Bing-Je_Wu_HW6

Bing-Je Wu

5/12/2019

Step 1: Load the data

```
Mydf <- airquality
```

Step 2: Clean the data

```
colSums(is.na(Mydf))
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

```
str(Mydf)
```

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

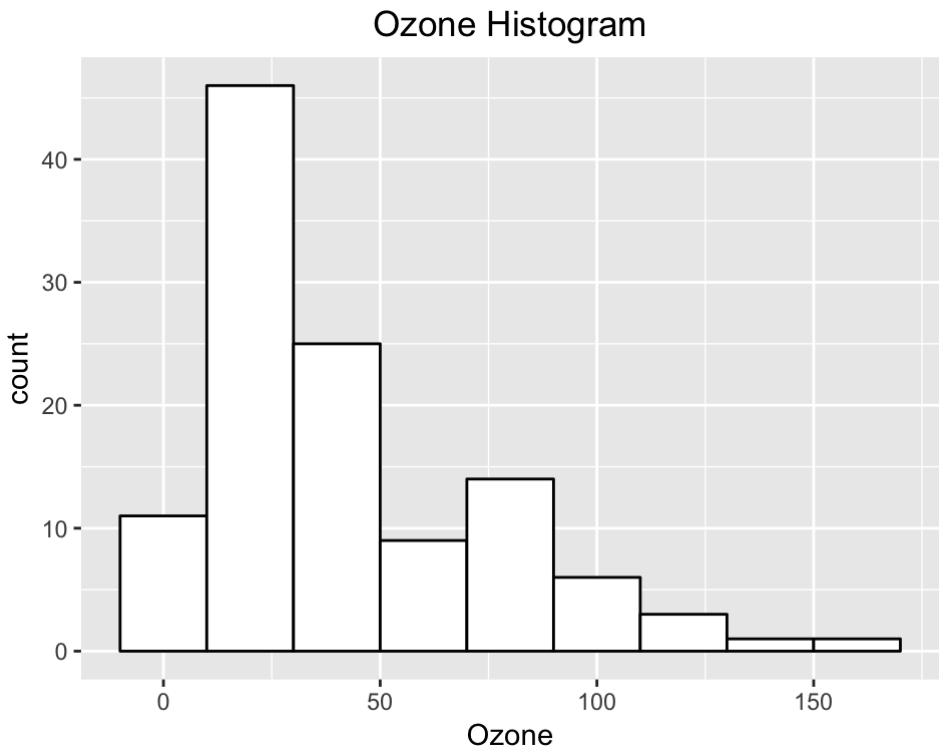
There are two options to deal with missing values
option 1. Using na.omit to remove records with missing values
option 2. Using na.rm on the each function to exclude missing value from analysis
