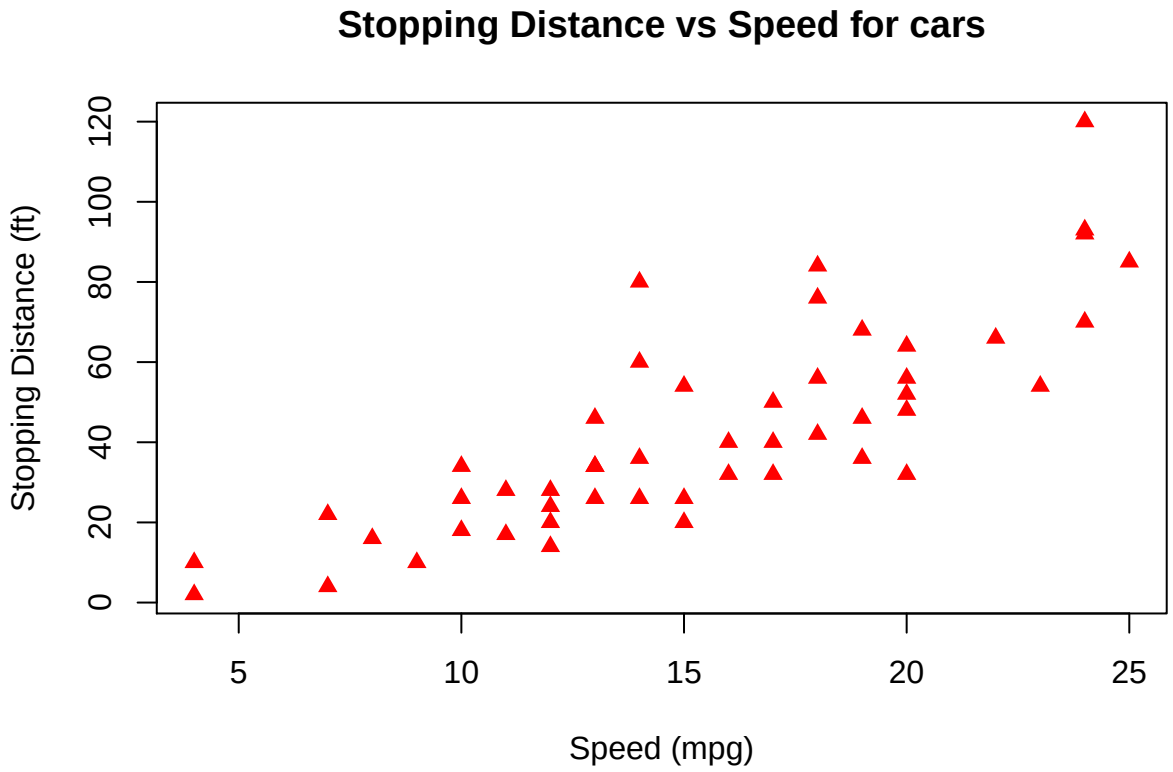
Homework 2

Chapter 4

Question 1

```
data(cars)
plot(
  dist ~ speed, data = cars,
  xlab = 'Speed (mpg)', ylab = 'Stopping Distance (ft)',
  main = 'Stopping Distance vs Speed for cars',
  col = 'red', pch = 17
  )
```

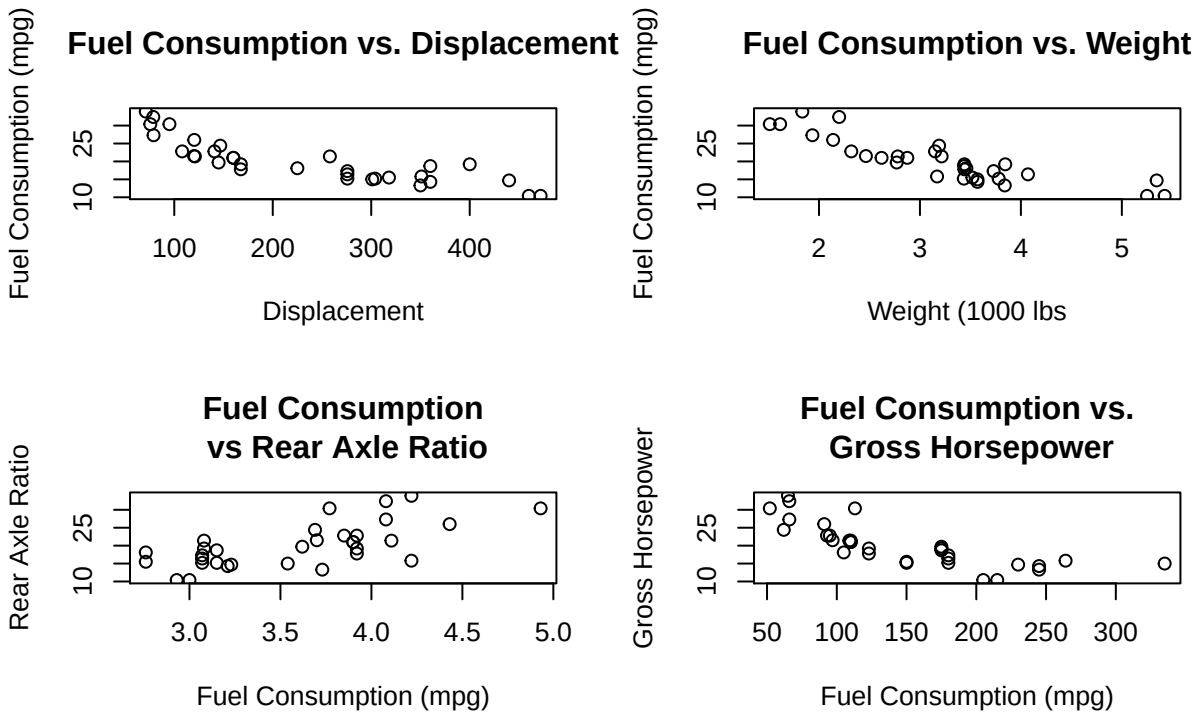


Based on the above graph, there is a possible linear relationship between stopping distance and speed.

Question 4

```
data(mtcars)
par(mfrow = c(2,2), oma = c(0,0,3,0))
plot(mpg ~ disp, data = mtcars, main = "Fuel Consumption vs. Displacement",
     xlab = "Displacement",
     ylab = "Fuel Consumption (mpg)")
plot(mpg ~ wt, data = mtcars, main = "Fuel Consumption vs. Weight",
     xlab = "Weight (1000 lbs", ylab = "Fuel Consumption (mpg)")
plot(mpg ~ drat, data = mtcars, main = "Fuel Consumption\n vs Rear Axle Ratio",
```

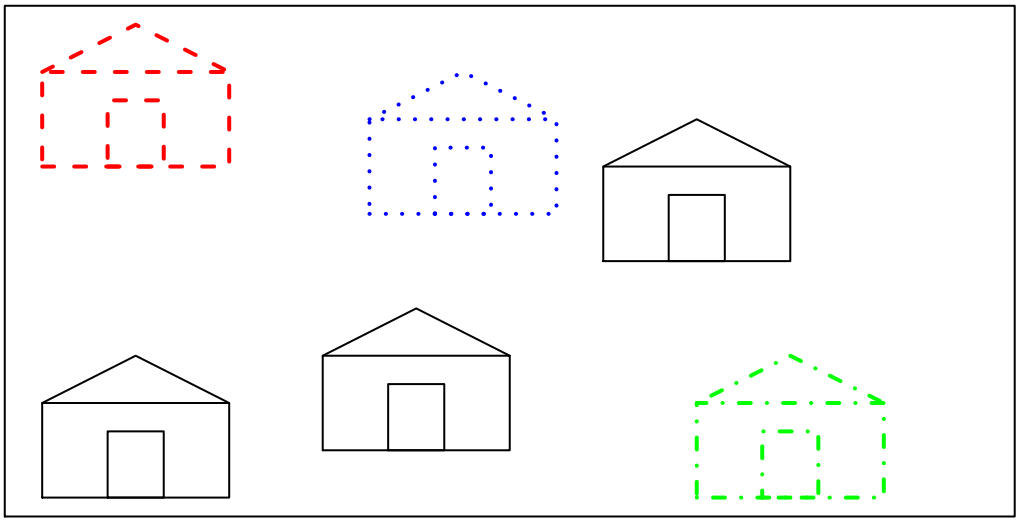
```
     xlab = "Fuel Consumption (mpg)", ylab = "Rear Axle Ratio")
plot(mpg ~ hp, data = mtcars, main = "Fuel Consumption vs.\n Gross Horsepower",
     xlab = "Fuel Consumption (mpg)", ylab = "Gross Horsepower")
```



Comparing the four graphs, the variable that appears to have the strongest relationship with mileage is weight which has a negative relation. While the weight increases the fuel consumption decreases.

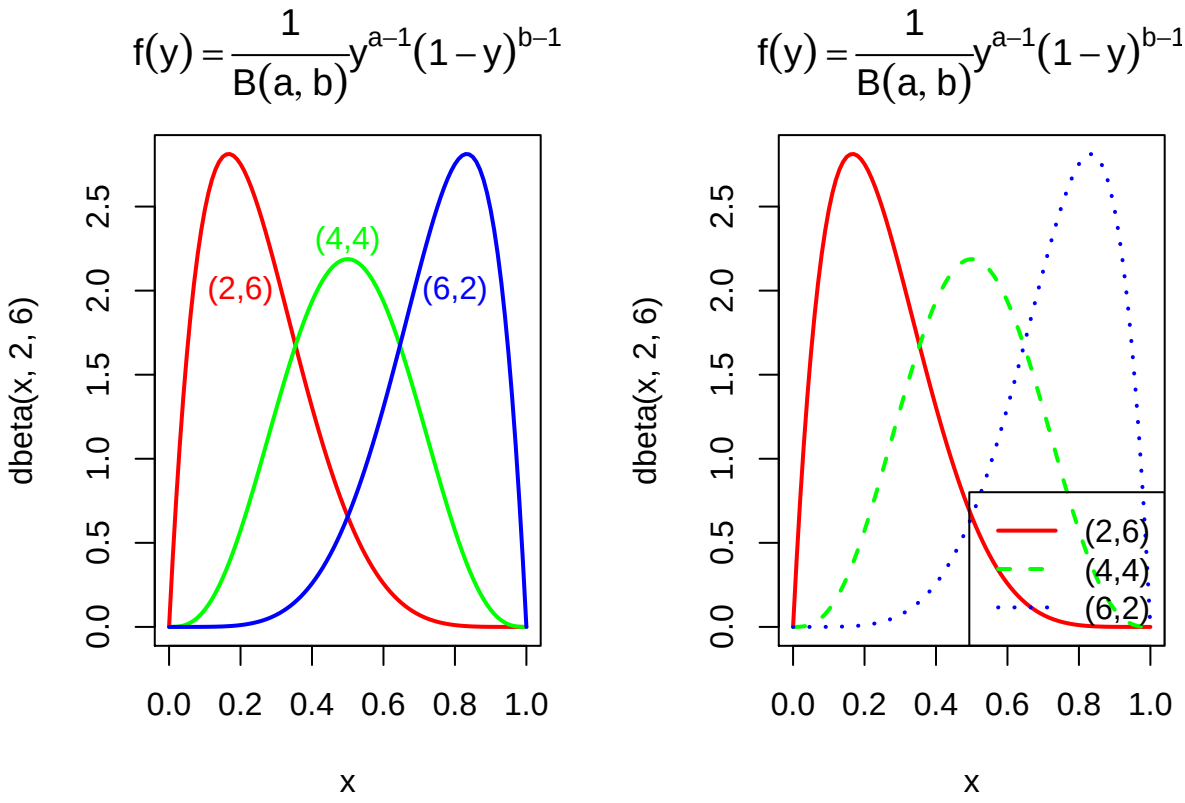
Question 5

```
house=function(x, y, ...){
  lines(c(x - 1, x + 1, x + 1, x - 1, x - 1),
  c(y - 1, y - 1, y + 1, y + 1, y - 1), ...)
  lines(c(x - 1, x, x + 1), c(y + 1, y + 2, y + 1), ...)
  lines(c(x - 0.3, x + 0.3, x + 0.3, x - 0.3, x - 0.3),
  c(y - 1, y - 1, y + 0.4, y + 0.4, y - 1), ...)
  }
plot.new()
plot.window(xlim=c(0,10),ylim=c(0,10))
house(1,1)
house(4,2)
house(7,6)
house(1,8,col='red',lwd=2,lty=2)
house(4.5,7,col='blue',lwd=2,lty=3)
house(8,1,col='green',lwd=2,lty=4)
box()
```



