Problem Set 1

Lars Lien Ankile

SECTION LEADER + TIME

Problem 1.

a) This method of selected gives no guarantees about the diverseness of the sample and whether or not this sample will be representative for the whole. E.g. if all the smallest firms happens to come before the larger firms in the phonebook, then the data collected might be very misleading.

b) The readers of Prevention magazine is not unlikely a self-selected group of people who care a lot about their health in the first place, and that is why they chose to subscribe to the magazine. This will in turn result in a very unrepresentative sample for the population at large.

c) Even though the clinics are accessible to anyone, that doesn't mean that all homeless people would actually go to a clinic. I find it more likely that the homeless who are actually suffering from mental illness are more likely to go to a clinic than those who are not suffering.

Problem 2.

a) It can be dangerous to mix up correlation with causation, i.e. just because two events often are observed together, that doesn't necessarily mean that one event is making the other happen. This headline makes it sound like people who are more optimistic suffer less from cardiovascular disease *because* they're more positive. If that is true is unknown and the only thing we can conclude from this is that people who are optimistic also *happen* to also suffer less from cardiovascular disease, for *some* reason.

b) To have a study actually say something about causation and not only correlation, one would have to design a study that controls for all spurius causes that could impact both positivity and cardiovascular disease. Somehow, if one could take a random sample of the population, divide it into two groups with roughly similar people in both groups, and then make one group think optimistically and make the other think pessimistically, while trying to control for any other effect, one could potentially say a little bit more about the effect, I'd imagine.

c) Just because someone is on average less likely to have some event occur it doesn't mean that all optimistic people have lower risk of cardiovascular events than all pessimistic people. Surely there will be pessimistic people who just happen to have a great cardiovascular system and will live to a hundred, while there are optimistic smokers who'll die from a hearth attack tomorrow.

Problem 3.

a) Both graphs shows the distribution of the ages that women had their first child for a given year. The year is 1980 for the first and 2016 for the second. In 1980, for example, we see that some, but very few people had their first child at 14, while having one's first child was the most common at roughly 19. The graph is not symmetric as it has a long tail to the right of the distribution and is therefore right-skewed. The graph for 2016 shows a shift of some of the population to the right. Also, there's two distinct tops on the graph. This graph also has a tail to the right which makes it right-skewed too, but not as much as the former one, though.

b)
   i. I think that the change in the distribution can be best explained by an increase in the percentage of women who take higher education.

   ii. There's a top at 19 as in the graph for 1980, but now there's also one at age 28. This suggests to me that many people self-select into two groups, presumably based on whether they take higher education or not. The second peak on the graph comes from people who have taken education and waited until they are more established before they have their first child.

Problem 4.

a)

```
#load the data
load("datasets/gun_deaths.Rdata")

#numerical summaries
# Females
fem_mean = mean(gun_deaths$age[gun_deaths$sex == 'F'], na.rm = TRUE)
fem_median = median(gun_deaths$age[gun_deaths$sex == 'F'], na.rm = TRUE)

sprintf('For females the mean age is %.2f and the median age is %.2f', fem_mean, fem_median)
```