In order to get the most valuable analysis, option 2 is the best solution.

Step 3 -1: Understand the data distribution

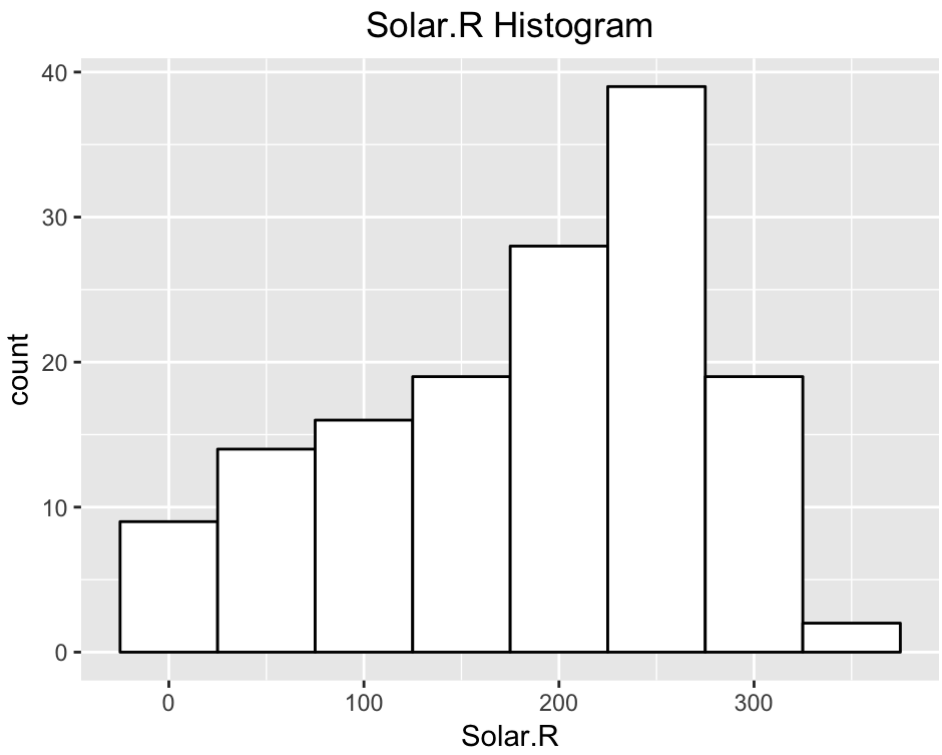
Create the following visualizations using ggplot:

• Histograms for each of the variables

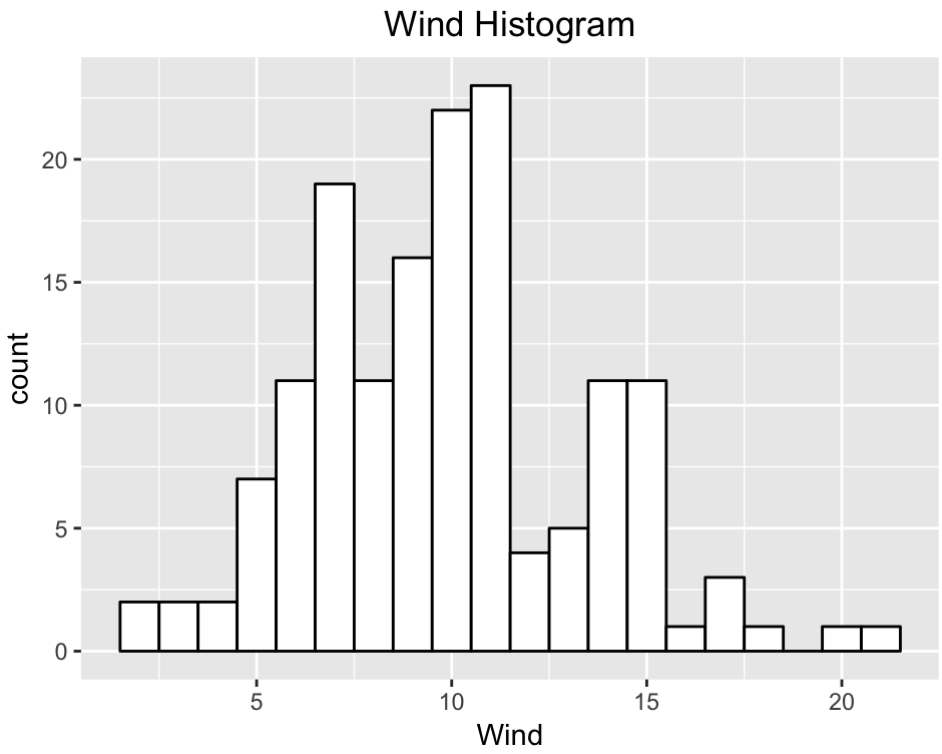
```
ggplot(Mydf, aes(x=Ozone)) +
  geom_histogram(binwidth =20,color="black",fill="white", na.rm = TRUE) +
  ggtitle("Ozone Histogram") + theme(plot.title = element_text(hjust = 0.5))
```



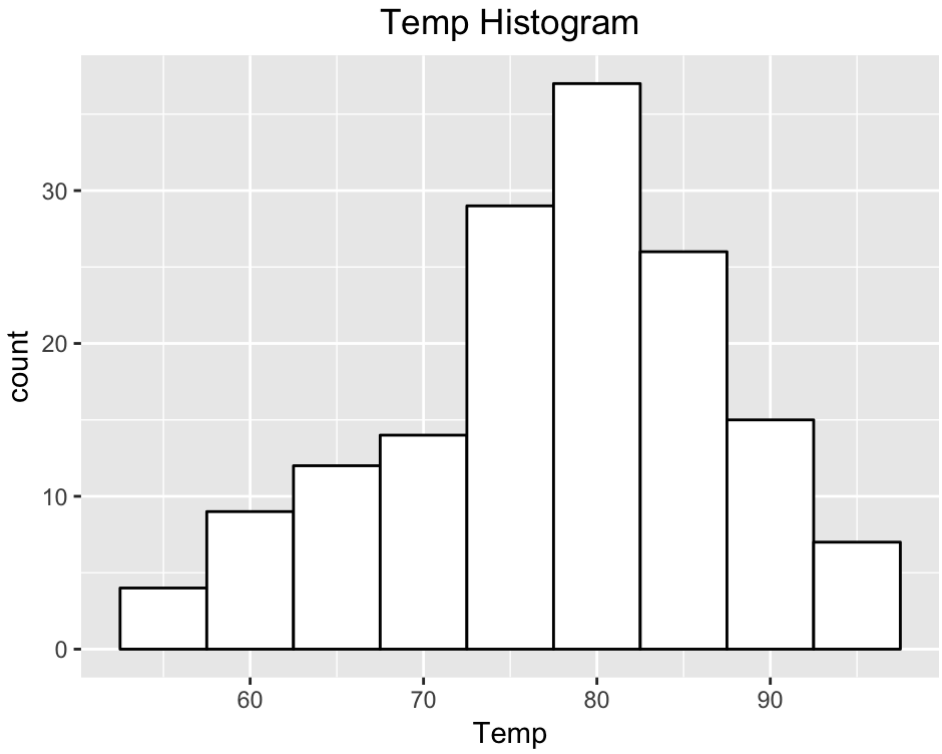

