**Quantitative Methods**

**List of Exercises N. 2**

**Selected Exercises from McClave (2014) – Chapter 2**

**2.4 Numerical Measures of Variability**

**Exercise 1. (66, HCOUGH). *Is honey a cough remedy?* Refer to the Archives of pediatrics and Adolescent Medicine (Dec. 2007) study of honey as a remedy of coughing. 105 ill children in the sample were randomly divided into 3 groups: those who received a dosage of an over-the-counter cough medicine (DM), those who received a dosage of honey (H), and those who received no dosage (control group). The coughing improvement scores (as determined by the children’s parents for the patients are reproduced in the table below:**



First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E1 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E1.xlsx")

View(L2E1)

attach(L2E1)

Create variables: Honey <- Honey

DM <- DM

Control <- Control

TotalScore <- TotalScore

Treatment <- Treatment

Before you begin the calculation of the standard deviations, try to look at your data. If you look, you will see that there is again missing values also called NA’s. You should remove these before you try to calculate the standard deviation. The code is:

na.omit(VARIABLE)

Insert the variables, and create new variables without the missing values:

Honey1 <- na.omit(Honey)

DM1 <- na.omit(DM)

Control1 <- na.omit(Control)

1. **Find the standard deviation of the improvement scores for the honey dosage group.**

The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values.A low standard deviation indicates that the data points tend to be very close to the mean of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

To calculate the standard deviation you need to use the following code:

sd(VARIABLE)

sd(Honey1)

Result: 2.855041

The standard deviation for Honey (Honey1) is equal to 2.855041.

1. **Find the standard deviation of the improvement scores for the DM dosage group.**

sd(DM1)

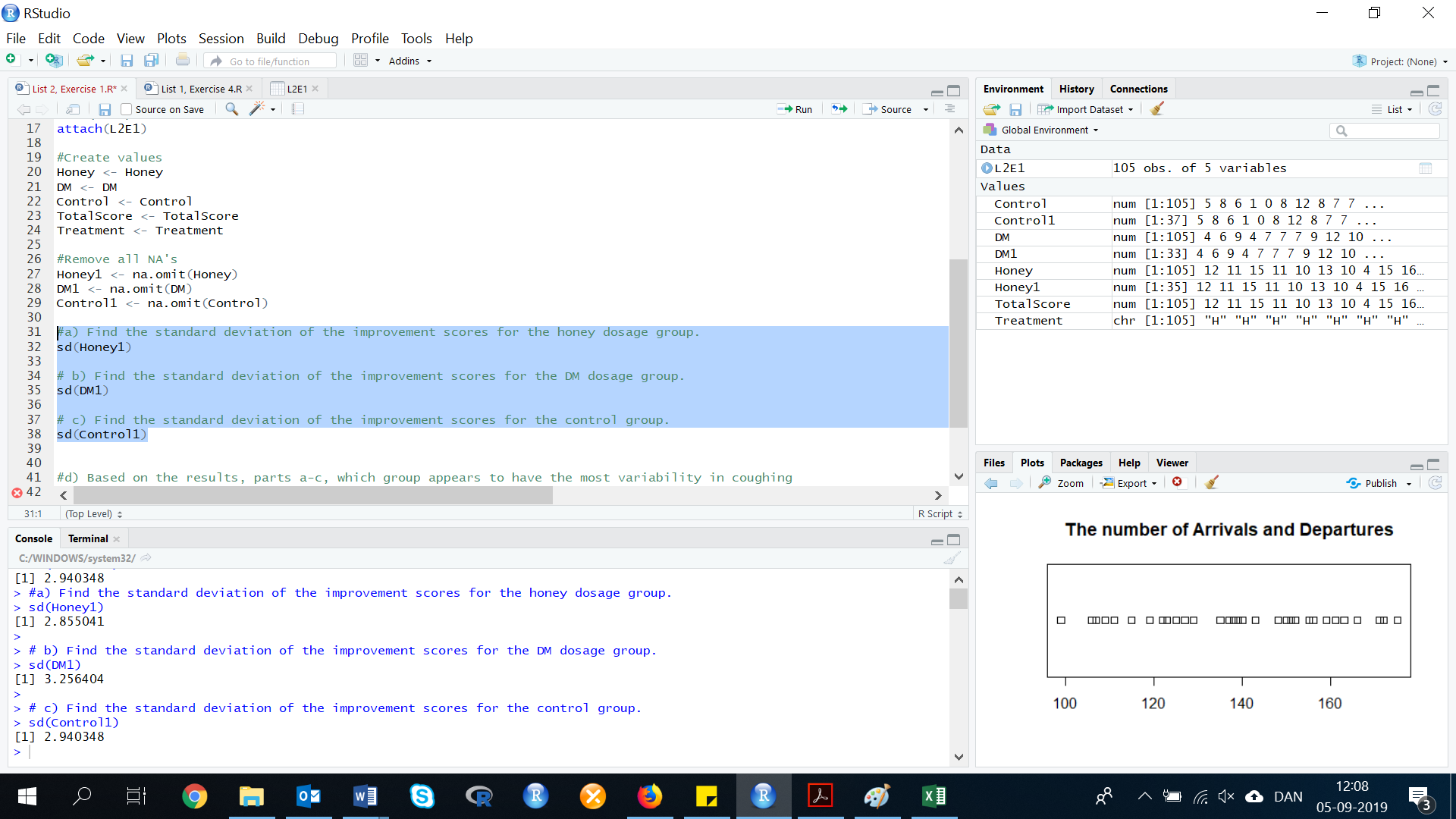
Result: 3.256404

1. **Find the standard deviation of the improvement scores for the control group.**

sd(DM1)

Result: 2.940348

You can see the results directly given in the R Console, below:



1. **Based on the results, parts a-c, which group appears to have the most variability in coughing improvement scores? The least variability?**

The group with the most variability is the one with the highest standard deviation, which is DM = 3.256404.

The group with the least variability is the group with the lowest standard deviation which is Honey = 2.855041

**Exercise 2. (67, CORSUS). *Corporate sustainability of CPA firms*. Refer to the Business and Society (march 2011) study on the sustainability behaviors of CPA corporations. Numerical measures of variation for level of support for the 992 senior managers are showed in the accompanying.**



First, we need to do as followed:

1. Clean your screen
2. Import data
3. Create variables
4. Run Library(mosaic)

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E2 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E2.xlsx")

View(L2E2)

attach(L2E2)

Create values: Support <- Support

1. **Locate the range in the printout. Comment on the accuracy of the statement: “The difference between the largest and the smallest values of the level of support for the 992 senior managers is 155 points”.**

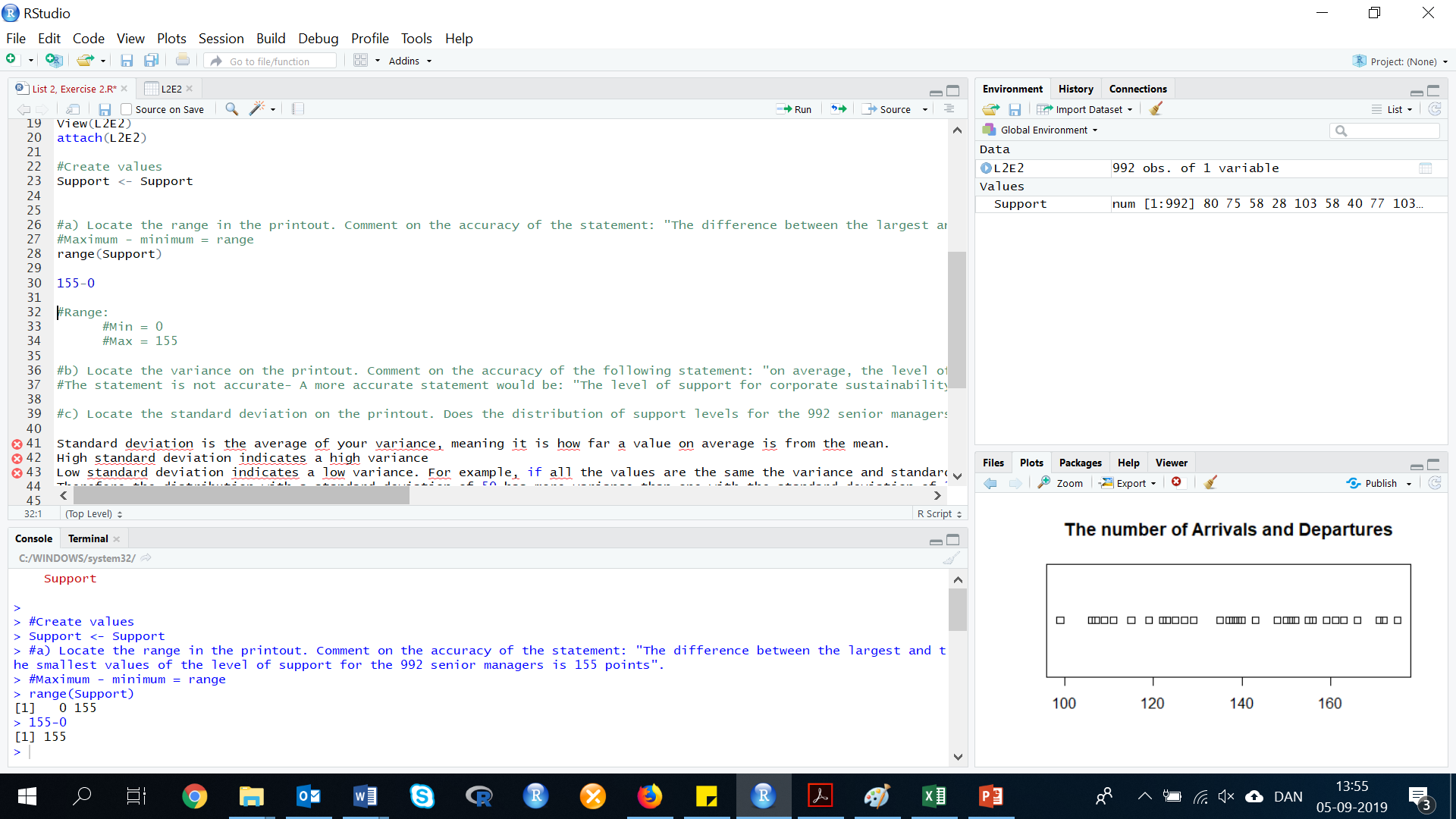
Definition of range: *the difference between the highest and lowest values.*

To calculate the range, you need to use the following code:

range(VARIABLE)

Insert the variable:

range(Support)



The R Console gives two values: 0 and 155. This means that the lowest value is 0 and the highest value is 155.

To calculate the range, you need to: highest value - lowest value = range

155-0= 155 The statement (in our definition above) is correct

Result: Range = 155

1. **Locate the variance on the printout. Comment on the accuracy of the following statement: “on average, the level of support for corporate sustainability for the 992 senior managers is 722 points.”**

The statement is not accurate. A more accurate statement would be: "The level of support for corporate sustainability for the 992 senior managers is 722.036 points".

1. **Locate the standard deviation on the printout. Does the distribution of support levels for the 992 senior managers have a more or less variation than another distribution with a standard deviation of 50? Explain.**

Standard deviation is the average of your variance, meaning it is how far a value on average is from the mean.

* High standard deviation indicates a high variance.
* Low standard deviation indicates a low variance.

For example, if all the values are the same the variance and standard deviation will be 0.

Therefore the distribution with a standard deviation of 50 has more variance than one with the standard deviation of 26.8. The data with a standard deviation of 50 has data that is more widely dispersed around the mean compared with the data with a standard deviation of 26.8

1. **Which measure of variation best describes the distribution of 992 support levels? Explain.**

Range does not tell us very much about the distribution of our data only that the data is within the range 0 to 155. For example, we could have data where only one value is 155 and the rest are: 0s, 1s and 2s

Variance tells us about the total distribution of the data, but standard deviation is even better because it is the average of your variance, meaning it is how far a value on average is from the mean.

The standard deviation is best to describe the distribution of the 992 support levels as it also takes into account the number of observations that contribute to the total variance.

