Sharat_Sripada_HW6.R

ssharat

2020-02-23

```
#
#       Course: IST-687
#       Name: Sharat Sripada
#       Homework #4
#       Due Date: 2/23/2020
#       Date Submitted: 2/23/2020
#       Topic: Using ggplot to visualize data (Air Quality Analysis)

# install.packages("ggplot2")
# install.packages("reshape2")

library("ggplot2")
library("reshape2")

air <- airquality

# Function to plot a histogram given a data-frame, column-name
my_ggplot_hist <- function(my_df, my_col){
  g <- ggplot(my_df, aes(x=my_col))
  g <- g + geom_histogram(binwidth = 5, color = 'white', fill = 'black')
  g

}

# Function to plot a box-plot given a data-frame, column-name
my_ggplot_box <- function(my_df, my_col){
  g <- ggplot(my_df, aes(x = ' ', y = my_col))
  g <- g + geom_boxplot()
  g
}

# Function to plot a line plot given a data-frame, x & y axes
my_ggplot_line <- function(my_df, my_x, my_y){
  g <- ggplot(my_df, aes(x = my_x, y = my_y, group = 1))
  g <- g + geom_line()
  g
}


# Step-2: Clean the Data
# Calculate mean for the Ozone column
```

```r
ozone_mean <- mean(air$Ozone, na.rm = TRUE)
cat("Replacing NAs in Ozone column with mean =", ozone_mean)
```

```
## Replacing NAs in Ozone column with mean = 42.12931
```

```r
air$Ozone[is.na(air$Ozone)] <- ozone_mean

# Calculate mean for the Solar.R column
solar_mean <- mean(air$Solar.R, na.rm = TRUE)
cat("Replacing NAs in Solar column with mean =", solar_mean)
```

```
## Replacing NAs in Solar column with mean = 185.9315
```

```r
air$Solar.R[is.na(air$Solar.R)] <- solar_mean

# Print the columns & examine for NAs
print(air$Ozone)
```

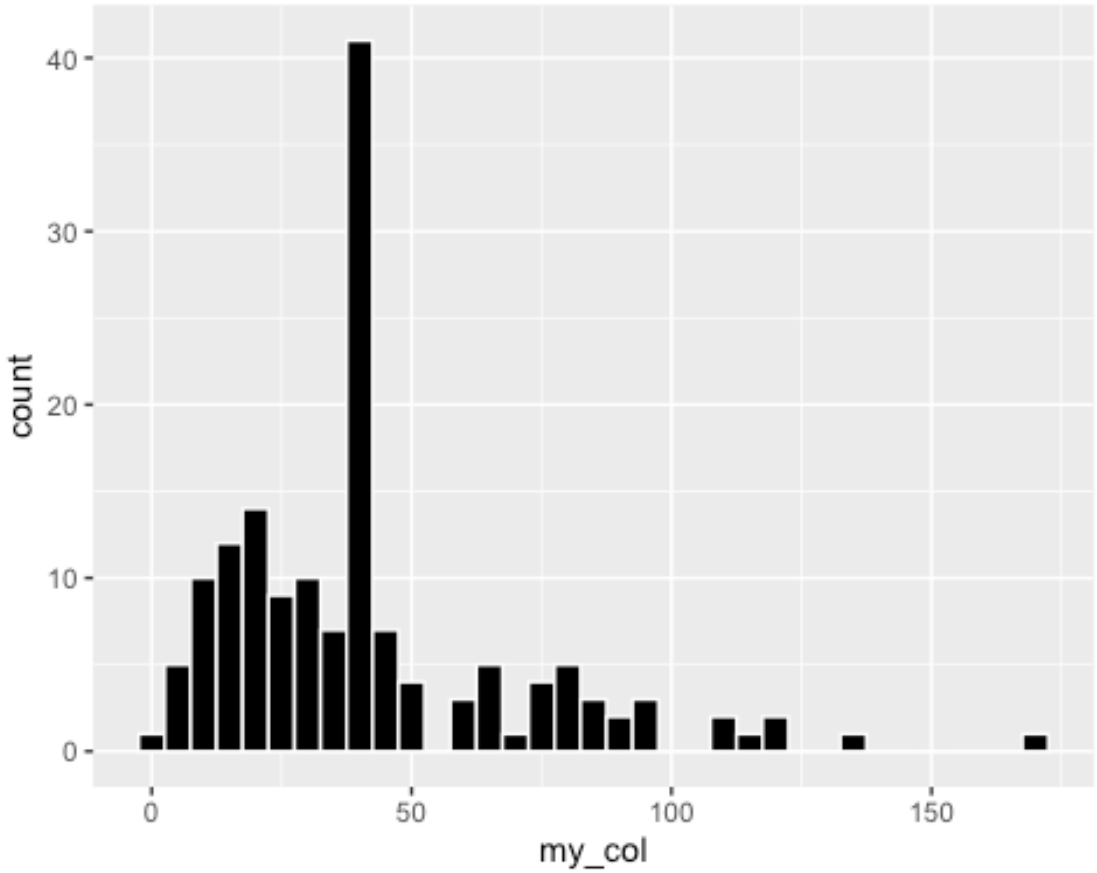
```
##   [1]  41.00000  36.00000  12.00000  18.00000  42.12931  28.00000
23.00000
##   [8]  19.00000   8.00000  42.12931   7.00000  16.00000  11.00000
14.00000
##  [15]  18.00000  14.00000  34.00000   6.00000  30.00000  11.00000
1.00000
##  [22]  11.00000   4.00000  32.00000  42.12931  42.12931  42.12931
23.00000
##  [29]  45.00000 115.00000  37.00000  42.12931  42.12931  42.12931
42.12931
##  [36]  42.12931  42.12931  29.00000  42.12931  71.00000  39.00000
42.12931
##  [43]  42.12931  23.00000  42.12931  42.12931  21.00000  37.00000
20.00000
##  [50]  12.00000  13.00000  42.12931  42.12931  42.12931  42.12931
42.12931
##  [57]  42.12931  42.12931  42.12931  42.12931  42.12931 135.00000
49.00000
##  [64]  32.00000  42.12931  64.00000  40.00000  77.00000  97.00000
97.00000
##  [71]  85.00000  42.12931  10.00000  27.00000  42.12931   7.00000
48.00000
##  [78]  35.00000  61.00000  79.00000  63.00000  16.00000  42.12931
42.12931
##  [85]  80.00000 108.00000  20.00000  52.00000  82.00000  50.00000
64.00000
##  [92]  59.00000  39.00000   9.00000  16.00000  78.00000  35.00000
66.00000
##  [99] 122.00000  89.00000 110.00000  42.12931  42.12931  44.00000
28.00000
## [106]  65.00000  42.12931  22.00000  59.00000  23.00000  31.00000
44.00000
## [113]  21.00000   9.00000  42.12931  45.00000 168.00000  73.00000
```

```
42.12931
## [120]  76.00000 118.00000  84.00000  85.00000  96.00000  78.00000
73.00000
## [127]  91.00000  47.00000  32.00000  20.00000  23.00000  21.00000
24.00000
## [134]  44.00000  21.00000  28.00000   9.00000  13.00000  46.00000
18.00000
## [141]  13.00000  24.00000  16.00000  13.00000  23.00000  36.00000
7.00000
## [148]  14.00000  30.00000  42.12931  14.00000  18.00000  20.00000
```

