Group Work on Graphics

Please write the names and NetIDs of your group members:

- Name: Ying Mufang ; NetID: mying4
- Name: Cai Naiqing ; NetID: ncai5
- Name: Meng Yuhan ; NetID: meng46
- Name: Pan Hongwei ; NetID: hpan55
- Name: Zhang Xueqian ; NetID: xzhang2278
- Name: Li Zihao ; NetID: zli873
- Name: Wei Haoxiang ; NetID: hwei64
- Name: Zhai Yibo ; NetID: yzhai28

Revise this graphics.Rmd (graphics.Rmd) file to produce the graphs and answer the questions below. Include reasonable labels (titles, axis labels, legends, etc.) with each graph. Please do not do statistical analysis of these questions–we'll get to that soon. Today's exercise just uses graphs. Please use **boldface** (by enclosing text in `** ... **`) when writing your answers so that we can find them easily.

At the end of class, one person from each group should submit a completed copy of this file. (Please don't submit multiple copies.)

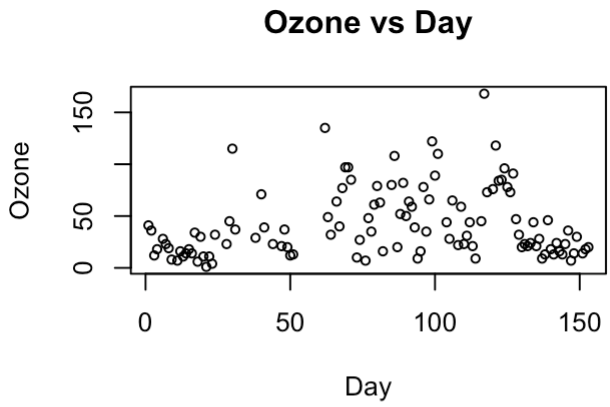
We'll grade this exercise by opening your "graphics.Rmd" file, clicking "Knit HTML", and reading the output. We'll assign points as follows:

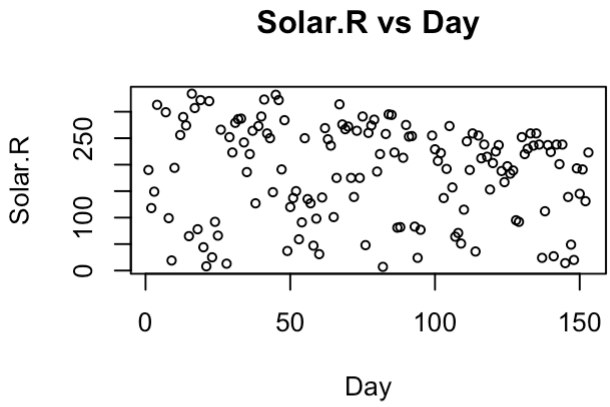
- No submission: 0/5
- Poor work: 3/5
- Good work: 5/5 (even if incomplete–there's more to do here than can be done in 75 minutes)

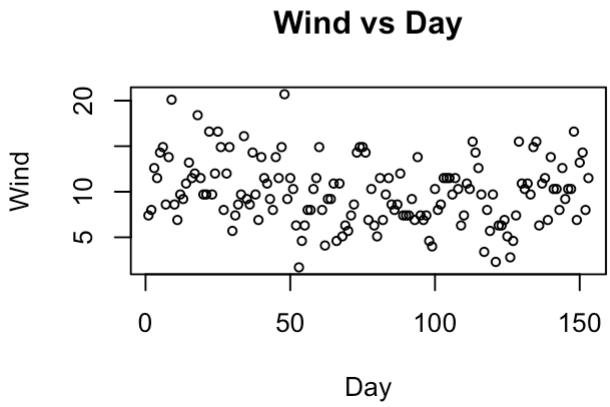
Air quality

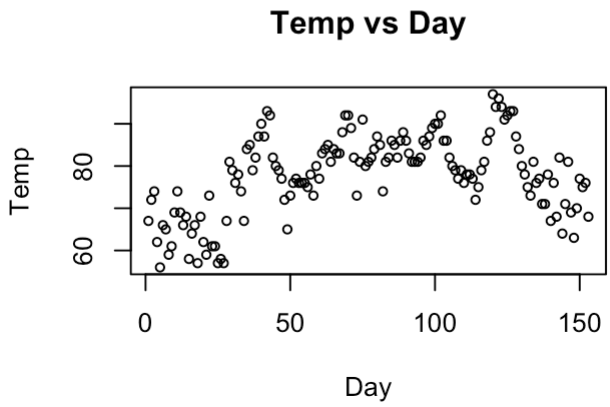
Consider the built-in data frame, `airquality`. Make a graph with four panels (two rows and two columns) to show each air quality variable against the day number (from 1 to 153) in the study. Which of the four variables seems to be affected most by the changing seasons?

```
par(mfrow=c(2,2))
attach(airquality)
Day=c(1:nrow(airquality))
plot(Day,Ozone,main = "Ozone vs Day",cex = 0.7)
plot(Day,Solar.R,main = "Solar.R vs Day",cex = 0.7)
plot(Day,Wind,main = "Wind vs Day",cex = 0.7)
plot(Day,Temp,main="Temp vs Day",cex = 0.7)
```







