Stat 107: Introduction to Business and Financial Statistics
Homework 1: Due Monday, Sept 12

Xiner Zhou

September 11, 2016

Note-much of this homework has you working with R. We will be covering the basics of R in week 2 of class and there are several guides on the course website we recommend for R including

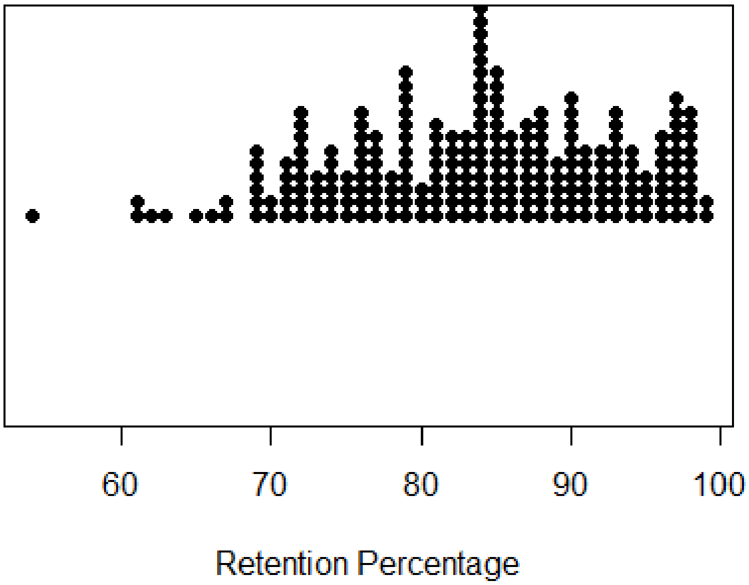
- The R Cookbook
- R by Example
- Best First R Tutorial
- Datacamp (https://www.datacamp.com/courses/free-introduction-to-r)
- CodeSchool (https://www.codeschool.com/courses/try-r)

1) This problem is absolutely goofy; we want you to read about R and start using it, but how does one force you to do that? By reproducing output of course. So for this problem we want you to reproduce Figures 5.1-5.5 in Chapter 5 of the book R by Example.

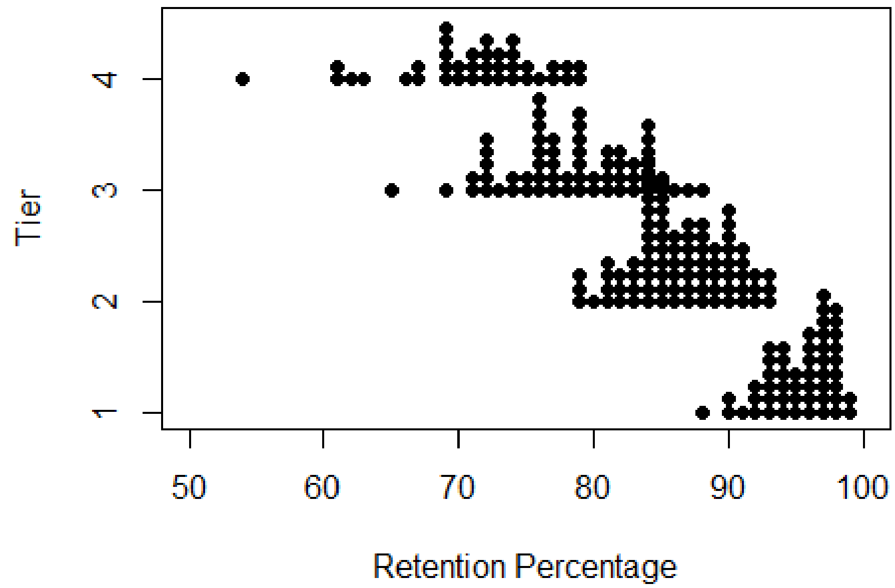
Note that we want you to read in the data set in a different way than at the start of Chapter 5. We will make your life easy by helping you get the data into R. Once you load up R, enter the following command to load the data
dat=read.csv("http://people.fas.harvard.edu/~mparzen/stat107/college.csv")

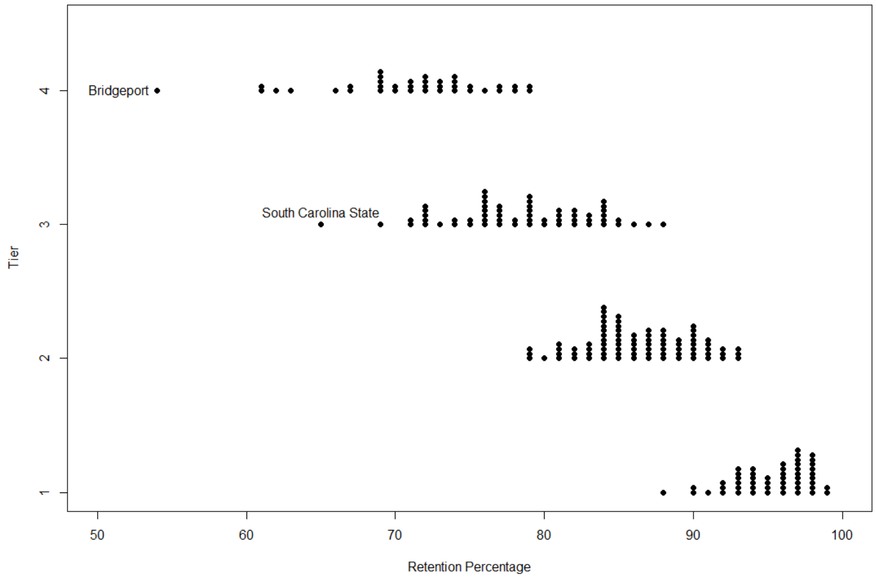
The deliverable for this question is a neat and organized print out (cut and paste the graphs) of figures 5.1-5.5 from the book R by Example.

```r
# read in data
dat<-read.csv("http://people.fas.harvard.edu/~mparzen/stat107/college.csv")
# subset to complete cases
college<-subset(dat,complete.cases(dat))

# re-produce figure 5.1
stripchart(college$Retention, method="stack", pch=19, xlab="Retention Percent
age")
```