```
ggplot(Mydf, aes(x=Solar.R)) +
  geom_histogram(binwidth =50,color="black",fill="white", na.rm = TRUE) +
  ggtitle("Solar.R Histogram") + theme(plot.title = element_text(hjust = 0.5))
```



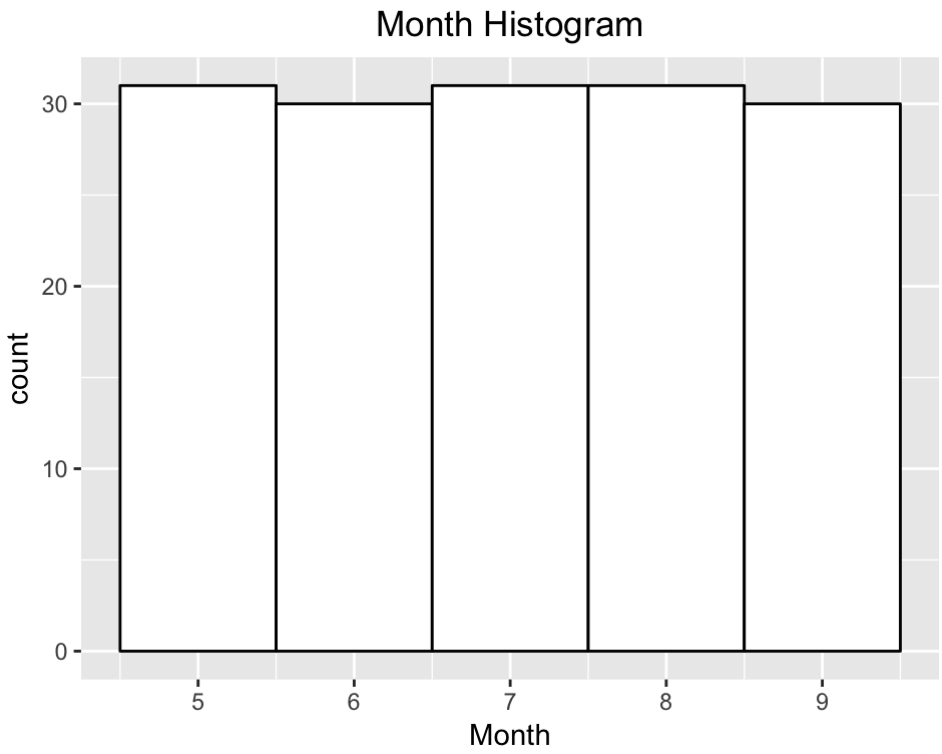

```
ggplot(Mydf, aes(x=Wind)) +
  geom_histogram(binwidth =1, color="black",fill="white", na.rm = TRUE) +
  ggtitle("Wind Histogram") + theme(plot.title = element_text(hjust = 0.5))
```



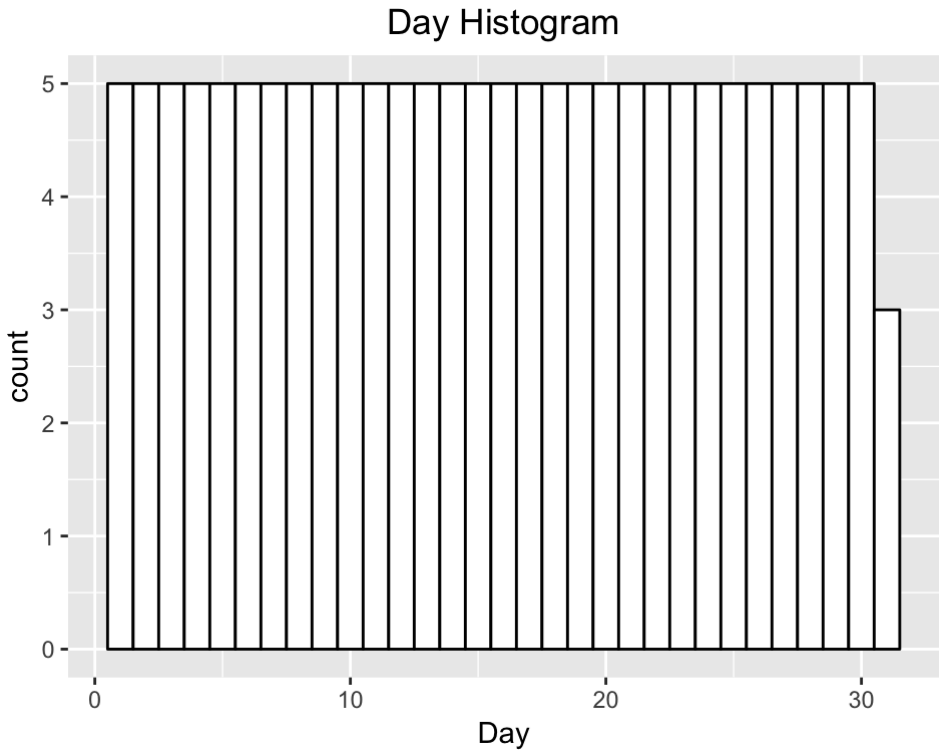

