

Chapter 2 Lab: Introduction to R

Your Name Goes Here

Instructions and time-saving hints. To begin this lab, download the archive and unpack it. Inside the folder, you will find a Rmarkdown file and csv files for lab data. Please keep the Rmarkdown file in the same directory as the data files.

The purpose of the first part of this lab is to quickly review the basics of R. To review the basics, you will reproduce the commands blocks in JW Section 2.3 in this Rmarkdown file. **Please do not retype all the R commands as shown in Section 2.3.** Instead, go to link to text for Lab 2, cut and paste the text into your Rmarkdown file. Then break up the R commands into the R chunks shown in the text. (R chunks are braced by triple backticks and a leading {r}, as before.) Do not put all commands into the same block! The idea is to imitate the code chunks in Section 2.3 with the additional plots and output omitted from the text. Below, I have given you the first two blocks for reference.

After breaking the text into the correct chunks, you may knit the document. You will notice two problems when knitting. The first is the appearance of a data editor window (3 times), due to the `fix()` commands in the blocks. You may simply quit these windows each time, or comment out the `fix()` statements. The second problem is an error message, due to the block at the top of p. 50 in JW. You may comment out this `plot()` statement to knit again.

Congratulations! When the document is successfully knitted, you should read Section 2.3 along with the Rmarkdown output. You should see the commands along with the output plots. Of course, the blocks with help commands, `fix()` commands, and intentional errors will not be reproduced. Together, JW2.3 and your Rmarkdown file will help you to review some basics of R. Note that the plotting is done with core R commands and not with the `ggplot2` package as in Homework 0.

Next, you will apply these core R basics to the `College.csv` dataset, as described in problem 8, JW p.54. Please complete the Rmarkdown document corresponding to problem statement. Submit your Rmd file along with an unzipped PDF of the result before the deadline.

Basic Commands

```
x <- c(1,3,2,5)
x

## [1] 1 3 2 5
x = c(1,6,2)
x

## [1] 1 6 2
y = c(1,4,3)

length(x)

## [1] 3
length(y)
```

```

## [1] 3
x+y

## [1] 2 10  5
ls()

## [1] "x" "y"
rm(x,y)
ls()

## character(0)
rm(list=ls())

#?matrix

x=matrix(data=c(1,2,3,4), nrow=2, ncol=2)
x

##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
x=matrix(c(1,2,3,4),2,2)

matrix(c(1,2,3,4),2,2,byrow=TRUE)

##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
sqrt(x)

##      [,1]      [,2]
## [1,] 1.000000 1.732051
## [2,] 1.414214 2.000000
x^2

##      [,1] [,2]
## [1,]    1    9
## [2,]    4   16
x=rnorm(50)
y=x+rnorm(50,mean=50,sd=.1)
cor(x,y)

## [1] 0.9958102
set.seed(1303)
rnorm(50)

## [1] -1.1439763145  1.3421293656  2.1853904757  0.5363925179  0.0631929665
## [6]  0.5022344825 -0.0004167247  0.5658198405 -0.5725226890 -1.1102250073
## [11] -0.0486871234 -0.6956562176  0.8289174803  0.2066528551 -0.2356745091
## [16] -0.5563104914 -0.3647543571  0.8623550343 -0.6307715354  0.3136021252
## [21] -0.9314953177  0.8238676185  0.5233707021  0.7069214120  0.4202043256
## [26] -0.2690521547 -1.5103172999 -0.6902124766 -0.1434719524 -1.0135274099
## [31]  1.5732737361  0.0127465055  0.8726470499  0.4220661905 -0.0188157917
## [36]  2.6157489689 -0.6931401748 -0.2663217810 -0.7206364412  1.3677342065

```

```

## [41] 0.2640073322 0.6321868074 -1.3306509858 0.0268888182 1.0406363208
## [46] 1.3120237985 -0.0300020767 -0.2500257125 0.0234144857 1.6598706557

set.seed(3)
y=rnorm(100)
mean(y)

## [1] 0.01103557

var(y)

## [1] 0.7328675

sqrt(var(y))

## [1] 0.8560768

sd(y)

## [1] 0.8560768

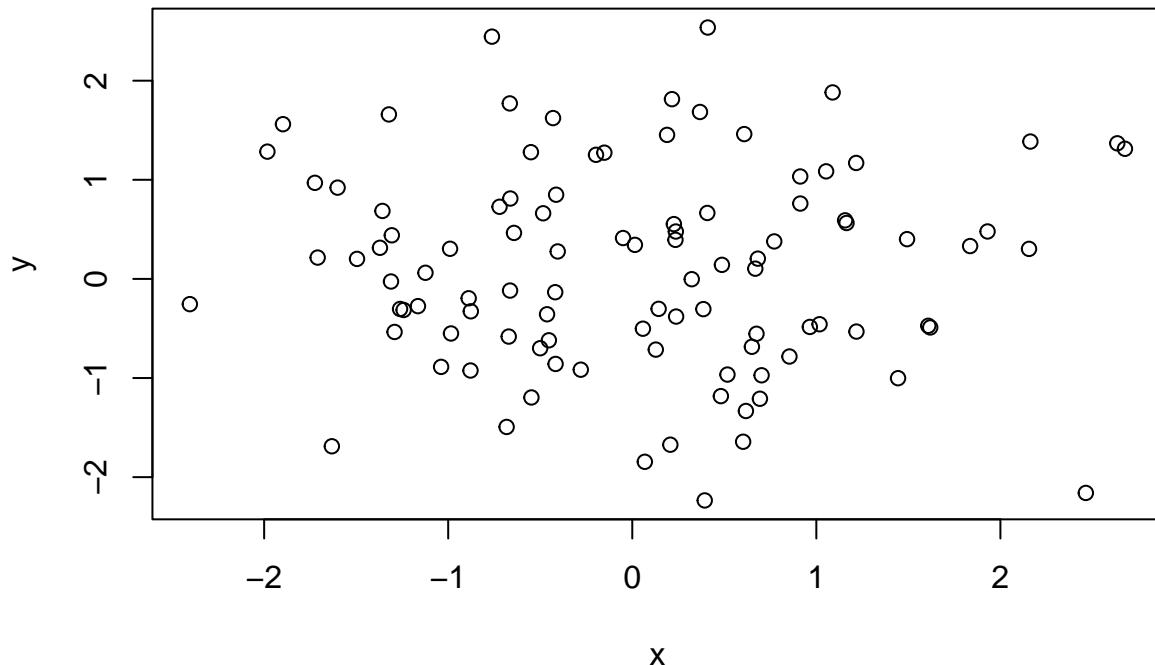
```

Graphics

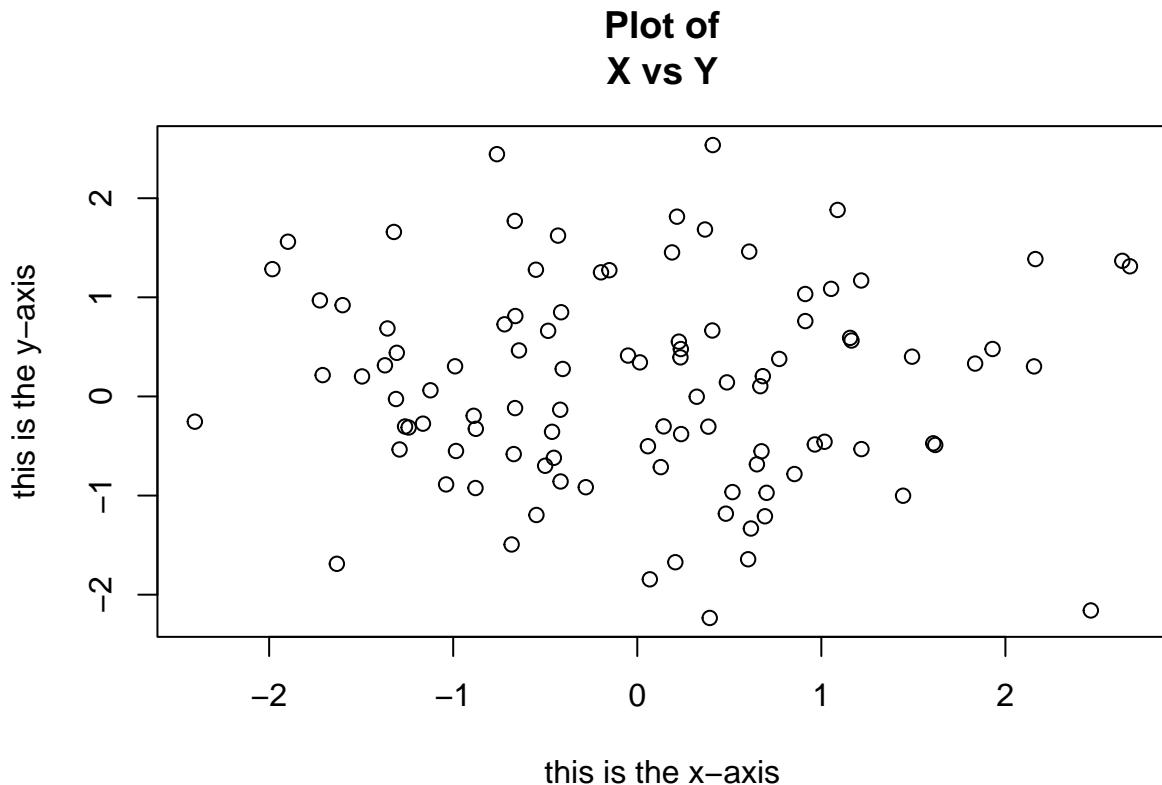
```

x=rnorm(100)
y=rnorm(100)
plot(x,y)

```



```
plot(x,y,xlab="this is the x-axis",ylab="this is the y-axis",main="Plot of X vs Y")
```



