# IST687 - Week 6 - Viz HW: air quality Analysis

*John Fields*

*5/7/2019*

### Step 1: Load the data

We will use the airquality data set, which you should already have as part of your R installation.

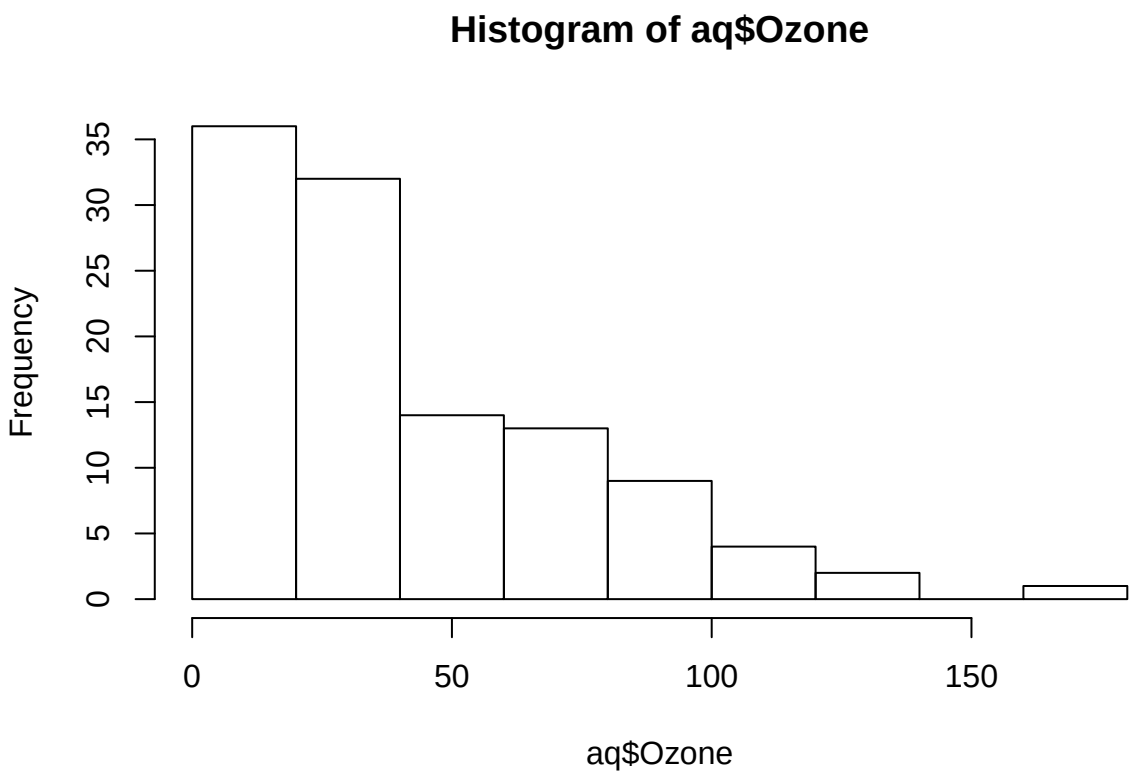
### Step 2: Clean the data

After you load the data, there will be some NAs in the data. You need to figure out what to do about those nasty NAs.

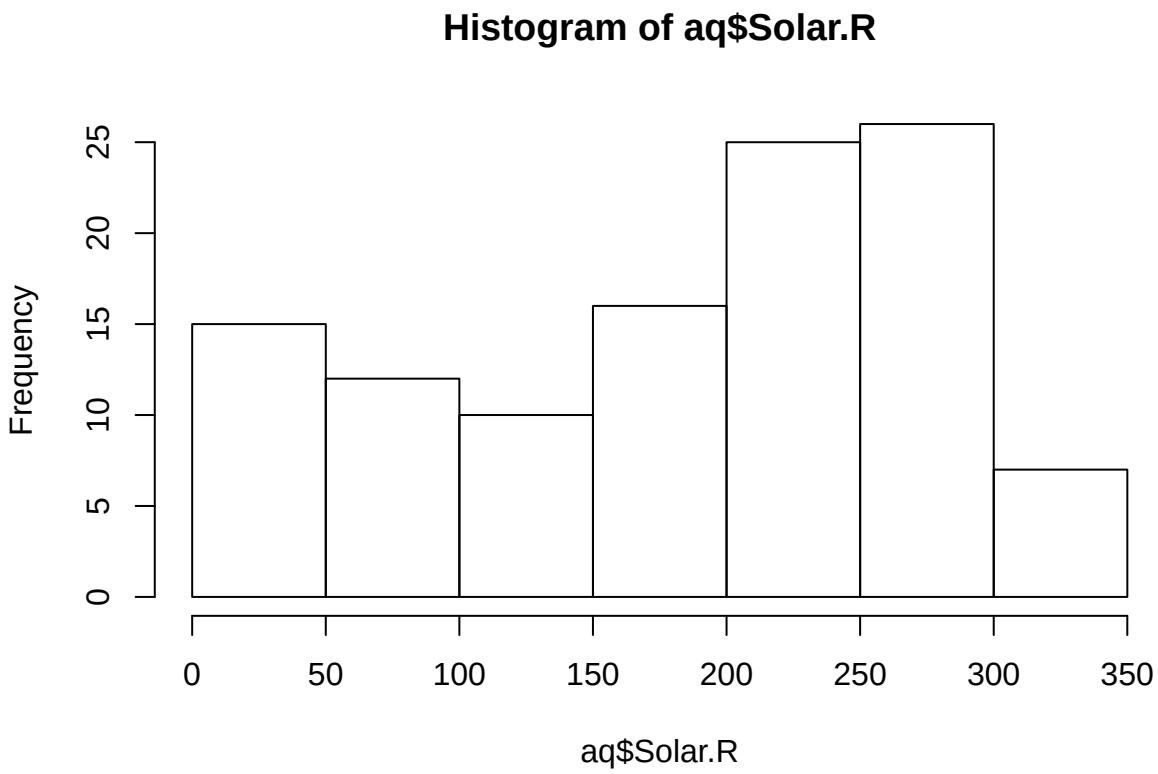
### Step 2: Understand the data distribution

Create the following visualizations using ggplot: • Histograms for each of the variables • Boxplot for Ozone • Boxplot for wind values (round the wind to get a good number of "buckets")