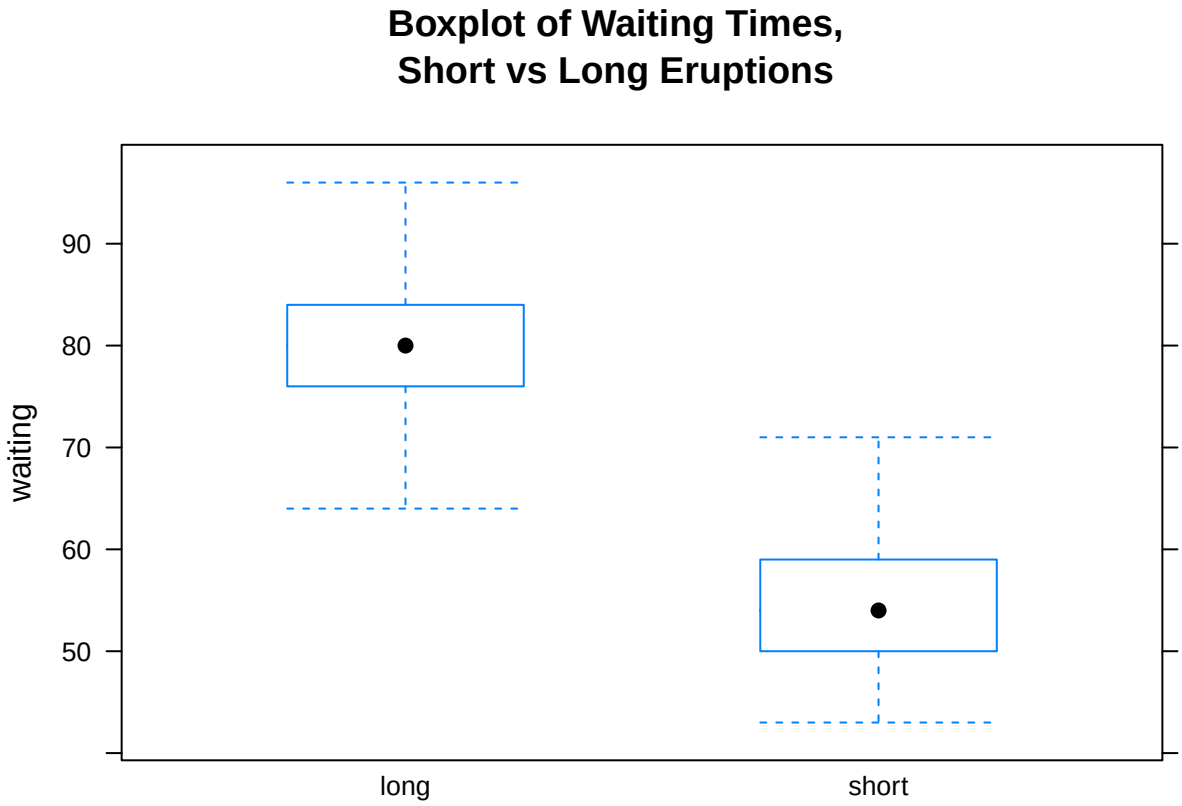
Question 6

```
par(mfrow=c(1,2))
curve(dbeta(x,2,6), from = 0, to = 1, col = 'red', lwd = 2)
curve(dbeta(x,4,4), from = 0, to = 1, col = 'green', lwd = 2, add = TRUE)
curve(dbeta(x,6,2), from = 0, to = 1, col = 'blue', lwd = 2, add = TRUE)
title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
text(x = 0.2, y = 2, labels = '(2,6)', col = 'red')
text(x = 0.5, y = 2.3, labels = '(4,4)', col = 'green')
text(x = 0.8, y = 2, labels = '(6,2)', col = 'blue')

curve(dbeta(x,2,6), from = 0, to = 1, col = 'red', lwd = 2)
curve(dbeta(x,4,4), from = 0, to = 1, col = 'green', lwd = 2, lty = 2, add = TRUE)
curve(dbeta(x,6,2), from = 0, to = 1, col = 'blue', lwd = 2, lty = 3, add = TRUE)
title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
legend(
  'bottomright',
  col = c('red','green','blue'),
  lty = c(1,2,3),
  lwd = 2,
  legend = c('(2,6)','(4,4)','(6,2)')
  )
```

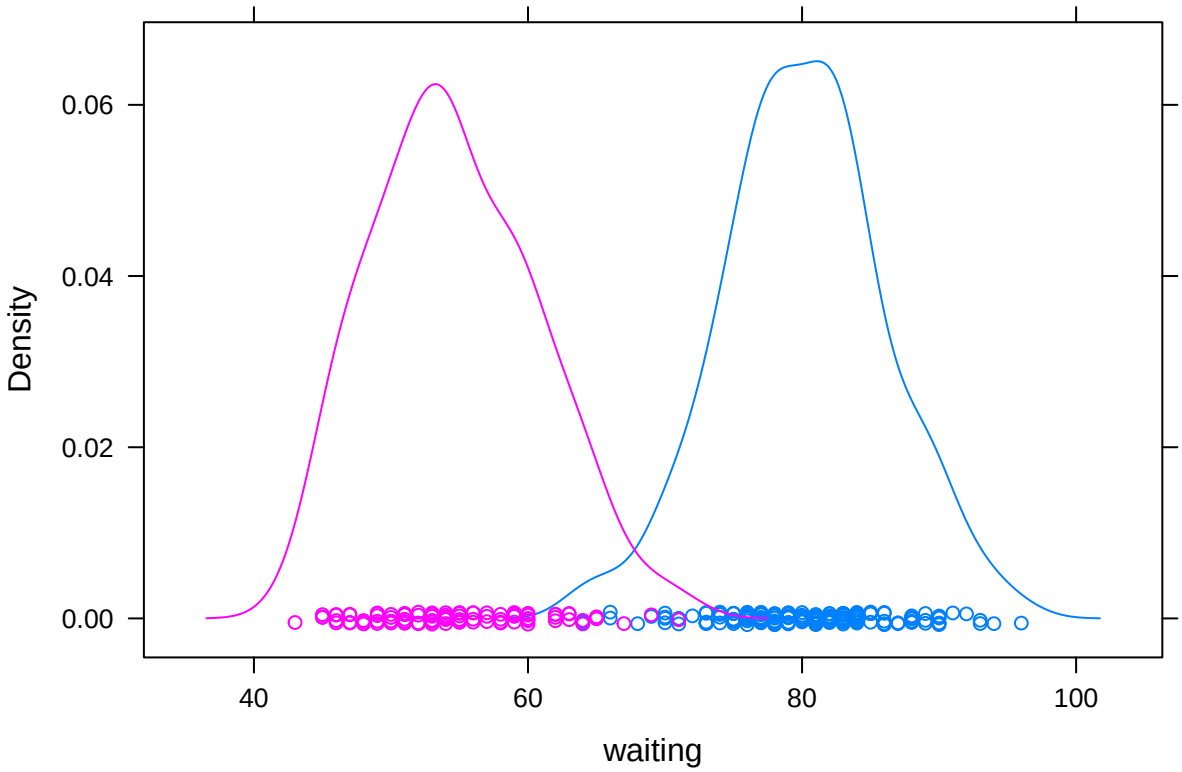


Question 7

```
data(faithful)
faithful$length = ifelse(faithful$eruptions < 3.2, 'short','long')
par(mfrow = c(1,2))
bwplot(waiting ~ length, data = faithful, main = 'Boxplot of Waiting Times,\nShort vs Long Eruptions')
```



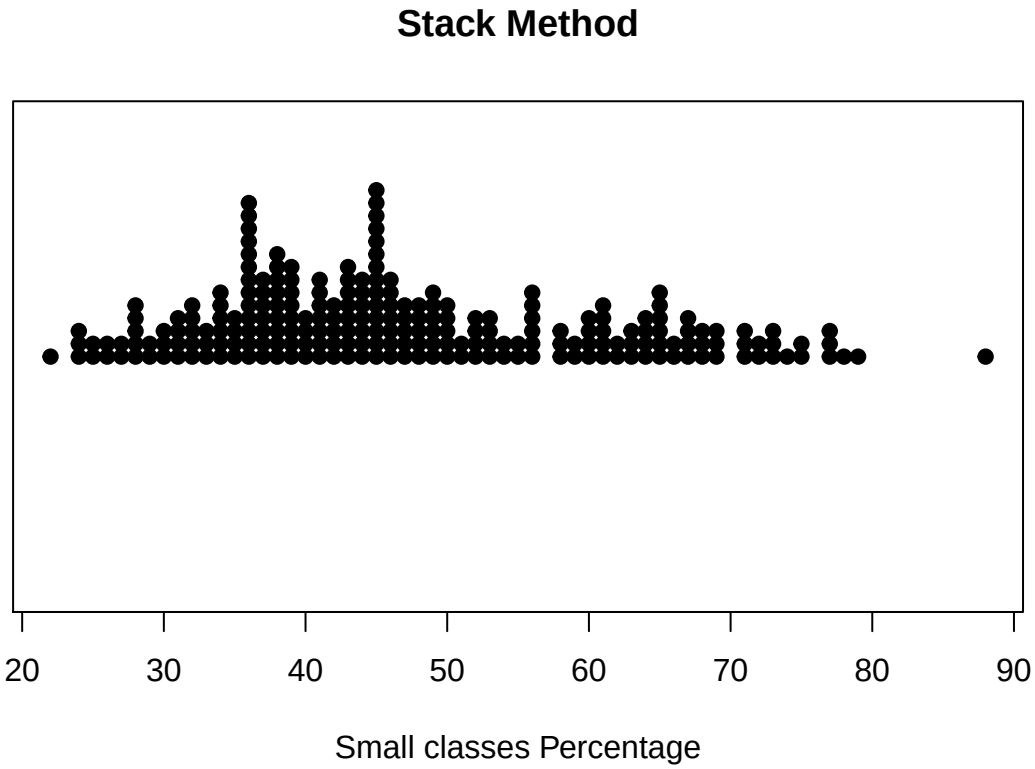

```
densityplot(~waiting, groups = length, data = faithful)
```



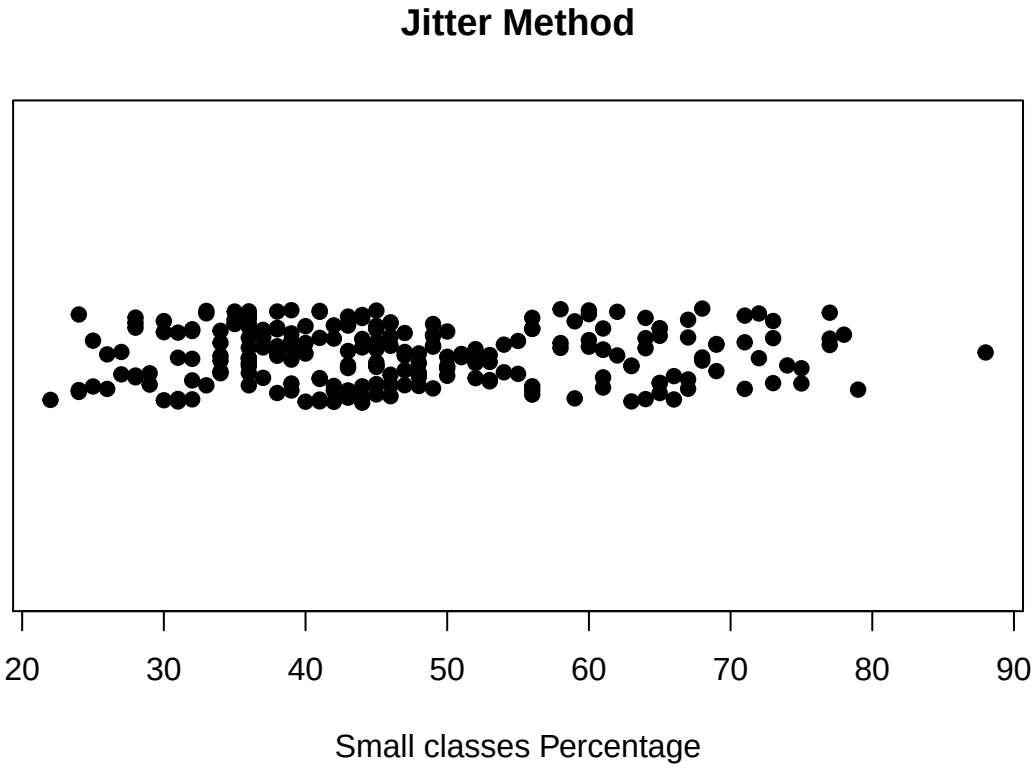
Chapter 5

Question 1

```
#(a)

dat=read.csv(url("http://personal.bgsu.edu/~mrizzo/Rx/Rx-data/college.txt"), sep="\t")

college = subset(dat, complete.cases(dat))

#college$Pct.20
stripchart(college$Pct.20, method="stack", pch=19, xlab="Small classes Percentage",
           main="Stack Method")
```




```
stripchart(college$Pct.20, method="jitter", pch=19, xlab="Small classes Percentage",
           main="Jitter Method")
```

