Josh Rhoades
Biostatistics
HW1

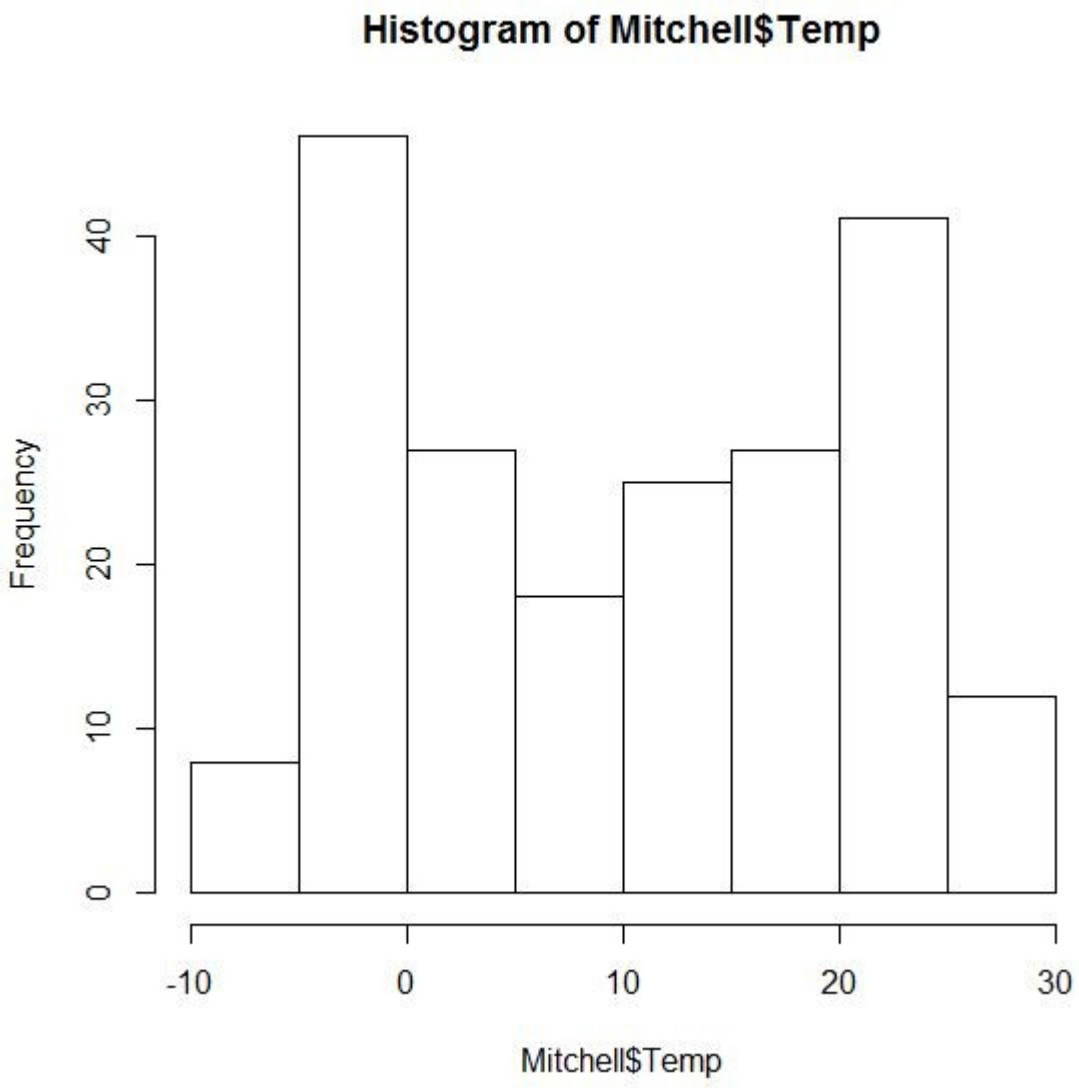
1. (a) The Experimental Unit for this study is one individual. The factors are the vulnerability to disease and the dosage amount. The response is the severity of the disease after exposure.
   (b) If testing for a diffrence based on gender the experiment can still be done, but more replicates would be needed. Doubling the total number of replicates would allow the same number of replicates per gender as was in the total original experiment. By keeping sampling size consistent, the experimental design is acceptible.
   (c) This would be a terrible experiment to perform in a typical chicken farm based on the inability to keep treatments isolated. Chicken houses are large and have high airflow which could act to vector the disease. Also, infection could spread to chickens outside of the experiment, which could be extremely costly.

2. Drug <-read.table('CholDrug.txt', header=T)

```
table(drug$gender,drug$dose)

   H L M
 F 5 3 2
 M 5 7 8

tapply(drug$Y,drug[,-1],mean)
     dose
gender      H        L        M
    F 142.70 144.3000 144.1000
    M 144.14 148.3714 143.3375
```

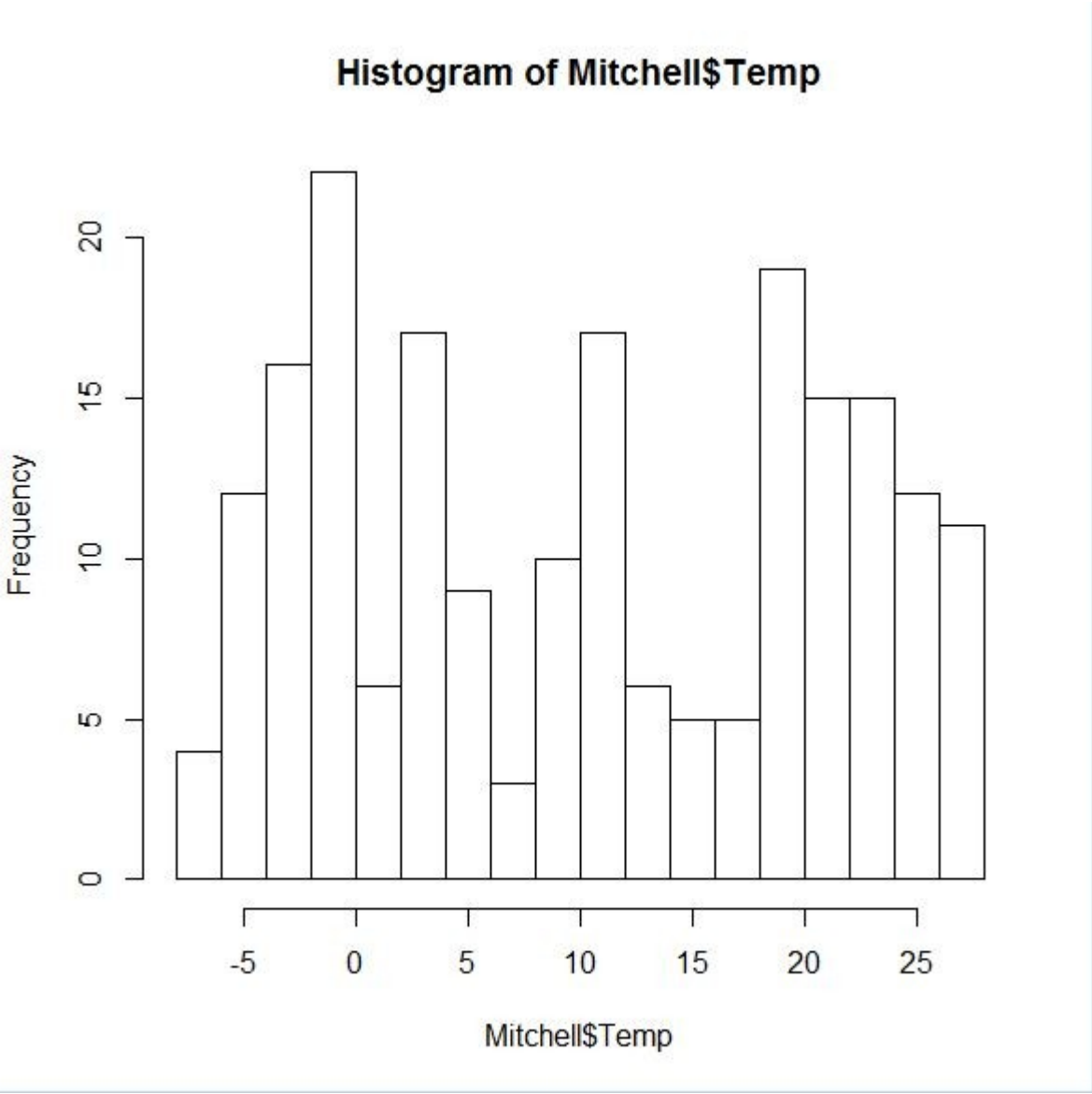
3.

```
Mitchell<-read.table('Mitchell.txt', header=T)

hist(Mitchell$Temp)
```