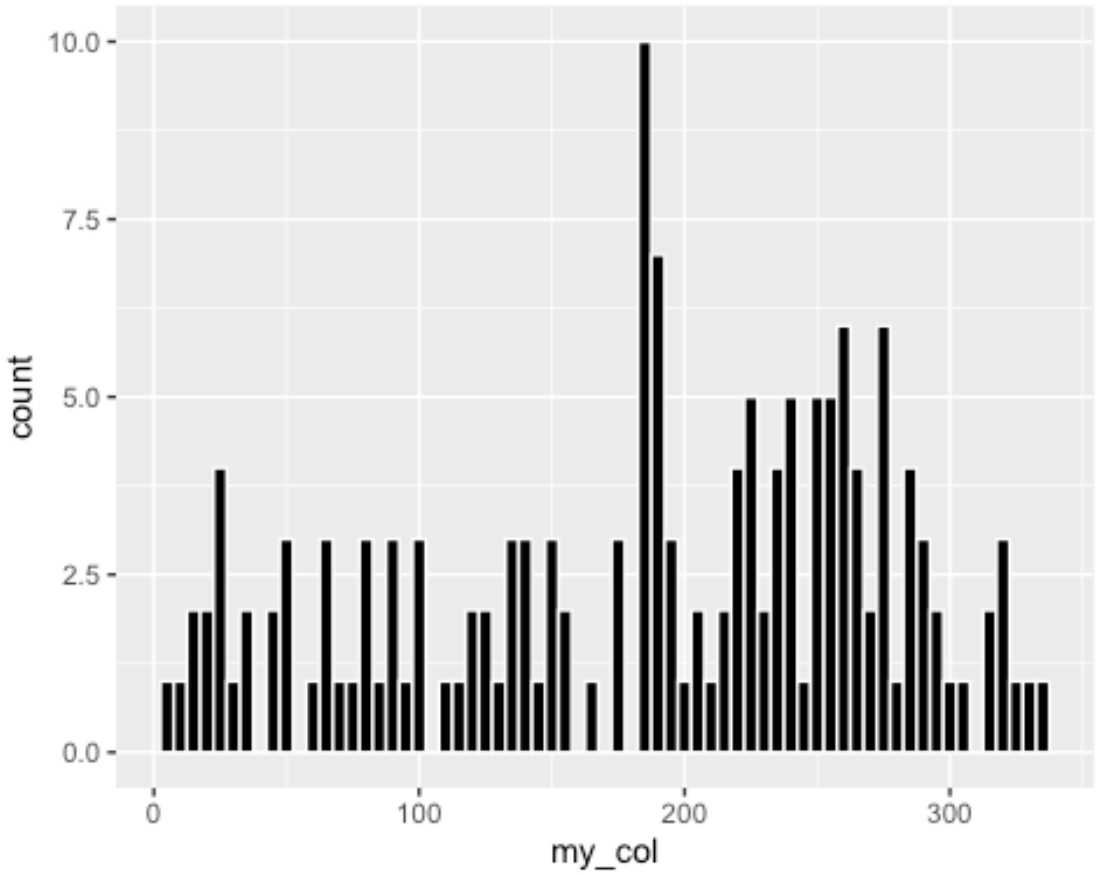
```
print(air$Solar.R)
```

```
##   [1] 190.0000 118.0000 149.0000 313.0000 185.9315 185.9315 299.0000
99.0000
##   [9]  19.0000 194.0000 185.9315 256.0000 290.0000 274.0000  65.0000
334.0000
##  [17] 307.0000  78.0000 322.0000  44.0000   8.0000 320.0000  25.0000
92.0000
##  [25]  66.0000 266.0000 185.9315  13.0000 252.0000 223.0000 279.0000
286.0000
##  [33] 287.0000 242.0000 186.0000 220.0000 264.0000 127.0000 273.0000
291.0000
##  [41] 323.0000 259.0000 250.0000 148.0000 332.0000 322.0000 191.0000
284.0000
##  [49]  37.0000 120.0000 137.0000 150.0000  59.0000  91.0000 250.0000
135.0000
##  [57] 127.0000  47.0000  98.0000  31.0000 138.0000 269.0000 248.0000
236.0000
##  [65] 101.0000 175.0000 314.0000 276.0000 267.0000 272.0000 175.0000
139.0000
##  [73] 264.0000 175.0000 291.0000  48.0000 260.0000 274.0000 285.0000
187.0000
##  [81] 220.0000   7.0000 258.0000 295.0000 294.0000 223.0000  81.0000
82.0000
##  [89] 213.0000 275.0000 253.0000 254.0000  83.0000  24.0000  77.0000
185.9315
##  [97] 185.9315 185.9315 255.0000 229.0000 207.0000 222.0000 137.0000
192.0000
## [105] 273.0000 157.0000  64.0000  71.0000  51.0000 115.0000 244.0000
190.0000
## [113] 259.0000  36.0000 255.0000 212.0000 238.0000 215.0000 153.0000
203.0000
## [121] 225.0000 237.0000 188.0000 167.0000 197.0000 183.0000 189.0000
95.0000
## [129]  92.0000 252.0000 220.0000 230.0000 259.0000 236.0000 259.0000
238.0000
## [137]  24.0000 112.0000 237.0000 224.0000  27.0000 238.0000 201.0000
238.0000
## [145]  14.0000 139.0000  49.0000  20.0000 193.0000 145.0000 191.0000
```

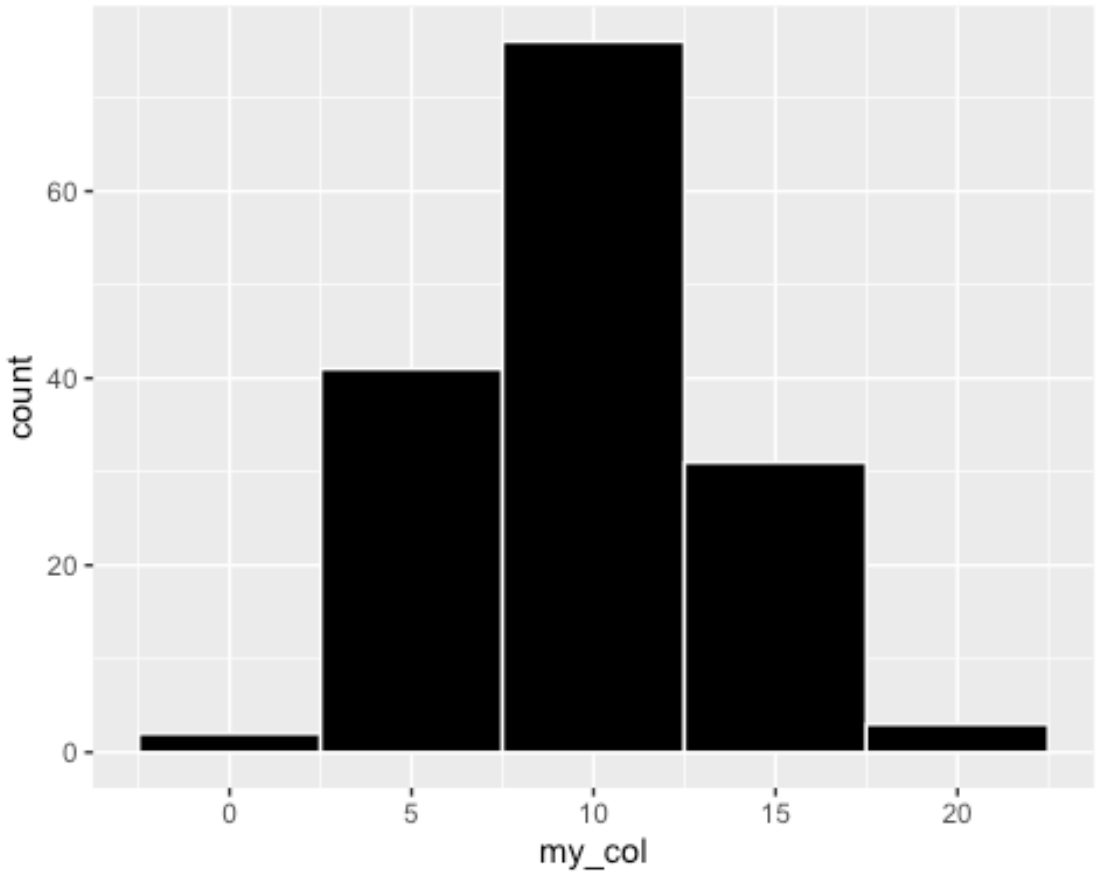
```
131.0000
## [153] 223.0000

# Step-3:
# Create a histogram for the Ozone column as X var.
my_ggplot_hist(air, air$Ozone)
```