**Exercise 3. (69, NUKES). *Active Nuclear power plants*. Refer to the US Energy Information Administration’s data on the number of nuclear power plants operating in each of the 20 states.**

First, we need to do as followed:

1. Clean your screen
2. Import data
3. Create variables
4. Run Library(mosaic)

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E3 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E3.xlsx")

View(L2E3)

attach(L2E3)

Create values: STATE <- STATE

PLANTS <- PLANTS

1. **Find the range, variance and standard deviation of this data set.**

You already know how to calculate the range and the standard deviation. You can see the explanation and codes in List 2, Exercise 1 and 2.

range(PLANTS)

The lowest value is = 1, and the highest value is = 11. This means that the range is = 10, as we minus the highest value with the lowest (11-1).

Result: Range = 10

sd(PLANTS)

Result: Standard deviation = 2.770142

To calculate the variance in R, you need the following code:

var(VARIABLE)

Insert the variable:

var(PLANTS)

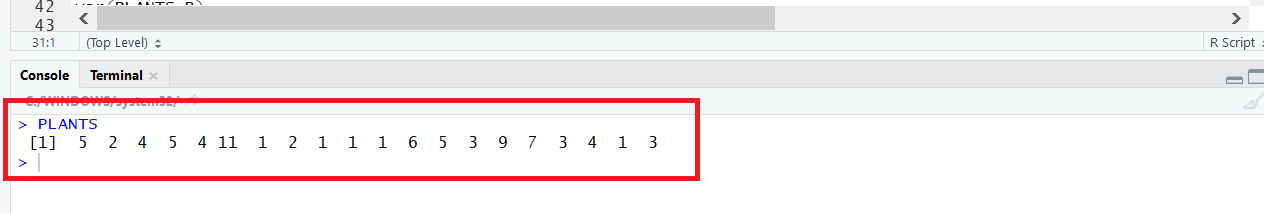
Result: Variance = 7.673684

1. **Eliminate the largest value from the data set and repeat part a. What effect does dropping this measurement have on the measures of variation found in part a?**

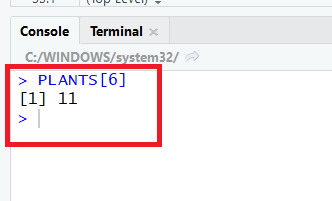
To remove the largest value from the data set, you need to select the values. However, there is many different ways to do this in R. One of the ways is as following:

First, you need to look at the variable. If you type in the name of the variable, and press run it will show you all of the data points:

PLANTS



By looking at the data, we can see that the highest value is number 11. If we count, we can see 11 is number 6 value in the row. To make sure, that we did count correctly, we write the following code:

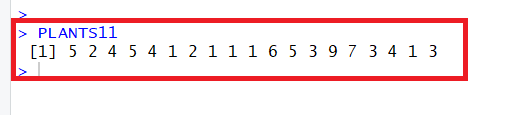
PLANTS[6]

This means, that we did count correctly. The highest value (11) is placed as number 6.

Those brackets [] illustrate another strong point of R. They represent a function that you can use to extract a value from that vector. You can get the fifth value of the preceding number vector like this:

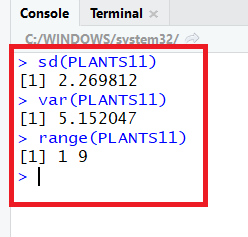
PLANTS[-6]

PLANTS11 <- PLANTS[-6]

Now PLANTS11 is a new variable without the highest value (11). Run the following code, if you want to check, that it is removed:

PLANTS11

Now we just have to calculate sd(), var(), and range(), just as in List 2, exercise 3a.

sd(PLANTS11)

var(PLANTS11)

range(PLANTS11)

range = 9-1 = 8.

Results: Standard deviation = 2.269812

Variance = 5.152047

Range = 8

1. **Eliminate the smallest and largest value from the data set and repeat part a. What effect does dropping both of these measurements have on the measures of variation found in part a?**

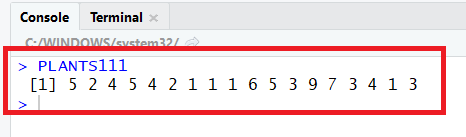
In this exercise, we have to remove both the lowest and highest value from the ‘PLANTS’ Variable. We will in this exercise reuse, the variable ‘PLANTS11’, which is already missing the highest value. Then we only need to remove the lowest.

Similar to exercise 3b, we look at the variable:

PLANTS11

We can see that the lowest value is 1, by counting we know that this value is present many times. However, we do only remove one of them. We chose to remove value number 6, which is equal to 1.

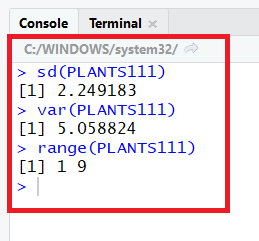
PLANTS11[-6]

We create a new variable:

PLANTS111 <-PLANTS11[-6]

If we write: PLANTS111

The R console gives the new variable.

Now we can calculate the range, variance, and standard deviation for the new variable ‘PLANTS111’.

sd(PLANTS111)

var(PLANTS111)

range(PLANTS111)

Result: Standard deviation = 2.249183

Variance = 5.058824

Range = 9-1 = 8

**2.5 Using the Mean and standard deviation to describe Data**

**Exercise 4. (75). Given a data set with a largest value of 760 and a smallest value of 135, what would you estimate the standard deviation to be? Explain the logic behind the procedure you used to estimate the standard deviation. Suppose the standard deviation is reported to be 25. Is this feasible? Explain.**

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Using Chebyshev's Rule, at least 8/9 of the measurements ill fall within 3 standard deviations of the mean. Thus, the range of data would be around 6 standard deviations. Using the empirical Rule approximately 95% of the observations are within 2 standard deviations of the mean. Thus, the range of the data would be around four standard deviations. We would expect the standard deviation to be somewhere between Range 6 and Range 4.

For our data, the range = 760 – 135 = 625.

Calculation of Range/6:

625/6

Result: 104.1667

Calculation of Range/4

625/4

Result: 156.25

Therefore, I would estimate that the standard deviation of the data set is between 104.17 and 156.25. It would not be feasible to have a standard deviation of 25. If the standard deviation were 25, the data would span: 625/25 = 25 standard deviations. This would be extremely unlikely.

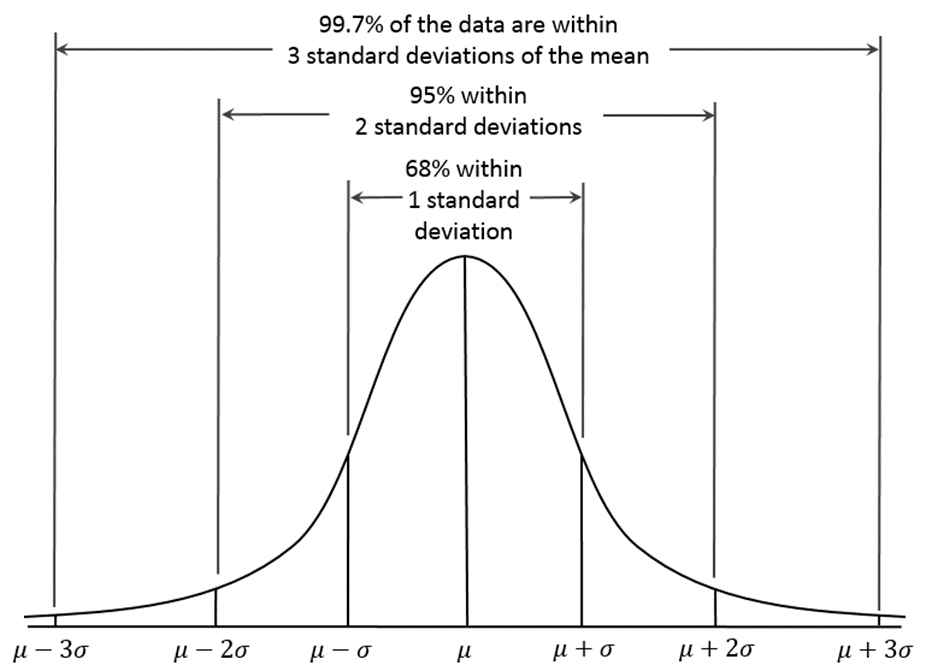
**Exercise 5. (78). *Blogs for Fortune 500 firms*. Refer to the Journal of Relationship Marketing (Vol. 7, 2008) study of the prevalence of blogs and forums at fortune 500 firms with both English and Chinese Web sites. In a sample of firms that provide blogs and forums as marketing tools, the mean number of blogs/forums per site was 4.25, with a standard deviation of 12.02.**

First, we need to do as followed:

1. Clean your screen

For explanation how to do clean your screen look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

1. **Provide an interval that is likely to contain the number of blogs/forums per site for at least 75% of the Fortune 500 firms in the sample.**

This is a bell curve which represents normally distributed data. We know that 68% of data is within 1 standard deviation, 95% of data is within 2 standard deviations and 99.7% of data is within 3 standard deviations for normal data.

For non-normally distributed data, there is another rule that you can use: Chebyshev’s inequality,

Where a minimum of just 75% of values must lie within two standard deviations of the mean and 89% within three standard deviations.

We have been provided with the mean (4.25) and standard deviation (12.02). From this information, we need to create an interval that contains minimum 75% of the data. Using our knowledge from above, we can calculate a range that includes 95% of the data. For both normal and non-normal data 75% will lay within 2 standard deviations.

4.25 + 12.02 + 12.02 = 28.29

or

4.25 - 12.02 - 12.02 = -19.79

Expect we can’t have a negative number of blogs….

Therefore, the interval is 0 to 28.8

Result: 28.8

1. **Do you expect the distribution of the number of blogs/forums to be symmetric, skewed right or skewed left? Explain.**

We expect the distribution to be skewed to the right, because we know we can’t have a negative number of blogs/forums. We can also assume that there are a few very large observations, because the standard deviation is so large compared to the mean. If the distribution is skewed to the right, it suggests that the data is not normally distributed, so we should use only Chebyshev's inequality for exercise a.

**2.6 Numerical Measures of Relative Standing**

**Exercise 6. (100, CORSUS). *Corporate sustainability of CPA firms*. Refer to the Business and Society (march 2011) study on the sustainability behaviors of CPA corporations. Numerical descriptive measures for the level of support for corporate sustainability for the 992 senior managers are repeated in the accompanying printout. One of the managers reported a support level of 155 points. Would you consider this support level to be typical of the study sample? Explain.**



First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E6 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E6.xlsx")

View(L2E6)

attach(L2E6)

Create Variables: Support <- Support

A Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores. A Z-score of 0 means the score is the same as the mean. A Z-score can also be positive or negative, indicating whether it is above or below the mean and by how many standard deviations.

To find the Z-Score you just simply use the:

* x (155)
* mean (67.755)
* Standard deviation (26.871)

In R we can calculate it, by:

First, calculating the mean and the standard deviation:

mean(VARIABLE)

sd(VARIABLE)

Insert the variable:

mean(Support)

sd(Support)

Create a variable for the mean value, so you can use the mean of Support to calculate the z score:

Mean\_Sup <- mean(Support)

Sd\_Sup <- sd(Support)

Now you can calculate the Z score:

z <- (155 - Mean\_Sup)/Sd\_Sup

Result: 3.246841

This solution is 3.246841 standard deviations from the mean, which in the extreme ends of a normal distribution.

**Exercise 7. (101, SPRFND). *Hazardous waste cleanup in Arkansas*. The Superfund Act was passed by congress to encourage state participation in the implementation of a law relating to the release and cleanup of hazardous substances. Hazardous waste sites financed by the Superfund Act are called Superfund sites. A total of 393 Superfund sites are operated by waste management companies in Arkansas (Tabor and Stanwick, Arkansas Business and Economic Review, Summer 1995). The number of these Superfund sites in each of Arkansas’s 75 counties are shown in the next table.**



First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

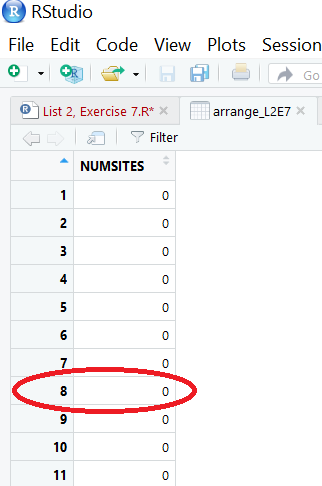
Library: library(mosaic)

Import data: library(readxl)

L2E7 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E7.xlsx")

View(L2E7)

attach(L2E7)

Create values: NUMSITES <- NUMSITES

1. **Find the 10th percentile of the data set. Interpret the result.**

The 10th percentile is the score that has at least 10% of the observations less than it. If we arrange the data in order form the smallest to the largest, the 10th percentile score will be the .10(75) = 7.5 or 8th observation. When the data are arrange in order, we need to look at the 8th observation.

