Lucas Faijdherbe(2594812) & Ruben van der Ham(2592271) – CS40
Assignment 1

Exercise 1.1
10. Sampling method: Voluntary response. This is a flawed sampling method, since respondents can decide themselves if they want to respond, which can lead to a 'biased sample.

12. Sampling method: Randomized sample. This can be considered a sound approach, since everybody has an equal chance of getting picked. The group has a substantial size of over a 1000 respondents, which also helps in making sure there will really be different kinds of people responding.

26. Everybody can respond to this poll. This means that not only students that follow a college major can respond, but also people that already have a job, etc. Now, the number says absolutely nothing. It might be that all college major students responded that their major prepared them for their chosen careers (so, 41%), but that the other 59% of the respondents did not even follow a college major, leading to a flawed conclusion. It would be better to only poll students.

Exercise 1.2
22. The level of measurement of the depth is ratio. The measurements can be ordered, and also have significance; the difference in depths can be useful for earthquake research.

32. This is a nominal level of measurement. There is nothing meaningful to the numbers; they just 'represent' the player. Because the numbers have no meaning, when ordered the mean of these numbers doesn't mean (pun not intended) anything.

Exercise 1.3
6. The study described corresponds to an experiment. The subjects were given a treatment and therefore they are modified. An observational study requires the subjects to be unmodified and only observed. Therefore the study is an experimental study instead of an observational study.

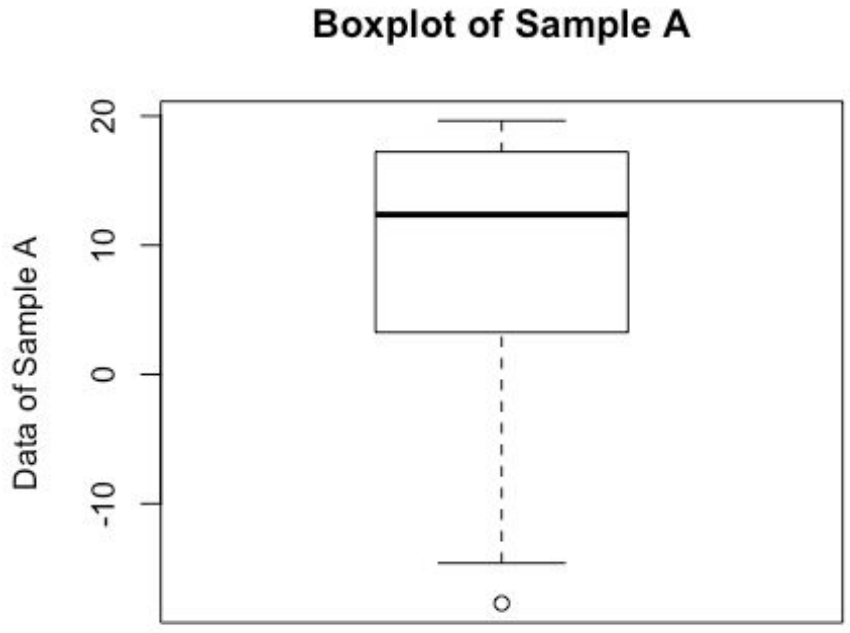
12. Type of sampling used: systematic sampling