```
ggplot(Mydf, aes(x=Temp)) +
  geom_histogram(binwidth=5, color="black",fill="white", na.rm = TRUE) +
  ggtitle("Temp Histogram") + theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(Mydf, aes(x=Month)) +
  geom_histogram(bins=5, color="black",fill="white", na.rm = TRUE) +
  ggtitle("Month Histogram") + theme(plot.title = element_text(hjust = 0.5))
```

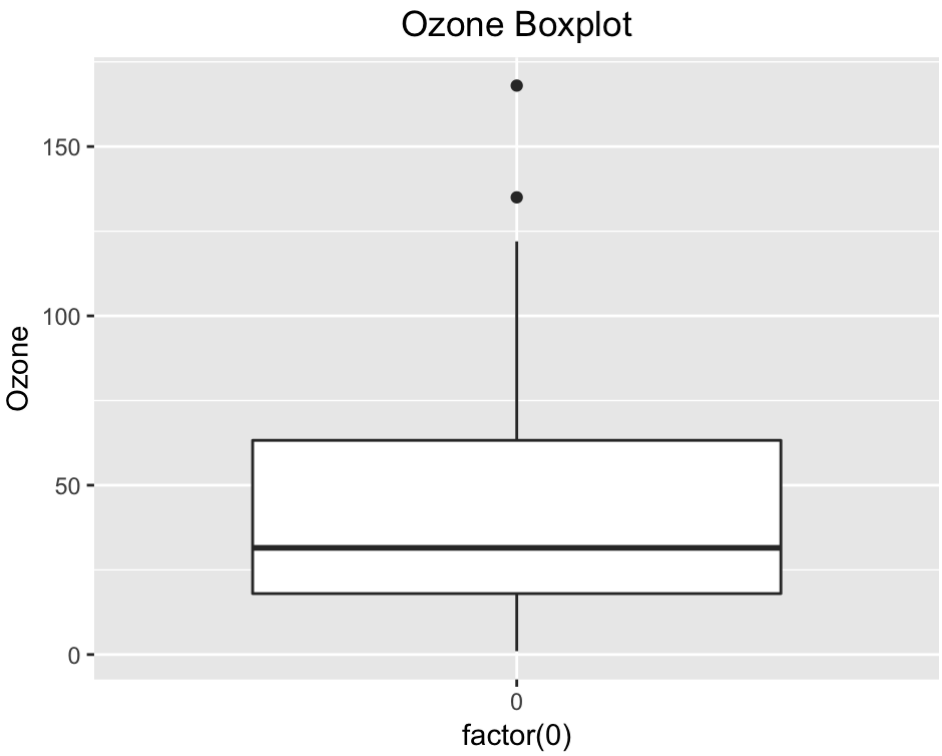



```
ggplot(Mydf, aes(x=Day)) +
  geom_histogram(bins=31,color="black",fill="white", na.rm = TRUE) +
  ggtitle("Day Histogram") + theme(plot.title = element_text(hjust = 0.5))
```



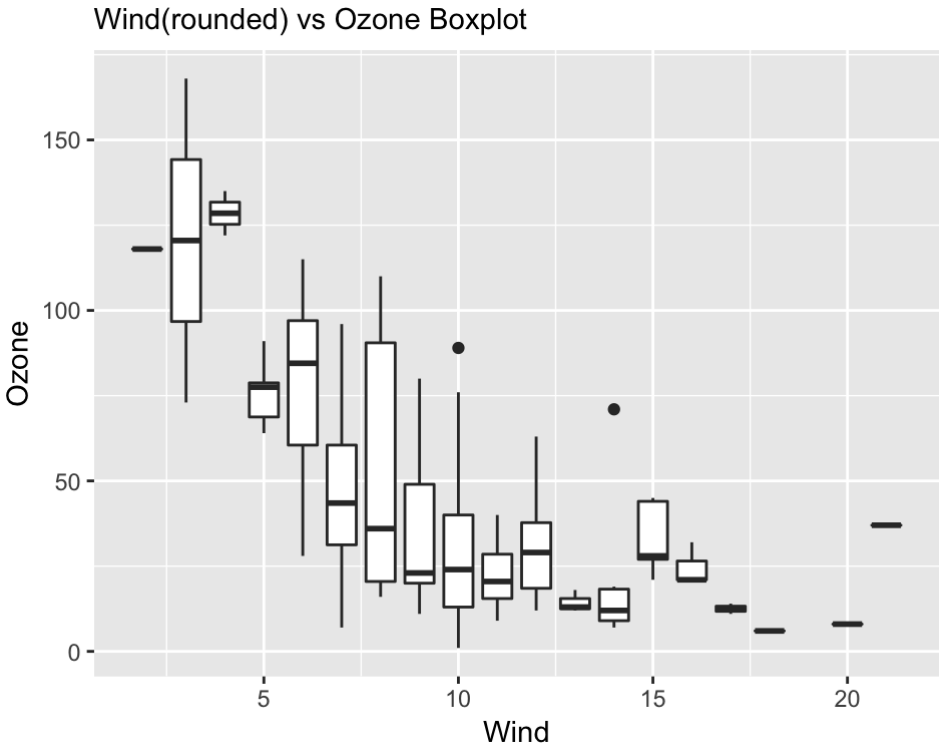
• Boxplot for Ozone

