Darrell_Nelson_HW6

Darrell Nelson II

February 26, 2019

Histograms for each variable

```
# Air Quality Analysis
# HW 06
# Darrell Nelson II

# Step 01: Load the data, ensure ggplot is on
air <- airquality
require(ggplot2)
```

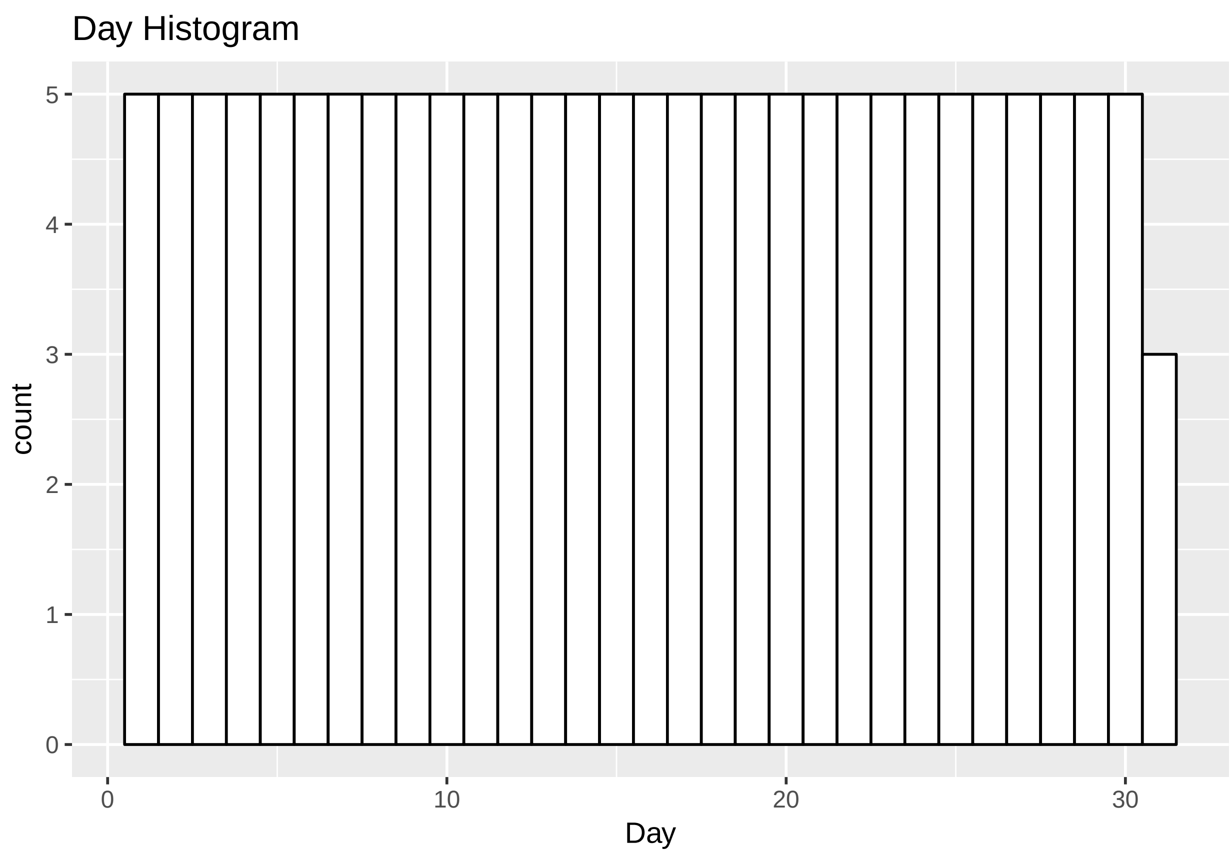
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
# Step 02: Clean the data
# Step 03: Understand the data distribution
# Cleaning and creating graphs in parallel will maximize the amount
# of data

# Check out which column contains the most NAs
colSums(is.na(air))
```

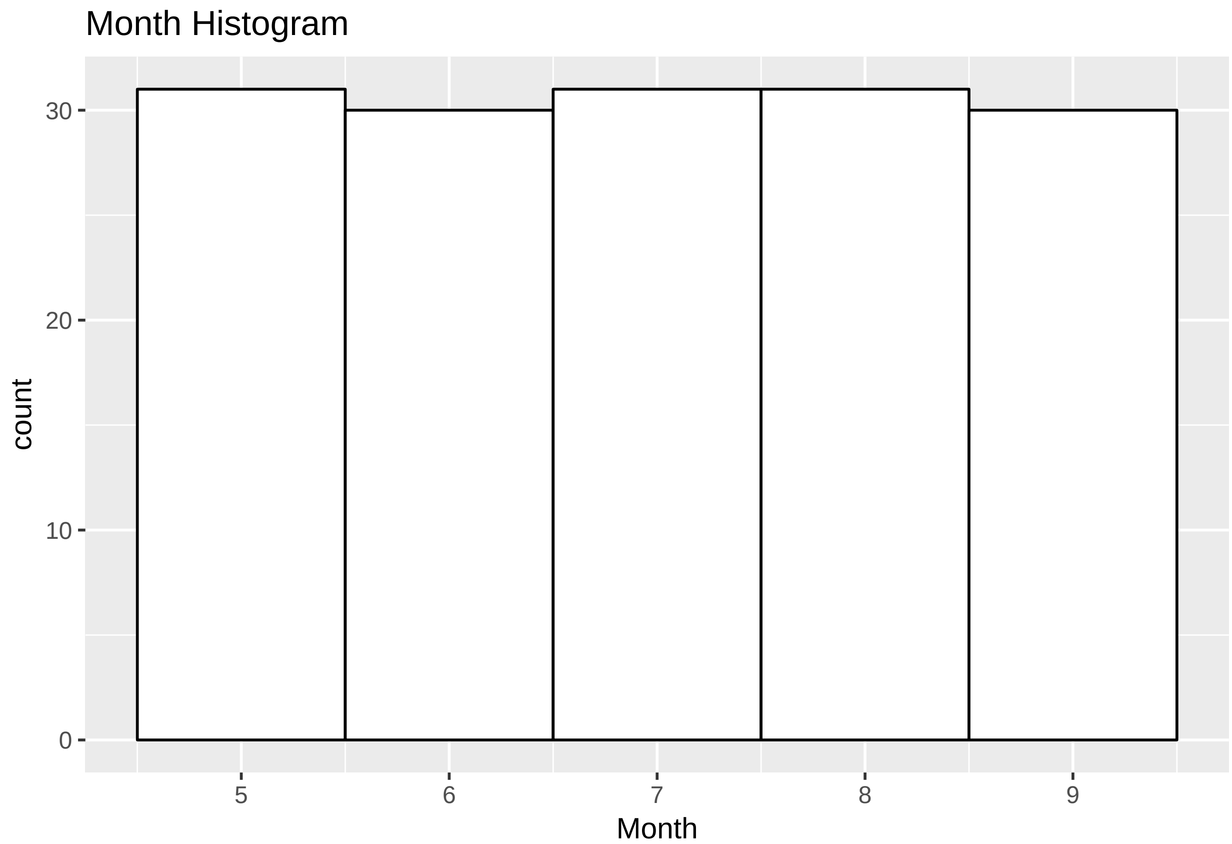
```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

```
# Generate Histograms
ggplot(air, aes(x=Day)) + geom_histogram(bins = 31, color = "black", fill =
            "white") + ggtitle("Day Histogram")
```