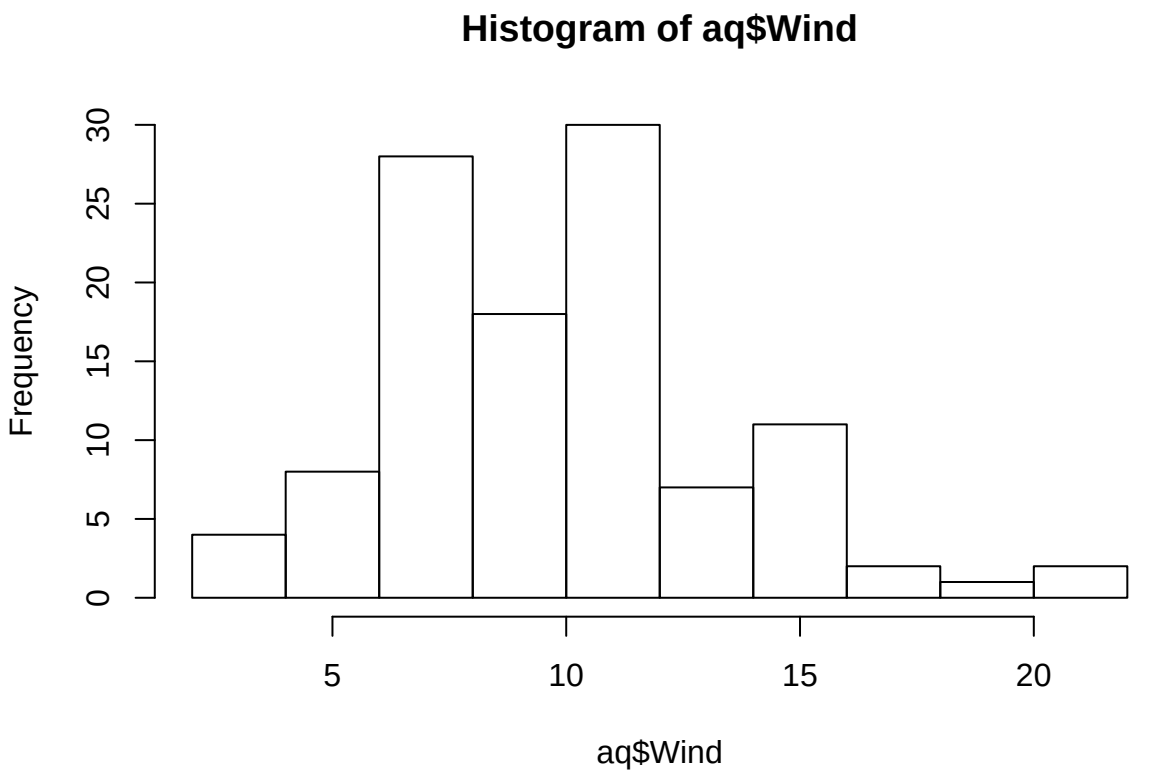
```
hist(aq$Ozone)
```



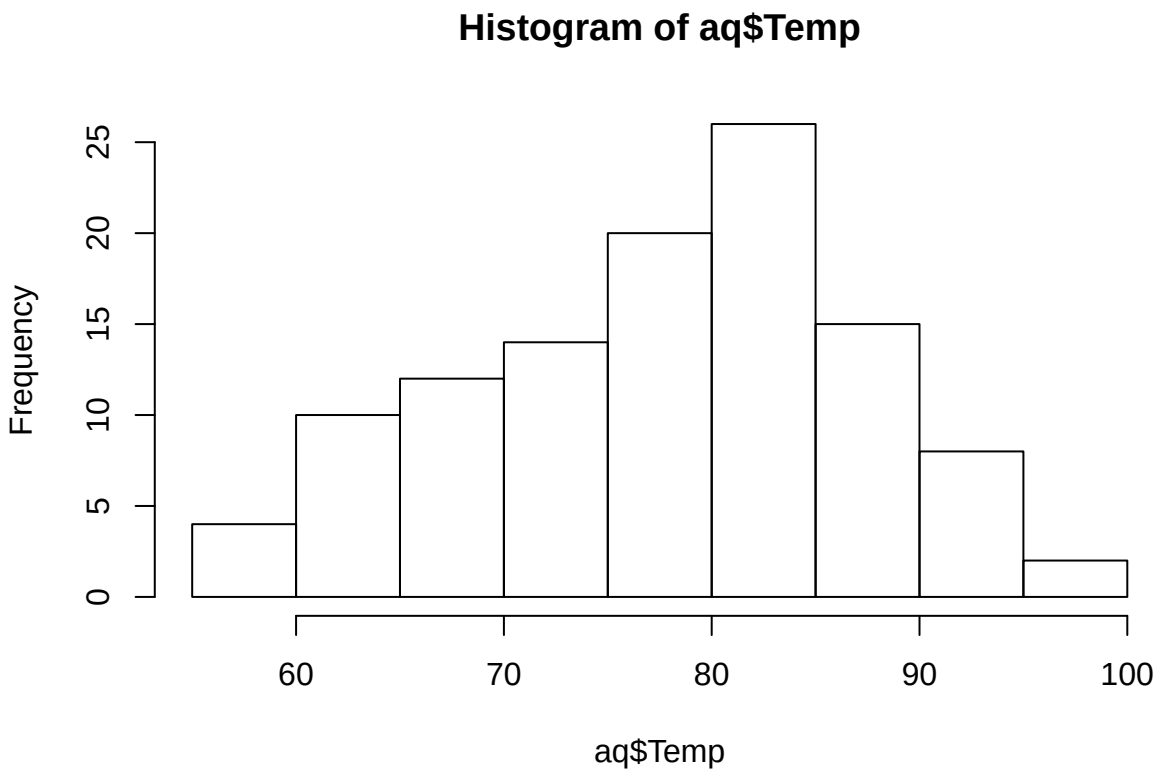

```
hist(aq$Solar.R)
```



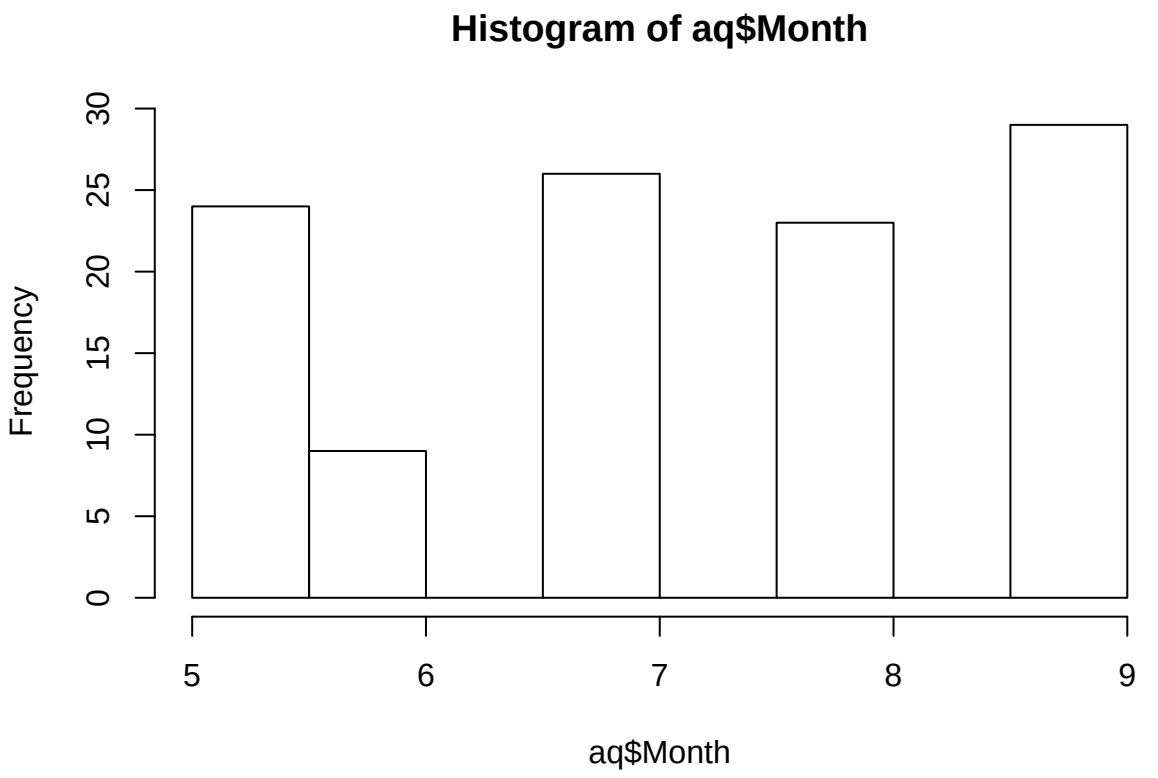

```