```
ggplot(Mydf, aes(x=factor(0), y=Ozone)) +geom_boxplot(na.rm = TRUE) +
  ggtitle("Ozone Boxplot") + theme(plot.title = element_text(hjust = 0.5))
```



• Boxplot for wind values (round the wind to get a good number of “buckets”)

```
Mydf$Wind_new <- round(Mydf$Wind)
ggplot(Mydf, aes(x=Wind_new, y=Ozone, group=Wind_new)) + geom_boxplot(na.rm = TRUE) +
  labs(subtitle = "Wind(rounded) vs Ozone Boxplot") + xlab("Wind")
```



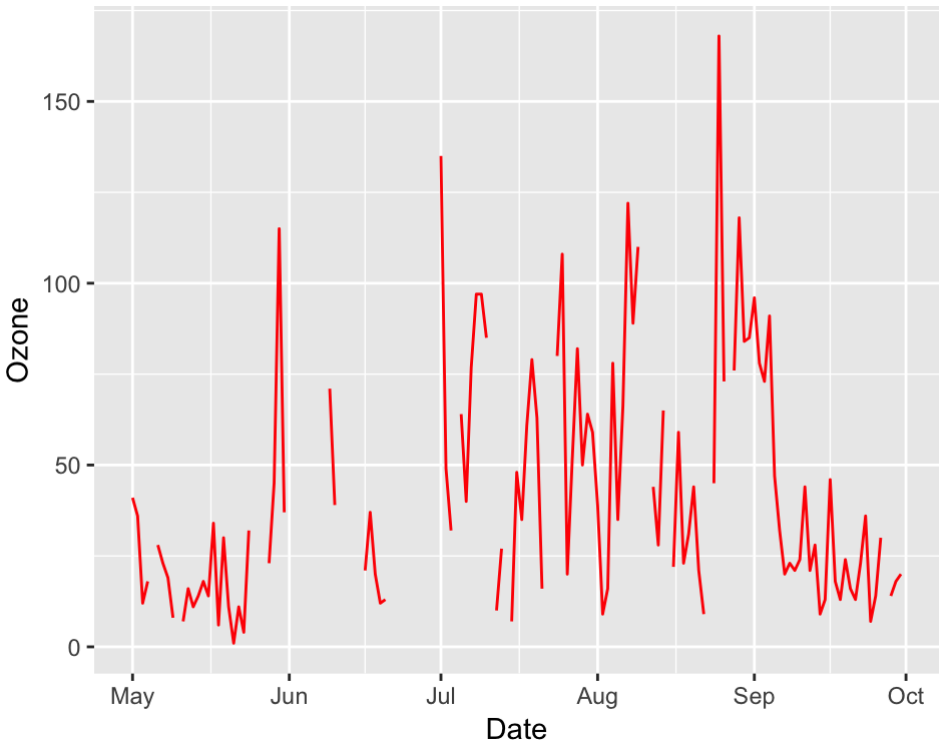
Step 3-2: Explore how the data changes over time

Combine Month and Day as Date

```
Mydf$Date <- as.Date(with(Mydf,paste("1973",Month,Day,sep='-')), "%Y-%m-%d")
```

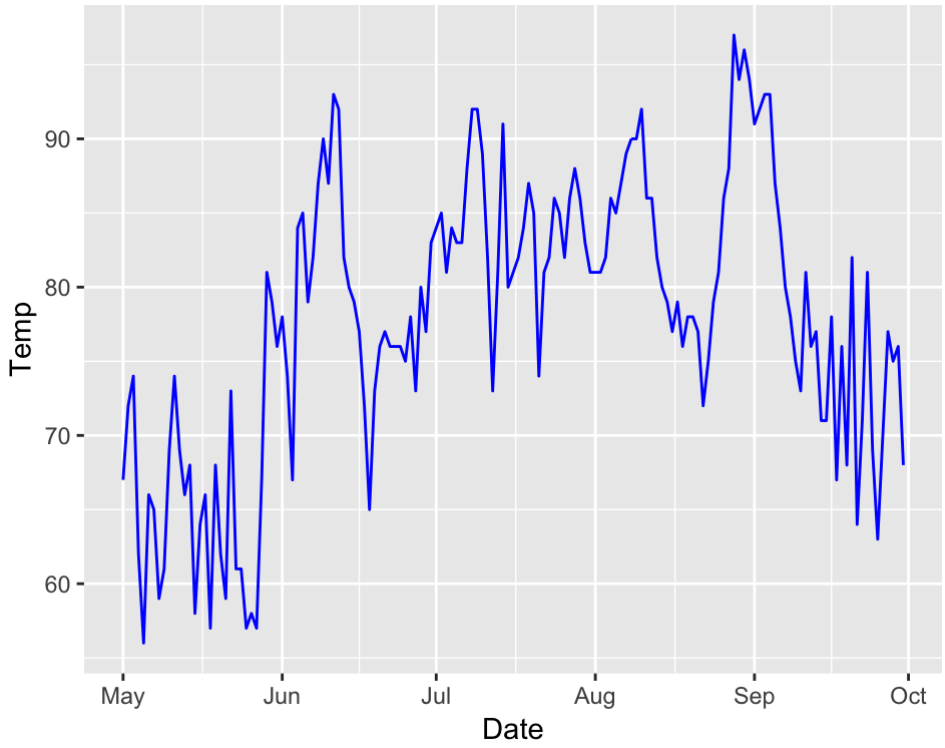
Create line chart for ozone

