HW6

Diego Valdes

February 19, 2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
rm(list=ls()) # clear work space
# dev.off(dev.list()["RStudioGD"]) # clear plots

suppressWarnings(require(ggplot2))
```

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

```r
suppressWarnings(require(lubridate))
```

```
## Loading required package: lubridate
```

```r
#suppressWarnings(require(plyr))
#suppressWarnings(require(scales))
#suppressWarnings(require(zoo))
#theme_set(theme_bw())
#theme_set(theme_classic())

# get data and clean it up
dataAQ = airquality
str(dataAQ)
```

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

```r
summary(dataAQ)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```

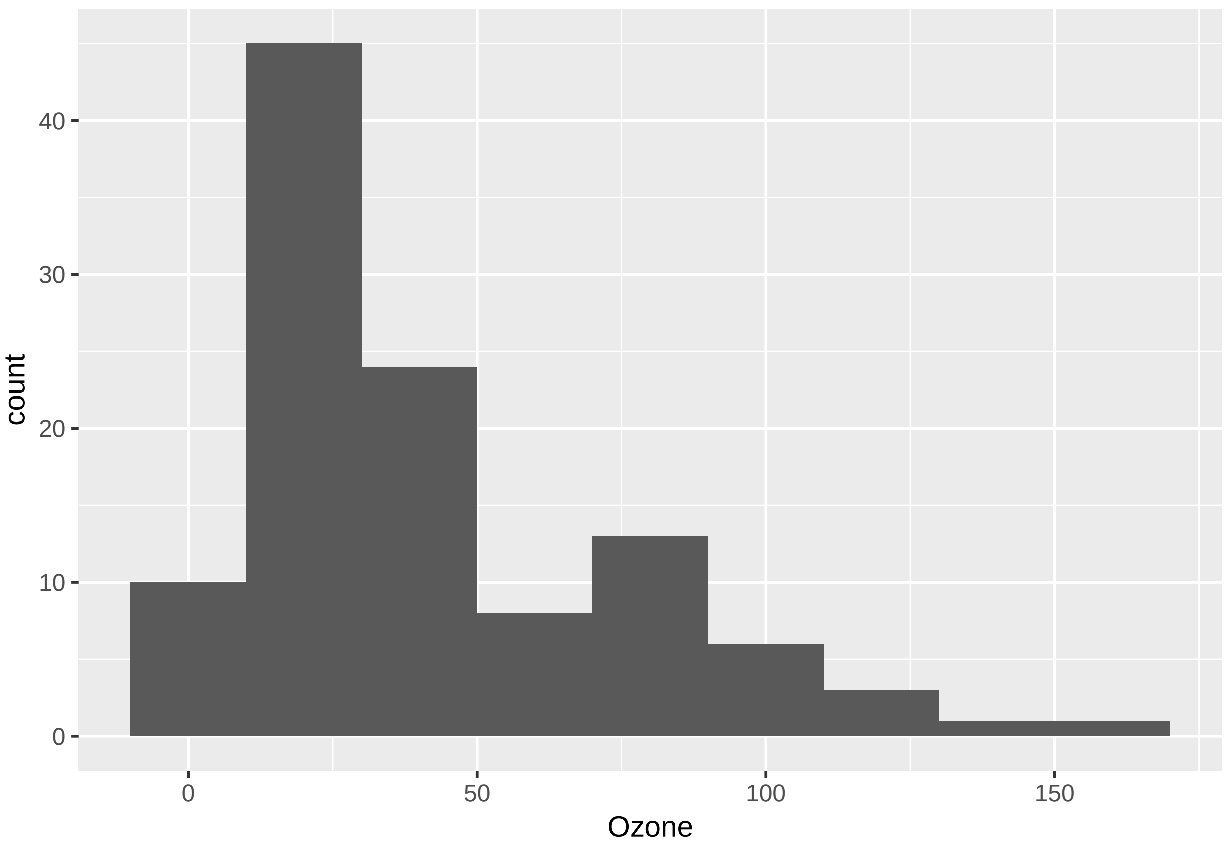
```
dataAQ = na.omit(dataAQ)
str(dataAQ)
```

```
## 'data.frame':    111 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 23 19 8 16 11 14 ...
##  $ Solar.R: int  190 118 149 313 299 99 19 256 290 274 ...
##  $ Wind   : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
##  $ Temp   : int  67 72 74 62 65 59 61 69 66 68 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 7 8 9 12 13 14 ...
##  - attr(*, "na.action")= 'omit' Named int  5 6 10 11 25 26 27 32 33 34 ...
##   ..- attr(*, "names")= chr  "5" "6" "10" "11" ...
```

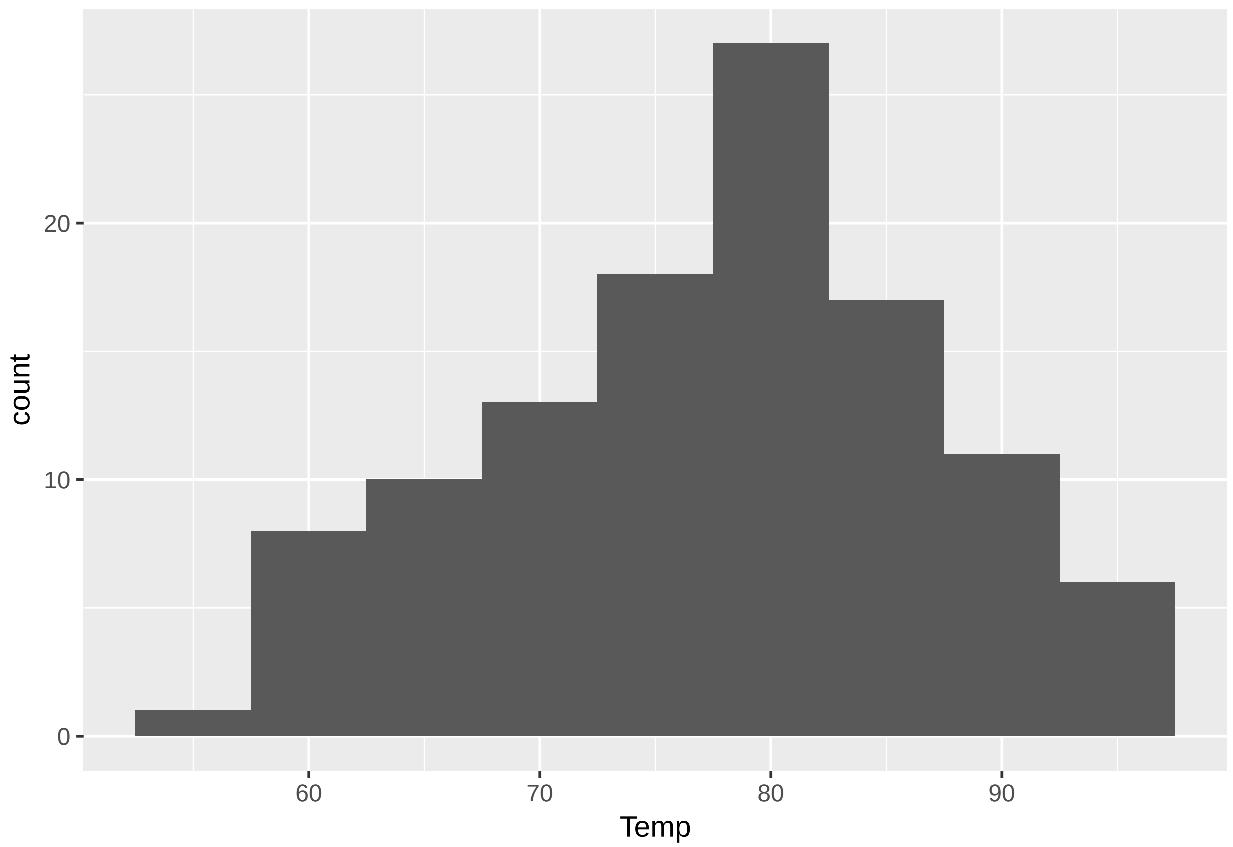
```
summary(dataAQ)
```

```
##      Ozone          Solar.R           Wind            Temp      
##  Min.   :  1.0   Min.   :  7.0   Min.   : 2.30   Min.   :57.00  
##  1st Qu.: 18.0   1st Qu.:113.5   1st Qu.: 7.40   1st Qu.:71.00  
##  Median : 31.0   Median :207.0   Median : 9.70   Median :79.00  
##  Mean   : 42.1   Mean   :184.8   Mean   : 9.94   Mean   :77.79  
##  3rd Qu.: 62.0   3rd Qu.:255.5   3rd Qu.:11.50   3rd Qu.:84.50  
##  Max.   :168.0   Max.   :334.0   Max.   :20.70   Max.   :97.00  
##      Month            Day       
##  Min.   :5.000   Min.   : 1.00  
##  1st Qu.:6.000   1st Qu.: 9.00  
##  Median :7.000   Median :16.00  
##  Mean   :7.216   Mean   :15.95  
##  3rd Qu.:9.000   3rd Qu.:22.50  
##  Max.   :9.000   Max.   :31.00
```