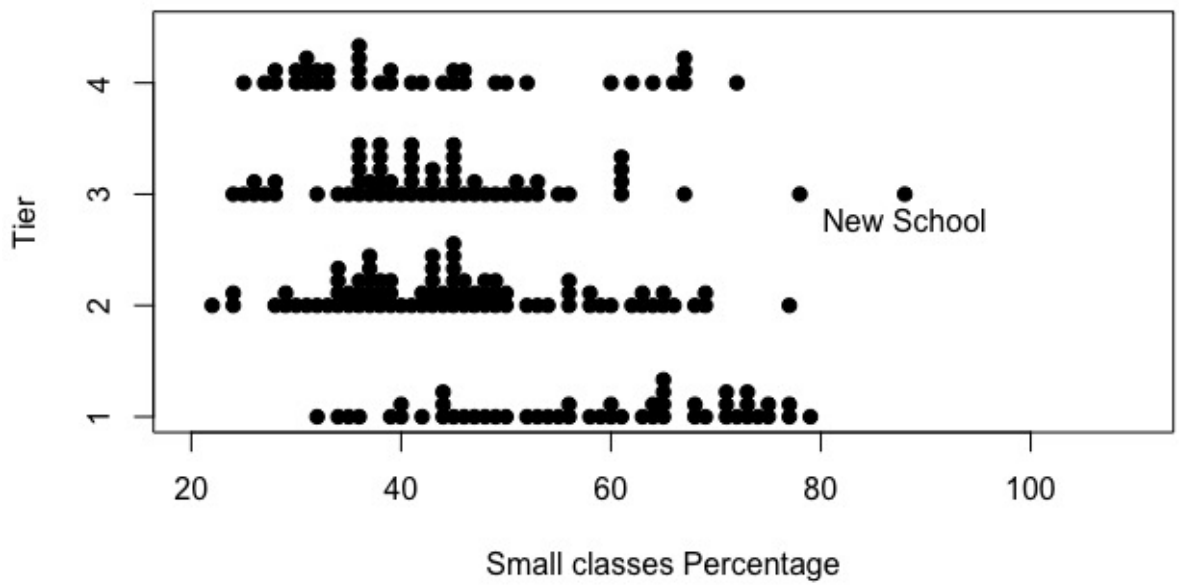


The stack method will stack the dots and we can see that there are much universities with the percentage of small classes between 30 and 50 %.

The jitter method plot the dots of the data with a sligth amount of irregular movement. And we can see that the points are more concentrated at the same interval of percentage.

It might be helpful to construct parallel stripcharts of percentage of small classes based on Tier.

```
#(b)
stripchart(Pct.20 ~ Tier, method="stack", pch=19, xlab="Percentage of Small Classes",
           ylab="Tier", xlim=c(20, 100), data=college)
identify(college$Pct.20, college$Tier, n=1, labels=college$School)
```

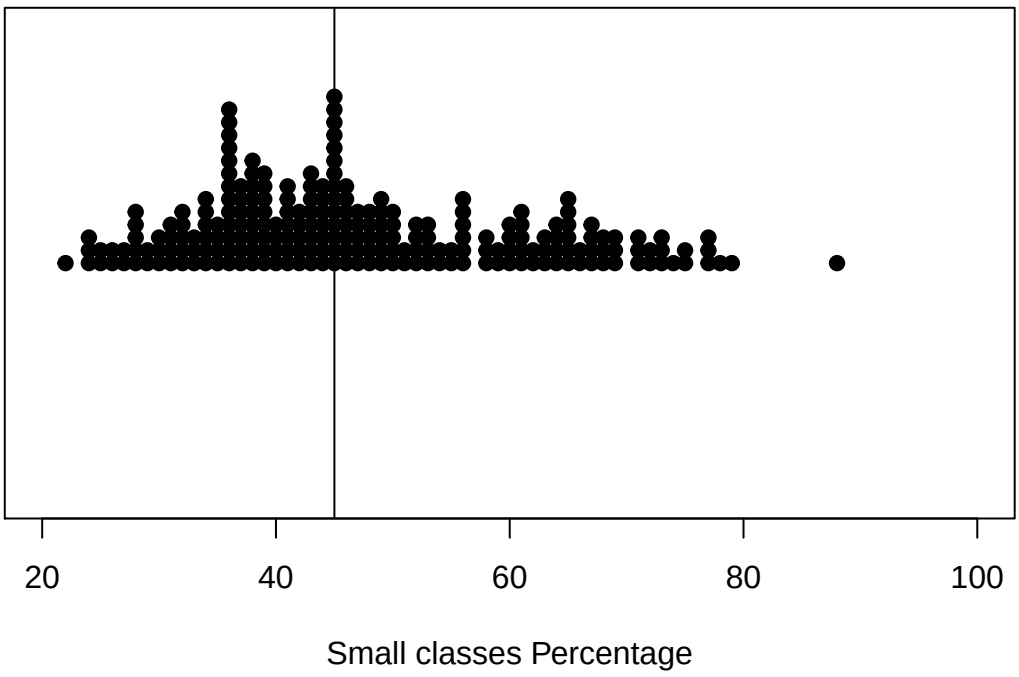


We can see that the school with an unusually high percentage of small classes is called "New School."

```
#(c)
mc=median(college$Pct.20)
mc
```

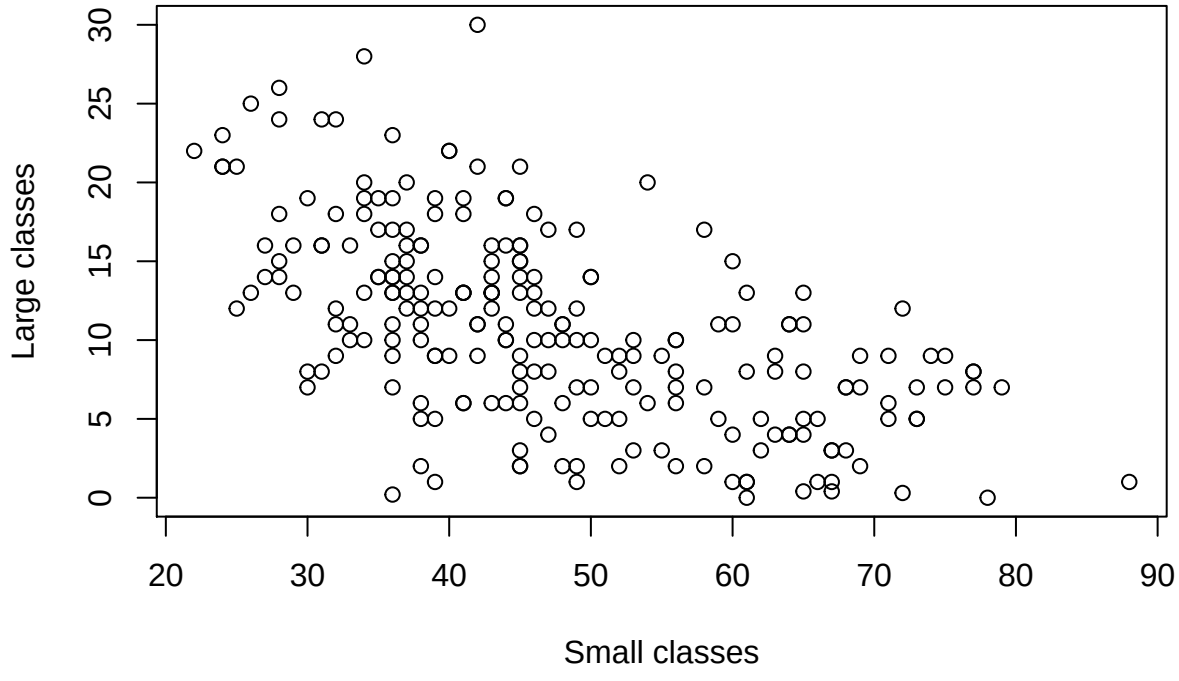
```
## [1] 45
```

```
stripchart(college$Pct.20, method="stack", pch=19, xlab="Small classes Percentage",
           , xlim=c(20, 100))
abline(v=mc)
```



Question 2

```
#(a)
plot(college$Pct.20, college$Pct.50, xlab="Small classes", ylab="Large classes")
```

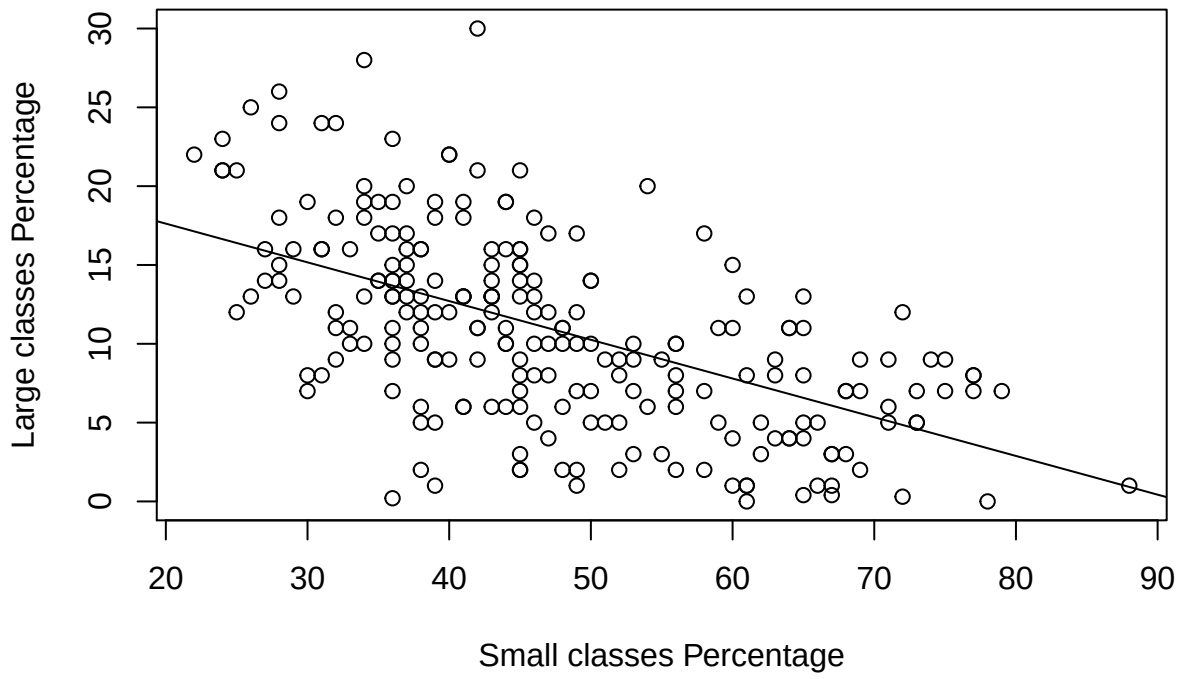



```
#(b)
plot(college$Pct.20, college$Pct.50, xlab="Small classes Percentage",
     ylab="Large classes Percentage")
fit = line(college$Pct.20, college$Pct.50)
```

```
fit
```

```
##
## Call:
## line(college$Pct.20, college$Pct.50)
##
## Coefficients:
## [1]  22.5351  -0.2456
```