Arrange the data:

arrange(VARIABLE)

arrange(L2E7)

arranged\_L2E7 <- arrange(L2E7)

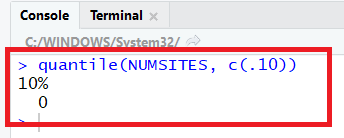
If we now look at over data, we will see that the 8th observation is: 0

This means, that according to our statement the 10th percentile is equal to 0. However, in R we can also calculate this:

quantile(DATA$VARIABLE, c(.quantile))

The quantile() function display the quartiles, but it can also be given a vector of quantiles to display.

quantile(arrange\_L2E7$NUMSITES, c(.10))



The R Console gives the result. The 10th quantile of the variable Numsites is equal to 0. This means at least 10% of the observations are less than 0, and we can confirm our result and statement above.

1. **Find the 95th percentile of the data set. Interpret the result.**

Now we will use the same method as before:

The 95th percentile is the score, that has at least 95% of the observations, less than it. If we arrange the data in order from smallest to largest, the 95th percentile score will be: .95(75) = 71.25 or the 72nd observation, which mean between 17 and 21.

Again, we therefore assume that the result will be 21. However, we can still calculate it in R to check if we are right.

quantile(arrange\_L2E7$NUMSITES, c(.95))

Result: 18.25

Our R Console gives a different results, than we assumed, but its because R is precise. The 71st observation = 17, and the 72nd observation = 21.

Our quick calculation above showed, that we are looking for the 71.25 observation, and this observation, according to R, is equal to 18.25. Therefore, the precise answer is, that the 95th quantile = 18.25. However, is it not wrong to write 21.

1. **Find the mean and standard deviation of the data. Then, use these values to calculate z-score for an Arkansas county with 48 Superfund sites.**

First, you need to calculate the mean and standard deviation:

mean()

mean(NUMSITES)

mean\_num <- mean(NUMSITES)

The mean is equal to 5.24.

sd()

sd(NUMSITES)

sd\_num <- sd(NUMSITES)

The standard deviation is equal to 7.244457.

For definition of z-scores, look at List 2 Exercise 6.

In R:

z <- (48- mean\_num)/sd\_num

Result: 5.902443

1. **Based on your answer c, would you classify 48 as an extreme number of Superfund sites?**

Yes. The score 48 is almost 6 standard deviations from the mean. We know that for any set of data almost all of the observations are within 3 standard deviations of the mean (Look at List 2 exercise 5a for explanation of this).

Almost 6 standard deviations from the mean is very unusual!

**Exercise 8. (103, ECOPHD). *Ranking PhD programs in economics*. Thousands of students apply for admission to graduate schools in economics each year with the intention of obtaining a PhD. The Southern Economic Journal (Apr. 2008) published a guide to graduate study in economics by ranking the PhD programs at 129 colleges and universities. Each program was evaluated according to the number of publications published by faculty teaching in the PhD program and by the quality of the publications. Data obtained from the Social Science Citation Index (SSCI) were used to calculate an overall productivity score for each PhD program. The mean and standard deviation of these 129 productivity scores were then used to compute a z-score for each economics program. Harvard University had the highest z-score (5.08) and, hence, was the top ranked school. Howard University was ranked last because it had the lowest z-score (-0.81). The data (z-scores) for all 129 economic programs are saved in the data file.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E8 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E8.xlsx")

View(L2E8)

attach(L2E8)

Create variables: School <- School

Rank <- Rank

ZScore <- ZScore

TopField<- TopField

1. **Interpret the z-score for Harvard University**

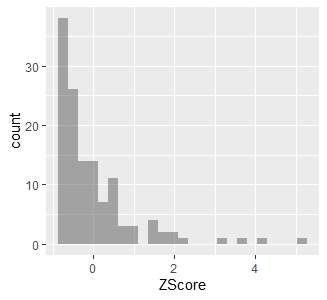
The z-score for Harvard is z = 5.08.

This means that Harvard’s productivity scores was 5.08 standard deviations above the mean. This is an extremely high and unusual z-score.

1. **Interpret the z-score for Howard University**

The z-score for Howard University is z=-.85.

This means that Howard University’s productivity was .85 below the mean. This is not an unusual z-score.

1. **The authors of the Scandinavian Economic Journal article note that “only 44 of the 129 school have positive z-scores, indicating that the distribution of overall productivity is skewed to the right.” Do you agree? Check your answer by constructing a histogram for the z-scores in the file.**

First, you need to make a histogram:

gf\_histogram( ~ VARIABLE, data = DATA)

gf\_histogram(~ZScore, data=L2E8)

Yes. Other indicators that the distribution is skewed to the right are the values of the highest and lowest z-scores. We could use the R code: Range(), to see the lowest and highest z-scores of for the universities.

range()

range(ZScore)

Result: Min (-0.81), Max (5.08)

So the lowest z-score is less than 1 standard deviation below the mean, while the highest score is 5.08 standard deviations above the mean. Therefore, the distribution will be skewed to the right.

**2.7 Methods for Detecting Outliers: Box Plots and z-Scores**

**Exercise 9. (115, BNKRPT). *Time in Bankruptcy*. Refer to the Financial Management (Spring 1995) study of 49 firms filing for prepackaged bankruptcies. 3 types of firms exist: (1) Those who hold no prefiling vote, (2) those who vote their preference for a joint solution and (3) Those who vote their preference for a prepack.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E9 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E9.xlsx")

View(L2E9)

attach(L2E9)

Create variables: TIME <- TIME

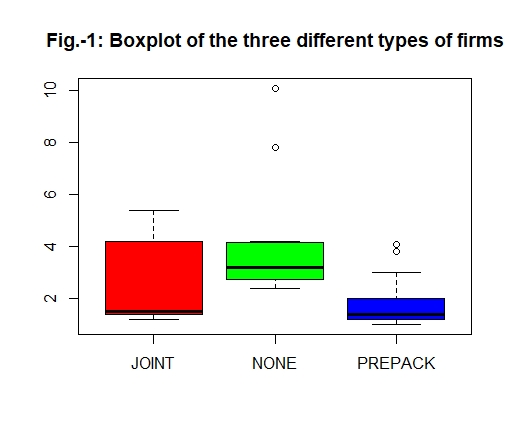
PLAN <- PLAN

1. **Construct a box plot for the time in bankruptcy (months) for each type of firm.**

There are two ways to solve this exercise. We can make a figure with all the boxplots, and we can make a boxplot for each of the types of firms.

Method 1: ALL BOXPLOTS IN ONE FIGURE

boxplot(TIME~PLAN, main="Fig.-1: Boxplot of the three different types of Plans", col= rainbow(3))



Method 2: A BOXPLOT FOR EACH PLAN

First, we need to filter our data, so we can work with each of the different firms:

dplyr::filter(DATA, Variable ==”VARIABLE”)

dplyr::filter(L2E9, PLAN ==”Joint”)

dplyr::filter(L2E9, PLAN ==”None”)

dplyr::filter(L2E9, PLAN ==”Prepack”)

Make these filters into new dataframes:

JOINT <- dplyr::filter(L2E9, PLAN ==”JOINT”)

NONE <- dplyr::filter(L2E9, PLAN ==”NONE”)

PREPACK <- dplyr::filter(L2E9, PLAN ==”PREPACK”)

Now we need to make a boxplot. A boxplot is a particular helpful graphical display to compare distributions. There are many different ways to make a boxplot. However, we suggest you to use this R code:

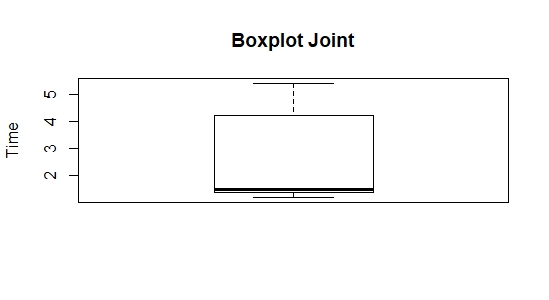
boxplot(DATA$VARIABLE)

If you want the boxplot to have a main title and title on the y-axis, then you can use the following codes:

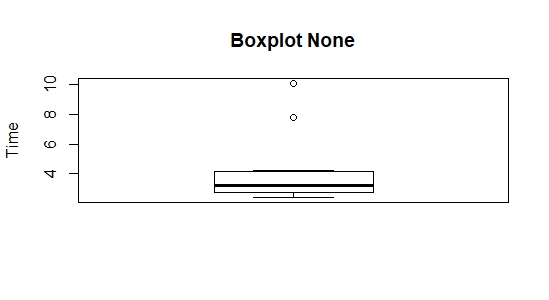
boxplot(DATA$VARIABLE, ylab='name', main='name')

Insert the data:

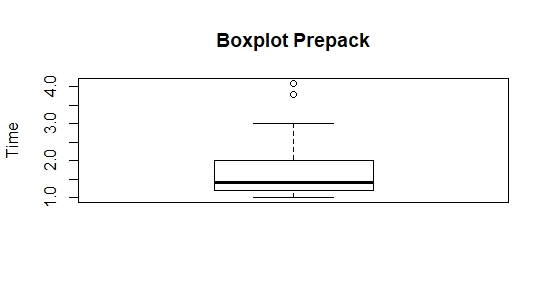
boxplot(JOINT$TIME, ylab='Time', main='Boxplot Joint')



boxplot(NONE$TIME, ylab='Time', main='Boxplot None')



boxplot(PREPACK$TIME, ylab='Time', main='Boxplot Prepack')



1. **Find the median bankruptcy times for the 3 types of firms.**

In this exercise, we need to calculate the median for each of the firms. To calculate the median, you need to use the following code:

median(DATA$VARIABLE)

median(JOINT$TIME)

median(NONE$TIME)

median(PREPACK$TIME)

Results: Join = 1.5

None = 3.2

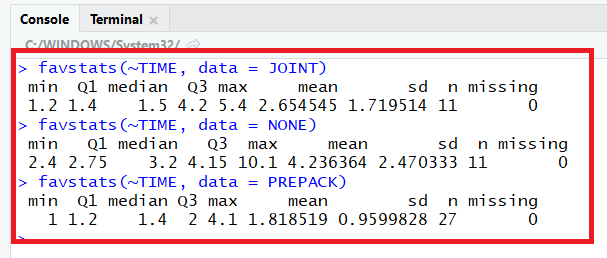
Prepack = 1.4

1. **How do the variabilities of the bankruptcy times compare for these 3 types of firms?**

In this exercise, we need to calculate the highest and lowest values, the first and third quantile (.25 and .75). To help us explain the variability of times of bankruptcy in the three types of firms. To help us calculate this, we use the following code:

favstats(~VARIABLE, data = DATA)

The favstats() function in this mosaic package, provides with a concise summary of some useful statistics.

favstats(~TIME, data = JOINT)

favstats(~TIME, data = NONE)

favstats(~TIME, data = PREPACK)

The only calculations, that we are missing is the range (max-min) and the Interquartile range (Q3-Q1):

Range:

Range\_Joint <- 5.4-1.2

Range\_None <- 10.1-2.4

Range\_Prepack <- 4.1-1

Range: Joint = 4.2, None = 7.7 and Prepack = 3.1

Interquartile range:

InterQ\_Joint <- 4.2-1.4

InterQ\_None <- 4.15-2.75

InterQ\_Prepack <- 2-1.2

Interquartile: Joint = 2.8, None = 1.4 and Prepack = 0.8

Using the boxplots and the data we calculated to create we can say that:

* The range of the "Prepack" (3.1) firms is less than the other two, while the range of the "None" (7.7) firms is the largest.
* The interquartile range of the "Prepack" (0.8) firms is less than the other two, while the interquartile range of the "Joint" (2.8) firms is larger than the other two.