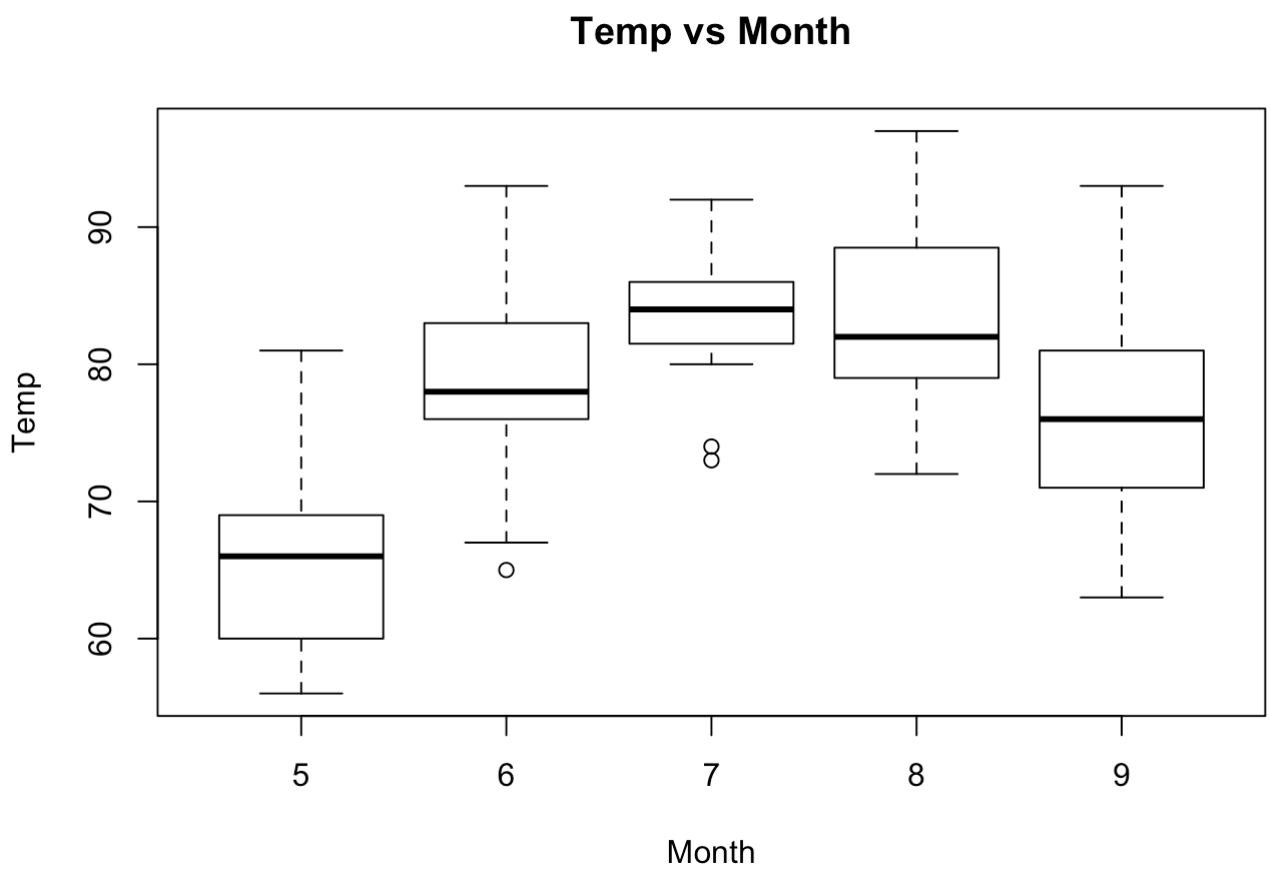



```
detach(airquality)
```

Ozone seems to be affected most by the changing seasons.

Make one graph of the temperatures grouped by month. According to your graph, which month was the warmest? Which month had the most uniform temperature?

```
boxplot(Temp ~ Month,data  = airquality,xlab = "Month", ylab = "Temp",main = "Temp vs
 Month")
```