hist(aq$Wind)
```



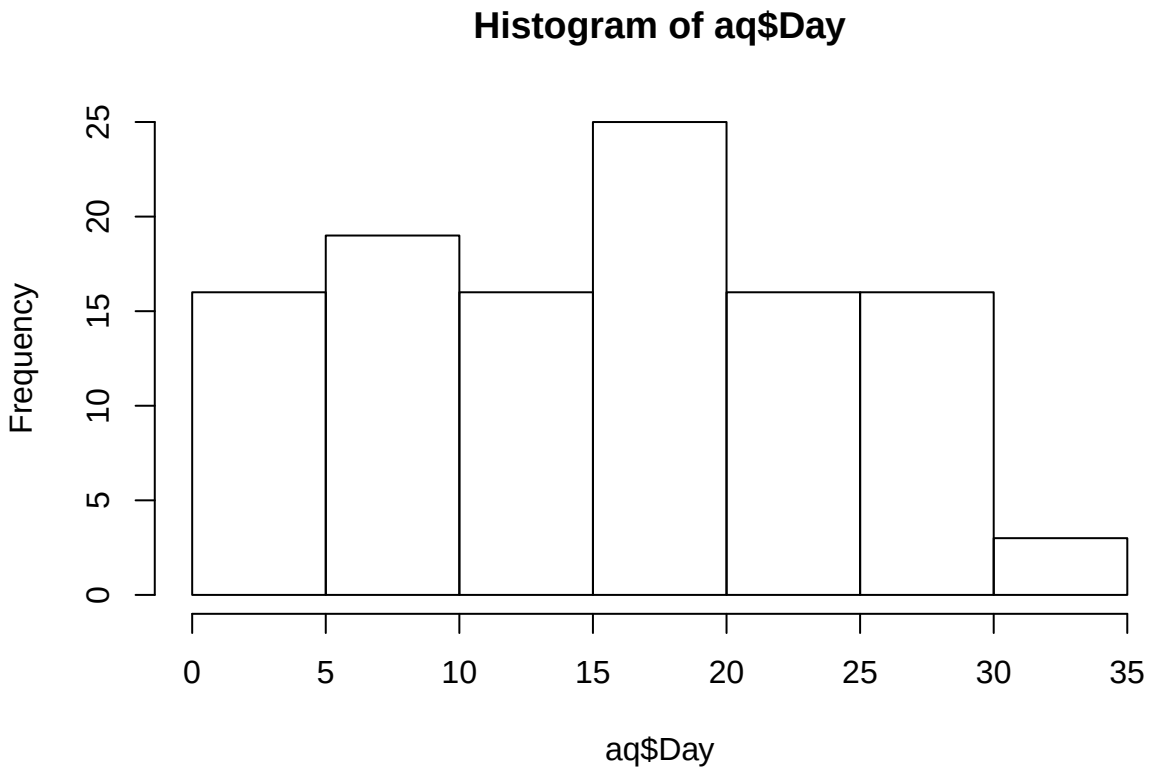

```
hist(aq$Temp)
```



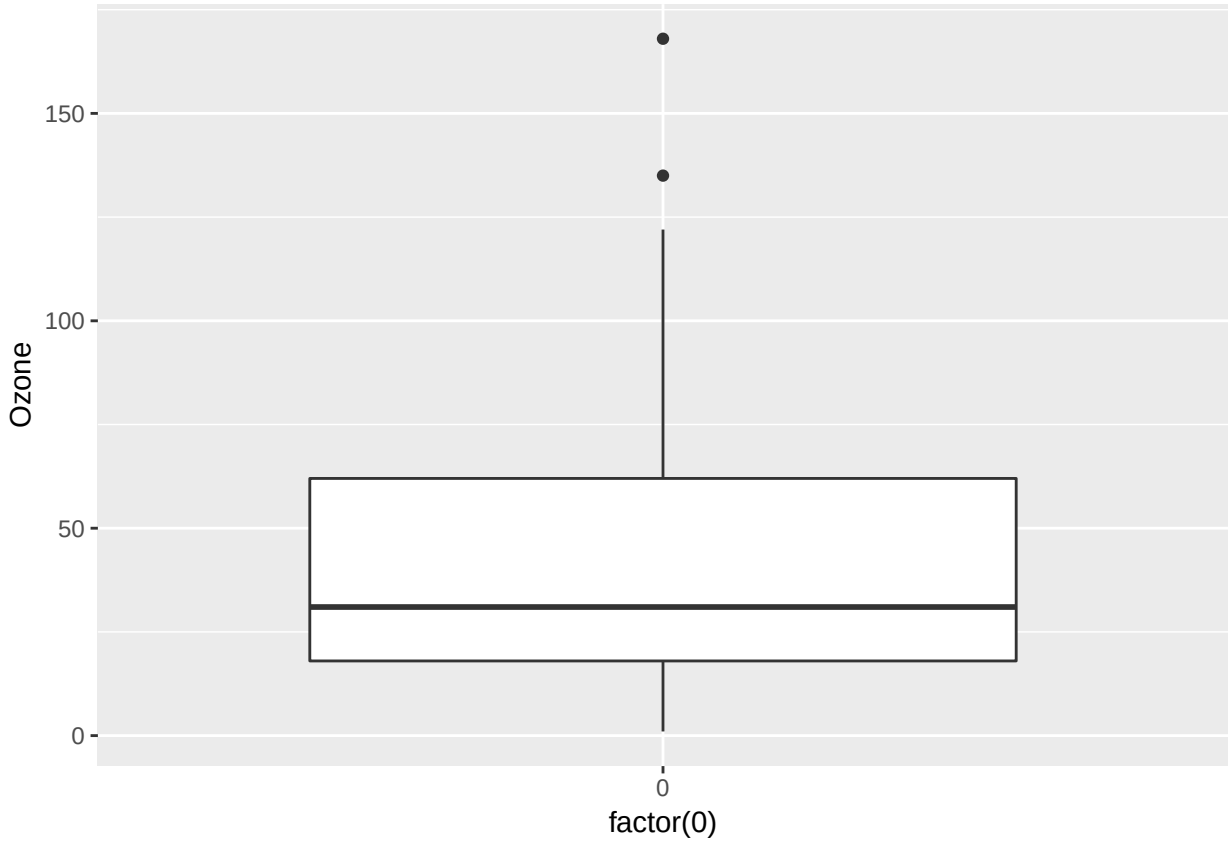

```
hist(aq$Month)
```



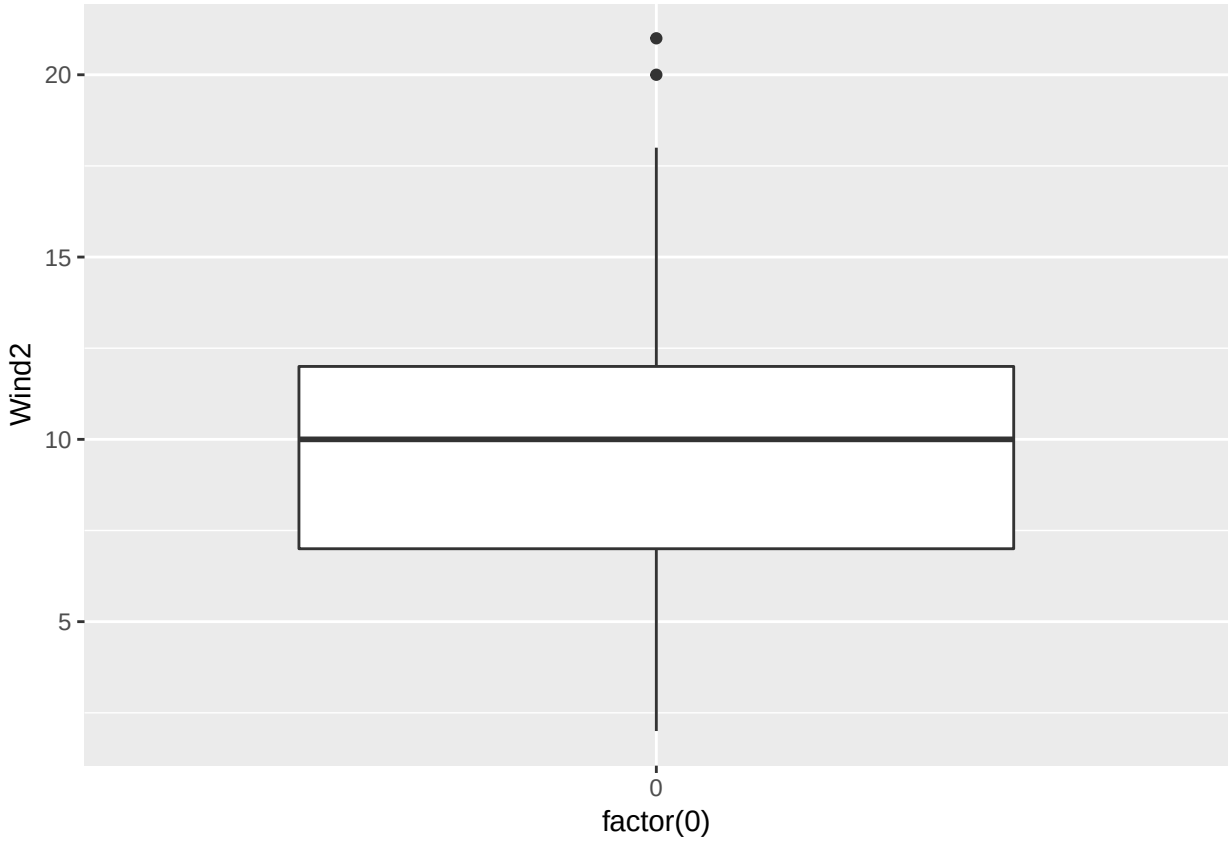

```
hist(aq$Day)
library("ggplot2")
```




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



```
aq$Wind2<-(round(aq$Wind))
ggplot(aq,aes(x=factor(0),Wind2))+geom_boxplot()
```