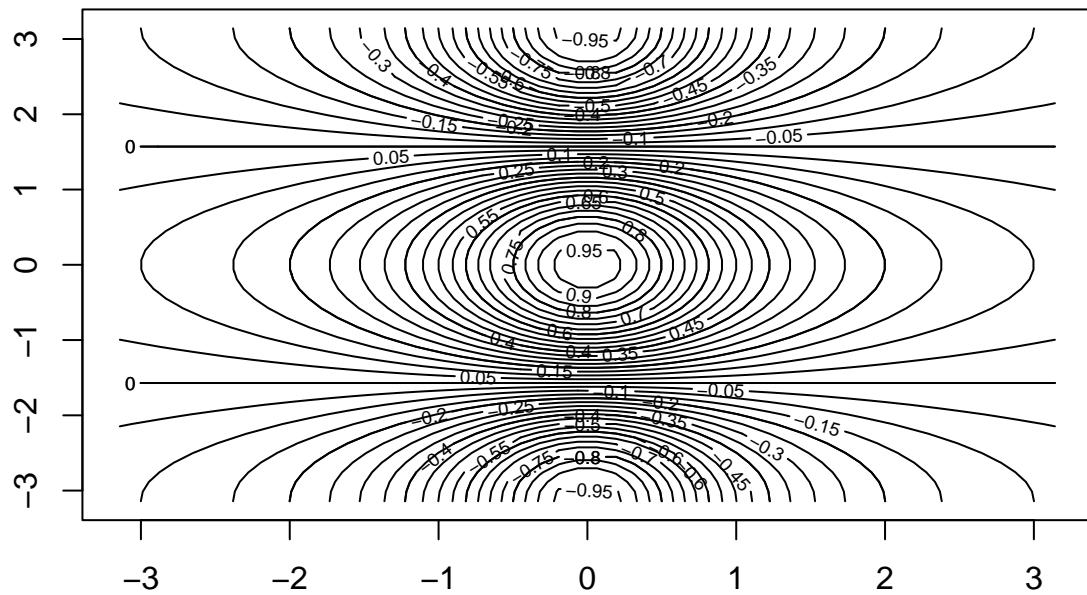
```
pdf("Figure.pdf")
plot(x,y,col="green")
dev.off()
```

```
## pdf
## 2
x=seq(1,10)
x

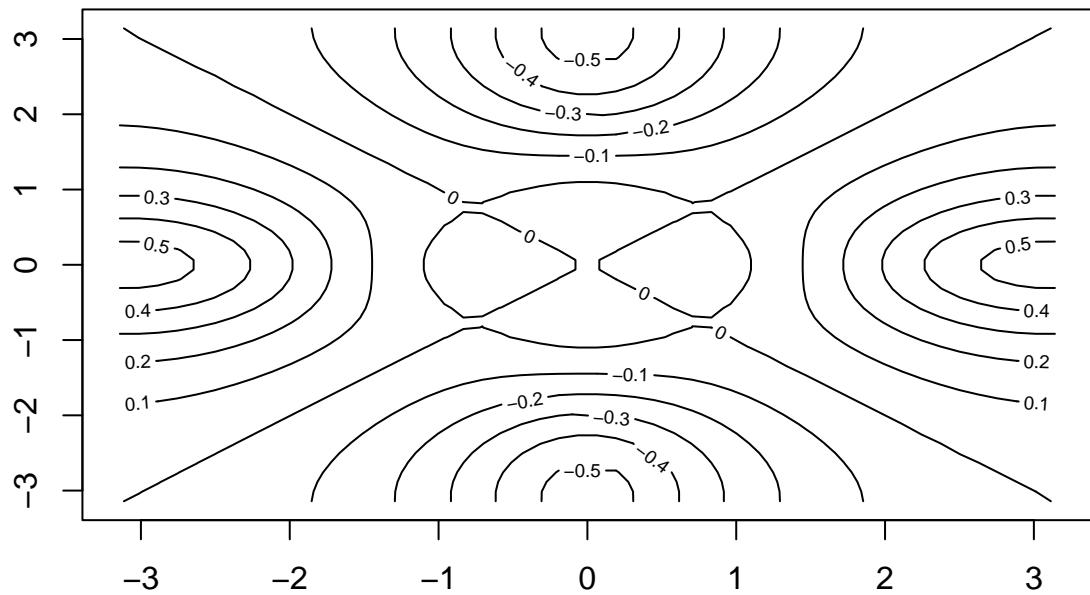
## [1] 1 2 3 4 5 6 7 8 9 10
x=1:10
x

## [1] 1 2 3 4 5 6 7 8 9 10
x=seq(-pi,pi,length=50)

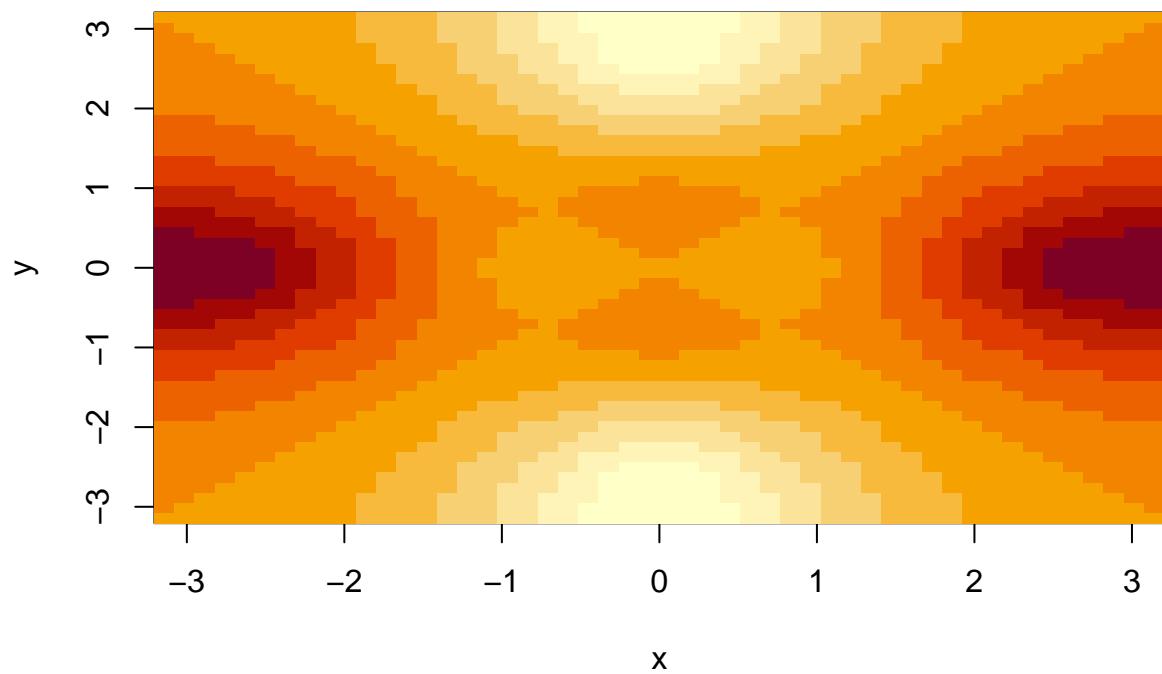
y=x
f=outer(x,y,function(x,y)cos(y)/(1+x^2))
contour(x,y,f)
contour(x,y,f,nlevels=45,add=T)
```



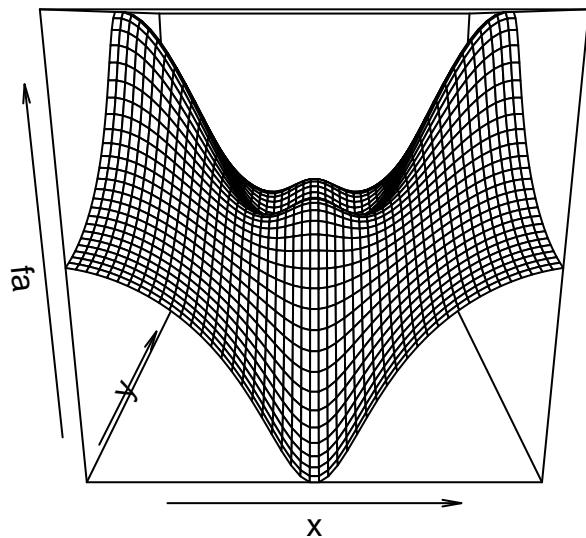
```
fa=(f-t(f))/2  
contour(x,y,fa,nlevels=15)
```



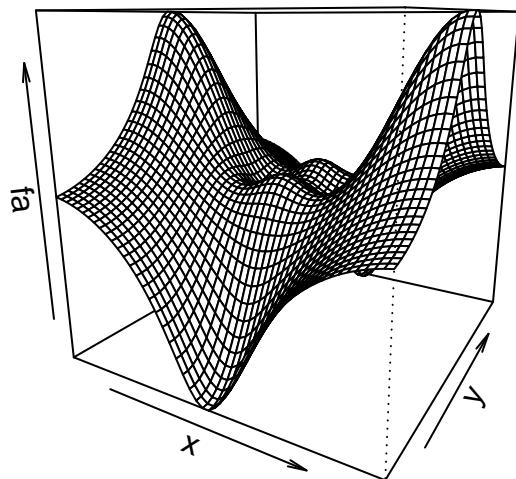
```
image(x,y,fa)
```



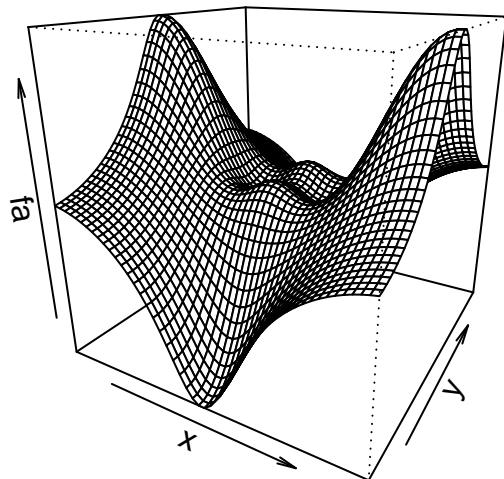
```
persp(x,y,fa)
```



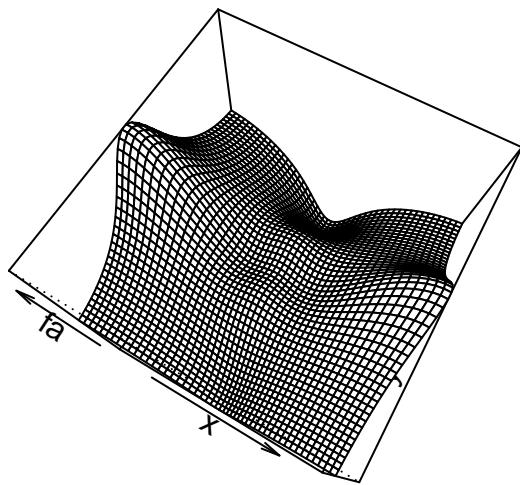
```
persp(x,y,fa,theta=30)
```



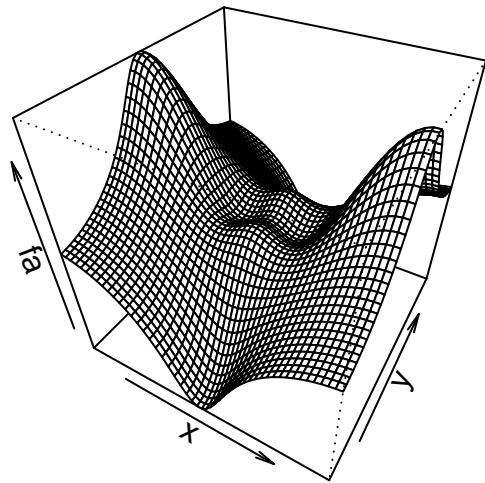
```
persp(x,y,fa,theta=30,phi=20)
```



```
persp(x,y,fa,theta=30,phi=70)
```



```
persp(x,y,fa,theta=30,phi=40)
```



```
# Indexing Data
```

```
A=matrix(1:16,4,4)
A
```