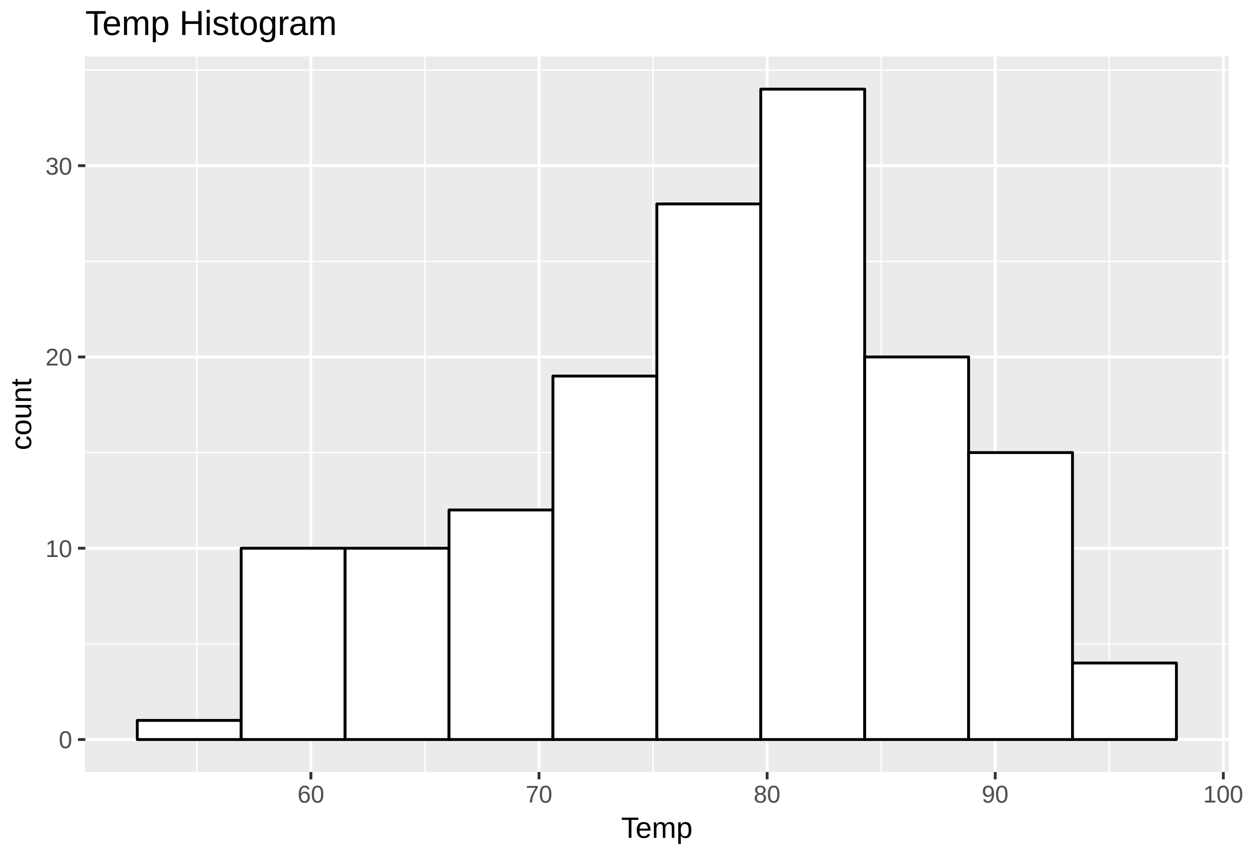
A 31 bin histogram best clarifies that the airquality dataset was collected over 5 months and 3 of those months had 31 days in them.

```
ggplot(air, aes(x=Month)) + geom_histogram(bins = 5, color = "black", fill = "white") + ggtitle("Month H
```



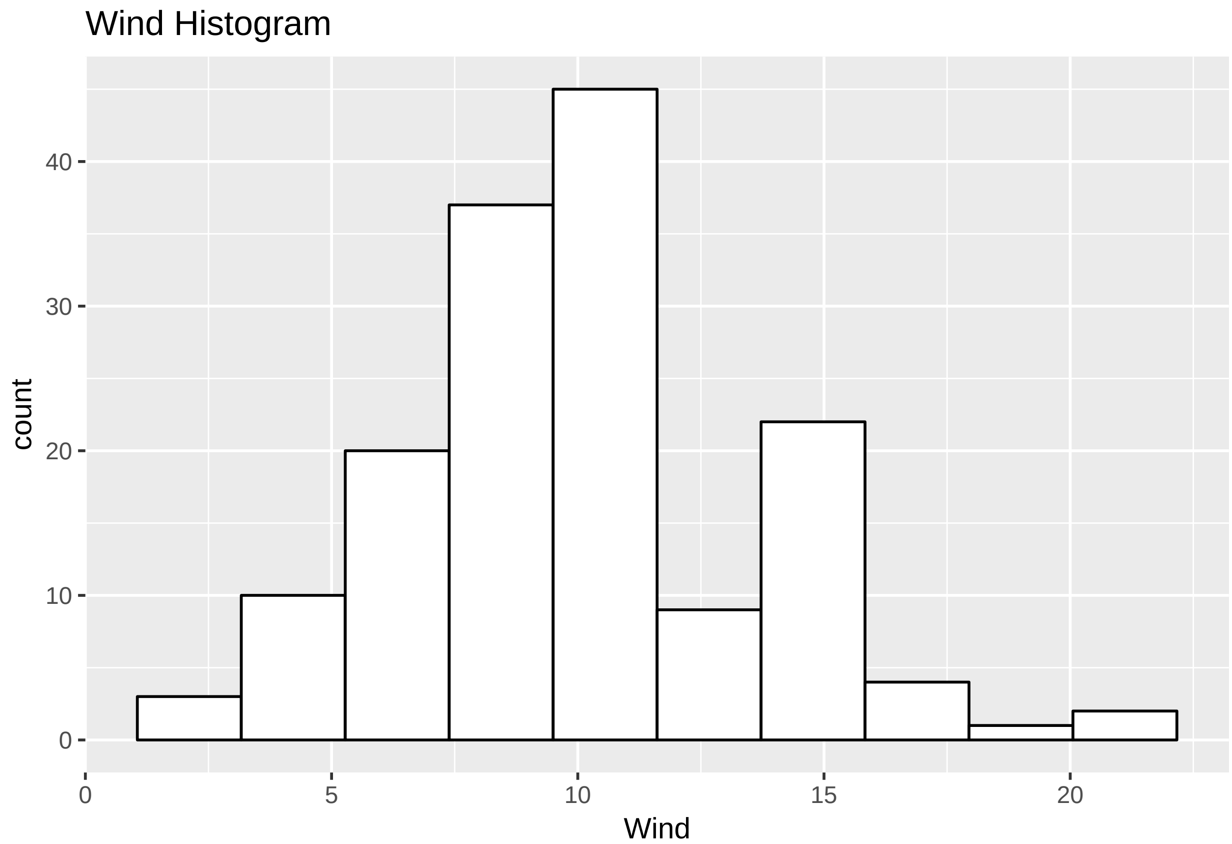
A 5 bin histogram best clarifies that the airquality dataset was collected over 5 months and 3 of those months had 31 days in them.

```
ggplot(air, aes(x=Temp)) + geom_histogram(bins = 10, color = "black", fill = "white") + ggtitle("Temp Hi
```