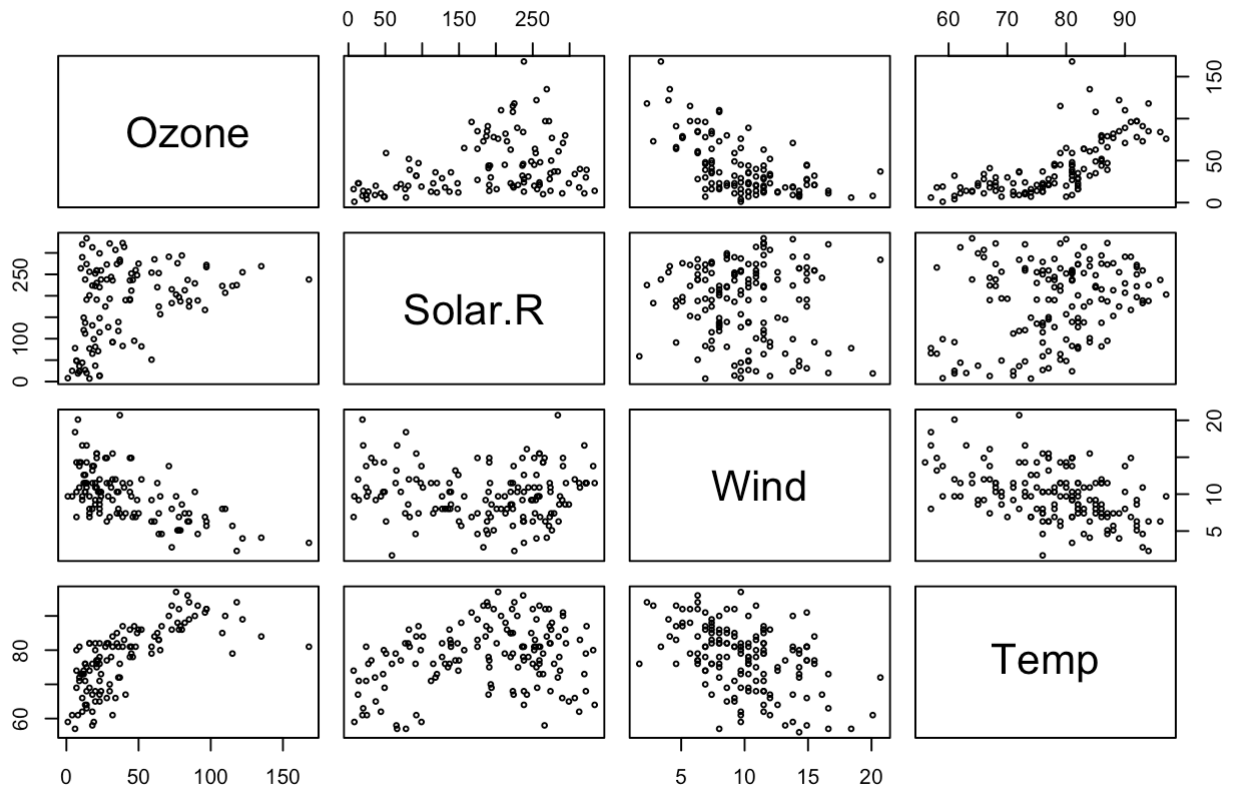
The 7th was the warmest month and the 9th had the most uniform temperature.

Does Ozone tend to increase, decrease, or stay the same as Solar.R increases? Does Ozone tend to increase, decrease, or stay the same as Wind increases? Does Ozone tend to increase, decrease, or stay the same as Temp increases? Make one figure to support your answer to all three questions.

```
attach(airquality)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     Day
```

```
pairs(airquality[,c(1,2,3,4)],cex = 0.5,main="Air Quality Variables Relationship")
```




```
detach(airquality)
```

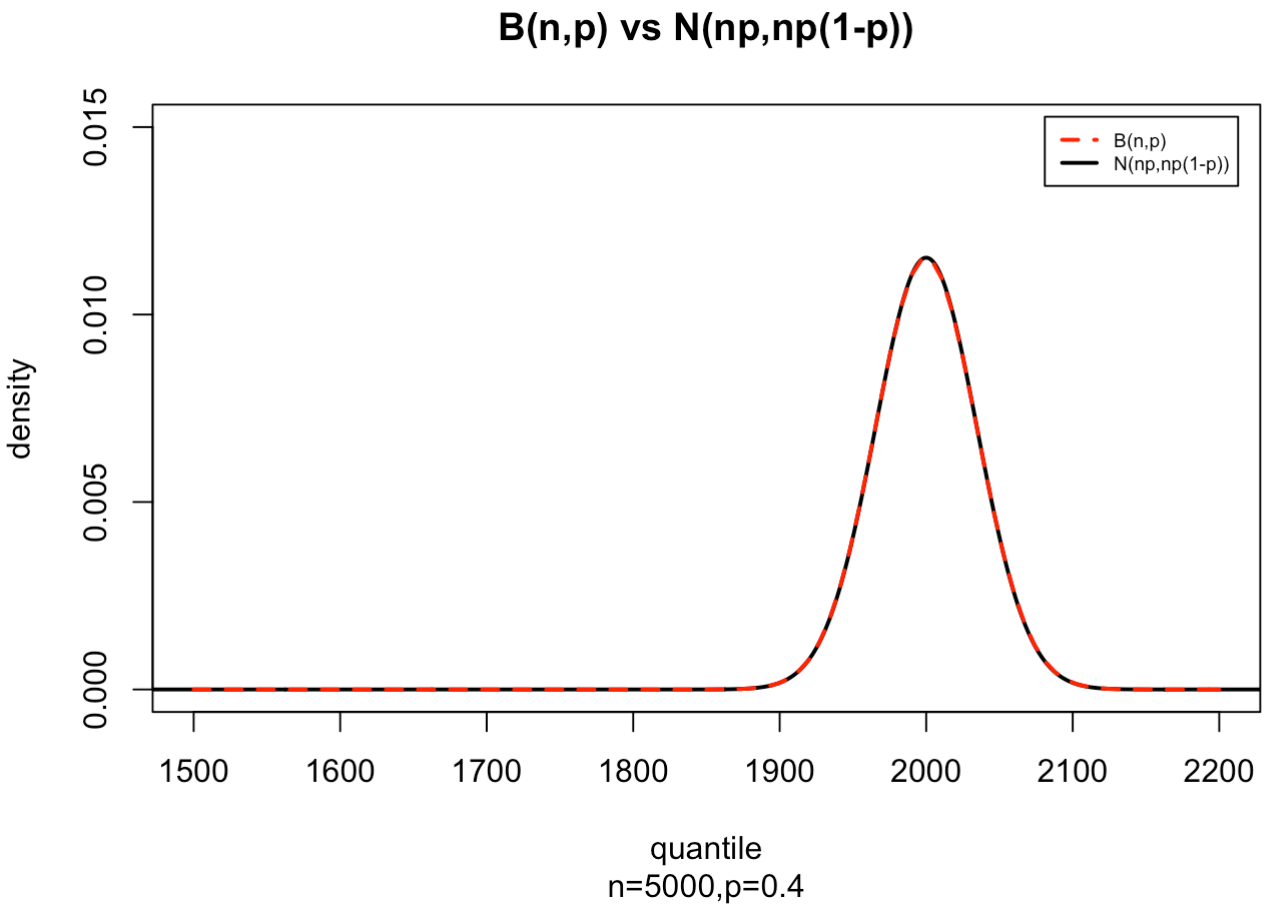
Ozone tends to decrease as Wind increases and tends to increase as Temp increases. As for Solar.R, the range of Ozone is quite large when Solar.R reaches around 250.