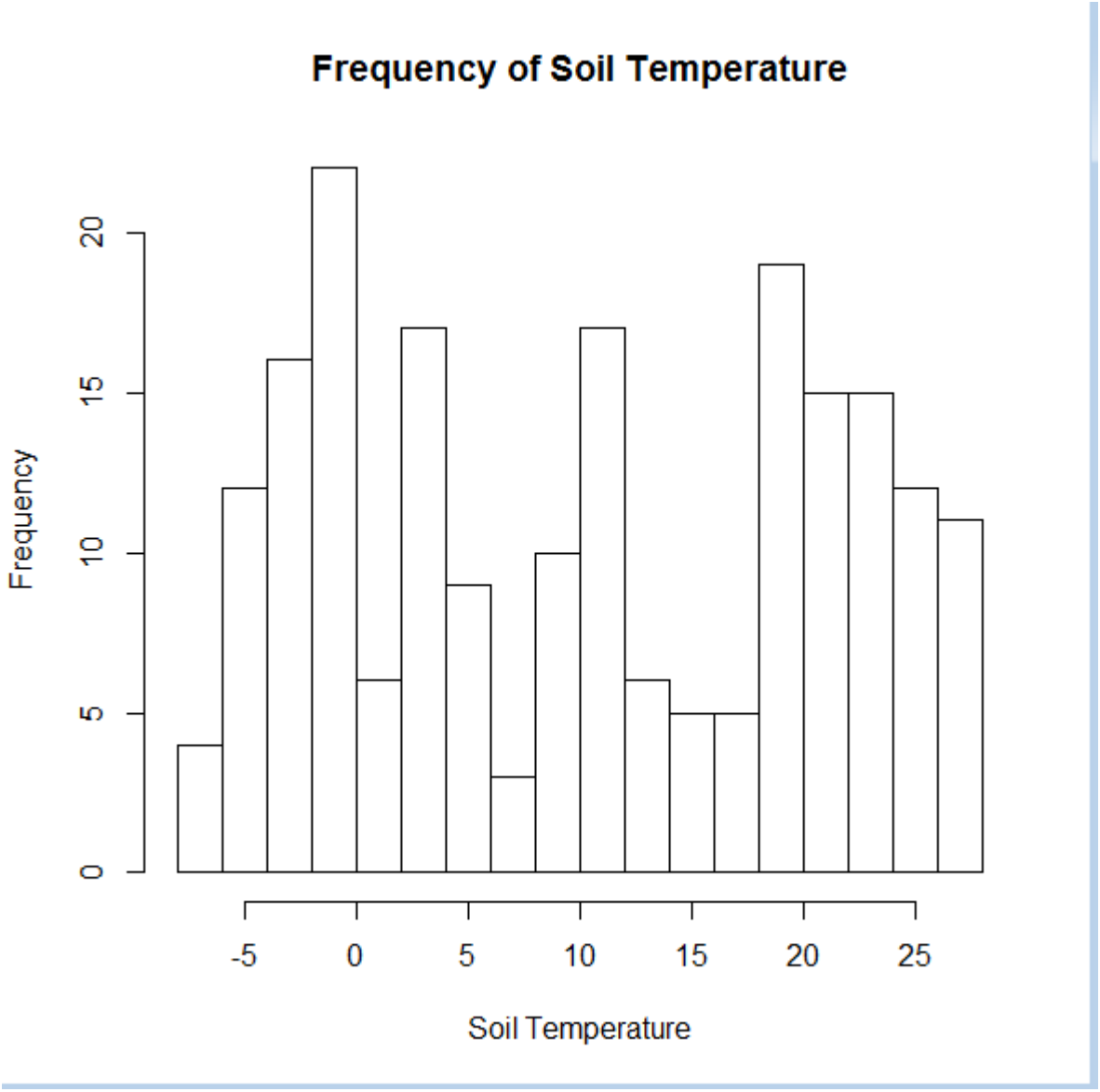
Default Histogram

hist(Mitchell$Temp, breaks=15)