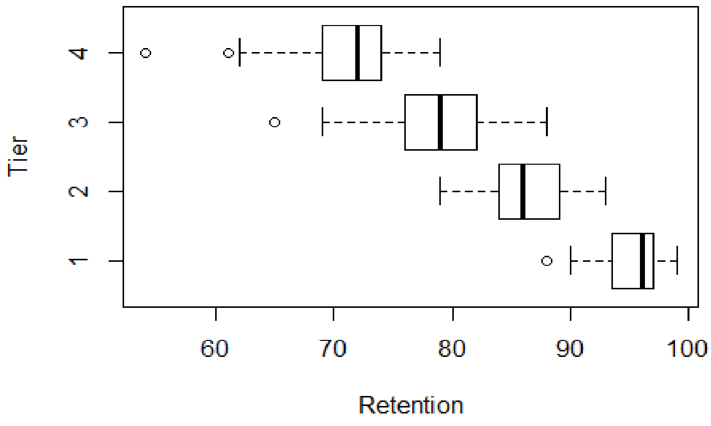


```r
# re-produce figure 5.2
stripchart(Retention~Tier, method="stack", pch=19, xlab="Retention Percentage
", ylab="Tier", xlim=c(50, 100), data=college)

# re-produce figure 5.3
identify(college$Retention, college$Tier, n=2, labels=college$School)
```

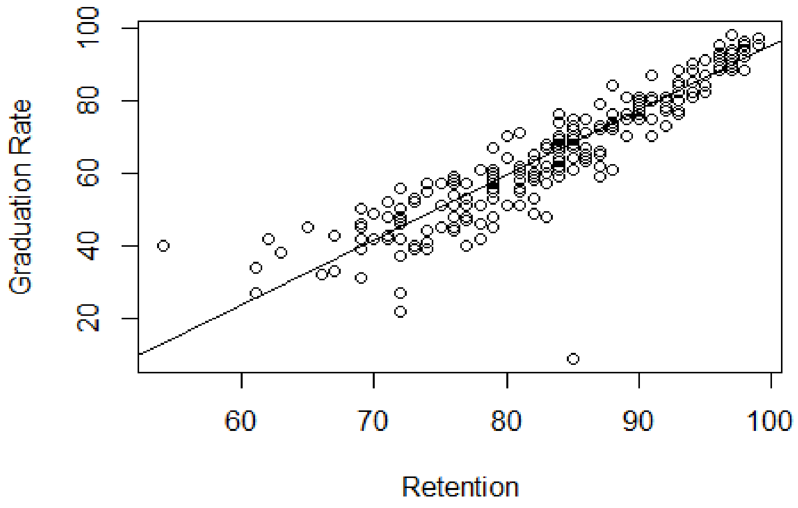




```
# re-produce figure 5.4
boxplot(Retention ~ Tier, data=college, horizontal=TRUE, ylab="Tier", xlab="R
etention")
```



```
# re-produce figure 5.5
plot(college$Retention, college$Grad.rate, xlab="Retention", ylab="Graduation
Rate")
fit<-line(college$Retention, college$Grad.rate)
coef(fit)

## [1] -83.657895   1.789474

abline(coef(fit))
```



2) Read the handout on the website 'BestFirstRTutorial' and use R to do the following problems. Neatly write up your solution to each part, clearly showing how R was used and the resulting output. Highlight the answer to each question so it is easy to find in your output.

a) Problems 1 and 2 on page 4

```
x<-c(4,2,6)
y<-c(1,0,-1)
length(x)

## [1] 3

sum(x)

## [1] 12

sum(x^2)

## [1] 56

x+y

## [1] 5 2 5

x*y

## [1]  4  0 -6

x-2

## [1] 2 0 4

x^2

## [1] 16  4 36

7:11

## [1]  7  8  9 10 11

seq(2,9)

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

seq(4,10,by=2)

## [1]  4  6  8 10

seq(3,30,length=10)

##  [1]  3  6  9 12 15 18 21 24 27 30

seq(6,-4,by=-2)

## [1]  6  4  2  0 -2 -4
```

b) Problem 1 page 5

```
x<- c(5,9,2,3,4,6,7,0,8,12,2,9)
x[2]

## [1] 9

x[2:4]

## [1] 9 2 3

x[c(2,3,6)]

## [1] 9 2 6

x[c(1:5,10:12)]

## [1]  5  9  2  3  4 12  2  9

x[-(10:12)]

## [1] 5 9 2 3 4 6 7 0 8
```