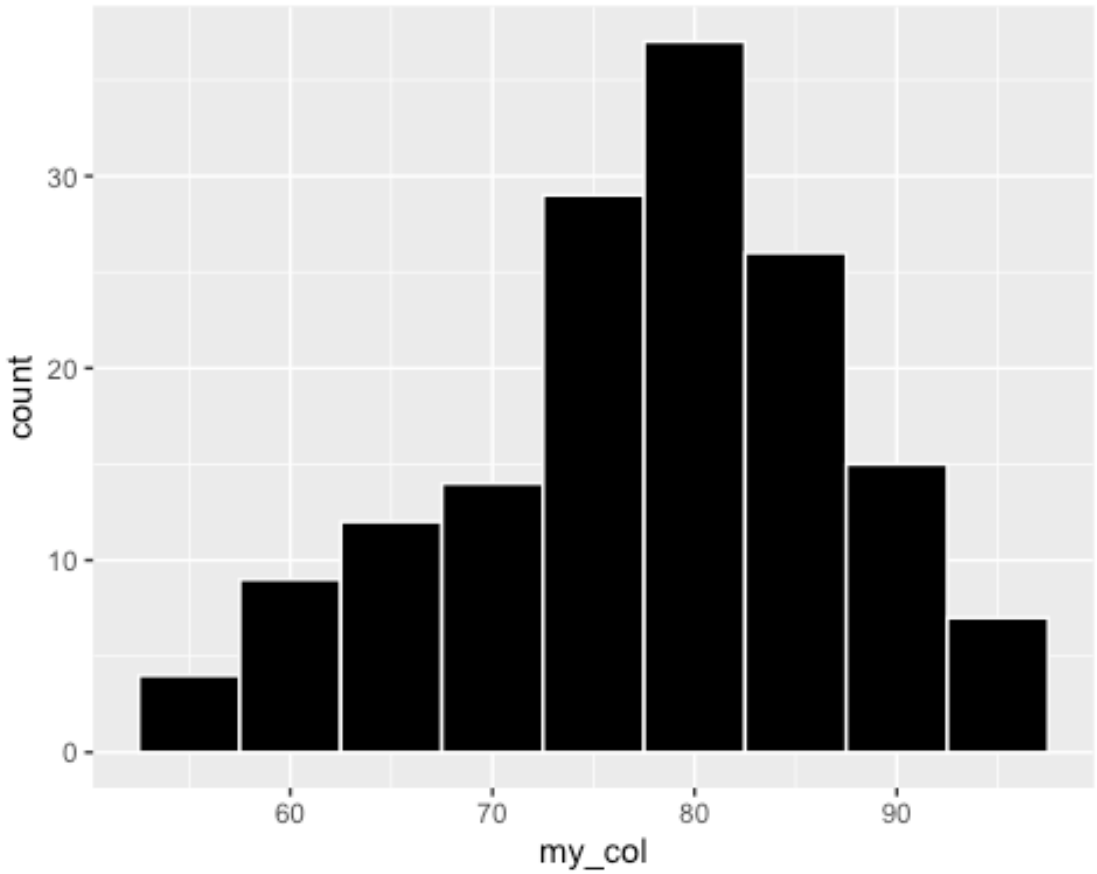

```
# Create a histogram for the Solar column as X var.
my_ggplot_hist(air, air$Solar.R)
```



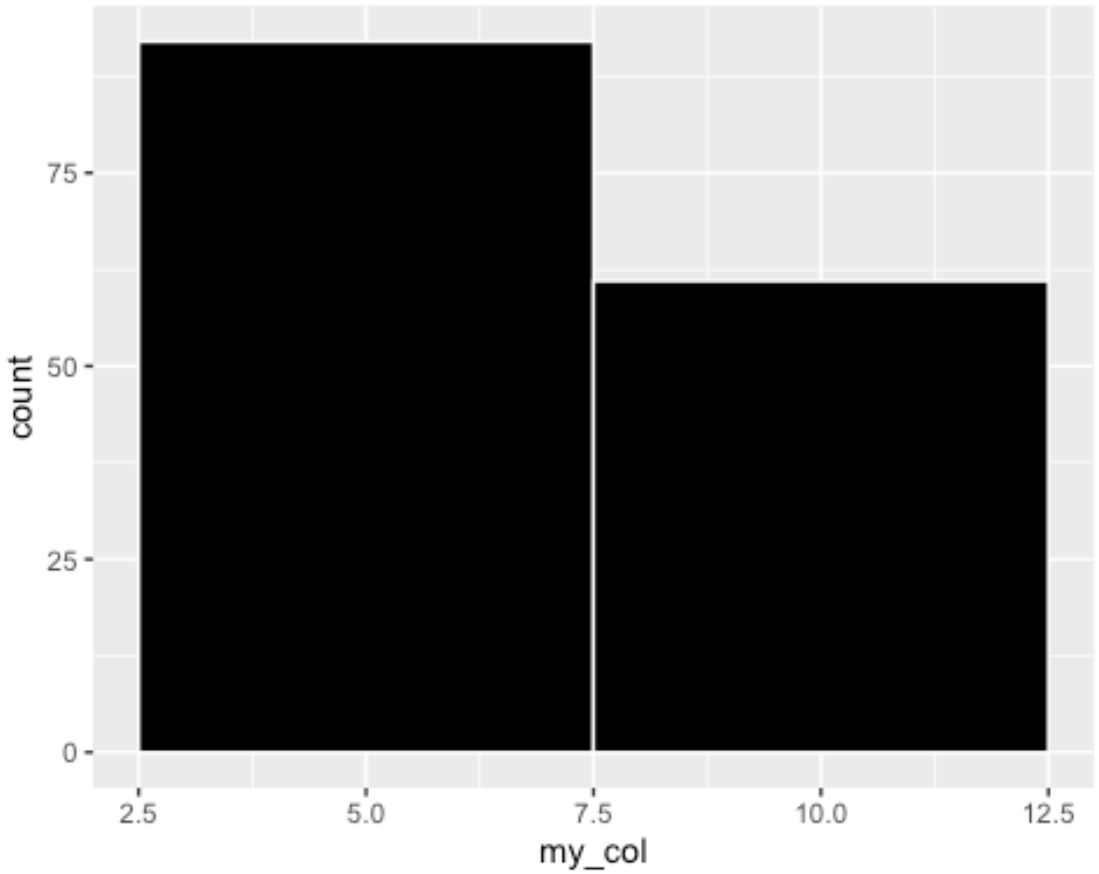

```
# Create a histogram for the Wind column as X var.
my_ggplot_hist(air, air$Wind)
```



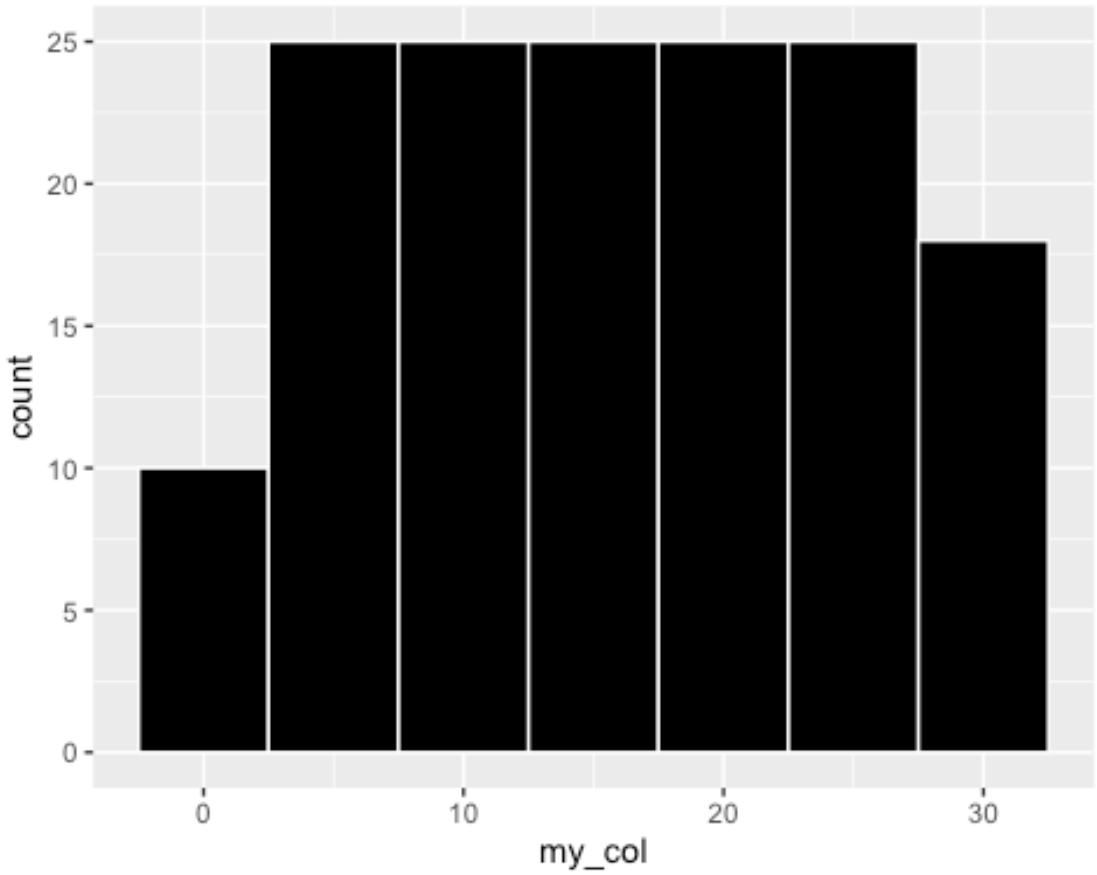

```
# Create a histogram for the Temp column as X var.
my_ggplot_hist(air, air$Temp)
```