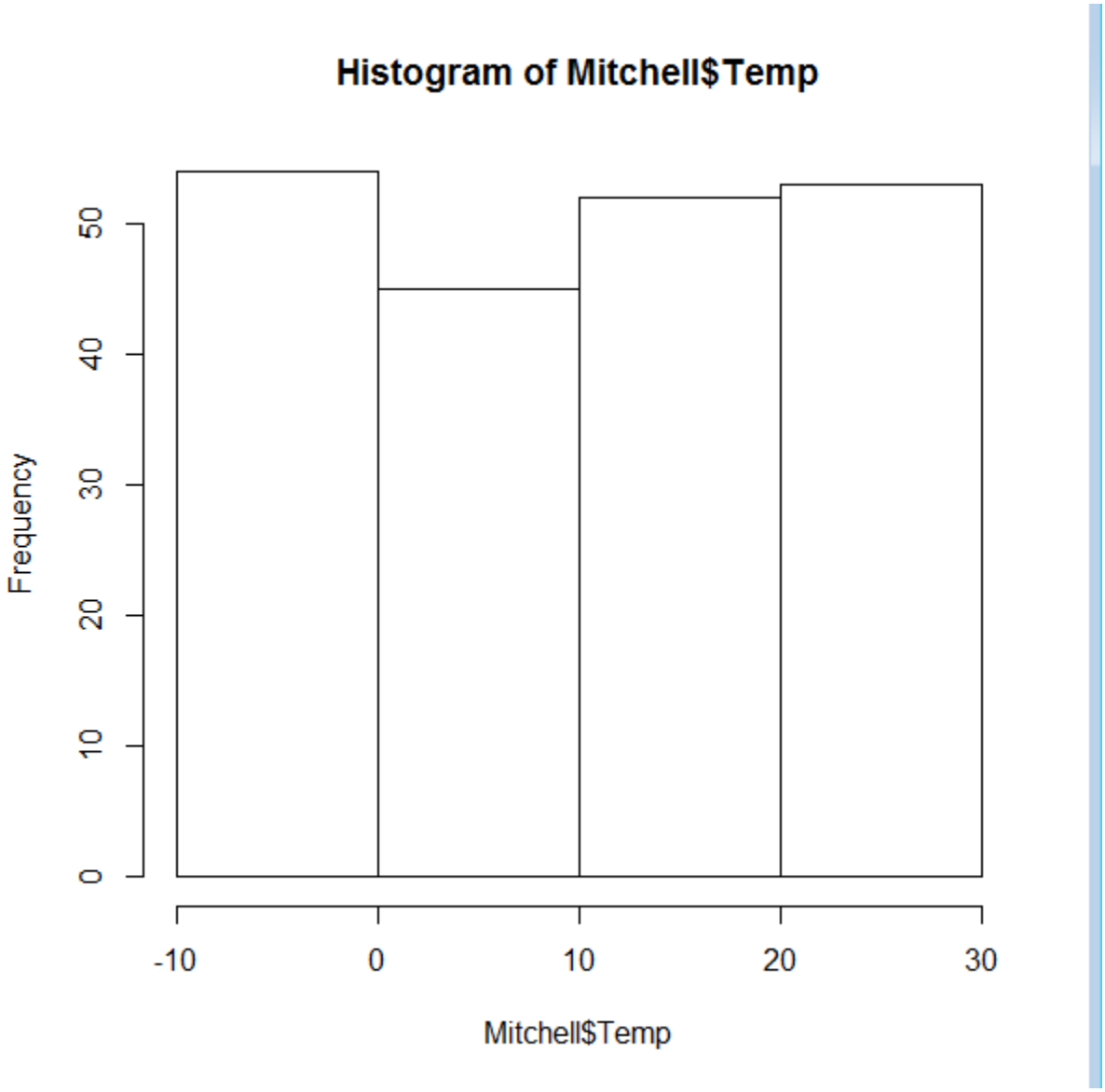
15 breaks

hist(Mitchell$Temp, breaks=15,xlab='Soil Temperature',main='Frequency of Soil Temperature')

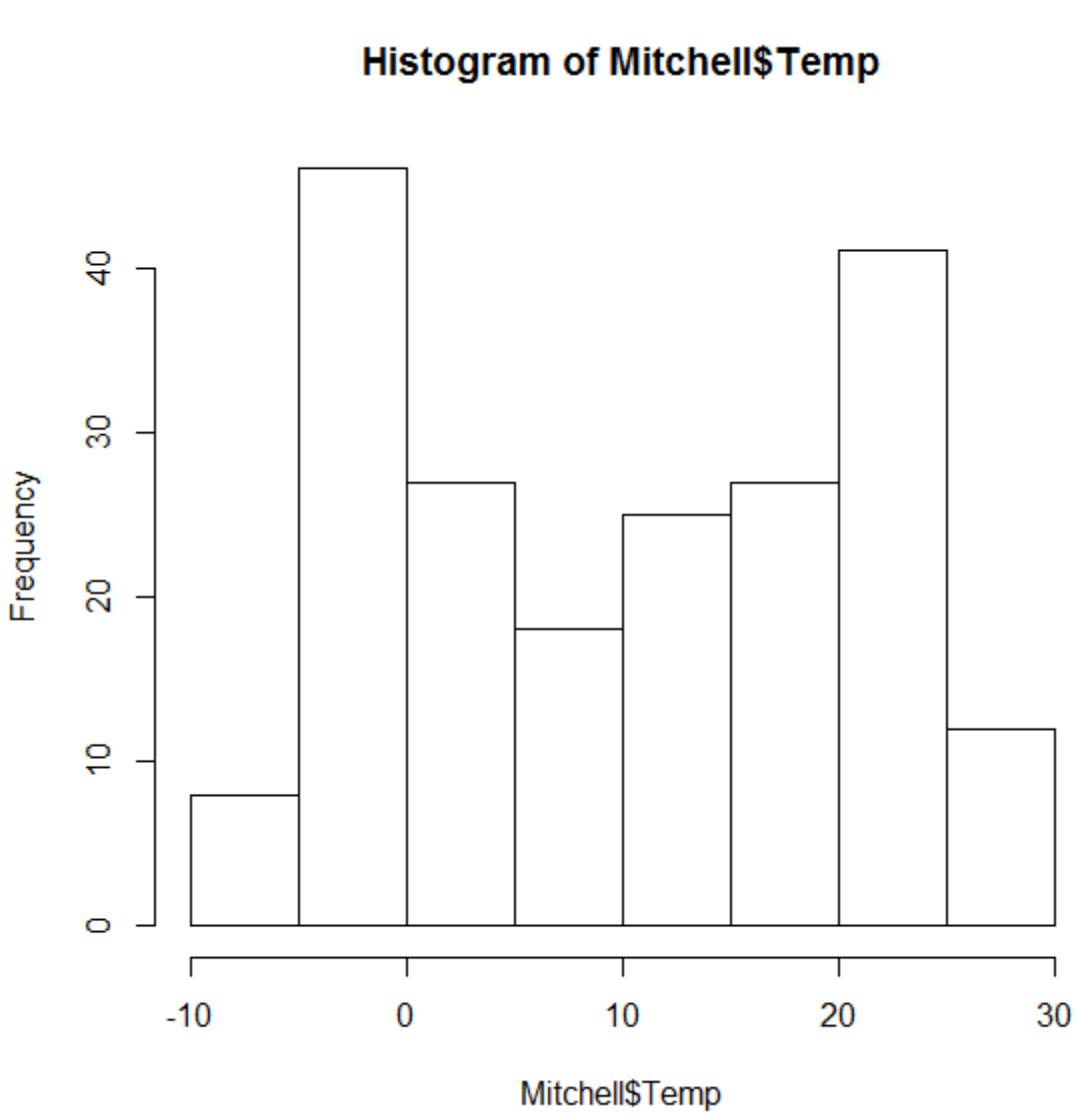


15 breaks, graph with Titles added

hist(Mitchell$Temp, breaks=c(-10,0,10,20,30))

