IST 687
Homework 6
Due Date: 11/16

Code requires the following packages to run

```
library(ggplot2) #install.packages('ggplot2')
library(devtools) #install.packages('devtools')
library(dplyr) #install.packages('dplyr')
library(reshape2) #install.packages('reshape2')
```

Step 1: Load the data

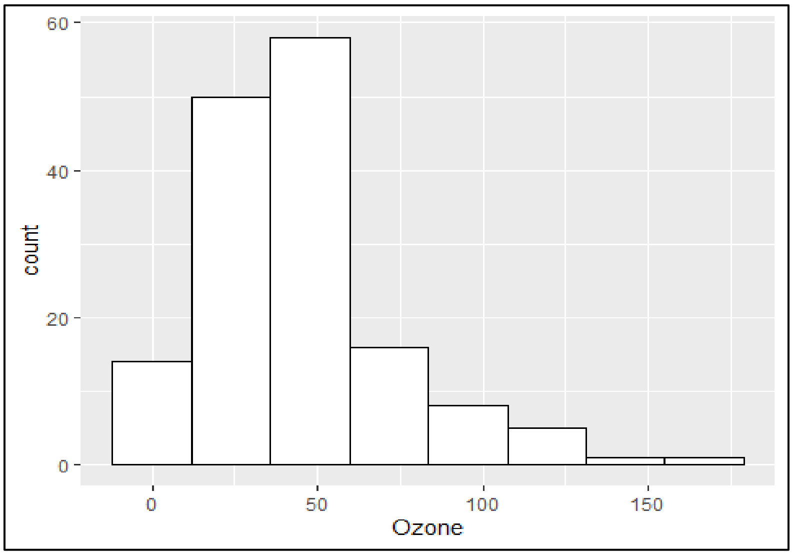
```
aq <- airquality #Load the data
```

Step 2: Clean the data

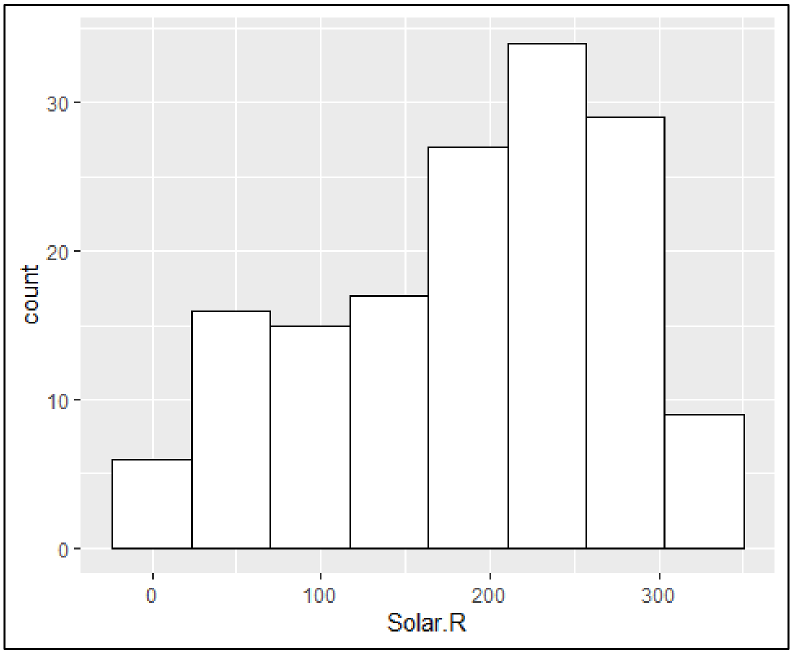
```
#Replace NA's to the mean for Ozone column
aq$Ozone[which(is.na(aq$Ozone))] <- as.integer(mean(aq$Ozone, na.rm = TRUE))

#Replace NA's to the mean for Solar.R column
aq$Solar.R[which(is.na(aq$Solar.R))] <- as.integer(mean(aq$Solar.R, na.rm = T
RUE))
```