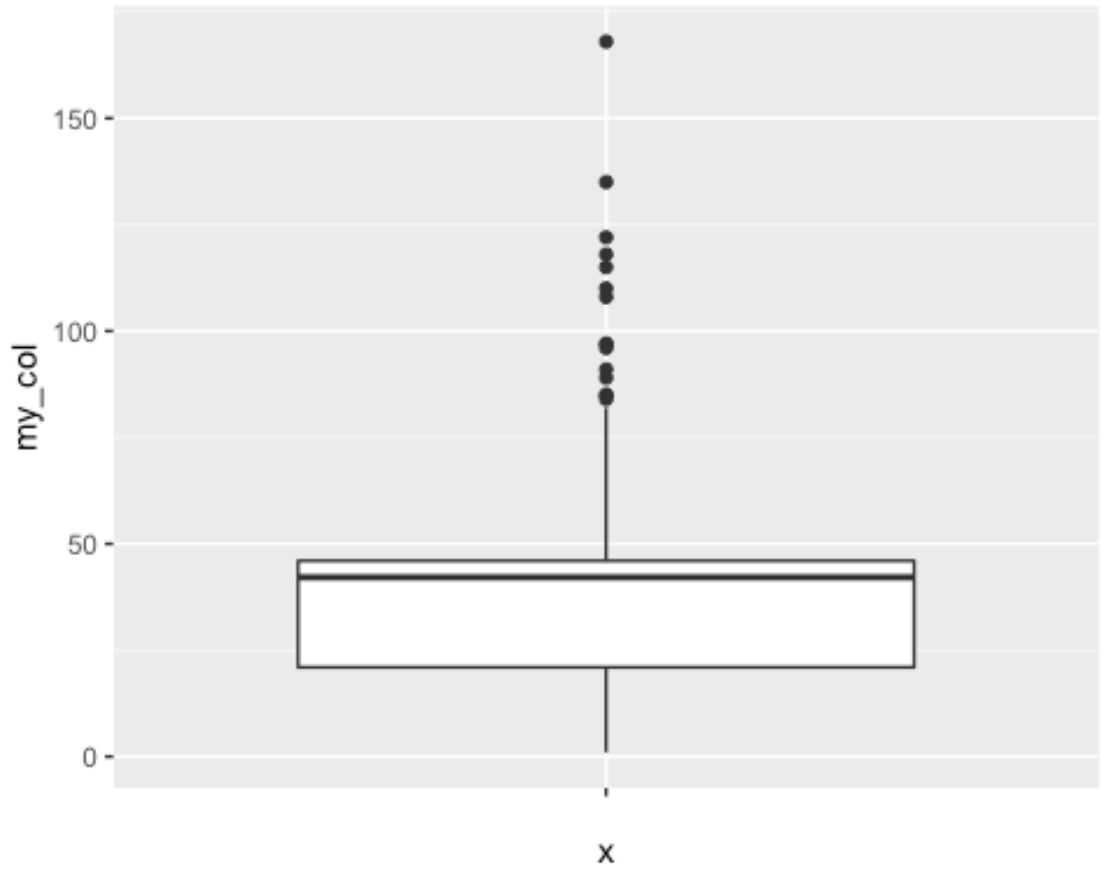

```
# Create a histogram for the Month column as X var.
my_ggplot_hist(air, air$Month)
```



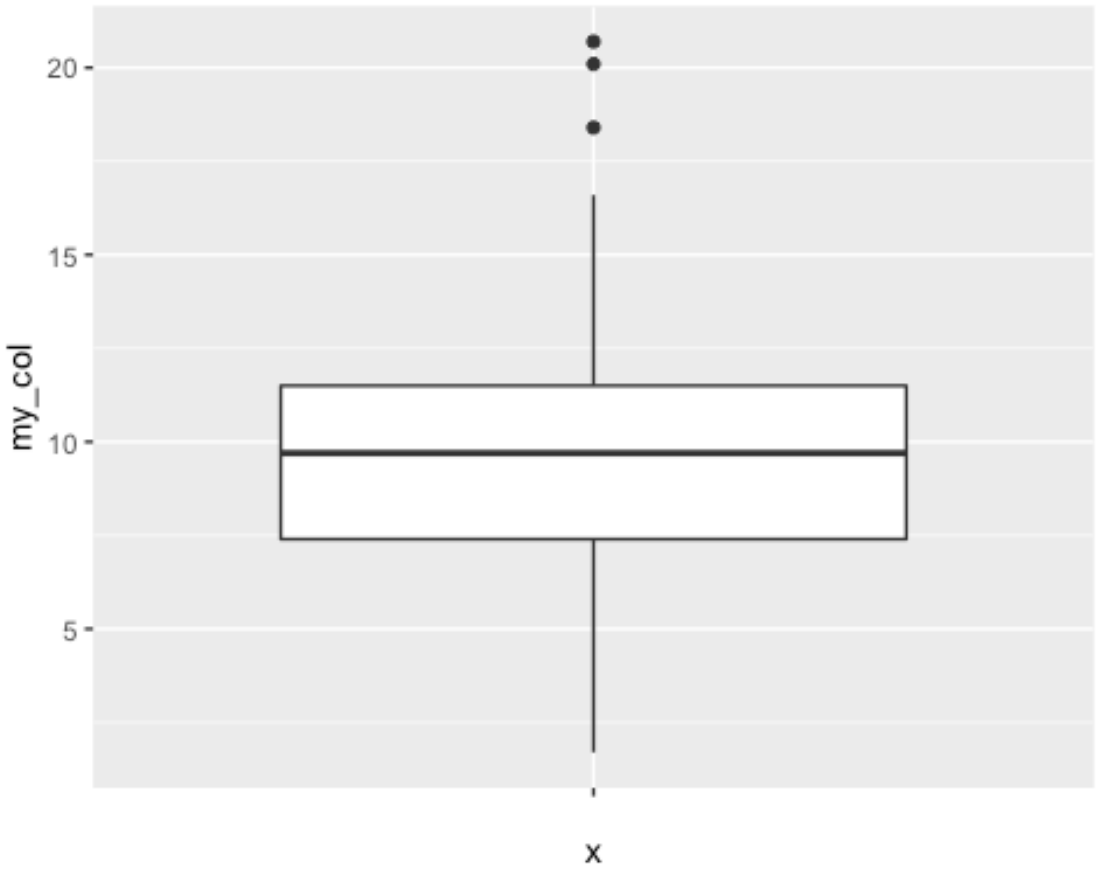

```
# Create a histogram for the Day column as X var.
my_ggplot_hist(air, air$Day)
```




```
# Create a box-plot for Ozone
my_ggplot_box(air, air$Ozone)
```




```
# Create a box-plot for Wind
my_ggplot_box(air, air$Wind)
```



