Chapter 2 Lab: Introduction to R

VRAJ DIYORA

3/17/2024

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

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

```
## [1] 1 3 2 5
```

```r
x = c(1,6,2)
x
```

```
## [1] 1 6 2
```

```r
y = c(1,4,3)
```

```r
length(x)
```

```
## [1] 3
```

```r
length(y)
```

```
## [1] 3
```

```r
x+y
```

```
## [1]  2 10  5
```

```r
ls()
```

```
## [1] "x" "y"
```

```r
rm(x,y)
ls()
```

```
## character(0)
```

```r
rm(list=ls())
```

```r
?matrix
```

```
## starting httpd help server ... done
```

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

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

```r
x=matrix(c(1,2,3,4),2,2)
```

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

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

```r
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] 0.9953808
```

```
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))
```

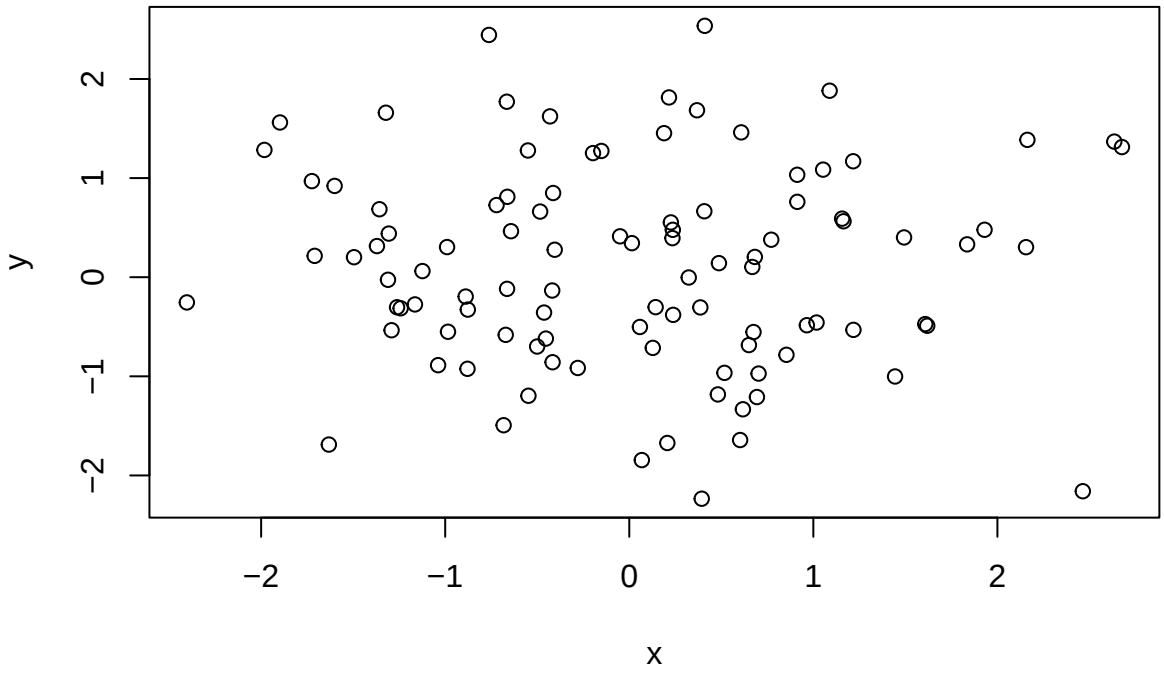
```
## [1] 0.8560768
```

```
sd(y)
```

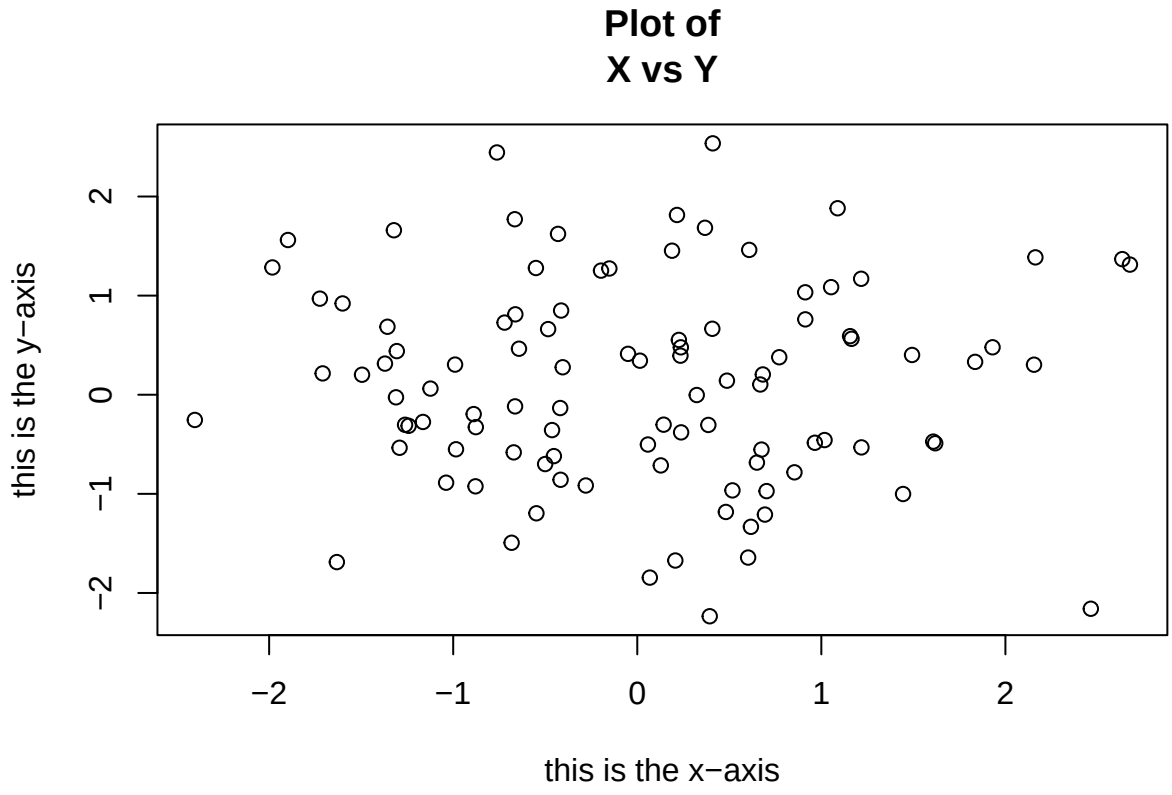
```
## [1] 0.8560768
```

Graphics

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




```r
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
```

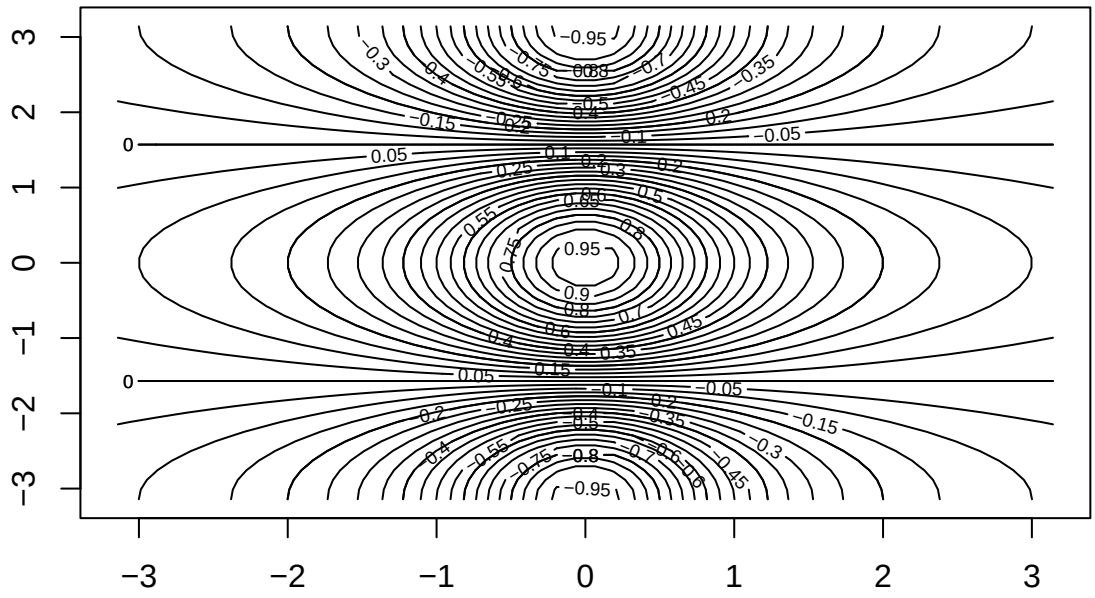
```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
x=1:10
x
```

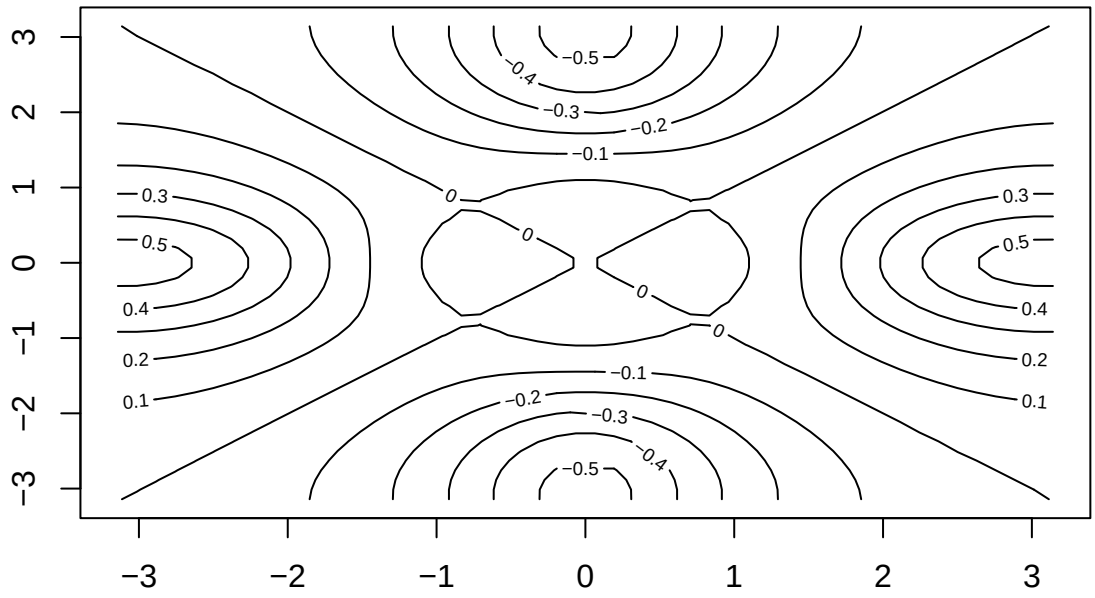
```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
x=seq(-pi,pi,length=50)
```