1. **The standard deviation of the bankruptcy times are 2.47 for “none”, 1.72 for “joint” and 0.96 for “prepack” Do the standard deviations agree with the interquartile ranges with regard to the comparison of the variabilities of the bankruptcy times?**

First, check if the numbers is correct:

sd(JOINT$TIME)

sd(NONE$TIME)

sd(PREPACK$TIME)

Standard deviation: Joint = 1.719514

None = 2.470333

Prepack = 0.9599828

The numbers are correct. Now we can answer the question.

No. The interquartile range for the "Prepack" firms is the smallest, which corresponds to the smallest standard deviation. However, the second smallest interquartile range corresponds to the "None" firms. The second smallest standard deviation corresponds to the "Joint" firms.

Since the standard deviation is a measure of average spread of data, it makes most sense that the firms with the smallest interquartile range is also the firm with the lowest standard deviation. Similarly the firm with the highest interquartile range should be the firm with the highest standard deviation.

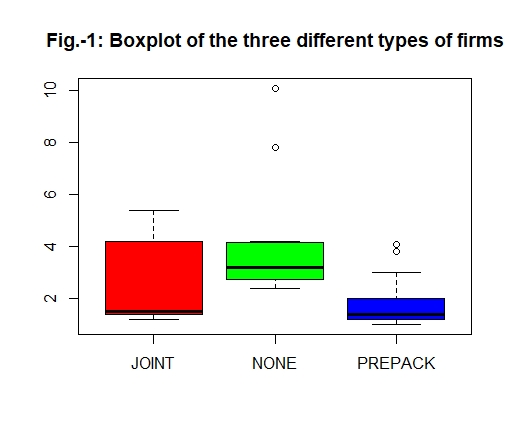
1. **Is there evidence of outliers in any of the 3 distributions?**

As in part a, there is two ways to solve this.

Method 1: ALL BOXPLOTS IN ONE FIGURE

First recall, how the boxplot looks:

We can see there is some outliers in ‘None’ and ‘Prepack’.

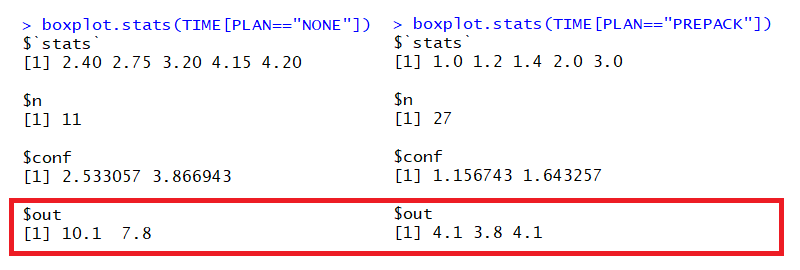


We can detect exactly the outliers by using the following codes:

boxplot.stats(TIME[PLAN=="NONE"])

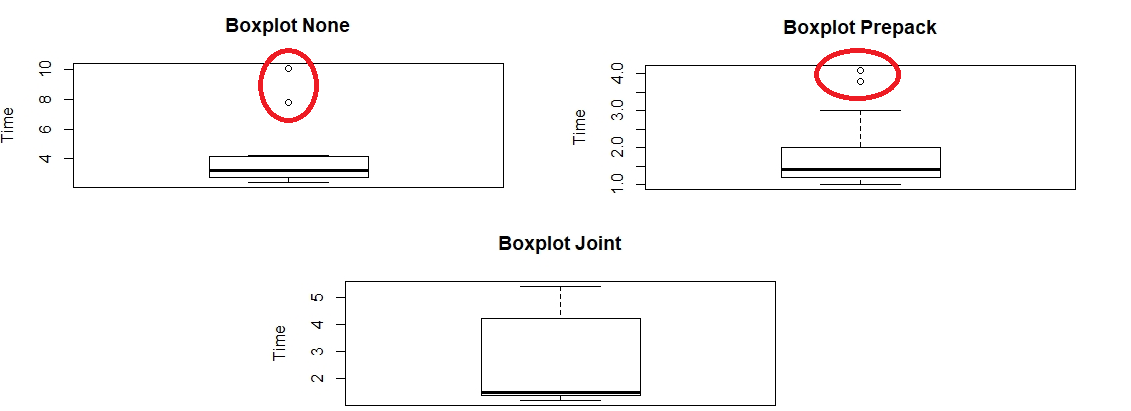
boxplot.stats(TIME[PLAN=="PREPACK"])

The R console will then give us:



By looking at the results, that the R console gave us, we can see that ‘NONE’ has two outliers (10.1 and 7.8), and ‘PREPACK’ has three outliers (4.1, 3.8, and 4.1).

Method 2: ONE BOXPLOT FOR EACH TYPE OF FIRM

If we look at the boxplots from an earlier exercise, we will see, that there is some outliers in two of the three distributions:

These boxplots shows us, just as in Method 1, that Prepack and None has outliers. We can see that the outliers for ‘Prepack’ is 3.8, and 4.1. However, we can not see that there is two outliers on the point 4.1. The outliers for ‘None’ is 7.8 and 10.1. However, you could also look at the data, and see if some of the values are quite low or high in comparison of all the other values for each of the three types of firms.

**Exercise 10. (117, SANIT). *Sanitation Inspection of Cruise Ships*. Refer to the sanitation levels of passenger cruise ships.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the Script: rm(list=ls())

Library: library(mosaic)

Import Data from Excel: library(readxl)

L2E10 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E10.xlsx")

View(L2E10)

attach(L2E10)

Create Variables: Ship\_Name <- Ship\_Name

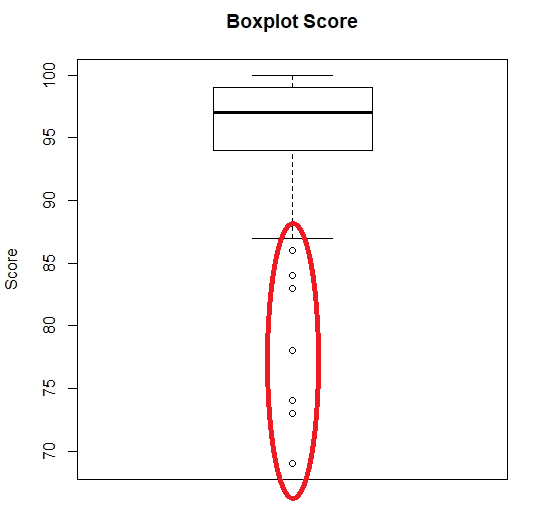
Cruise\_Line <- Cruise\_Line

Score <- Score

1. **Use the box plot method to detect any outliers in the data set.**

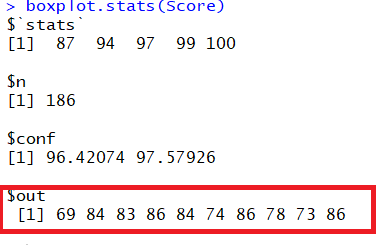
In this exercise, we will use the same codes as in Exercise 9, to calculate the boxplot. For further explanation of what a Boxplot is, look at exercise 9. As score is the only variable will numerical values, it is only possible to detect the outliers in this variable.   
Insert the data and variables:

boxplot(L2E10$Score, ylab='Score', main='Boxplot Score')



If we look at the boxplot, we can see that there is at least 7 outliers, which has a value below 90. By running the boxplot.stats() function we can detect the outliers:

boxplot.stats(Score)



By looking at this, we can see that there are 10 outliers (69, 84, 83, 86, 84, 74, 86, 78, 73, and 86) in ‘Score’.

1. **Use the z-score method to detect any outliers in the data set.**

Look at List 2, Exercise 6, for further information about, how to calculate the z-score. However, we need to calculate the mean and the standard deviation:

mean(VARIABLE)

sd(VARIABLE)

Insert the variable and create the new variable (so we can work with this mean and standard deviation value):

mean\_Score <- mean(Score)

sd\_Score <- sd(Score)

Now you can calculate the Z score for each outlier:

Insert the data:

Outlier\_69 <- (69-mean\_Score)/sd\_Score

Outlier\_73 <- (73-mean\_Score)/sd\_Score

Outlier\_74 <- (69-mean\_Score)/sd\_Score

Outlier\_78 <- (78-mean\_Score)/sd\_Score

Outlier\_83 <- (83-mean\_Score)/sd\_Score

Outlier\_84 <- (84-mean\_Score)/sd\_Score

Outlier\_86 <- (86-mean\_Score)/sd\_Score

Z-scores for the outliers: Outlier\_69 = -5.379395

Outlier\_73 = -4.573461

Outlier\_74 = -5.379395

Outlier\_78 = -3.566043

Outlier\_83 = -2.558625

Outlier\_84 = -2.357141

Outlier\_86 = -1.954174

Based on the empirical rule, any z-score greater than 3 or less than – 3 is considered to be an outlier. This is the rule of thumb on the empirical rule. This means that it is only the following values, that are considered as outliers: 69, 73, 74, and 78 as they are less than -3.

1. **Do the 2 methods agree? Explain why.**

Using the boxplot, we can observe 10 outliers (69, 73, 74, 78, 83, 84, 84, 86, 86, 86). Using the z-score method, only 4 observations (69, 73, 74, and 78) were identified as outliers. However, the 6 additional points (83, 84, 84, 86, 86, 86) that were not identified as outliers using the z-score method were very close to the cutoff value.

**2.8 Graphing Bivariate Relationships**

**Exercise 11. (129, PARS). *Performance ratings of government agencies*. The U.S. Office of Management and Budget (OMB) requires government agencies to produce annual performance and accounting reports (PARS) each year. A research team at George Mason University evaluated the quality of the PARS for 24 government agencies (The Public Manager, Summer 2008). Evaluation scores ranged from 12 (lowest) to 60 (highest). The PARS evaluation scores for 2 consecutive years are shown in the table below:**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E11 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E11.xlsx")

View(L2E11)

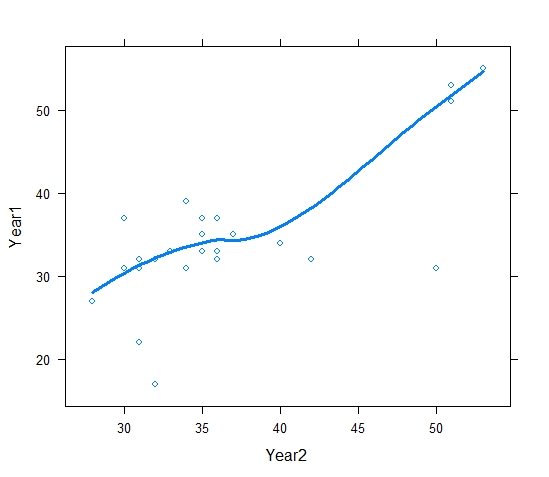
attach(L2E11)

Create variables: Agency <- Agency

Year1 <- Year1

Year2 <- Year2

1. **Construct a scatterplot for the data. DO you detect a trend in the data?**

Look at List 1, Exercise 3 for explanation of a scatterplot. However, the code for this scatterplot is different, in comparison to the one used in List 1, Exercise 3. In this exercise, we have to detect if there is a trend between Year 1 and Year 2. Therefore, the code is as following:

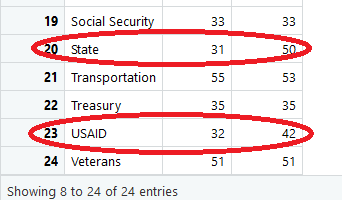
xyplot(X ~ Y, data=DATA)

Insert data:

xyplot(Year1 ~ Year2, data=L2E11)

We can even get R to make us a line, that can help us assess linearity of the relationship between Year 1 and Year 2. This is added by specifying both points (using ‘p’) and a lowess smoother. The lowess smoother is a smooth line through a timeplot or scatter plot, that can help us see the relationship between variables and foresee trends.