Normal approximation to binomial

The binomial distribution of the number, x, of successes in n independent trials, each having probability p of success, is approximated by the normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. That is, Bin$(n,p) \approx N(np, np(1-p))$. Make a graph showing the Bin(n,p) probability mass function and the $N(np, np(1-p))$ probability density curve. Choose values of n and p for which the approximation looks good.

Choose n=5000, p=0.4

```
par(mfrow=c(1,1))
n=5000
p=0.4
x=seq(0,3000,length.out = 3000)
mu=n*p
sd=sqrt(n*p*(1-p))
density.norm=dnorm(x,mu,sd)
plot(x,density.norm,type ="l",xlim=c(1500,2200),ylim = c(0,0.015),main="B(n,p) vs N(n
p,np(1-p))",xlab ="quantile",ylab="density",lty=1,lwd=2,sub="n=5000,p=0.4")
curve(dbinom(x,n,p),from=1500,to=2200,add=TRUE,col="red",lty=2,lwd=2)
legend("topright",inset = 0.02,c("B(n,p)","N(np,np(1-p))"),col = c("red","black"),lty
 = c(2,1),lwd=c(2,2),cex = 0.6)
```



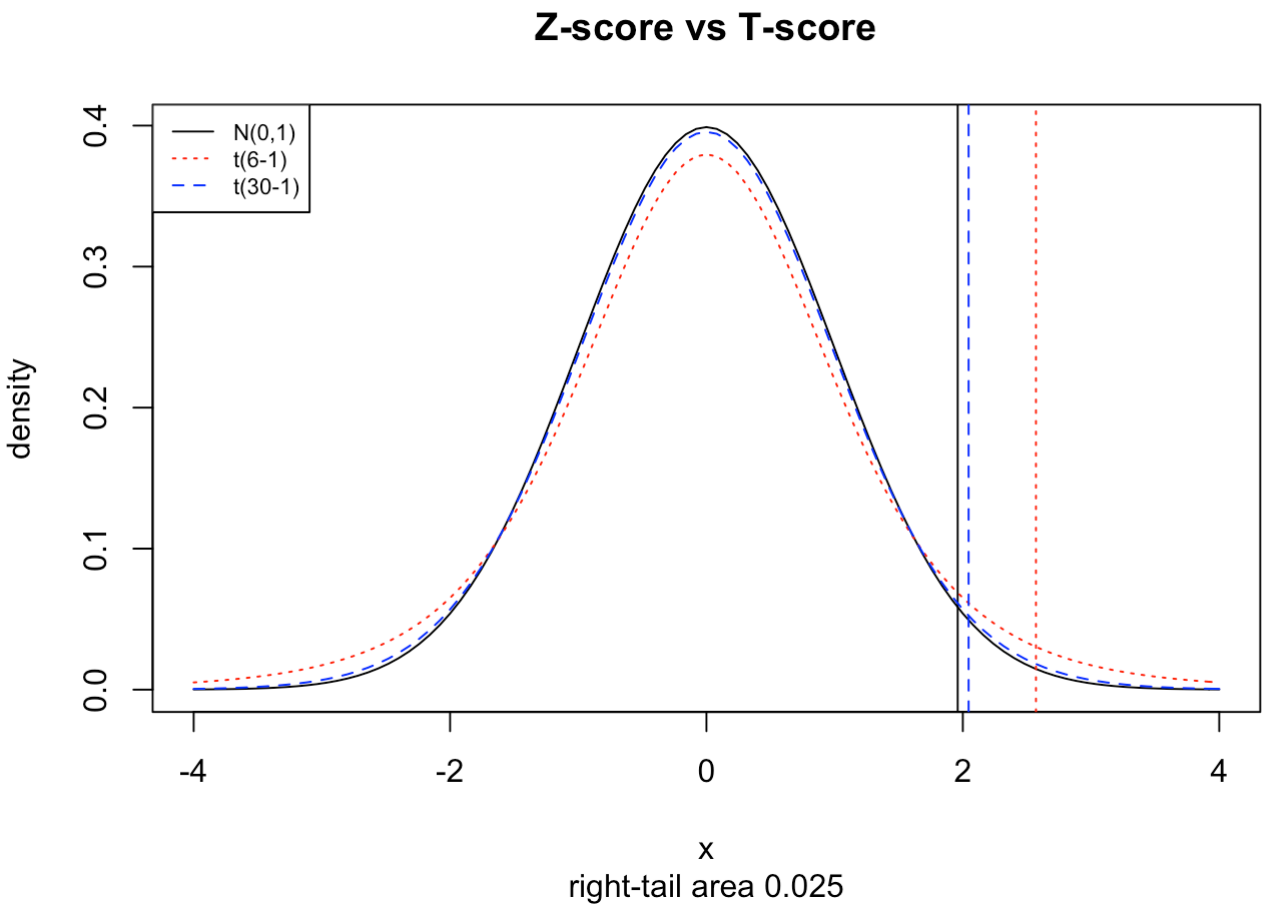
Z-score vs. T-score