18. Type of sampling used: cluster sampling

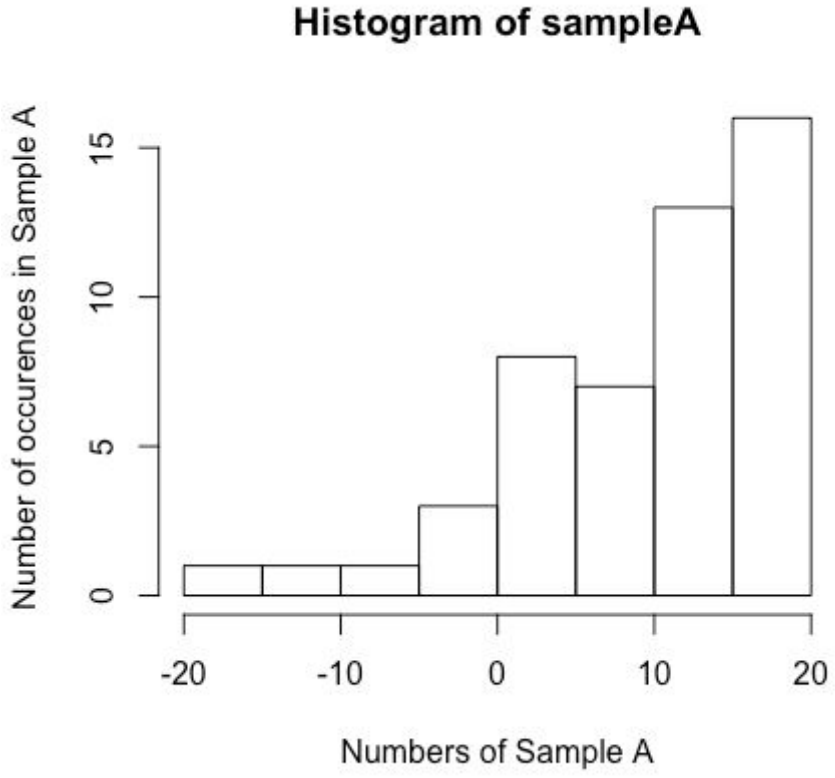
Exercise 1.4
a. The Y-axis starts at 50 instead of 0. It seems like CDU has 10, maybe even 20 times more seats than the CSU, but in reality, they only have 4x more seats. The representation exaggerates the difference between the parties.

b. The bar graph: The bar graph can showcase different categories, and their corresponding 'value', in this case the average number of posts per day; this is something a histogram cannot do, since histograms represent some range of numbers, and the frequency of their occurrence, something we do not need for this kind of data.

Exercise 1.5



a.



b. We used the function summary(sampleA) in R and got the following results:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-17.710   3.653  12.370   9.819  17.100  19.640
```

We used the function sd(sampleA) to get the standard deviation of A, which yielded the following result:

```
8.911
```

c. To start off, we have a minimal value of -17.71, and a maximal value of 19.64. This gives us a range of 37.35.

The shape of the data is left-skewed when plotted in a histogram (see 1.5a), so we have more occurrences of high numbers (> 0), than we have of 'low' numbers (< 0).

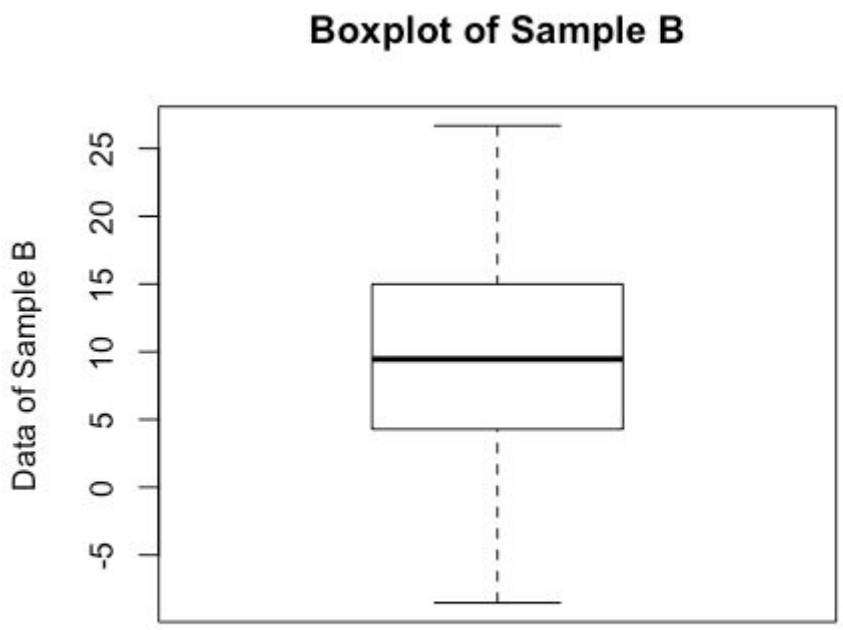
The data is located around the 0 on the x-axis (see 1.5a).

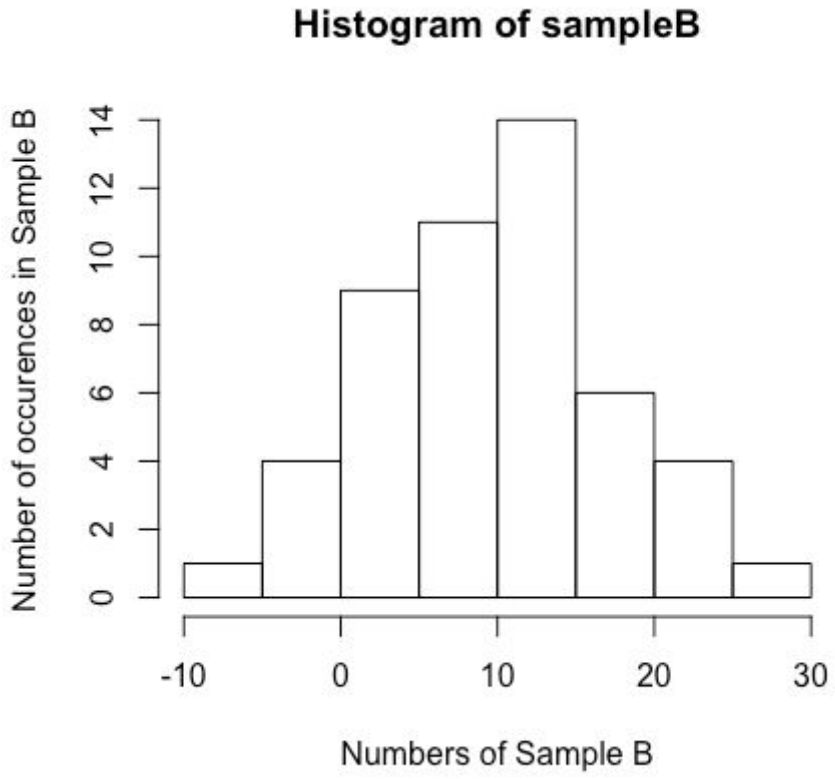
The median of the set is 9.82, and the standard deviation is 8.91, as mentioned above.

Overview of the quartiles:
1st quartile: 25% = 3.65
2nd quartile: 50% = 12.37 (mean)
3nd quartile: 75% = 17.097
So we see, even though our lowest value is -17.710, the numbers start getting positive quite fast. After the first quarter, we already have positive numbers (3.65). This shows again that we have a lot more occurrences of numbers greater than 0, then we have smaller than 0, something that can also be seen in the boxplot.

It seems like a big part of the numbers in the data set can be found in the range [10,20], which have a high number of occurrences in the histogram.

d. Boxplot and histogram of sample B:





The function summary(sampleB) gives us:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -8.526   4.319   9.458   9.577  14.930  26.680
```

The standard deviation of sampleB, with function sd(sampleB) gives us:

```
7.690578
```

The minimal value of the set is -8.53, and the maximal value is 26.68. This gives us a range of 35.21.

The shape of the data is almost symmetrical, with a little skew to the right (see histogram above). The most numbers are somewhere in the range [0,15].

The data is located around the 10 on the x-axis.

We have a median of 9.577, and a standard deviation of 7.69.