```
ggplot(Mydf, aes(x=Date, y=Ozone, group=1)) +
  geom_line(color="red", na.rm = TRUE)
```



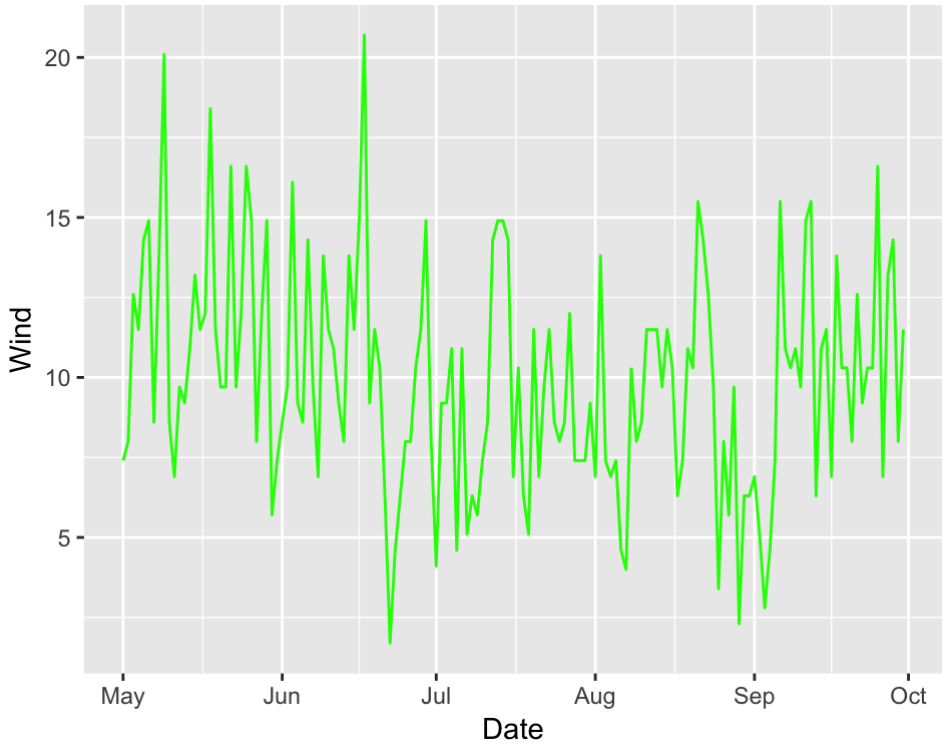
Create line chart for temp

```
ggplot(Mydf, aes(x=Date, y=Temp, group=1)) +
  geom_line(color="blue", na.rm = TRUE)
```



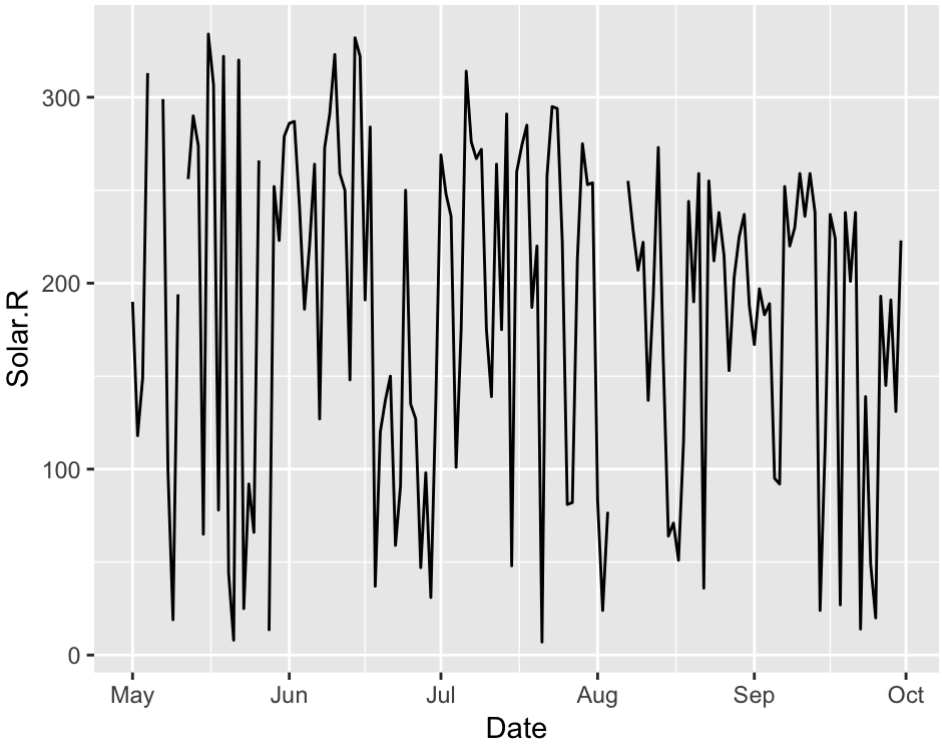
Create line chart for wind