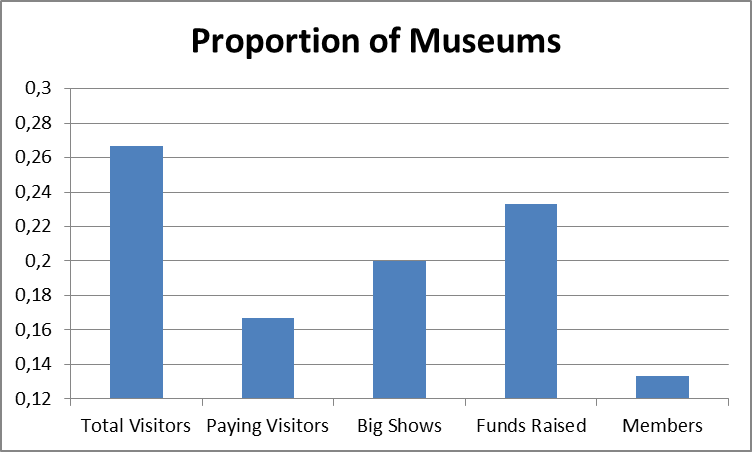
xyplot(Year1 ~ Year2, type=c("p","smooth"), pch=1, Cex=0.6, lwd=3, data=L2E11)

1. **Based on the graph, identify one or two agencies that had greater than expected PARS evaluation scores for year 2.**

From the scatteplot and lowess line, we can see that two agencies that had greater than expected PARS evaluation scores for Year 2 were USAID and State.

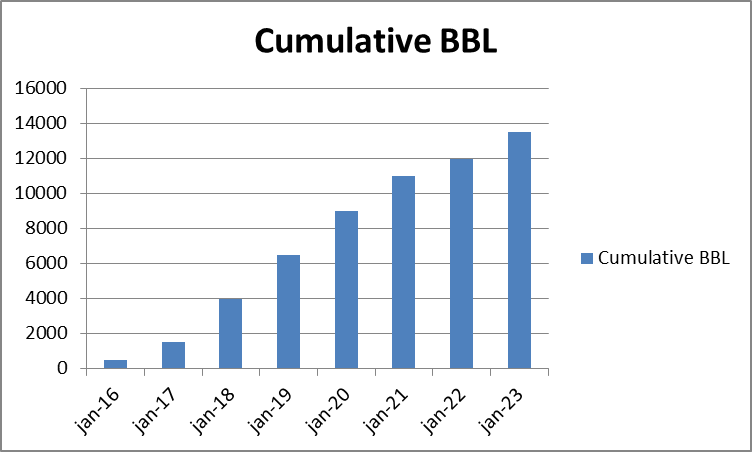
**2.9 Distorting the Truth with Descriptive Techniques**

**Exercise 12. (133, MMC). *Museum Management*. Refer to the Museum Management and Curatorship (June 2010) study of how museums evaluate their performance. Managers of 30 museums of contemporary art identified the performance measure used most often. A summary of the results are reproduced in the table. Consider the bar graph shown. Identify two ways in which the bar graph might mislead the viewer by overemphasizing the importance of one of the performance measures.**



One way the bar graph can mislead the viewer is that the vertical axis has been cut off. Instead of starting at 0, the vertical axis starts at 12. Another way the bar graph can mislead the viewer is that as the bars get taller, the widths of the bars also increase.

**Exercise 13. (136, BPOIL) *BP oil leak*. In the summer of 2010, an explosion on the Deepwater Horizon oil drilling rig caused a leak in one of British Petroleum (BP) Oil Company’s wells in the Gulf of Mexico. Crude oil rushed unabated for 3 straight months into the Gulf until BP could fix the leak. During the disaster, BP used suction tubes to capture some of the gushing oil. In May of 2011, in an effort to demonstrate the daily improvement in the process, a BP representative presented a graphic on the daily number of 42-gallon barrels (bbl) of oil collected by the suctioning process. A graphic similar to the one used by BP is shown below:**



First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E13 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E13.xlsx")

View(L2E13)

attach(L2E13)

Create variables: DAY <- DAY

BBL <- BBL

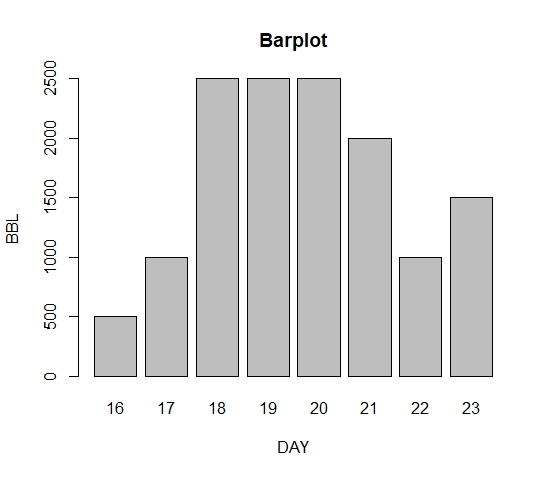
1. **Note that the vertical axis represents the cumulative number of barrels collected per day. This is calculated by adding the amounts of the previous days’ oil collection to the current days’ oil collection. Explain why this graph is misleading.**

This graph is misleading because it looks like as the days are increasing, the number of barrels collected by day are also increasing. However, the bars are the cumulative number of barrels collected. So BP is trying to make us think that they collect more and more barrels of oil each day when they actually don't.

1. **Estimates of the actual number of barrels of oil collected per day for each of the 8 days are listed in the table below. Construct a graph for this data that accurately depicts BP’s progress in its daily collection of oil. What conclusions can you draw with your graphic?**

In this exercise, we need to use a new code:

barplot(VARIABLE)



Insert variable:

barplot(BBL)

We can use a lot of different parameters, to make the plot look better, such as:

main = "Maximum Temperatures in a Week",

xlab = "Degree Celsius",

ylab = "Day",

names.arg = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"),

col = "darkred",

horiz = TRUE

The code with some of the parameters:

barplot(BBL,xlab="DAY", ylab="BBL",col="grey", main="Barplot", border="black", names.arg = c("16", "17", "18", "19", "20", "21", "22", "23"))

This graph shows that there has not been steady improvement in the collecting of barrels as the other graph lead us to believe. There was an increase for 3 days, then a levelling off for 3 days and then a decrease.

**Supplementary Exercises on Methods for Describing Sets of Data**

**14. (137). Construct a relative frequency histogram for the data summarized in the following table:**



First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E14 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E14.xlsx")

View(L2E14)

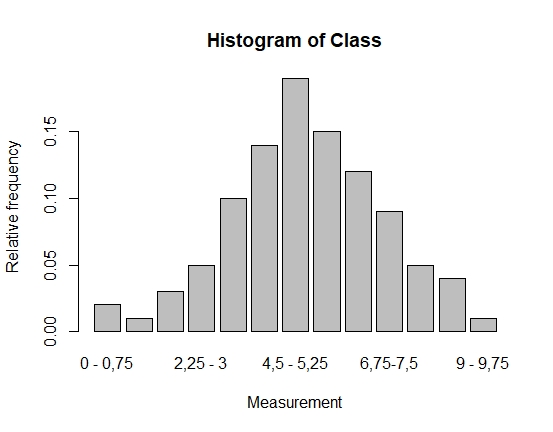
attach(L2E14)

Create variables: Measurement\_Class <- Measurement\_Class

Relative\_frequency <- Relative\_frequency

In this exercise, you have to make a relative frequency histogram for the Measurement Class. A relative frequency histogram is a type of graph that shows how often something happens, in percentages. However, in R the histograms are made with RAW data, and we do not have this data. Therefore, in this exercise you should make a barplot instead. We are using the barplot() function:

barplot(Relative\_frequency,xlab="Measurement", ylab="Relative frequency",col="grey", main="Histogram of Class", border="black", names.arg = Measurement\_Class)



This is the relative frequency histogram of Measurement Class.

**Exercise 15. (142). For each of the following data sets, compute the mean, the variance and the standard deviation. If appropriate, specify the units in which your answers are expressed.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

1. **4, 6, 6, 5, 6, 7.**

In this exercise, we do not have an Excel file. However, in R we can create a vector by our self. By using this code:

NAME <- c(VALUE, VALUE, VALUE, VALUE….)

Insert the values:

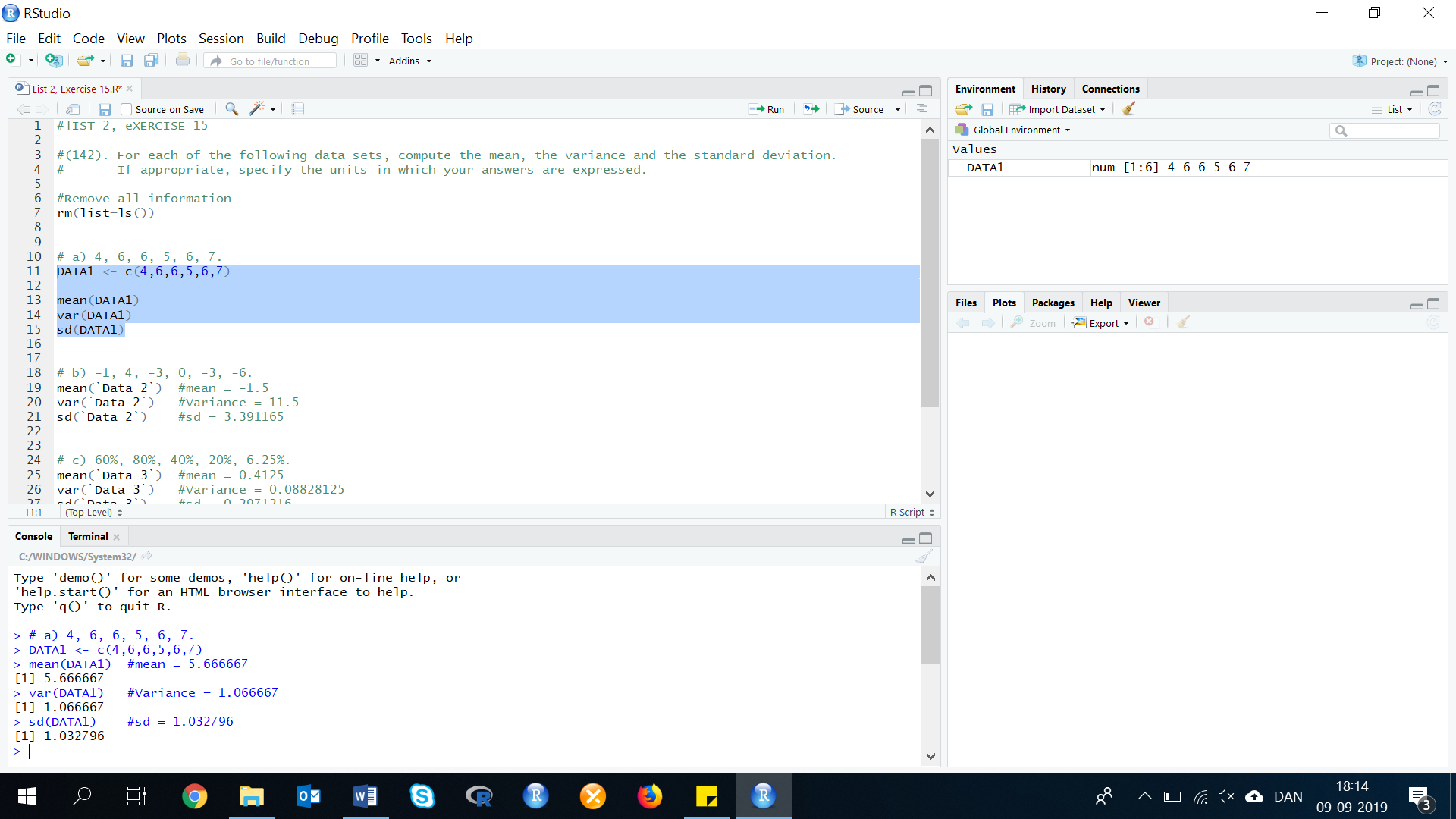
DATA1 <- c(4,6,6,5,6,7)

Now we can use DATA1, to calculate the mean, variance and standard deviation:

mean(DATA1)

var(DATA1)

sd(DATA1)