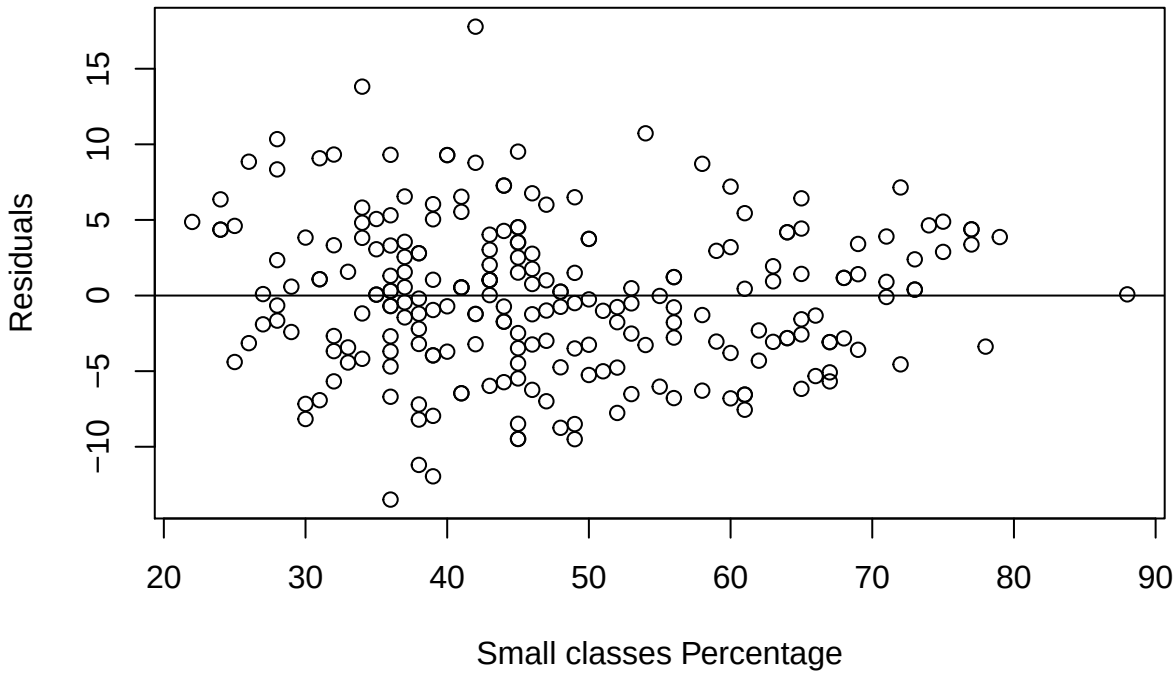
```
abline(coef(fit))
```




```
#(c)
CP20=60
CP50=fit$coefficients[1]-(fit$coefficients[2]*CP20)
CP50
```

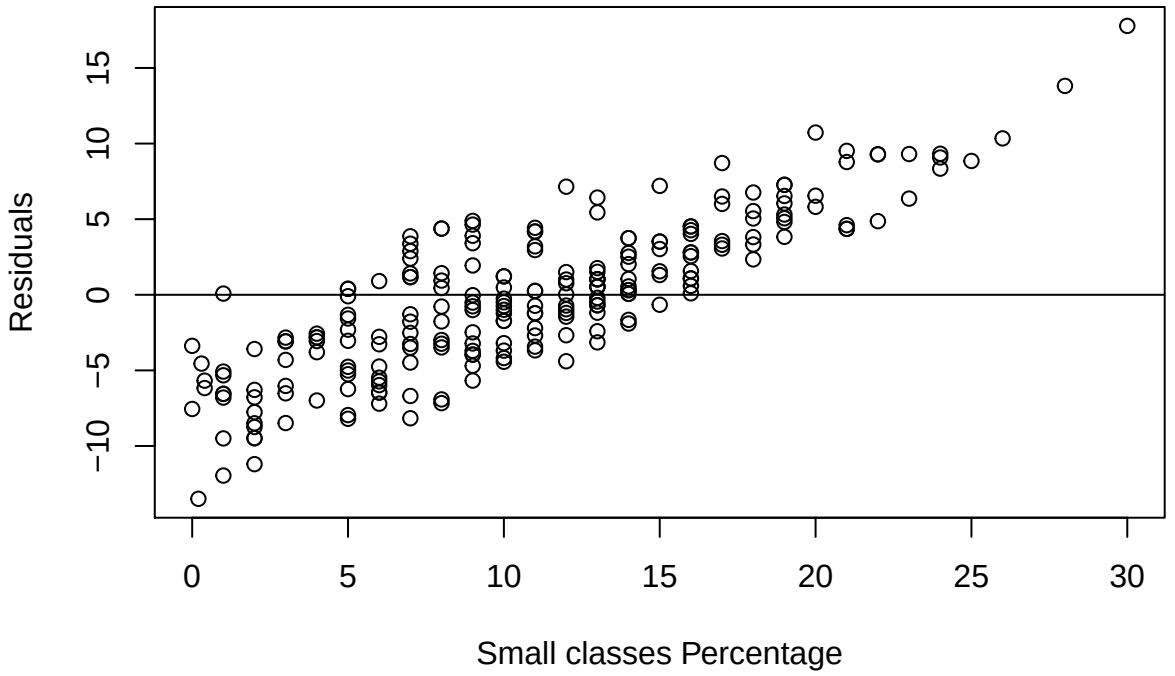
```
## [1] 37.27193
```

```
#(d)
plot(college$Pct.20, fit$residuals, xlab="Small classes Percentage", ylab="Residuals")
abline(h=0)
```



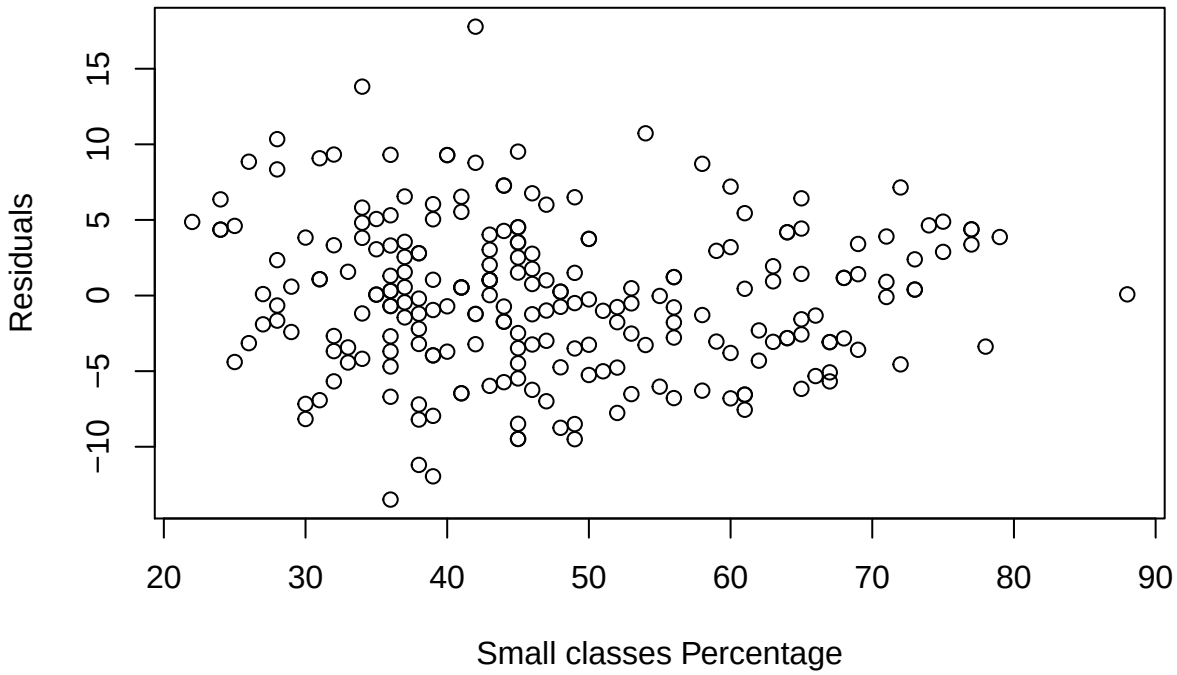
The residuals for a small Pct.20 does not show a evident pattern, the values seems to be random.

```
plot(college$Pct.50, fit$residuals, xlab="Small classes Percentage", ylab="Residuals")
abline(h=0)
```



The residuals for a large Pct.50 does show a evident pattern, the values seems not be random. There is a patter of growth in a linear way.

```
#(e)
plot(college$Pct.20, fit$residuals, xlab="Small classes Percentage", ylab="Residuals")
identify(college$Pct.20, fit$residuals, n=7, labels=college$School)
```




```
## integer(0)
```

The seven positive residuals that exceed 10 in values are University of California, San Diego, University California, Davis, UCLA, Texas - Dallas. This indicate that their percentage of large classes is large given their percentage of small classes.

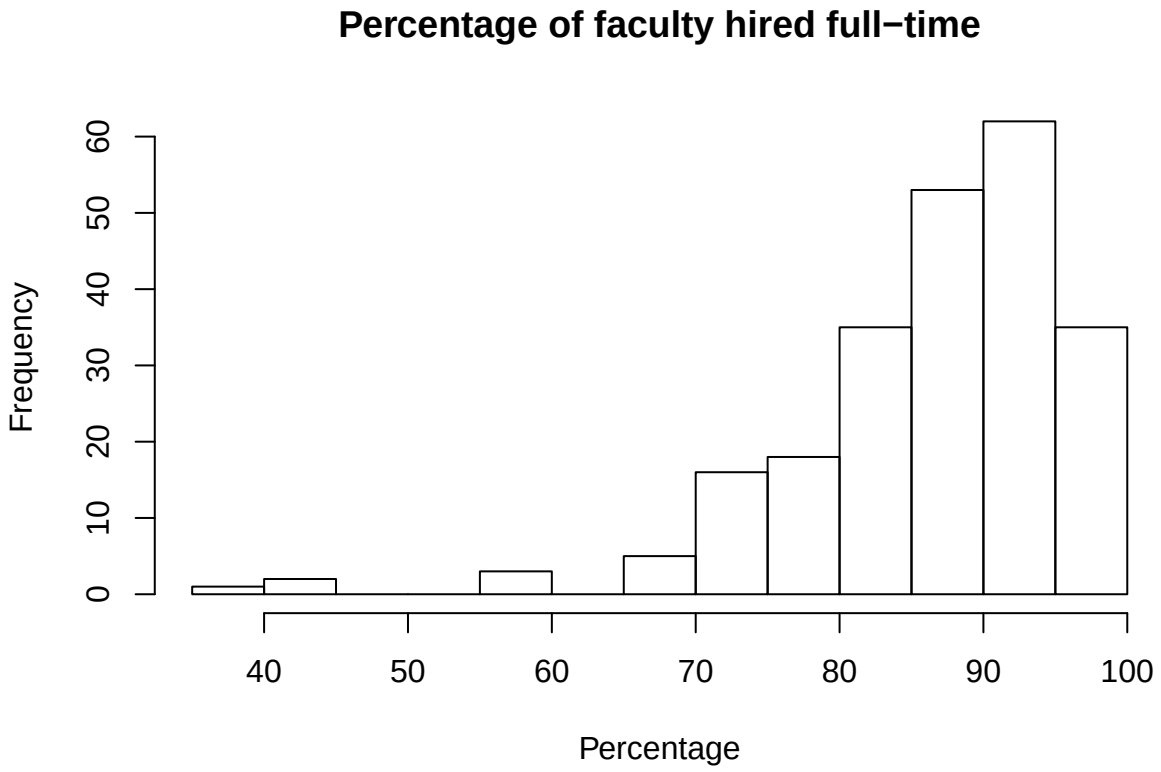
The University San Diego, DePaul are St Thomas have the large negative residuals. These school's with large classes percentage is lower than would predict from their small classes percentage.