Overview of the quartiles:
1st quartile: 25% = 4.32
2nd quartile: 50% = 9.46 (mean)
3rd quartile: 75% = 14.93
We mostly have positive numbers in this set, given that the minimal value is -8.53, and the maximal is 26.68; so it is no surprise that the first quartile starts somewhere just above 0. If we look at the boxplot, we see that everything is distributed nicely (which can also be seen on the histogram). The 'middle' of the data set (numbers from the first quartile to the third quartile), seems to be really lying in the middle.

e. We think the two data sets *do* originate from the same population distribution. Both samples have almost the same mean: 9.819 for sample A, and 9.577 for sample B. The standard deviations of the two do differ a bit though: 8.91 vs 7.69, which means the data in sample A is a bit more dispersed. However, if we look at the boxplots (or just at the quartiles), we can see that the first and the third quartile are very close to each other; 3.65 to 17.09 in sample A, versus 4.31 to 14.93 in sample B. This means that we have do have a very similar data dispersion from the first to the third quartile.

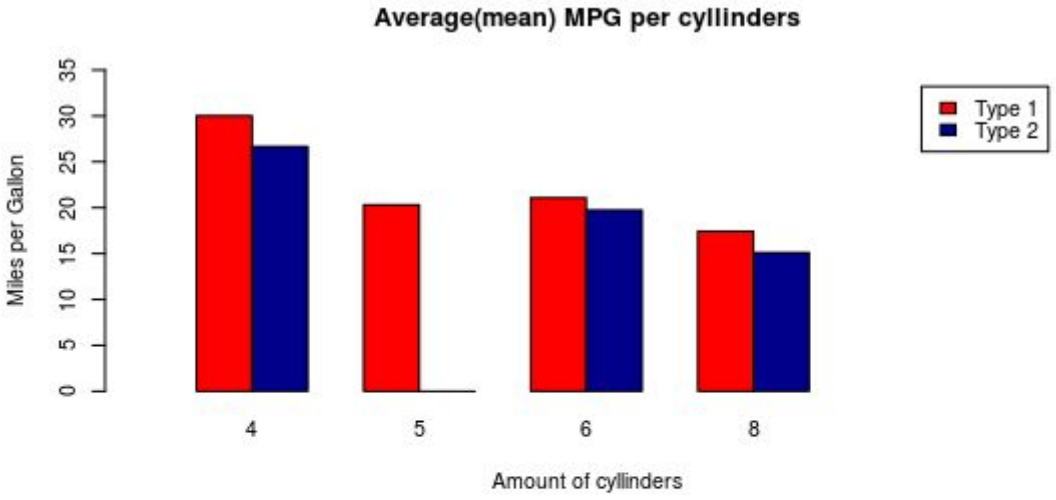
If we look at the extremes and ranges of both samples, we see that we have a very similar range (37.35 for sample A, 35.21 for sample B), but that the minimal and maximal values do differ a bit; this can of course happen when we take a sample: those are the extreme cases. It seems that sample A has some more numbers < - 5, and sample B has some more numbers > 20. It does even seem like sample A is the same histogram as sample B, but it stopped after the number 19.

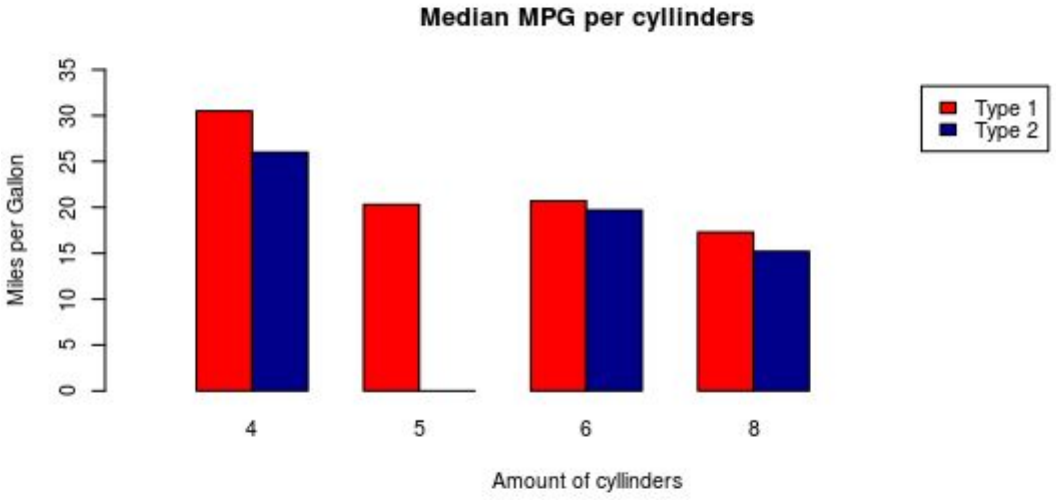
It seems safe to say that even though there are some differences in the data, the data has a very similar mean, a very similar range and a very similar dispersion of numbers in the first to third quartile, that the data does come from the same population.