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

```
A[2,3]
```

```
## [1] 10
A[c(1,3),c(2,4)]
```

```
##      [,1] [,2]
## [1,]     5    13
## [2,]     7    15
```

```
A[1:3,2:4]
```

```
##      [,1] [,2] [,3]
## [1,]     5     9    13
## [2,]     6    10    14
## [3,]     7    11    15
```

```
A[1:2,]
```

```
##      [,1] [,2] [,3] [,4]
```

```

## [1,]    1    5    9   13
## [2,]    2    6   10   14
A[,1:2]

##      [,1] [,2]
## [1,]    1    5
## [2,]    2    6
## [3,]    3    7
## [4,]    4    8
A[1,]

## [1] 1 5 9 13
A[-c(1,3),]

##      [,1] [,2] [,3] [,4]
## [1,]    2    6   10   14
## [2,]    4    8   12   16
A[-c(1,3),-c(1,3,4)]

## [1] 6 8
dim(A)

## [1] 4 4

```

Loading Data

```

Auto=read.table("C:/Users/chase/OneDrive/Northeastern Code/DS 5220/Lab2/project2/Auto.data")
#fix(Auto)

Auto=read.table("C:/Users/chase/OneDrive/Northeastern Code/DS 5220/Lab2/project2/Auto.data",header=T,na
#fix(Auto)

Auto=read.csv("C:/Users/chase/OneDrive/Northeastern Code/DS 5220/Lab2/project2/Auto.csv",header=T,na.st
#fix(Auto)
dim(Auto)

## [1] 397   9
Auto[1:4,]

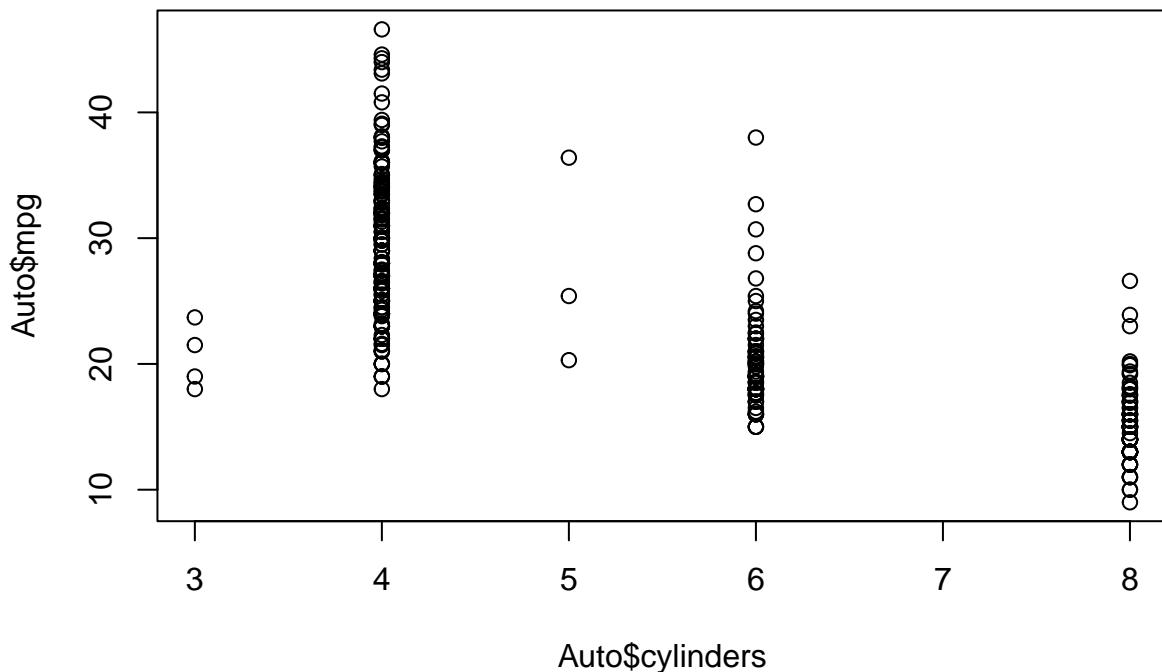
##      mpg cylinders displacement horsepower weight acceleration year origin
## 1    18         8          307        130    3504       12.0     70      1
## 2    15         8          350        165    3693       11.5     70      1
## 3    18         8          318        150    3436       11.0     70      1
## 4    16         8          304        150    3433       12.0     70      1
##                               name
## 1 chevrolet chevelle malibu
## 2           buick skylark 320
## 3      plymouth satellite
## 4           amc rebel sst
Auto=na.omit(Auto)
dim(Auto)

```

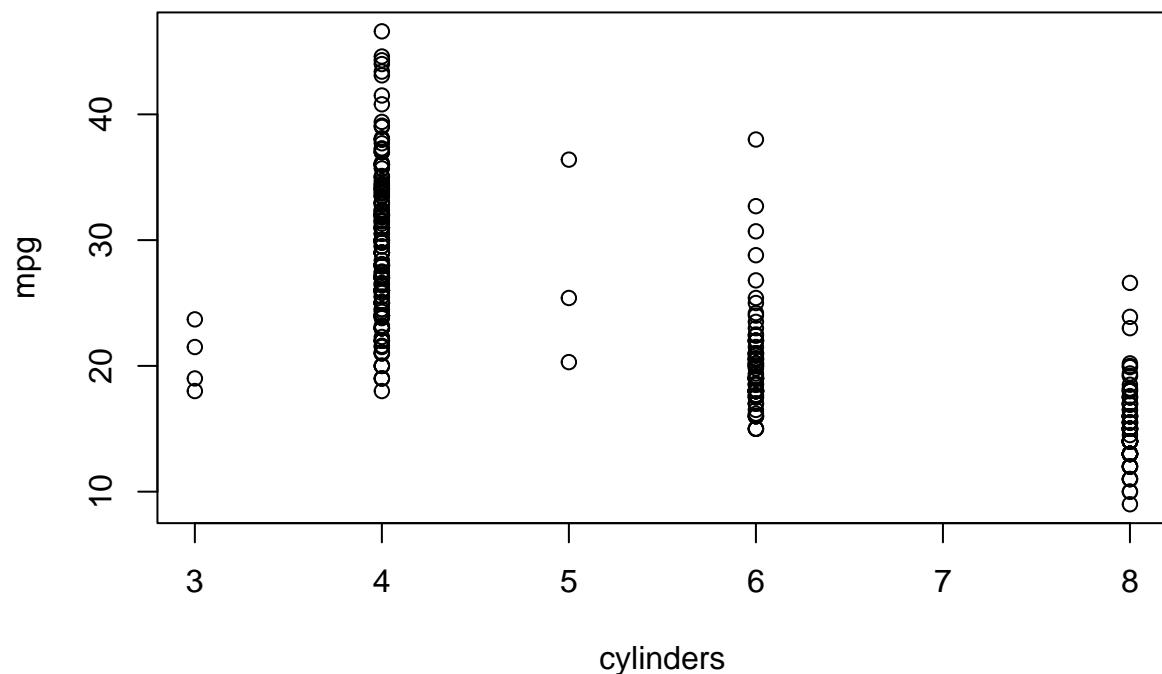
```
## [1] 392    9  
names(Auto)  
  
## [1] "mpg"          "cylinders"     "displacement"  "horsepower"   "weight"  
## [6] "acceleration"  "year"         "origin"        "name"
```

Additional Graphical and Numerical Summaries

```
#plot(cylinders, mpg)  
plot(Auto$cylinders, Auto$mpg)
```

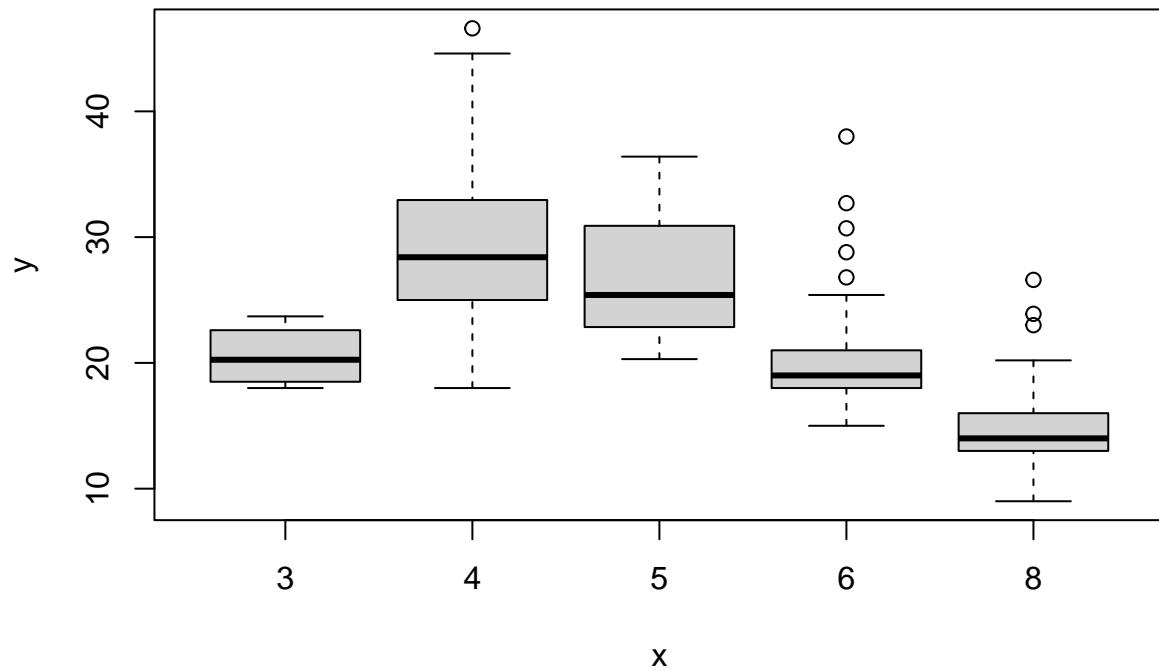


```
attach(Auto)  
plot(cylinders, mpg)
```