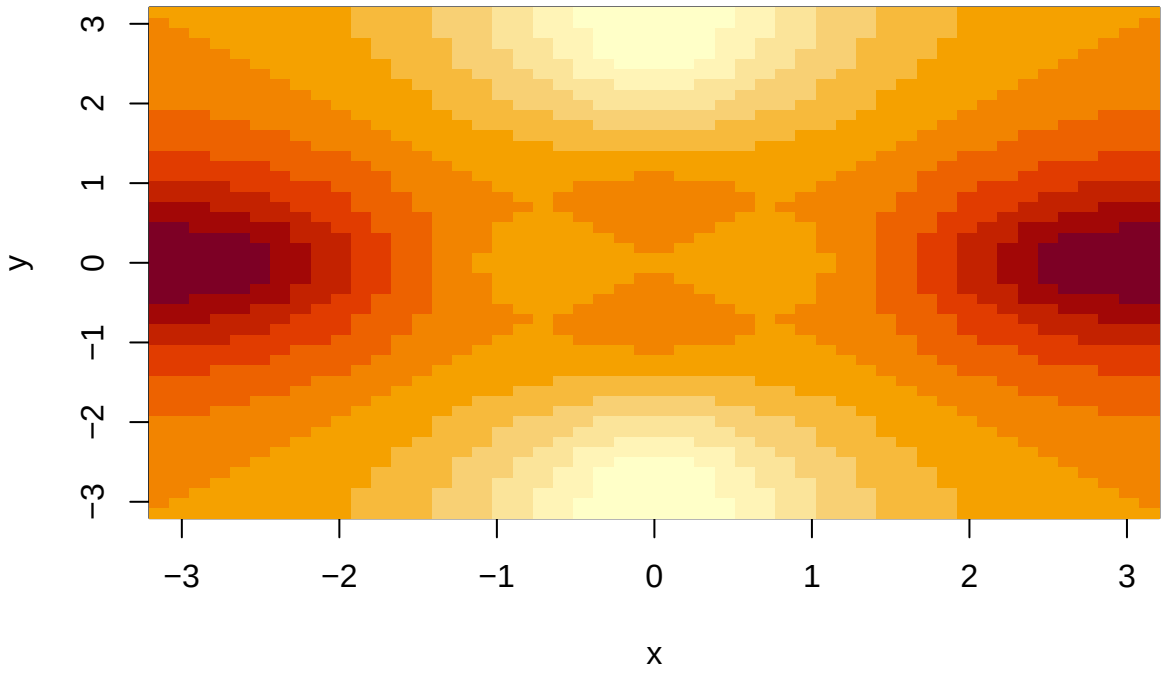
```
y=x
f=outer(x,y,function(x,y)cos(y)/(1+x^2))
contour(x,y,f)
contour(x,y,f,nlevels=45,add=T)
```



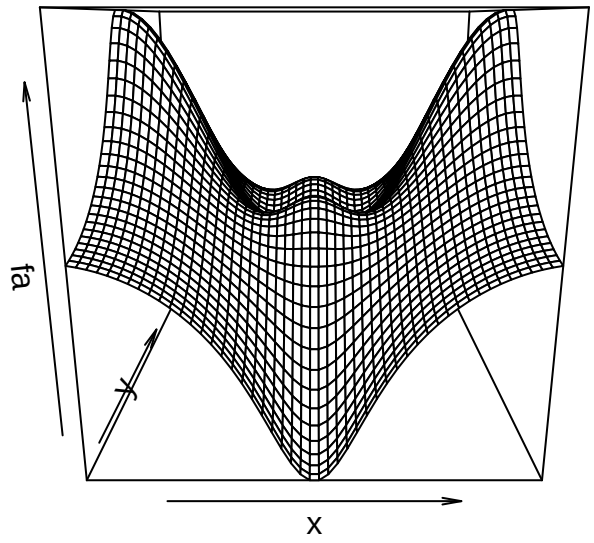
```
fa=(f-t(f))/2
contour(x,y,fa,nlevels=15)
```



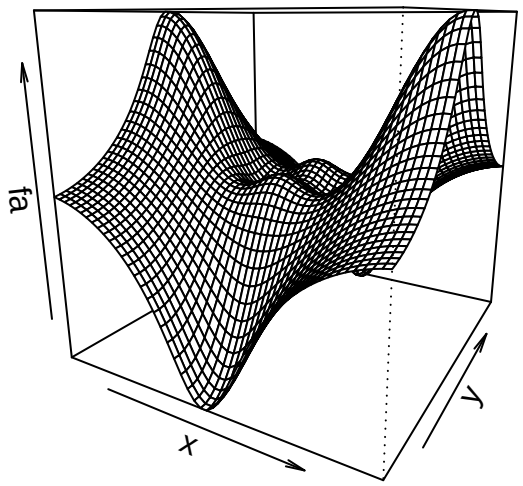

```
image(x,y,fa)
```



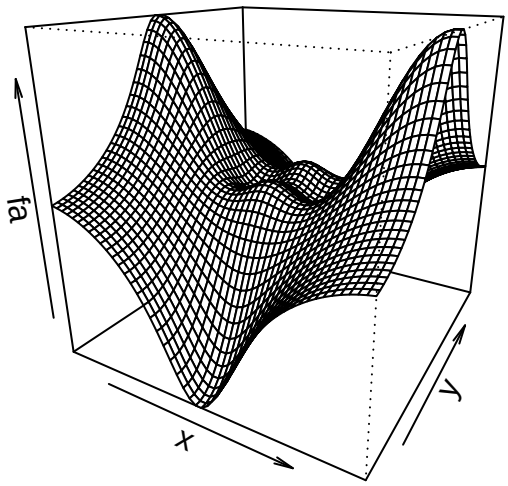

```
persp(x,y,fa)
```



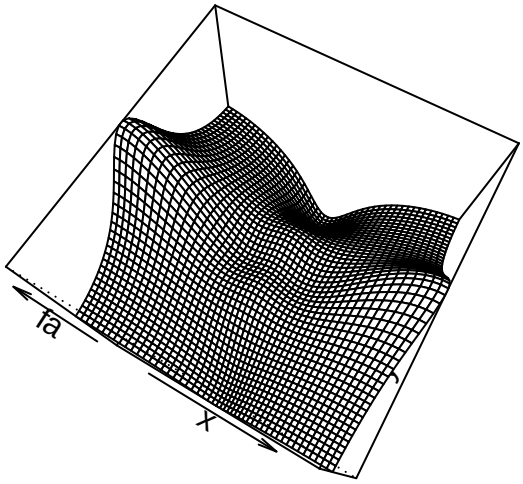

```
persp(x,y,fa,theta=30)
```



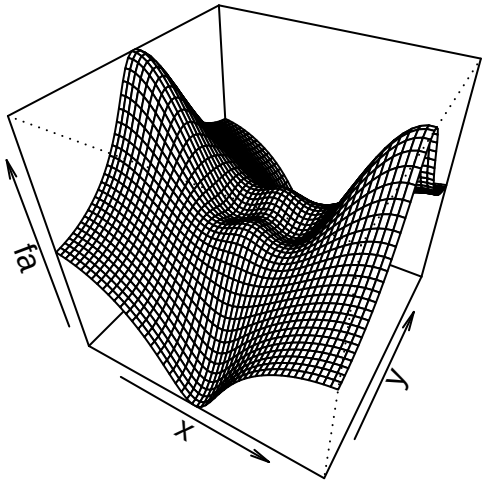

```
persp(x,y,fa,theta=30,phi=20)
```




```
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("Auto.data")
#fix(Auto)
```

```
 Auto=read.table("Auto.data",header=T,na.strings="?")
#fix(Auto)
```

```
Auto=read.csv("Auto.csv",header=T,na.strings="?")
#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
```