Result: The R console gives us:

Mean = 5.66667

Variance = 1.066667

Standard deviation = 1.032796

1. **-1, 4, -3, 0, -3, -6.**

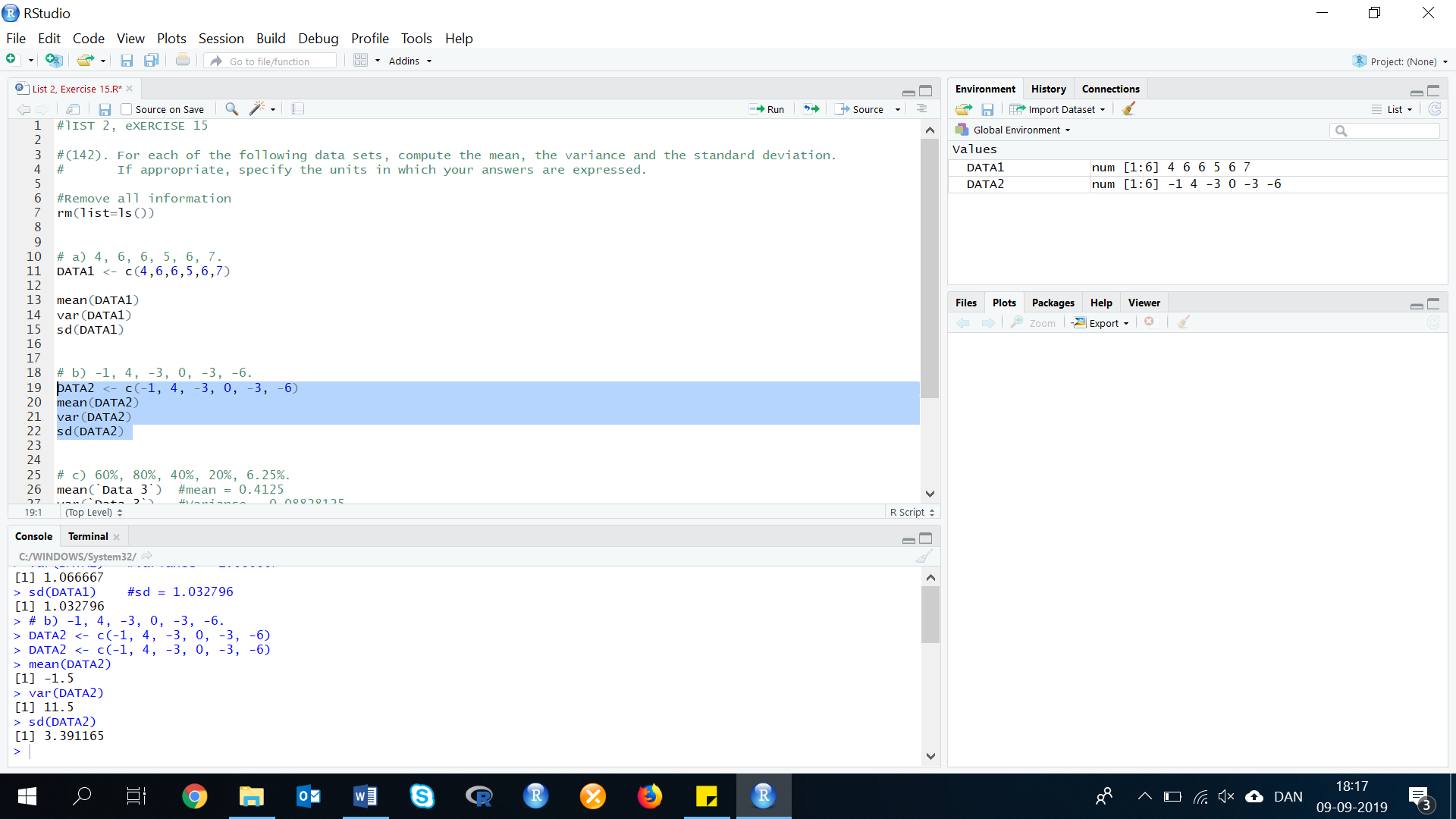
We are using the same method as in exercise a.

DATA2 <- c(-1, 4, -3, 0, -3, -6)

mean(DATA2)

var(DATA2)

sd(DATA2)

Results:

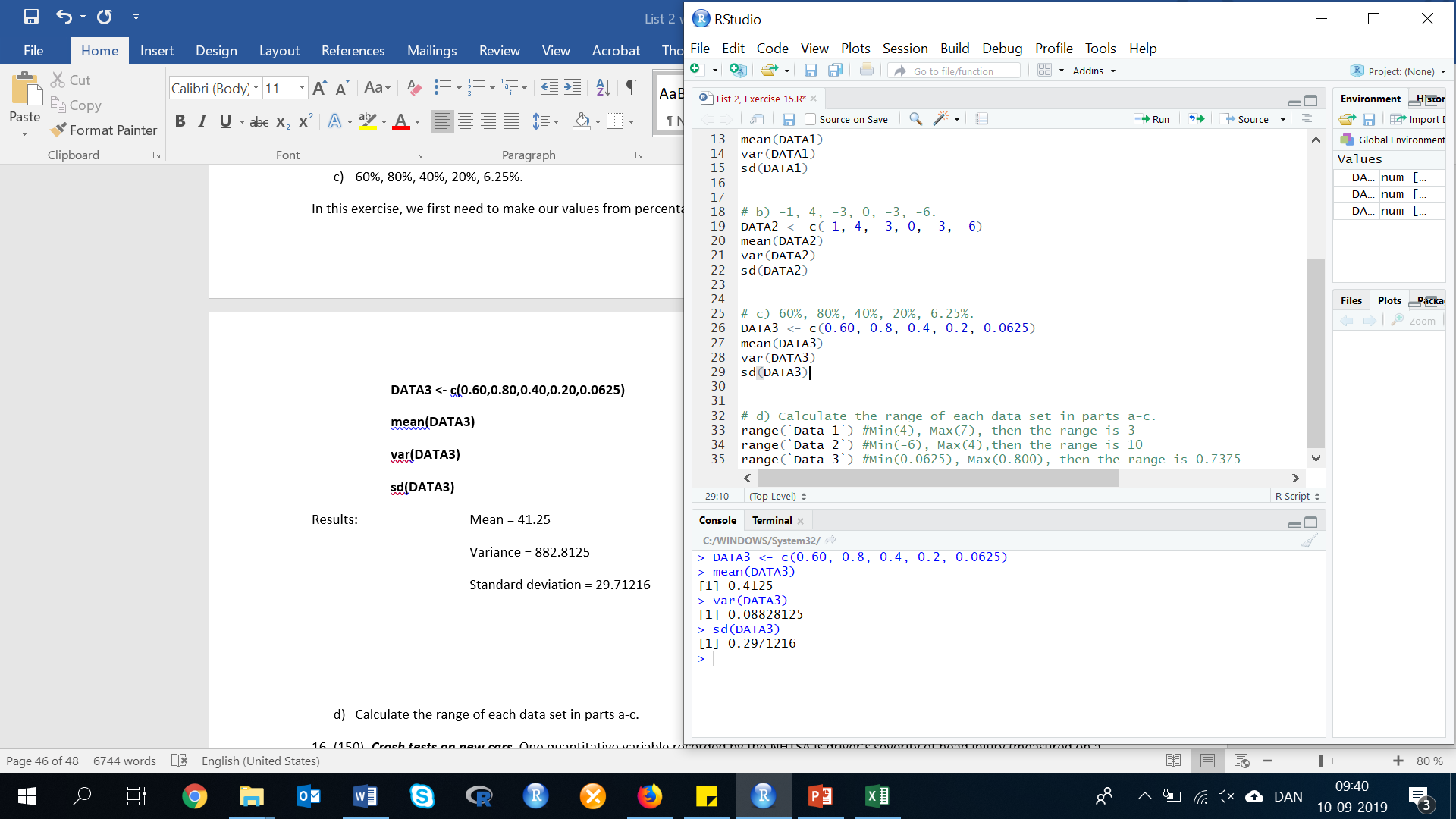
Mean = -1.5

Variance = 11.5

Standard deviation = 3.391165

1. **60%, 80%, 40%, 20%, 6.25%.**

In this exercise, we first need to make our values from percentage to numbers. Therefore, our data that should be created is now:

DATA3 <- c(0.60,0.80,0.40,0.20,0.0625)

mean(DATA3)

var(DATA3)

sd(DATA3)

Results: Mean = 0.4125 🡪 41.25 %

Variance = 0.08828125

Standard deviation = 0.2971216

1. **Calculate the range of each data set in parts a-c.**

To calculate the range of each of the datasets, we need to use the following code:

range(DATA)

Insert data:

range(DATA1)

range(DATA2)

range(DATA3)

Result: range(DATA1) = 7-3 = 3

range(DATA2) = 4-(-6)=4+6 = 10

range(DATA3) = 0.80-0.0625 = 0.7375

**Exercise 16. (150). *Crash tests on new cars*. One quantitative variable recorded by the NHTSA is driver’s severity of head injury (measured on a scale from 0 to 1500). The mean and standard deviation for the 98 driver head-injury ratings are displayed in the printout below.**



First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

1. **Give a practical interpretation of the mean.**

The mean(average) driver's head injury serverity in head-on collisions is 603.7

1. **Use the mean and the standard deviation to make a statement about where most of the head-injury ratings fall.**

The mean and the median is close in value, which indicates that the data should be symmetric. In this case we can apply, the empirical rule. We know, that 95% of all observations fall within 2 standard deviations of the mean.

1. **Find the z-score for a driver head-injury rating of 408. Interpret the result.**

To calculate the z-score we simply need the following values:

x = 408

mean = 603.7

standard deviation = 185.4

Calculation:

z<-(408-603.7)/185.4

Result: z = -1.055556

The value (-1.06) is fairly low and is not an unusual value to observe.

**17. (159, HULL). *Hull failures of oil tankers*. Owing to several major ocean oil spills by tank vessels, Congress passed the 1990 Oil Pollution Act, which requires all tankers to be designed with ticker hulls. Further improvements in the structural design of a tank vessel have been proposed since then, each with the objective of reducing the likelihood of oil spill and decreasing the amount of outflow in the event of a hull puncture. To aid in this development, Marine Technology (Jan. 1995) reported on the spillage amount (in thousands of metric tons) and cause of puncture for 42 major oil spills from tankers and carriers. Cause of puncture is classified as either collision (C), fire / explosion (FE), hull failure (HE) or grounding (G). The data are saved in the accompanying file.**

First, we need to do as followed:

1. Clean your screen
2. Run Library(mosaic)
3. Import data
4. Create variables

For explanation how to do this look at List 1, Exercise 1.