c) Problem 1 parts a and b page 8

```
x<-matrix(c(3,-1,2,1),ncol=2)
y<-matrix(c(1,0,4,1,0,-1),ncol=3)
2*x

##      [,1] [,2]
## [1,]    6    4
## [2,]   -2    2

x*x

##      [,1] [,2]
## [1,]    9    4
## [2,]    1    1
```

d) Problem 2 page 10 (the first problem 2 on page 10)

```
attach(mtcars)
apply(mtcars[,c(6,1)],2,mean)

##       wt      mpg
##  3.21725 20.09062

apply(mtcars[,c(6,1)],2,median)

##     wt    mpg
##  3.325 19.200
```

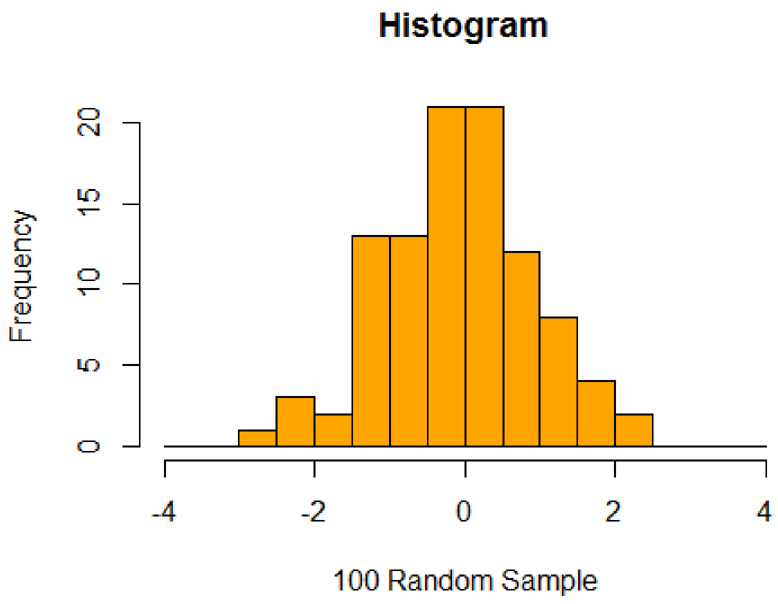
e) Problem 1, page 11

```
dnorm(0.5, mean=2, sd=0.5)

## [1] 0.008863697
```

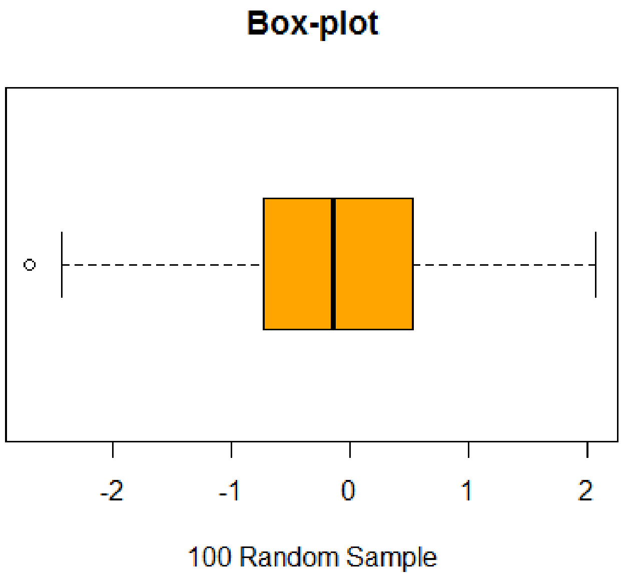
```
pnorm(2.5, mean=2, sd=0.5)

## [1] 0.8413447

qnorm(0.95, mean=2, sd=0.5)

## [1] 2.822427

pnorm(3, mean=2, sd=0.5)-pnorm(1, mean=2, sd=0.5)

## [1] 0.9544997
```

f) Problems 1 and 2, page 13-14

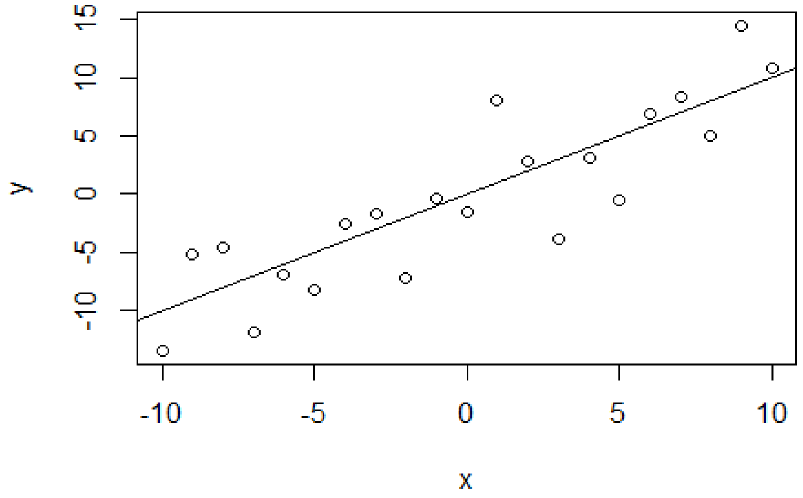
```
set.seed(02138)
x<-rnorm(100)
hist(x,
     breaks = seq(-4,4,0.5),
     main="Histogram",
     xlab="100 Random Sample",
     col = "orange")
```



```
boxplot(x,horizontal = T,
          main="Box-plot",
          xlab="100 Random Sample",
          col = "orange" )
```



```
x<-(-10):10 # generate a sequence from -10 to 10 with increment=1
n<-length(x) # store the length of vector x into n
y<-rnorm(n,x,4) # randomly draw 1 sample from each of the n normal distributi
ons with mean equals to the respective element in x
plot(x,y) # plot random samples versus its underlying mean
abline(0,1) # add a linear regression line, which is supposed to have interce
pt=0 and slope=1
```