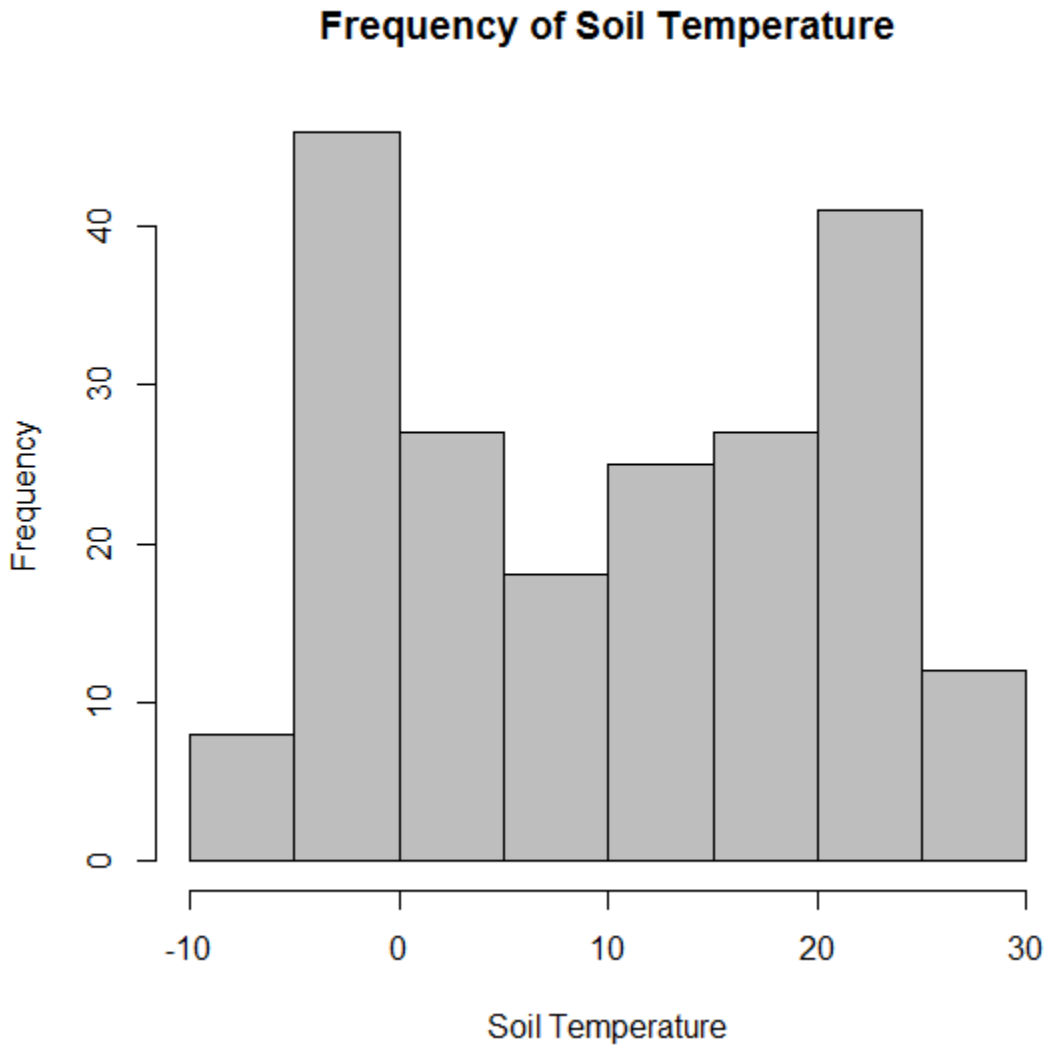


breaks at -10,0,10,20,30C



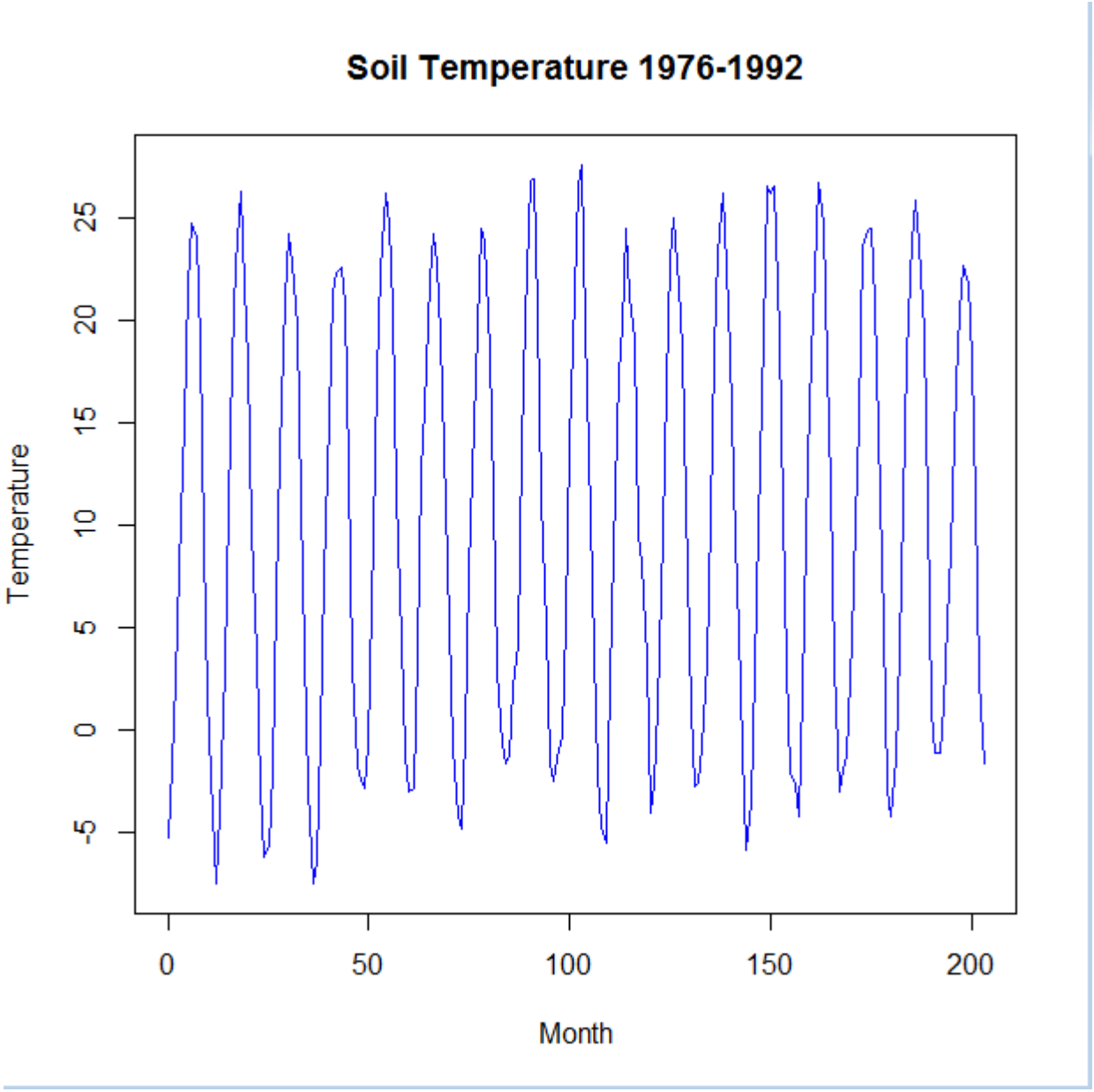
hist(Mitchell$Temp, breaks=c(-10,-5,0,5,10,15,20,,25,30))Breaks at -10,-5,0,5,10,15,20,25,30C

```
hist(Mitchell$Temp, col = "grey", breaks=c(-10,-5,0,5,10,15,20,25,30) ,xlab='Soil
Temperature',main='Frequency of Soil Temperature')
```



