Homework 1

Shun-Lung Chang, Dilip Hiremath

1. Before you address your real task you just want to get a bit acquainted with the data. Read the data into R and use the function summary to get a first overview on the data.

```
load("data/Forbes2016.Rdata") # read the data into R
summary(Forbes2016)
```

```
      Rank                            Company              Country
 Min.   :   1   Merck                     :   2   United States :586
 1st Qu.: 500   3i Group                  :   1   Japan         :219
 Median :1000   3M                        :   1   China         :200
 Mean   :1001   77 Bank                   :   1   United Kingdom: 92
 3rd Qu.:1501   AAC Technologies Holdings:   1   South Korea   : 67
 Max.   :2000   Aareal Bank               :   1   France        : 61
                (Other)                   :1994   (Other)       :776
     Sales            Profits            Assets          Market.Value
 Min.   :  0.24   Min.   :-23.100   Min.   :   0.98   Min.   :  0.44
 1st Qu.:  4.30   1st Qu.:  1.500   1st Qu.:  10.40   1st Qu.:  6.30
 Median :  8.90   Median :  3.380   Median :  21.70   Median : 10.90
 Mean   : 17.84   Mean   :  3.562   Mean   :  80.81   Mean   : 22.35
 3rd Qu.: 17.90   3rd Qu.:  5.570   3rd Qu.:  49.15   3rd Qu.: 21.90
 Max.   :482.10   Max.   : 53.700   Max.   :3420.30   Max.   :586.00
 NA's   :4                          NA's   :1         NA's   :1
```

(a) How many companies in the data set are located in China?

There are 200 companies located in China, as one can see from the summary table.

(b) How many companies in the data set have missing values (i.e. are listed as NA)? For which variables are data missing?

If we look at *Assets* and *Market.Value*, there is 1 missing value in both variables. The row containing missing values is shown below.

```
kable(Forbes2016[is.na(Forbes2016$Assets), ])
```

	Rank	Company	Country	Sales	Profits	Assets	Market.Value
1840	1840	Banque nationale de Belgique	Belgium	1.7	9.03	NA	1.3

```
kable(Forbes2016[is.na(Forbes2016$Market.Value), ])
```

	Rank	Company	Country	Sales	Profits	Assets	Market.Value
685	685	Gree Electric Appliances	China	19.7	2.3	26.8	NA

In addition, there exist four rows containing missing values in *Sales*. These rows are shown below.

```
kable(Forbes2016[is.na(Forbes2016$Sales), ])
```

	Rank	Company	Country	Sales	Profits	Assets	Market.Value
672	672	Porsche Automobil Holding	Germany	NA	3.50	34.9	16.60
1072	1072	Kyushu Financial Group	Japan	NA	8.12	73.0	2.60
1398	1398	RHB Capital	Malaysia	NA	3.87	53.7	4.90
1988	1988	Ivanhoe Mines	Canada	NA	7.28	1.0	6.34

(c) Which four countries have the top most companies in the Forbes' list?

As the summary table indicates, the first top countries are United States, Japan, China, and United Kingdom.

(d) Which share of companies in the data set have a market value smaller than or equal to 6.30 billion USD?

The share is 25.24%, and the value can be obtained with the following code.

```
sum(Forbes2016$Market.Value <= 6.3, na.rm = TRUE) / nrow(Forbes2016)
```

```
[1] 0.2523738
```

(e) Which share of companies in the data set have assets that are larger than or equal to 10.40 billion USD?

The share is 75.01%, and the value can be obtained with the following code.

```
sum(Forbes2016$Assets >= 10.4, na.rm = TRUE) / nrow(Forbes2016)
```

```
[1] 0.7501249
```

2. Let us have a closer look at the variable types involved here. Looking at the content of each variable:

(a) Which of the variables in the data set are discrete?

The discrete variables are *Rank*, *Company*, and *Country*.

(b) Which of the variables are continuous?

The discrete variables are *Sales*, *Profits*, *Assets*, and *Market.Value*.

(c) Which discrete ones can be ordered?

The variable *Rank* can be orderd.