g) Problem 2, page 16 [to make life easier write two functions, one to return the sum of x and another function to return the sum of x squared-you might want to do problem 3 for experience in writing simple functions then come back to this part.]

```
sum.of.x <-function(x){
  return(sum(x))
}

sum.of.square.x<-function(x){
  return(sum(x^2))
}
```

---

3) We will revisit optimization after we start portfolio theory, but since it is easy enough to do in R [for the most basic problems] let's spend a little time with the idea since it is a good example of writing a function, plotting and then calling an R routine [we use the phrase R routine and R function interchangeably, but don't want you to get confused between the function you write as opposed to a built in R function]. Broadly speaking, optimization is the process of finding the minimum or maximum of an objective function.

Finally, what you have to do for this problem:

A farmer has 2400 feet of fencing and wants to fence off a rectangular field that borders a straight river. He needs no fence along the river. What are the dimensions of the field that has the largest area?

a) Write a function in R for the area of the field as a function of x [what is x? Why only x?]

Answer: Define x as the length of the rectangular border that is parallel to the river (but not along the river), because the total length of fence is set to 2400, by the rule of simple math, the other two side of the rectangular is (2400-x)/2. Therefore, the area is x*(2400-x)/2 .

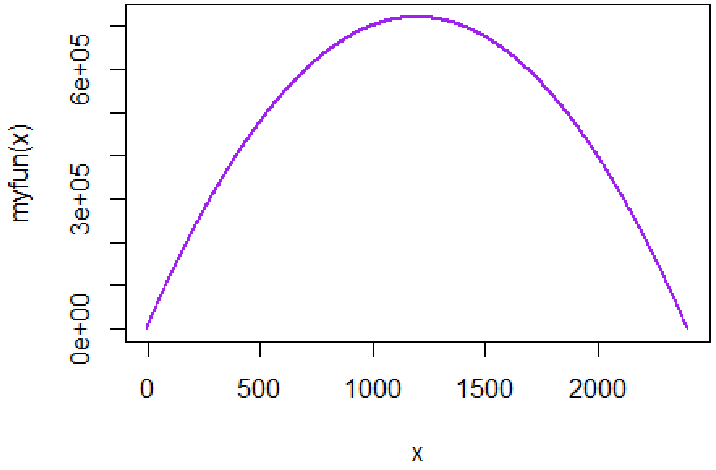
```
myfun<-function(x){
  return(x*(2400-x)/2)
}
```

b) What is the range of possible values for x?

Answer: Because all sides of the rectangular have to be larger than 0, i.e., x>0 and (2400-x)/2>0, which gives the range of possible values of x to be 0<x<2400.

c) Plot the function in part (a) .

```
x<-seq(0,2400,0.1)
plot(x,myfun(x), col="purple", cex=.2 )
```



d) Find the maximum value [set maximum=TRUE in the optimize function]

```
optimize(myfun, lower=0,upper=2400, maximum = TRUE)
```

```
## $maximum
## [1] 1200
##
## $objective
## [1] 720000
```

e) Write up your solution clearly.

Answer: The farmer should make a rectangular fence with 1200 as the length of the side parallel to the river, and 600 as the length of two sides verticle to the river, to make the area of the field maximized which is 720,000 square feet.

---

4) Remember integration from your high school calculus class? Ok, neither do I. Luckily R can do integration for us. Now I haven't used an integral in a long time so not sure how useful this is, but it's educational thing to do.

a) Write a function in R to generate f(x)=0.2+25x. Call your function myfun1. Deliverable is your code.