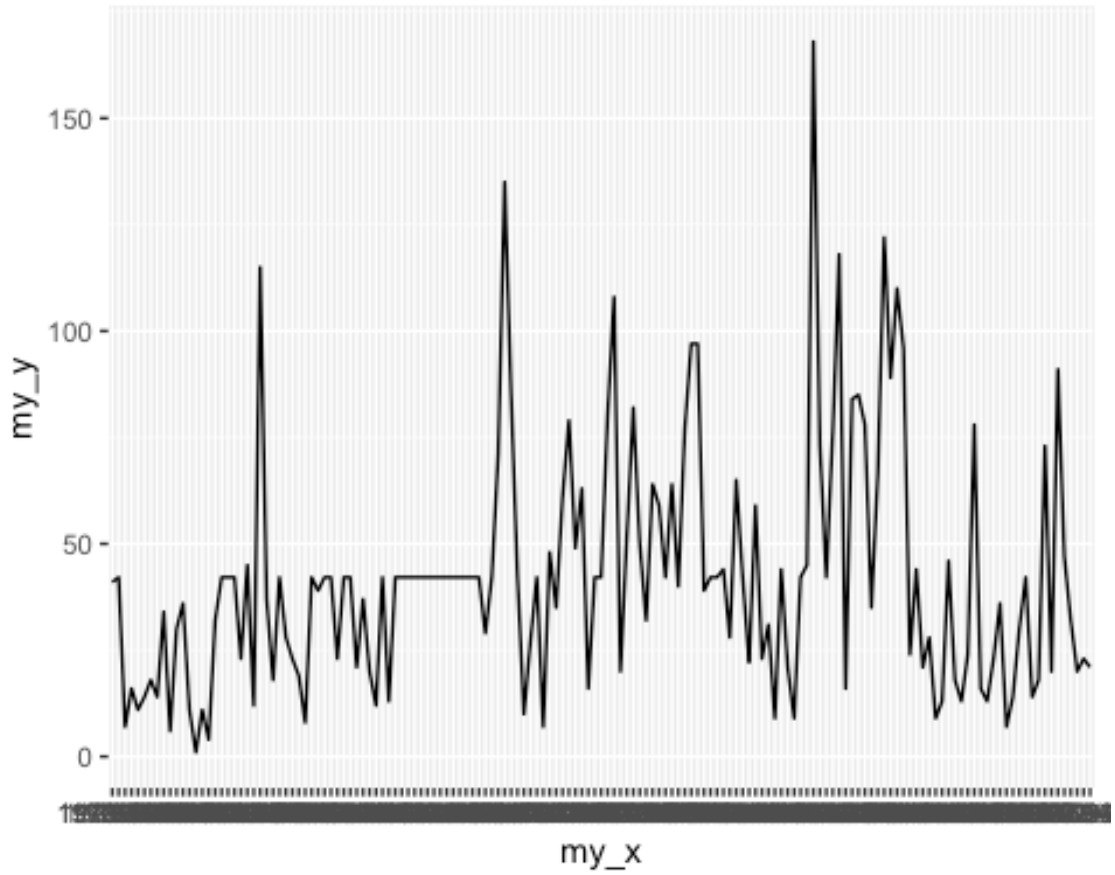

```
# Reading the box-plot visually:
# min = 1.25
# max = 20.625
# Q1 = 7.5
# Q3 = 11.25
# Median = 9.375
summary(air$Wind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.700   7.400   9.700   9.958  11.500  20.700
```

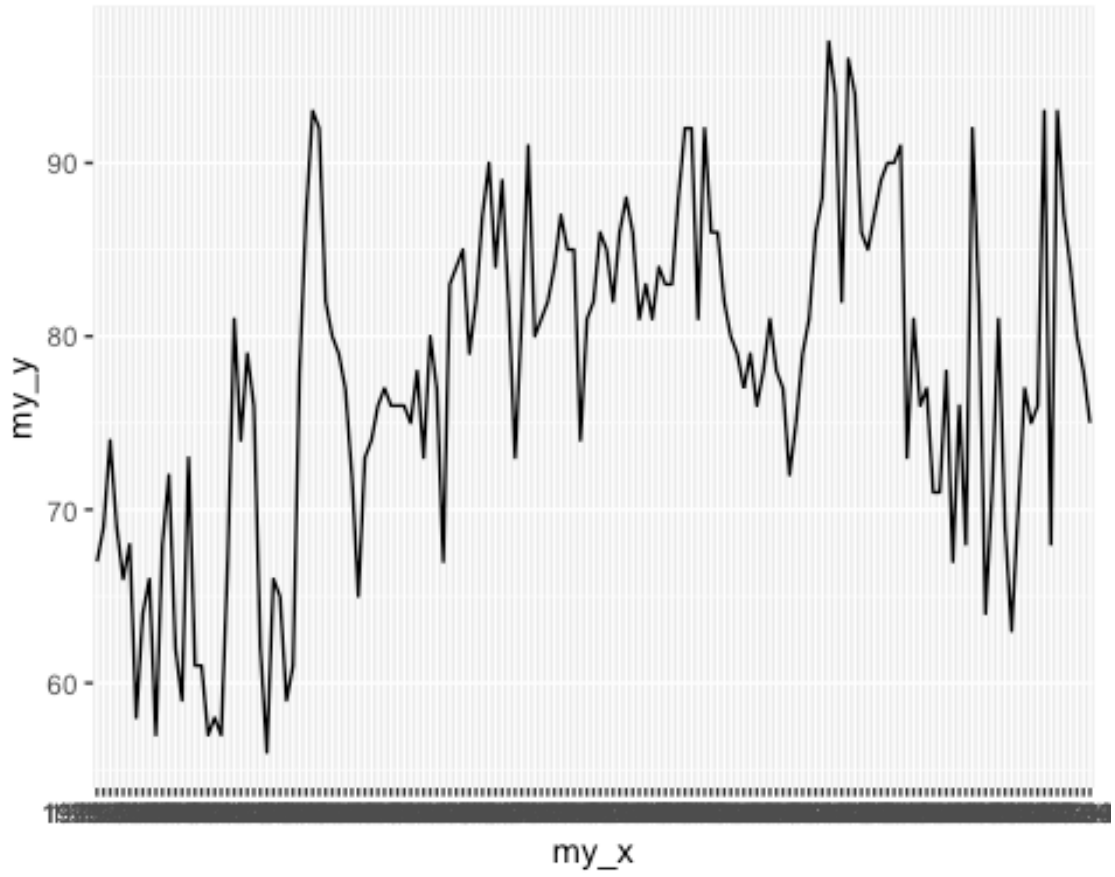
```
# TO-DO: Work on reading box-plots more accurately - Can you change Axis to
granular values?

# Step-3:
# Create a new column in df air called Date
air$Date <- gsub(" ", "", paste("1973", "-", air$Month, "-", air$Day))
print(air$Date)
```

```
##   [1] "1973-5-1"  "1973-5-2"  "1973-5-3"  "1973-5-4"  "1973-5-5"  "1973-5-
6"
##   [7] "1973-5-7"  "1973-5-8"  "1973-5-9"  "1973-5-10" "1973-5-11" "1973-5-
12"
##  [13] "1973-5-13" "1973-5-14" "1973-5-15" "1973-5-16" "1973-5-17" "1973-5-
```