Step 3: Understand the data distribution

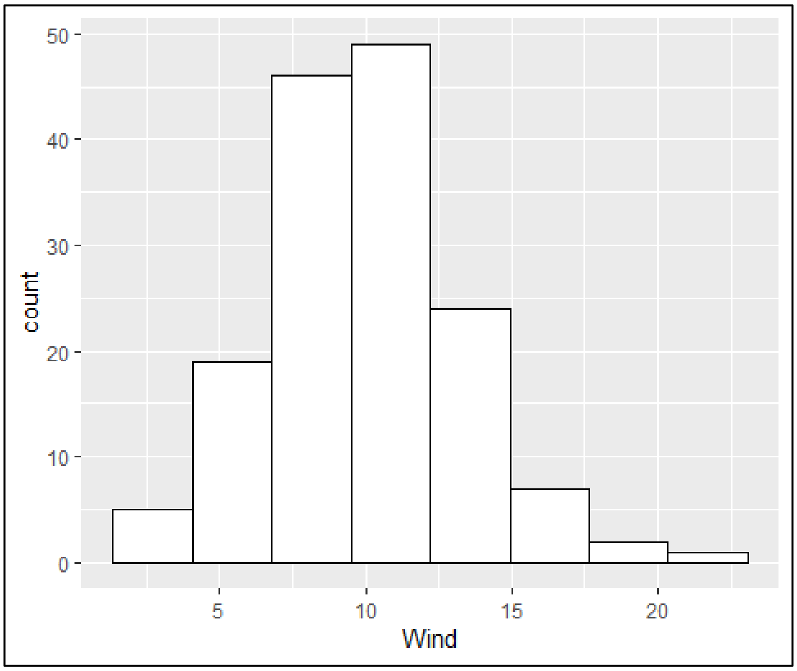
```
#Generate the following visualizations using ggplot
ggplot(aq, aes(x=Ozone)) + geom_histogram(bins=8, color = "black", fill = "wh
ite") #histogram of ozone
```



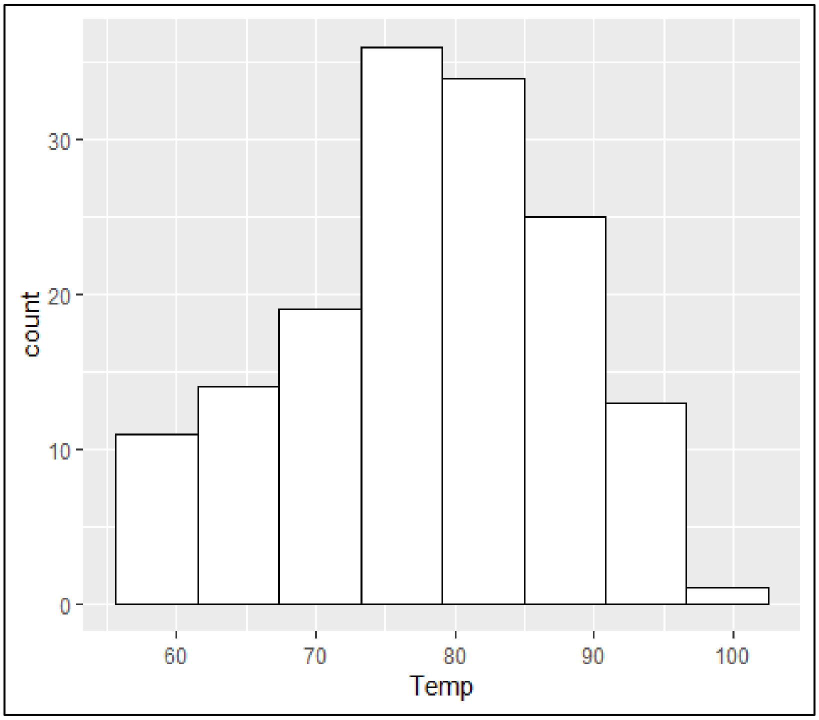
```
ggplot(aq, aes(x=Solar.R)) + geom_histogram(bins=8, color = "black", fill = "
white") #histogram of Solar.R
```