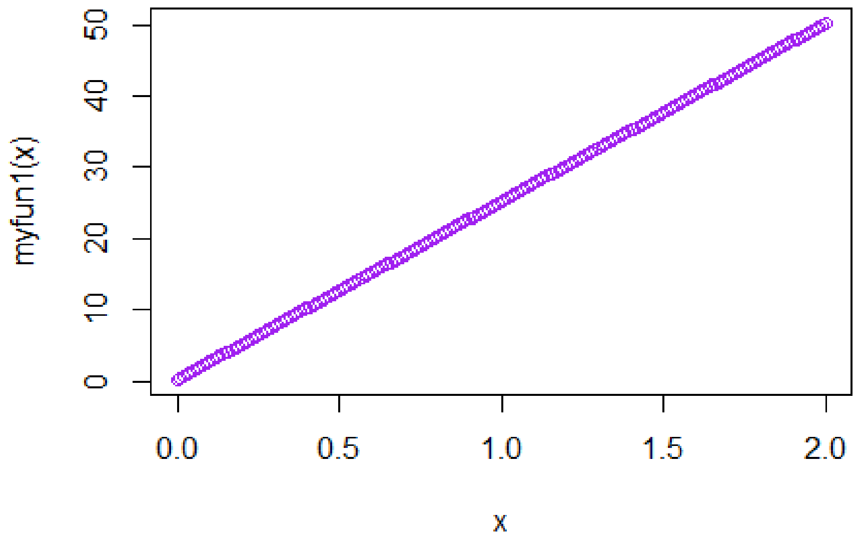
```
myfun1<-function(x){
  return(0.2+25*x)
}
```

b) R can vectorize commands. That is it can compute the function on a vector at once. Generate a sequence in R from 0 to 2 by x=seq(0,2,.01).

i. Deliverable. In words, what does the seq command do? > Answer: the command seq generates a sequence from 0 to 2 with increment=0.01.

ii. Deliverable. Generate a plot of x versus f(x).

```
x<-seq(0,2,0.01) # the command seq generates a sequence from 0 to 2 with incr
ement=0.01
plot(x,myfun1(x), col="purple")
```



iii. Deliverable. Calculate the area under the curve by hand. > Answer: (0.2+50.2)*2/2=50.4

iv. Deliverable. Show the function call integrate(myfun1,0,1).

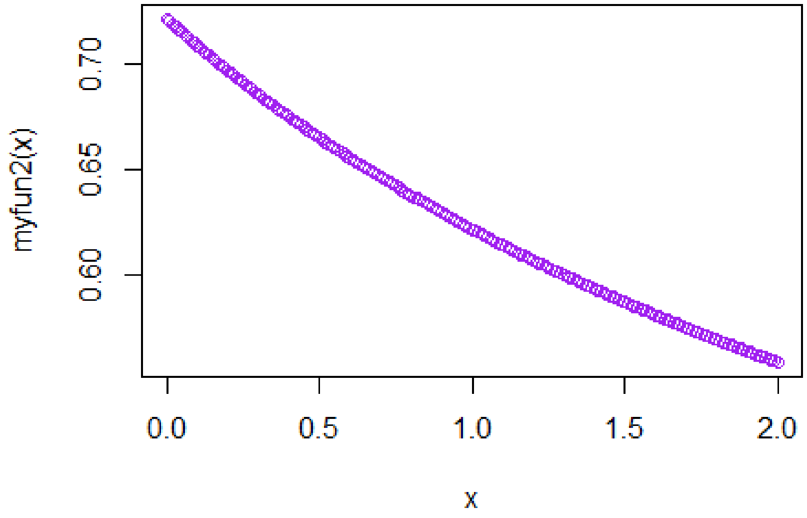
```
integrate(myfun1, 0,2)

## 50.4 with absolute error < 5.6e-13
```

v. Deliverable. Do your answers to (iii) and (iv) agree? > Answer: Yes, answers to (iii) and (iv) agree!

c) Write a function in R to generate f(x)=1/log(x+4). Call your function myfun2.

i. Deliverable. Generate a plot of x versus f(x).

```
myfun2<-function(x){
  return(1/log(x+4))
}
plot(x,myfun2(x), col="purple")
```



ii. Deliverable. Show the function call integrate(myfun2,-1,3).

```
integrate(myfun2, -1, 3)

## 2.593463 with absolute error < 6.9e-14
```

iii. Go to wolframalpha.com and enter this code integrate(1/log(x+4),-1,3). Does the answer agree with R?

Answer: Yes, wolframalpha.com is 2.59346, agree with R!

d) In R, execute the command integrate(dnorm,0,Inf). What is the answer? What is it calculating?

```
integrate(dnorm, 0, Inf)

## 0.5 with absolute error < 4.7e-05
```

Answer: It's calculating the integral of standard normal density funtion (mean=0, sd=1) from 0 to infinite, which is the probability of drawing a random sample from Normal(0,1) greater than 0, which is 50% of chance.

5) Install the quantmod package in R using the command install.packages("quantmod")