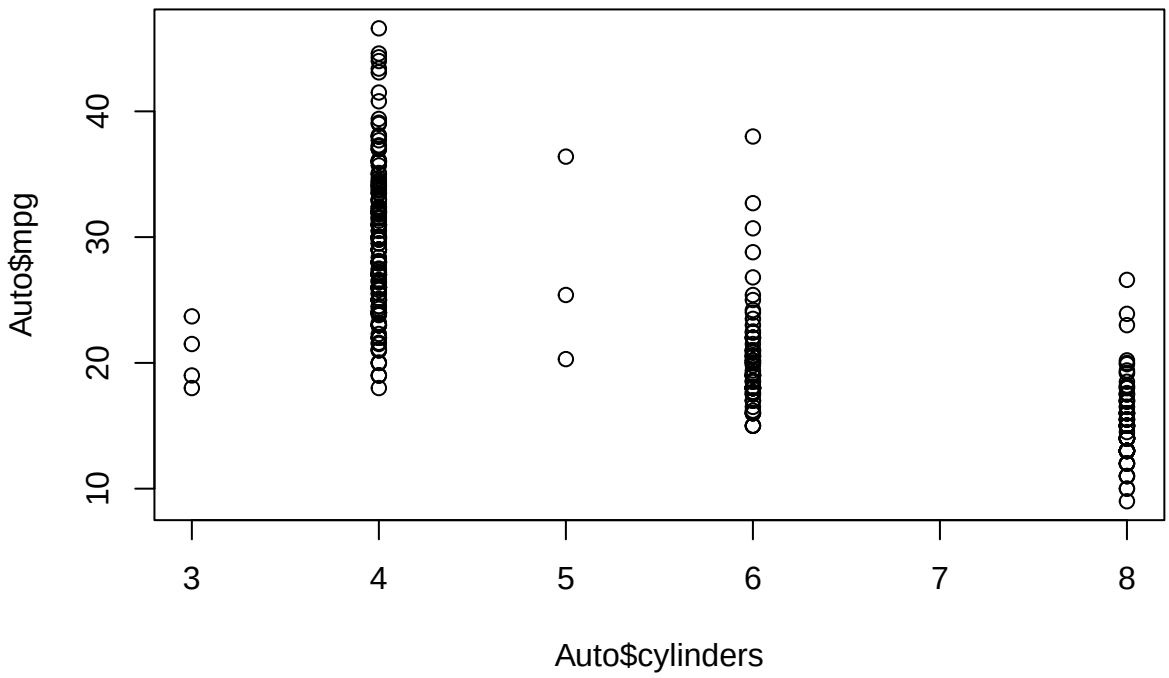
```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

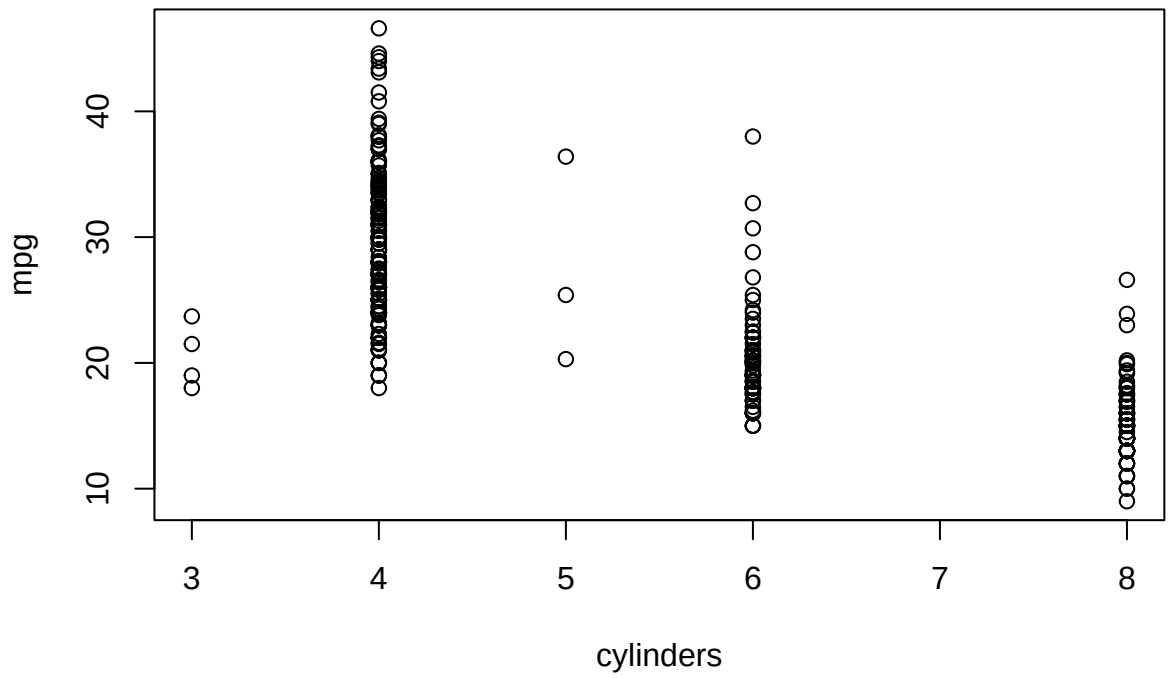
Additional Graphical and Numerical Summaries

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

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

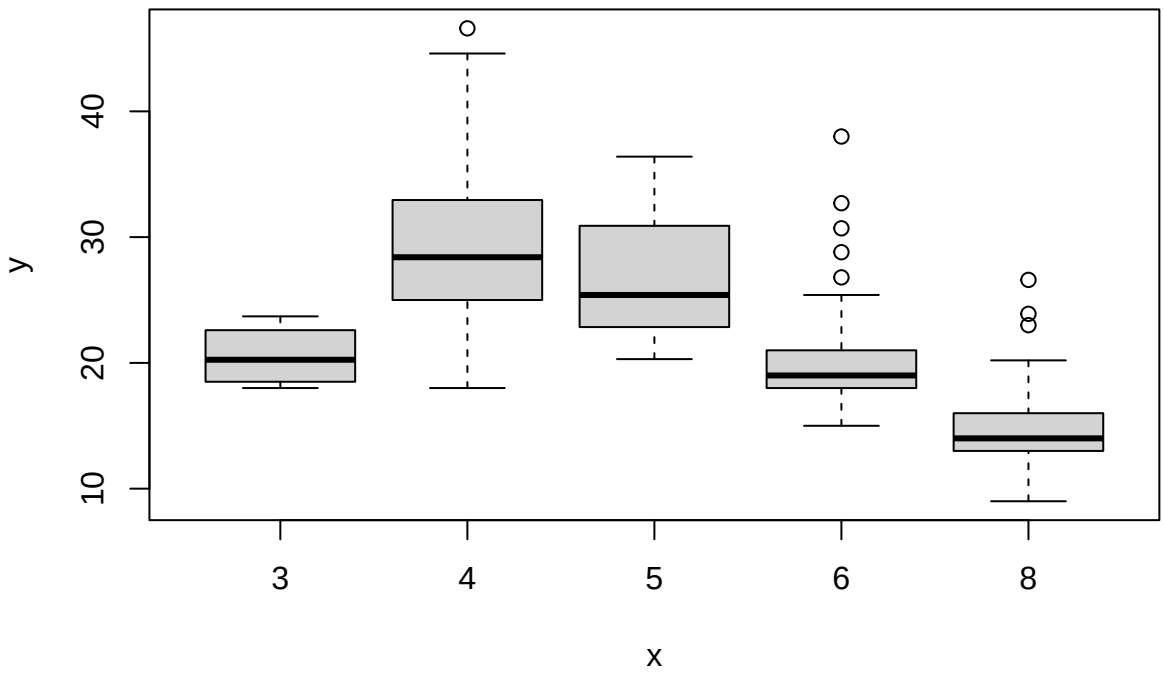



```
attach(Auto)
plot(cylinders, mpg)
```

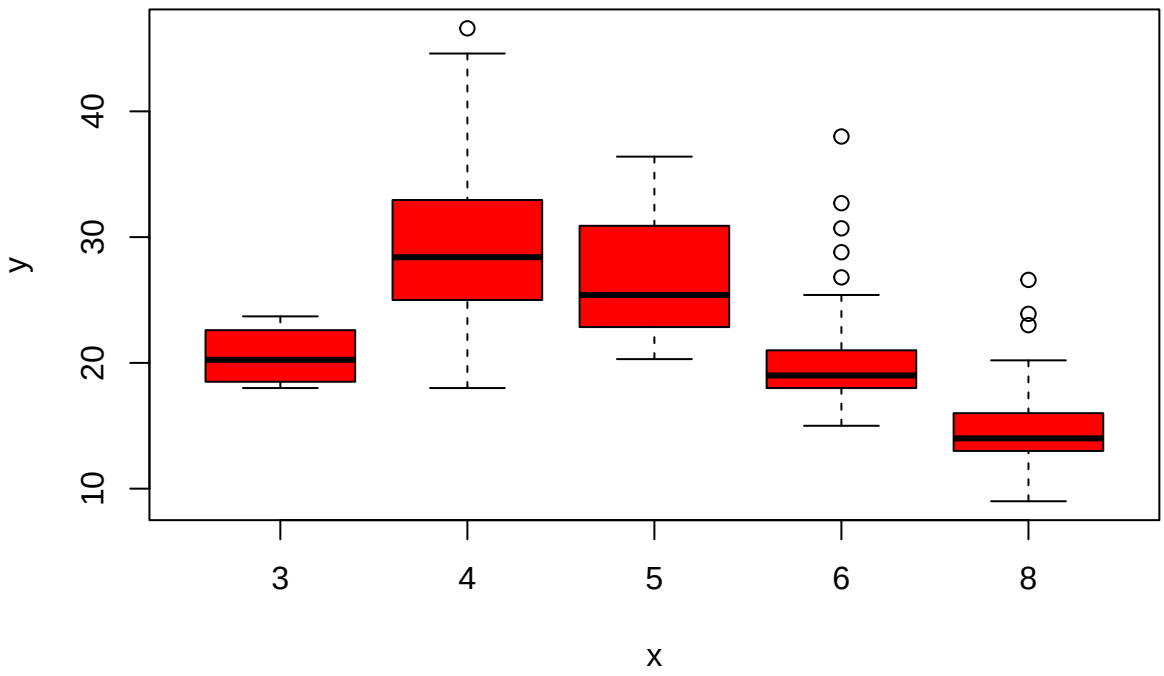



```
cylinders=as.factor(cylinders)
```

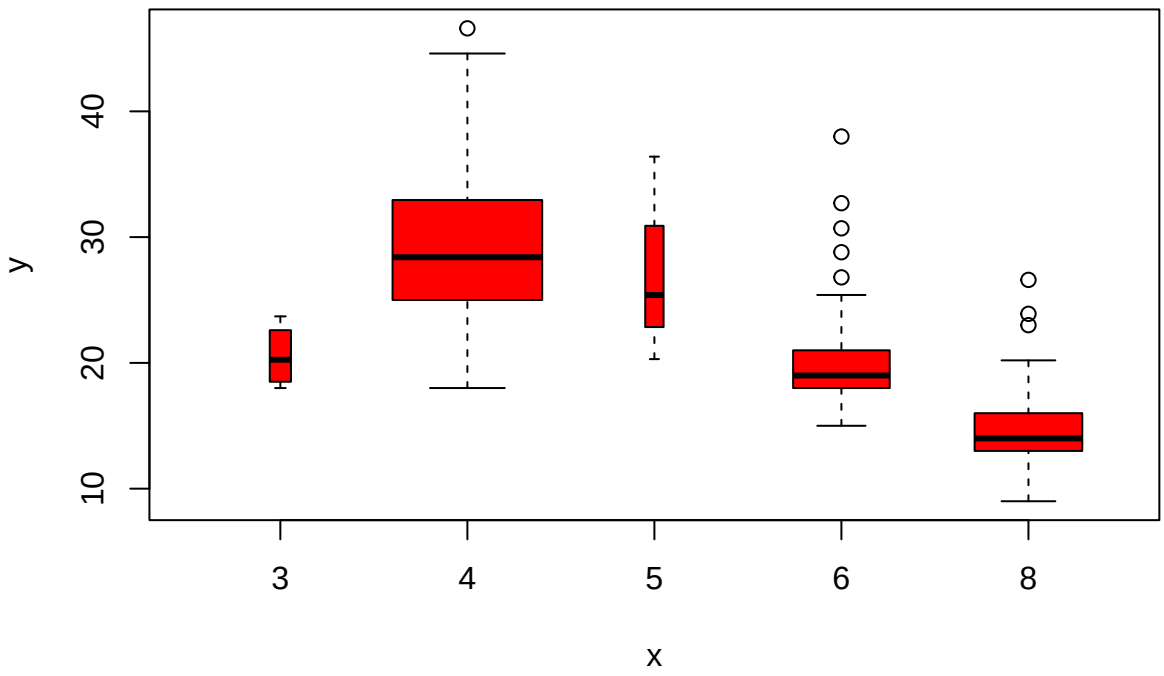
```
plot(cylinders, mpg)
```



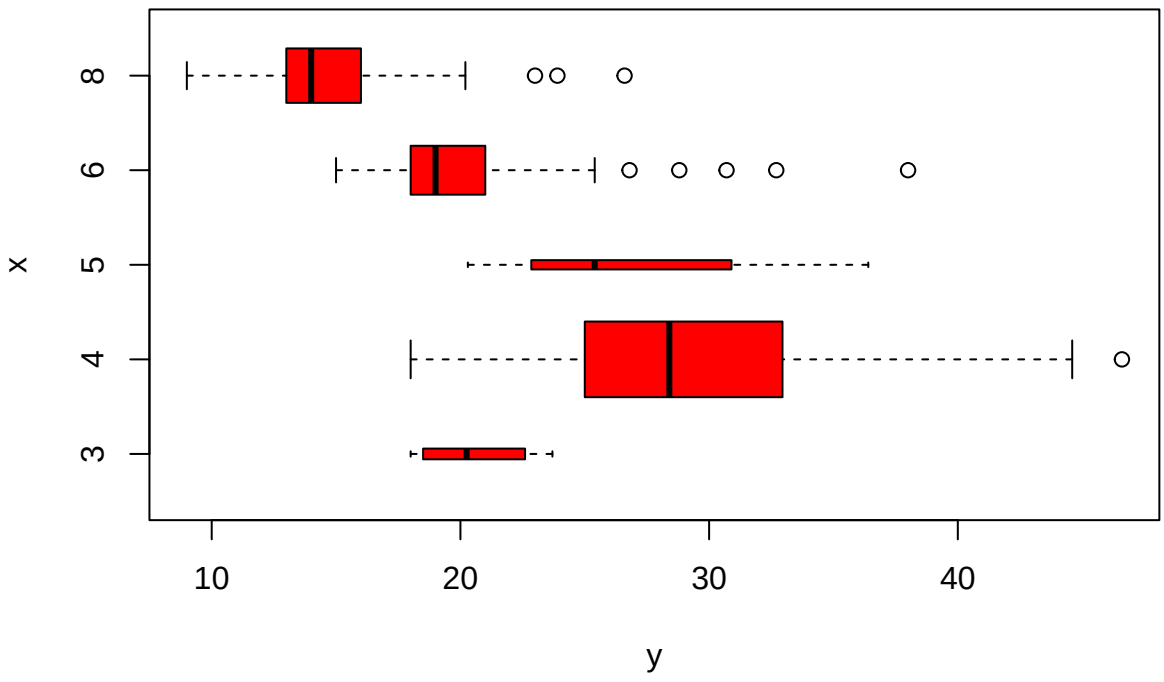