A Z-score calculated as $Z = \frac{X-\mu}{\sigma/\sqrt{n}}$ has the normal distribution with mean 0 and standard deviation 1: $Z \sim N(0,1)$. A T-score calculated as $T = \frac{X-\mu}{s/\sqrt{n}}$ has the Student's t distribution with $n-1$ degrees of freedom: $T \sim t_{n-1}$. The t_{n-1} density curve is shorter with thicker tails than the $N(0,1)$ density because s varies more than σ (a constant). However, the former density approaches the latter as n increases. Make a graph of three probability density curves:

- $N(0,1)$ (a solid line)
- t_{6-1} (a dotted line)
- t_{30-1} (a dashed line)

For each curve, make a vertical line (of the same type as the curve's line type) from the x-axis to the curve at the point x that cuts off off right-tail area 0.025.

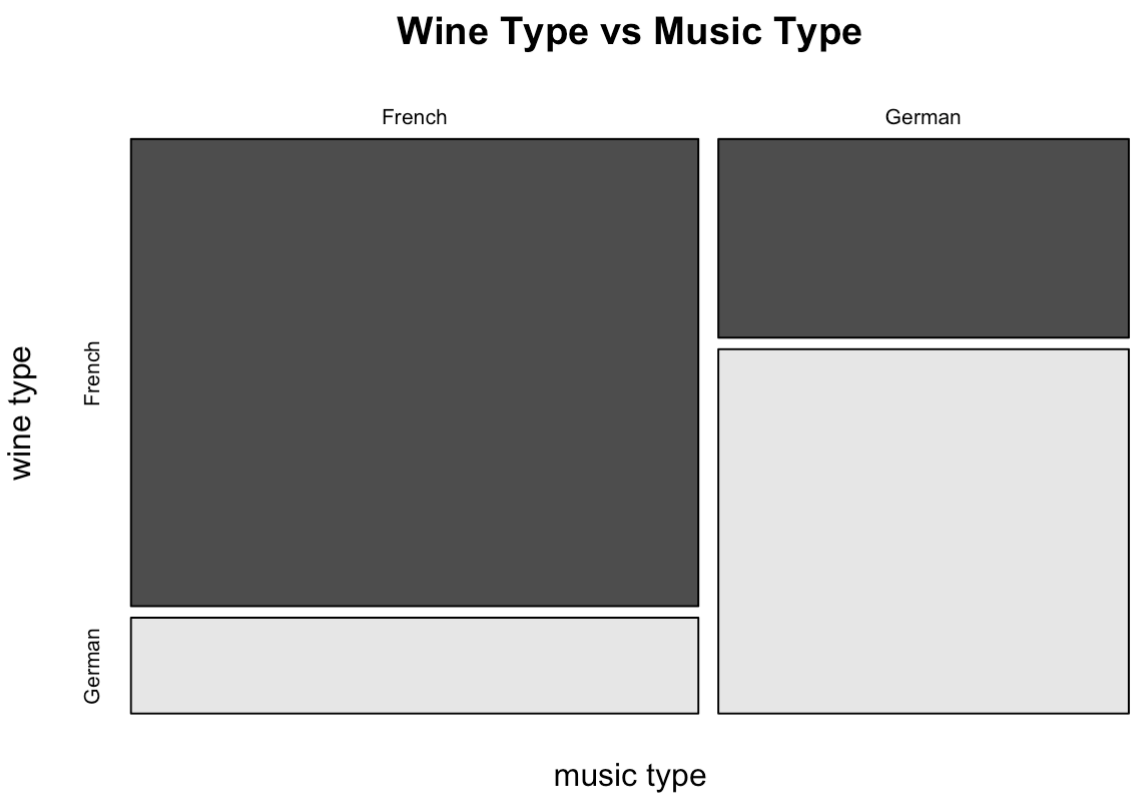
```
curve(dnorm(x,0,1),-4,4,ylab="density",main="Z-score vs T-score",sub="right-tail area
 0.025")
abline(v=qnorm(0.975,0,1))
curve(dt(x,5),-4,4,add=T,col="red",lty=3)
abline(v=qt(0.975,5),col="red",lty=3)
curve(dt(x,29),-4,4,add=T,col="blue",lty=2)
abline(v=qt(0.975,29),col="blue",lty=2)
legend("topleft",legend=c("N(0,1)","t(6-1)","t(30-1)"),col=c("black","red","blue"),lt
y=c(1,3,2),cex = 0.7)
```