3. Which types are used in R to store the variables in the data frame?

The following code returns the class of each variable.

```
sapply(Forbes2016, class)
```

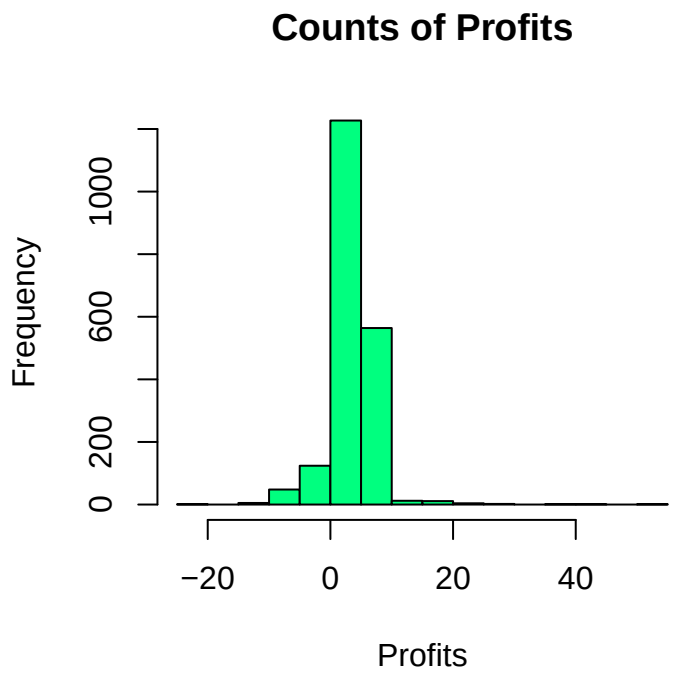
```
        Rank      Company      Country        Sales      Profits
   "integer"     "factor"     "factor"    "numeric"    "numeric"
      Assets Market.Value
   "numeric"    "numeric"
```

4. Look at the summary statistics to see how the variable class affects the output. Which variable classes produce the same summary output? Which produce a different one?

The variables belonging to class **integer** or **numeric** produce the measures of center and dispersion. On the other hand, variables in class **factor** produce the counts for each categories.

5. Draw a histogram for the variable Profit using the default number of bins and color the bins in 'springgreen'. Add a meaningful title to the plot.

```
hist(Forbes2016$Profits, col = "springgreen",
     main = "Counts of Profits", xlab = "Profits")
```



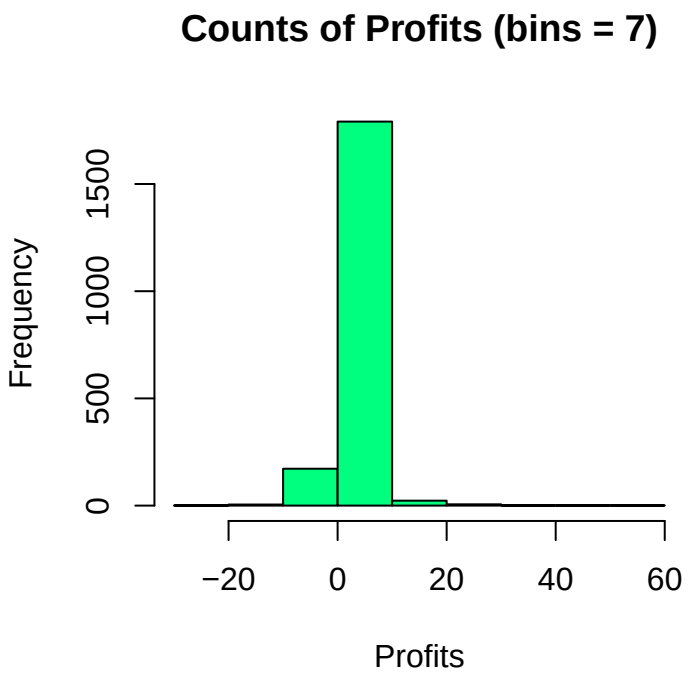
(a) Is the distribution of Profit in the data set symmetric?

As can been from the plot, the distributions of *Profits* is not quitely symmetric. There exist more companies with positive profits.