```
plot(cylinders, mpg, col="red")
```



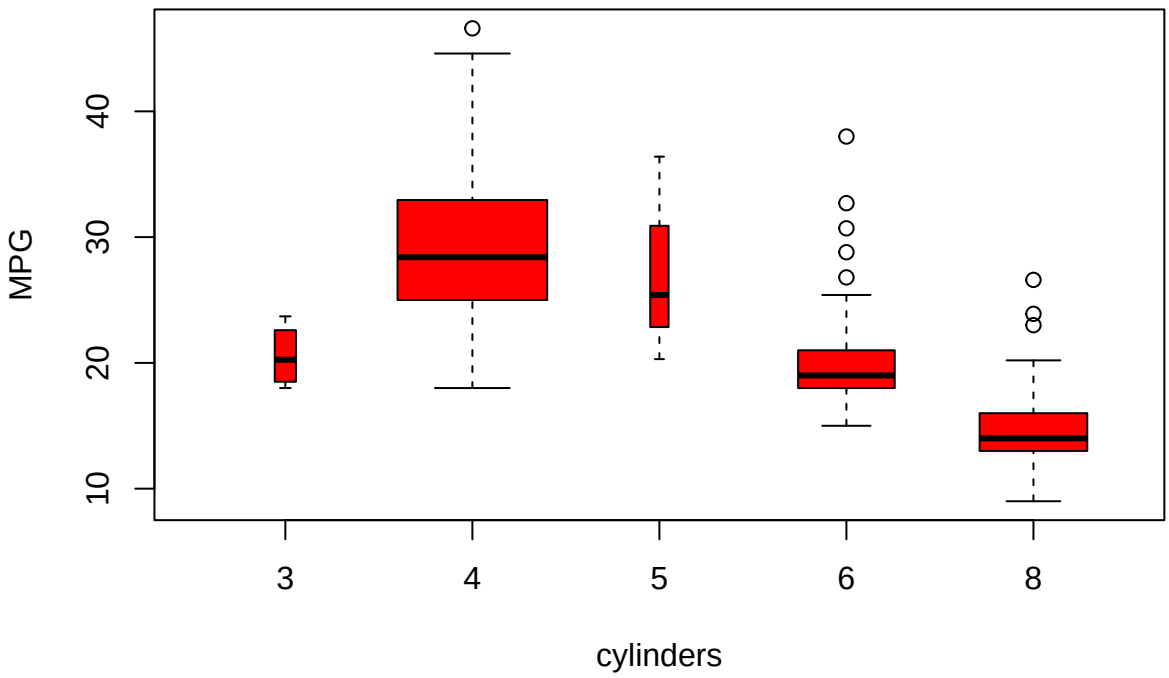

```
plot(cylinders, mpg, col="red", varwidth=T)
```



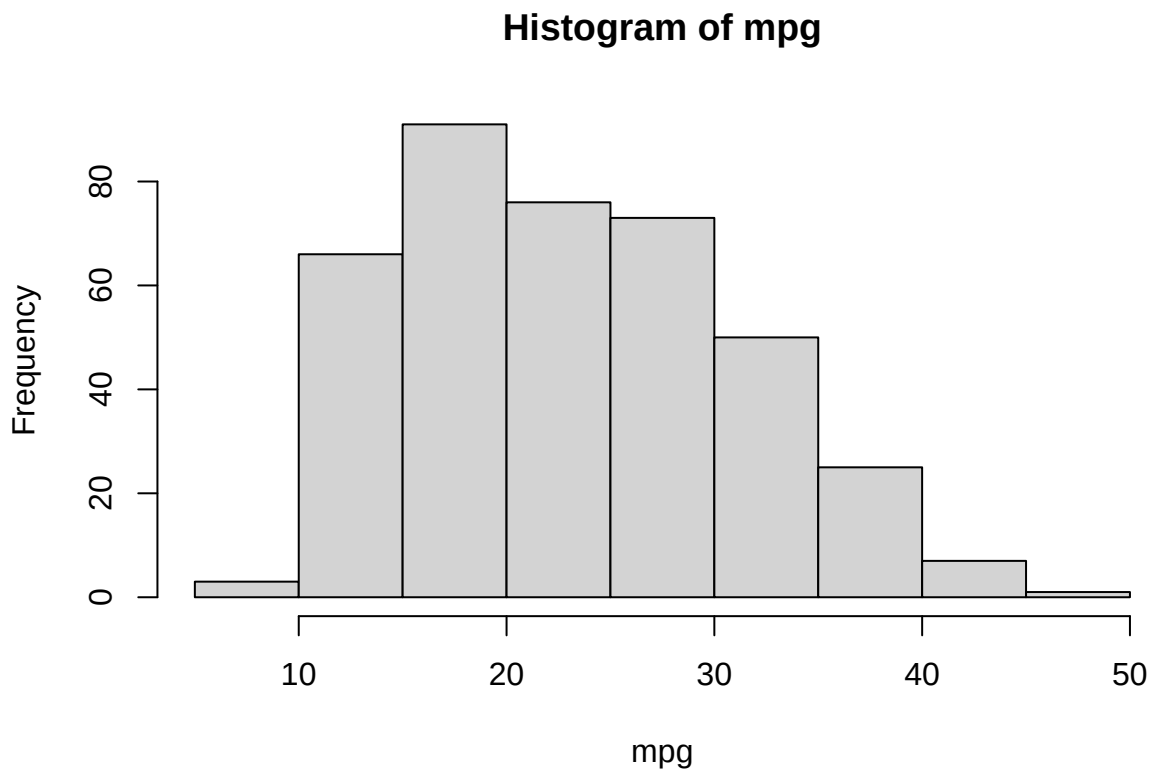

```
plot(cylinders, mpg, col="red", varwidth=T,horizontal=T)
```



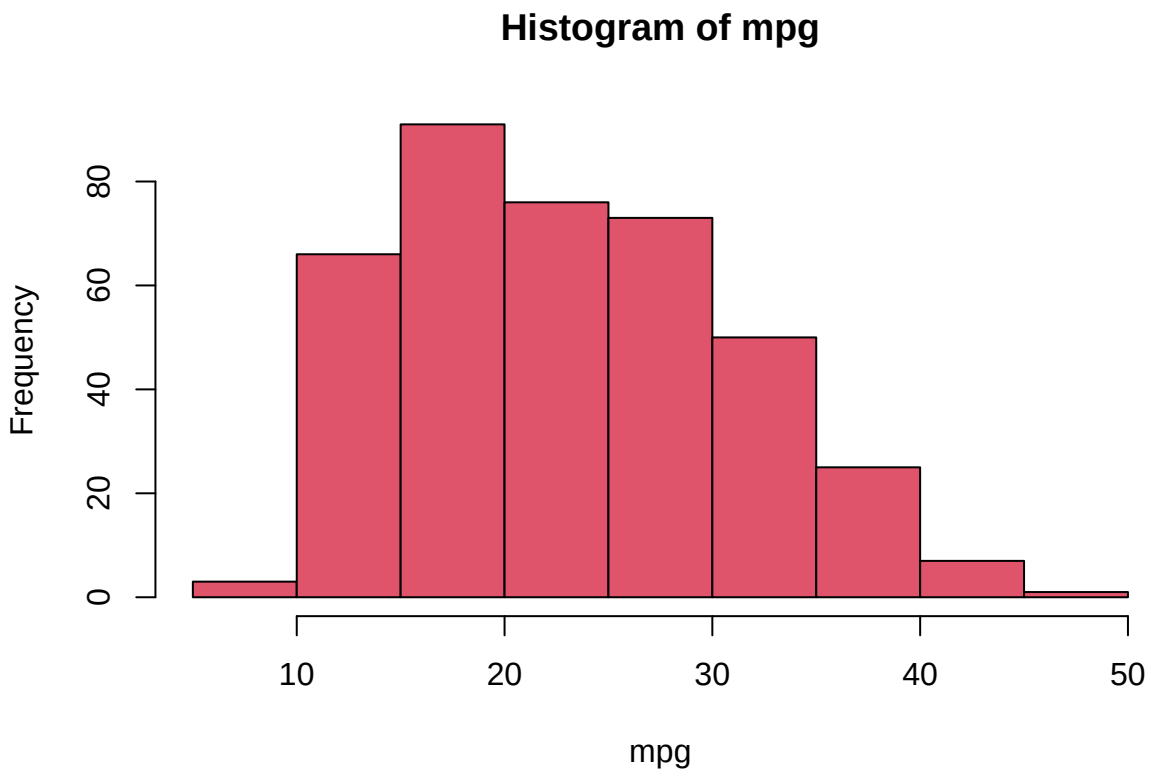