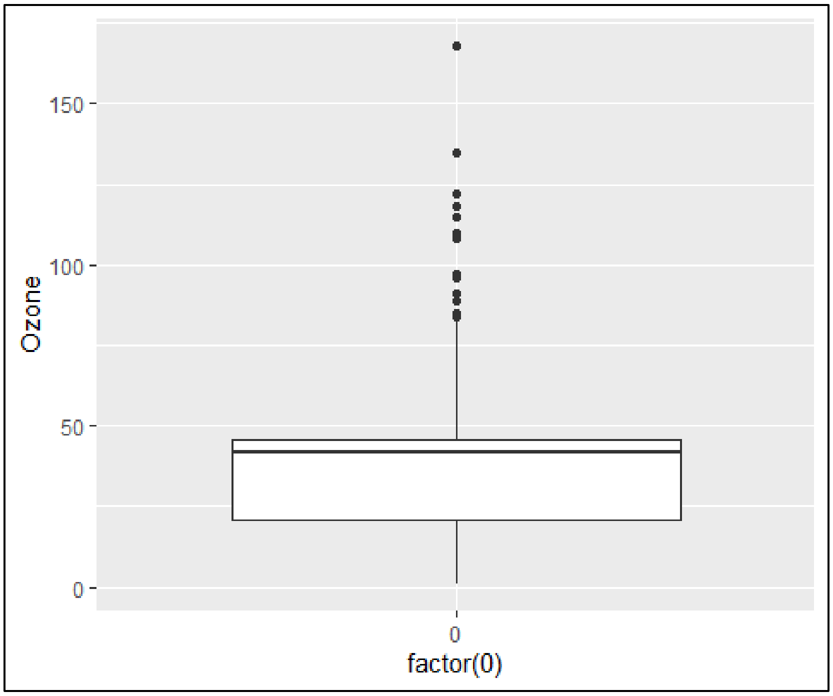
```r
ggplot(aq, aes(x=Wind)) + geom_histogram(bins=8, color = "black", fill = "white") #histogram of Wind
```



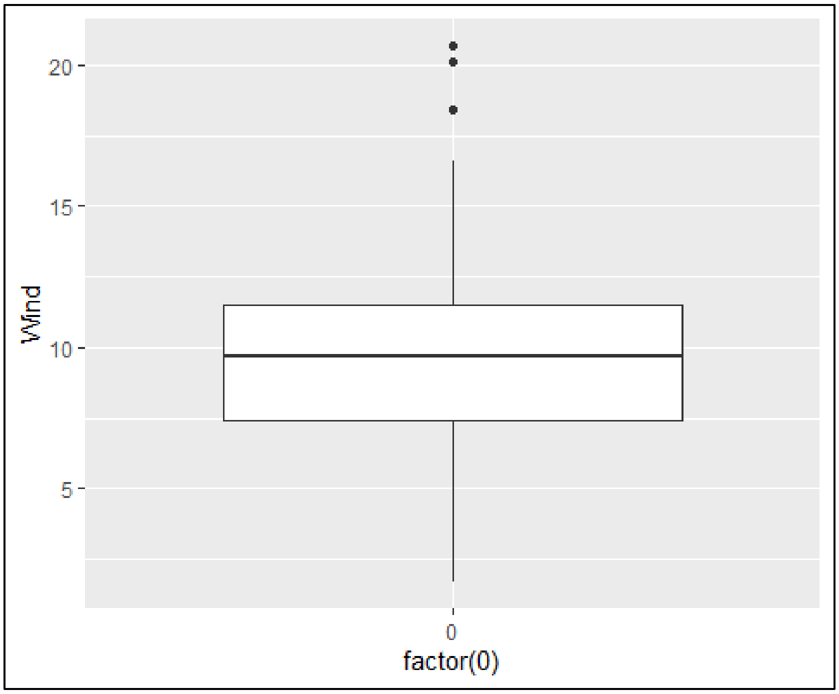
```r
ggplot(aq, aes(x=Temp)) + geom_histogram(bins=8, color = "black", fill = "white") #histogram of Temp
```



```
ggplot(aq, aes(x=factor(0),Ozone))+ geom_boxplot() #Boxplot for Ozone
```