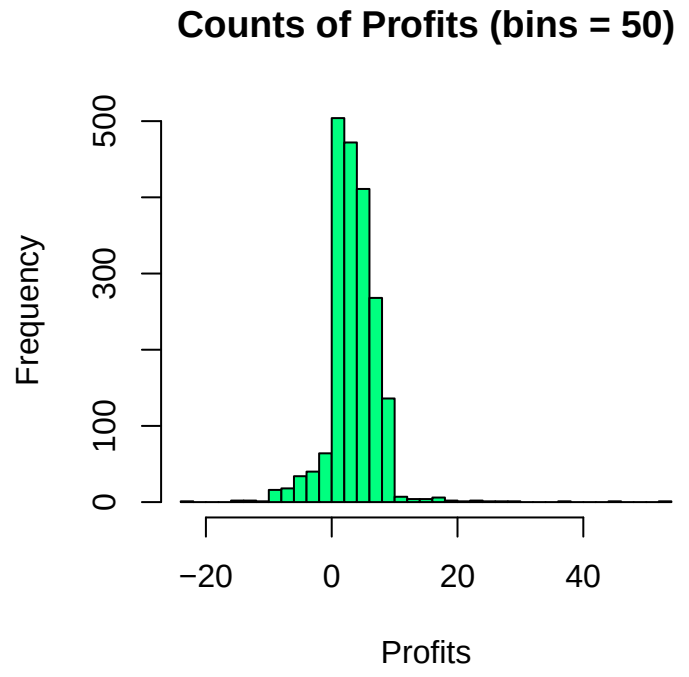
(b) Does the selected number of bins yield a good graphical representation of the data? Try out 7, 50 and 100 bins! Is any of these three options more suited?

In fact, there is no optimal number for bins. But in this case, we would say that 50 is better since the bins are not too small or too large.

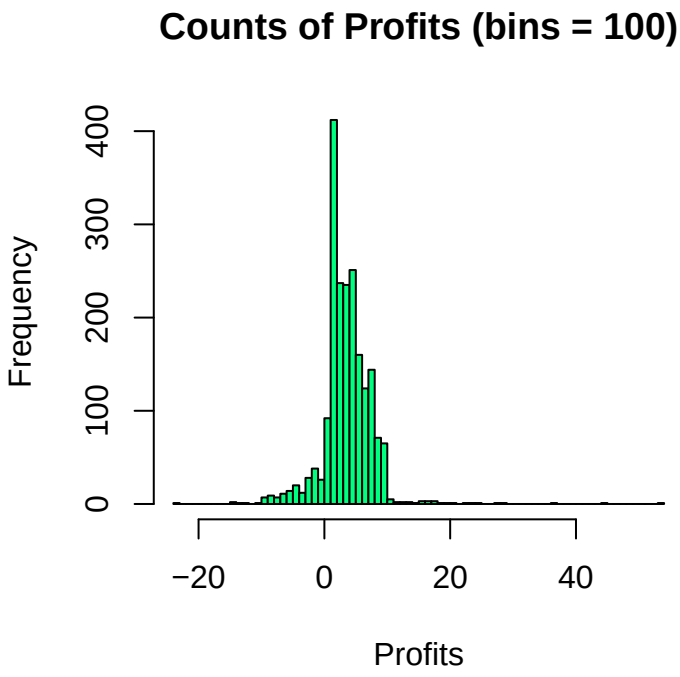
```
hist(Forbes2016$Profits, col = "springgreen", breaks = 7,
     main = "Counts of Profits (bins = 7)", xlab = "Profits")
```




```
hist(Forbes2016$Profits, col = "springgreen", breaks = 50,
     main = "Counts of Profits (bins = 50)", xlab = "Profits")
```

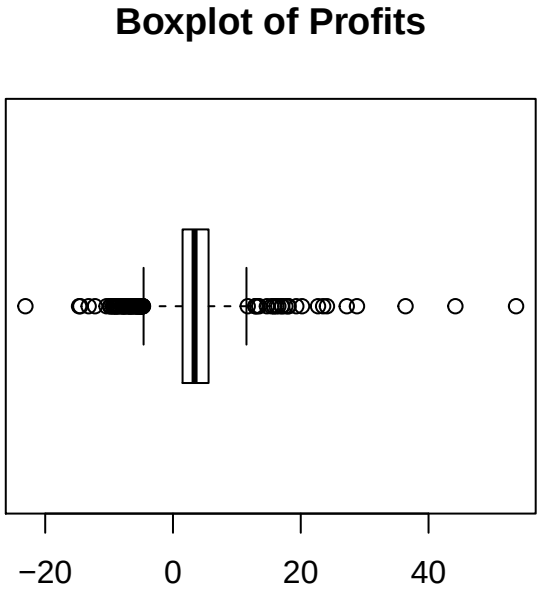


```r
hist(Forbes2016$Profits, col = "springgreen", breaks = 100,
     main = "Counts of Profits (bins = 100)", xlab = "Profits")
```