```
plot(cylinders, mpg, col="red", varwidth=T, xlab="cylinders", ylab="MPG")
```



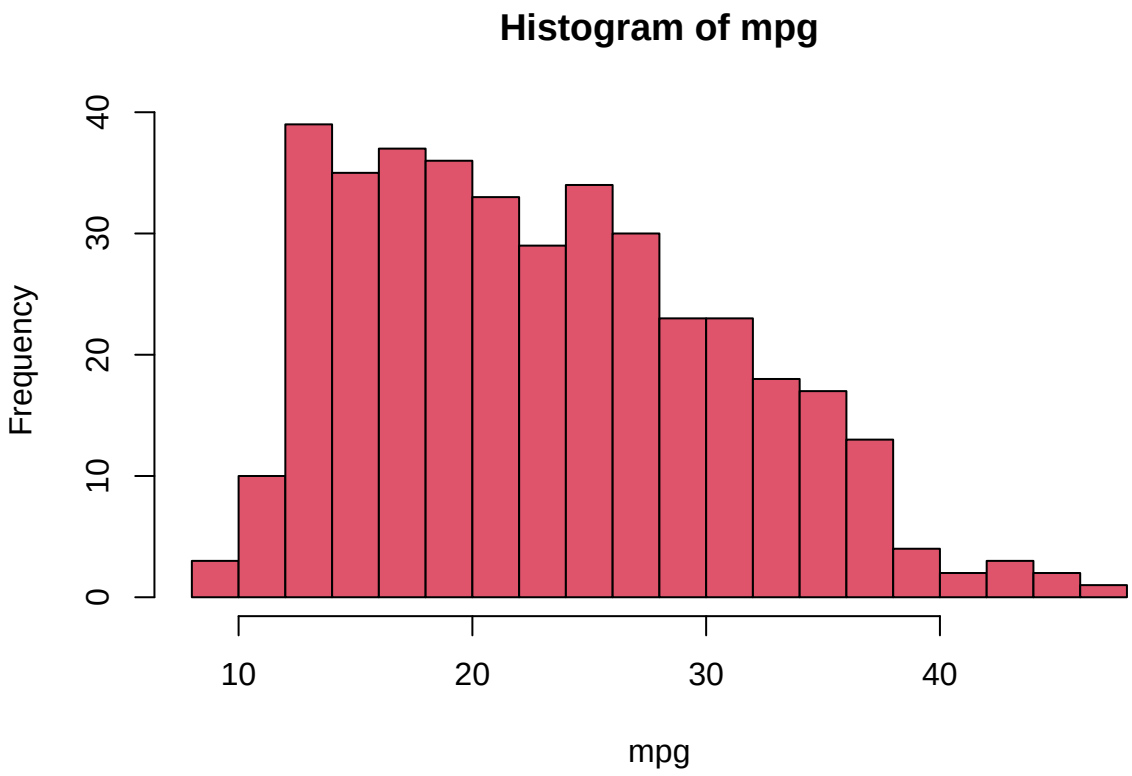

```
hist(mpg)
```




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




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




```
# Check the structure of Auto dataset
str(Auto)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel
##  - attr(*, "na.action")= 'omit' Named int [1:5] 33 127 331 337 355
##   ..- attr(*, "names")= chr [1:5] "33" "127" "331" "337" ...
```

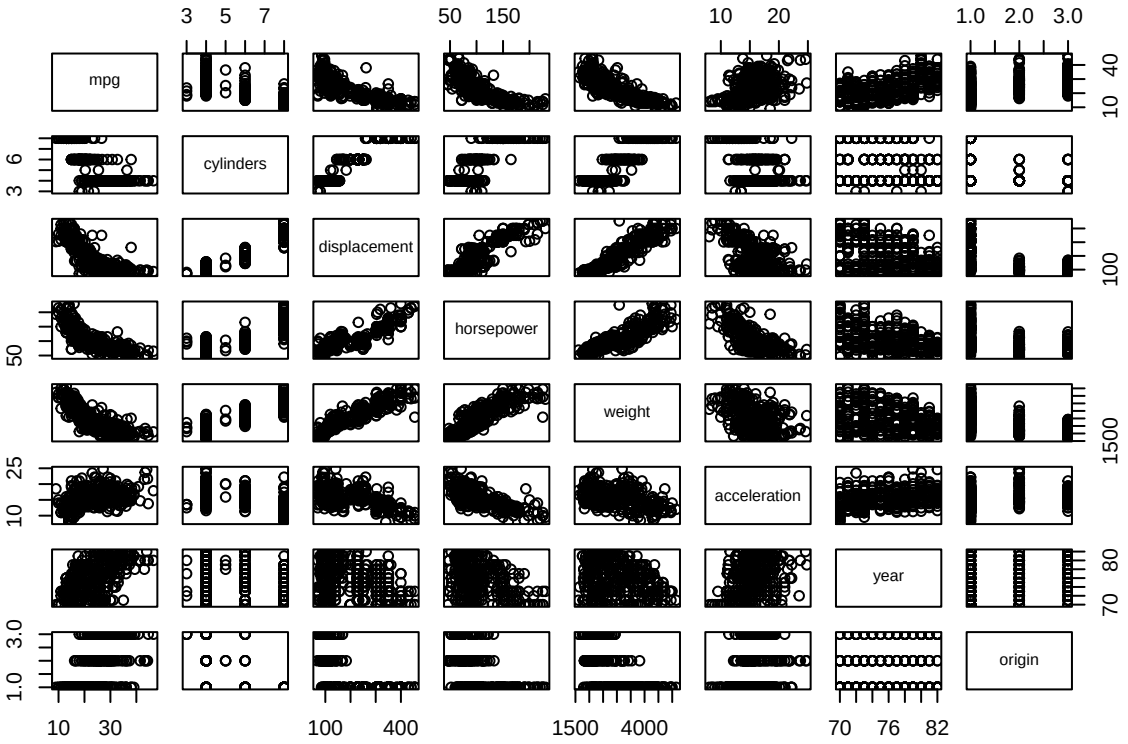
```
# Subset Auto dataset with only numeric columns
numeric_auto <- Auto[, sapply(Auto, is.numeric)]

# Check the structure of numeric_auto
str(numeric_auto)
```

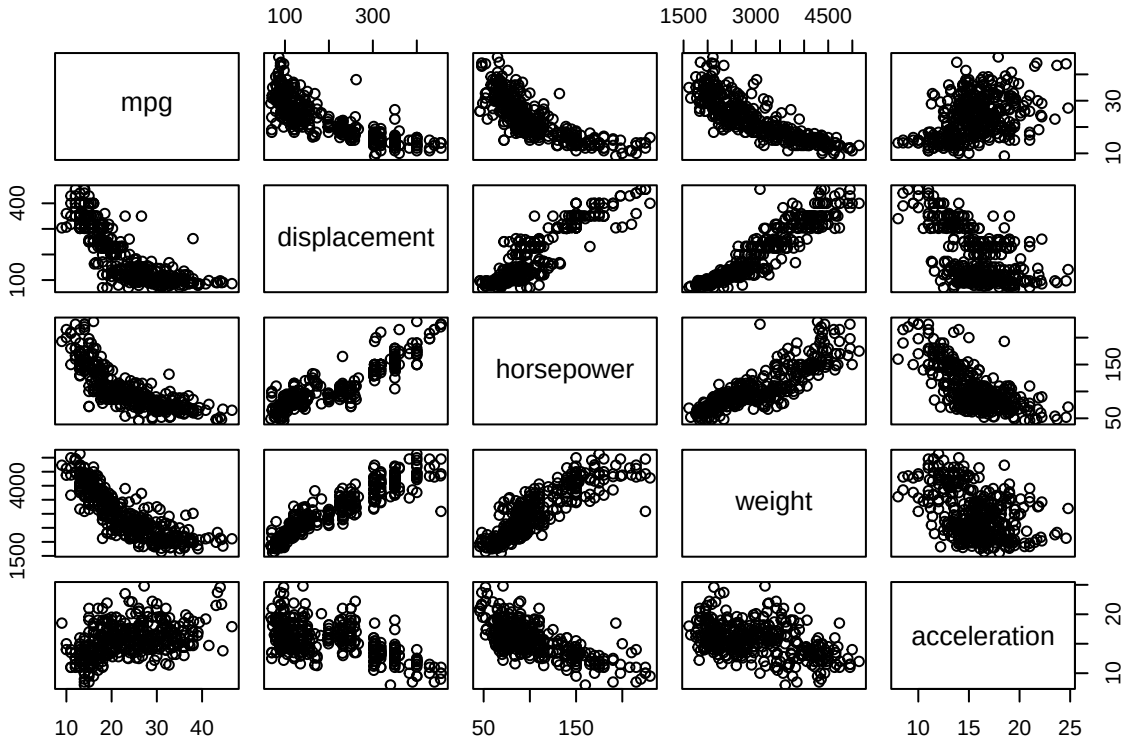
```
## 'data.frame':    392 obs. of  8 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
```

```
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
```

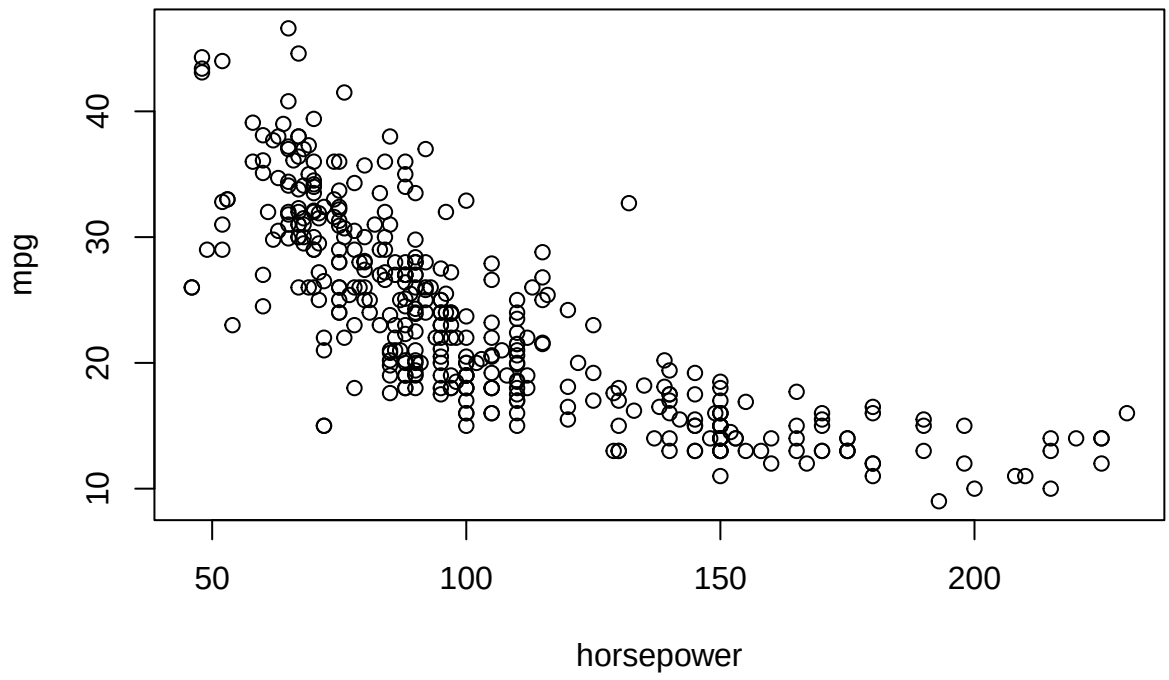
```
pairs(numeric_auto)
```




```
pairs(~ mpg + displacement + horsepower + weight + acceleration, numeric_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.

```
library(MASS)

library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.3.3
```

```
##
## Attaching package: 'ISLR'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     Auto
```

```
college<-read.csv("college.csv")
```

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.