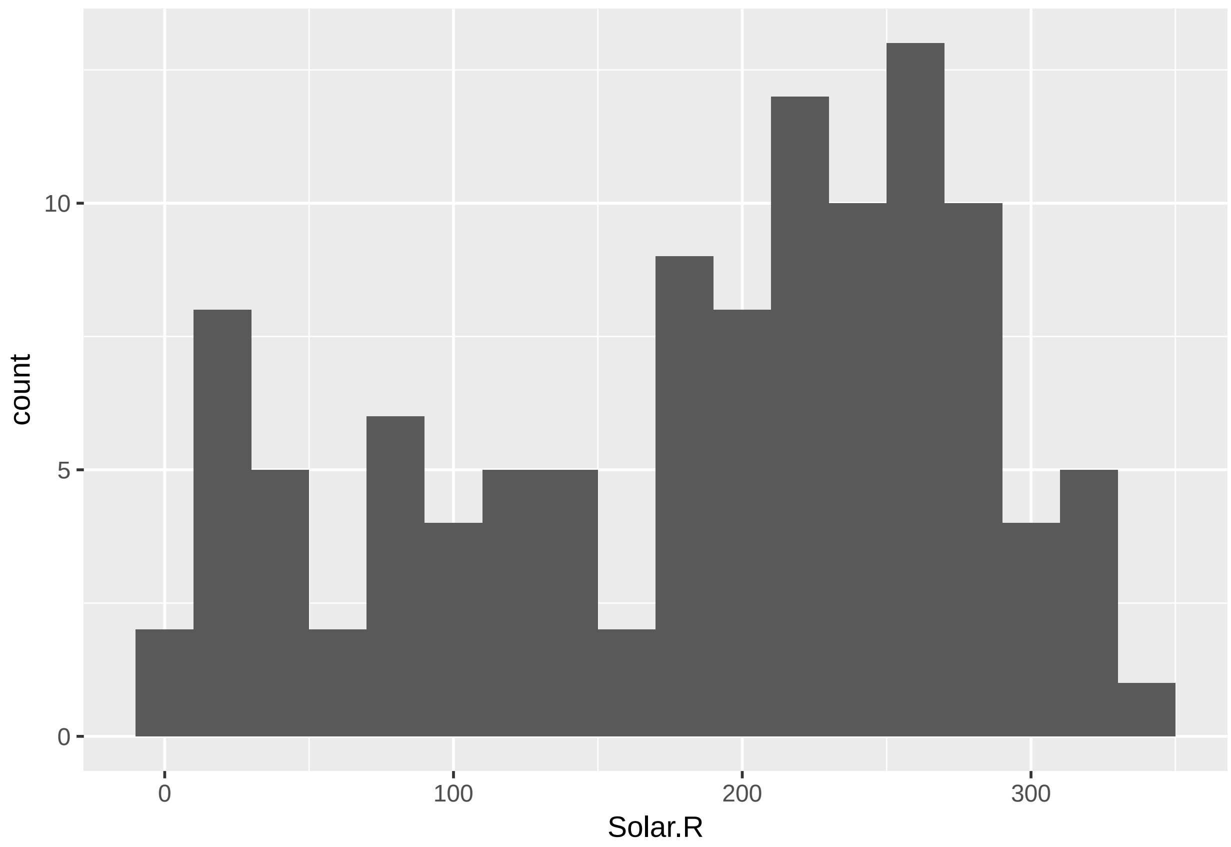
```
# add year
dataAQ$Year = c(rep(1973, nrow(dataAQ)))
#dataAQ$Time = with(dataAQ, ISOdate(dataAQ$Year, dataAQ$Month, dataAQ$Day))
dataAQ$Time = ISOdate(dataAQ$Year, dataAQ$Month, dataAQ$Day)

# histogram for each variable
ggplot(dataAQ, aes(Ozone) ) +  geom_histogram(binwidth = 20)
```