Question 5

```
#(a)
college$Full.time
```

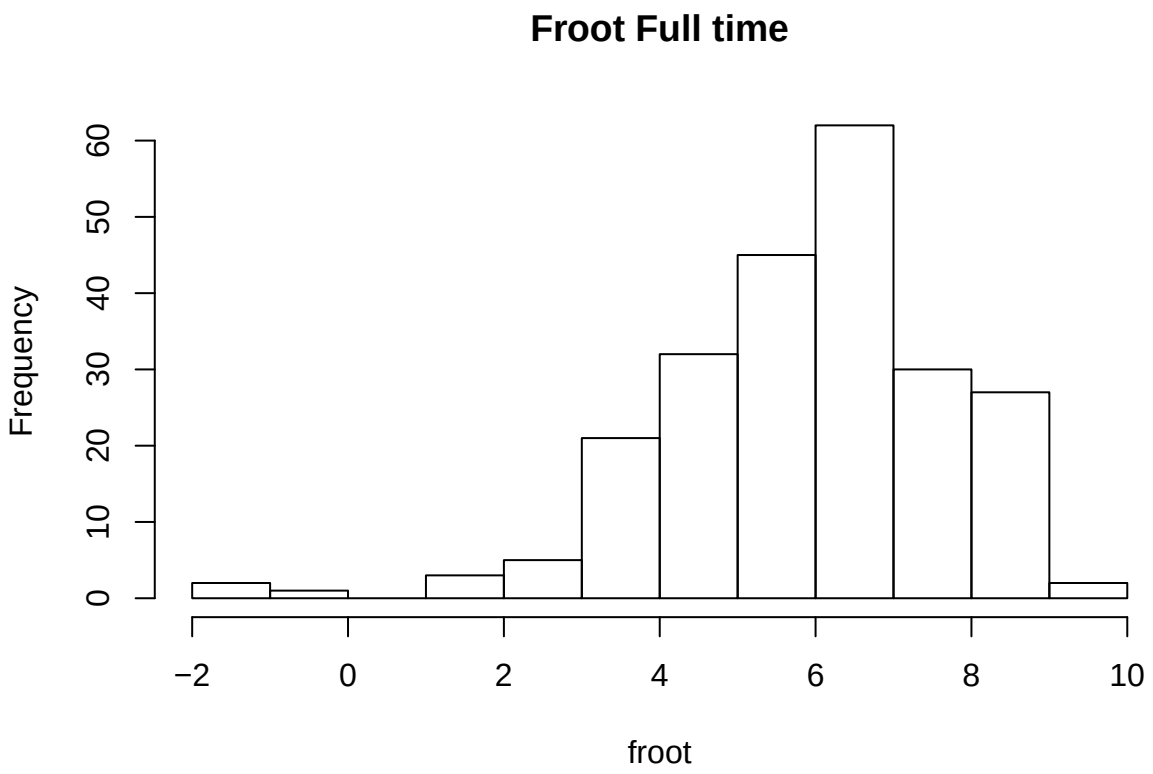
```
##   [1]  93  92  88  97  90  99  86  92  86  97  93  96  94  98  98  93  95
##  [18]  93  93  96  90  94  77  90  98  82  93  86  97  92  88  73  92  78
##  [35] 100  87  93  85  99  94  93  93  94  94  93  91  95  99  97  73  87
##  [52]  89  68  89  89  81  90  78  84  93  95  95  92  86  85  90  84  93
##  [69]  92  90  95  95  88  98  95  89  87  85  90  88  86  94  78  80  80
##  [86]  74  96  83  83  95  96  81  92  92  88  84  90  94  98 100  91  91
## [103]  98  91  81  95  97  89  83  80  84  72  88  82  90  87  79  95  70
## [120]  95  76  94  87  87  86  85  99  74  96  93  97  91  58  89  92  78
## [137]  94  74  89  71  78  72  89  80  92  94  86  37  94  92  82  92  66
## [154]  81  72  83  73  90  95  71  42  73  98  76  97  86  89  99  98  93
## [171]  91  74  84  89  84  82  82  90  99  85  86  97  95  97  73  86  98
## [188]  92  83  85  89  72  83  84  82  96  87  97  88  58  88  78  91  67
## [205]  84  77  89  67  77  73  58  45  91  87  92  83  78  84  90  88  82
## [222]  79  92  88  82  88  91  84  97  99
```

```
hist(college$Full.time, main="Percentage of faculty hired full-time", xlab="Percentage")
```

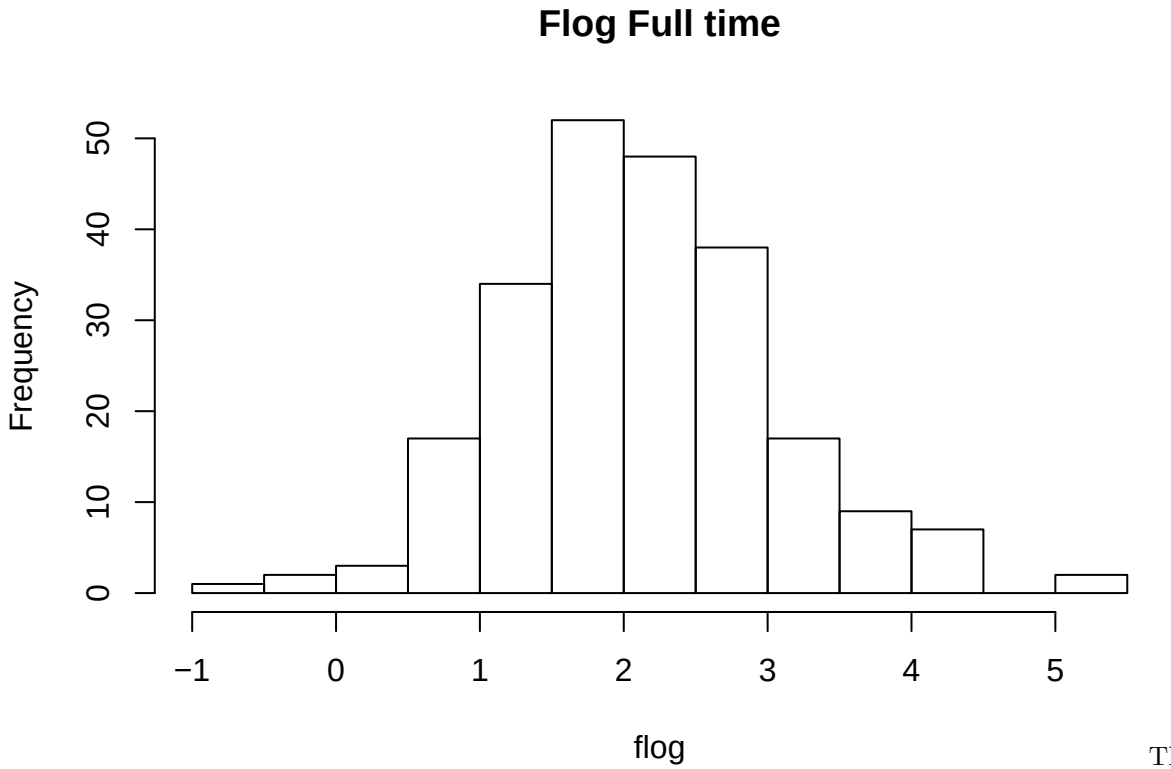


It can be seem that there is a larger amount of faculty that are hired full-time given that the distribution of the data is concentrated mostly in the rigth side of the graph, with higher frenquecy (left-skewed).

```
#(b)
froot=sqrt(college$Full.time)-sqrt(100-college$Full.time)
flog = log(college$Full.time + 0.5) - log(100 - college$Full.time + 0.5)

hist(froot, main="Froot Full time")
```




```
hist(flog, main="Flog Full time")
```



The flog transformation makes the full-time percentage approximately symmetric, but the froot transformation does not.

```
#(c)
```

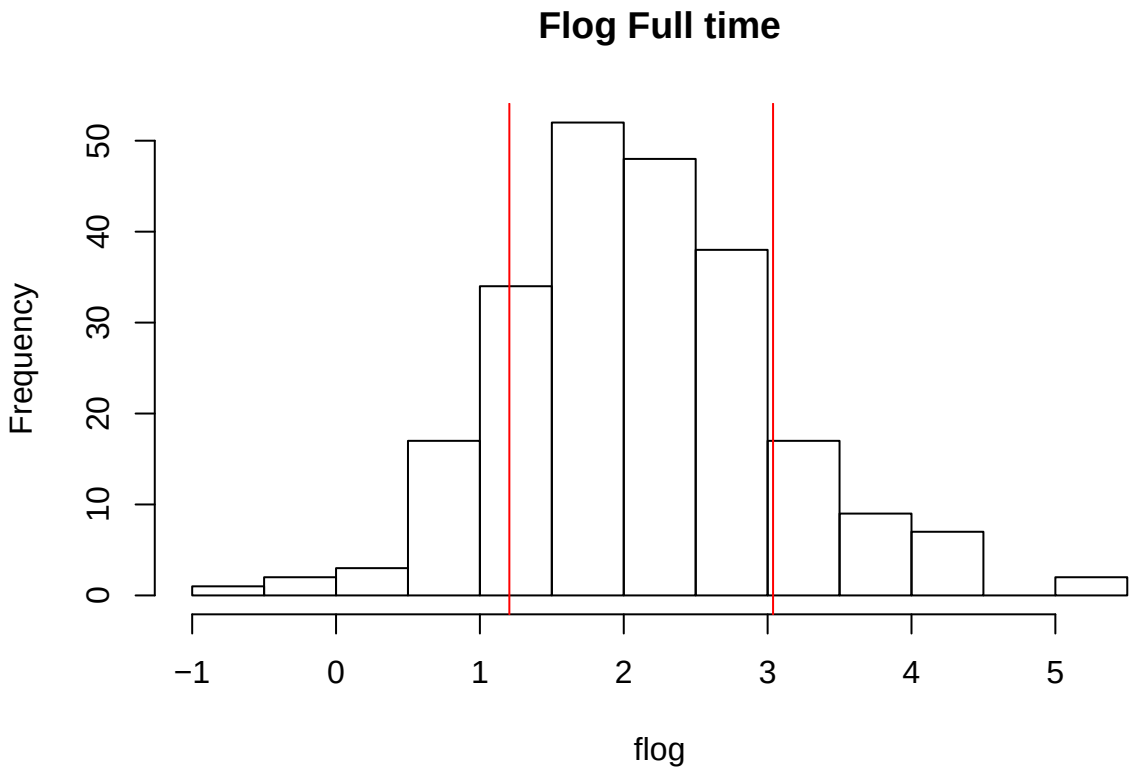
```r
mean(flog)-sd(flog)
```

```
## [1] 1.20476
```

```r
mean(flog)+sd(flog)
```

```
## [1] 3.037732
```

```r
hist(flog, main="Flog Full time")
abline(v=1.20476, col=2)
abline(v=3.037732, col=2)
```