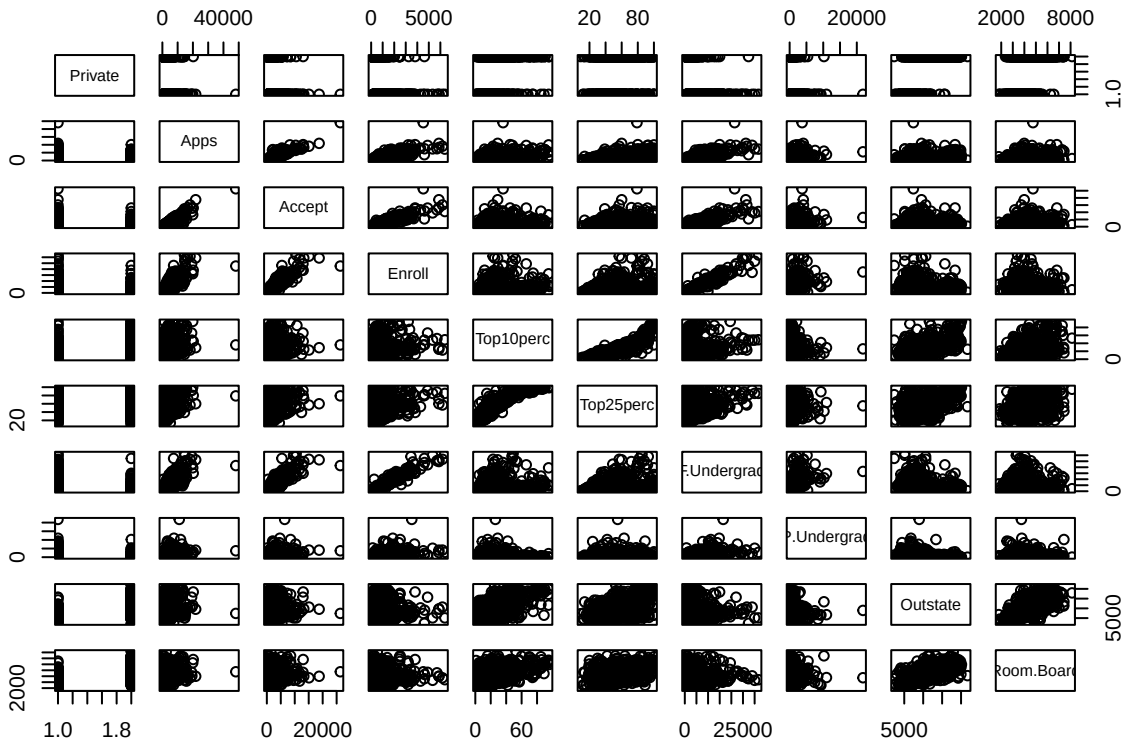
`Use the 'summary()' function to produce a numerical summary of the variables in the data set.`

```
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]`.

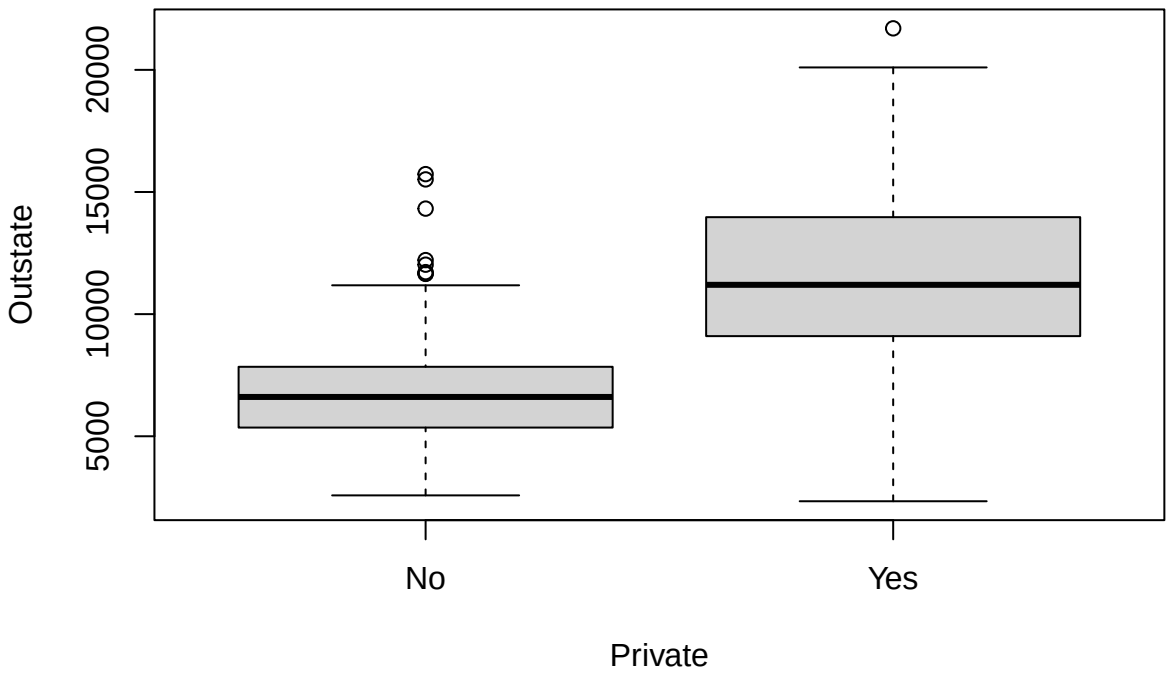
```
college$Private <- as.factor(college$Private)
pairs(college[,1:10])
```




```
#we first convert the private variable of the college dataset into categorical data, and once that is d
```

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

```
plot(college$Private, college$Outstate,
     xlab = "Private", ylab = "Outstate")
```




```
#We are plotting the graph of Outstate versus Private as requested. This graph compares the tuition fee
```

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)
```

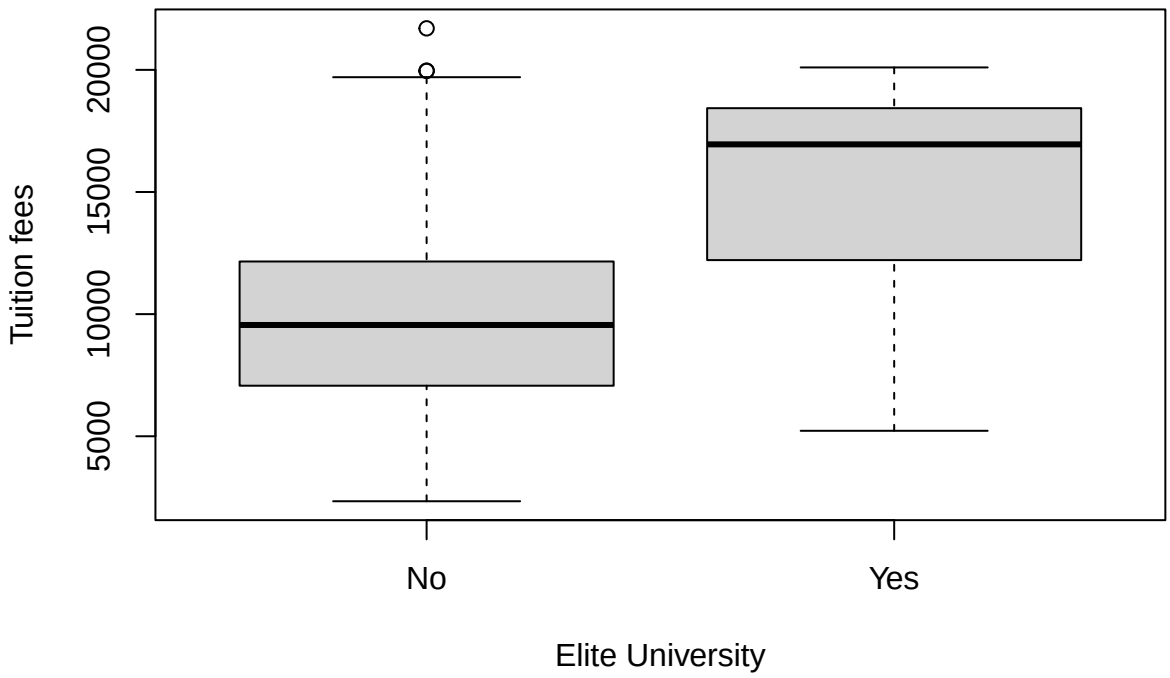
```
#We start by creating a vector called 'Elite' with the same number of rows as our college dataset, fill
```

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(Elite)
```

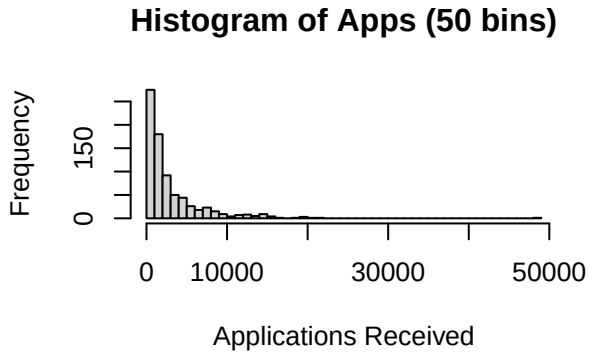
```
##  No Yes
## 699  78
```

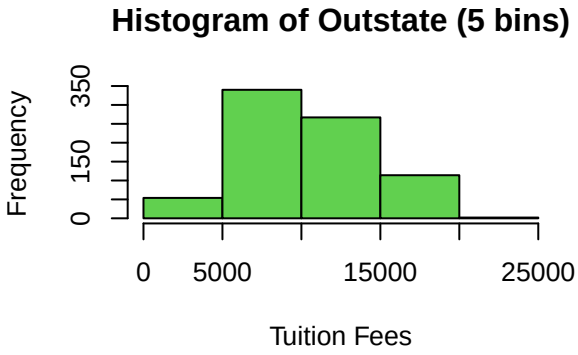
```
plot(college$Elite, college$Outstate,
     xlab = "Elite University", ylab = "Tuition fees")
```

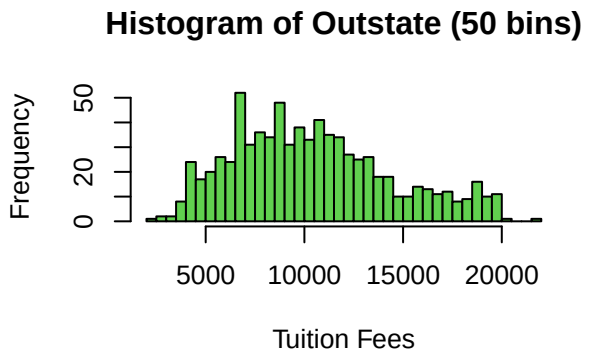


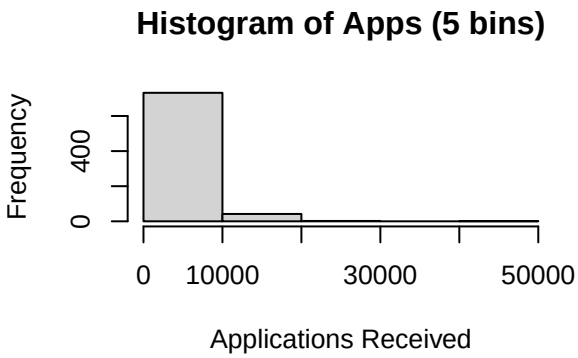
Use the `hist()` function to produce some histograms with 5 and 50 bins for `Outstate` and `Apps`. Use the command `par(mfrow=c(2,2))`: it will divide the print window into four regions so that four plots can be made simultaneously.