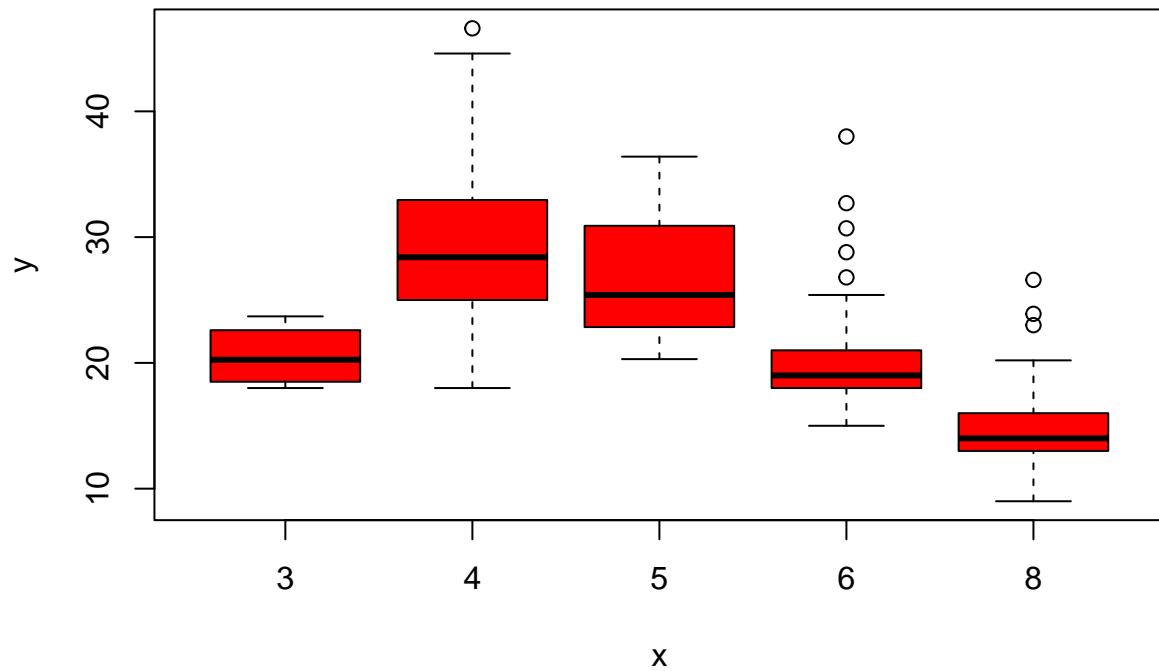


```
cylinders=as.factor(cylinders)
```

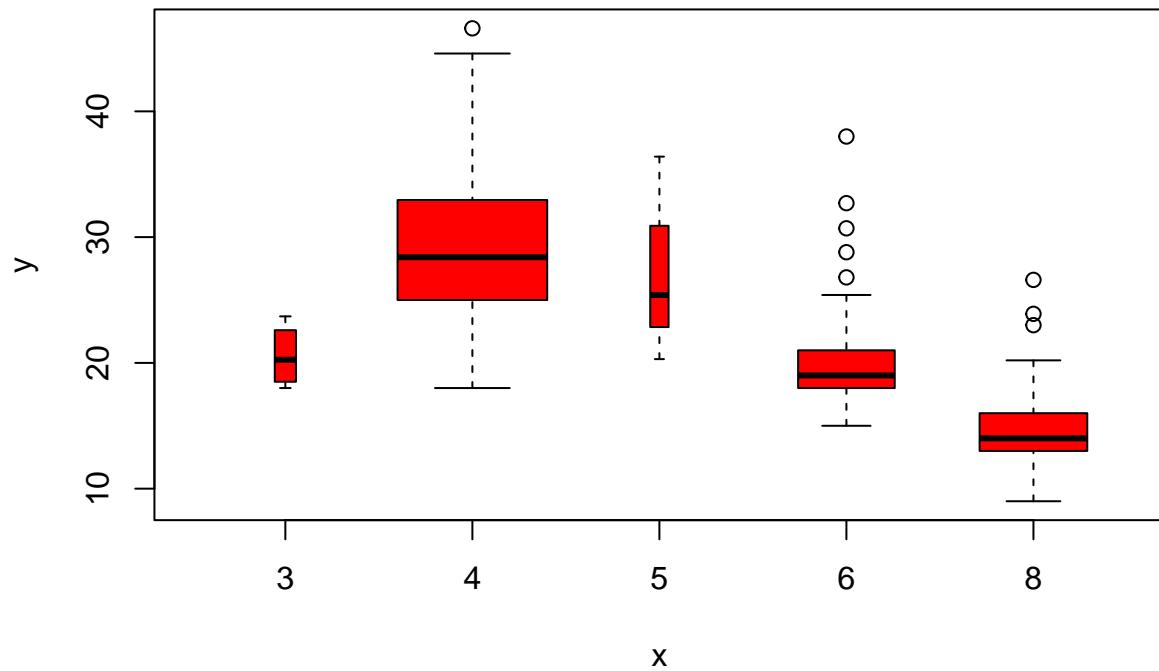
```
plot(cylinders, mpg)
```



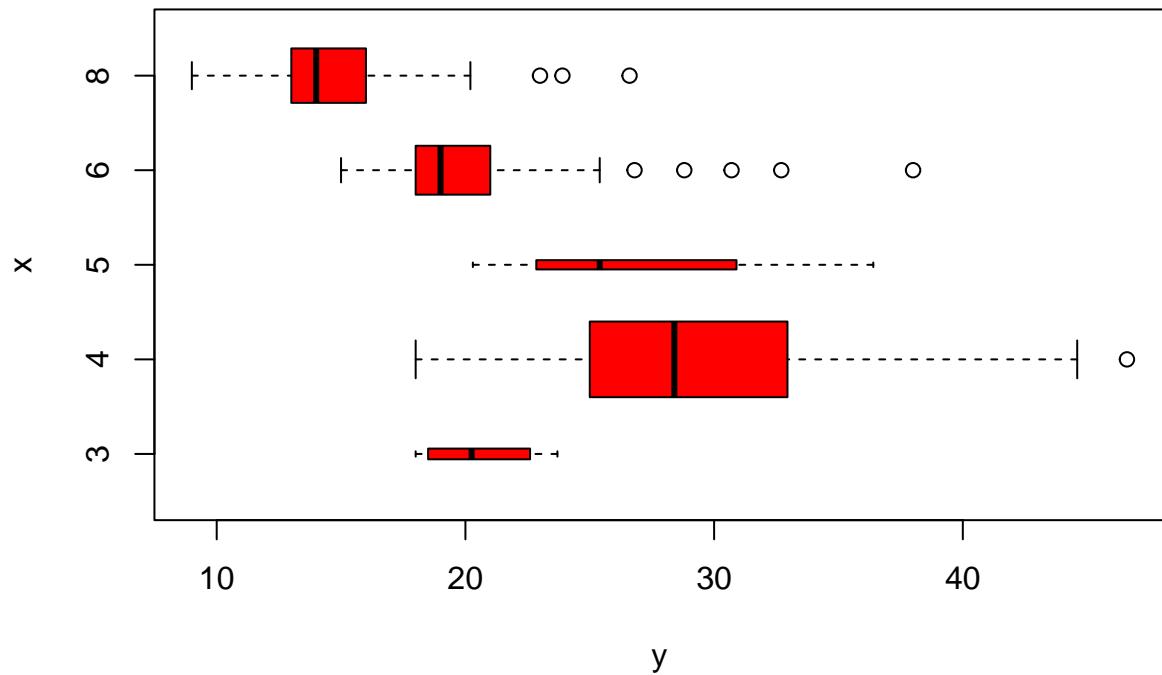
```
plot(cylinders, mpg, col="red")
```



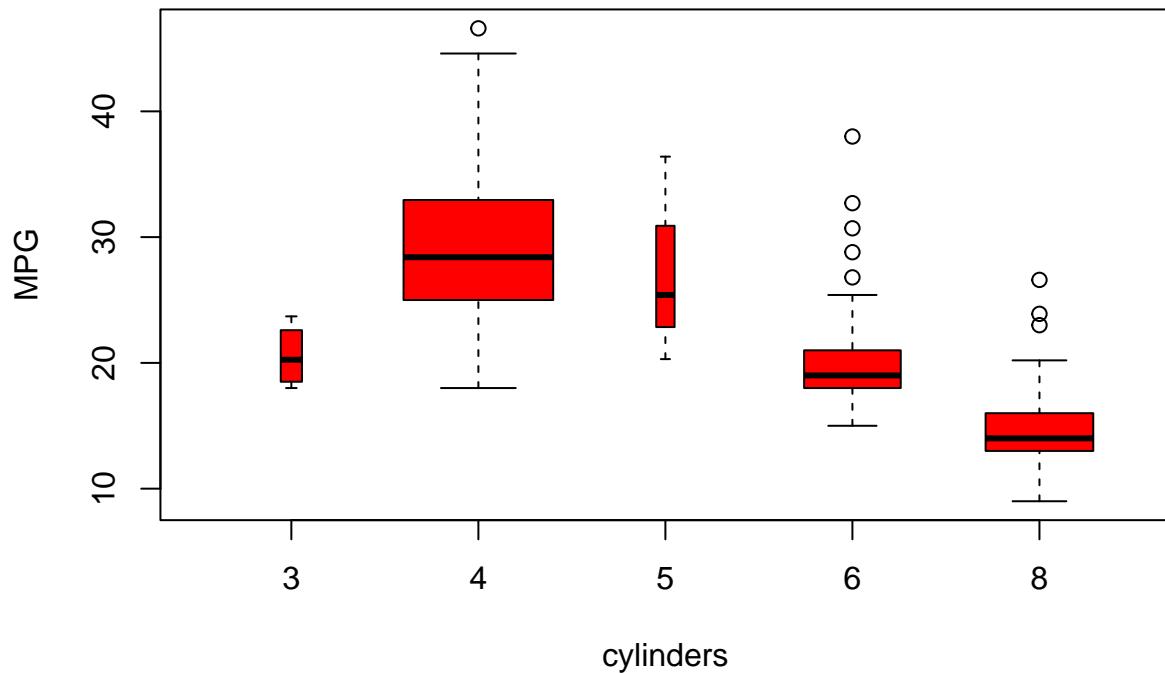
```
plot(cylinders, mpg, col="red", varwidth=T)
```



```
plot(cylinders, mpg, col="red", varwidth=T, horizontal=T)
```