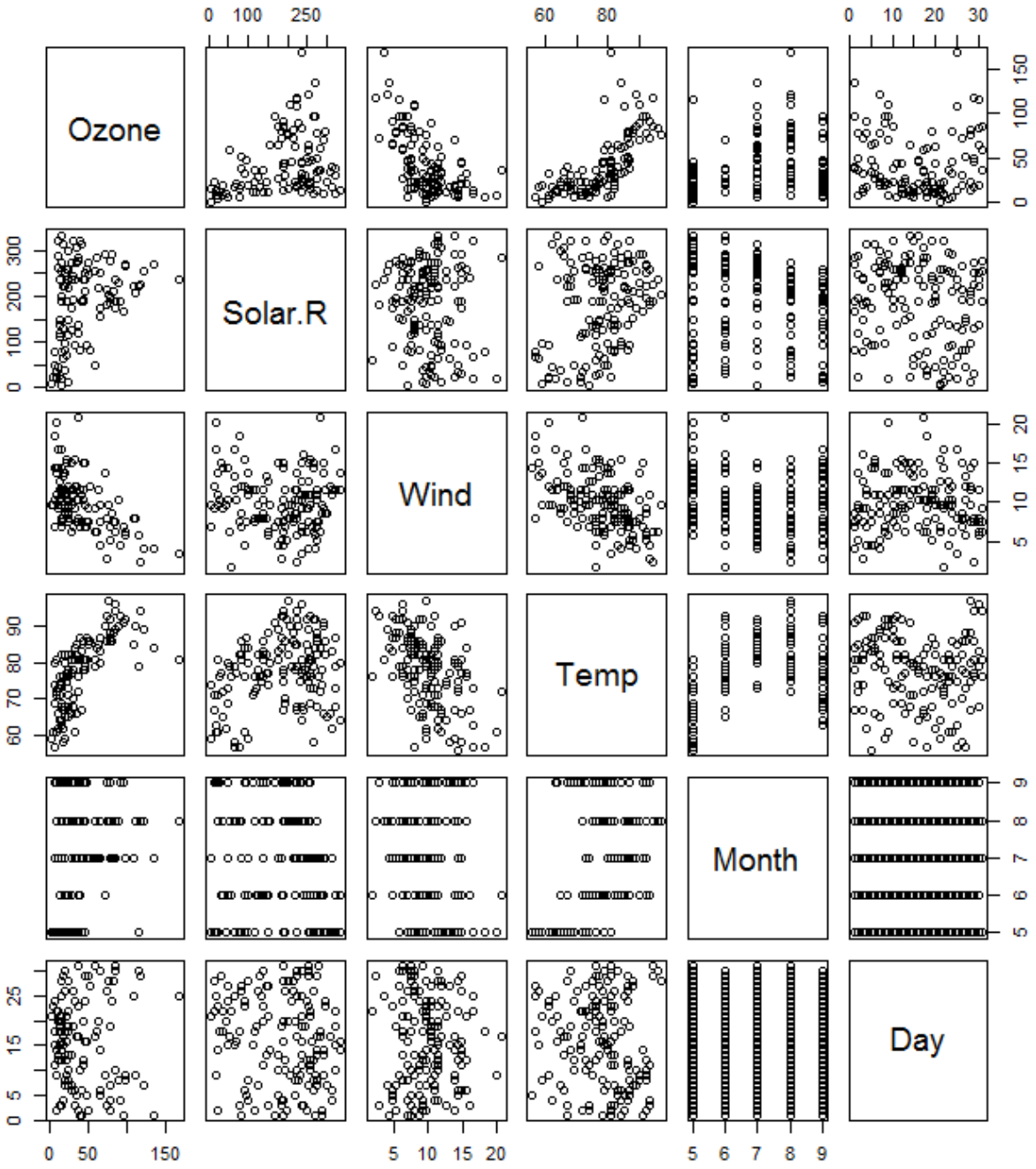
By varying the size of the bins you can get a better idea of the data. Instead of grouping the data into large ranges, smaller ranges are used and are more descriptive.

plot(Mitchell, type='l', col='blue', xlab='Month',ylab='Temperature',main='Soil Temperature 1976-1992')



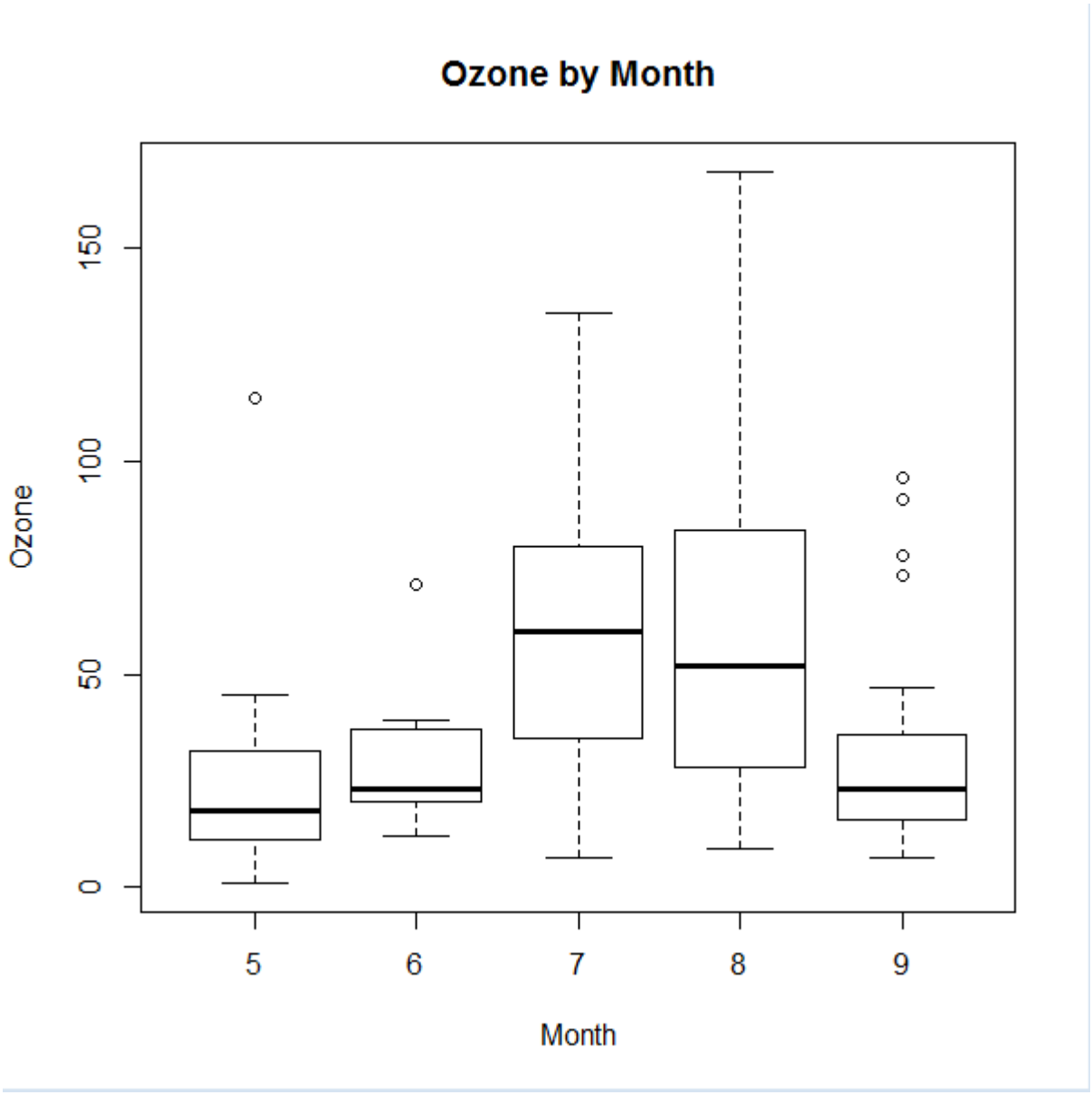
The graph shows the seasonal trends in temperature. The peaks represent summer, the bottom represents the winter and the appropriate spaces in between represents spring and fall.