```
par(mfrow=c(2,2))
hist(college$Apps, xlab = "Applications Received",breaks= 50,main = "Histogram of Apps (50 bins)")
hist(college$Outstate, col=3, xlab = "Tuition Fees",breaks=5, main = "Histogram of Outstate (5 bins)")
hist(college$Outstate, col=3, xlab = "Tuition Fees",breaks=50, main = "Histogram of Outstate (50 bins)"
hist(college$Apps, xlab = "Applications Received",breaks= 5,main = "Histogram of Apps (5 bins)")
```









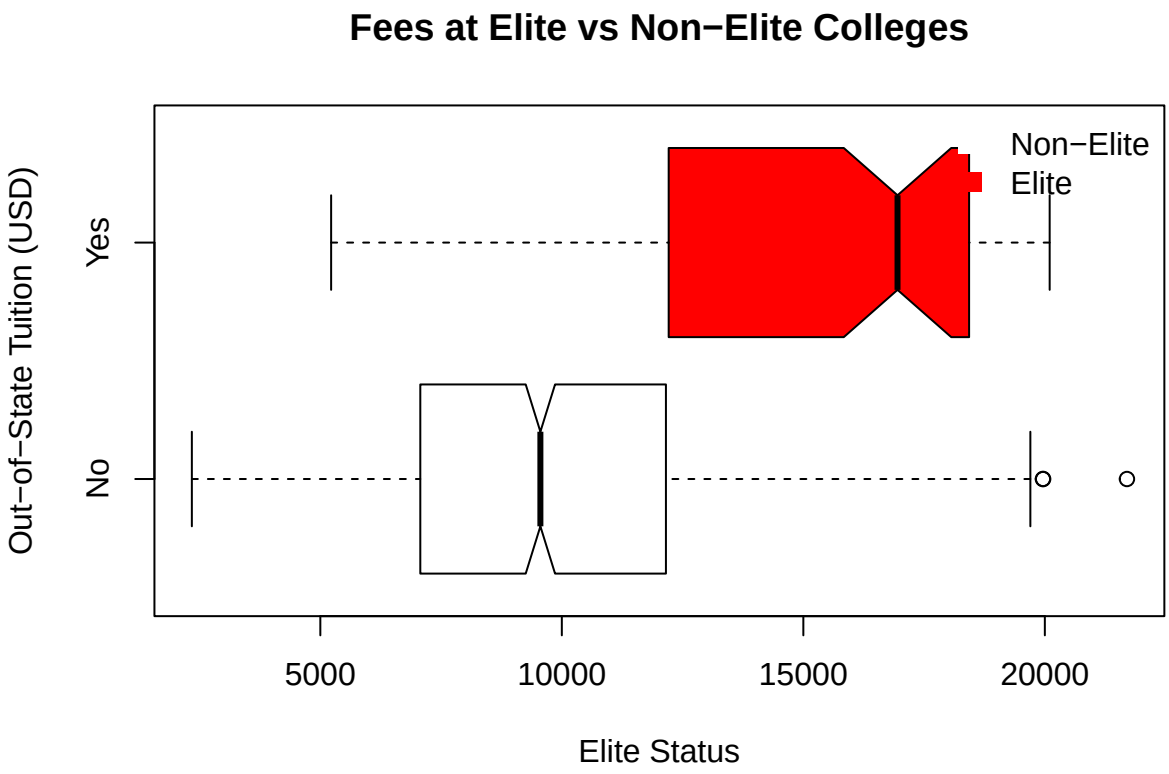
Continue exploring the data, and provide a brief summary of what you discover.

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* ?

```
boxplot(college$Outstate ~ college$Elite,horizontal = TRUE,
        xlab = "Elite Status", ylab = "Out-of-State Tuition (USD)",
        main = "Fees at Elite vs Non-Elite Colleges",
        col = c("white", "red"), # specify the colors for the boxes
        # specify the colors for the box borders
        notch = TRUE) # add notches to the boxes
legend("topright", legend=c("Non-Elite", "Elite"), fill=c("white", "red"),
       border=FALSE, bty="n", box.lty=0)
```




```
summary_hypo1 <- aggregate(Outstate ~ Elite, data = college, FUN = summary)
print(summary_hypo1)
```

```
##   Elite Outstate.Min. Outstate.1st Qu. Outstate.Median Outstate.Mean
## 1    No      2340.000         7070.000        9556.000      9904.166
## 2   Yes      5224.000        12219.000       16950.000     15248.564
##   Outstate.3rd Qu. Outstate.Max.
## 1        12155.000     21700.000
## 2        18411.500     20100.000
```

```
#The results of Hypothesis 1 indicate a noticeable difference in median fees between elite and non-elit
```

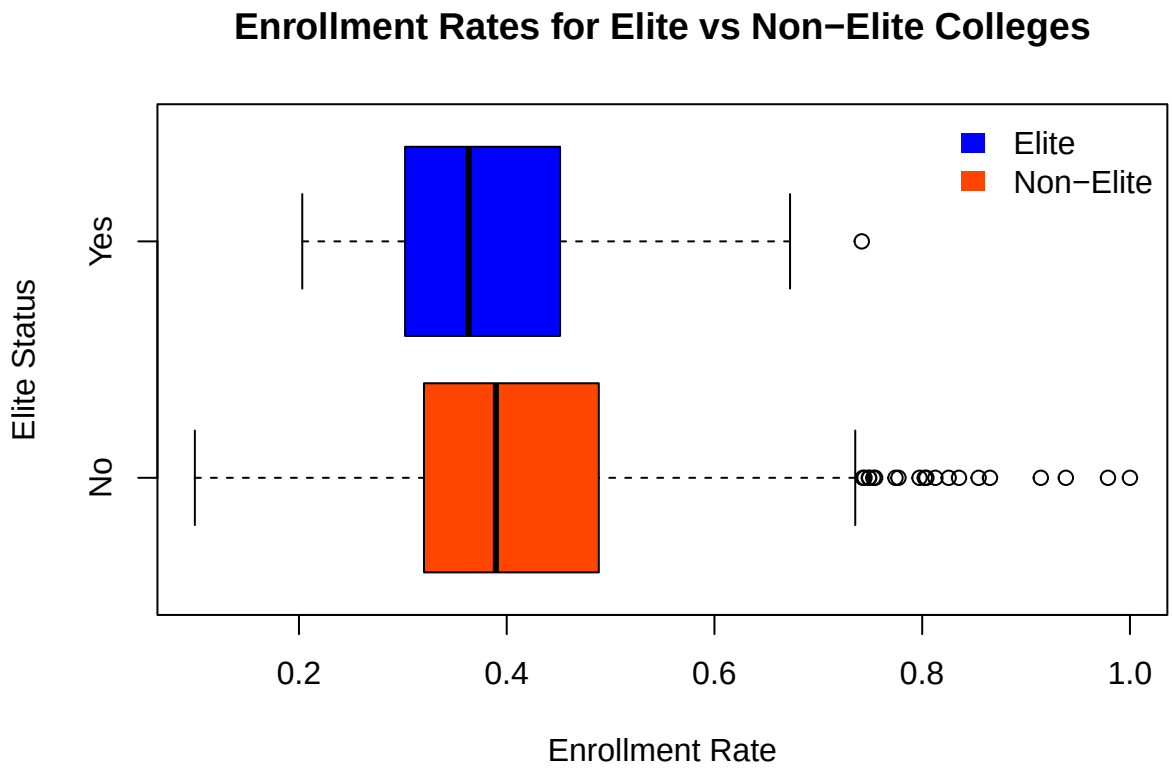
The enrollment rate is the fraction of accepted students who enrolled. *Hypothesis 2. The enrollment rate at the best colleges, as indicated by* ***Elite****, is higher than that at other colleges.* **Methodology:** Create a new variable called `EnrollRate`, using `Enroll` and `Accept`. Use the `attach()` 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?

```
college$EnrollRate <- college$Enroll/college$Accept
# Attach data
attach(college)
```

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

```
##
##     Elite
```

```
# Plot boxplots
boxplot(college$EnrollRate ~ college$Elite, horizontal = TRUE,
        col = c("#FF4500", "blue"),
        xlab = "Enrollment Rate", ylab = "Elite Status",
        main = "Enrollment Rates for Elite vs Non-Elite Colleges")
legend("topright", legend=c("Elite", "Non-Elite"), fill=c("blue", "#FF4500"),
       border=FALSE, bty="n", box.lty=0)
```



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?

```
college$AppsPerEnrolled <- college$Apps / college$Enroll
attach(college)
```

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

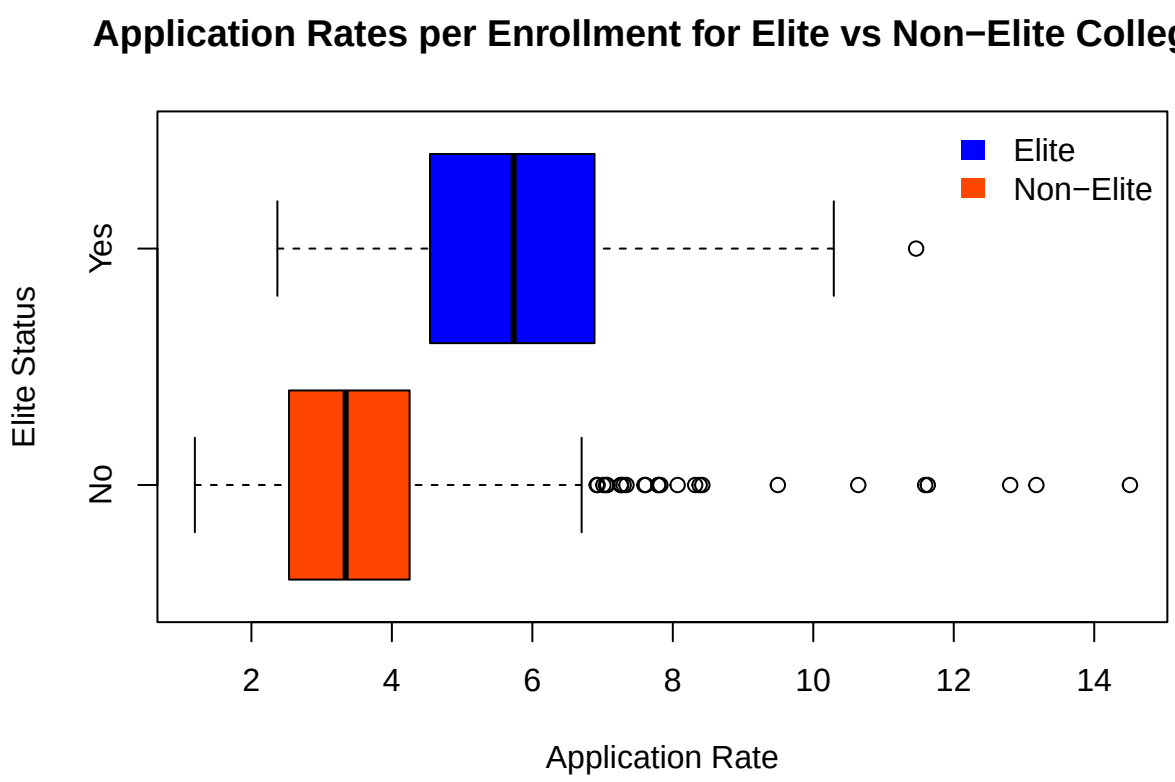
```
## The following objects are masked from college (pos = 3):
```