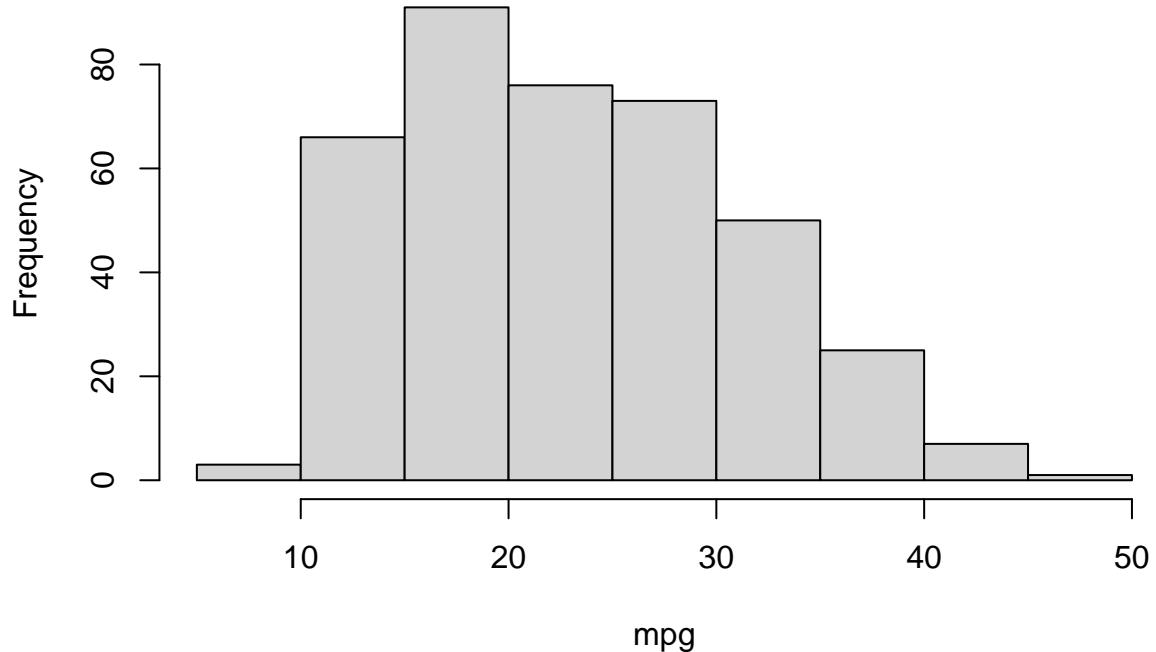


```
plot(cylinders, mpg, col="red", varwidth=T, xlab="cylinders", ylab="MPG")
```