```
ggplot(Mydf, aes(x=Date, y=Wind, group=1)) +
  geom_line(color="green", na.rm = TRUE)
```



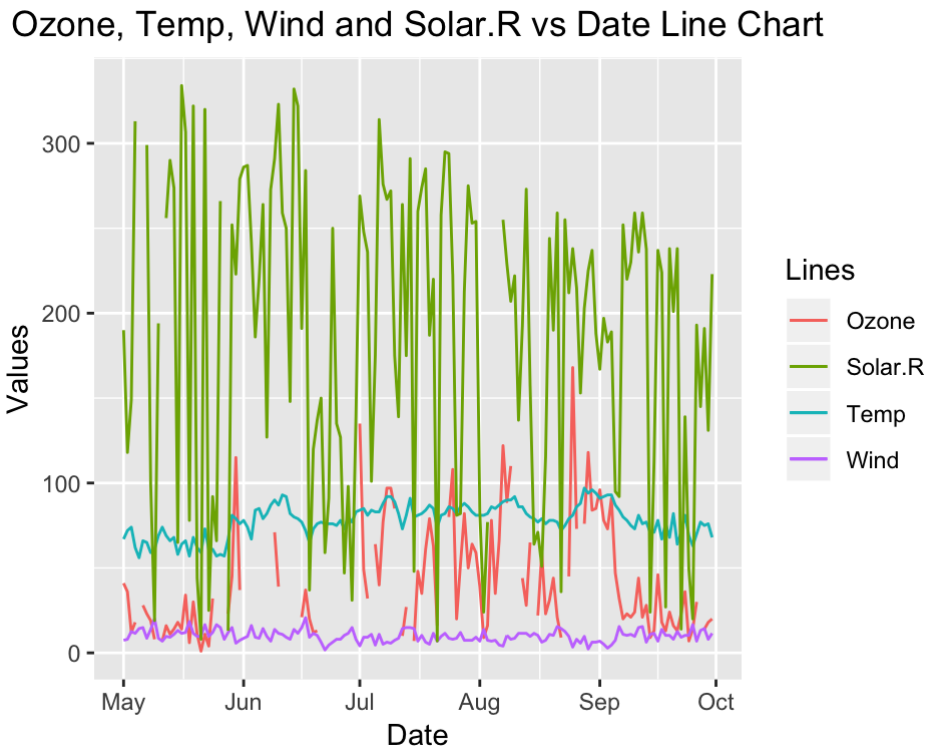
Create line chart for solar.R

```
ggplot(Mydf, aes(x=Date, y=Solar.R, group=1)) +
  geom_line(color="black", na.rm = TRUE)
```



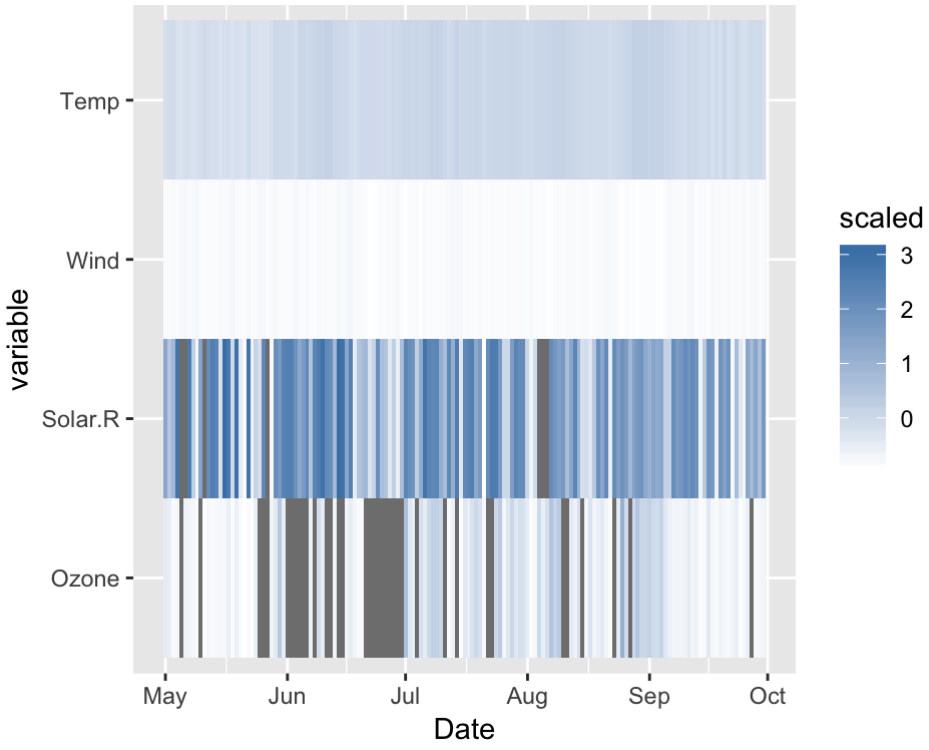
Combine 4 line charts