```
##
##     Accept, Apps, Books, Elite, Enroll, EnrollRate, Expend,
##     F.Undergrad, Grad.Rate, Outstate, P.Undergrad, perc.alumni,
##     Personal, PhD, Private, Room.Board, S.F.Ratio, Terminal, Top10perc,
##     Top25perc
```

```
# Check the first few rows again
head(college)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate Elite EnrollRate AppsPerEnrolled
## Abilene Christian University        60    No  0.5852273        2.302358
## Adelphi University                  56    No  0.2661123        4.269531
## Adrian College                      54    No  0.3062899        4.250000
## Agnes Scott College                 59   Yes  0.3925501        3.043796
## Alaska Pacific University           15    No  0.3767123        3.509091
## Albertson College                   55    No  0.3298539        3.715190
```

```
# Plot boxplots
boxplot(college$AppsPerEnrolled ~ college$Elite, horizontal = TRUE,
        col = c("#FF4500", "blue"),
        xlab = "Application Rate", ylab = "Elite Status",
        main = "Application Rates per Enrollment for Elite vs Non-Elite Colleges")
legend("topright", legend=c("Elite", "Non-Elite"), fill=c("blue", "#FF4500"),
       border=FALSE, bty="n", box.lty=0)
```

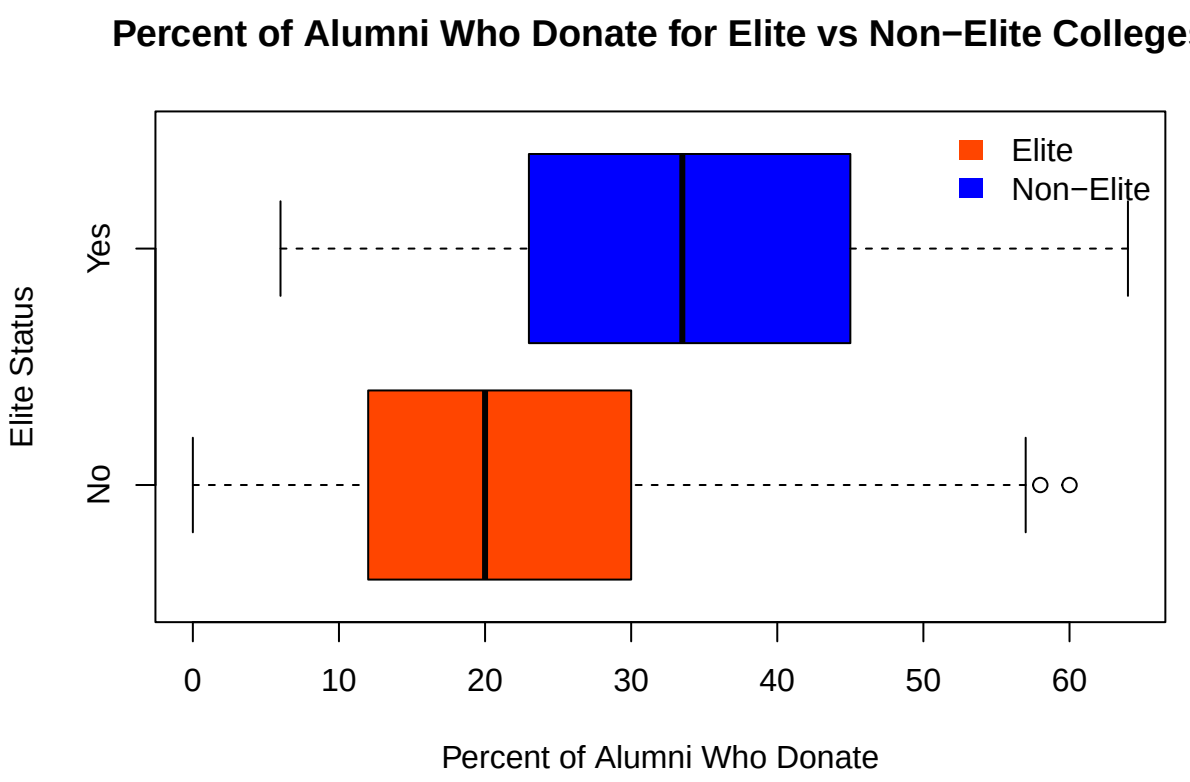



```
#The results of Hypothesis 3 are in line with expectations. It is evident that the applications per enr

#The median for non-elite colleges would have been considerably lower had it not been for the presence
```

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* ?

```
boxplot(college$perc.alumni ~ college$Elite, horizontal = TRUE,
        col = c("#FF4500", "blue"),
        xlab = "Percent of Alumni Who Donate", ylab = "Elite Status",
        main = "Percent of Alumni Who Donate for Elite vs Non-Elite Colleges")
legend("topright", legend=c("Elite", "Non-Elite"), fill=c("#FF4500", "blue"),
       border=FALSE, bty="n", box.lty=0)
```




```
#The boxplot analysis clearly illustrates that alumni donations are significantly higher for elite coll

#Moreover, the presence of a few higher-end outliers in the non-elite group further accentuates the dis
```

Hypothesis 5. Total other miscellaneous expenses other than tuition made by students in elite colleges are higher than the expenses made by non-elite group.

```
college$Totalexpenses <- college$Room.Board+college$Books+college$Personal
# Attach data
attach(college)
```

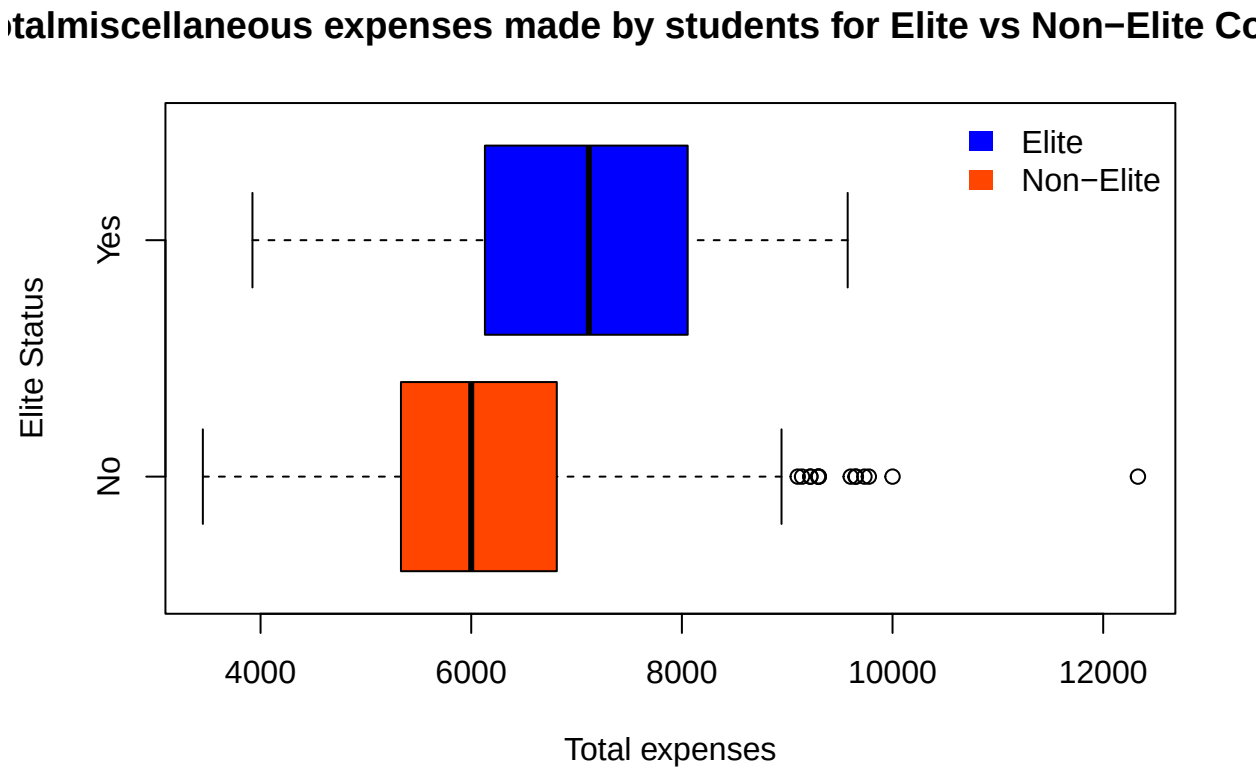
```
## The following object is masked _by_ .GlobalEnv:
##
##     Elite

## The following objects are masked from college (pos = 3):
##
##     Accept, Apps, AppsPerEnrolled, Books, Elite, Enroll, EnrollRate,
##     Expend, F.Undergrad, Grad.Rate, Outstate, P.Undergrad, perc.alumni,
##     Personal, PhD, Private, Room.Board, S.F.Ratio, Terminal, Top10perc,
##     Top25perc

## The following objects are masked from college (pos = 4):
##
```

```
##     Accept, Apps, Books, Elite, Enroll, EnrollRate, Expend,
##     F.Undergrad, Grad.Rate, Outstate, P.Undergrad, perc.alumni,
##     Personal, PhD, Private, Room.Board, S.F.Ratio, Terminal, Top10perc,
##     Top25perc
```

```
# Plot boxplots
boxplot(college$Totalexpenses ~ college$Elite, horizontal = TRUE,
        col = c("#FF4500", "blue"),
        xlab = "Total expenses", ylab = "Elite Status",
        main = "Totalmiscellaneous expenses made by students for Elite vs Non-Elite Colleges")
legend("topright", legend=c("Elite", "Non-Elite"), fill=c("blue", "#FF4500"),
       border=FALSE, bty="n", box.lty=0)
```




```
#The boxplot visualization distinctly illustrates that the total expenses of students in elite colleges

#This disparity would likely have been even more pronounced, but the presence of outliers in the non-el
```