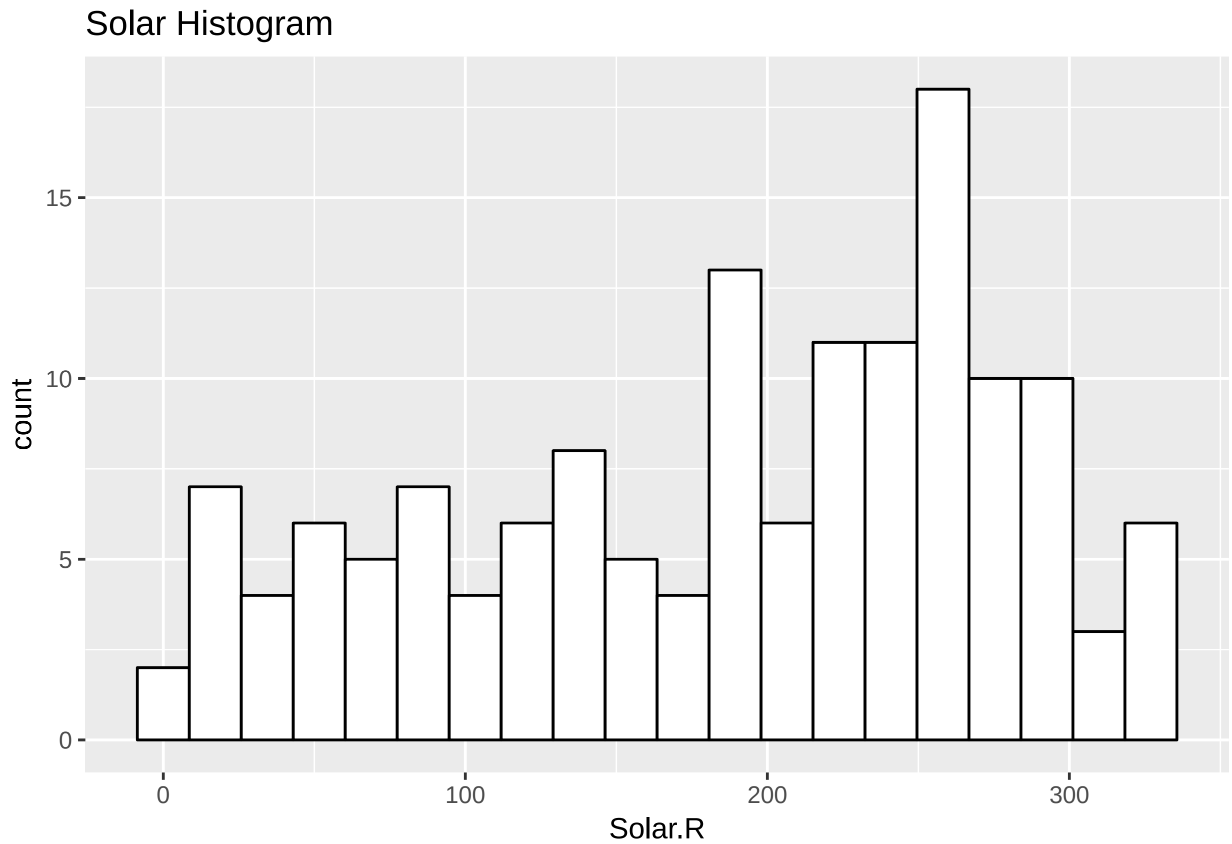
A 10 bin histogram shows that the airquality dataset collected temperatures ranging from the mid 50s to mid 90s with a low of 65 F on one day and a high of 96 F on about 4 days. The median temperature is in the low 80s (81-83 F).

```
ggplot(air, aes(x=Wind)) + geom_histogram(bins = 10, color = "black", fill = "white") + ggtitle("Wind H
```



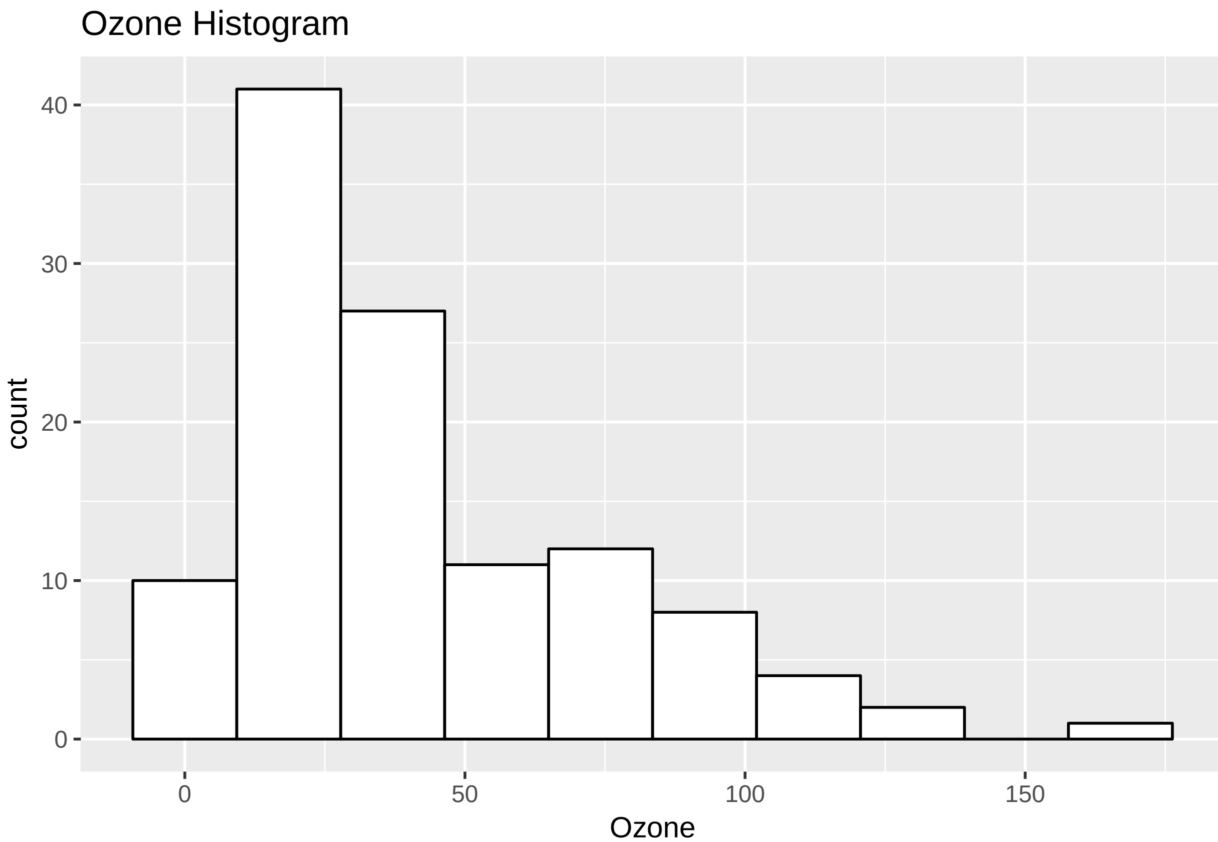
10 bin histogram shows that the median windspeed in this dataset is 11-13 mph."

```
# Omit all rows with NA in Solar R, while omitting the Ozone column from the new dataframe.
# That way we can create a histogram with the most data from the Solar R column
air1 <- na.omit(air[-1])
ggplot(air1, aes(x=Solar.R)) + geom_histogram(bins = 20, color = "black", fill = "white") + ggtitle("Sol
```



A 20 bin histogram shows that the median solar radiation is around 250-270 L."