4. (a) ny<-read.csv('ozone.csv',header=TRUE)

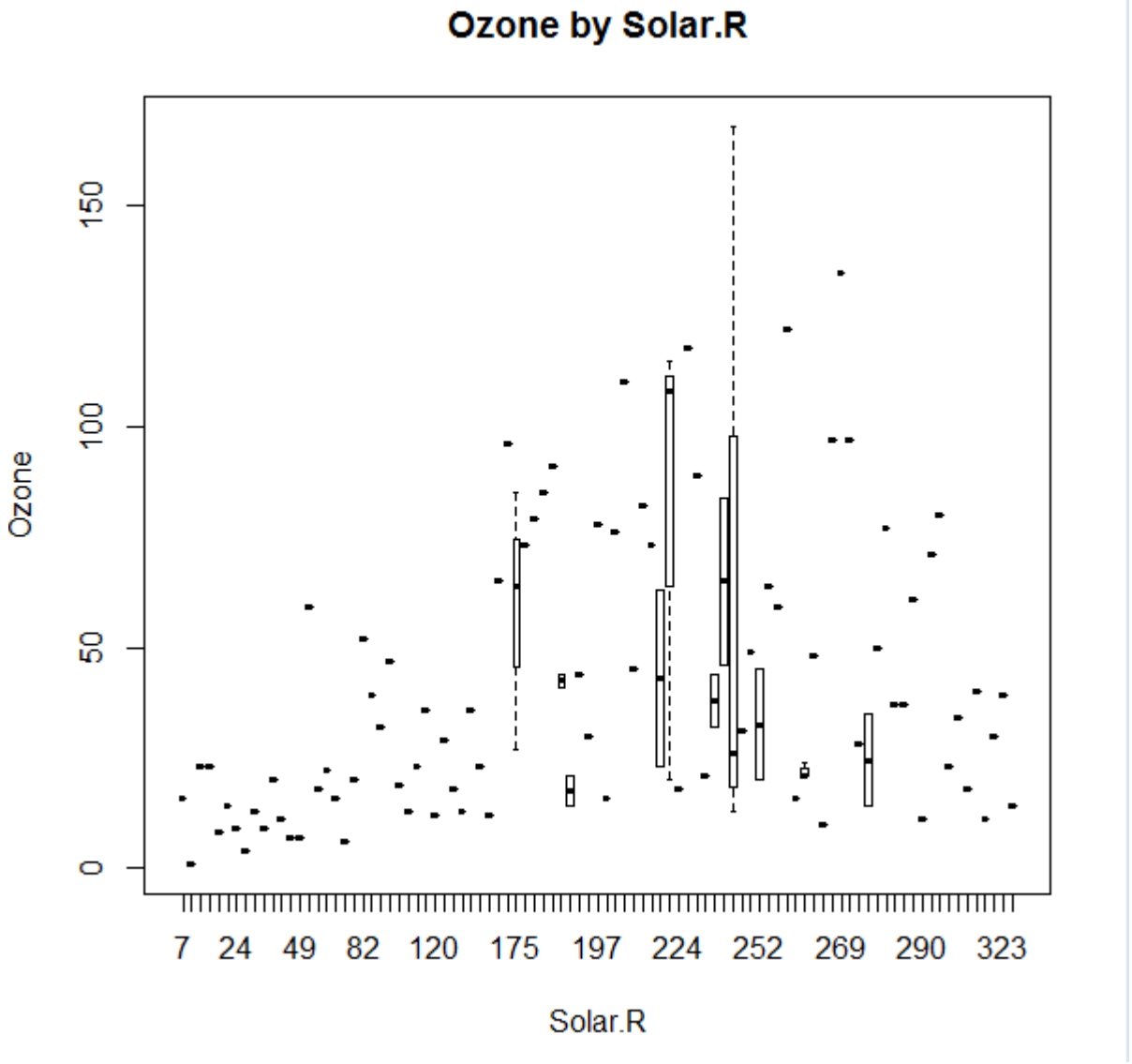

Ozone
Solar.R
Wind
Temp
Month
Day
0 100 250
60 80
0 10 20 30
0 50 100 150
0 100 200 300
5 10 15 20
60 70 80 90
5 6 7 8 9
0 5 15 25
0 50 150
5 10 15 20
5 6 7 8 9

(b)

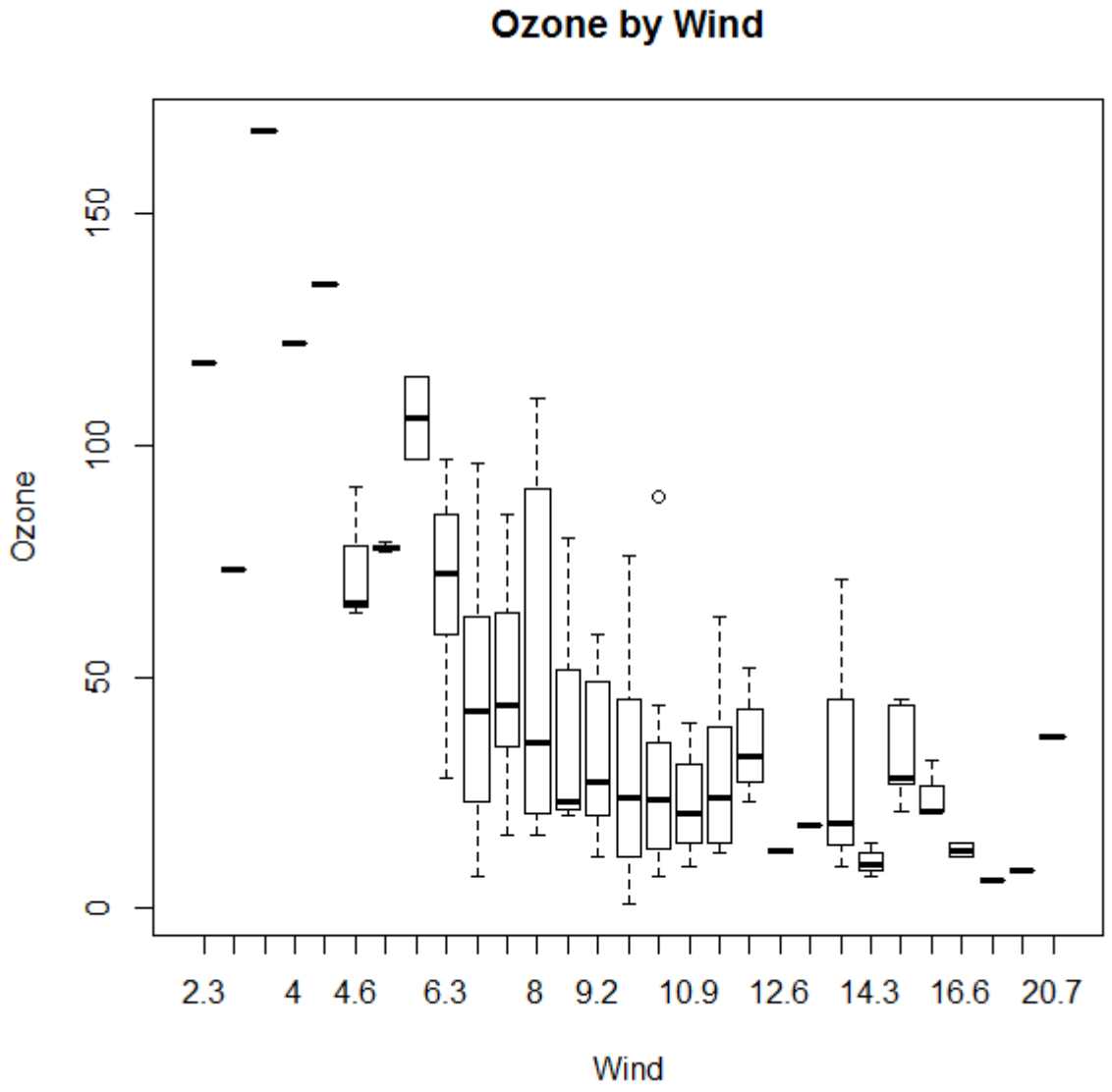
boxplot(Ozone~Month,data=ny, xlab='Month', ylab='Ozone',main='Ozone by Month')