Clean the R Script: rm(list=ls())

Library: library(mosaic)

library(dplyr)

Import data from Excel: library(readxl)

L2E17 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E17.xlsx")

View(L2E17)

attach(L2E17)

Create variables: Spillage<-Spillage

Cause<-Cause

1. **Use a graphical method to describe the cause of oil spillage for the 42 tankers. Does the graph suggest that any one cause is more likely to occur than any other? How is this information of value to the design engineers?**

In this exercise, we will start with filtering the data, so we know how many observations there are of each of the 5 different types of causes of oil spillage:

COLLISION <- dplyr::filter(L2E17, Cause =="Collision")

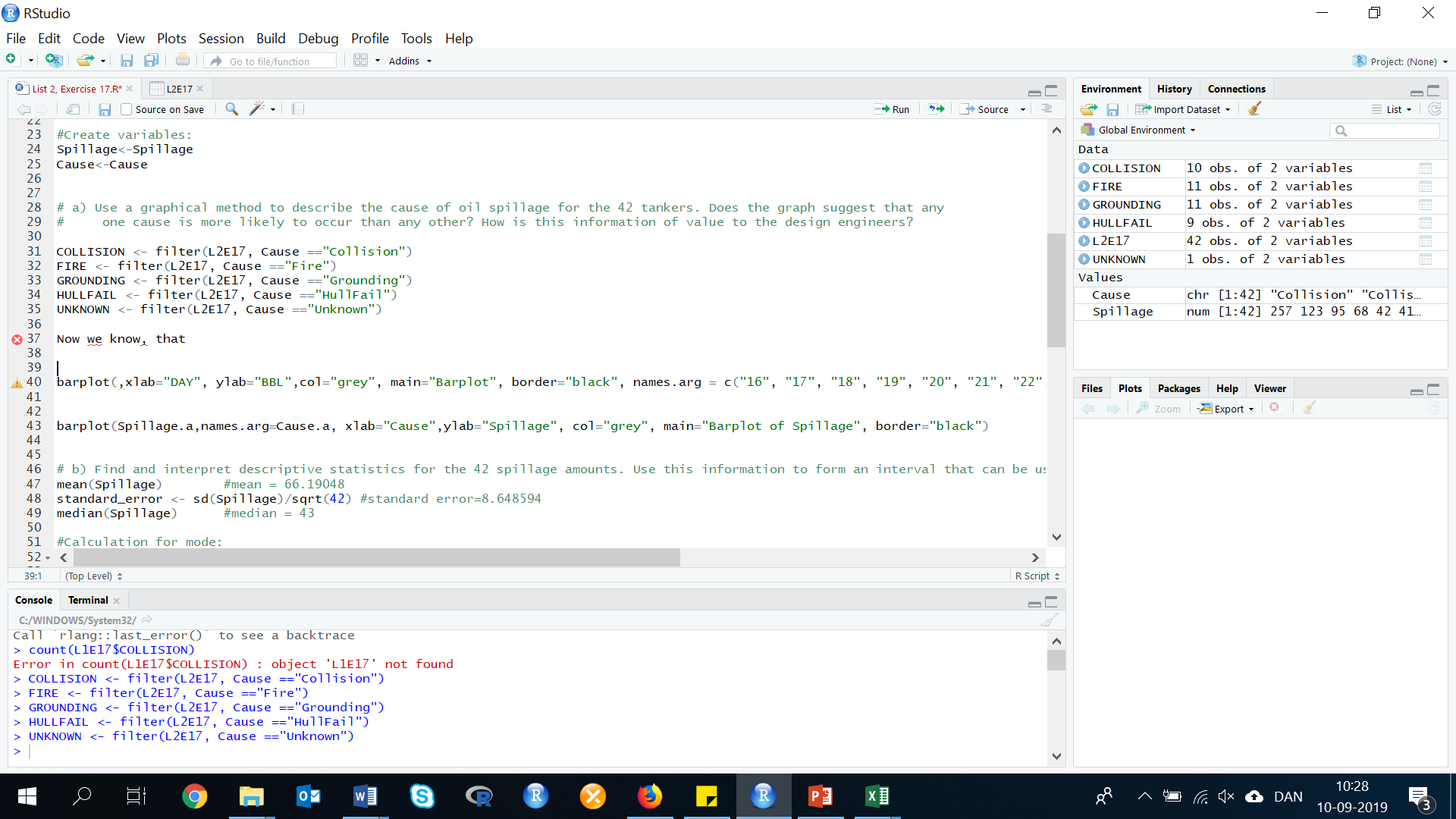
FIRE <- dplyr::filter(L2E17, Cause =="Fire")

GROUNDING <- dplyr::filter(L2E17, Cause =="Grounding")

HULLFAIL <- dplyr::filter(L2E17, Cause =="HullFail")

UNKNOWN <- dplyr::filter(L2E17, Cause =="Unknown")

We now know, that each variable have the following number of observations:

Collision = 10

Fire = 11

Grounding 11

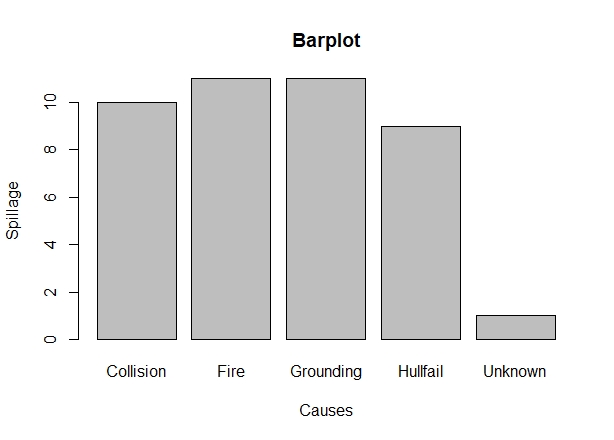
Hullfail = 9

Unknown = 1

With this information, we can make a new variable/vector/matrix. This dataset can we use to make a barplot of the causes of oil spillage for the 42 tankers.

NEW\_DATA <- c(10,11,11,9,1)

Insert this new variable, in the barplot code:

barplot(NEW\_DATA,xlab="Causes", ylab="Spillage",col="grey", main="Barplot", border="black", names.arg = c("Collision", "Fire", "Grounding", "Hullfail", "Unknown"))

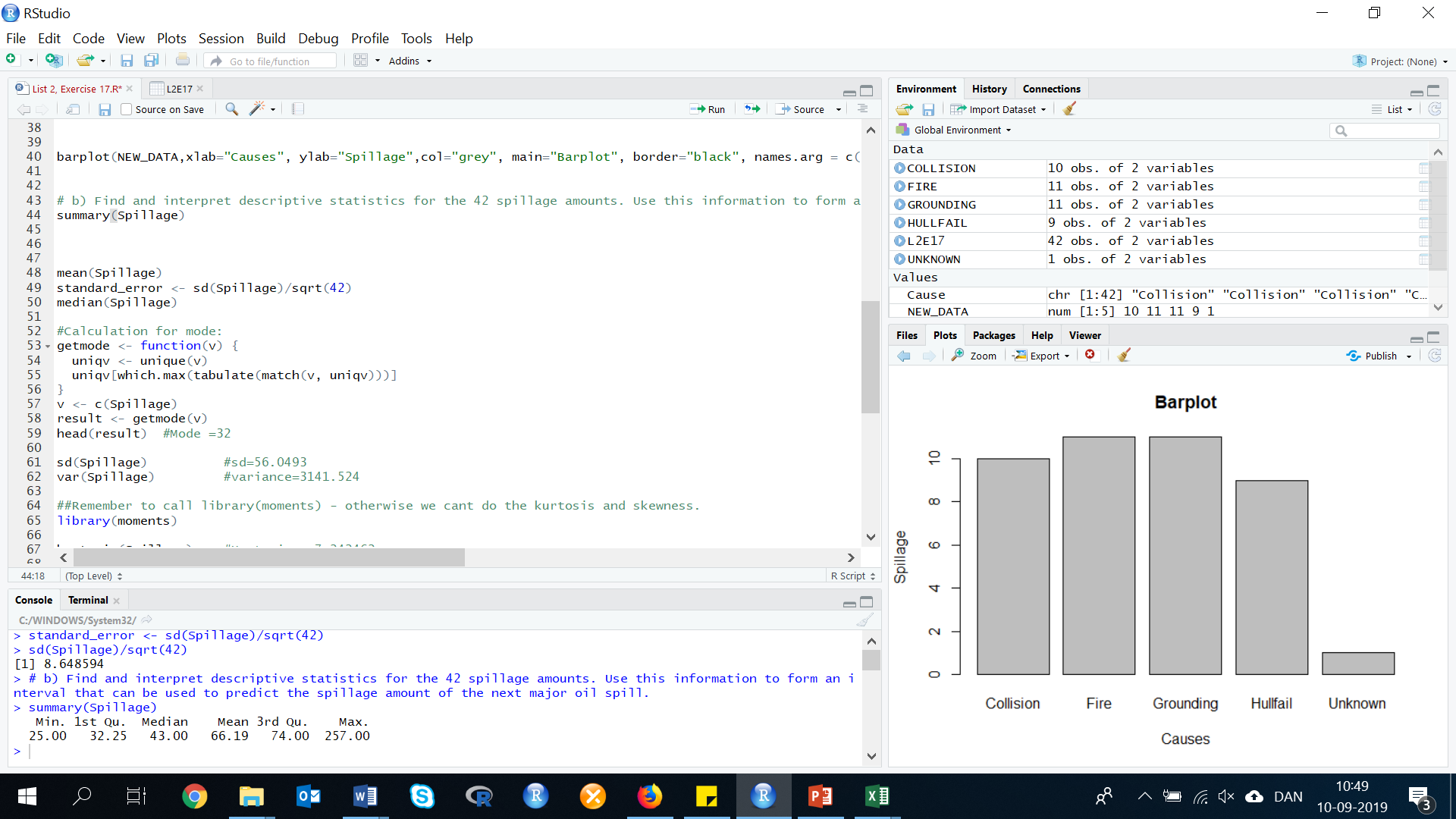
1. **Find and interpret descriptive statistics for the 42 spillage amounts. Use this information to form an interval that can be used to predict the spillage amount of the next major oil spill.**

In R, there is not a specific code, as in excel, where we can calculate all the needed descriptive statistic. However, there is some codes. We recommend you to either calculate it manually, or by using the same codes as we are in this exercise.

First, we use the R code below. We use this to calculate, the median, mean, range, minimum, maximum, Q1 and Q3:

summary(VARIABLE)

Insert variable: summary(Spillage)



Now we need to use the individual R codes, for each of the missing descriptive statistics:

|  |  |  |
| --- | --- | --- |
|  | **R Code** | **Result** |
| Standard Error | sd(Spillage)/sqrt(42) | 8.648594 |
| Mode | getmode <- function(v) {  uniqv <- unique(v)  uniqv[which.max(tabulate(match(v, uniqv)))]  }  mode <- getmode(Spillage)  print(mode) | 32 |
| Standard Deviation | sd(Spillage) | 56.0493 |
| Sample Variance | var(Spillage) | 3141.524 |
| Kurtosis | Install.package(fBasics)  library(fBasics)  kurtosis(Spillage) | 3.901689 |
| Skewness | library(fBasics)  skewness(Spillage) | 2.133413 |
| Sum | sum(Spillage) | 2780 |
| Count | length(Spillage) | 42 |
| Confidence Level(95,0%) | C\_error<-qnorm(0.975)\*standard\_error | 16.95093 |

Our level of certainty about the true mean is 95% in predicting that the true mean is within:

Confident1 <- mean(Spillage)+C\_error = 83.14141

Confident2 <- mean(Spillage)-C\_error = 49.23954

The interval between 49.23954 and 83.14141 assuming that the original random variable is normally distributed, and the samples are independent.

The mean spillage amount is 66.19 thousand metric tons, while the median is 43.00. Since the median is so much smaller than the mean, it indicates that the data are skewed to the right. The standard deviation is 56.05. Again, since this value is so close to the value of the mean it indicates that the data are skewed to the right.

Since the data are skewed to the right, we cannot use the Empirical Rule (the one about normal data and the bell curve, see excersize 5a and 5b) so we use Chebyshev's rule. Using this rule we know that 8/9 of the observations fall within 3 standard deviations of the mean.