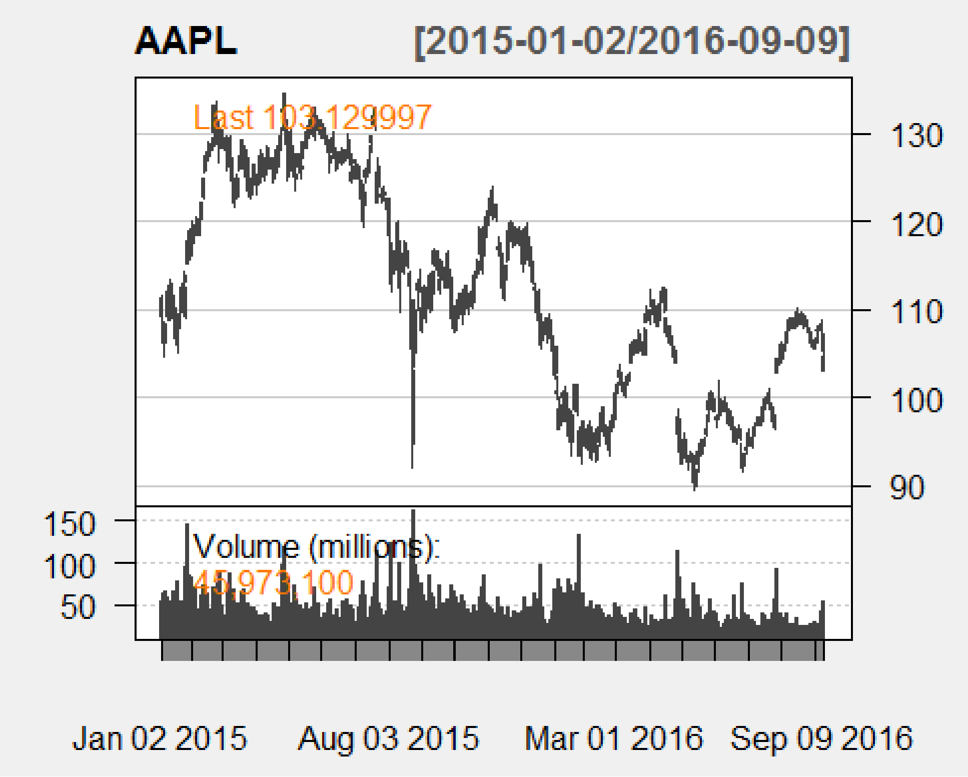
Note: we want to always work with adjusted closing prices (adjusted for stock splits and dividends) instead of the usual closing prices. This is done as follows: getSymbols("AAPL", from = "2015-01-01") AAPL.a =adjustOHLC(AAPL)

a) Using R, get AAPL's data from 2015-01-01 (see above). There is no deliverable for this part of the question.

```
library(quantmod)
getSymbols("AAPL", from="2015-01-01")

## [1] "AAPL"

AAPL.a<-adjustOHLC(AAPL)
```

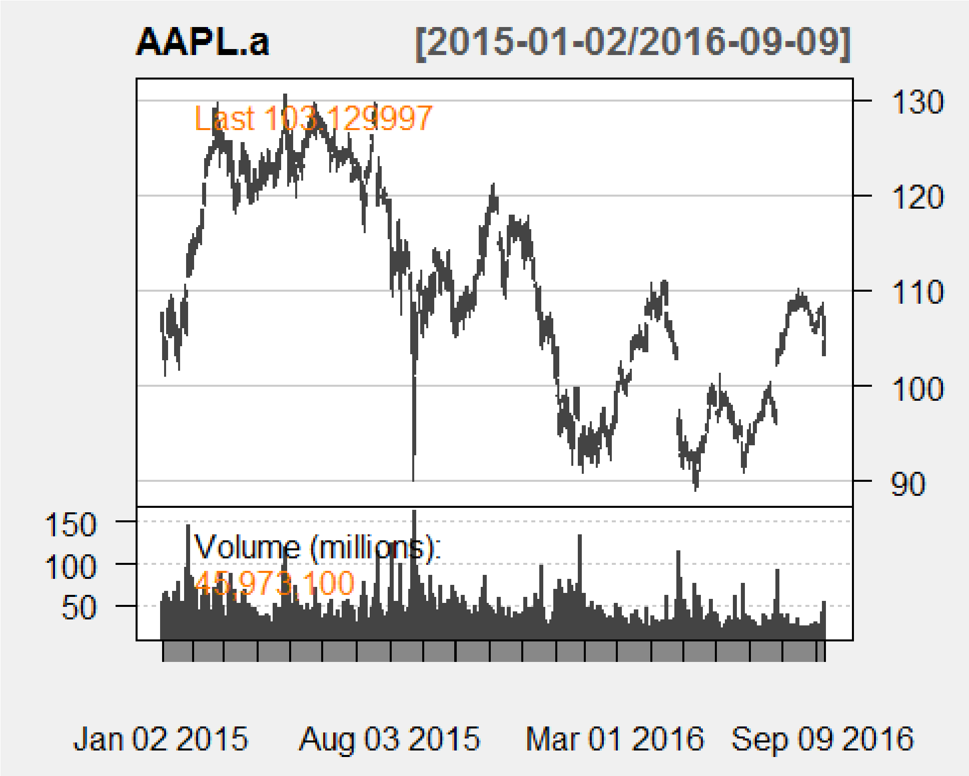
b) Produce a plot of AAPL's (Apple) unadjusted closing price for the last year using the command candleChart(AAPL,theme = "white")

```
candleChart(AAPL, theme = "white")
```



c) Produce a plot of AAPL's (Apple) adjusted closing price for the last year using the command candleChart(AAPL.a,theme = "white")

```
candleChart(AAPL.a, theme = "white")
```



d) How do the plots in (b) and (c) differ?