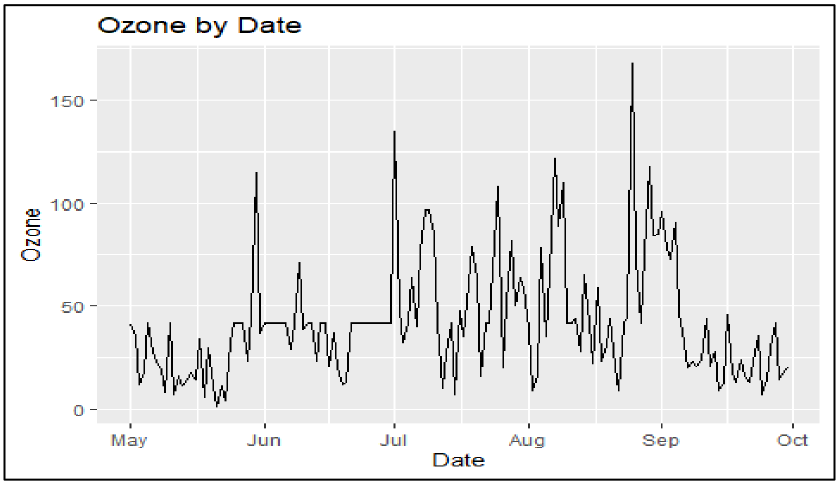


```
ggplot(aq, aes(x=factor(0),Wind))+ geom_boxplot() #Boxplot for wind
```