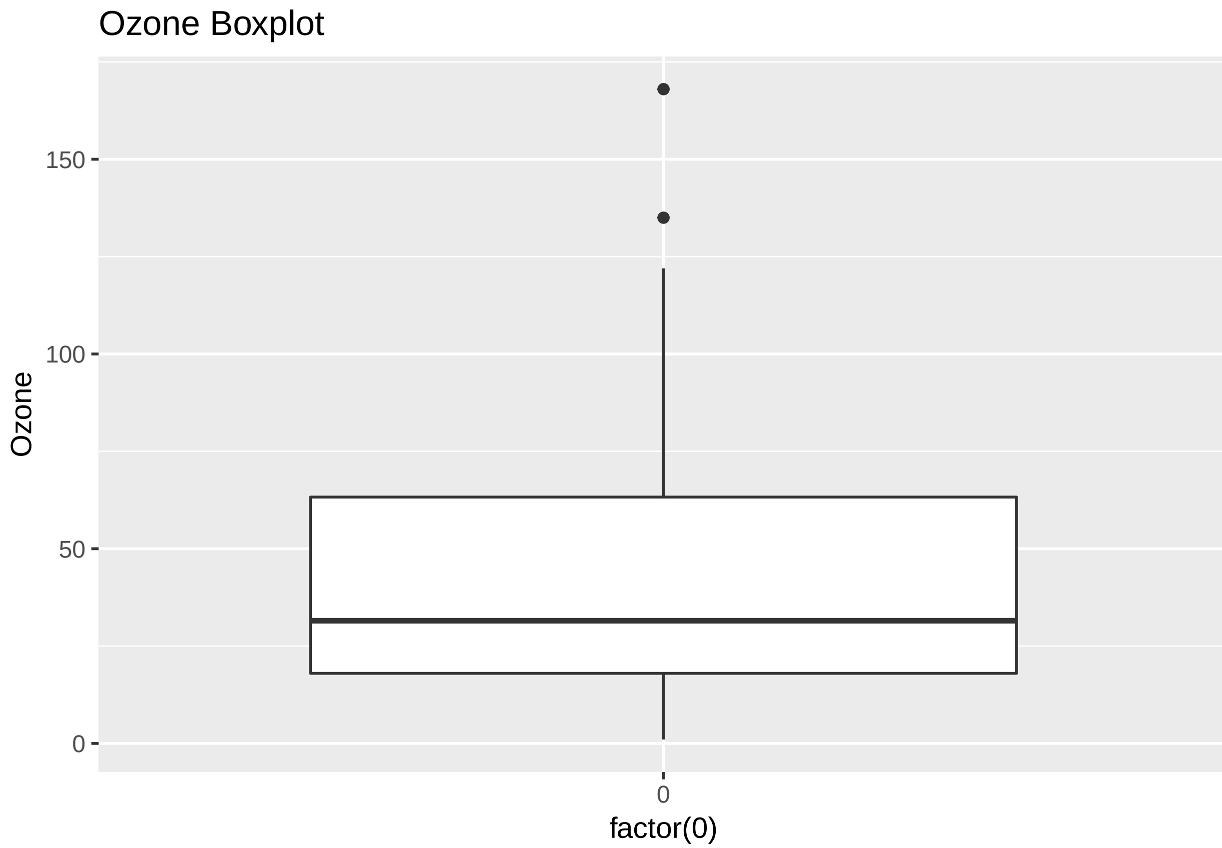
```
# Omit all rows with NA in Ozone, while omitting the Solar.R column from the new dataframe.
# That way we can create a histogram with the most data from the Ozone column
air2 <- na.omit(air[-2])
ggplot(air2, aes(x=Ozone)) + geom_histogram(bins = 10, color = "black", fill = "white") + ggtitle("Ozone
```



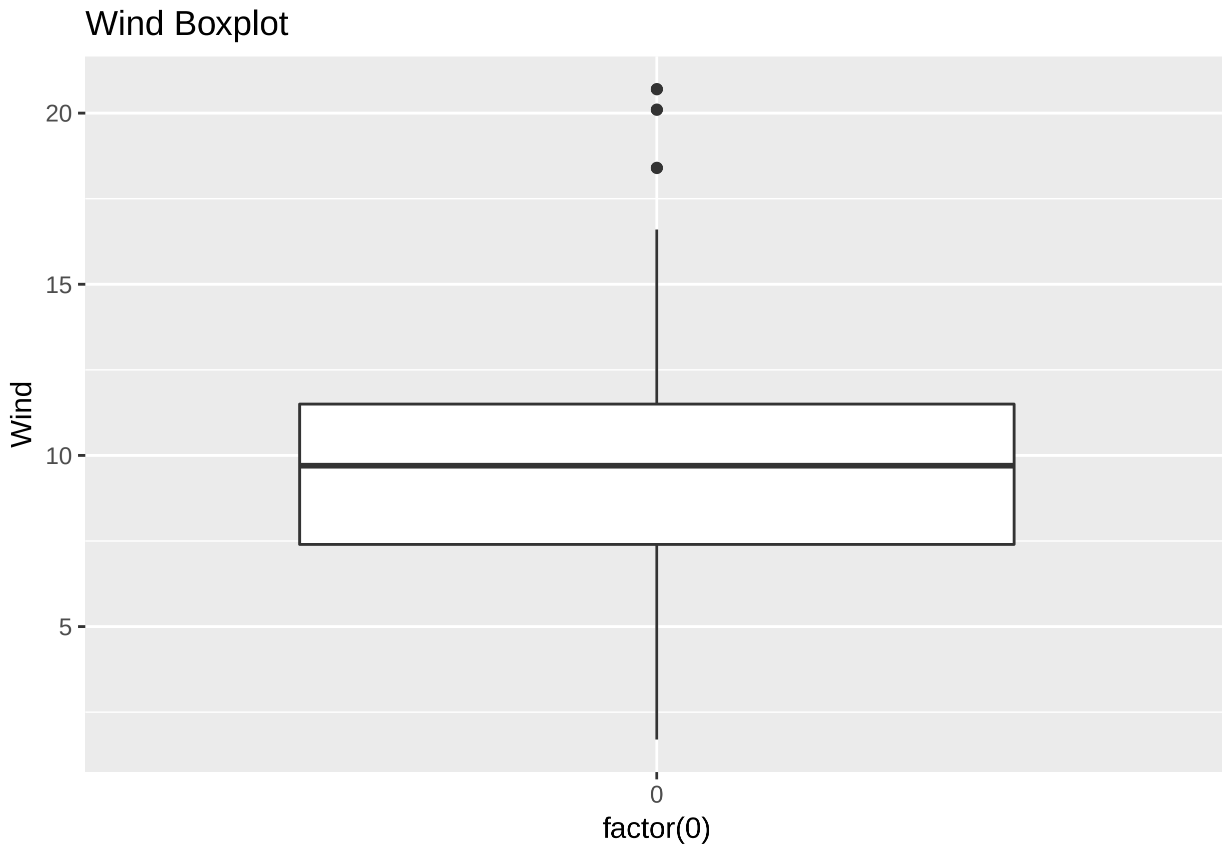
A 10 bin histogram shows that the median ozone is around 20-30 ppb."