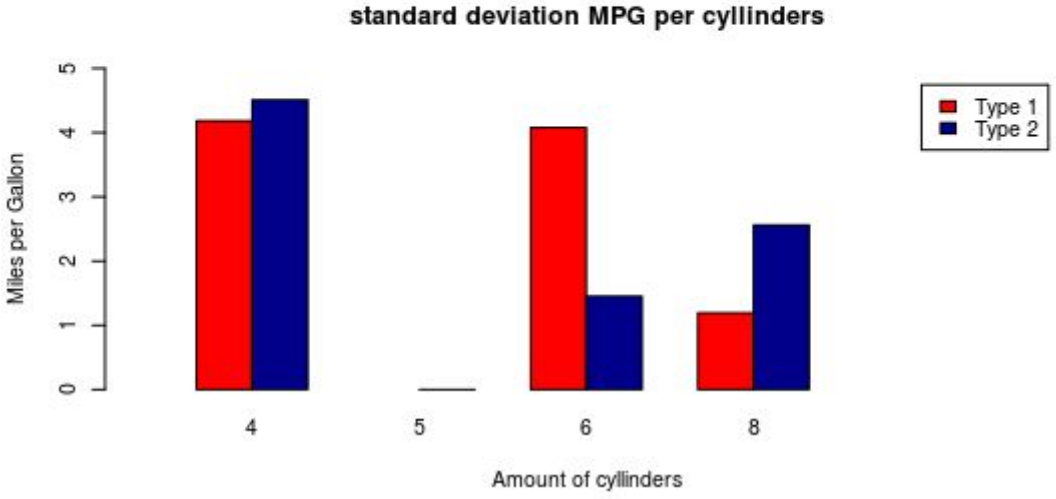
Exercise 1.6

Graphical Summary:

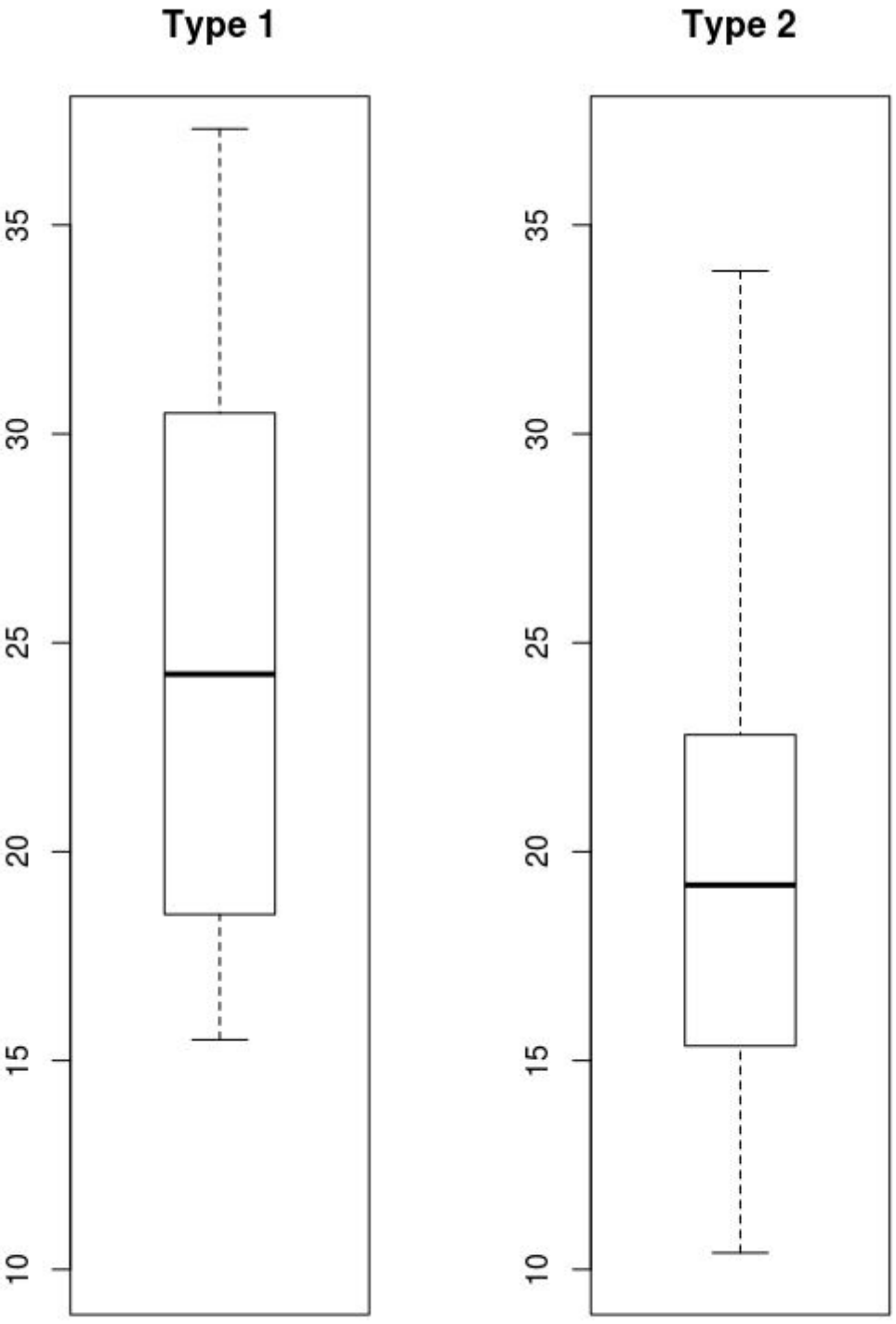
mean, median and Standard Deviation:







Boxplots:



Numerical Summary:

MPG's sample type 1

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15.50	18.52	24.25	24.76	30.38	37.30

standard deviation: 6.55

MPG's sample type 2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.40	15.42	19.20	20.09	22.80	33.90

Standard deviation: 6.03

As seen in the first 3 plots the mean and median MPG is for every amount of cylinder higher for car type 1 than car type 2. However, the standard deviation shows for 6 cylinders the deviation is higher for car type 1, so type 2 performs more consistent with 6 cylinder engines than type 1 with the same amount of cylinders.

Bottom line from the median, mean and standard deviation:
-Type 1 performs more consistent MPG rates, with the exception of 6 cylinder engines. Type 2 6cyl is more accurate than type 1 6cyl.
-Type 1 performs overall on mean and median values for each amount of cylinders better than type 2 does.

From the boxplots we can conclude from the IQR and median, the type 1 car performs overall a higher MPG compared to the type 2 car.
However, the type 1 car sample may have more 4 cylinder, thus more efficient engines in its sample which make this boxplot an inaccurate summary.