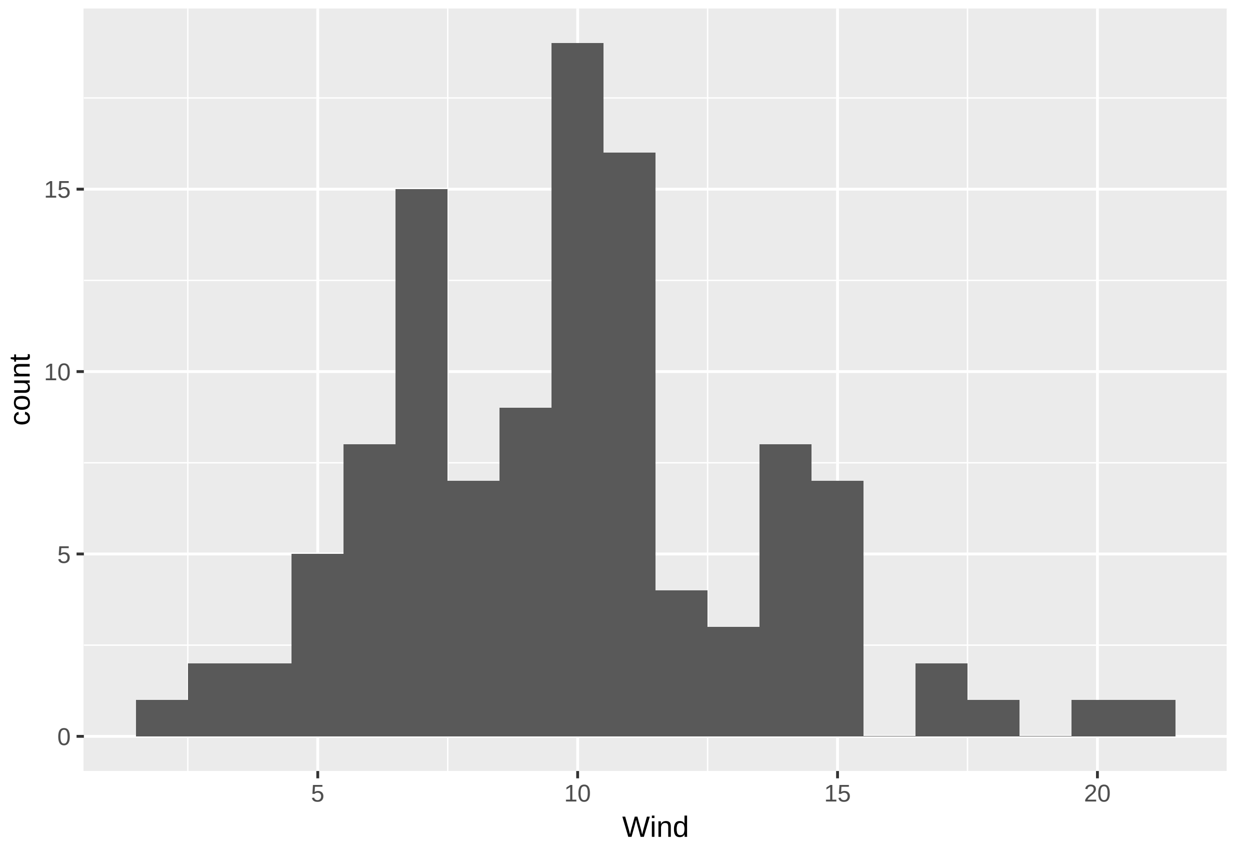
```
ggplot(dataAQ, aes(Temp) ) +  geom_histogram(binwidth = 5)
```



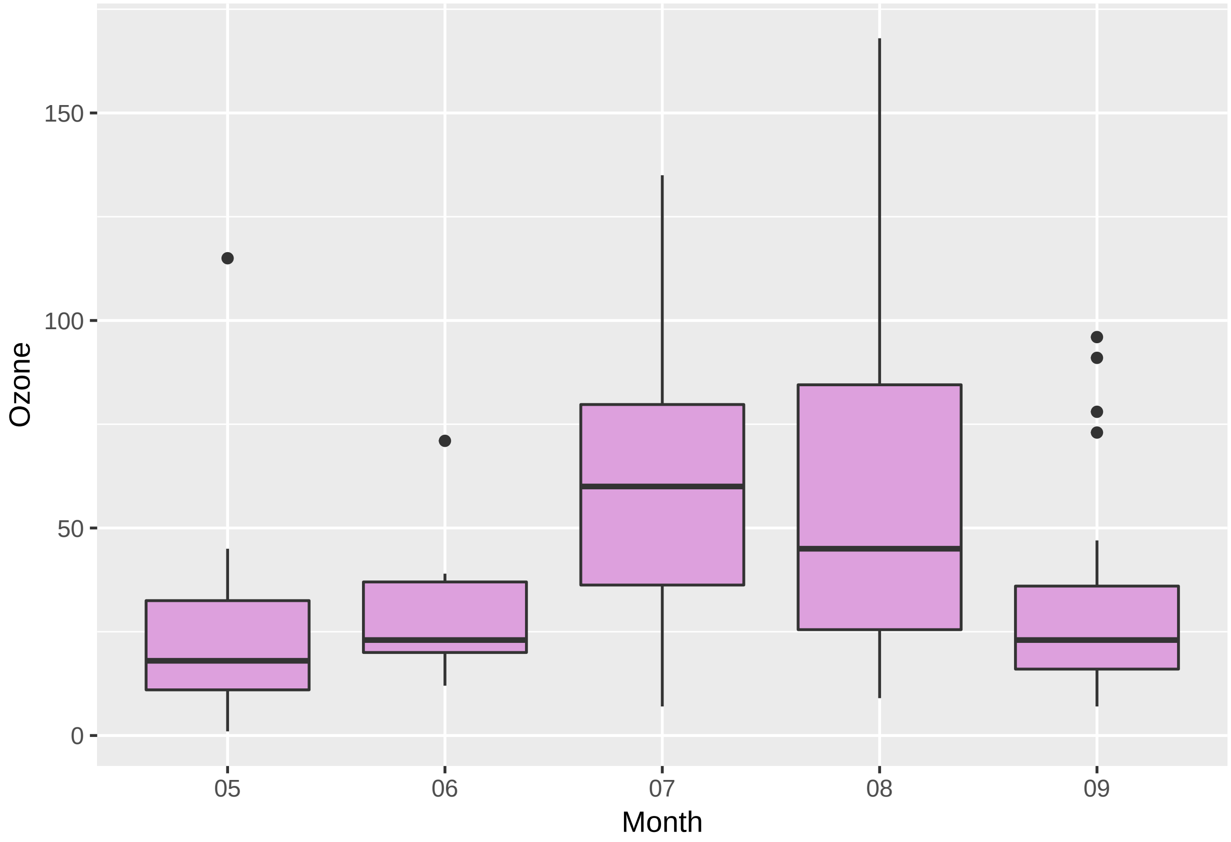
```r
ggplot(dataAQ, aes(Solar.R) ) +  geom_histogram(binwidth = 20)
```