Answer: Generally speaking, the unadjusted closing price and adjusted price during this period are similar, with minor difference, especially we can observe that the maximum of adjusted is larger than unadjusted.The correlation between adjusted and unadjusted is 0.9979007.These difference are due to all corporate actions that are beyond the supply and demand of market participants, such as stock splits, dividends/distributions and rights offerings.

e) What day of the year had the lowest trading volume for AAPL? Does this make sense? [see page 318 of the R Cookbook for how to find the min and max of a vector].

```
AAPL[which.min(AAPL$AAPL.Volume),]
```

```
##            AAPL.Open AAPL.High AAPL.Low AAPL.Close AAPL.Volume
## 2015-11-27    118.29    118.41    117.6     117.81    13046400
##            AAPL.Adjusted
## 2015-11-27      115.8376
```

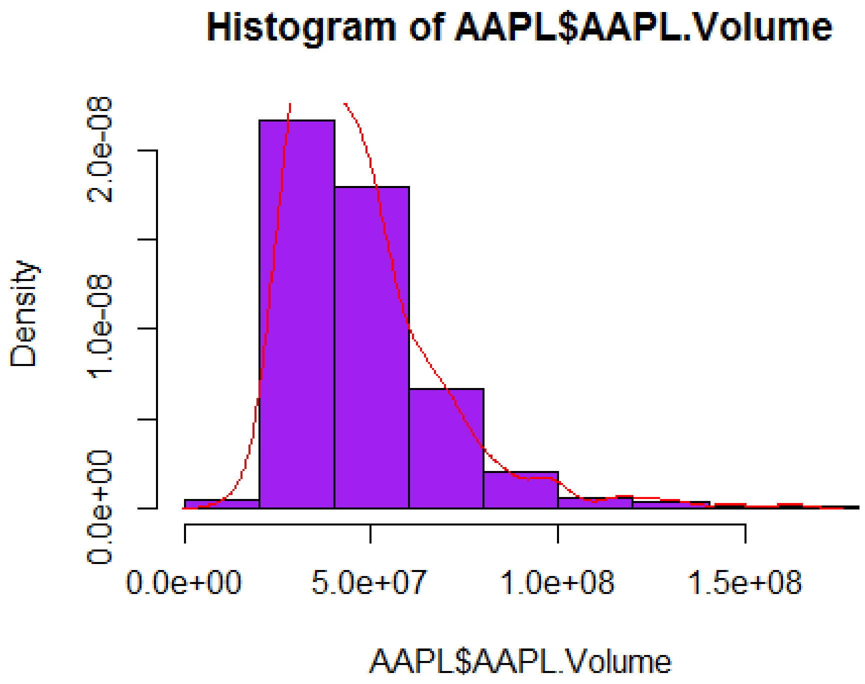
Answer: The lowest daily volumn at 2015-11-27 makes sense, it might be the evidence of the lack of long-term confidence on Apple's CEO Tim Cook among investors, and partially due to Steve Job's death.

f) Go to finance.yahoo.com and find the mean volume (over the last three months) for AAPL.

Answer: 33.203M

g) Make a histogram of daily volume and overlay a density estimate on it (see Recipe 10.19 from the R Cookbook). Doe the data look Normal? Should it be Normal?

```
hist(AAPL$AAPL.Volume, prob=T, col="purple") # using a probability scale
lines(density(AAPL$AAPL.Volume), col="red") # Graph the approximate density
```



Answer: The data doesn't look like Normal, it's highly right skewed. The distribution of volume is not expected to be normally distributed either, because the volume is a non-negative number, it's value should only be on the positive side, therefore, should have a heavy tail on the right.

h) Find the mean and median of daily volume over the last three months for AAPL from R [get the data using [Use getSymbols("AAPL", from = "2016-06-01")]

```
getSymbols("AAPL", from="2016-06-01")

## [1] "AAPL"

mean(AAPL$AAPL.Volume)

## [1] 32635628

median(AAPL$AAPL.Volume)

## [1] 28912100
```

i) Compare the mean values from (f) and (g). Are they similar?

Answer: 32.636M from g) and 33.203M from f), they are similar but with discernible difference which is about 1M.

j) Are the mean and median from (h) the same or different? What does that imply?

Answer: The mean and median are not the same, the difference implies that the distribution is asymmetrical, and hence certainly not normal. And the mean is greater than median also suggests that there are outliers on the right.

k) What is the correlation between AAPL's daily closing price and the daily closing price of QQQ (the Nasdaq 100 Index)? The relevant R command is cor(Ad(QQQ),Ad(AAPL))

```
getSymbols("AAPL", from="2015-01-01")
```

```
## [1] "AAPL"

getSymbols("QQQ", from="2015-01-01")

## [1] "QQQ"

cor(Ad(QQQ),Ad(AAPL))

##              AAPL.Adjusted
## QQQ.Adjusted      0.181376
```

Answer: Very weak correlation between AAPL's daily closing price and the daily closing price of QQQ (the Nasdaq 100 Index).

l) What is the correlation between AAPL's daily volume and the daily volume of QQQ?

