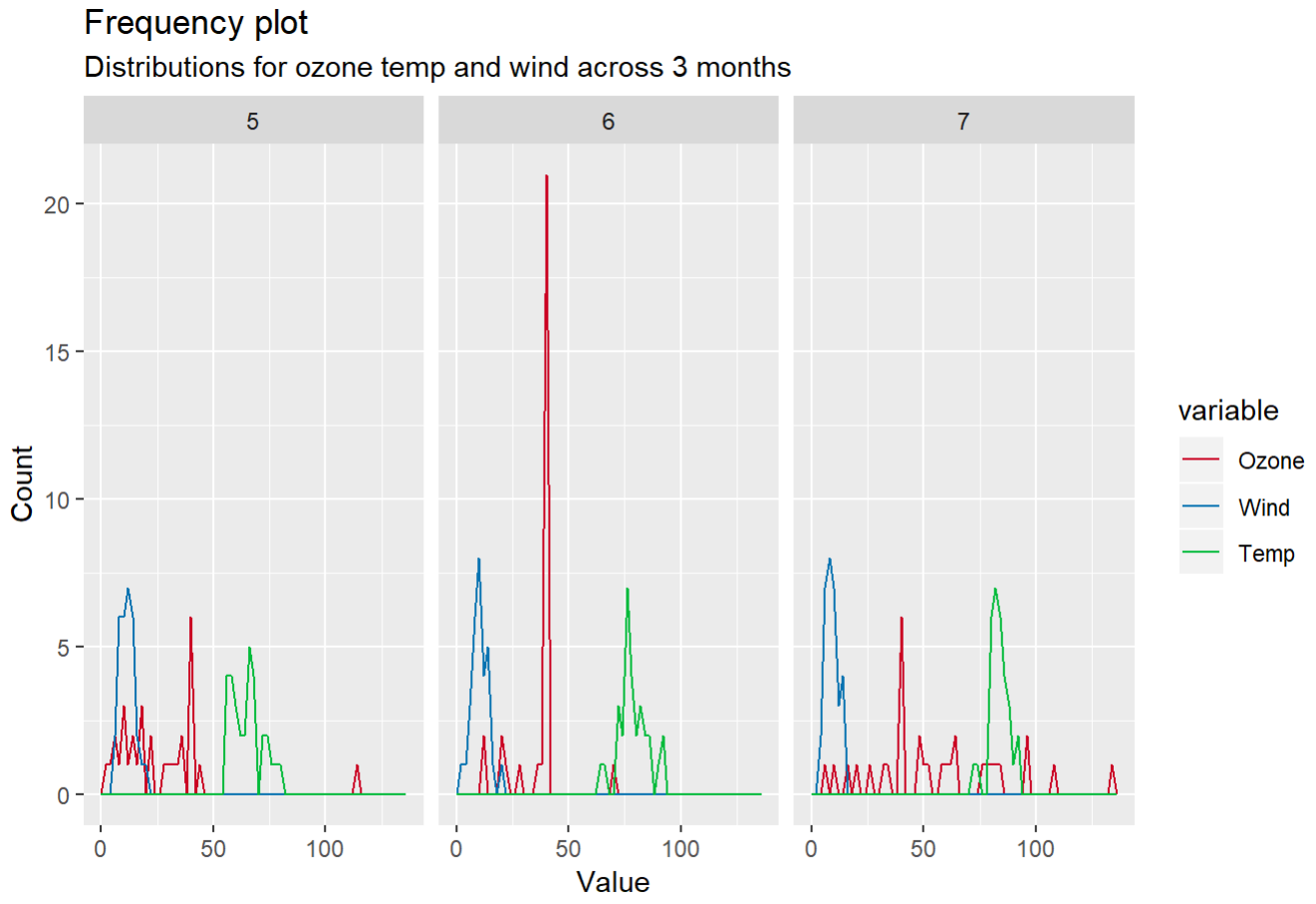


Plot 12

For all 3 variables

Now,the same is given for distribution of all the three cases. The colour is given based on the variable namely ozone,temperature and wind. The colours of red,blue and green and classic theme appears to be simple and differentiating the three.Therefore, it is kept.

```
ggplot(g, aes(x=value)) +
  geom_freqpoly(binwidth=2,aes(color=variable),size = 0.5)+
  facet_wrap(~Month)+
  scale_colour_manual(values= c("#ca0020","#0571b0","#00ba38")) +
  labs(title="Frequency plot", subtitle="Distributions for ozone temp and wind across 3 months",
y="Count", x=" Value",caption = "Plot 13")
```