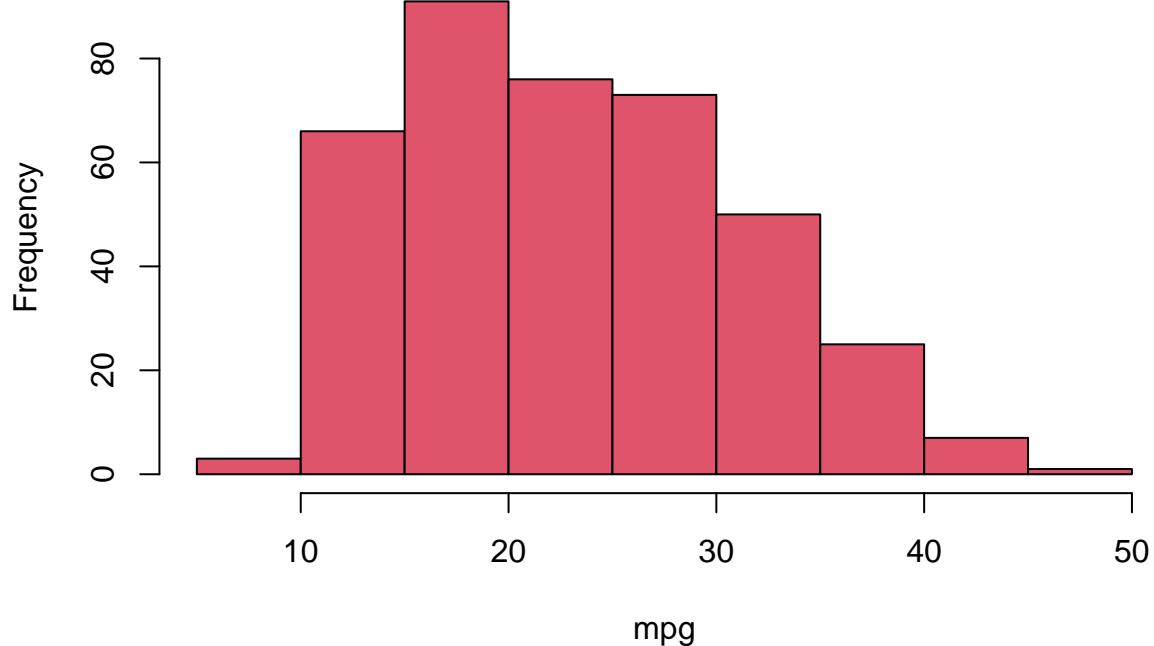
```
hist(mpg)
```

Histogram of mpg



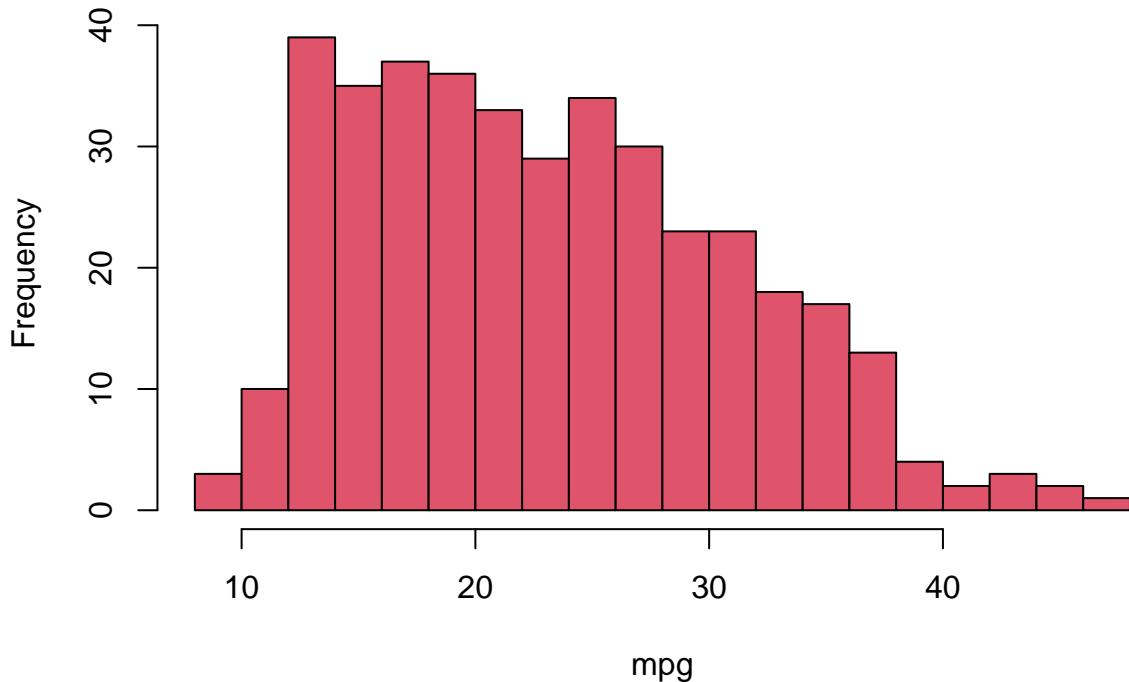
```
hist(mpg, col=2)
```

Histogram of mpg

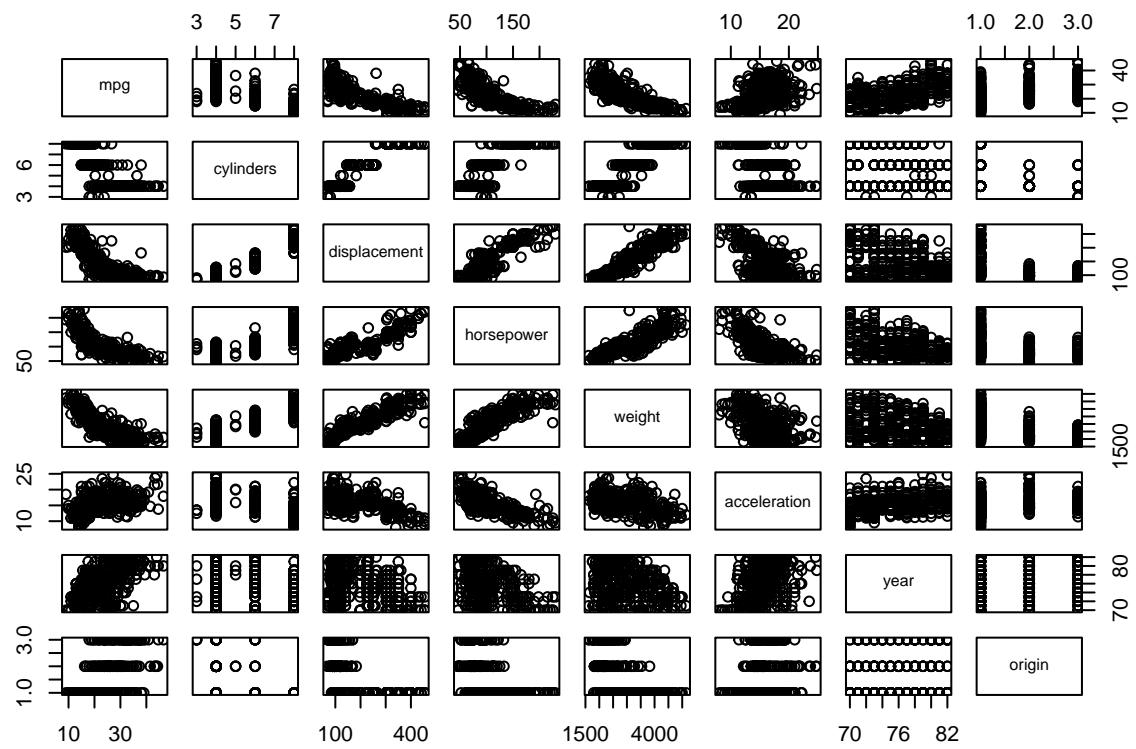


```
hist(mpg, col=2, breaks=15)
```

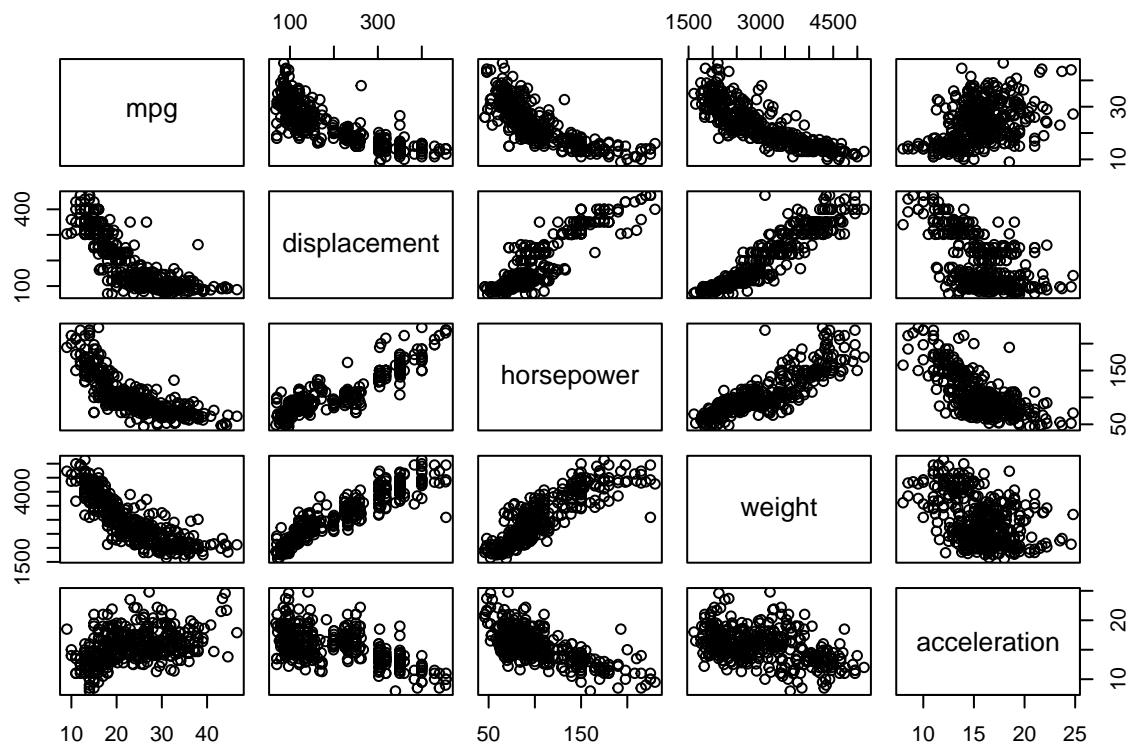
Histogram of mpg



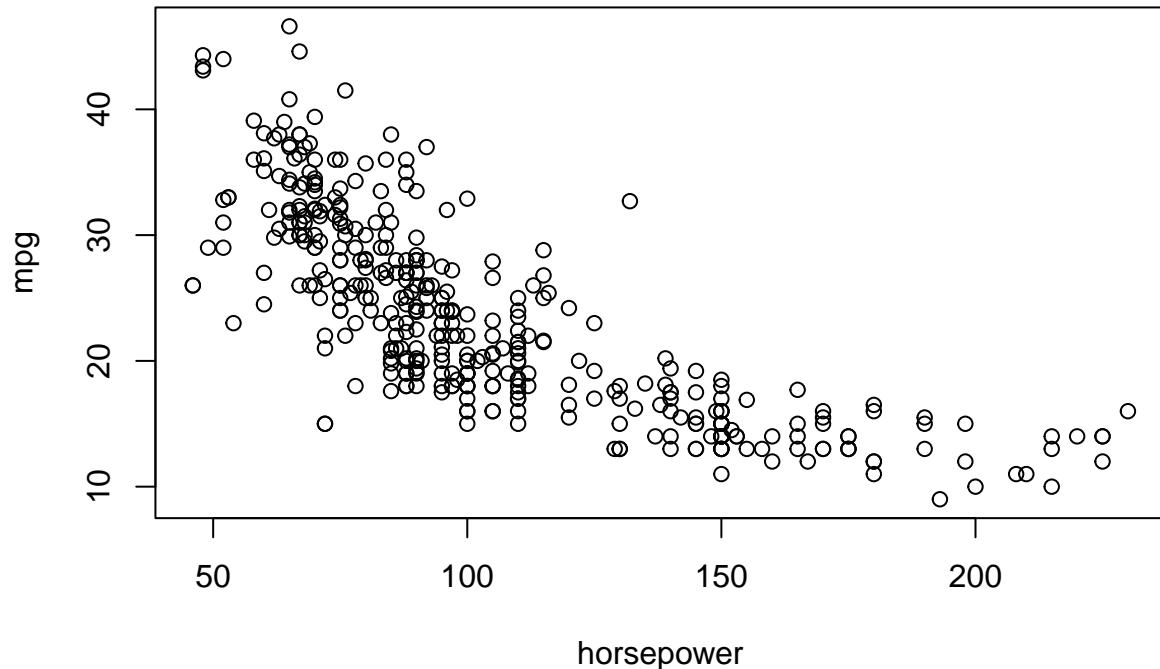
```
pairs(Auto[1:8])
```



```
pairs(~ mpg + displacement + horsepower + weight + acceleration, Auto)
```



```
plot(horsepower,mpg)
identify(horsepower,mpg,name)
```



```

## integer(0)
summary(Auto)

##      mpg      cylinders      displacement      horsepower      weight
##  Min.   : 9.00  Min.   :3.000  Min.   :68.0  Min.   :46.0  Min.   :1613
##  1st Qu.:17.00  1st Qu.:4.000  1st Qu.:105.0  1st Qu.:75.0  1st Qu.:2225
##  Median :22.75  Median :4.000  Median :151.0  Median :93.5  Median :2804
##  Mean   :23.45  Mean   :5.472  Mean   :194.4  Mean   :104.5  Mean   :2978
##  3rd Qu.:29.00  3rd Qu.:8.000  3rd Qu.:275.8  3rd Qu.:126.0  3rd Qu.:3615
##  Max.   :46.60  Max.   :8.000  Max.   :455.0  Max.   :230.0  Max.   :5140
##      acceleration      year      origin      name
##  Min.   : 8.00  Min.   :70.00  Min.   :1.000  Length:392
##  1st Qu.:13.78  1st Qu.:73.00  1st Qu.:1.000  Class :character
##  Median :15.50  Median :76.00  Median :1.000  Mode  :character
##  Mean   :15.54  Mean   :75.98  Mean   :1.577
##  3rd Qu.:17.02  3rd Qu.:79.00  3rd Qu.:2.000
##  Max.   :24.80  Max.   :82.00  Max.   :3.000

summary(mpg)

##      Min. 1st Qu. Median  Mean 3rd Qu.  Max.
##  9.00 17.00 22.75 23.45 29.00 46.60

#Applied JW p.54 (8.)##

```

Here, we are using the College data set, found in `College.csv`. Before reading the data into R, it can be viewed in Excel or a text editor.

1. Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

```
path = "C:/Users/chase/OneDrive/Northeastern Code/DS 5220/Lab2/project2/College.csv"
college = read.csv(path)
```

2. Look at the data using the `fix()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
rownames(college)=college[,1]
#fix(college)
```

You should see that there is now a `row.names` column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
college=college[,-1]
#fix(college)
```

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that R is giving to each row.