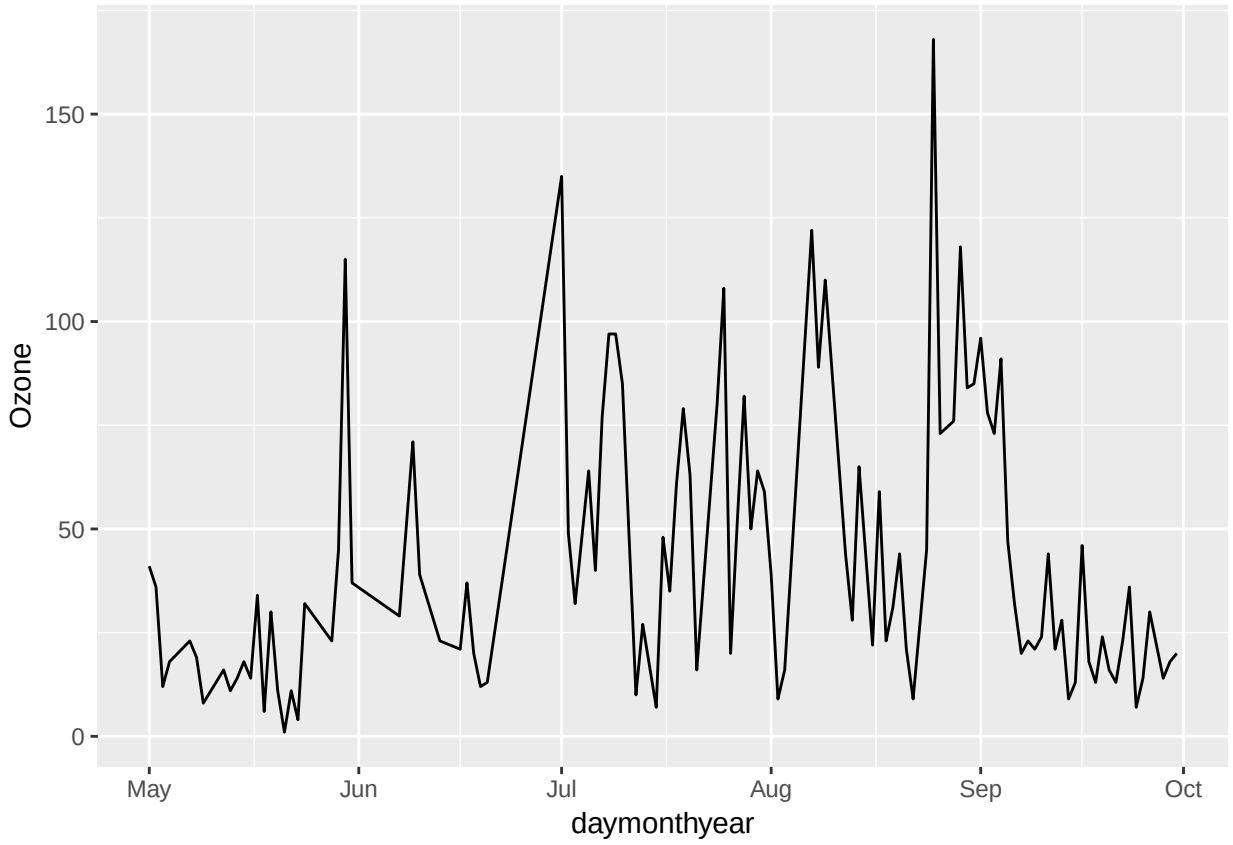

## Step 3: Explore how the data changes over time First, make sure to create appropriate dates (this data was from 1973). Then create line charts for ozone, temp, wind and solar.R (one line chart for each, and then one chart with 4 lines, each having a different color). Create these visualizations using ggplot. Note that for the chart with 4 lines, you need to think about how to effectively use the y- axis.

```
aq$Year<-(1973)
aq$Year<-as.integer(aq$Year)
aq$daymonthyear <- as.Date(paste(aq$Month,aq$Day,aq$Year,sep = "." ), format = "%m.%d.%Y")
aq$daymonthyear
```

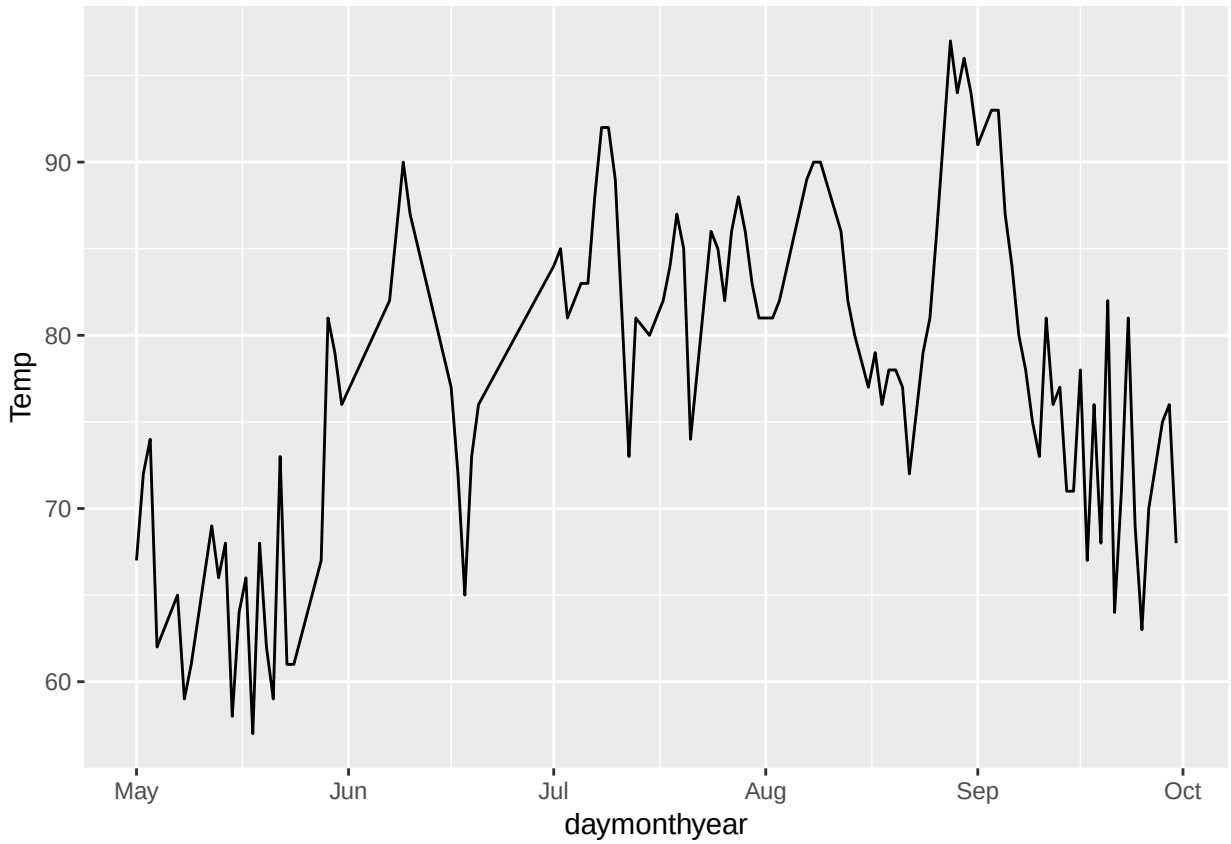
```
##   [1] "1973-05-01" "1973-05-02" "1973-05-03" "1973-05-04" "1973-05-07"
##   [6] "1973-05-08" "1973-05-09" "1973-05-12" "1973-05-13" "1973-05-14"
##  [11] "1973-05-15" "1973-05-16" "1973-05-17" "1973-05-18" "1973-05-19"
##  [16] "1973-05-20" "1973-05-21" "1973-05-22" "1973-05-23" "1973-05-24"
##  [21] "1973-05-28" "1973-05-29" "1973-05-30" "1973-05-31" "1973-06-07"
##  [26] "1973-06-09" "1973-06-10" "1973-06-13" "1973-06-16" "1973-06-17"
##  [31] "1973-06-18" "1973-06-19" "1973-06-20" "1973-07-01" "1973-07-02"
##  [36] "1973-07-03" "1973-07-05" "1973-07-06" "1973-07-07" "1973-07-08"
##  [41] "1973-07-09" "1973-07-10" "1973-07-12" "1973-07-13" "1973-07-15"
##  [46] "1973-07-16" "1973-07-17" "1973-07-18" "1973-07-19" "1973-07-20"
##  [51] "1973-07-21" "1973-07-24" "1973-07-25" "1973-07-26" "1973-07-27"
##  [56] "1973-07-28" "1973-07-29" "1973-07-30" "1973-07-31" "1973-08-01"
##  [61] "1973-08-02" "1973-08-03" "1973-08-07" "1973-08-08" "1973-08-09"
##  [66] "1973-08-12" "1973-08-13" "1973-08-14" "1973-08-16" "1973-08-17"
##  [71] "1973-08-18" "1973-08-19" "1973-08-20" "1973-08-21" "1973-08-22"
```

```
##  [76] "1973-08-24" "1973-08-25" "1973-08-26" "1973-08-28" "1973-08-29"
##  [81] "1973-08-30" "1973-08-31" "1973-09-01" "1973-09-02" "1973-09-03"
##  [86] "1973-09-04" "1973-09-05" "1973-09-06" "1973-09-07" "1973-09-08"
##  [91] "1973-09-09" "1973-09-10" "1973-09-11" "1973-09-12" "1973-09-13"
##  [96] "1973-09-14" "1973-09-15" "1973-09-16" "1973-09-17" "1973-09-18"
## [101] "1973-09-19" "1973-09-20" "1973-09-21" "1973-09-22" "1973-09-23"
## [106] "1973-09-24" "1973-09-25" "1973-09-26" "1973-09-28" "1973-09-29"
## [111] "1973-09-30"
```