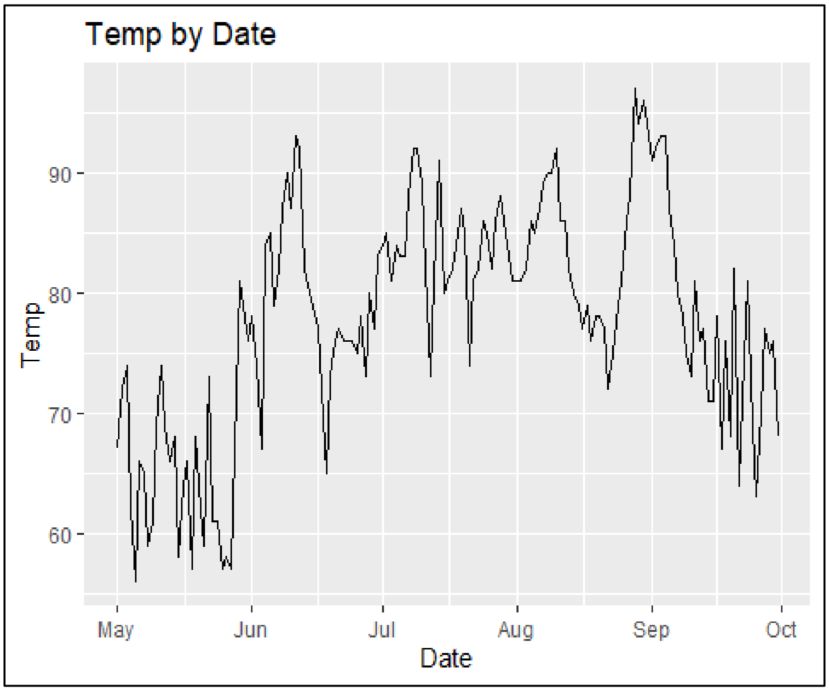


Step 3: Explore how the data changes over time

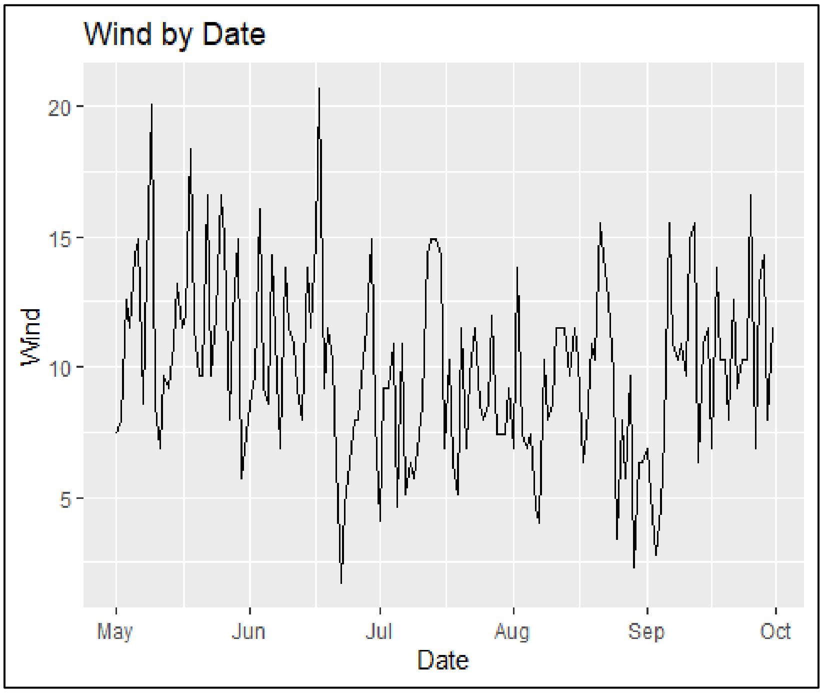
```
#Cleaning up the dates
aq$Year <- "1973" #Add a column for year
aq$Month <- paste(0,aq$Month, sep = "") #Add leading zero to months as needed
aq$Day[which(nchar(aq$Day) == 1)] <- paste(0, aq$Day[which(nchar(aq$Day) == 1
)], sep = "") #Add leading zero to day as needed
aq$Date <- paste(aq$Month, aq$Day, aq$Year, sep = "/")
aq$Date <- as.Date(aq$Date, "%m/%d/%Y") #reformat the date column as a date m
ode

#Ozone line chart
ggplot(aq,aes(x=Date, y=Ozone)) + geom_line() + ggtitle("Ozone by Date")
```



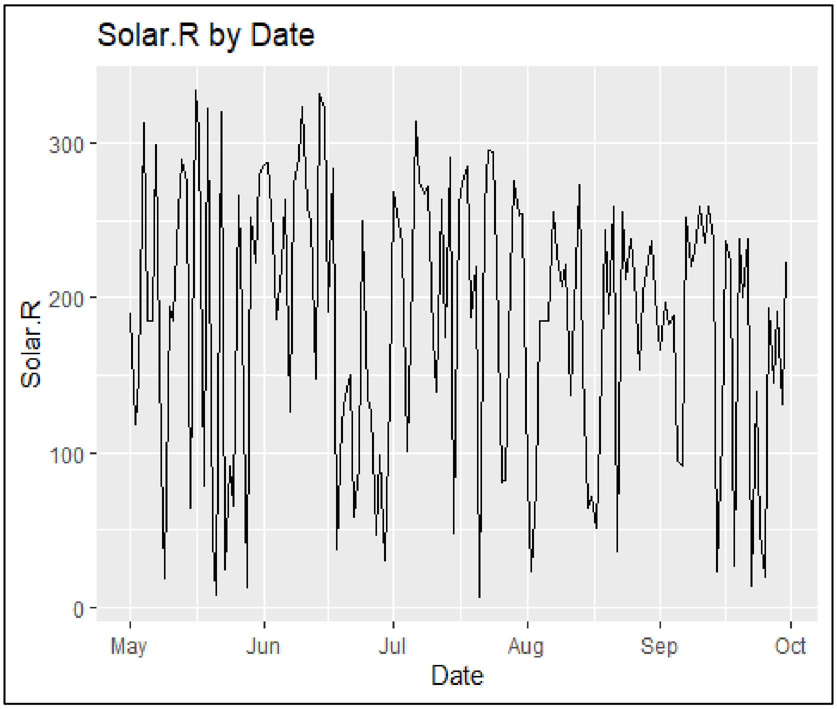
```
#Temp line chart
ggplot(aq,aes(x=Date, y=Temp)) + geom_line() + ggtitle("Temp by Date")
```