3. Please complete these parts.

Use the ``summary()`` function to produce a numerical summary of the variables in the data set.

```
summary(college)
```

	Private	Apps	Accept	Enroll
## Length:	777	Min. : 81	Min. : 72	Min. : 35
## Class :	character	1st Qu.: 776	1st Qu.: 604	1st Qu.: 242
## Mode :	character	Median : 1558	Median : 1110	Median : 434
##		Mean : 3002	Mean : 2019	Mean : 780
##		3rd Qu.: 3624	3rd Qu.: 2424	3rd Qu.: 902
##		Max. :48094	Max. :26330	Max. :6392
## Top10perc	Top25perc	F.Undergrad	P.Undergrad	
## Min. : 1.00	Min. : 9.0	Min. : 139	Min. : 1.0	
## 1st Qu.: 15.00	1st Qu.: 41.0	1st Qu.: 992	1st Qu.: 95.0	
## Median : 23.00	Median : 54.0	Median : 1707	Median : 353.0	
## Mean : 27.56	Mean : 55.8	Mean : 3700	Mean : 855.3	
## 3rd Qu.: 35.00	3rd Qu.: 69.0	3rd Qu.: 4005	3rd Qu.: 967.0	
## Max. : 96.00	Max. :100.0	Max. :31643	Max. :21836.0	
## Outstate	Room.Board	Books	Personal	
## Min. : 2340	Min. :1780	Min. : 96.0	Min. : 250	
## 1st Qu.: 7320	1st Qu.:3597	1st Qu.: 470.0	1st Qu.: 850	
## Median : 9990	Median :4200	Median : 500.0	Median :1200	
## Mean : 10441	Mean :4358	Mean : 549.4	Mean :1341	
## 3rd Qu.:12925	3rd Qu.:5050	3rd Qu.: 600.0	3rd Qu.:1700	
## Max. :21700	Max. :8124	Max. :2340.0	Max. :6800	
## PhD	Terminal	S.F.Ratio	perc.alumni	
## Min. : 8.00	Min. : 24.0	Min. : 2.50	Min. : 0.00	
## 1st Qu.: 62.00	1st Qu.: 71.0	1st Qu.:11.50	1st Qu.:13.00	
## Median : 75.00	Median : 82.0	Median :13.60	Median :21.00	
## Mean : 72.66	Mean : 79.7	Mean :14.09	Mean :22.74	
## 3rd Qu.: 85.00	3rd Qu.: 92.0	3rd Qu.:16.50	3rd Qu.:31.00	

```

##   Max.    :103.00   Max.    :100.0   Max.    :39.80   Max.    :64.00
##   Expend      Grad.Rate
##   Min.    : 3186   Min.    : 10.00
##   1st Qu.: 6751   1st Qu.: 53.00
##   Median  : 8377   Median  : 65.00
##   Mean    : 9660   Mean    : 65.46
##   3rd Qu.:10830   3rd Qu.: 78.00
##   Max.    :56233   Max.    :118.00

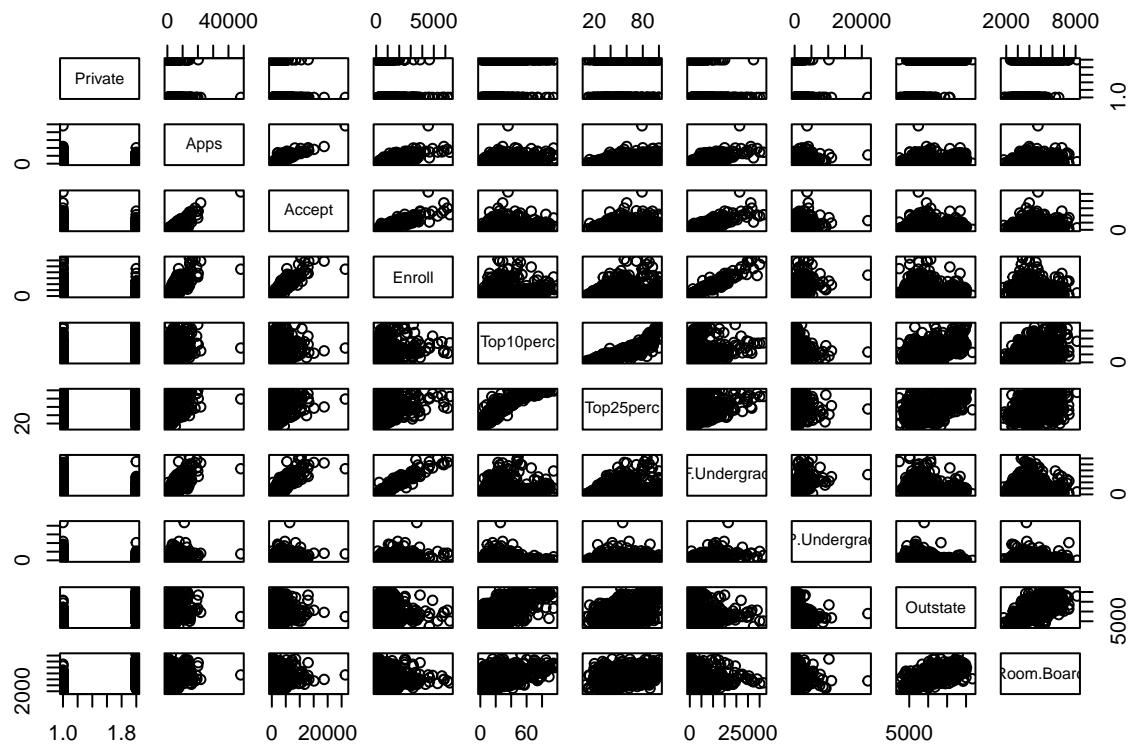
```

Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

```

college[,1] = as.numeric(factor(college[,1]))
pairs(college[,1:10])

```



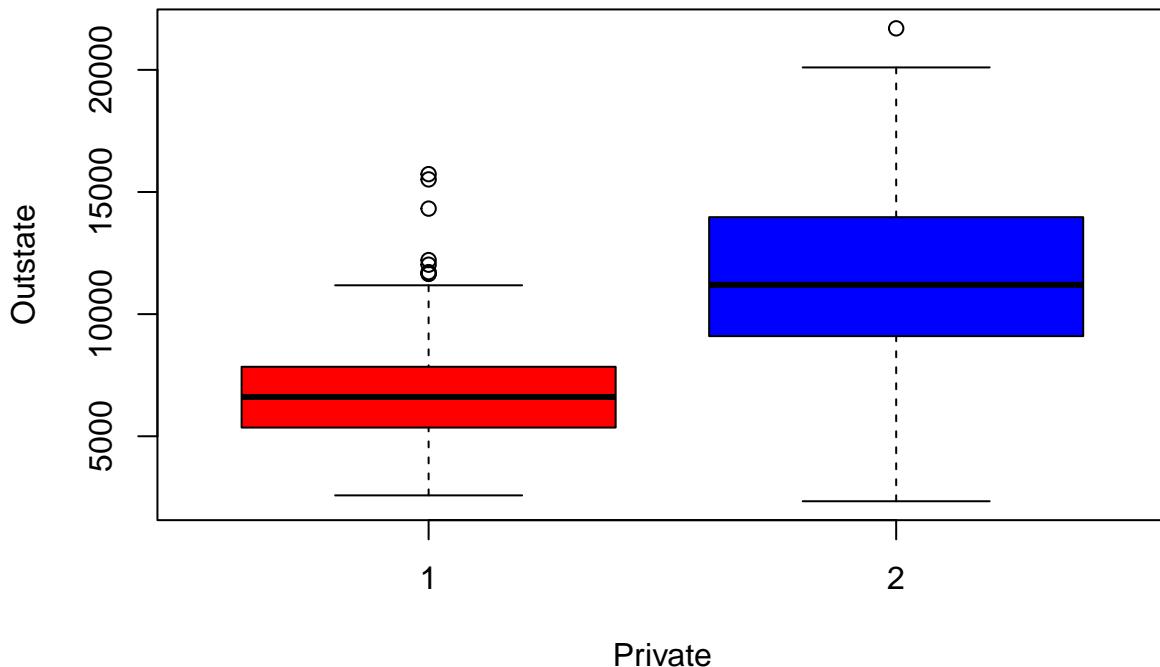
Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

```

boxplot(college$Outstate ~ college$Private, main = "Outstate vs Private",
       xlab = "Private", ylab = "Outstate", col = c("red", "blue"))

```

Outstate vs Private



Create a new qualitative variable, called `Elite`, by *binning* the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50 %.

```
Elite=rep("No",nrow(college))
Elite[college$Top10perc >50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college ,Elite)
```

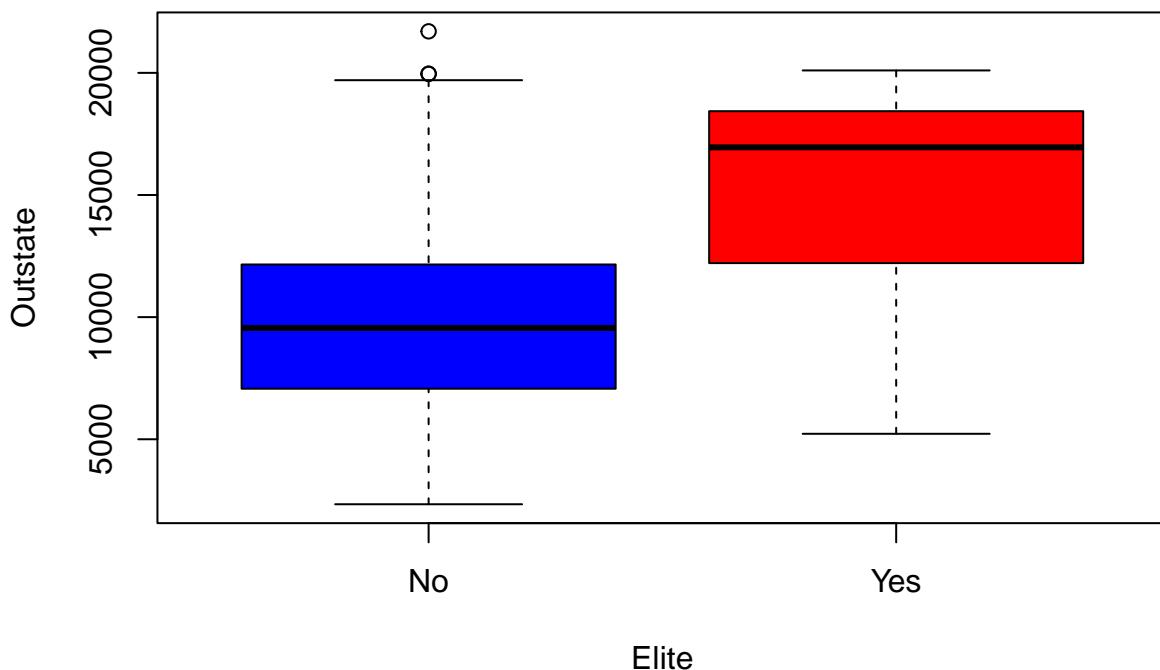
Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

```
summary(college$Elite)

##  No Yes
## 699  78

boxplot(college$Outstate ~ college$Elite, col = c("blue", "red"), main = "Outstate vs Elite", xlab = "E
```