6. Draw a boxplot for the variable Profit.

```r
boxplot(Forbes2016$Profits, horizontal = TRUE, main = "Boxplot of Profits")
```



(a) Compare the boxplot with the histograms above.

The histograms represent the distribution of *Profits* by frequencies in each bins. Besides, the boxplot depicts *Profits* through its quantiles.

(b) Which graphical representation of the data do you find more helpful?

It depends on what information one attempts to seek. If one is more intereted in the overall distribution of a numerical variable, the histogram is more helpful. Yet a boxplot visually displays the quatiles and can help one to simply detect the outliers of a variable.

(c) How many companies do have profits above 40 billion USD? Can you find out which companies these are?

Two company have rofits over 40 billion USD, and the companies are "ICBC" and "Apple".

```
Forbes2016[Forbes2016$Profits >= 40, "Company"]
```

```
[1] ICBC  Apple
2000 Levels: 3i Group 3M 77 Bank AAC Technologies Holdings ... Zurich Insurance Group
```

7. You have a closer look at profits of the companies.

(a) First, you want to know which profit is needed such that a company belongs to the top two percent.

As the results indicate below, a company with profits above 9.87 billion USD can be in the top two percent.

```
quantile(Forbes2016$Profits, probs = 0.98)
```

```
 98%
9.87
```

(b) Which percentage of companies has negative profits?

8.85% companies has negative profits.

```
sum(Forbes2016$Profits < 0) / nrow(Forbes2016)
```

```
[1] 0.08845577
```

(c) How large is the interquartile range in sales for the companies on Forbes' list?

The interquartile is 13.6 billion USD.

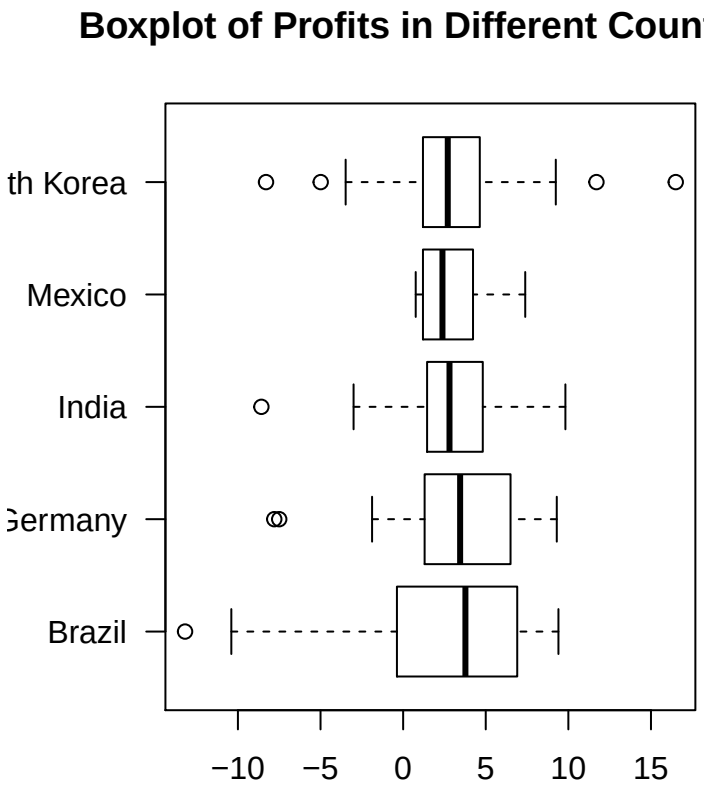
```
IQR(Forbes2016$Sales, na.rm = TRUE)
```

```
[1] 13.6
```

8. Now you start turning towards your real task. Create a subset of the data set comprising all companies that have their legal residence in one of the following countries: Brazil, India, South Korea, France, Germany or Mexico. Draw a boxplot for the variable Profits splitted by country for these companies.

```
sub_forbes2016 <- subset(Forbes2016,
                         Country %in% c("Brazil", "India", "South Korea", "Franc",
                                        "Germany", "Mexico"))
```

```r
boxplot(Profits ~ as.character(Country), data = sub_forbes2016, horizontal = TRUE,
        main = "Boxplot of Profits in Different Countries",
        las = 1)
```



(a) How much difference in medians between countries is shown in the plot?