Influence of music on wine purchase

The file wine.csv (http://www.stat.wisc.edu/~jgillett/327-1/graphics/wine.csv) contains data on wine purchased from a liquor store. Each row corresponds to a bottle of wine purchased. The first column indicates which type of music was playing in the store during the purchase. The second column indicates which type of wine was purchased. Make a graph that gives evidence about the question of whether type of music and type of wine are independent. Do you think they are independent?

```
wine = read.csv('wine.csv')
counts = table(wine)
mosaicplot(counts, xlab='music type', ylab='wine type',main="Wine Type vs Music Type"
,color = TRUE)
```



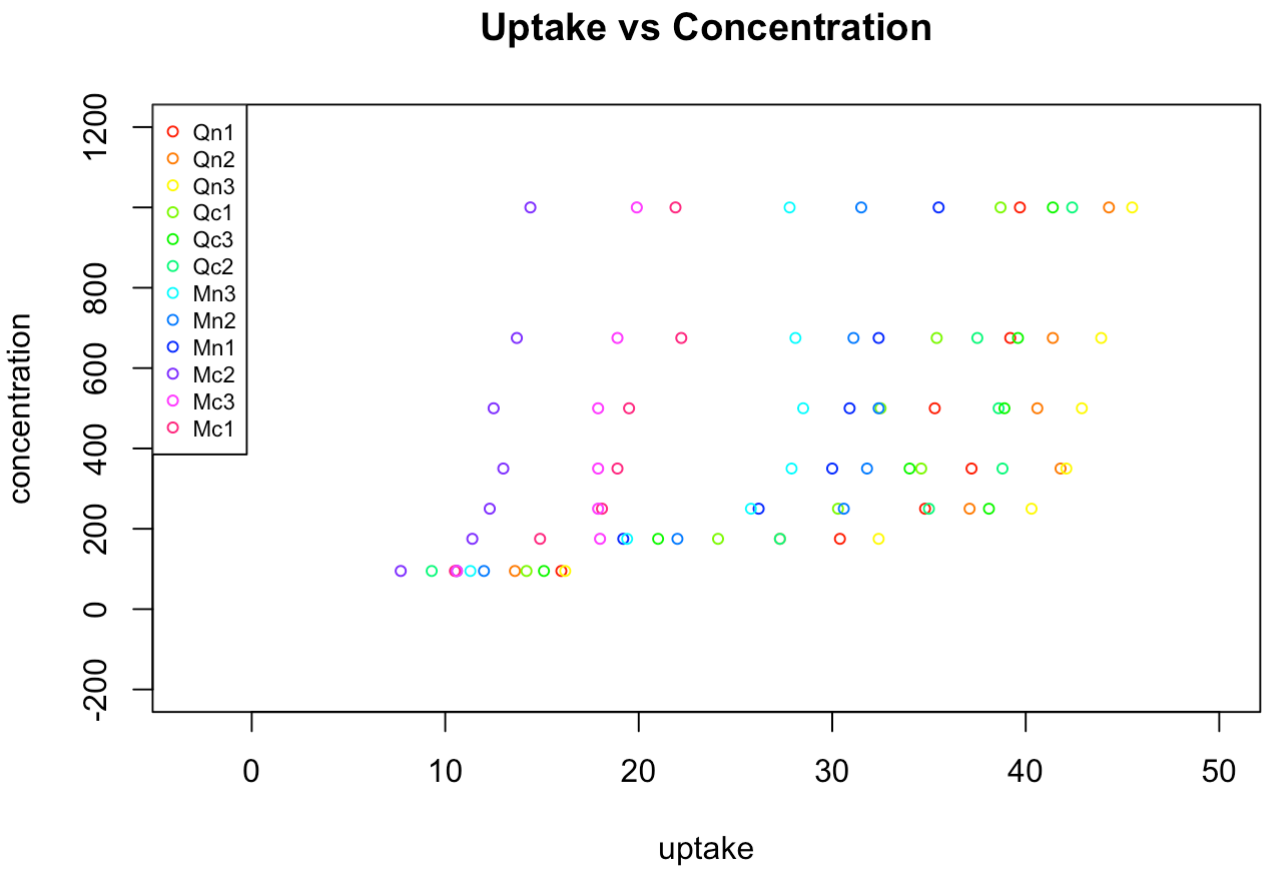
Based on the graph, four blocks' sizes are not similar. Therefore they are not independent.