Outstate vs Elite

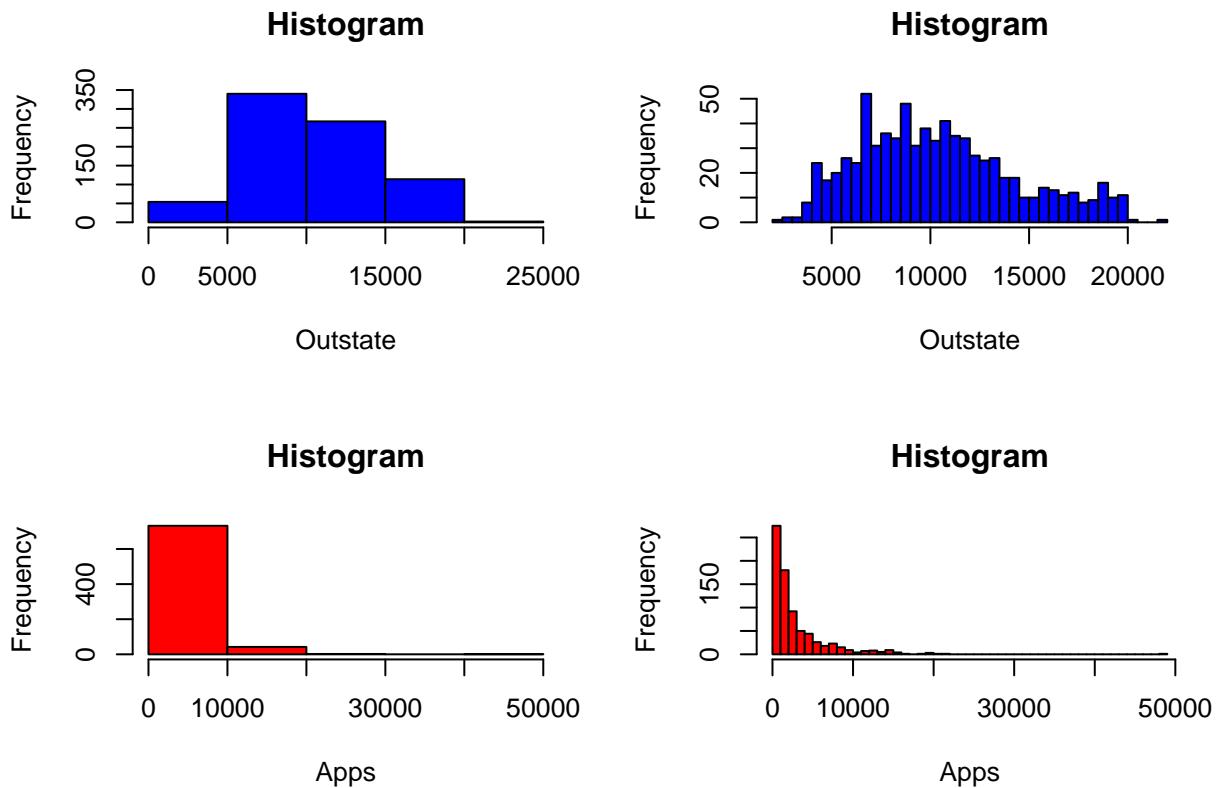


Use the `hist()` function to produce some histograms with 5 and 50 bins for `Outstate` and `Apps`. Use the command `par(mfrow=c(2,2))`: it will divide the print window into four regions so that four plots can be made simultaneously.

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

hist(college$Outstate, breaks = 5, col = "blue", xlab = "Outstate", main = "Histogram")
hist(college$Outstate, breaks = 50, col = "blue", xlab = "Outstate", main = "Histogram")

hist(college$Apps, breaks = 5, col = "red", xlab = "Apps", main = "Histogram")
hist(college$Apps, breaks = 50, col = "red", xlab = "Apps", main = "Histogram")
```



Continue exploring the data, and provide a brief summary of what you discover.

This statement is intentionally vague: as a data scientist, it will be your job to propose hypotheses about the data, and then to use the data to address your hypotheses. This is a creative, iterative, (and fun) process.

In this first lab I will propose four hypotheses for you. You must address these hypotheses with the suggested approaches. The approach will be to produce one plot for each hypothesis in a 2 by 2 set of plots. Discuss your findings for each case, including: whether the data supports or rejects the hypothesis, and to what degree the hypothesis is rejected or supported. After addressing the four hypotheses, I would like for you to propose at least one more hypothesis, and develop a methodology. Brevity matters, and less is best. Produce a single plot upon which to base your answer. Long answers will be given little credit. For this last part, you will be graded on the “interest” of the hypothesis, your approach to address it, and your discussion.

Hypothesis 1. The tuition at the best colleges, as indicated by Elite, is higher than that at other colleges. **Methodology:** Use one-over-other horizontal red boxplots of Outstate for elite and non-elite colleges. Title and label the axes of this plot thoughtfully. Carefully use this single plot to address this hypothesis. Can you be precise about the word *higher* ?

The enrollment rate is the fraction of accepted students who enrolled. *Hypothesis 2. The enrollment rate at the best colleges, as indicated by Elite, is higher than that at other colleges.* **Methodology:** Create a new variable called EnrollRate, using Enroll and Accept. Use the attach() function to make your code cleaner. Use one-over-other horizontal green boxplots of EnrollRate for elite and non-elite colleges. Title and label the axes of this plot thoughtfully. Carefully use this single plot to address this hypothesis. Can you be precise about the word *higher* ? Can you comment on the presence of outliers in the boxplots?

Hypothesis 3. The number of applications per enrolled student is higher at elite colleges than other colleges. **Methodology:** Use one-over-other horizontal blue boxplots of applications per enrolled student for elite and non-elite colleges. Title and label the axes of this plot thoughtfully. Carefully use this single plot to address

this hypothesis. Can you be precise about the word *higher* ? Can you comment on the presence of outliers in the boxplots?

Hypothesis 4. The fraction of alumni who donate is higher at elite colleges than other colleges. **Methodology:** Use one-over-other horizontal cyan boxplots of `perc.alumni` for elite and non-elite colleges. Title and label the axes of this plot thoughtfully. Carefully use this single plot to address this hypothesis. Can you be precise about the word *higher* ?

```
#Hypothesis 1
par(mfrow=c(2,2))
attach(college)

## The following object is masked _by_ .GlobalEnv:
##
##      Elite

boxplot(Outstate ~ Elite,
        main = "Tuition at Top Colleges Tends to be Higher",
        ylab = "Elite",
        xlab = "Outstate Tuition",
        col = "red",
        horizontal = TRUE)
# On average it seems that the tuition at 'Elite' colleges tends to
# cost more than at non-Elite colleges, with a slighter larger deviation.

#Hypothesis 2
EnrollRate = Accept / Enroll
college = data.frame(college, EnrollRate)

boxplot(EnrollRate ~ Elite,
        main = "Enrollment Rate at Colleges",
        horizontal = TRUE,
        xlab = "Enrollment Rate",
        col = "green")

# Enrollment Rate at Elite colleges appears to be very slightly more frequent on average,
# and with a similar spread, though there appear to be several outliers of very high enrollment
# rates for some non-Elite colleges, possibly some less exclusive colleges
# like community colleges or otherwise.

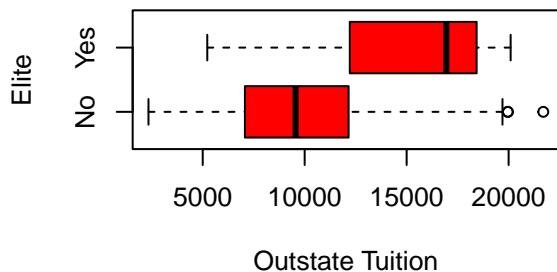
#Hypothesis 3
EnrollperApp = Apps / Enroll
college = data.frame(college, EnrollperApp)

boxplot(EnrollperApp ~ Elite,
        horizontal = TRUE,
        col = "blue",
        main = "Applications per Student at Elite Colleges",
        xlab = "Applications per Enrollment")

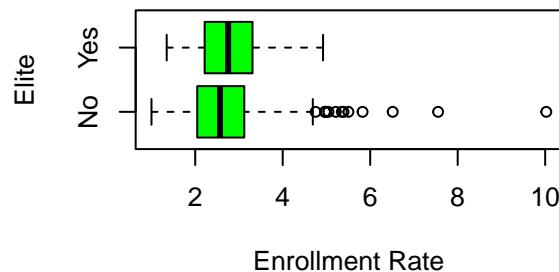
# The number of applications per enrolled student does appear to be higher on average at elite schools,
# and with a slightly larger spread. There are several outliers in the non-elite schools, however,
# possibly due to some more exclusive but 'non-elite' colleges.
```

```
#Hypothesis 4
boxplot(perc.alumni ~ Elite,
        horizontal = TRUE,
        col = "cyan",
        main = "Alumni Donations Higher at Elite Colleges",
        xlab = "% Alumni Donations")
```

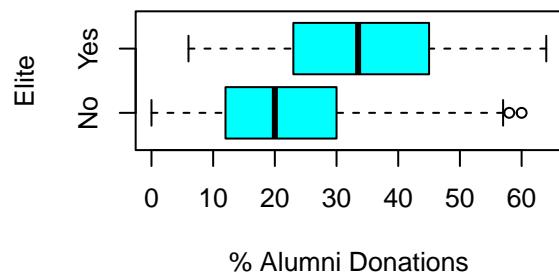
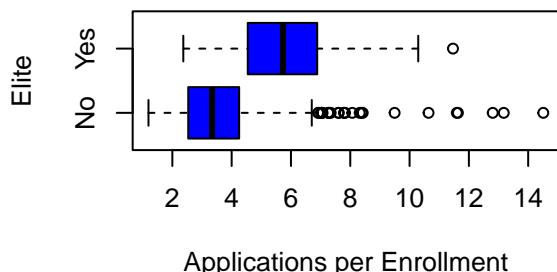
Tuition at Top Colleges Tends to be High



Enrollment Rate at Colleges



Applications per Student at Elite College **Alumni Donations Higher at Elite College**



```
# The percent of Alumni who donate to their college does appear to be higher
# on average at Elite colleges, with a similar spread to non-Elite colleges.
```

```
detach(college)
```