The differences in the medians are all smaller than 1.5 billion USD. As the plot and following table show, the largest difference is 1.390 (between Mexico and Brazil), and the smallest difference is 0.32 (between Germany and Brazil).

```r
grouped_median <- aggregate(sub_forbes2016$Profits, list(sub_forbes2016$Country), median)

grouped_median <- setNames(grouped_median$x, grouped_median$Group.1)

dist(grouped_median) # differeces of medians in different countries ()
```

```
            Brazil Germany India Mexico
Germany      0.320
India        0.965   0.645
Mexico       1.390   1.070 0.425
South Korea  1.070   0.750 0.105  0.320
```

(b) Which country shows the largest median profit?

The plot indicates that the largest median profit lies in Brazil.

(c) Which country/countries show the most symmetric distributions for profit?

South Korea shows the most symmetric distributions for profit in the plot.

(d) Create a frequency table for the variable Country sorted in decreasing order. Which three countries – among those selected – have the top most companies in Forbes' list?

The top three contries are "South Korea", "India" and "Germany", as the following table lists.

```
sort(table(as.character(sub_forbes2016$Country)), decreasing = TRUE)
```

```
South Korea       India     Germany      Brazil      Mexico
         67          56          50          19          15
```

9. Now you start with your real task for the subset data created in Question 8.

(a) What is the median profit of the selected companies?

The median profit is 2.95 billion USD.

```
median(sub_forbes2016$Profits, na.rm = TRUE)
```

```
[1] 2.95
```

(b) What is the average profit of the selected companies?

The average profit is 3.04 billion USD.

```
mean(sub_forbes2016$Profits, na.rm = TRUE)
```

```
[1] 3.037536
```

(c) What is the average distance for a company's profit to the average profit (i.e. what is the standard deviation of profits)?

The standard deviation of profit is 3.81 billion USD.

```
sd(sub_forbes2016$Profits, na.rm = TRUE)
```

```
[1] 3.81334
```

(d) Does the top third of selected companies have profits that are at least as high as the median profit for all companies on the Forbes list?

The median of profits in Germany (3.45) exceeds the median profit for all companies (2.95), while the medians of profits in South Korea (2.7) and India (2.805) do not.

```
kable(aggregate(sub_forbes2016$Profits, list(sub_forbes2016$Country), median))
```

Group.1	x
Brazil	3.770
Germany	3.450

Group.1	x
India	2.805
Mexico	2.380
South Korea	2.700

10. In order to fulfil your task you finally draw a random sample of size 80 from the list of your selected companies. Create the resulting data frame!

```
set.seed(42) # fix the random seed
sampled_forbes2016 <- sub_forbes2016[sample(1:nrow(sub_forbes2016), 80), ]
```

(a) Compute the mean profit for your sample?

The mean profit for the sample is 2.51 billion USD.

```
mean(sampled_forbes2016$Profits, na.rm = TRUE)
```

```
[1] 2.512125
```

(b) Compute the median profit for your sample?

The median profit for the sample is 2.25 billion USD.

```
median(sampled_forbes2016$Profits, na.rm = TRUE)
```

```
[1] 2.25
```

(c) Compute the minimum profit for your sample?

The minimum profit for the sample is -13.2 billion USD.

```
min(sampled_forbes2016$Profits, na.rm = TRUE)
```

```
[1] -13.2
```

(d) Compute the maximum profit for your sample?

The maximum profit for the sample is 16.5 billion USD.

```
max(sampled_forbes2016$Profits, na.rm = TRUE)
```

```
[1] 16.5
```

(e) Compute the 90%-quantile for your sample?