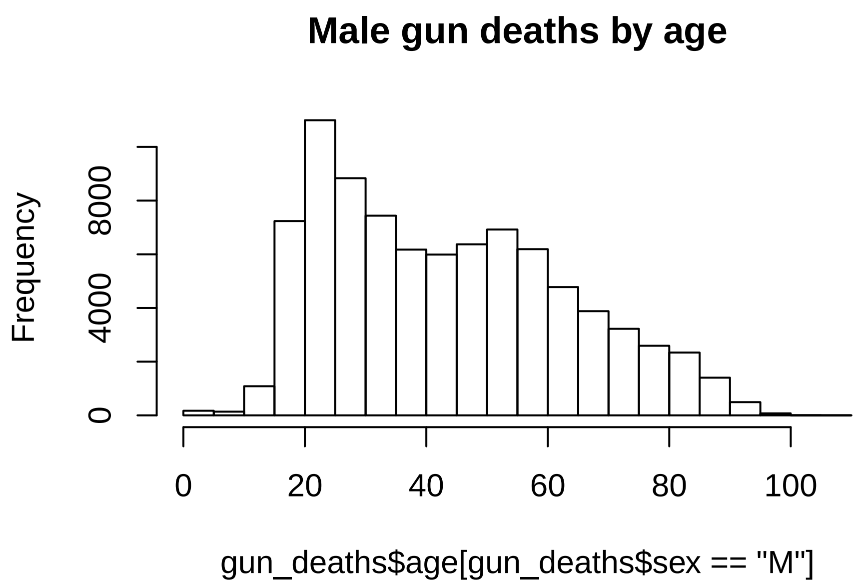
```
## [1] "For females the mean age is 43.70 and the median age is 44.00"
```

```
# Males
male_mean = mean(gun_deaths$age[gun_deaths$sex == 'M'], na.rm = TRUE)
male_median = median(gun_deaths$age[gun_deaths$sex == 'M'], na.rm = TRUE)

sprintf('For males the mean age is %.2f and the median age is %.2f', male_mean, male_median)
```

```
## [1] "For males the mean age is 43.88 and the median age is 41.00"
```

```
#graphical summaries

# Females
hist(gun_deaths$age[gun_deaths$sex == 'F'], main = "Female gun deaths by age")
```


```
# Males
hist(gun_deaths$age[gun_deaths$sex == 'M'], main = "Male gun deaths by age")
```



From the above we see that the female distribution is slightly left-skewed, while the male distribution is right-skewed (mean is smaller than median in the first case and mean is greater than median in the second case). The difference between the mean and median is relatively small in the case for the females, though. From the histograms we see that the most common age to be killed at as a man is right after 20, which is what one might expect. A little more surprising, to me at least, is that the highest rate of gun deaths for females happen between 40 and 60.

b)

```
table(gun_deaths$intent)
```

```
##
##   Accidental     Homicide      Suicide Undetermined
##         1639        35176        63175          807
```

We see that suicide is by far the most common category for cause of gun death. Homicides also contribute a lot, while accidental deaths and undetermined deaths are much less common.

c)

```
# Get info
table(gun_deaths$intent, gun_deaths$police)
```

```
##
##                    No   Yes
##   Accidental     1639     0
##   Homicide      33774  1402
##   Suicide       63175     0
##   Undetermined    807     0
```

```
# Calculate proportion
total = 1402 + 33774
1402 / (total)
```

```
## [1] 0.03985672
```

```r
total
```

```
## [1] 35176
```

1402 out of a total of 35176, or about 4.5%, of the gun deaths classified as homicides was caused by police intervention.

d)

```r
table(gun_deaths$month)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12
## 8273 7093 8289 8455 8669 8677 8989 8783 8508 8406 8243 8413
```

```r
num_spring = length(gun_deaths$month[gun_deaths$month >= 3 & gun_deaths$month < 6])
num_summer = length(gun_deaths$month[gun_deaths$month >= 6 & gun_deaths$month < 9])
num_fall = length(gun_deaths$month[gun_deaths$month >= 9 & gun_deaths$month < 12])
num_winter = length(gun_deaths$month[gun_deaths$month >= 12 | gun_deaths$month < 3])

sprintf("Gun death per season: Spring: %d, Summer: %d, Fall: %d, Winter: %d.", num_spring, num_
```

```
## [1] "Gun death per season: Spring: 25413, Summer: 26449, Fall: 25157, Winter: 23779."
```

The summer was the season with the most gun deaths with its 26449 gun deaths.