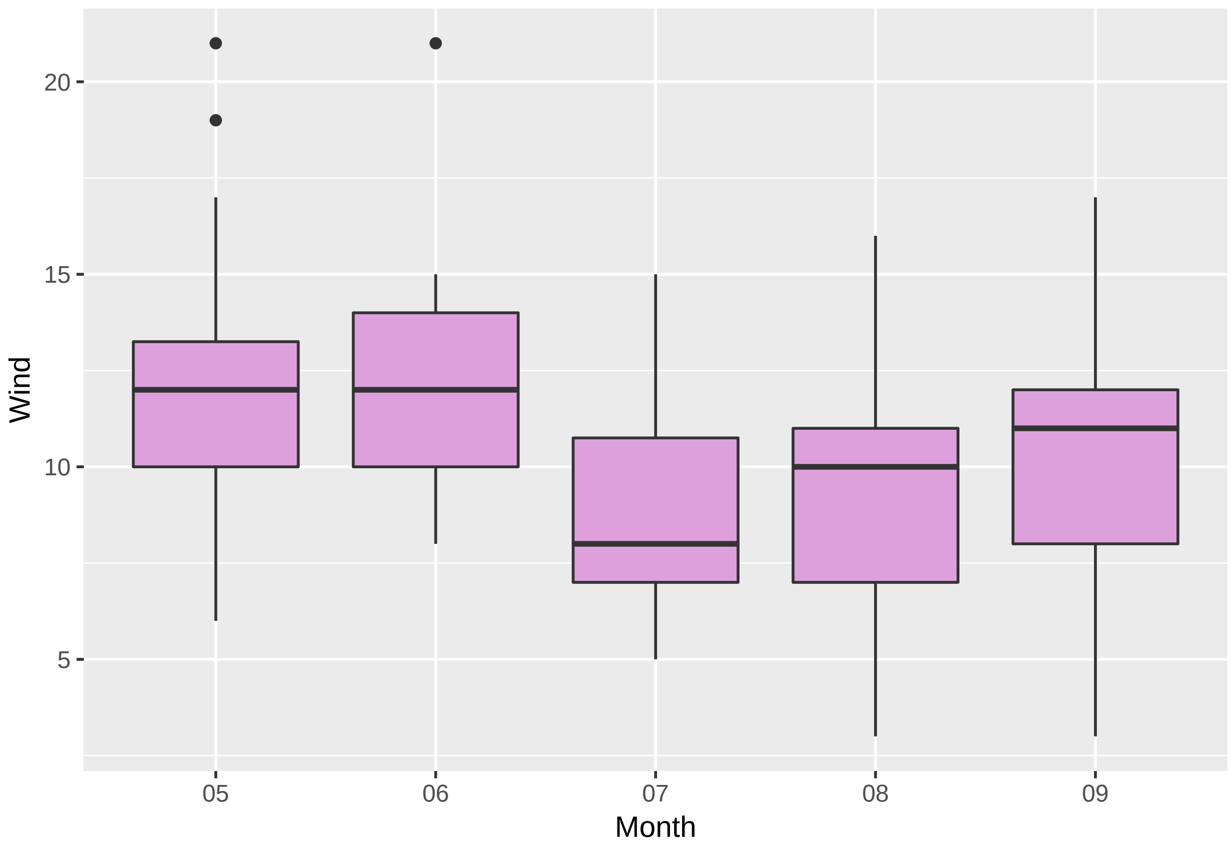
```r
ggplot(dataAQ, aes(Wind) ) +  geom_histogram(binwidth = 1)
```



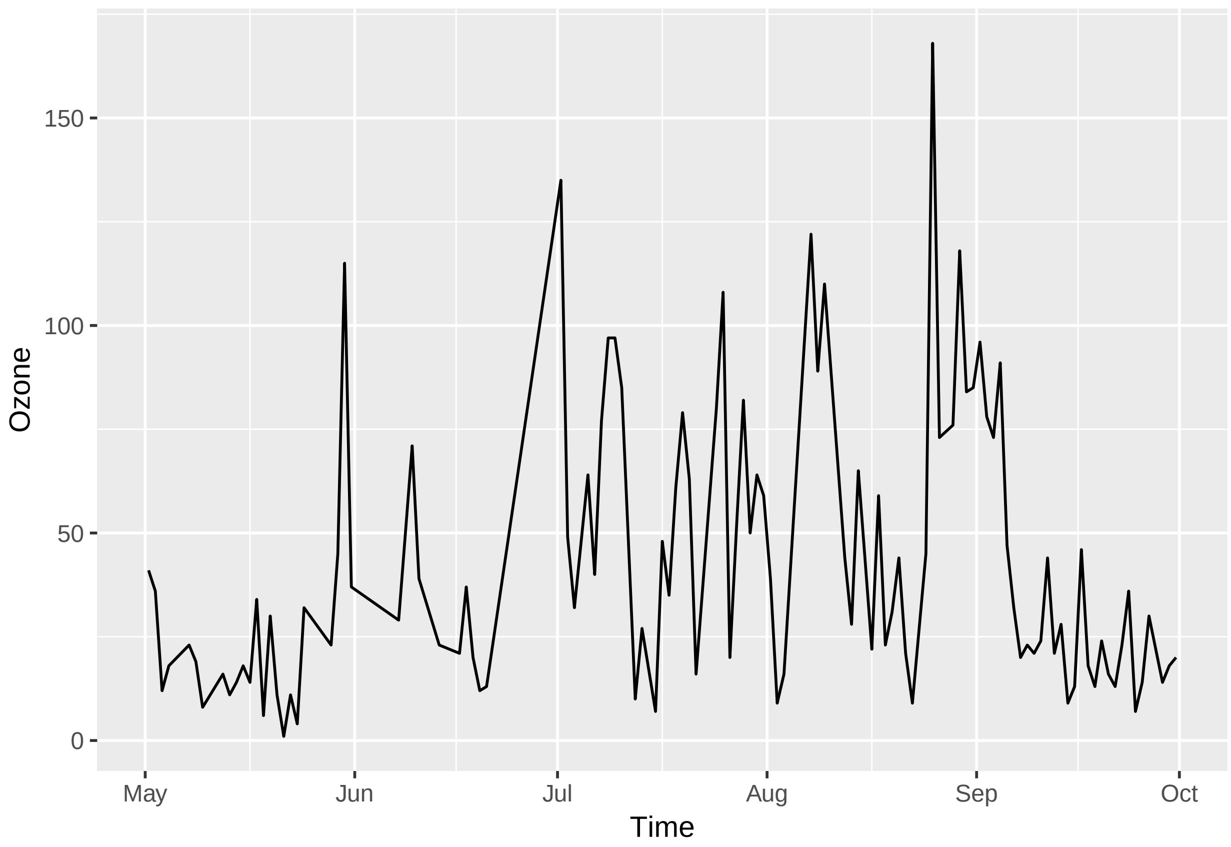
```
# box plots for ozone
ggplot(dataAQ, aes(format(Time, "%m"), Ozone)) +
  geom_boxplot(fill = "plum") +
  xlab("Month")
```