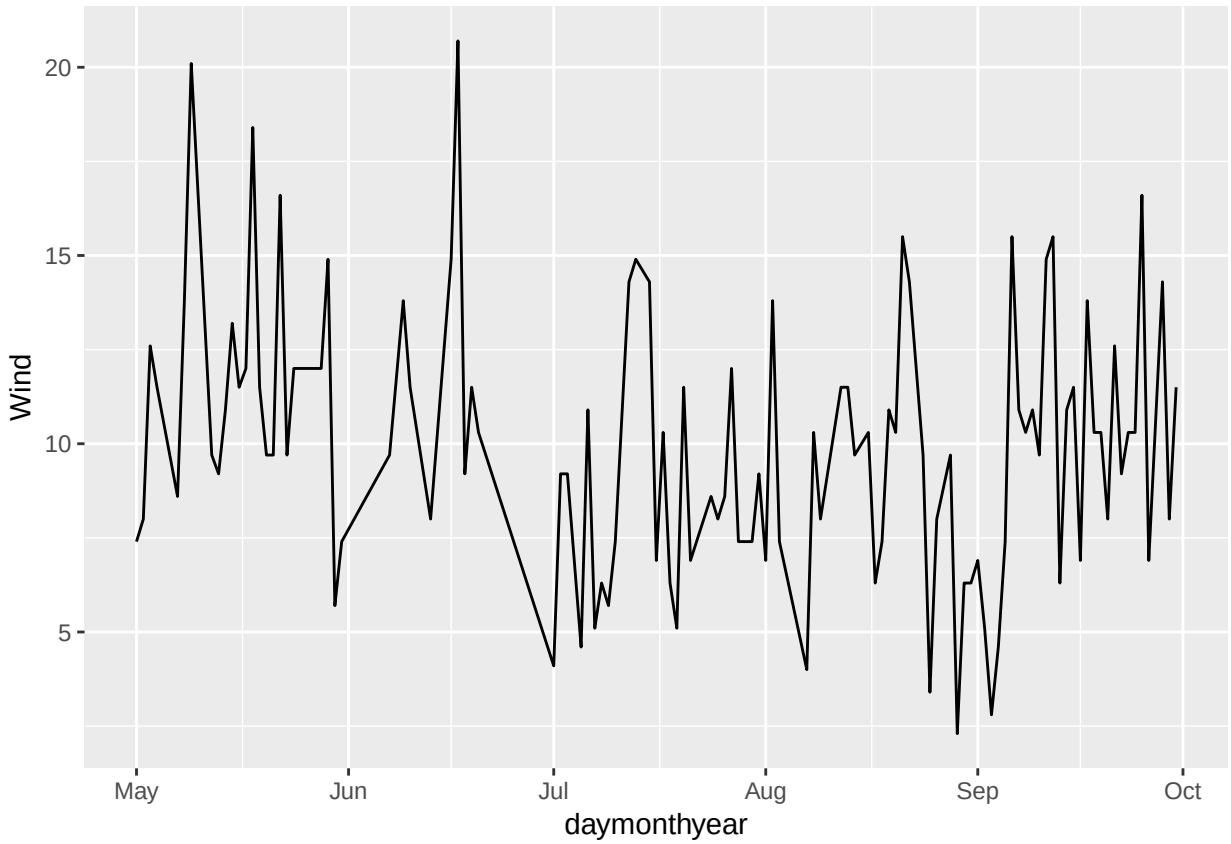
```
ggplot(aq,aes(x=daymonthyear,y=Ozone,group=1)) + geom_line()
```



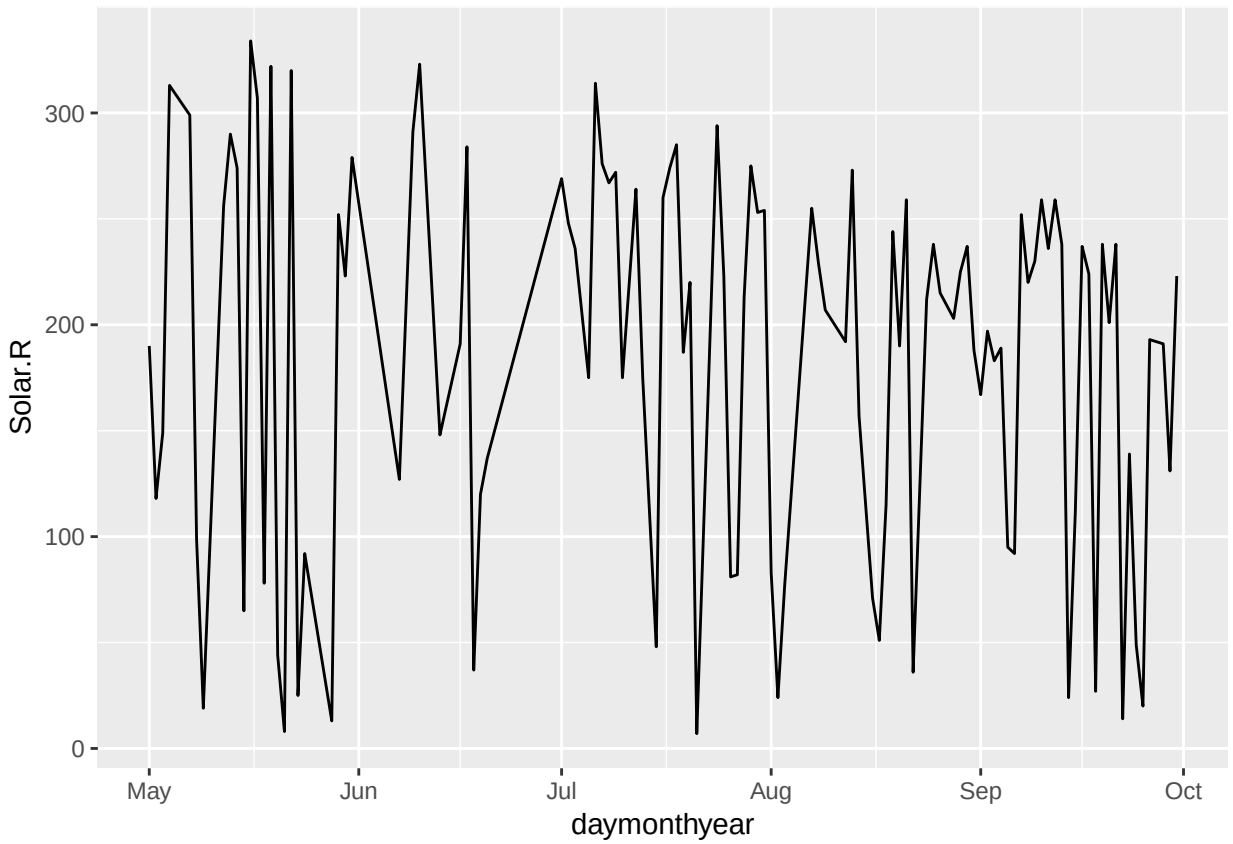

```
ggplot(aq,aes(x=daymonthyear,y=Temp,group=1)) + geom_line()
```