e)

```r
deaths.2012 = gun_deaths[gun_deaths$year == 2012 & gun_deaths$education %in% c("HS/GED", "Some
table(deaths.2012$race)
```

```
##
##         Asian/Pacific Islander                          Black
##                            372                           5207
##                       Hispanic Native American/Native Alaskan
##                           1651                            197
##                          White
##                          18121
```

```r
# Missing values tho?

white_deaths = nrow(deaths.2012[deaths.2012$race == 'White', ])
total = nrow(deaths.2012)

sprintf("Out of the total of %d deaths in 2012 among people with at least high school educatio
```

```
## [1] "Out of the total of 25548 deaths in 2012 among people with at least high school educati
```

```r
sprintf("I.e. %.2f%% of the people killed were white.", white_deaths * 100 / total)
```

```
## [1] "I.e. 70.93% of the people killed were white."
```

As we can se from the above, the majority of the people killed were white with roughly 71% being white.

f)

```
table(gun_deaths$race, gun_deaths$intent)
```

```
##
##                                  Accidental Homicide Suicide Undetermined
##   Asian/Pacific Islander                 12      559     745           10
##   Black                                 328    19510    3332          126
##   Hispanic                              145     5634    3171           72
##   Native American/Native Alaskan         22      326     555           14
##   White                                1132     9147   55372          585
```

From the above I notice especially two glaring things from the data. First, for black people, there are almost 6 times as many homicides as there are suicides. Second, for white people, there are almost 6 times as many suicides than homicides, i.e. the exact opposite of black people.

Problem 5.

a) i.

```r
#load the data
cps = read.csv("datasets/CPSData.csv", header = TRUE)

#explore age
hist(cps$Age)
```


```r
#explore sex
table(cps$Sex)
```

```
##
## Female   Male
##  67481  63821
```

```r
#explore race
table(cps$Race)
```

```
##
##  American Indian              Asian              Black        Multiracial
##             1433               6520              13913               2897
## Pacific Islander              White
##              618             105921
```

From the age histogram above we see that there's a relatively even distribution of people in the ages between 0 and 60. After 60 there's a dropoff of number of people, which makes sense. From the first table above we see that there's a little more women than men in the sample, which makes sense given that there's more women in society at large as well, presumably because women live longer. When it comes to race, we see that there's by far most white people with almost 106 thousand people. Black people comes second with almost 14 thousand, which is significantly less.

ii. One thing I find slightly unusual about the age of the people is that there's a relatively large spike in the number of 75-80 year-olds as compared to those 70-75 and 80-85. One would suspect a more "linear" descent. However, these people are the people who would've been born after the second world war, which makes sense because we know that there was a big baby boom after the war.

b)

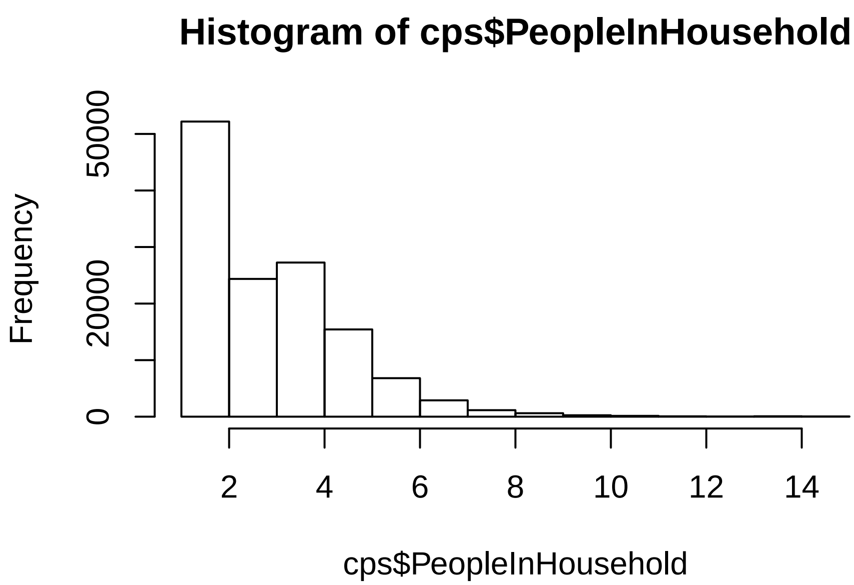
```
#numerical summaries
summary(cps$PeopleInHousehold)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.284   4.000  15.000
```