Boxplots

```
# Boxplot for Ozone
ggplot(air2, aes(x=factor(0), Ozone)) + geom_boxplot() + ggtitle("Ozone Boxplot")
```

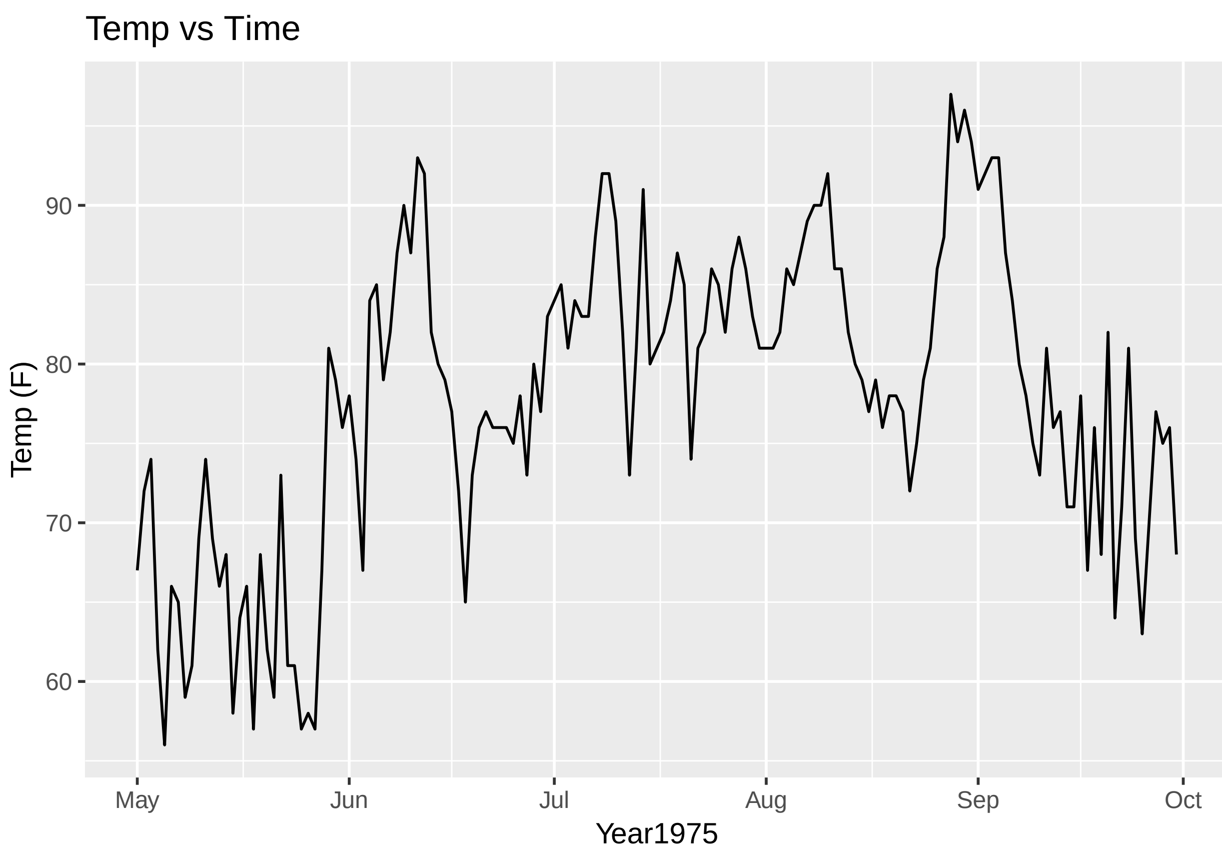



```r
# Boxplot for Wind
ggplot(air, aes(x=factor(0), Wind)) + geom_boxplot() + ggtitle("Wind Boxplot")
```