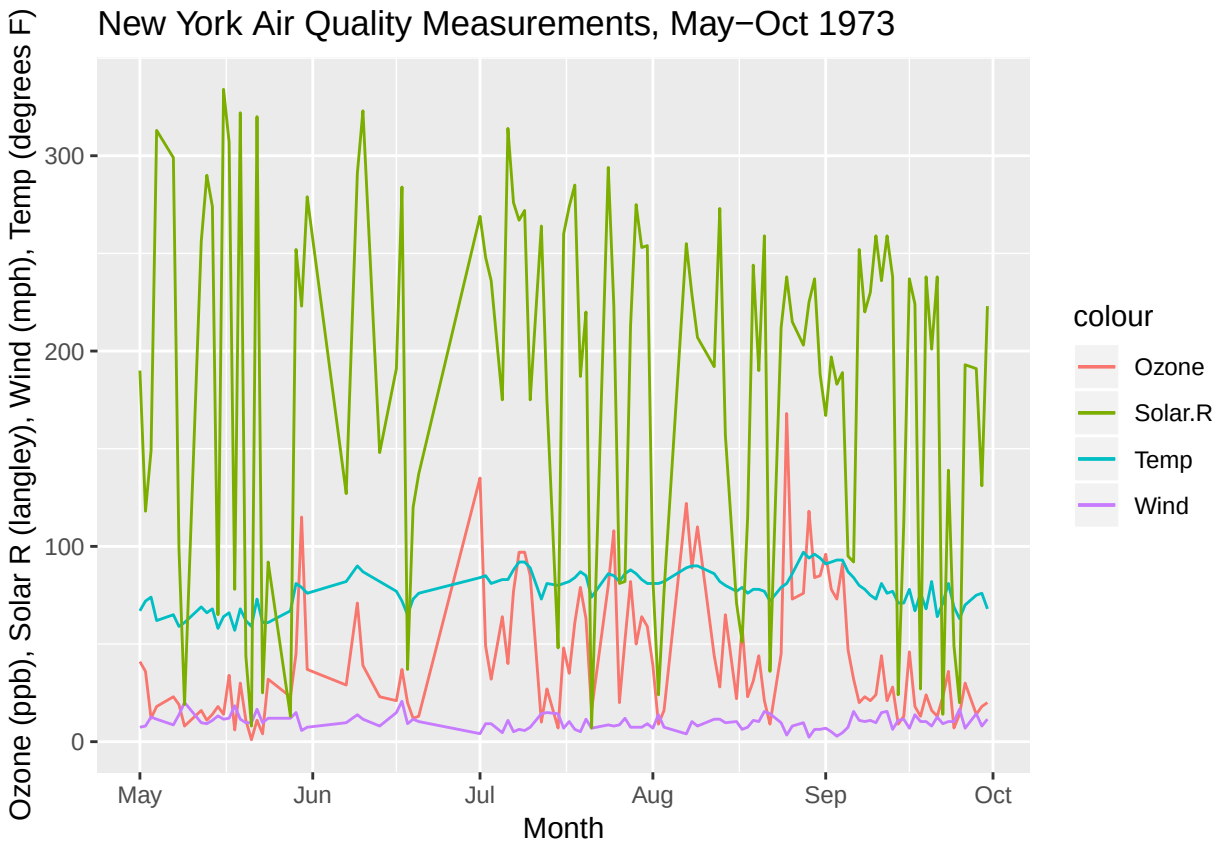

```
ggplot(aq,aes(x=daymonthyear,y=Wind,group=1)) + geom_line()
```




```
ggplot(aq,aes(x=daymonthyear,y=Solar.R,group=1)) + geom_line()
```




```
g<- ggplot(aq,aes(daymonthyear)) + geom_line(aes(y=Ozone,color="Ozone")) + geom_line(aes(y=Temp,color="
g <- g + ylab("Ozone (ppb), Solar R (langley), Wind (mph), Temp (degrees F)") + ggtitle("New York Air Qu
g
```