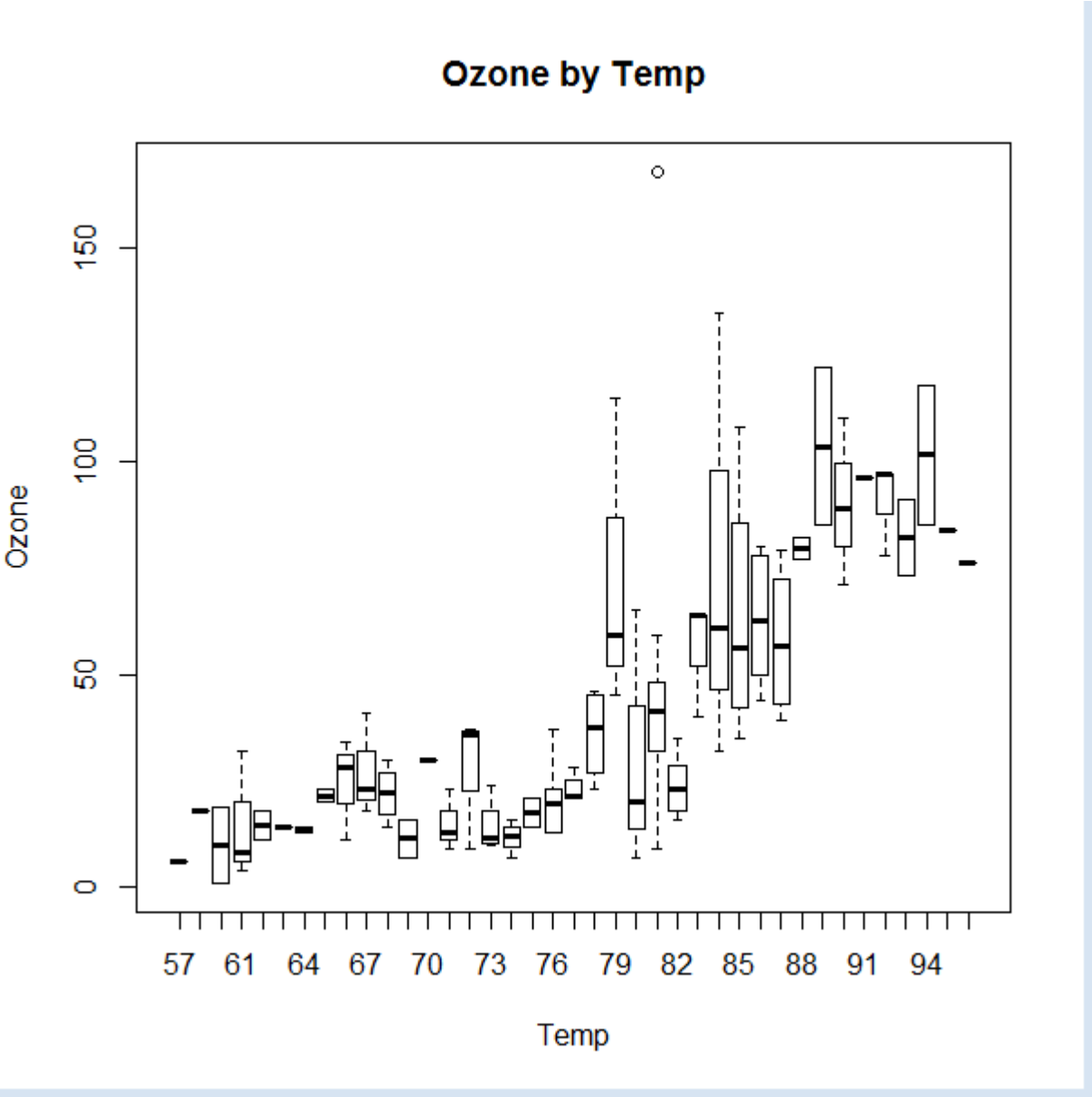
boxplot(Ozone~Solar.R,data=ny, xlab='Solar.R', ylab='Ozone',main='Ozone by Solar.R')



boxplot(Ozone~Wind,data=ny, xlab='Wind', ylab='Ozone',main='Ozone by Wind')


Ozone by Wind
Ozone
0
50
100
150
2.3
4
4.6
6.3
8
9.2
10.9
12.6
14.3
16.6
20.7
Wind

boxplot(Ozone~Temp,data=ny, xlab='Temp', ylab='Ozone',main='Ozone by Temp')

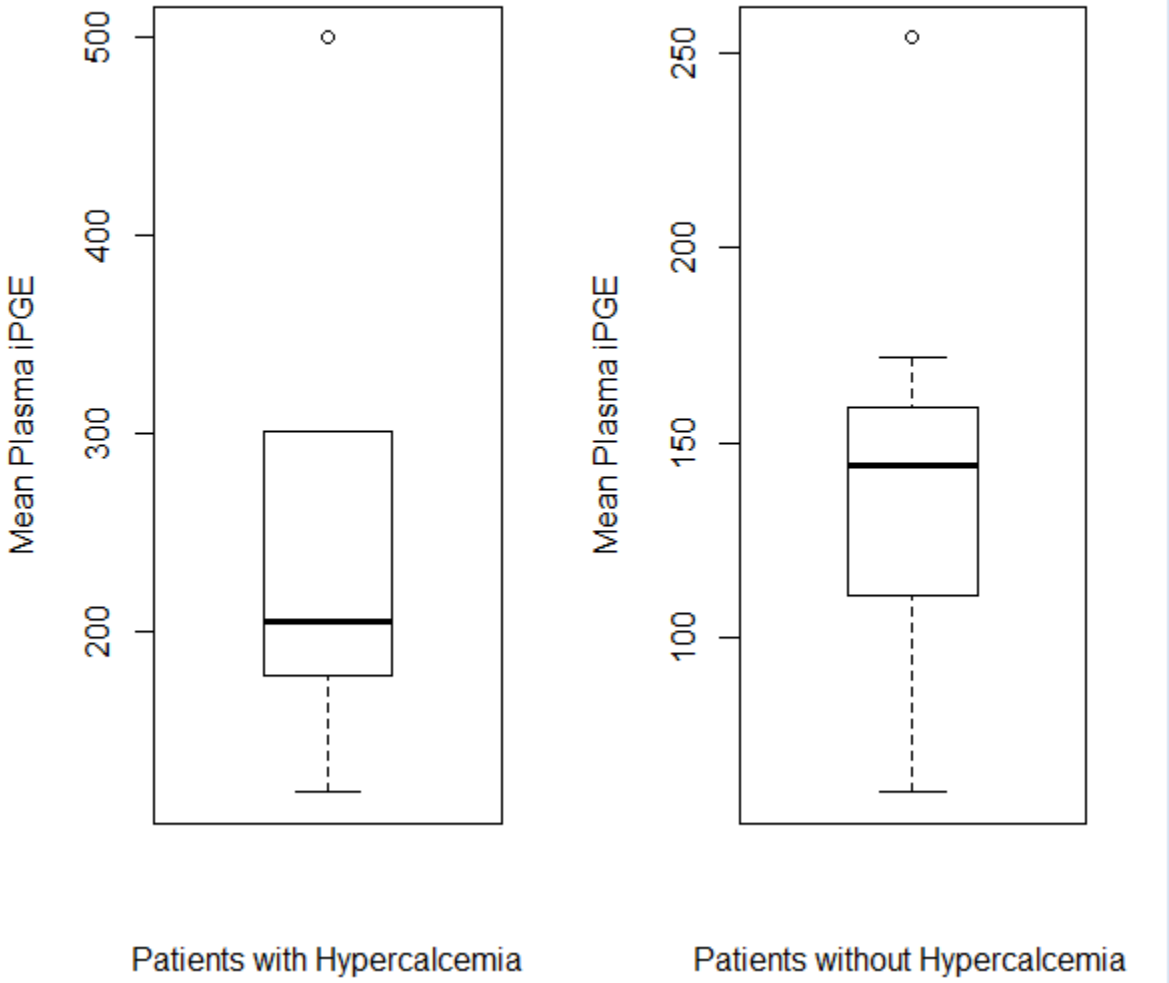


It is interesting that the wind and ozone have a inverse relationship. Ozone tends to peak when Solar Radiation is between 170 and 260, this is interesting because I would expect Solar Radiation to have a linear relationship, not bell shaped. Ozone is constant across most months, but has large variation in months 7 and 8 which correspond with the summer months in North America when the Solar radiation and Temperature are highest.