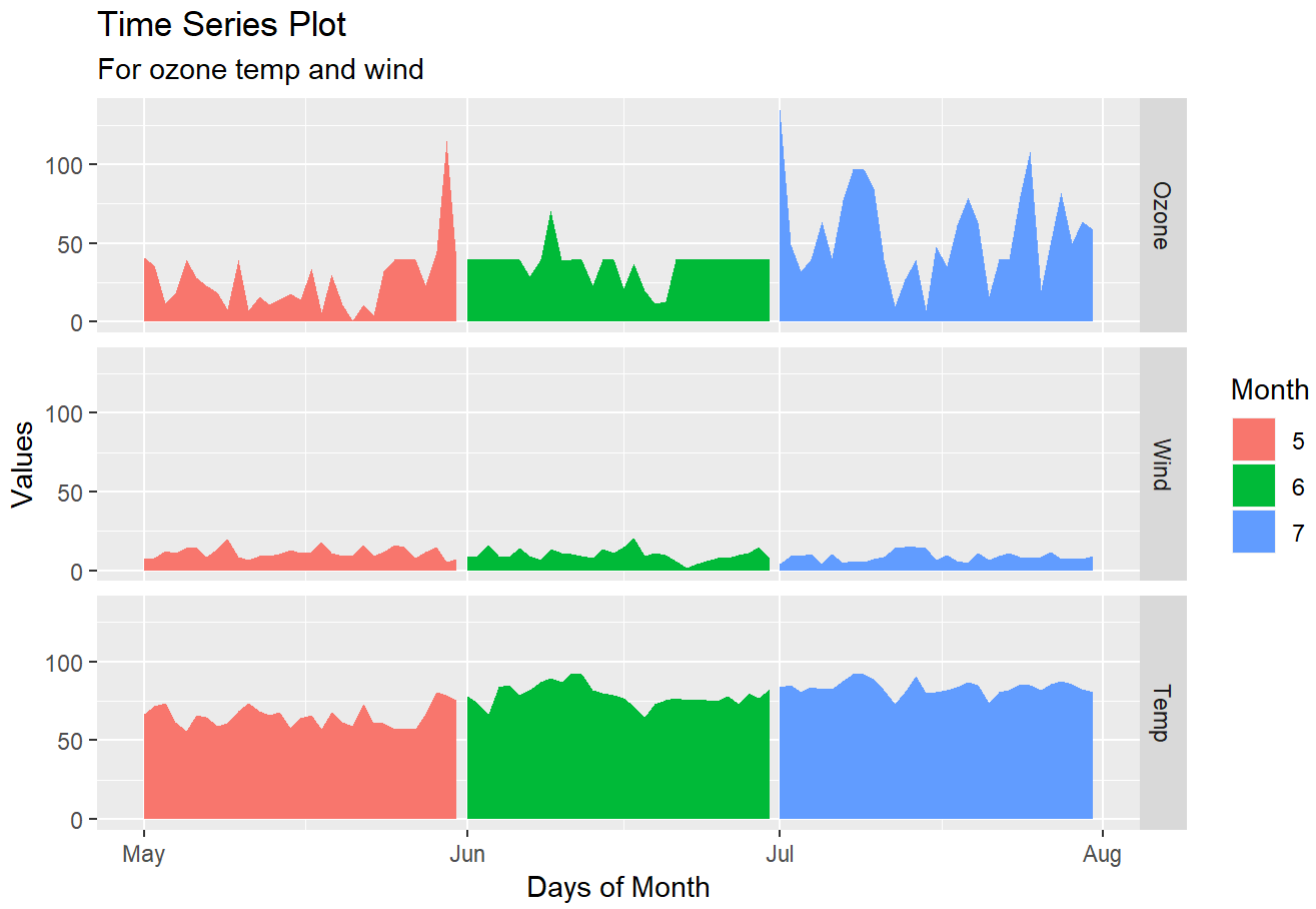


Plot 13

Part-5

Inorder to obtain the visualisation to show the temporal data aspect, geom_area is used. In this, the values of ozone,temperature and wind is found for all the days for each month of May,June and July. The classic theme and default colours are retained as it seems the visual simple and easily differentiating.

```
ggplot(g) +
  geom_area(aes(x = Date, y = value ,fill = Month))+
  labs(title="Time Series Plot", subtitle="For ozone temp and wind", y="Values", x="Days of Mont
h",caption = "Plot 14")+
  facet_grid(variable ~ .)
```



Plot 14

References

1)https://stackoverflow.com/questions/25835643/replace-missing-values-with-column-mean (https://stackoverflow.com/questions/25835643/replace-missing-values-with-column-mean)

2)https://stackoverflow.com/questions/3443687/formatting-decimal-places-in-r (https://stackoverflow.com/questions/3443687/formatting-decimal-places-in-r)

3)https://www.statmethods.net/management/reshape.html (https://www.statmethods.net/management/reshape.html)

4)https://stackoverflow.com/questions/37704212/extract-month-and-year-from-date-in-r (https://stackoverflow.com/questions/37704212/extract-month-and-year-from-date-in-r)

5)http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Slope%20Chart (http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Slope%20Chart)

6)http://mlbernauer.github.io/R/20150309_treemaps_with_ggplot2.html (http://mlbernauer.github.io/R/20150309_treemaps_with_ggplot2.html)

7)https://bookdown.org/lyzhang10/lzhang_r_tips_book/how-to-plot-data.html (https://bookdown.org/lyzhang10/lzhang_r_tips_book/how-to-plot-data.html)

8)Tutorial 2 materials