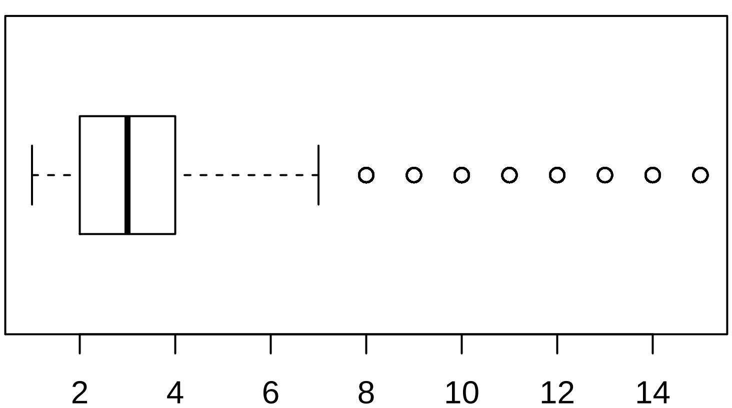
```
IQR(cps$PeopleInHousehold)
```

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

```
#graphical summaries
hist(cps$PeopleInHousehold)
```




```
boxplot(cps$PeopleInHousehold, horizontal = TRUE)
```



The median number of people in a household is 3, while the mean is 3.284, which means our data is right-skewed. This makes sense to me because most people are either living alone or in a family of 2, 3 or 4 people. Still, there are people who have much larger households. This is confirmed by the histogram. The boxplot confirmes that there are a lot of outliers who have a lot of people in their households.

c)

```r
summary(cps$Citizenship)
```

```
##      Citizen, Native Citizen, Naturalized          Non-Citizen
##               116639                 7073                 7590
```

```r
nrow(cps[cps$Citizenship == "Citizen, Native", ]) / nrow(cps)
```

```
## [1] 0.8883261
```

The vast majority of people are native citizens, with 89% of people belonging to that category. The remaining 11% is more or less evenly divided between naturalized citizens and non-citizens.

d)

```r
prop.table(table(cps$Race, cps$Hispanic), 1)
```

```
##
##                            0          1
##   American Indian  0.78785764 0.21214236
##   Asian            0.98266871 0.01733129
##   Black            0.95536549 0.04463451
##   Multiracial      0.84535727 0.15464273
##   Pacific Islander 0.87540453 0.12459547
##   White            0.84204265 0.15795735
```

From the above we see that American Indians, Multiracials, and Whites are the races where at least 15% identify as hispanic.