The 90%-quantile for the sample is 7.33 billion USD.

```
quantile(sampled_forbes2016$Profits, probs = 0.9)
```

```
  90%
7.335
```

11. Are median and mean profits of your sample below or above the median and mean profits for all companies in the selected countries?

The median profit of sampled data (2.25) is smaller than that of all companies (2.95).

The mean profit of sampled data (2.51) is also smaller than that of all companies (3.04).

12. Compute the standard deviation of profits for your sample. Is it smaller or larger than the standard deviation in profits for all selected companies?

The standard deviation of profits of sampled data (4.2) is larger than that of all companies (3.81).

```
sd(sampled_forbes2016$Profits, na.rm = TRUE)
```

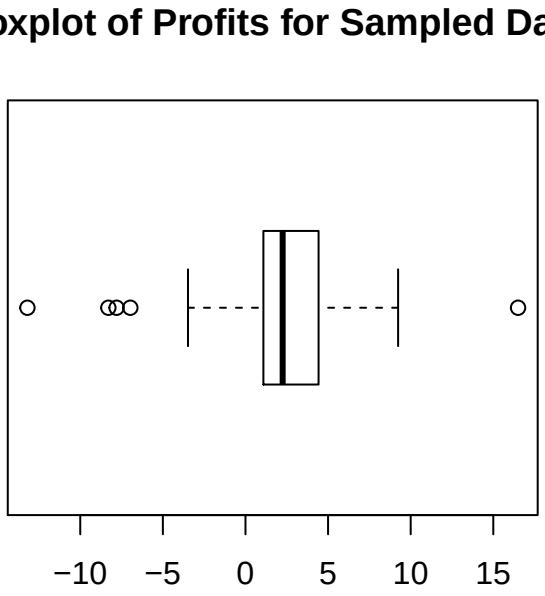
```
[1] 4.201943
```

13. Draw a boxplot, a histogram, and a density plot of the profit for your sample. Can you draw any substantial information from your plot? (e.g about skewness of the distribution, outliers, quartiles)

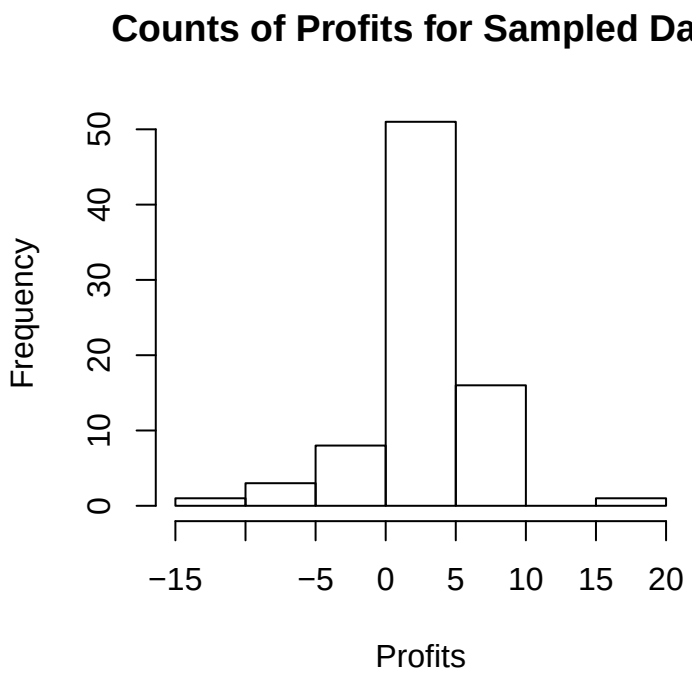
As the plots display, there exists no extreme skewness but slight left skewness. That is, the outliers appear more often in the left side of the distribution. The boxplot also shows that 4 companies has profits less than lower whisker (Q1 - 1.5IQR).

The histogram suggests that most companies have profits in 0 to 10 billion USD, and the interquatile range is around 3 billion USD in the boxplot.

```
boxplot(sampled_forbes2016$Profits, horizontal = TRUE,
        main = "Boxplot of Profits for Sampled Data")
```




```
hist(sampled_forbes2016$Profits,
     main = "Counts of Profits for Sampled Data", xlab = "Profits")
```




```
d <- density(sampled_forbes2016$Profits)
plot(d, main = "Density of Profits for Sampled Data", xlab = "Profits")
```