Plants

Look at the built-in data frame `CO2` . Describe the data set (in English and the kind of language used in an introductory statistics course, not in R language). Mention whether it's from an experiment or an observational study and mention which are independent/explanatory variables and which are dependent/response variables.

Make a graph of uptake vs. concentration, coloring the points according to Plant. The relationship between uptake and concentration is roughly the same for each Plant: describe it.

```
palette(rainbow(12))
color<- palette()
plot(CO2$uptake, CO2$conc, col=CO2$Plant,xlim=c(-3,50),ylim =c(-200,1200),main = "Upt
ake vs Concentration",xlab='uptake', ylab='concentration',cex = 0.7)
legend("topleft",legend = levels(CO2$Plant),col = 1:12,pch = 1 ,cex= 0.7)
```




```
palette("default")
```

Description

The relationship between uptake and concentration is that with the increase of uptake of plants, the concentration also increases.

Break the previous graph into two parts, one for Quebec and one for Mississippi. Which Type has greater variability in uptake for a given concentration?

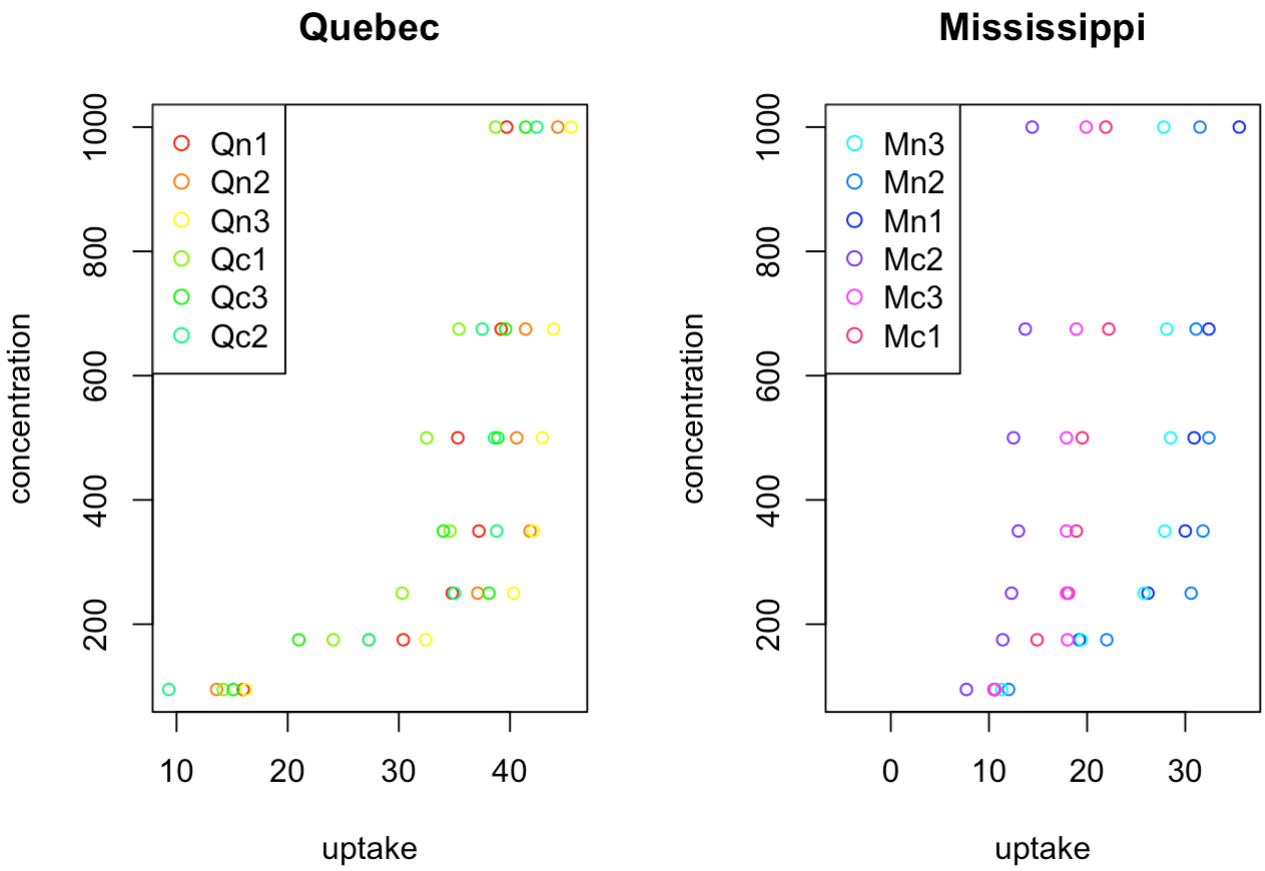
```
par(mfrow = c(1,2))

palette(rainbow(12)[1:6])
Q<-CO2[CO2$Type == "Quebec",]
Q$Plant <- factor(Q$Plant)
plot(Q$uptake,Q$conc,col = Q$Plant,xlab='uptake', ylab='concentration',cex = 0.8,main
 = "Quebec")

legend("topleft",legend = levels(Q$Plant),col = 1:6,pch =1)

palette(rainbow(12)[7:12])
M<-CO2[CO2$Type == "Mississippi",]
M$Plant <- factor(M$Plant)
plot(M$uptake,M$conc,col =M$Plant,xlab='uptake', ylab='concentration',cex = 0.8,main
 = "Mississippi",xlim = c(-5,36))

legend("topleft",legend = levels(M$Plant),col = 1:6 , pch =1)
```




```
palette("default")
```