e)

```r
table(cps$Age, cps$Married)
```

```
##
##      Divorced Married Never Married Separated Widowed
##   0         0       0             0         0       0
##   1         0       0             0         0       0
##   2         0       0             0         0       0
##   3         0       0             0         0       0
##   4         0       0             0         0       0
##   5         0       0             0         0       0
##   6         0       0             0         0       0
##   7         0       0             0         0       0
##   8         0       0             0         0       0
##   9         0       0             0         0       0
##   10        0       0             0         0       0
##   11        0       0             0         0       0
##   12        0       0             0         0       0
##   13        0       0             0         0       0
```

```
## 14     0    0     0   0  0
## 15     6   19  1753  16  1
## 16    10    8  1721   8  4
## 17     5   19  1729  10  1
## 18     6   21  1559   9  1
## 19     8   47  1448  14  0
## 20     6   56  1328   5  3
## 21    11  118  1378  16  2
## 22    22  187  1306  17  4
## 23    18  266  1333  18  3
## 24    27  323  1245  30  2
## 25    31  411  1132  29  1
## 26    54  484  1064  32  9
## 27    60  587   975  27  8
## 28    80  761   857  35  3
## 29    75  814   731  21  4
## 30    82  930   791  46  5
## 31    96  922   704  37  3
## 32   129 1037   573  41 10
## 33   135 1089   534  42  4
## 34   130 1000   483  34  6
## 35   161 1047   433  62 13
## 36   142 1022   443  48  8
## 37   138 1016   325  43  9
## 38   191  958   323  46 12
## 39   161 1060   264  45 12
## 40   204 1034   276  46 11
## 41   211 1126   266  54 16
## 42   242 1149   257  50 13
## 43   256 1218   270  50 25
## 44   232 1199   255  49 29
## 45   272 1115   278  58 26
## 46   251 1106   243  42 23
## 47   241 1090   240  56 20
## 48   290 1198   228  45 30
## 49   312 1339   242  60 36
## 50   341 1290   243  47 45
## 51   312 1269   256  53 41
## 52   326 1285   231  41 52
## 53   350 1275   266  53 50
## 54   357 1228   216  55 56
## 55   329 1279   187  36 64
## 56   309 1312   191  38 85
## 57   324 1211   186  41 65
## 58   312 1232   200  41 89
## 59   286 1194   161  37 80
## 60   279 1195   158  34 80
## 61   297 1142   165  34 97
```

```
##   62      252    1097           136        28      82
##   63      260    1065           118        28     125
##   64      268    1011           119        22      99
##   65      251    1057           103        20     138
##   66      228    1073           101        23     152
##   67      173     837            73        14     130
##   68      179     741            71        11     128
##   69      140     705            52        13     152
##   70      165     779            59        12     180
##   71      135     655            56        18     167
##   72      107     624            38        12     160
##   73       98     583            27         8     180
##   74       98     533            39         9     163
##   75       82     456            28        10     187
##   76       63     431            28         8     199
##   77       69     424            26         7     172
##   78       73     372            22         4     188
##   79       54     352            36         3     216
##   80      219    1269            91        15    1070
##   85      120     757           102        11    1456
```

From the above table we see that the amount of married people increases with age up to a certain point, i.e. around 50. After this point we see a lot of divorces in the years following, but the number of divorces is decreasing as people get older, which makes sense, I think. Also, the number of widowed people increase as people age, which also makes a lot of sense. What I found surprising, though, was the amount of 15-year olds that were divorced, a whopping 6 poeple, not to mention the one 15-year old who is widowed and the 16 who are separated.

Problem 6.

```r
#load quantmod package
library(quantmod)

#load AAPL and MSFT data
getSymbols("AAPL", from = "2018-01-01", to = "2019-07-22")
```

```
## [1] "AAPL"
```

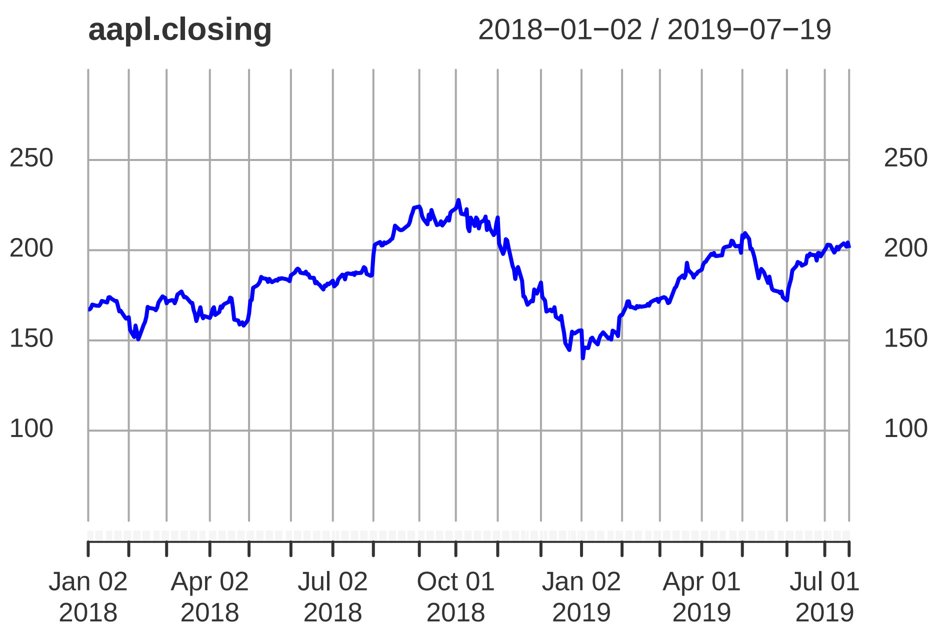
```r
getSymbols("MSFT", from = "2018-01-01", to = "2019-07-22")
```

```
## [1] "MSFT"
```