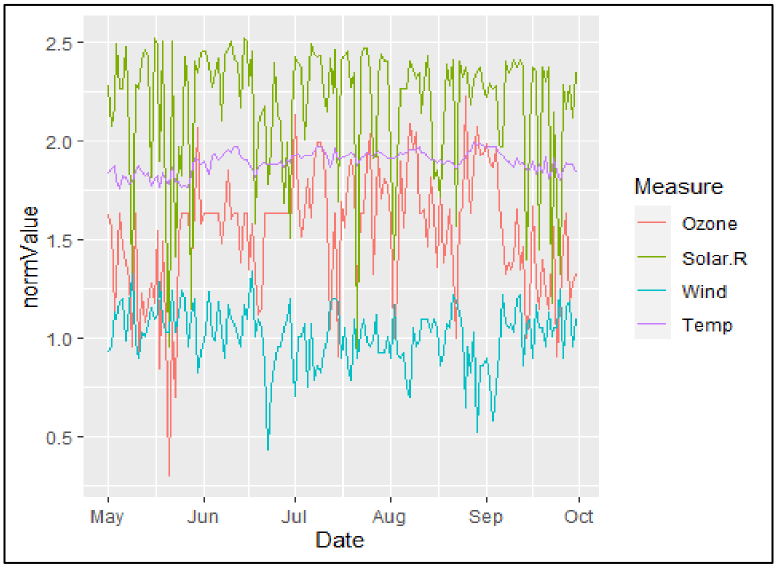
```
#Wind line chart
ggplot(aq,aes(x=Date, y=Wind)) + geom_line() + ggtitle("Wind by Date")
```



```
#Solar.R line chart
ggplot(aq,aes(x=Date, y=Solar.R)) + geom_line() + ggtitle("Solar.R by Date")
```

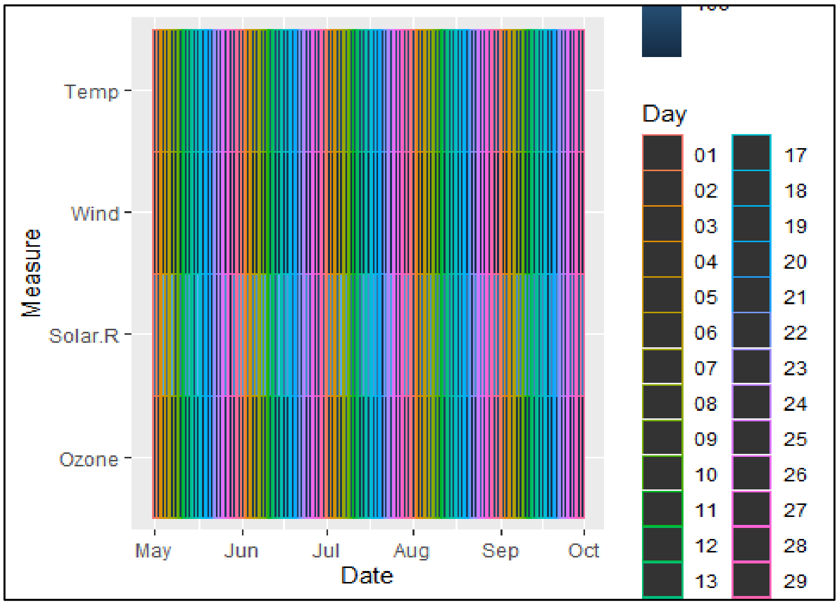


```
#Combined line chart
aqflat <- melt(aq, c("Date","Month","Day","Year"), variable.name = "Measure",
value.name = "Value") #new dataframe that is in tabular form
aqflat$normValue <- log10(aqflat$normValue <- aqflat$Value+1) #rescaled  valu
es
ggplot(aqflat, aes(x=Date, group=Measure, color=Measure)) + geom_line(aes(y=n
ormValue)) #Combined line chart
```