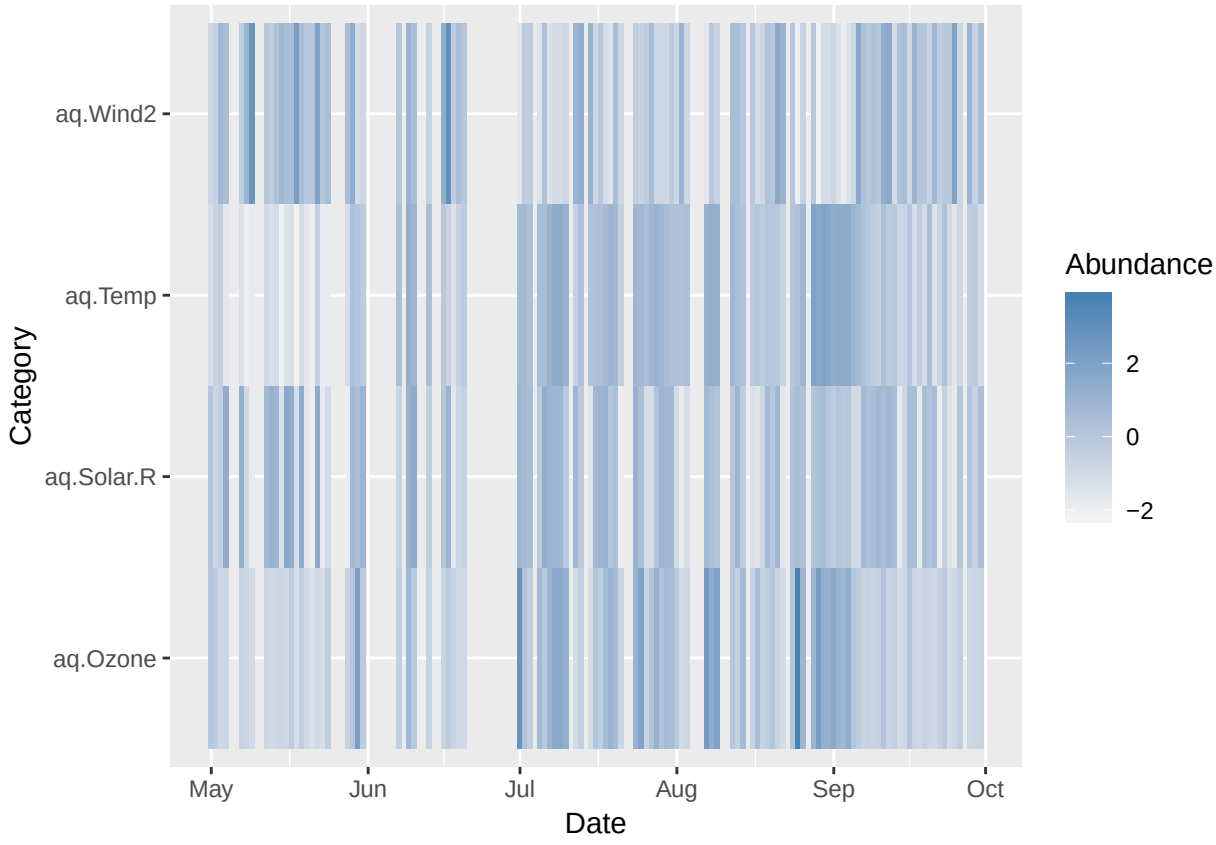

### Step 4: Look at all the data via a Heatmap

Create a heatmap, with each day along the x-axis and ozone, temp, wind and solar.r along the y-axis, and days as rows along the y-axis. Great the heatmap using geom_tile (this defines the ggplot geometry to be 'tiles' as opposed to 'lines' and the other geometry we have previously used). Note that you need to figure out how to show the relative change equally across all the variables.

```
#The heatmap was one of the most challenging exercises in all of the homework so far.
#I consulted the following websites and Courtney Smith to figure out how to scale the
#data and put into key-value pairs using the Gather function.  Many, many hours spent
#on this and I'm pleased with the results!
#https://www.r-bloggers.com/how-to-create-a-fast-and-easy-heatmap-with-ggplot2/
#https://jcoliver.github.io/learn-r/006-heatmaps.html
#https://www.gastonsanchez.com/visually-enforced/how-to/2014/01/15/Center-data-in-R/
#I also collaborated with Courtney Smith on the heatmap as mentioned above.  I tried many
#different color combinations and found that her choice of gray95 and steelblue worked best.
#I used the code --> + scale_fill_gradient(low = "gray95", high = "steelblue")
#with Courtney's permission.

# centering with 'scale()'
mine.data <- data.frame(aq$Ozone,aq$Solar.R,aq$Wind2,aq$Temp)
center_scale <- function(x)
  {
    scale(x, scale = FALSE)
  }
# apply it
```

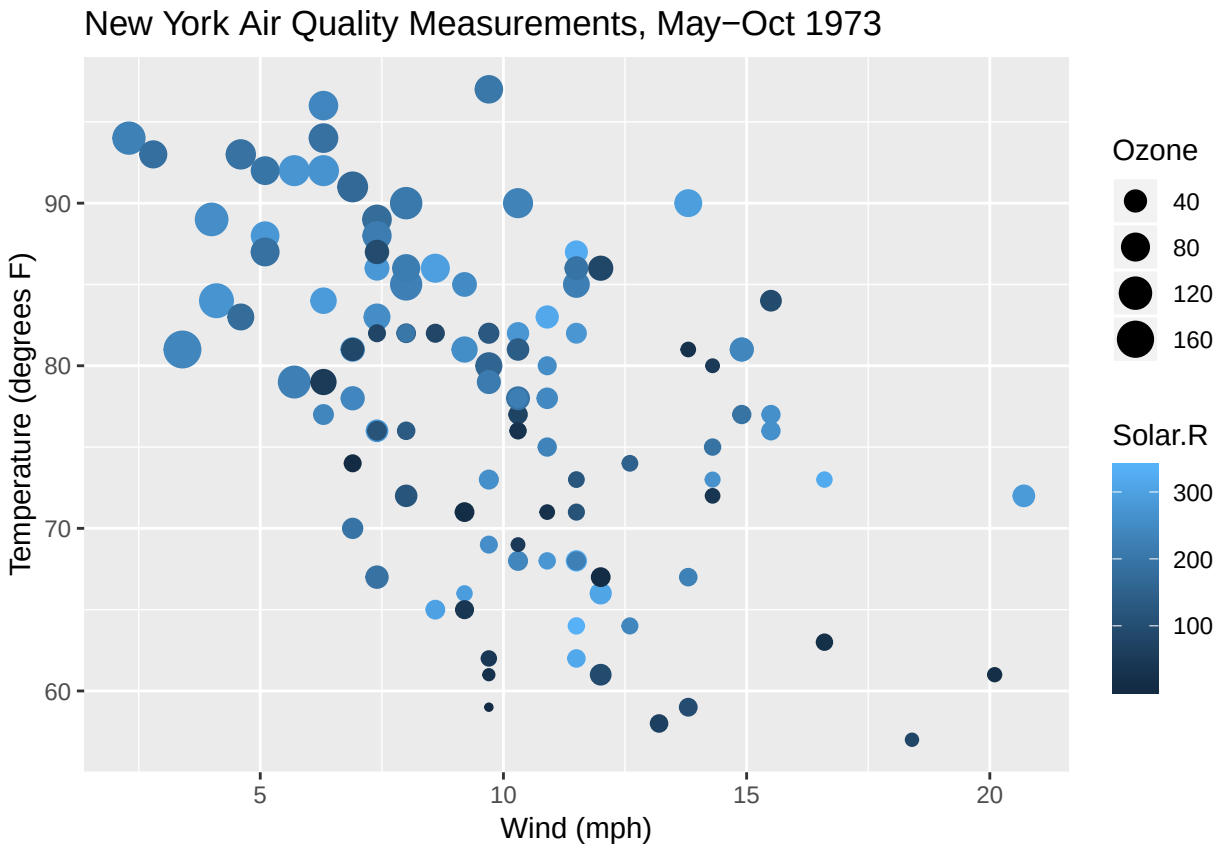
```
mine.data<-scale(mine.data)
mine.data<-data.frame(mine.data,aq$daymonthyear)
library("tidyr")
mine.long <- gather(data = mine.data, key = Category, value = Abundance,-c(5))
#plot it
library("ggplot2")
mine.heatmap <- ggplot(data = mine.long, mapping = aes(x = aq.daymonthyear,y = Category,fill = Abundanc
mine.heatmap
```




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

Create a scatter chart (using ggplot geom_point), with the x-axis representing the wind, the y-axis representing the temperature, the size of each dot representing the ozone and the color representing the solar.R

```
library("ggplot2")
g <- ggplot(aq, aes(x=Wind,y=Temp)) + geom_point(aes(size=Ozone, color=Solar.R)) + ggtitle("New York Ai
g <- g + ylab("Temperature (degrees F)") + xlab("Wind (mph)")
g
```




## Step 6: Final Analysis

- Do you see any patterns after exploring the data?

1. The ozone levels tend to be higher in the summer when there are higher temperatures and less wind.
2. The solar.r levels have a wide range from a minimum of 7 to a maximum of 334. I would question the accuracy of these measurements and determine if the equipment used to collect this information is calibrated and accurate.

```
min(aq$Solar.R)
```

```
## [1] 7
```

```
max(aq$Solar.R)
```

```
## [1] 334
```

- What was the most useful visualization?

When beginning the analysis, the line charts were helpful to understand each of the variables. After this initial analysis, the heat map and scatter plots were helpful to validate some of my theories about the correlation between attributes like higher temperatures/less wind and increased ozone.