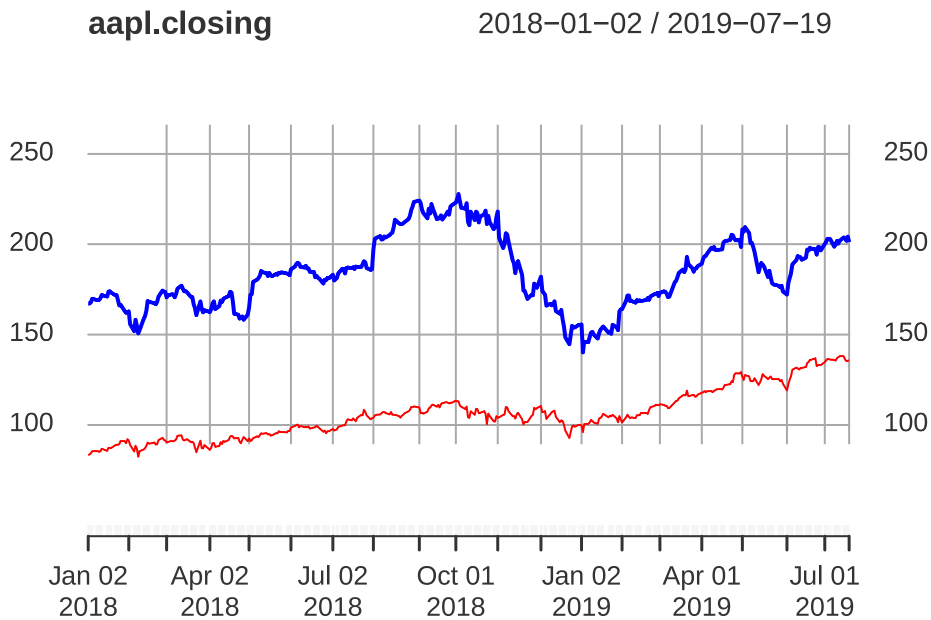
```r
#obtain adjusted closing prices
aapl.closing = Ad(AAPL)
msft.closing = Ad(MSFT)

#obtain daily volume
aapl.volume = Vo(AAPL)
msft.volume = Vo(MSFT)

#obtain daily returns
aapl.return = as.numeric(dailyReturn(Ad(AAPL)))
msft.return = as.numeric(dailyReturn(Ad(MSFT)))
```

a)

```r
#plot prices
plot(aapl.closing, col = "blue", ylim = c(50, 300))
```



```r
lines(msft.closing, col = "red")
```



Both companies had a positive price change over the period. It looks like Microsoft had a slightly better return. One big difference between them is that Apple looks much more volatile and more risky. It grew more from Jan '18 to Oct '18 than Microsoft, but Microsoft didn't have the drastic drop in price between Oct '18 and Jan '19 as Apple had. There's a similar effect in May '19. Microsoft might've been a better investment in this period.

b)

```r
#example: summary of aapl.closing
summary(as.numeric(aapl.closing))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   140.1   169.2   183.1   184.2   198.7   227.8
```

```r
#return for aapl
100 * (as.numeric(aapl.closing)[length(aapl.closing)] - as.numeric(aapl.closing)[1]) / as.nume
```

```
## [1] 20.34641
```

```r
#return for msft
100 * (as.numeric(msft.closing)[length(msft.closing)] - as.numeric(msft.closing)[1]) / as.nume
```

```
## [1] 62.99628
```