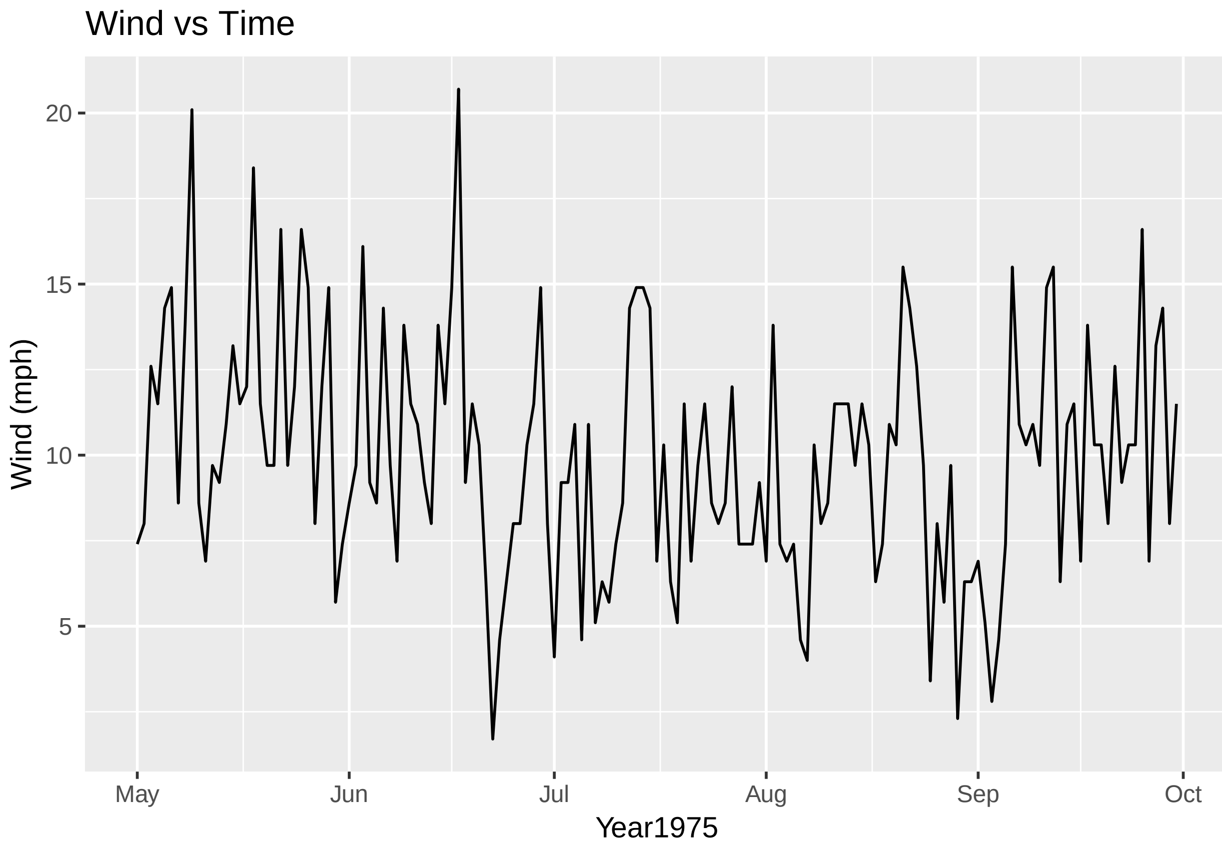


Step 03: Explore how the data changes over time

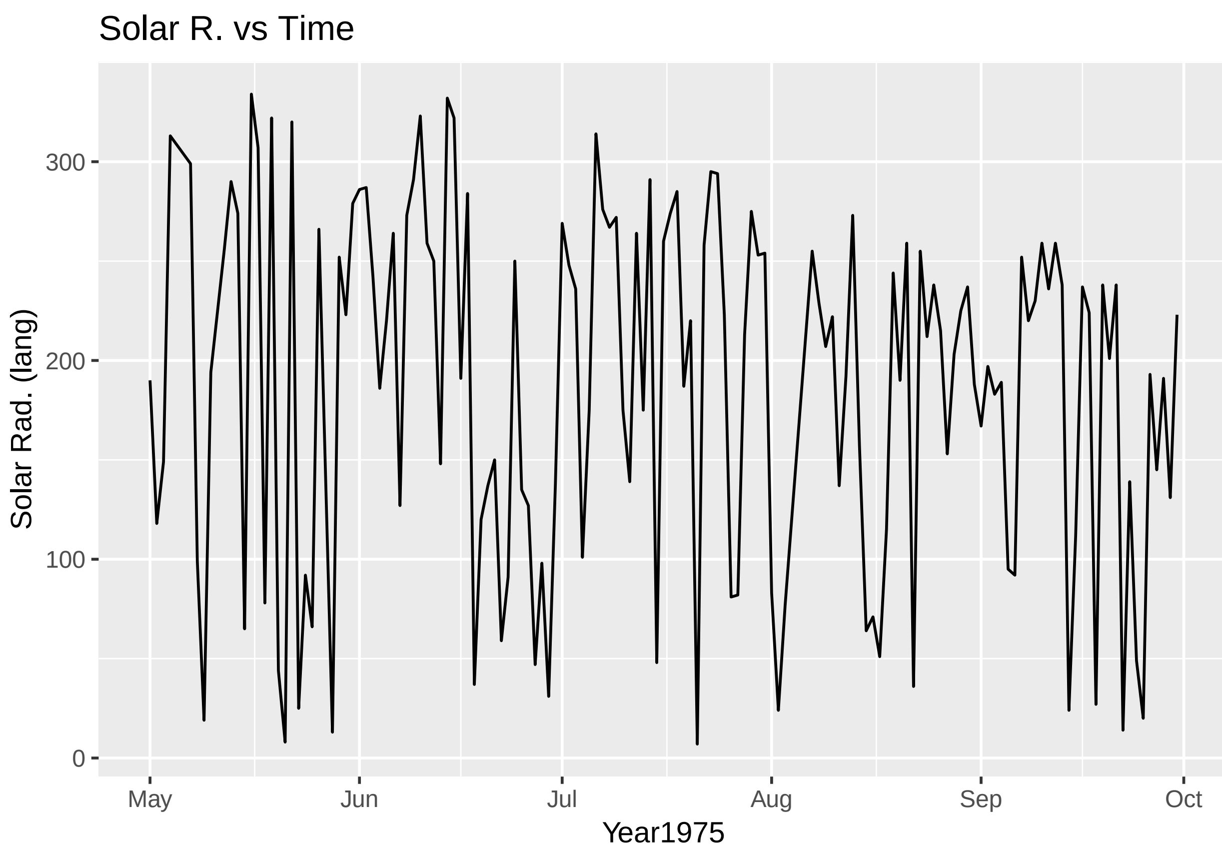
```
# Step 03: Explore how the data changes over time
Year1975 <- seq(from = as.Date("1973-05-01"), to = as.Date("1973-09-30"), by = 'day')
air3 <- cbind(air, Year1975)

# Create a line graph for ozone, temp, wind and solar.R over time
ggplot(air3, aes(x=Year1975)) + geom_line(aes(y=Temp)) + ylab("Temp (F)") + ggtitle("Temp vs Time")
```



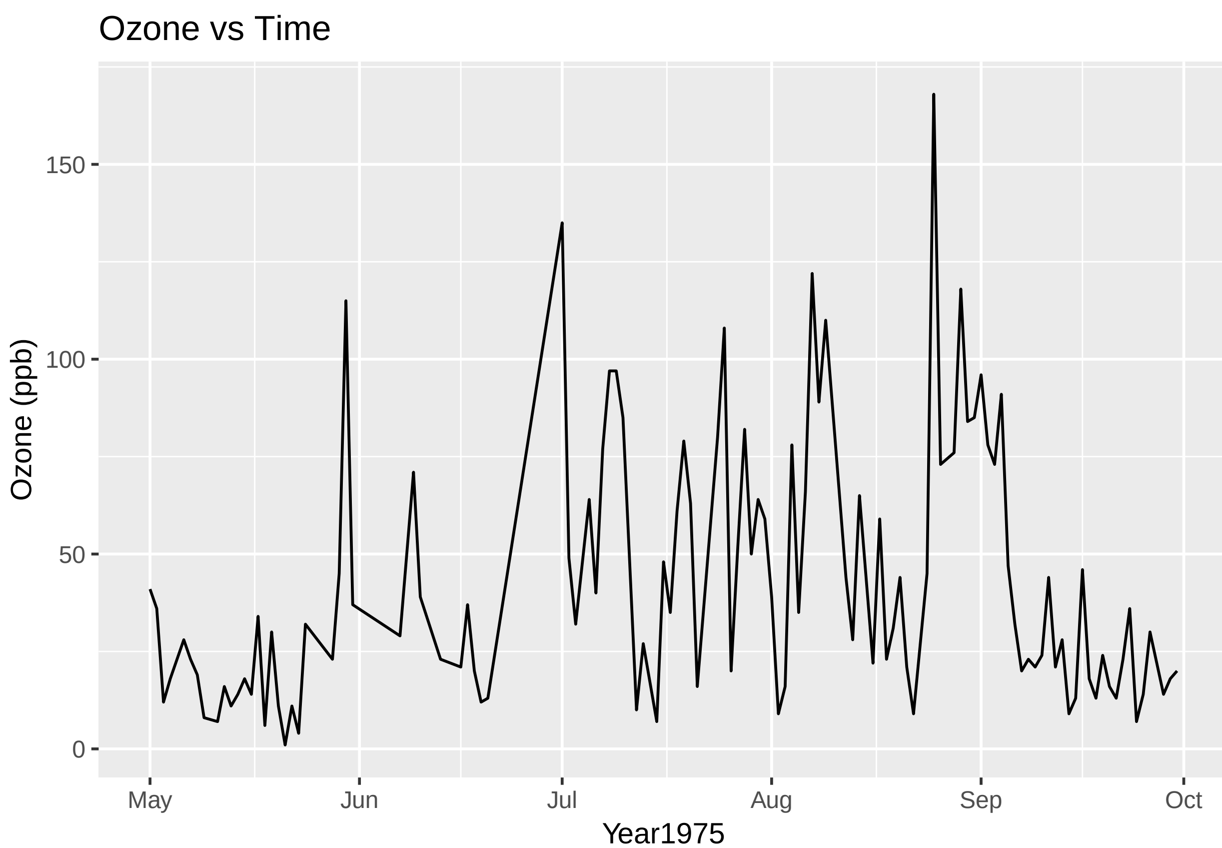

```
ggplot(air3, aes(x=Year1975)) + geom_line(aes(y=Wind)) + ylab("Wind (mph)") + ggtitle("Wind vs Time")
```




```
air4 <- na.omit(air3[-1])
ggplot(air4, aes(x=Year1975)) + geom_line(aes(y=Solar.R)) + ylab("Solar Rad. (lang)") + ggtitle("Solar R
```




```
air5 <- na.omit(air3[-2])
ggplot(air5, aes(x=Year1975)) + geom_line(aes(y=Ozone)) + ylab("Ozone (ppb)") + ggtitle("Ozone vs Time")
```