The Mississippi group has greater variability in uptake for a given concentration

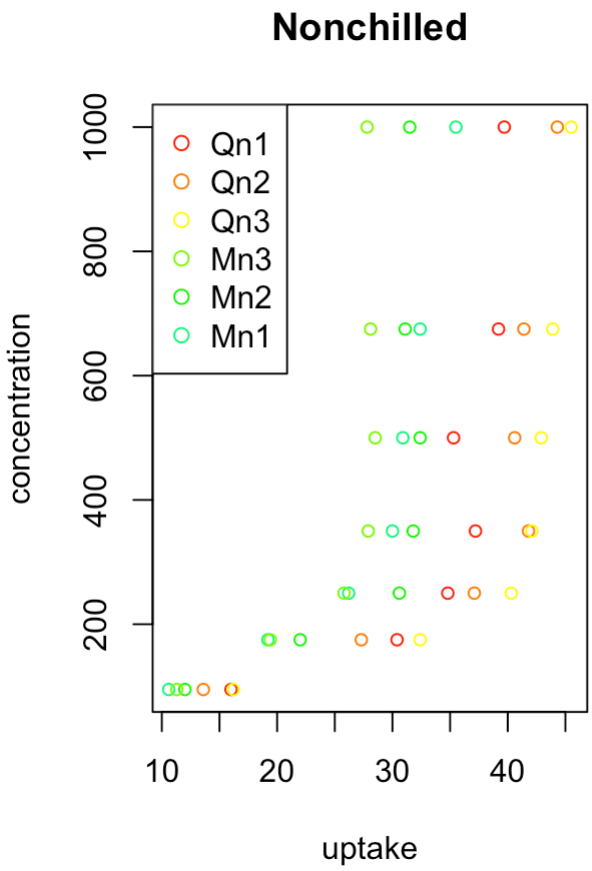
Break the first graph into two parts, one for nonchilled and one for chilled. Broadly speaking, what is the effect of chilling on uptake?

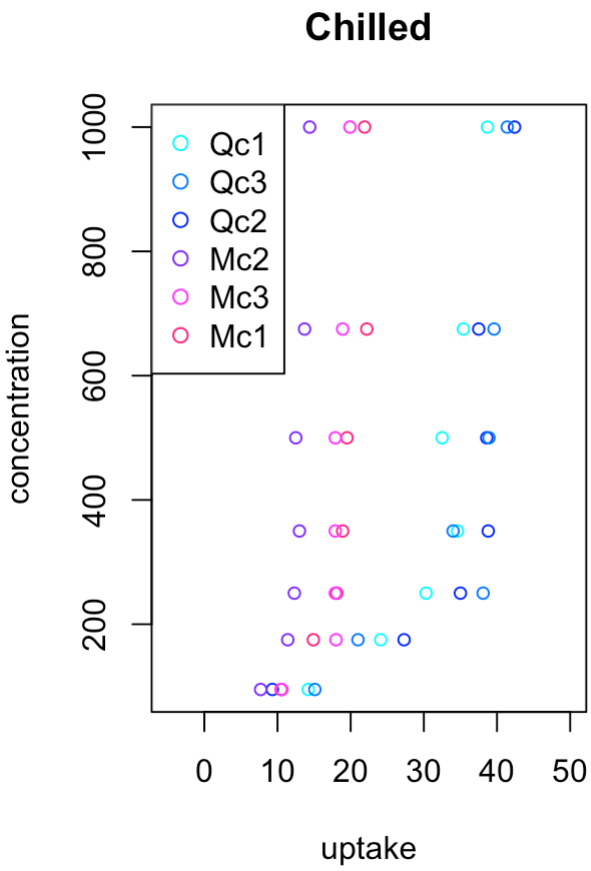
```
par(mfrow = c(1,2))
palette(rainbow(12)[1:6])
N<-CO2[CO2$Treatment == "nonchilled",]
N$Plant <- factor(N$Plant)
plot(N$uptake,N$conc,col = N$Plant,xlab='uptake', ylab='concentration',cex = 0.8,main
 = "Nonchilled")

legend("topleft",legend = levels(N$Plant),col = 1:6,pch =1)

palette(rainbow(12)[7:12])
C<-CO2[CO2$Treatment == "chilled",]
C$Plant <- factor(C$Plant)
plot(C$uptake,C$conc,col =C$Plant,xlab='uptake', ylab='concentration',cex = 0.8,main
 = "Chilled",xlim = c(-5,50))

legend("topleft",legend = levels(C$Plant),col = 1:6 , pch =1)
```






```
palette("default")
```

Chilling on uptake leads to uptake's dispersion,which means the values of uptake becomes more dispersal.