5. (a) The intervals are 10 lbs with the ends of the distribution open, so less that 50 lbs and greater than 100 lbs.
   (b) summary(data)
   summary(data)
   Pull.Strength Cases.Observed
   <100:1 Min. : 10.00
   <50 :1 1st Qu.: 23.00
   <60 :1 Median : 42.00
   <70 :1 Mean : 74.14
   <80 :1 3rd Qu.:126.50
   <90 :1 Max. :168.00
   >100:1
   The median is 42lbs, the 25th percentile is 23lbs and the 75th percentile is 126.5lbs.
   (d) The mean is 74.14lbs, the standard deviation is 64.59lbs.
   (e) The empirical rule states that 68% of observations should lie within one standard deviation of the mean, given a normal distribution.
6. withH<-read.csv('withH.csv',header=TRUE)
   withoutH<-read.csv('withoutH.csv',header=TRUE)
   summary(withH)
   Mean.Plasma Mean.Serum.Calcium
   Min. :118.0 Min. :11.20
   1st Qu.:178.5 1st Qu.:11.53
   Median :204.5 Median :11.90
   Mean :259.6 Mean :12.09
   3rd Qu.:293.8 3rd Qu.:12.43
   Max. :500.0 Max. :13.40
   summary(withoutH)
   Mean.Plasma Mean.Serum.Calcium
   Min. : 60.0 Min. : 6.40
   1st Qu.:110.5 1st Qu.: 9.25
   Median :144.0 Median :10.10
   Mean :139.5 Mean :10.25
   3rd Qu.:159.0 3rd Qu.:10.40
   Max. :254.0 Max. :18.00
   mean for patients with Hypercalcemia is 259.6
   mean for patients without Hypercalcemia 144.0
   var(withH)
   Mean.Plasma Mean.Serum.Calcium
   Mean.Plasma 19162.93333 16.3733333
   Mean.Serum.Calcium 16.37333 0.5965556
   > sqrt(19162.93333)
   [1] 138.4302
   The standard deviation for mean plasma of patients with Hypercalcemia is 138.43.

var(withoutH)
Mean.Plasma Mean.Serum.Calcium
Mean.Plasma 2614.67273 -72.612727
Mean.Serum.Calcium -72.61273 8.006727
> sqrt(2614.67273)
[1] 51.13387

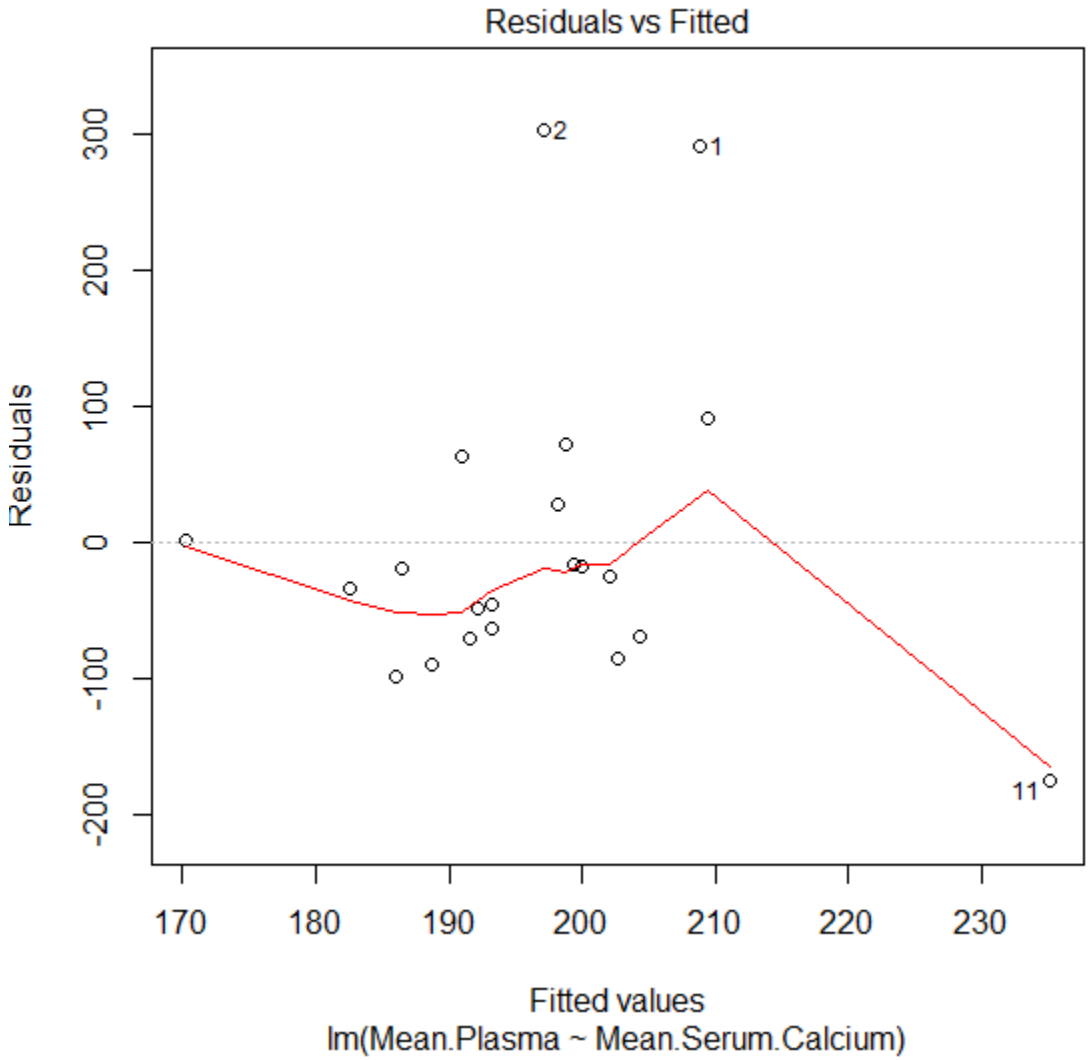
The standard deviation for mean plasma of patients without Hypercalcemia is 51.13.




```
attach(mtcars)
par(mfrow=c(1,2))
boxplot(withH[,1], xlab='Patients with Hypercalcemia',ylab='Mean Plasma iPGE')
boxplot(withoutH[,1], xlab='Patients without Hypercalcemia',ylab='Mean Plasma iPGE')
```

The two box plots do show large differences in Mean Plasma between the two groups, patients with hypercalcemia have a much higher level than patients without. Also the patients with hypercalcemia have much higher variance than the group without.

```
lm(data3)

Call:
lm(formula = data3)

Coefficients:
      (Intercept)  Mean.Serum.Calcium
          134.588               5.583

fitted.model<-lm(Mean.Plasma~Mean.Serum.Calcium, data=data3)
plot(fitted.model)
```



No strong relationship between Mean Plasma and Mean Serum Calcium is seen when a regression line is fit to the data. There is a strong clustering of values, which does suggest correlation between the two variables.