```
ggplot(Mydf, aes(x=Date, group=1)) +
  geom_line(aes(y=Ozone, color="Ozone"), na.rm = TRUE) +
  geom_line(aes(y=Temp, color="Temp"), na.rm = TRUE) +
  geom_line(aes(y=Wind, color="Wind"), na.rm = TRUE) +
  geom_line(aes(y=Solar.R,color="Solar.R"), na.rm = TRUE) +
  scale_color_discrete(name="Lines") +
  ggtitle("Ozone, Temp, Wind and Solar.R vs Date Line Chart") +
  theme(plot.title=element_text(hjust=0.5)) +
  ylab("Values")
```



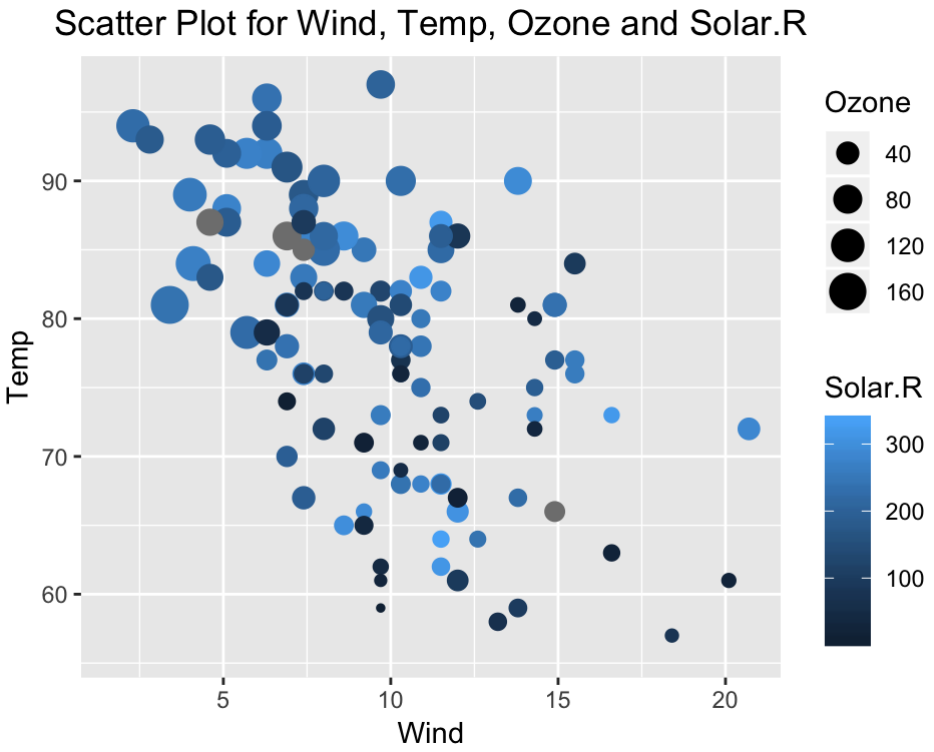
Step 4: Look at all the data via a Heatmap

```
library(reshape2)
mydf.m<- melt( Mydf, id.vars=c("Date"), names(Mydf)[1:4] )
mydf.m$scaled<- scale(mydf.m$value)
ggplot( data=mydf.m, aes(Date, variable)) +
  geom_tile(aes(fill = scaled), na.rm = TRUE) +
  scale_fill_gradient(low = "white", high = "steelblue")
```



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

```
ggplot(Mydf,aes(x=Wind, y=Temp)) +
  geom_point(aes(color=Solar.R, size=Ozone), na.rm = TRUE) +
  ggtitle("Scatter Plot for Wind, Temp, Ozone and Solar.R") +
  theme(plot.title=element_text(hjust=0.5))
```



Step 6: Final Analysis

• Do you see any patterns after exploring the data?
Patterns:
1. When the temperature goes up, ozone goes up as well.
2. Temperature goes up when Solar Radiation is high.
3. Ozone and temperature are high in September

• What was the most useful visualization?
Scatter plot is the most useful visualization. Because it can not only show the relationship between two variables but also give the distribution information.