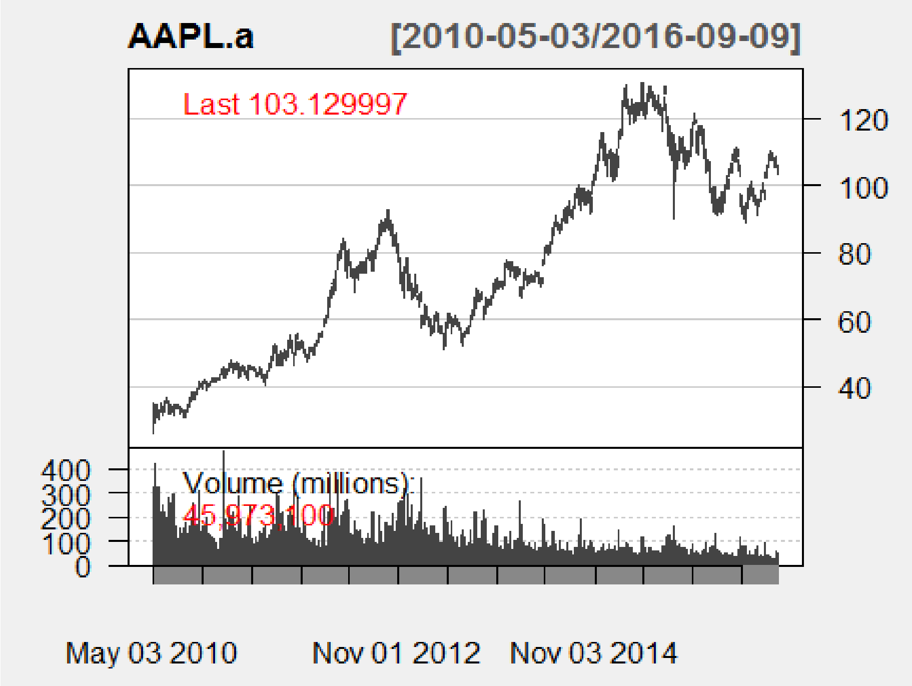
```
cor(QQQ$QQQ.Volume,AAPL$AAPL.Volume)

##            AAPL.Volume
## QQQ.Volume   0.5662741
```

Answer: Very strong correlation between AAPL's daily volume and the daily volume of QQQ.

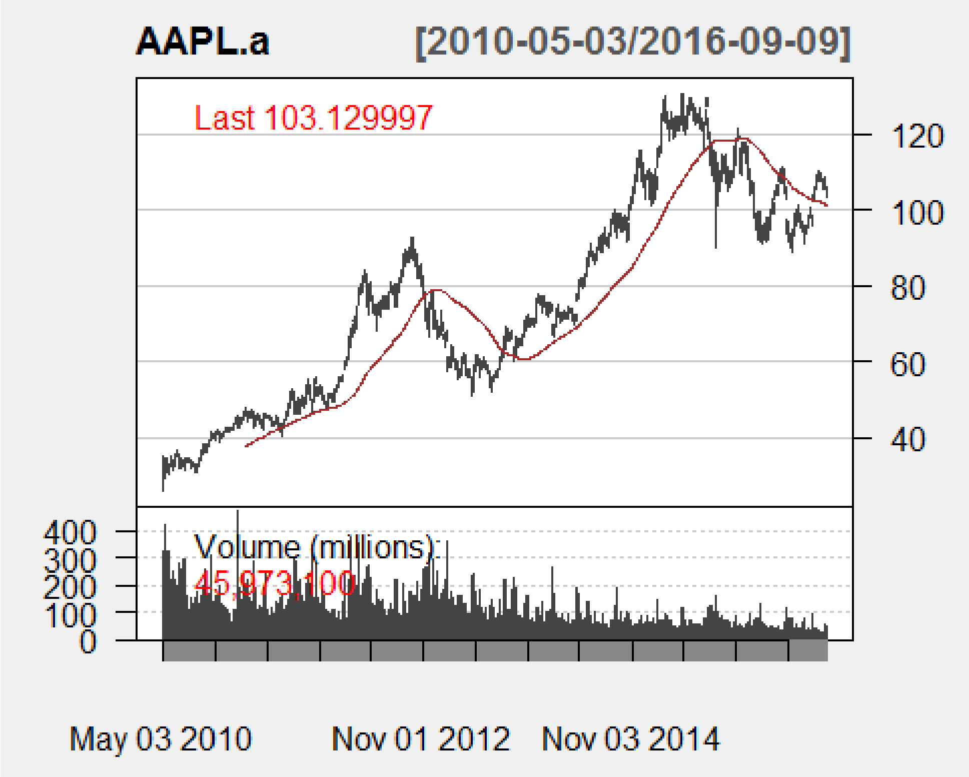
m) People who use technical analysis on the stock market often look to the 200 day moving average (often call the simple moving average) to measure long term trends. If a stock price is above its 200 day moving average it's a buy and above the long term trend. If its below its 200 day moving average it's a sell. Run the following R code and discuss the resultant graph in terms of buying and selling AAPL.

getSymbols("AAPL",from="2010-05-01") AAPL.a=adjustOHLC(AAPL) candleChart(AAPL.a, multi.col = TRUE, theme = "white") addSMA(n=200)

```
getSymbols("AAPL", from="2010-05-01")

## [1] "AAPL"

AAPL.a<-adjustOHLC(AAPL)
candleChart(AAPL.a, multi.col = TRUE, theme="white")
```



```
addSMA(n=200)
```



Answer: During (2010-05-01, 2012-11-01) and (2015-05-01, up till now), the stock price is above its 200 day moving average and therefore it's a buy; between 2012-11-01 and 2015-05-01, the stock price is below it's 200 day moving average and therefore it's sell.