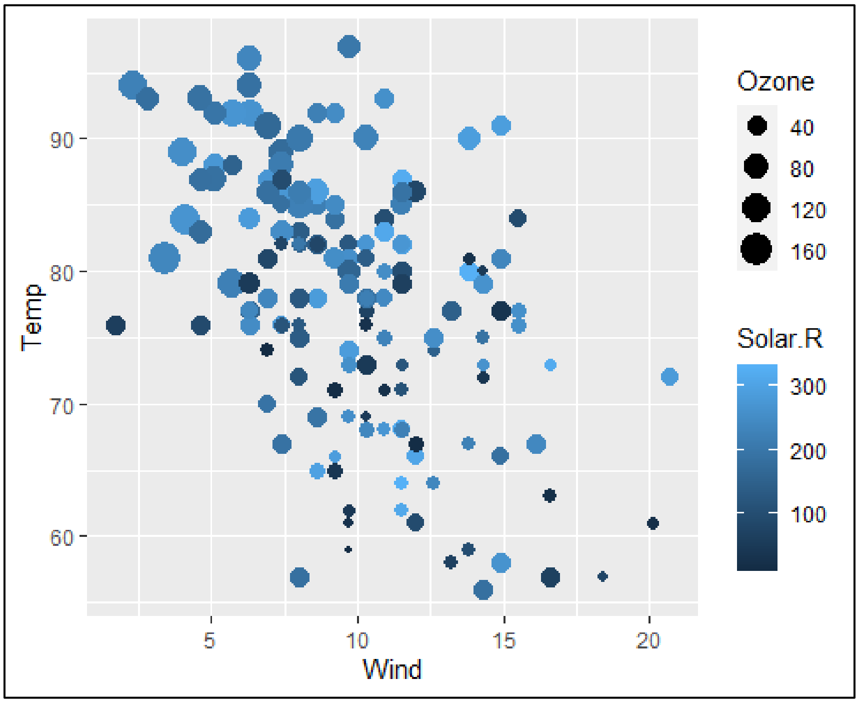
Step 4: Look at all the data via a Heatmap

```
ggplot(aqflat, aes(x=Date, y=Measure, group=Day, color=Day, fill=Day)) + geom
_tile(aes(fill=Value))
```



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

```
#scatter plot for airquality measures
ggplot(aq, aes(Wind, Temp)) + geom_point(aes(size=Ozone, color=Solar.R))
```



Step 6: Final Analysis

Do you see any patterns after exploring the data?

I noticed that there is a negative correlation between the temperature and wind variables. In addition, there is a positive correlation between the temperature and ozone variables. Another interesting observation was that the wind and ozone variables have a very similar looking frequency distribution. Something else that stood out to me was that temperatures over 95 degrees are extremely rare as only a couple of these cases were recorded.

What was the most useful visualization?

I felt that the scatter plot was the most effective visualization because I was able to very quickly draw meaningful conclusion from the data. For example, it only took me a moment to see that were relationships between the variables. I also like the scatter plot because it demarcates the shape of the data while staying easily readable. Although the other visualizations provided some information as well, I believe that the scatter plot helped me get the best understanding of the data in the shortest amount of time.