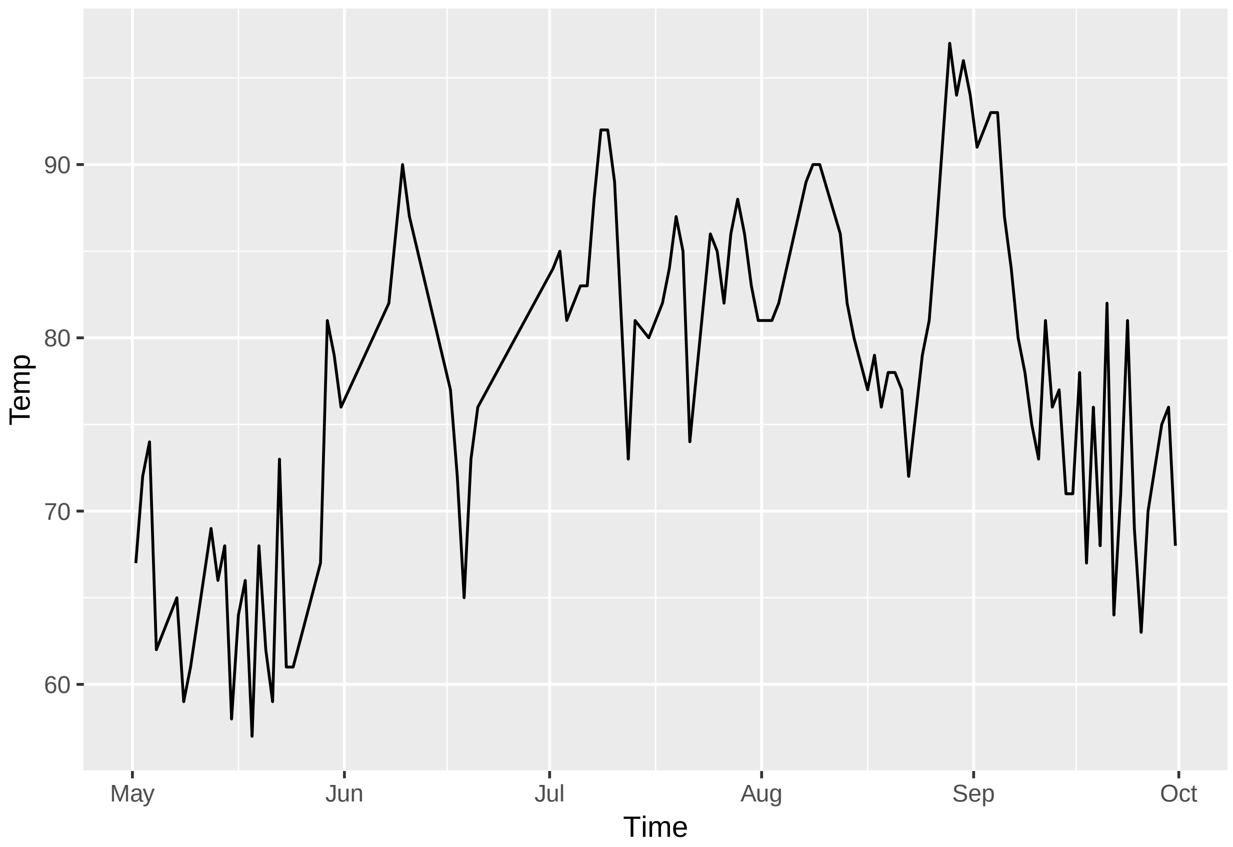
```
# box plot for wind
ggplot(dataAQ, aes(format(Time, "%m"), ceiling(Wind))) +
  geom_boxplot(fill = "plum") +
  xlab("Month") + ylab("Wind")
```



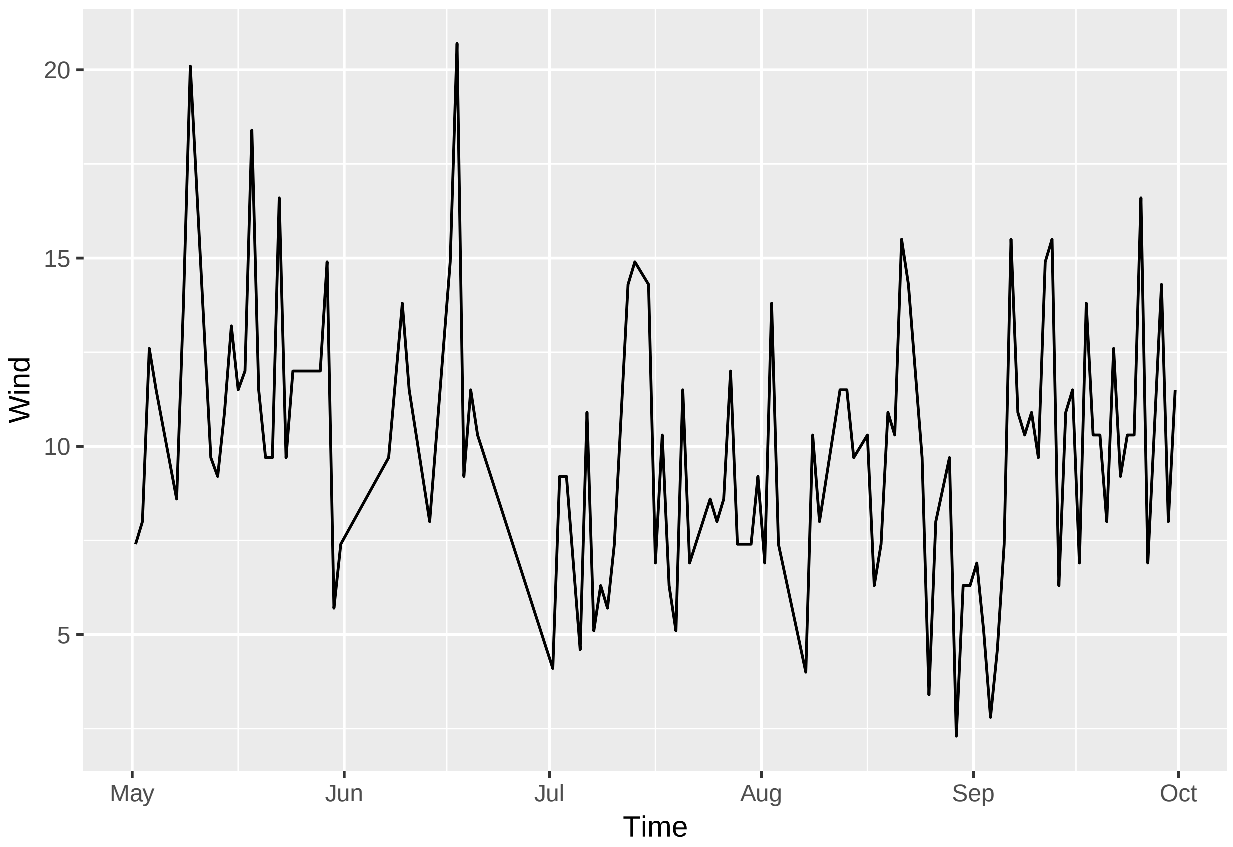
```
# line charts for variables
ggplot(dataAQ, aes(x = Time)) + geom_line(aes(y = Ozone))
```