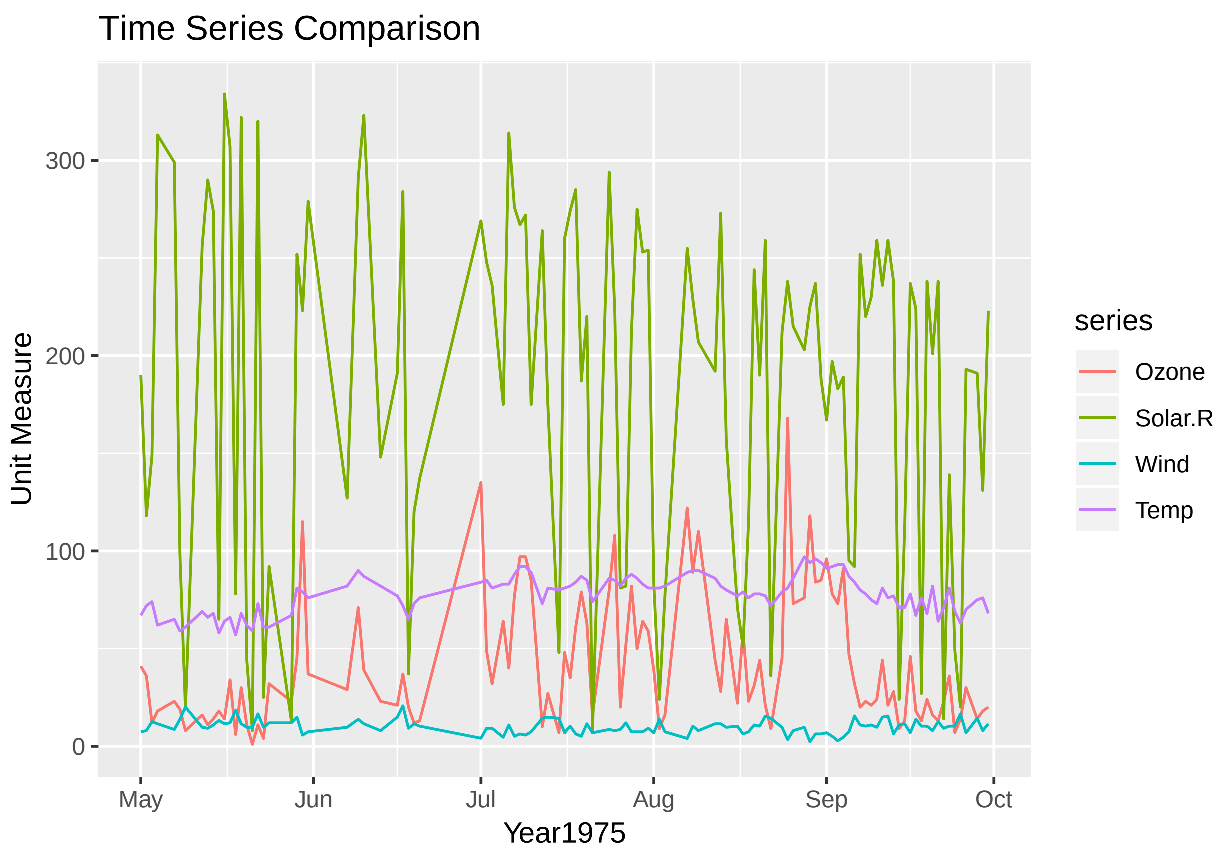



```
# Omit all rows with NA in any cell to ensure fair comparisons across ozone, temp, wind, and solar R.
# Also remove any columns that won't be graphed
air5 <- na.omit(air3[-5:-6])
require(ggplot2)
require(reshape2)
```

```
## Loading required package: reshape2
```

```
## Warning: package 'reshape2' was built under R version 3.5.2
```

```
air6 <- melt(air5,  id.vars = 'Year1975', variable.name = 'series')

# plot on same grid, each series colored differently --
# good if the series have same scale
ggplot(air6, aes(Year1975,value)) + geom_line(aes(colour = series)) + ylab("Unit Measure") + ggtitle("Ti
```



Step04: Look at all the data via a Heatmap

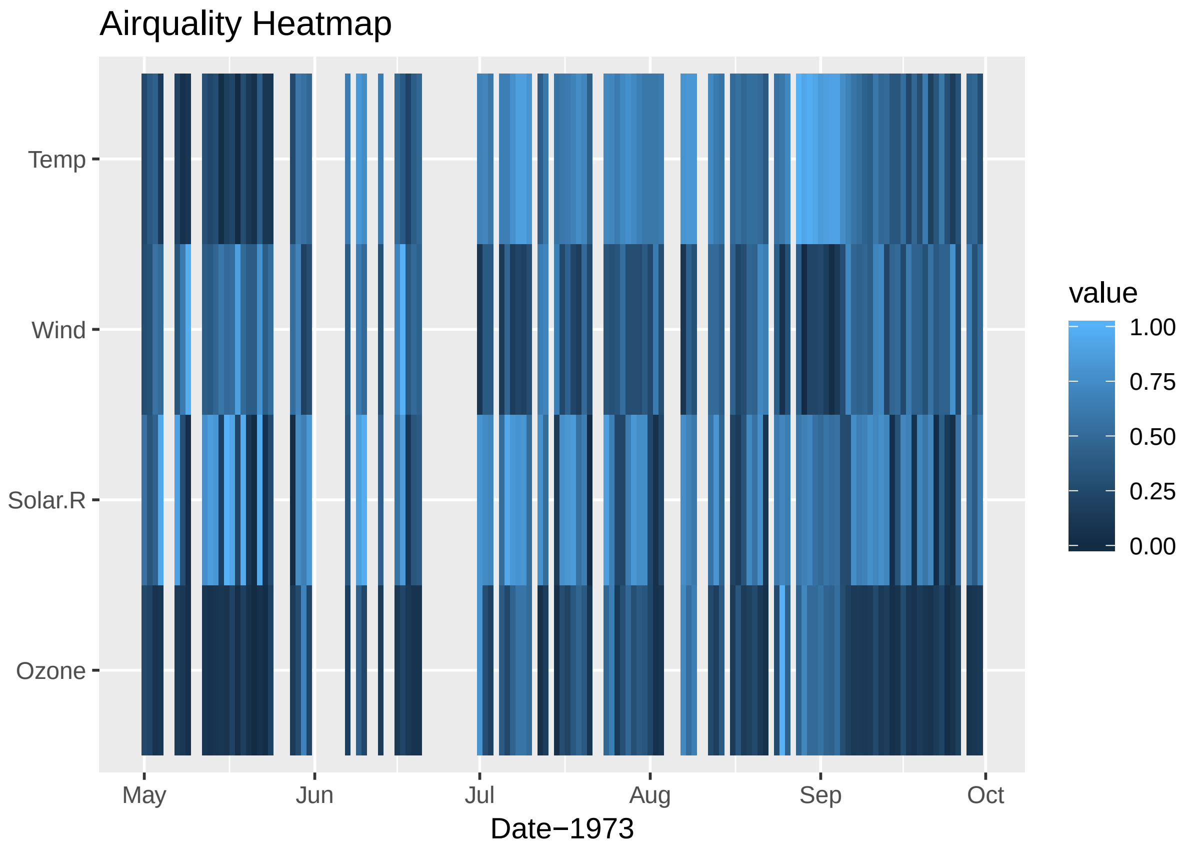
```
# Step 04: Look at all the data via a Heatmap
air7 <- na.omit(air3[-5])
air8 <- air7[-5:-6]

# Scale the factors from 0 to 1
library(scales)
```

```
## Warning: package 'scales' was built under R version 3.5.2
```

```
air8$Ozone <- rescale(air8$Ozone)
air8$Solar.R <- rescale(air8$Solar.R)
air8$Wind <- rescale(air8$Wind)
air8$Temp <- rescale(air8$Temp)
# Melt data
meltair8 <- melt(air8)
```