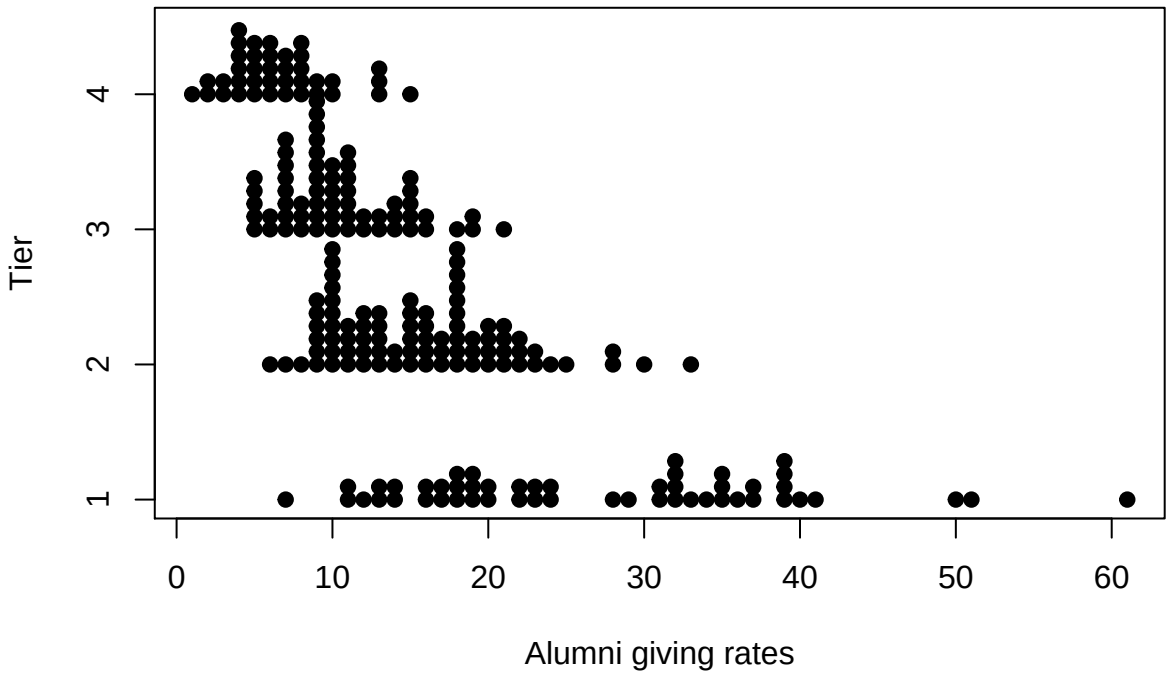


Question 7

```r
#(a)
college$Alumni.giving
```

```
##   [1] 40 61 41 31 37 35 39 36 33 39 51 31 35 35 32 39 37 34 24 50 13 20 28
##  [24] 14 23 39 17 24 22 32 32 11 22 23 29 32  7 18 14 13 19 16 11 18 18 12
##  [47] 20 17 16 19 19 22 10 16 12  9 15 12 18 14 28 16 16 15 15 19 17 10 17
##  [70] 20 15 15 10 13 18 25 18  9 18 10 10 28 13 18 22 10 33 22 18 17 16 16
##  [93] 24 18 13 13 30  7 20 23 21  6 18 21 11 12 12 18 21 23 10  8 20 21 10
## [116] 15 11 19 10 11  9 19  9 10 11  9  9 14 13 20 18 12  7 11 12 11  9  6
## [139] 21  8  5  9  7  5 19 12 11  8 16 15  8 14  6 13  9 11  9 10  7 18 19
## [162]  9 16  9 11  5 11  9  9 10  7 10 14 10  5  9  7 15 13  9  7 15 14 15
## [185] 15  5 11 10 10  7  7  9  9  6  4 10 13  9 10  4  7  5  8  3  8  1  4
## [208]  7 13 15  4  2  4  8  6  7  8  5  2  5  5  7 13  3  6  6  5  6  8  4
```

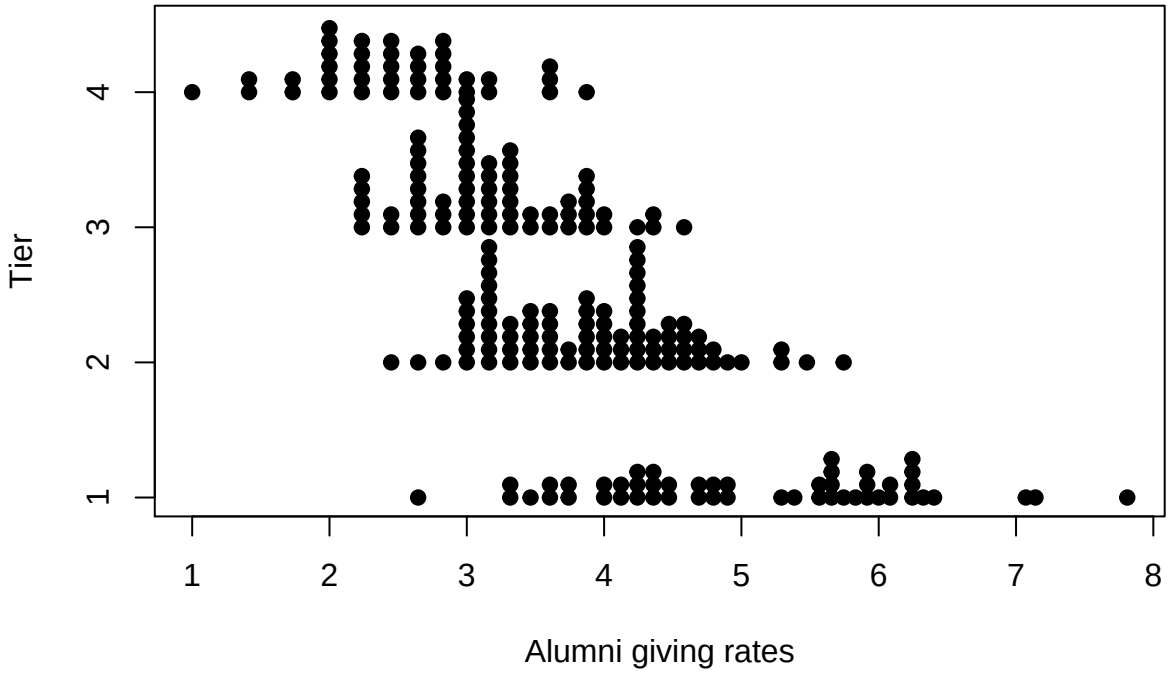
```r
stripchart(Alumni.giving ~ Tier, method="stack", pch=19, xlab="Alumni giving rates",
           ylab="Tier", data=college)
```



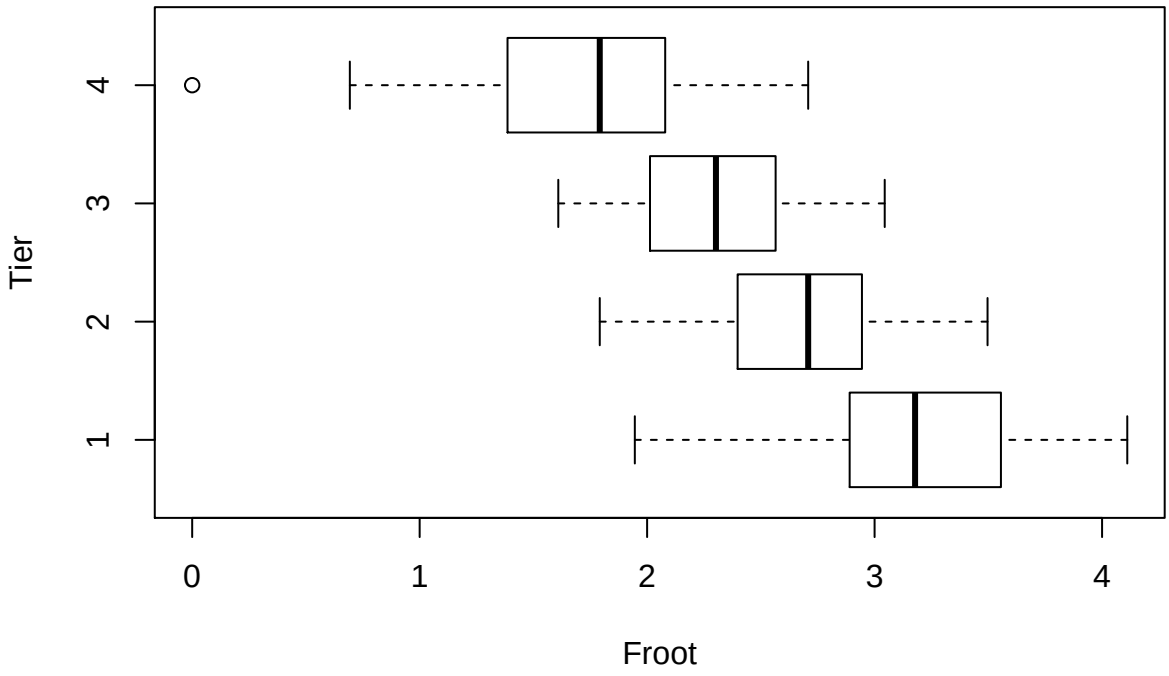

```
#(b and c)
```

As one moves from Tier 4 to Tier 1, the average of giving tends to increase. And the spread also increase.

```
#(d)
roots = sqrt(college$Alumni.giving)
logs = log(college$Alumni.giving)

stripchart(roots ~ Tier, method="stack", pch=19, xlab="Alumni giving rates",
           ylab="Tier", data=college)
```



```
boxplot(logs ~ Tier, data=college, horizontal=TRUE,
        xlab="Froot", ylab="Tier")
```




```
#(e)
```

It can be seen that the square root transformation make the data spread approximately the same between groups. However, the log transformation did not make the spread approximately the same.

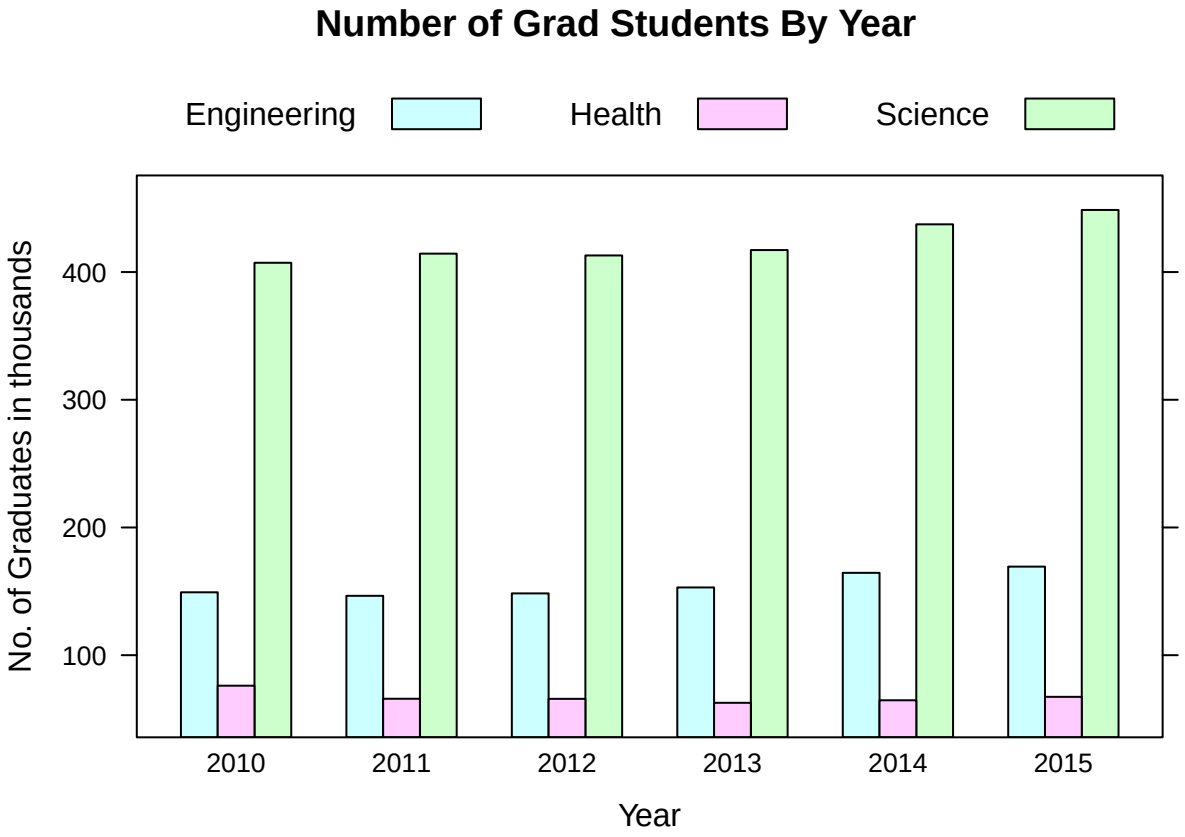
Additional Problem

Part A

```
setwd("~/Desktop")
dat = read.table("GS.csv", header=T, as.is = T, sep="\t")
subj_df = dat[,-6]
names(subj_df) = c("Field", "Yr2010", "Yr2011", "Yr2012", "Yr2013", "Yr2014", "Yr2015")
head(subj_df)
```