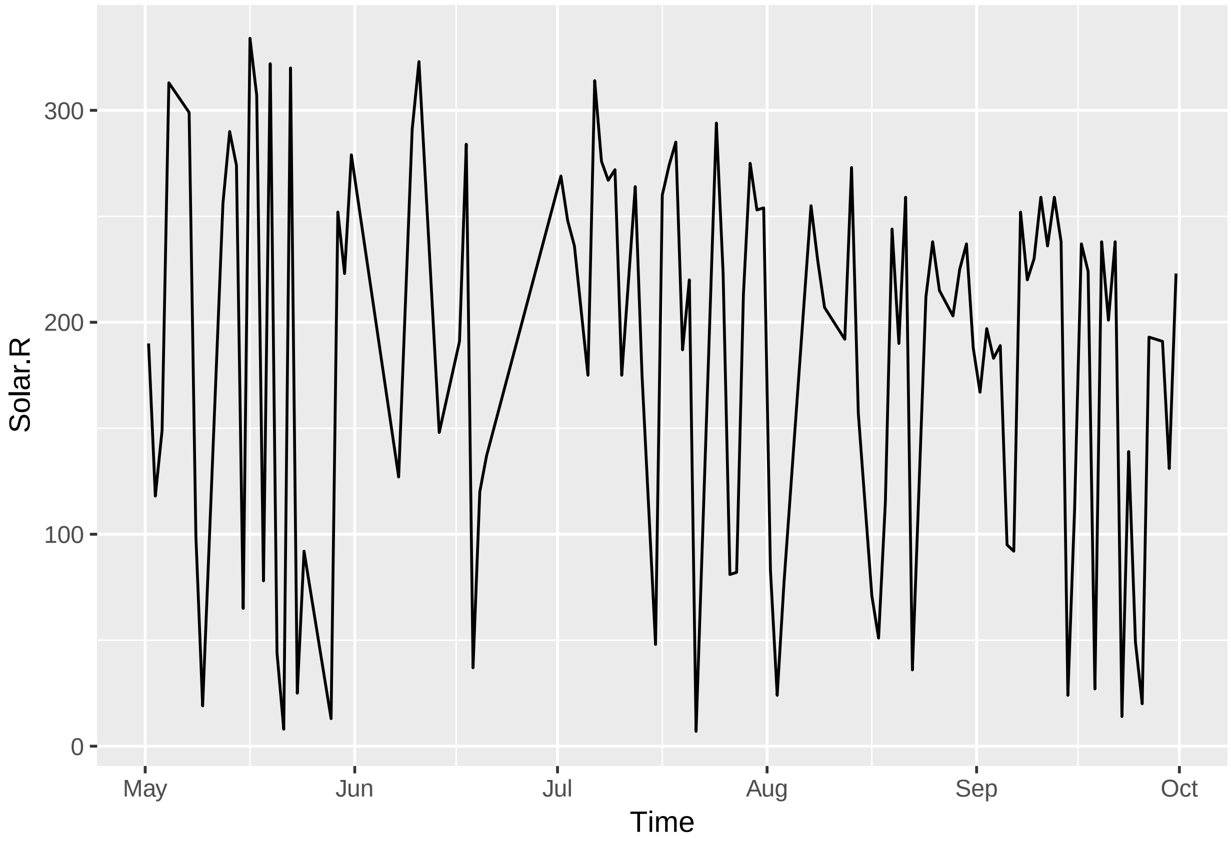
```
ggplot(dataAQ, aes(x = Time)) + geom_line(aes(y = Temp))
```



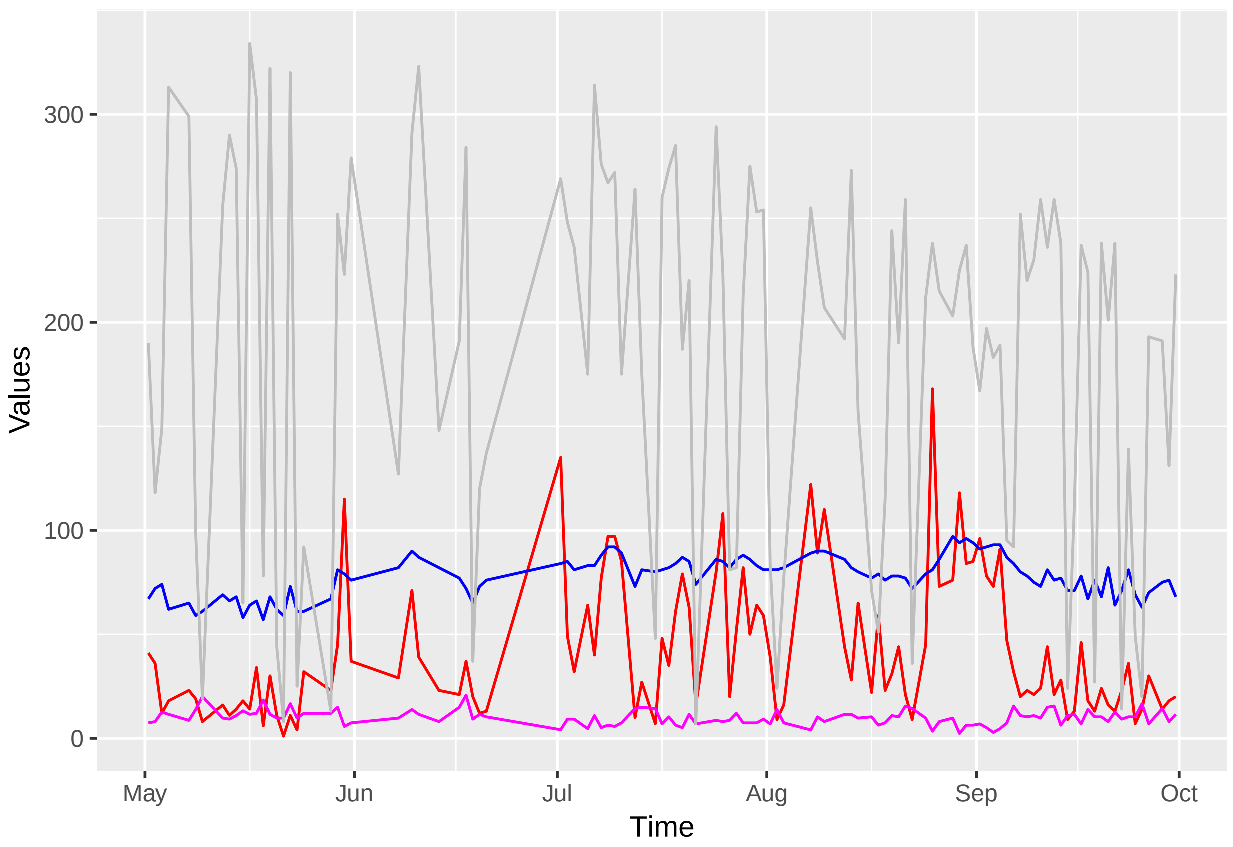
```r
ggplot(dataAQ, aes(x = Time)) + geom_line(aes(y = Wind))
```



```
ggplot(dataAQ, aes(x = Time)) + geom_line(aes(y = Solar.R))
```