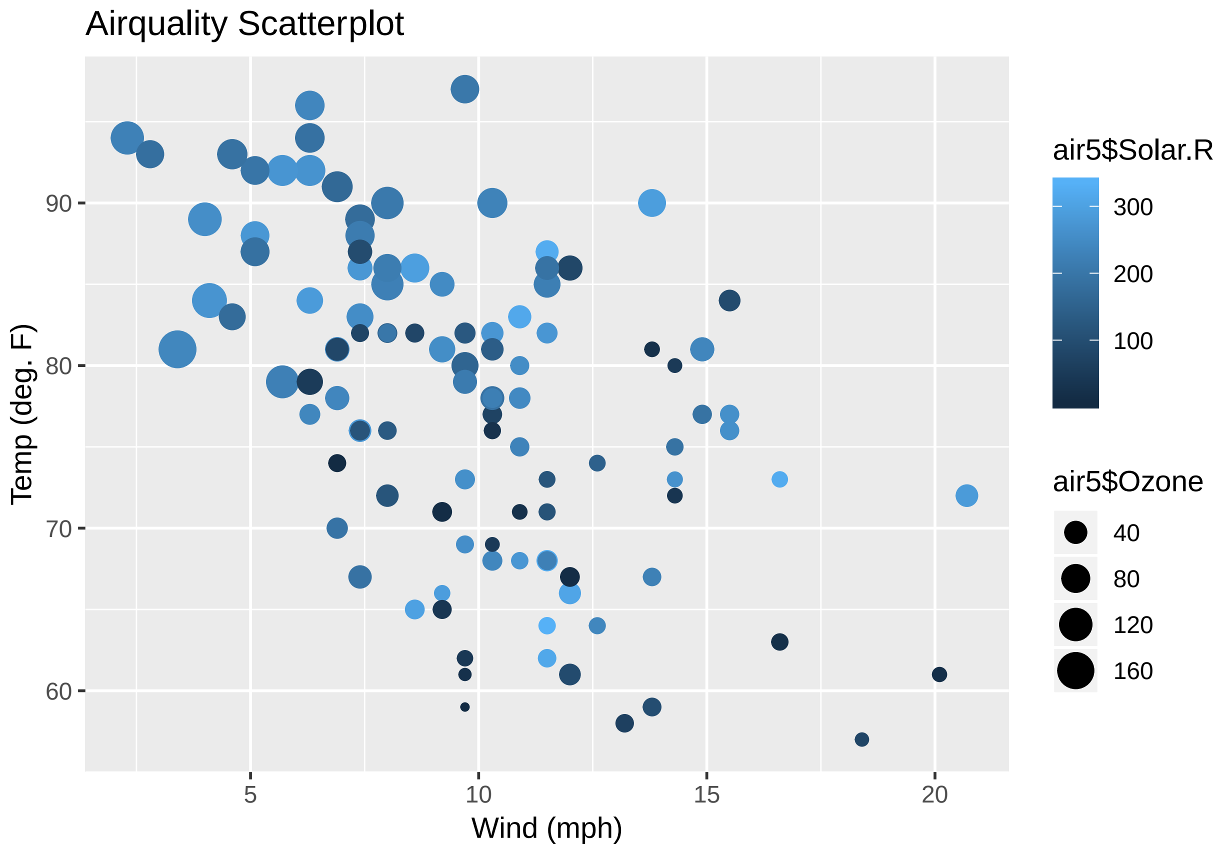
```
## No id variables; using all as measure variables
```

```
# Add date column
hey <- rep(air7$Year1975, 4)
meltair8[3] <- hey
# Plot the heatmap!
ggplot(meltair8, aes(x=V3, y=variable, fill=value)) + geom_tile() + ylab("") + xlab("Date-1973") + ggti
```



Step04: Look at all the data via a scatter chart

```
# Look at all the data via a scatter chart
ggplot(air5) + geom_point(data = air5, aes(x=air5$Wind, y=air5$Temp, size=air5$Ozone, color=air5$Solar.R
```



Step 06: Do you see any patterns after exploring the data

There appears to be a significant correlation between temperature, solar radiation, wind, and concentration of ozone in the New York atmosphere. Initially I believed ozone is indicative of good air quality since stratospheric ozone protects the planet from harmful UV radiation. However, ground-level ozone (which is what this dataset is measuring) is categorized as a pollutant that is harmful to breathe, damages crops and other vegetation, and is the main ingredient in smog. Ground-level ozone is created pollutants from industrial machinery and car exhausts when they react with sunlight. Therefore, it makes sense that on days with very little wind there is a higher concentration of ozone in New York. Intensified sunlight and temperatures exacerbate the problem. The radiation from sunlight combined with hot temperatures creates a more reactive atmospheric environment that worsens the air quality. It is clear to see from the "Airquality Scatterplot" graph that high temperatures and solar radiation coupled with low wind speeds results in a very ozone thick (smog) environment.

Digging deeper you can also tell that there's a stronger correlation between low wind speed and high temperatures resulting in thicker smog regardless of how much solar radiation is present that day. This drives home the fact that smog is an entity that is constantly being generated by fossil fuel combustion processes. Without the wind to blow these particulates away the city seeps in its own chemical cocktail of hazardous gas.

Judging by the "Airquality Heatmap" it is clear to see that the hottest months from May 1st - September 30th, 1975 are July and August, which aligns with the general thinking of summer months being the hottest. You can also see that for the first two weeks of September there was very little wind, and that New York received the most solar radiation in July. Throughout this timeframe ozone layers were relatively low except for a few very high points happening on July 1st and in late August.

The temperature line graph shows that May was a relatively cool month; the heat started to pick up as New York approached summer time and started to cool off again in mid-September.

What the most useful visualization?

The most useful visualization was the scatterplot. Being able to plot 4 different variables in a 2D space made it very easy to digest a lot of different information at the same time. I was able to see trends between multiple variables that I didn't see before I created the scatterplot. This is a tool that I will definitely be using going forward.