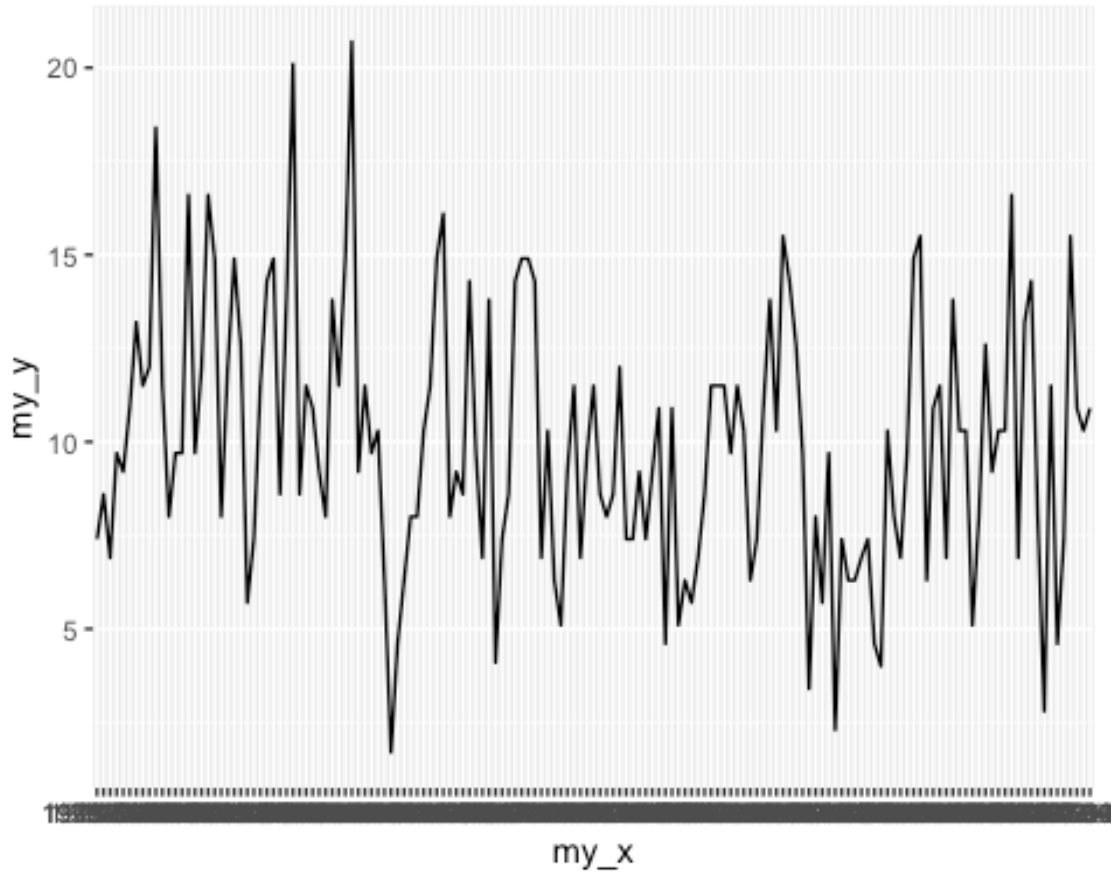
```
18"
##  [19] "1973-5-19" "1973-5-20" "1973-5-21" "1973-5-22" "1973-5-23" "1973-5-
24"
##  [25] "1973-5-25" "1973-5-26" "1973-5-27" "1973-5-28" "1973-5-29" "1973-5-
30"
##  [31] "1973-5-31" "1973-6-1"  "1973-6-2"  "1973-6-3"  "1973-6-4"  "1973-6-
5"
##  [37] "1973-6-6"  "1973-6-7"  "1973-6-8"  "1973-6-9"  "1973-6-10" "1973-6-
11"
##  [43] "1973-6-12" "1973-6-13" "1973-6-14" "1973-6-15" "1973-6-16" "1973-6-
17"
##  [49] "1973-6-18" "1973-6-19" "1973-6-20" "1973-6-21" "1973-6-22" "1973-6-
23"
##  [55] "1973-6-24" "1973-6-25" "1973-6-26" "1973-6-27" "1973-6-28" "1973-6-
29"
##  [61] "1973-6-30" "1973-7-1"  "1973-7-2"  "1973-7-3"  "1973-7-4"  "1973-7-
5"
##  [67] "1973-7-6"  "1973-7-7"  "1973-7-8"  "1973-7-9"  "1973-7-10" "1973-7-
11"
##  [73] "1973-7-12" "1973-7-13" "1973-7-14" "1973-7-15" "1973-7-16" "1973-7-
17"
##  [79] "1973-7-18" "1973-7-19" "1973-7-20" "1973-7-21" "1973-7-22" "1973-7-
23"
##  [85] "1973-7-24" "1973-7-25" "1973-7-26" "1973-7-27" "1973-7-28" "1973-7-
29"
##  [91] "1973-7-30" "1973-7-31" "1973-8-1"  "1973-8-2"  "1973-8-3"  "1973-8-
4"
##  [97] "1973-8-5"  "1973-8-6"  "1973-8-7"  "1973-8-8"  "1973-8-9"  "1973-8-
10"
## [103] "1973-8-11" "1973-8-12" "1973-8-13" "1973-8-14" "1973-8-15" "1973-8-
16"
## [109] "1973-8-17" "1973-8-18" "1973-8-19" "1973-8-20" "1973-8-21" "1973-8-
22"
## [115] "1973-8-23" "1973-8-24" "1973-8-25" "1973-8-26" "1973-8-27" "1973-8-
28"
## [121] "1973-8-29" "1973-8-30" "1973-8-31" "1973-9-1"  "1973-9-2"  "1973-9-
3"
## [127] "1973-9-4"  "1973-9-5"  "1973-9-6"  "1973-9-7"  "1973-9-8"  "1973-9-
9"
## [133] "1973-9-10" "1973-9-11" "1973-9-12" "1973-9-13" "1973-9-14" "1973-9-
15"
## [139] "1973-9-16" "1973-9-17" "1973-9-18" "1973-9-19" "1973-9-20" "1973-9-
21"
## [145] "1973-9-22" "1973-9-23" "1973-9-24" "1973-9-25" "1973-9-26" "1973-9-
27"
## [151] "1973-9-28" "1973-9-29" "1973-9-30"

# Create a line plot for Ozone
my_ggplot_line(air, air$Date, air$Ozone)
```



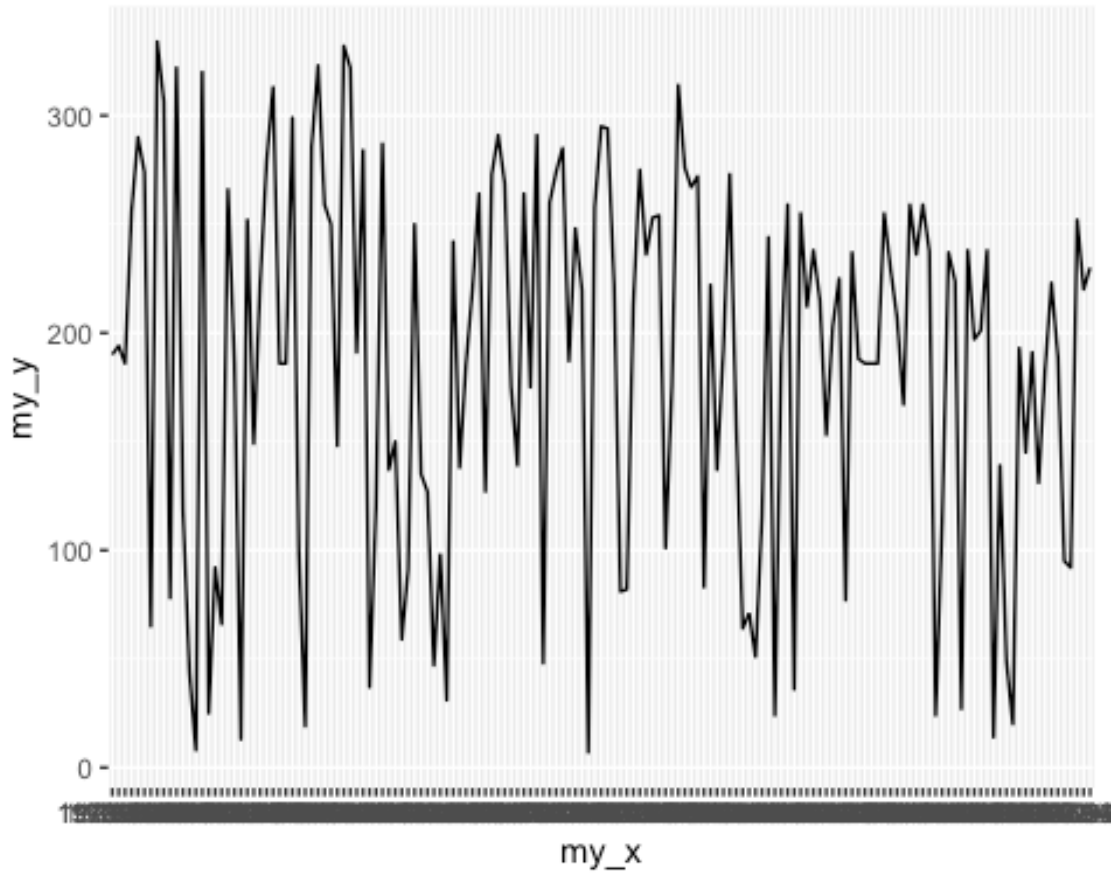

```
# Create a line plot for Temp
my_ggplot_line(air, air$Date, air$Temp)
```