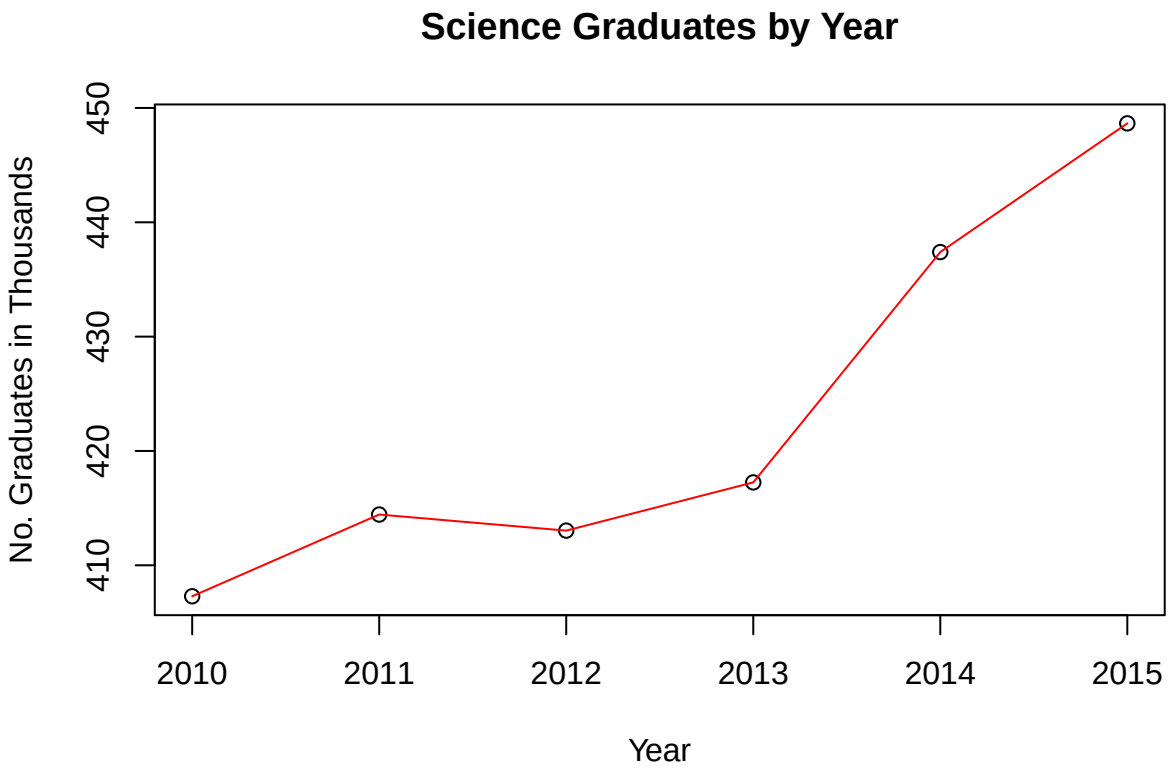
```
##                     Field Yr2010 Yr2011 Yr2012 Yr2013 Yr2014 Yr2015
## 1     All surveyed fields 632652 626820 627243 633010 666586 685397
## 2 Science and engineering 556532 560941 561418 570300 601883 618008
## 3                 Science 407291 414440 413033 417251 437395 448654
## 4   Agricultural sciences  15656  16129  16234  16429  17505  18610
## 5     Biological sciences  74928  75423  76447  76649  78490  80096
## 6                 Anatomy    849    762    700    527    554    594
```

```
barchart((SciHeEng$SEHt.Freq*.001) ~ as.factor(SciHeEng$SEHt.Year),
         groups=SciHeEng$SEHt.Field, data = SciHeEng, auto.key = list(columns = 3),
         main = "Number of Grad Students By Year\n", xlab = "Year",
         ylab = "No. of Graduates in thousands")
```

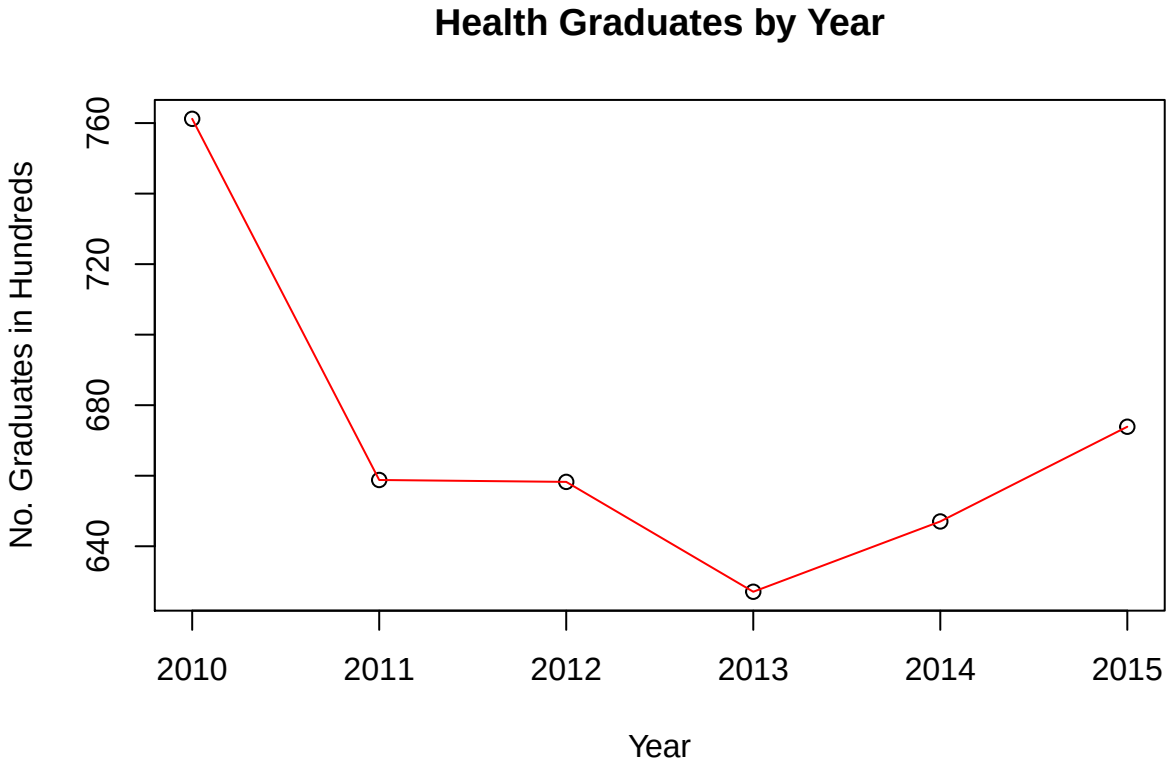


From the above bar chart, it appears that Engineering and Science have increased slightly, but Health may have decreased slightly. Based on the scaling, the increases/decreases are not obvious. Next, we examine plots of the three subjects individually over time.

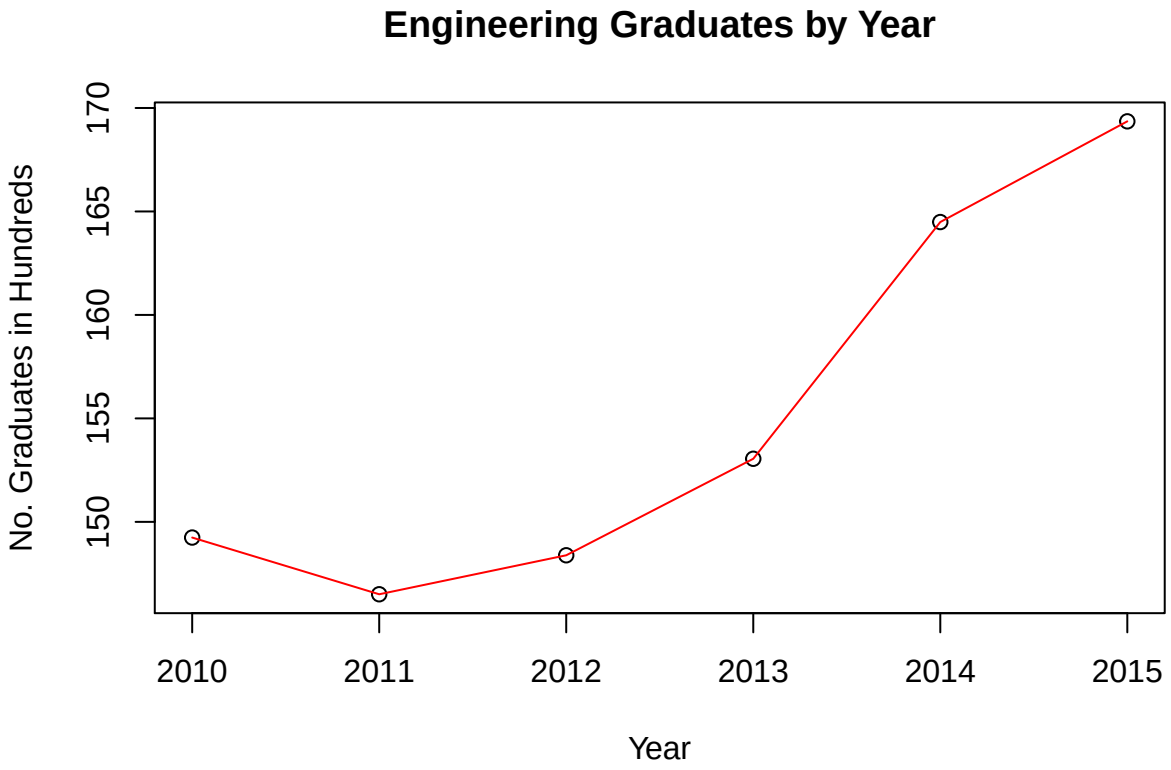
```
Science = Science/1000
plot(Science ~ Year, main = "Science Graduates by Year",
     ylab = "No. Graduates in Thousands",
     xlab = "Year")
lines((Science ~ Year), lwd = 1, col = 'red')
```



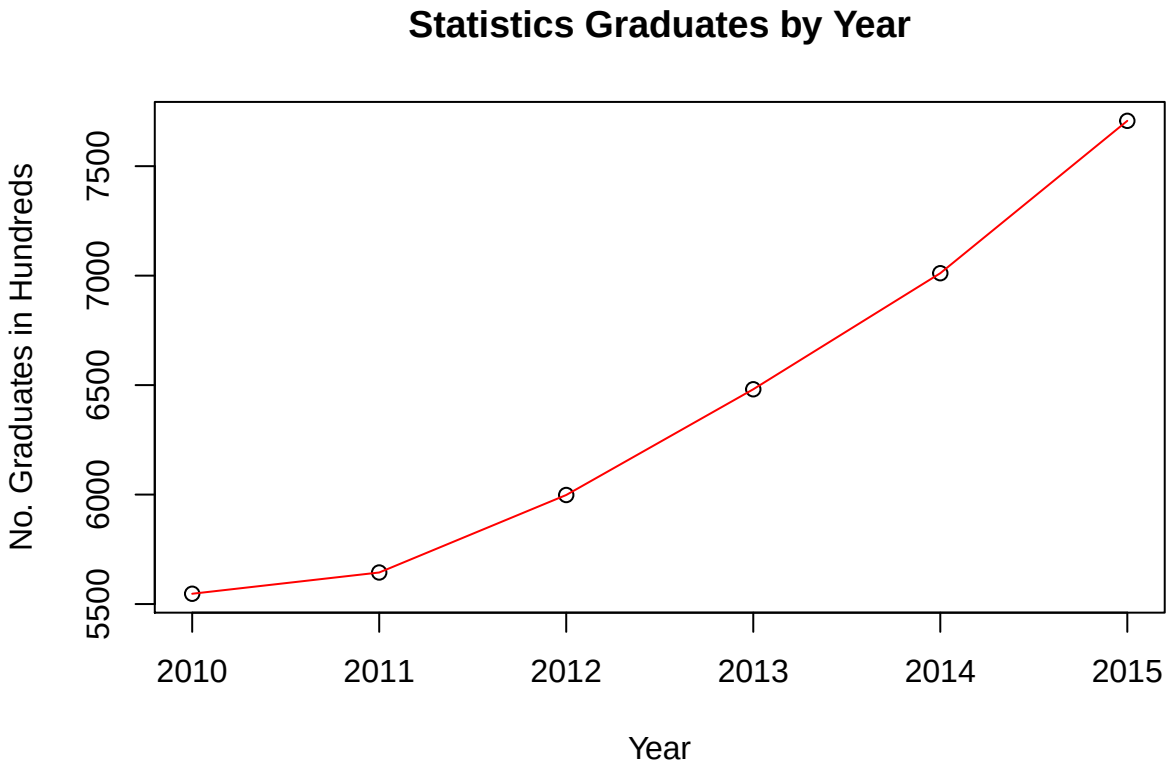
From the above plot, it is confirmed that the number of Science graduates from 2010 to 2015 increases; however there does appear to be a very small dip from 2011 to 2012.



From the above plot, it is clearer that the number of health graduates decreases. We see that from 2010 to 2011 there is a sharp decrease of over 10,000 graduates followed by a small increase from 2013 to 2015.