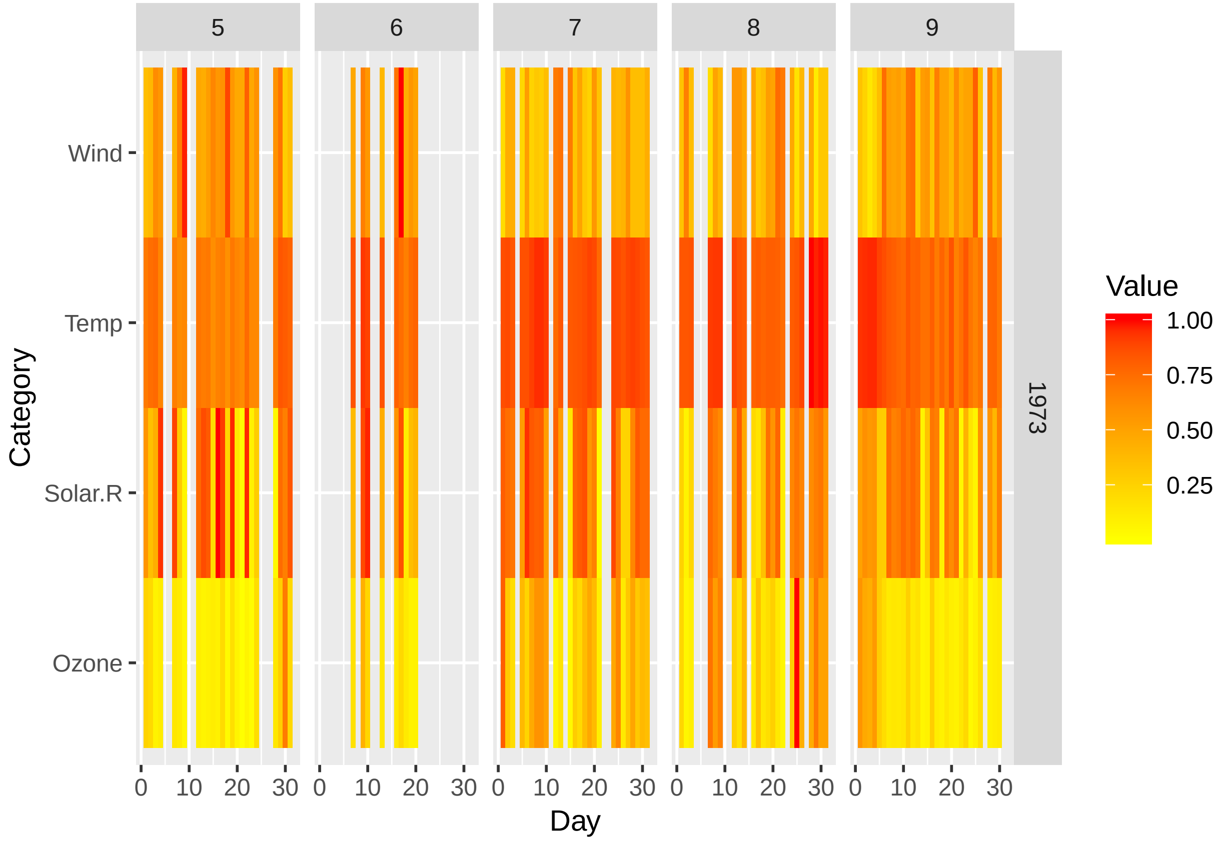


```
# all of them and with colors
ggplot(dataAQ, aes(x = Time)) +
  geom_line(aes(y = Ozone), color = 2) +
  geom_line(aes(y = Temp), color = 4) +
  geom_line(aes(y = Wind), color = 6) +
  geom_line(aes(y = Solar.R), color = 8) +
  ylab("Values") + xlab("Time")
```

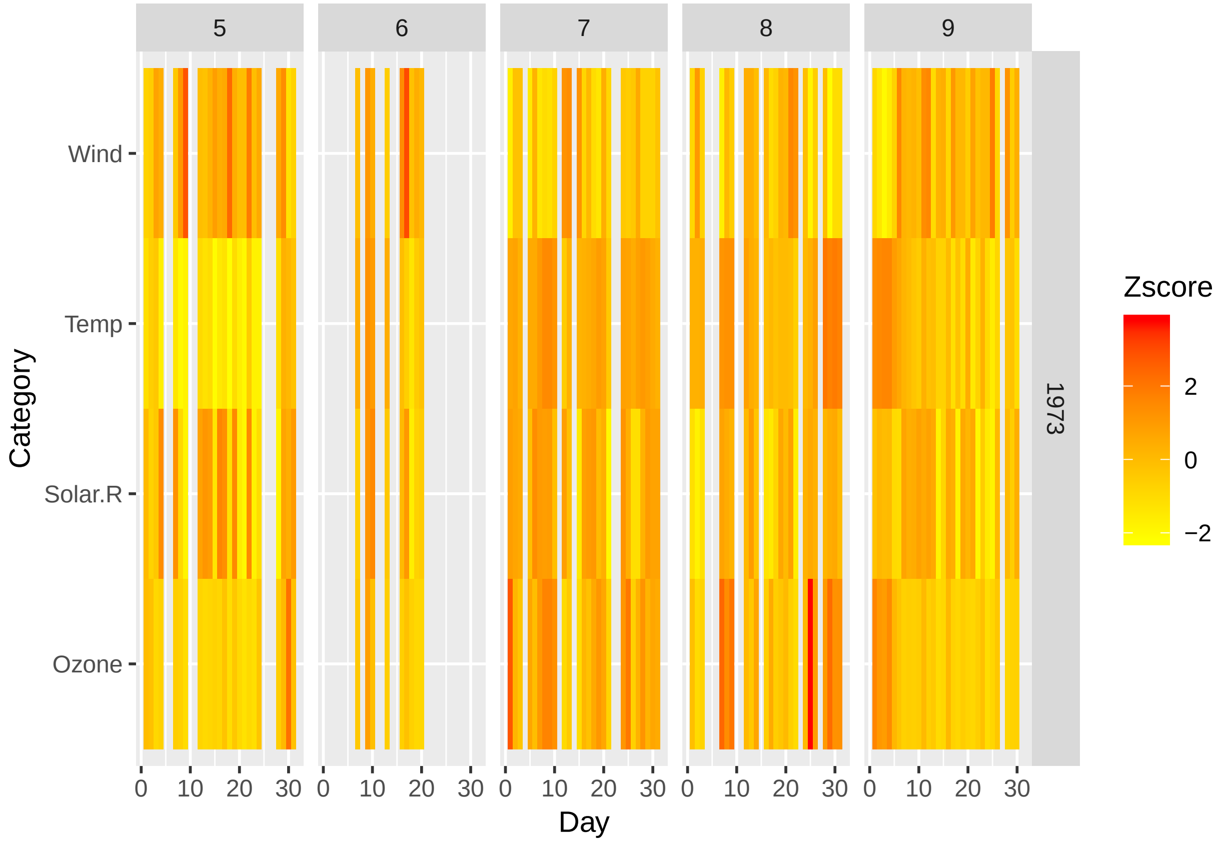


```
# mutating data to fit heat map
# filtering data so that the values Ozone, Temp, Wind, and Solar are categories
# Going to make the data fit the graph I need to create
OzoneDf = dataAQ[ , c(1,5,6,7,8)]
OzoneDf$Category = c(rep("Ozone", 111))
newColNames = colnames(OzoneDf) # col names
newColNames[1] = "Value"
colnames(OzoneDf) = newColNames # Standardize columns across dfs
OzoneDf$Value = OzoneDf$Value/max(dataAQ$Ozone) # convert to percent
OzoneDf$Zscore = scale(OzoneDf$Value)[ , 1] # z scores

# repeat process for each value
SolarDf = dataAQ[ , c(2,5,6,7,8)]
SolarDf$Category = c(rep("Solar.R", 111))
colnames(SolarDf) = newColNames
SolarDf$Value = SolarDf$Value/max(dataAQ$Solar.R)
SolarDf$Zscore = scale(SolarDf$Value)[ , 1]

WindDf = dataAQ[ , c(3,5,6,7,8)]
WindDf$Category = c(rep("Wind", 111))
colnames(WindDf) = newColNames
WindDf$Value = WindDf$Value/max(dataAQ$Wind)
WindDf$Zscore = scale(WindDf$Value)[ , 1]

TempDf = dataAQ[ , c(4,5,6,7,8)]
TempDf$Category = c(rep("Temp", 111))
```