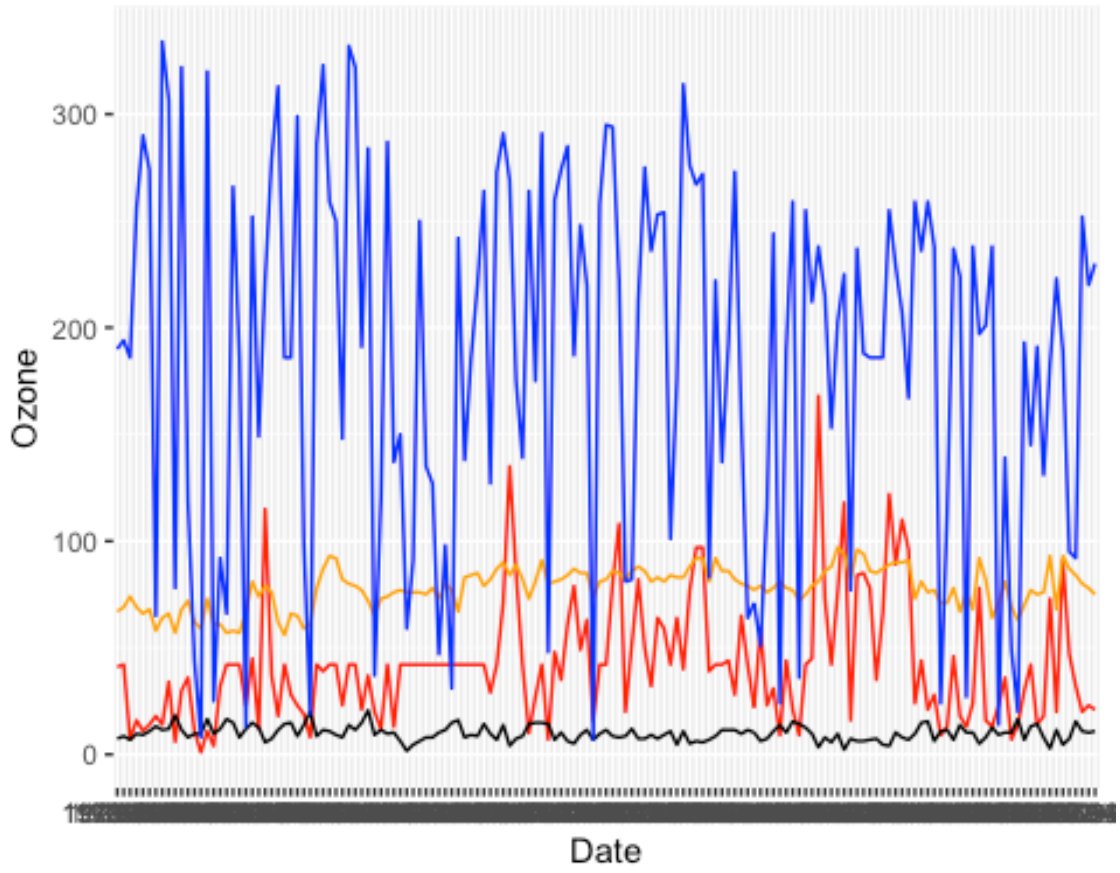

```
# Create a line plot for Wind
my_ggplot_line(air, air$Date, air$Wind)
```



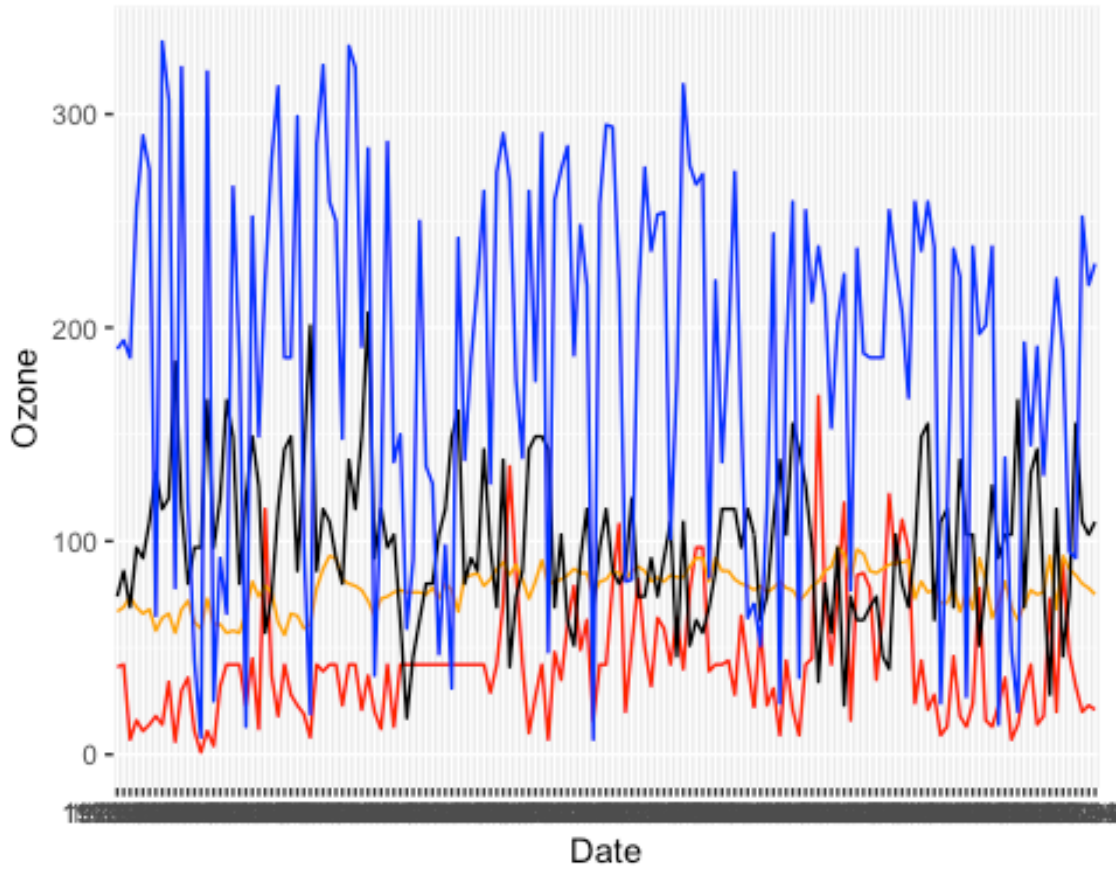

```r
# Create a line plot for Solar.R
my_ggplot_line(air, air$Date, air$Solar.R)
```



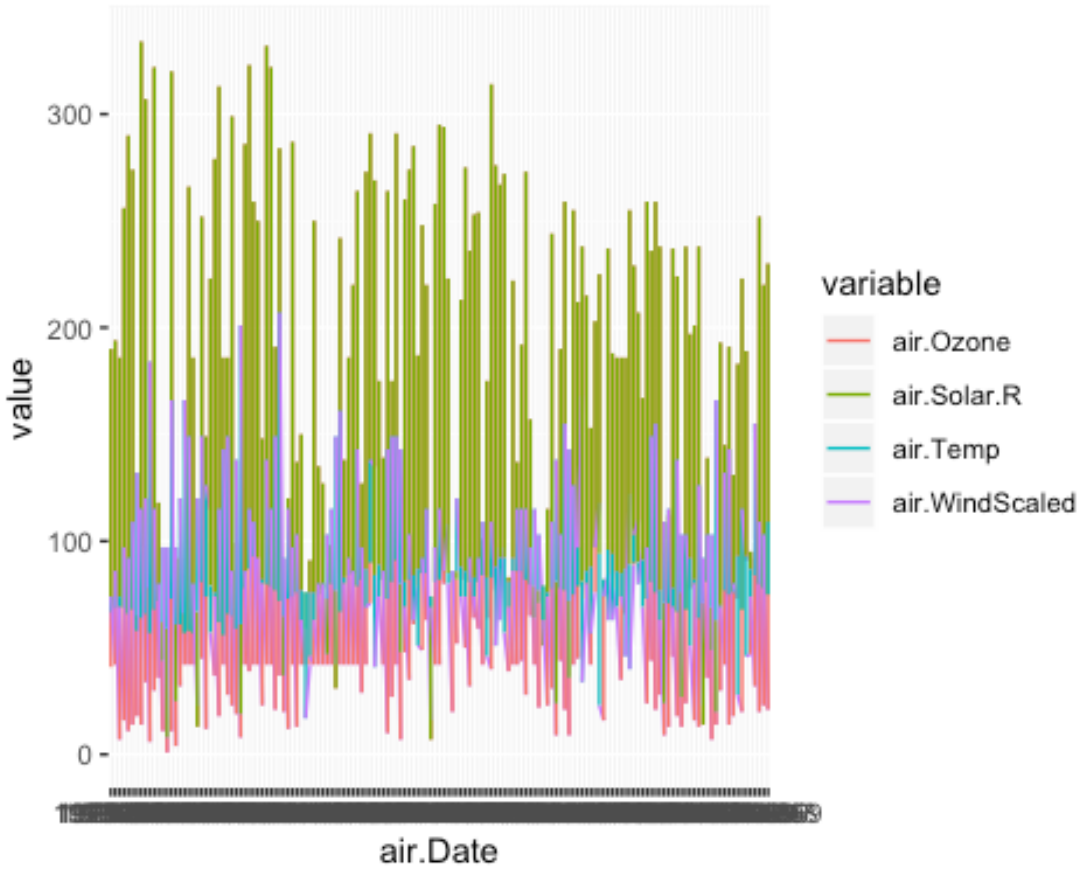

```r
# Create one chart comprising all 4x lines in different colors
g <- ggplot(air, aes(x = Date, group=1))
g <- g + geom_line(aes(y = Ozone, group=1), color = "red")
g <- g + geom_line(aes(y = Temp, group=1), color = "orange")
g <- g + geom_line(aes(y = Wind, group=1), color = "black")
g <- g + geom_line(aes(y = Solar.R, group=1), color = "blue")
g
```