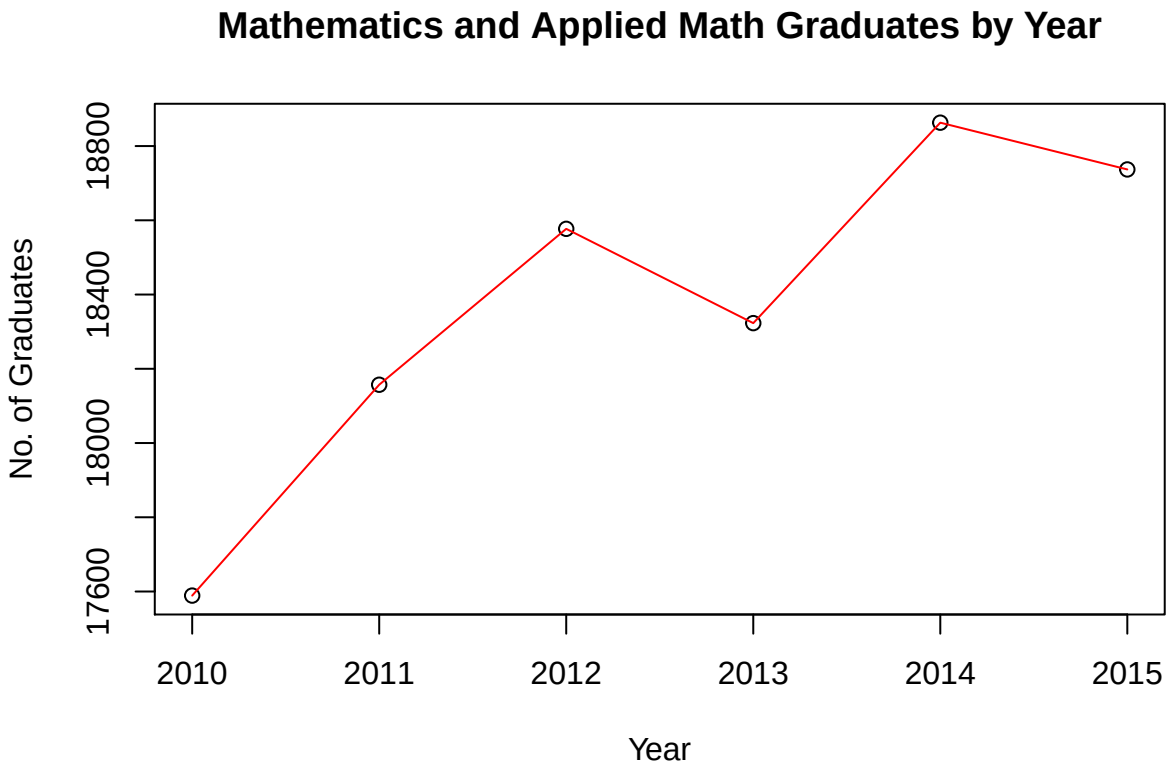


Similar to the number of science graduates, we see an overall increase from 2010 to 2015. It is more clear in this plot that there was a small dip in 2011, followed by a continuous increase through 2015.

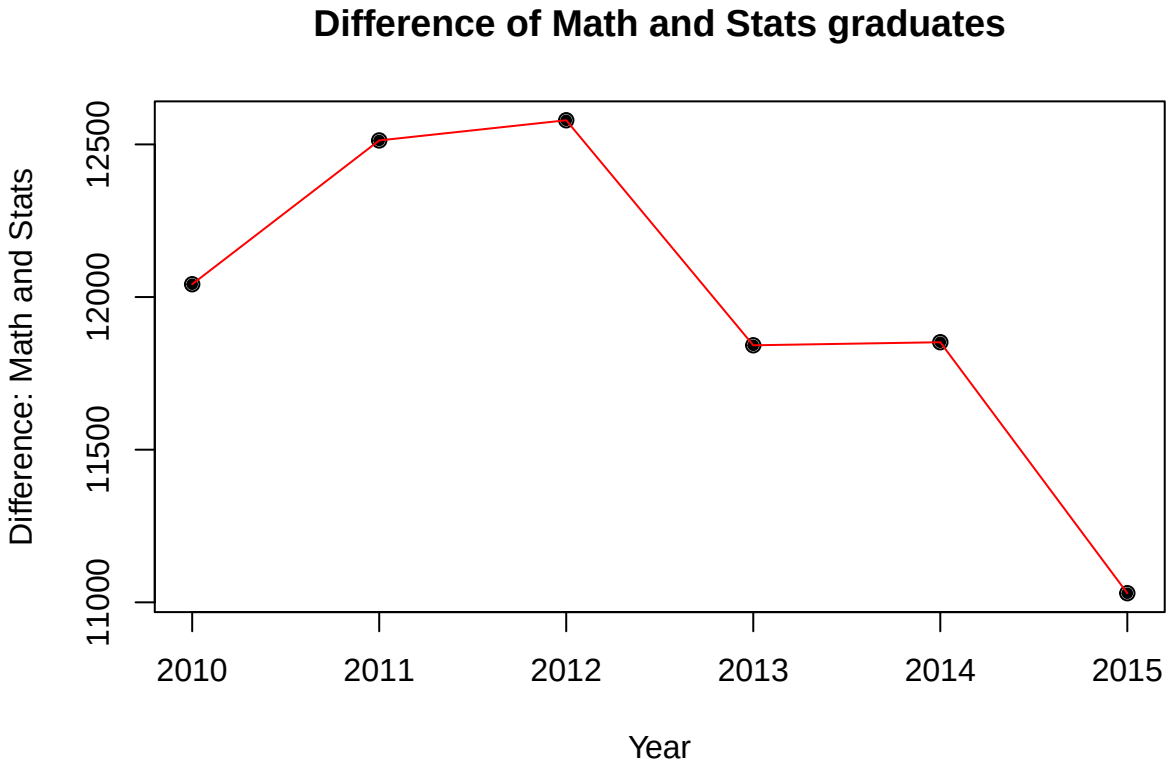


Statistics shows an increase in graduate students from 2010 to 2015, similar to Science and Engineering.

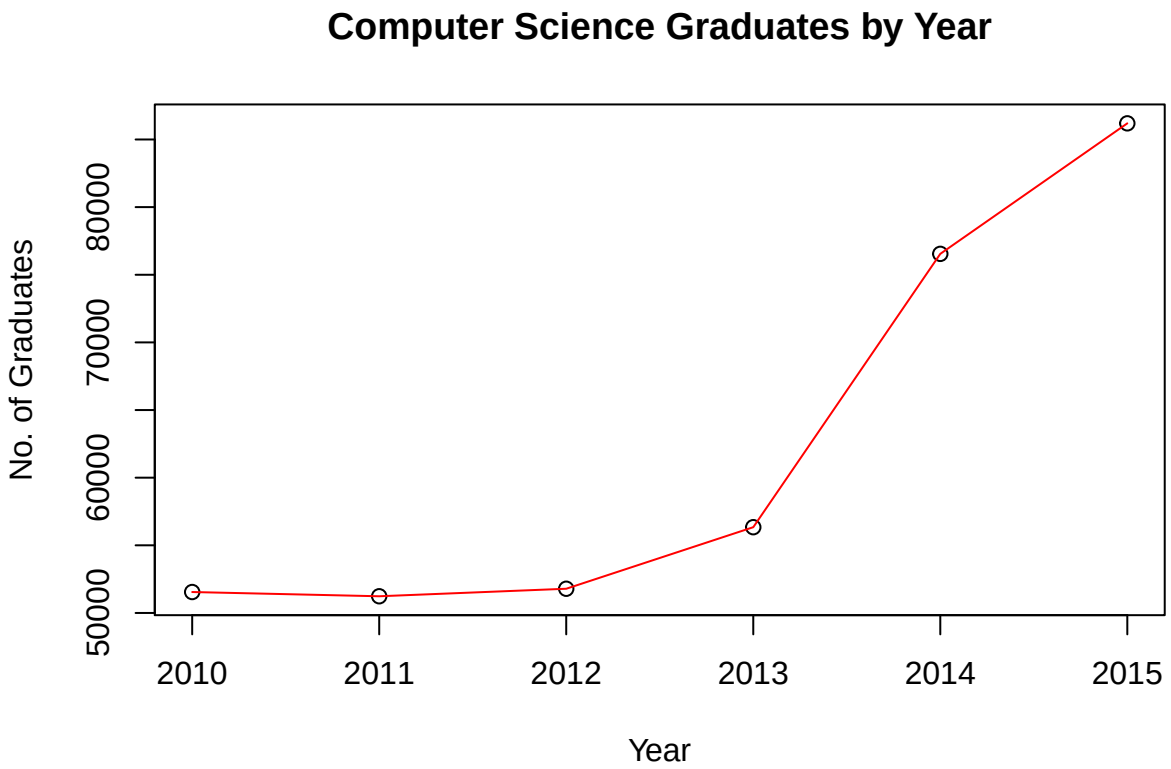
Part B



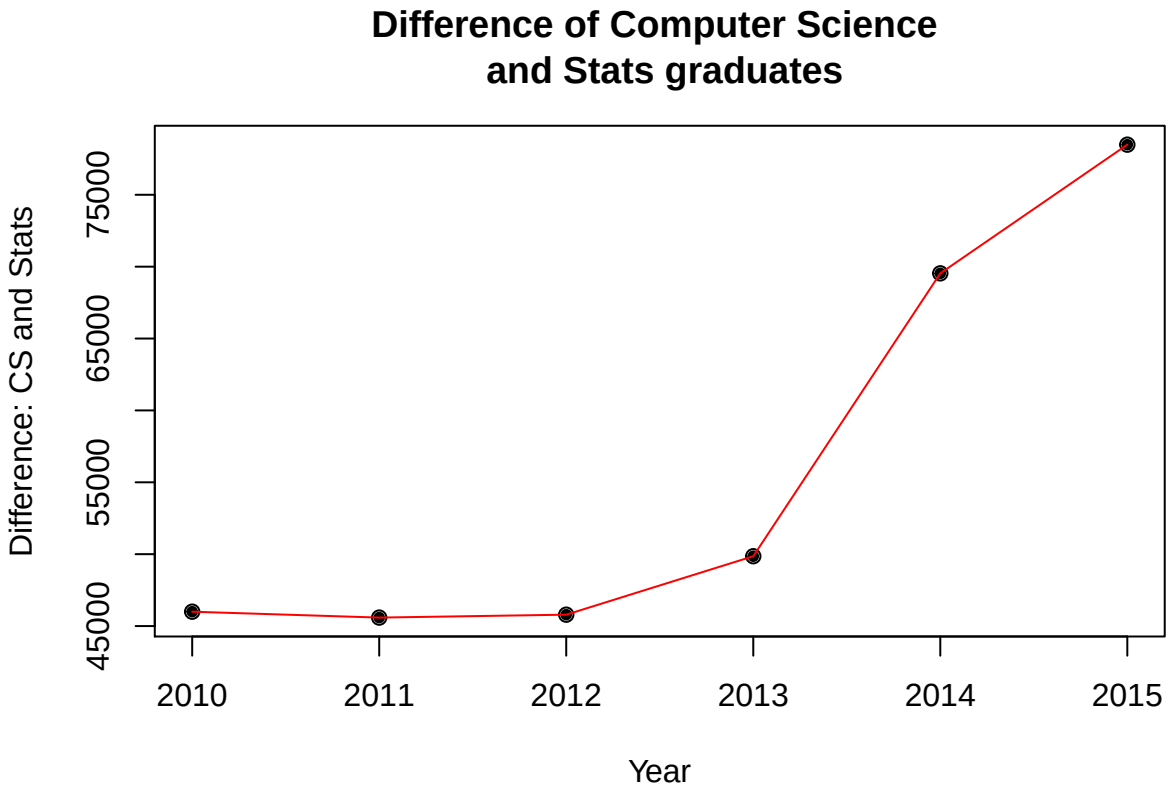
Mathematics and applied math shows an increase from 2010 to 2015, with an unusual dip from 2012 to 2013.



Between 2010 and 2015, the number of math and applied math graduates was greater than the number of stats graduates but we can see that the difference decreases steadily from 2012 to 2015.



The number of computer science graduates remains steady from 2010 to 2012, then sharply increases from 2012 to 2015.



From 2010 to 2015, the number of computer science graduates was greater than the number of statistics graduates, but from 2012 to 2015 the difference increases.