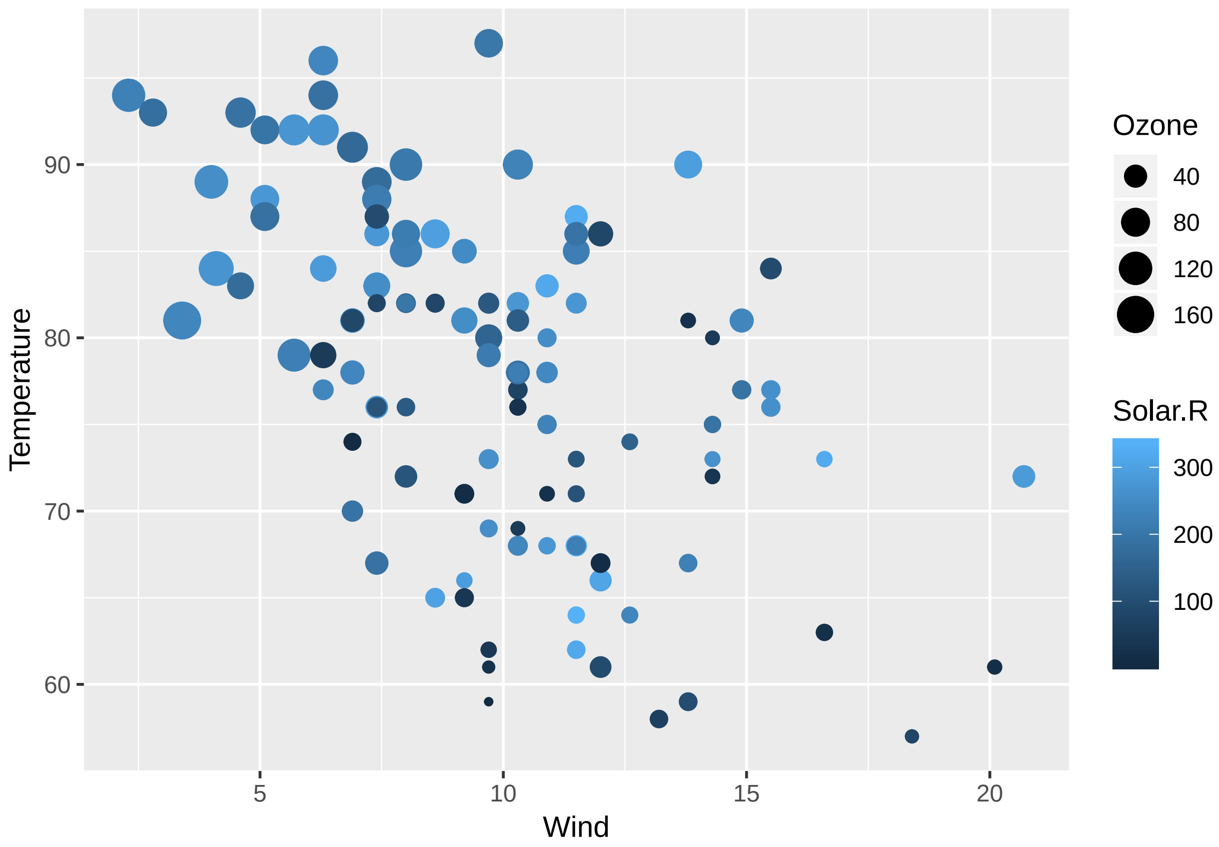
```r
colnames(TempDf) = newColNames
TempDf$Value = TempDf$Value/max(dataAQ$Temp)
TempDf$Zscore = scale(TempDf$Value)[ , 1]

newDataAQ = rbind(OzoneDf, SolarDf, WindDf, TempDf) # combine all into one df

# plot heat map w/ percents
ggplot(newDataAQ, aes(Day, Category)) + geom_tile(aes(fill = Value)) +
  facet_grid(Year ~ Month) +
  #scale_fill_gradientn(colours = heat.colors(500))
  scale_fill_gradient(low = "yellow", high = "red")
```



```r
# plot heat map 2 z scores
ggplot(newDataAQ, aes(Day, Category)) + geom_tile(aes(fill = Zscore)) +
  facet_grid(Year ~ Month) +
  #scale_fill_gradientn(colours = heat.colors(500))
  scale_fill_gradient(low = "yellow", high = "red")
```



```
# scatter plot
ggplot(dataAQ ,aes(Wind, Temp)) +
  geom_point(aes(color = Solar.R, size = Ozone)) +
  xlab("Wind") + ylab("Temperature")
```



```
##################################################################################################
# FINAL ANALYSIS
# Some of the patterns I see are that higher ozone values seem to be correlated with higher temperature
# month of July had some of the highest values overall amoung all 4 variables.
#
# I found the scatter plot to be the most useful because it had all the data on one graph and the repre
# for each value was different and easy to understand.  I also found the box plots and heat map to be u
# for different reasons.  The box plots where a good visual to see where and when most of the data was 
# while the heat map also had all the values on one chart for correlation, however, I didn't find it as
# as the scatter plot.
```