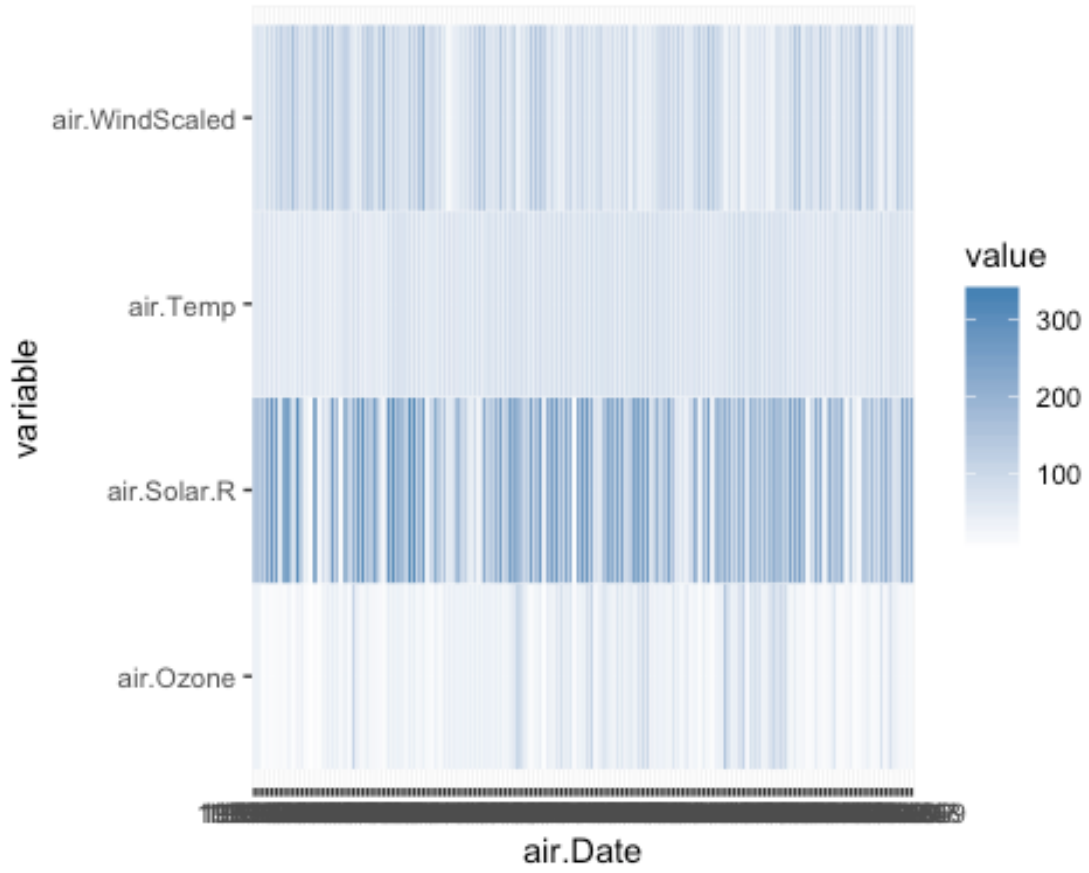

```r
# Scale the Wind data/column - Create a new column WindScaled in air df
air$WindScaled <- air$Wind * 10
g <- ggplot(air, aes(x = Date, group=1))
g <- g + geom_line(aes(y = Ozone, group=1), color = "red")
g <- g + geom_line(aes(y = Temp, group=1), color = "orange")
g <- g + geom_line(aes(y = WindScaled, group=1), color = "black")
g <- g + geom_line(aes(y = Solar.R, group=1), color = "blue")
g
```



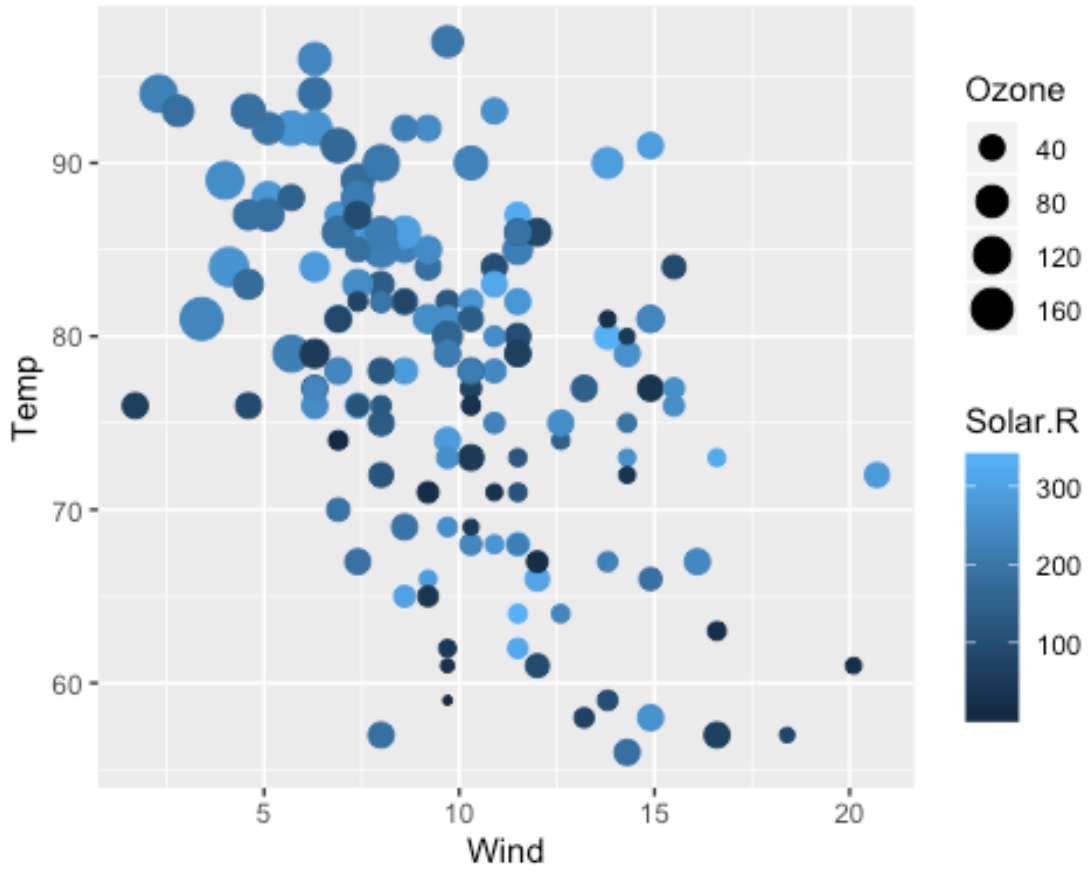

```
# Create a new data-frame picking Date, Ozone, Temp, WindScaled, Solar.R from
air
new_air <- data.frame(air$Date, air$Ozone, air$Solar.R, air$Temp,
air$WindScaled)

# Using the melt function from reshape2 shrink/collapse table
melt_new_air <- melt(new_air, id.vars=1)
g <- ggplot(melt_new_air, aes(x = air.Date, y = value, col = variable, group
= 1))
g <- g + geom_line()
g
```




```
# Step-4: Create a heap-map using the geom_tile function
g <- ggplot(melt_new_air, aes(air.Date, variable))
g <- g + geom_tile(aes(fill = value) , color = "white" ) +
scale_fill_gradient(low = "white", high = "steelblue")
g
```




```
# Step-5: Create a scatter plot
g <- ggplot(air, aes(x=Wind, y=Temp))
g <- g + geom_point(aes(size=Ozone, color=Solar.R))
g
```


Temp
90
80
70
60
5
10
15
20
Wind
Ozone
40
80
120
160
Solar.R
300
200
100