From this we see that Microsoft had a much better return than Apple in this period. The fact that the scale on the y-axis is linear and Microsoft's price started out lower might be a little misleading in the graph above.

c)

```r
#sd for aapl
sd(aapl.closing)
```

```
## [1] 19.54567
```

```r
#sd for msft
sd(msft.closing)
```

```
## [1] 13.54294
```

From the above we see that Microsoft is indeed the least volatile of the two stocks.

d)

```r
# Min and max for Apple
sprintf('Apple: Max value: %.2f on %s, and min value: %.2f on %s',
        max(aapl.closing),
        index(aapl.closing[which.max(aapl.closing)]),
        min(aapl.closing),
        index(aapl.closing[which.min(aapl.closing)]))
```

```
## [1] "Apple: Max value: 227.84 on 2018-10-03, and min value: 140.09 on 2019-01-03"
```

```r
# Min and max for Microsoft
sprintf('Microsoft: Max value: %.2f on %s, and min value: %.2f on %s',
        max(msft.closing),
        index(msft.closing[which.max(msft.closing)]),
        min(msft.closing),
        index(msft.closing[which.min(msft.closing)]))
```

```
## [1] "Microsoft: Max value: 137.97 on 2019-07-12, and min value: 82.35 on 2018-02-08"
```

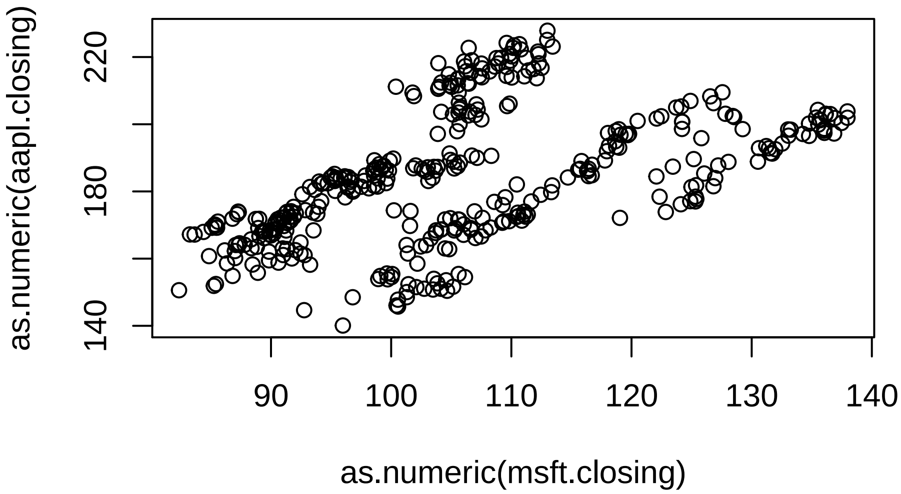
See above.

e)
  i. From the graph in a) above, it seems like Microsoft and Apple are pretty correlated, at least when it comes to the direction of the price movement, i.e. when apple goes up, microsoft does too, and vice versa. However, we see that Apple's movements are much more extreme than Microsoft's.

  ii. Used cor to calculate how correlated the stock prices are. See below.

  iii. The correlation between the stock prices is at 0.5153. This is a positive number which means that there is a positive correlation between them. This means that the price of one tends to move in the same direction as the other over time. However, we see that the correlation is at roughly 0.5, which means that this is by no means a perfect correlation, i.e. there is a lot of variability between the stocks, too.

```r
#graphical summary (part i.)
plot(as.numeric(msft.closing), as.numeric(aapl.closing))
```




```
#numerical summary (part ii.)
cor(as.numeric(msft.closing), as.numeric(aapl.closing))
```

```
## [1] 0.5152774
```