Amount of Cylinders sample type 1
Mean: 5.394
Median:4.5

Amount of Cylinders sample type 2
Mean: 6.188
Median:6

The boxplot of both sample MPG's does not give an accurate "side by side" insight of the MPG per type of car because of the type 2 data contains much more generically inefficient high cylinder engines.

Appendix: R code

1.5a)

```
> sampleA = scan("/Users/lucasfaijdherbe/Library/Mobile
Documents/com~apple~CloudDocs/Computer Science/Statistical
Methods/Assignments/Assignment 1/Exercises/sampleA.txt")
Read 50 items
> boxplot(sampleA,main="Boxplot of Sample A",ylab="Data of
Sample A")
> hist(sampleA, xlab="Numbers of Sample A", ylab = "Number of
occurences in Sample A")
```

b)

```
> summary(sampleA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-17.710   3.653  12.370   9.819  17.100  19.640
> quantile(sampleA,0.5)
     50%
12.37158
> sd(sampleA)
[1] 8.910884
```

d)

```
> sampleB = scan("/Users/lucasfaijdherbe/Library/Mobile
Documents/com~apple~CloudDocs/Computer Science/Statistical
Methods/Assignments/Assignment 1/Exercises/sampleB.txt")
Read 50 items
> boxplot(sampleB,main="Boxplot of Sample B",ylab="Data of
Sample B")
> hist(sampleB, xlab="Numbers of Sample B", ylab = "Number of
occurences in Sample B" )
> summary(sampleB)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -8.526   4.319   9.458   9.577  14.930  26.680
> quantile(sampleB,0.5)
     50%
9.457853
> sd(sampleB)
[1] 7.690578
```

1.6:

```
par(mfrow=c(3,1))

means = list(means[[1]],means[[2]],c(), c())
means[[4]][3:4]  = means[[2]][2:3]
means[[4]][1:2] = c(means[[2]][1],0)
means[[3]] = rbind(means[[1]],means[[4]])
rownames(means[[3]]) = c("Type 1","Type 2")
means[[3]]
barplot(means[[3]],names.arg = groupedPerCyl1[,"Cylls"],ylim =
c(0,35),col=c("red","darkblue"),width = (0.5),xlim = c(0,8),ylab =
"Miles per Gallon", xlab = "Amount of cyllinders", main =
"Average(mean) MPG per cyllinders",beside = T,legend =
rownames(means[[3]]))

medians = list(medians[[1]],medians[[2]],c(), c())
medians[[4]][3:4]  = medians[[2]][2:3]
medians[[4]][1:2] = c(medians[[2]][1],0)
medians[[3]] = rbind(medians[[1]],medians[[4]])
rownames(medians[[3]]) = c("Type 1","Type 2")
medians[[3]]
barplot(medians[[3]],names.arg = groupedPerCyl1[,"Cylls"],ylim =
c(0,35),col=c("red","darkblue"),width = (0.5),xlim = c(0,8),ylab =
"Miles per Gallon", xlab = "Amount of cyllinders", main = "Median
MPG per cyllinders",beside = T,legend = rownames(medians[[3]]))

sds = list(sds[[1]],sds[[2]],c(), c())
sds[[4]][3:4]  = sds[[2]][2:3]
sds[[4]][1:2] = c(sds[[2]][1],0)
sds[[3]] = rbind(sds[[1]],sds[[4]])
rownames(sds[[3]]) = c("Type 1","Type 2")
sds[[3]]
barplot(sds[[3]],names.arg = groupedPerCyl1[,"Cylls"],ylim =
c(0,5),col=c("red","darkblue"),width = (0.5),xlim = c(0,8),ylab =
"Miles per Gallon", xlab = "Amount of cyllinders", main =
"standard deviation MPG per cyllinders",beside = T,legend =
rownames(sds[[3]]))


par(mfrow=c(1,2))

boxplot(mileage$mpg1, main= "Type 1", ylim = c(10,37))
boxplot(mileage$mpg2, main= "Type 2", ylim = c(10,37))
```