**66.19 + 56.05 + 56.05 + 56.05 = 234.34** (or -101.96 except you can't have negative spillage) 8/9 of all the spillage is within 0 and 234.34 thousand metric tons

**Exercise 18. (162, WHEELS). *Time to develop price quotes*. A manufacturer of industrial wheels is losing many profitable orders because of the long time it takes the firm’s marketing, engineering, and accounting departments to develop price quotes for potential customers. To remedy this problem, the firm’s management would like to set guidelines for the length of time each department should spend developing price quotes. To help develop these guidelines, 50 requests for price quotes were randomly selected from the set of price quotes made last year: the processing time (in days) was determined for each price quote for each department. Several observations are displayed in the table below. The price quotes are also classified by whether or not they were “lost” (i.e., whether or not the customer placed an order after receiving the price quote).**



Clean the R Script: rm(list=ls())

Library: library(mosaic)

Import data from Excel: library(readxl)

L2E18 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List2/L2E18.xlsx")

View(L2E18)

attach(L2E18)

Create variables: REQUEST <- REQUEST

MARKET <- MARKET

ENGINEER <- ENGINEER

ACCOUNT <- ACCOUNT

TOTAL<- TOTAL

LOST<- LOST

1. **Construct a stem-and-leaf display for the total processing time for each department. Shade the leaves that correspond to “lost” orders in each of the displays, and interpret each of the displays.**

First, we must compute the total processing times by adding the processing times of the three departments. To do this, we use the code:

cbind(VARIABLE1, VARIABLE2, VARIABLE3, VARIABLEn)

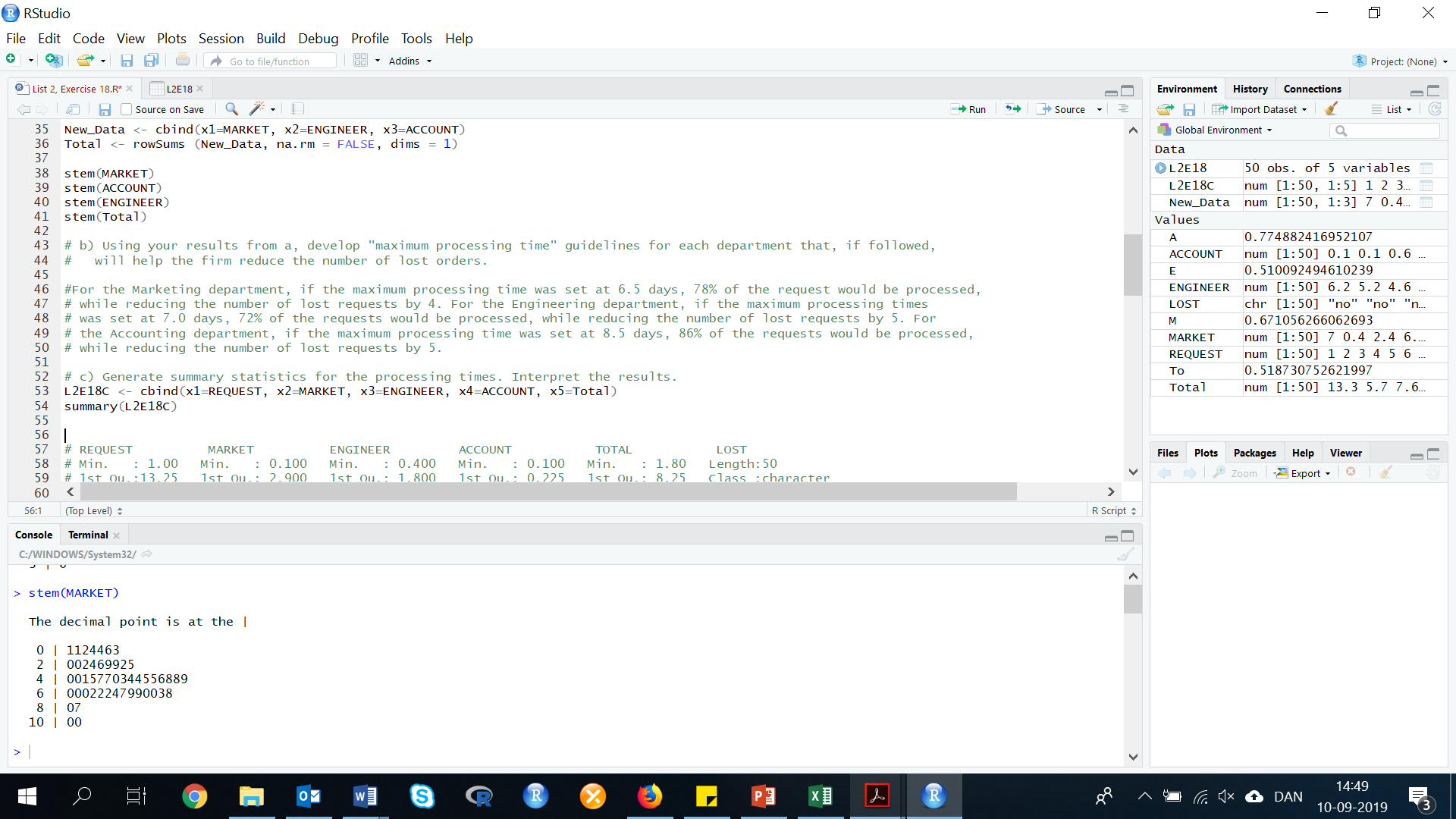
New\_Data <- cbind(x1=MARKET, x2=ENGINEER, x3=ACCOUNT)

Now we have created the data, that we need to calculate the total processing times. We use the code:

rowSums (DATA, na.rm = FALSE, dims = 1)

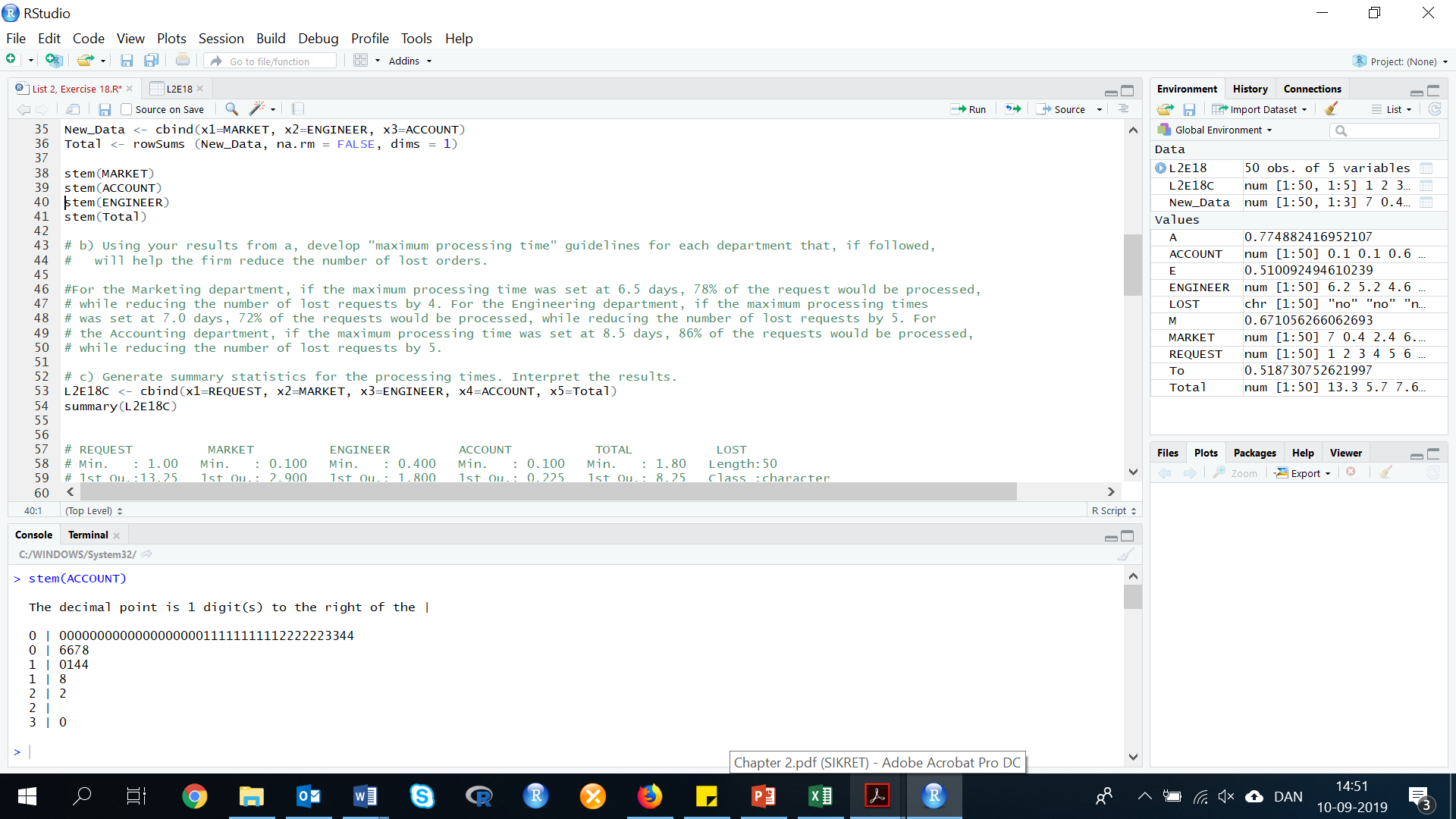
The total processing times are as follows:

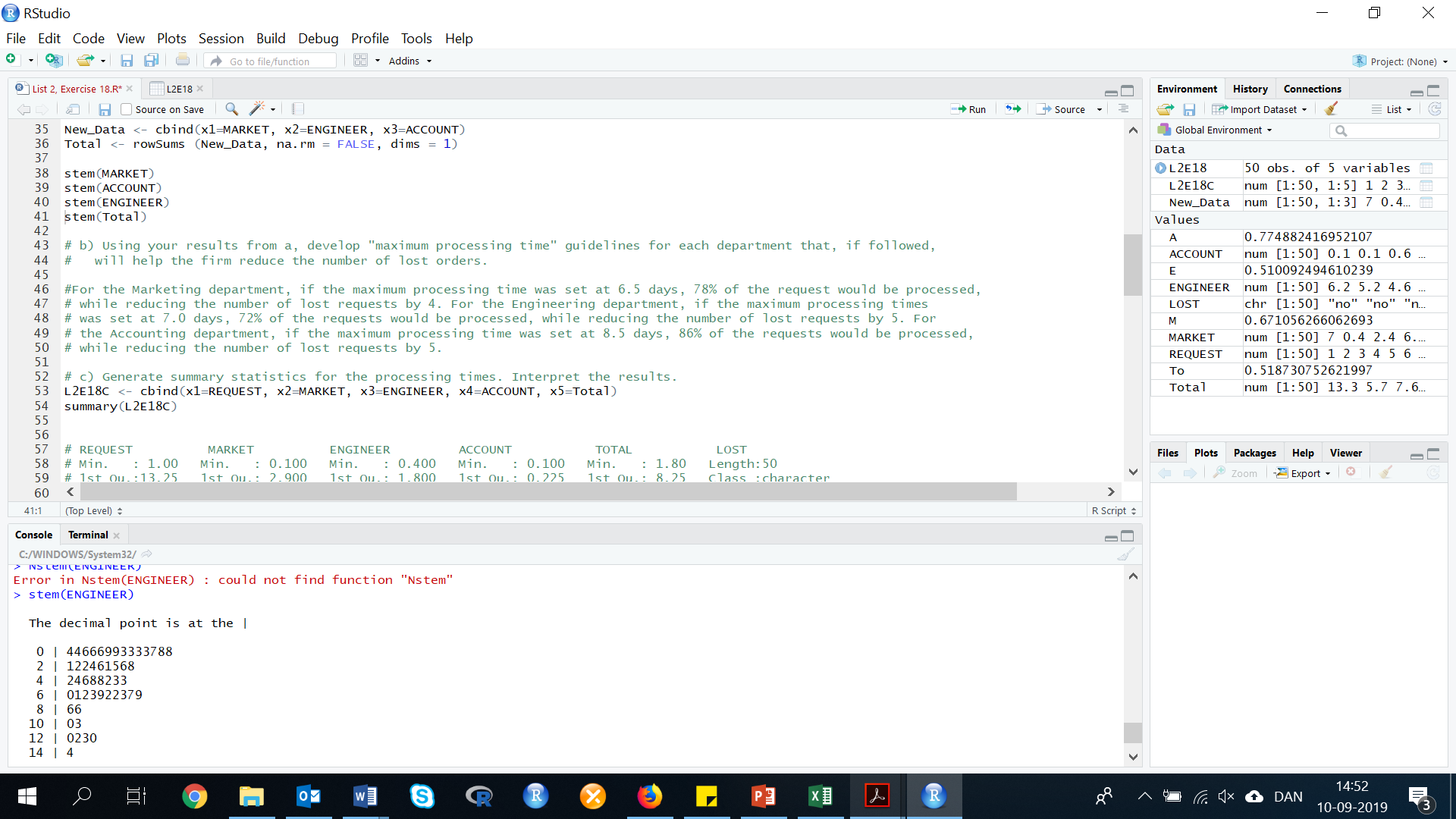
Total <- rowSums (New\_Data, na.rm = FALSE, dims = 1)

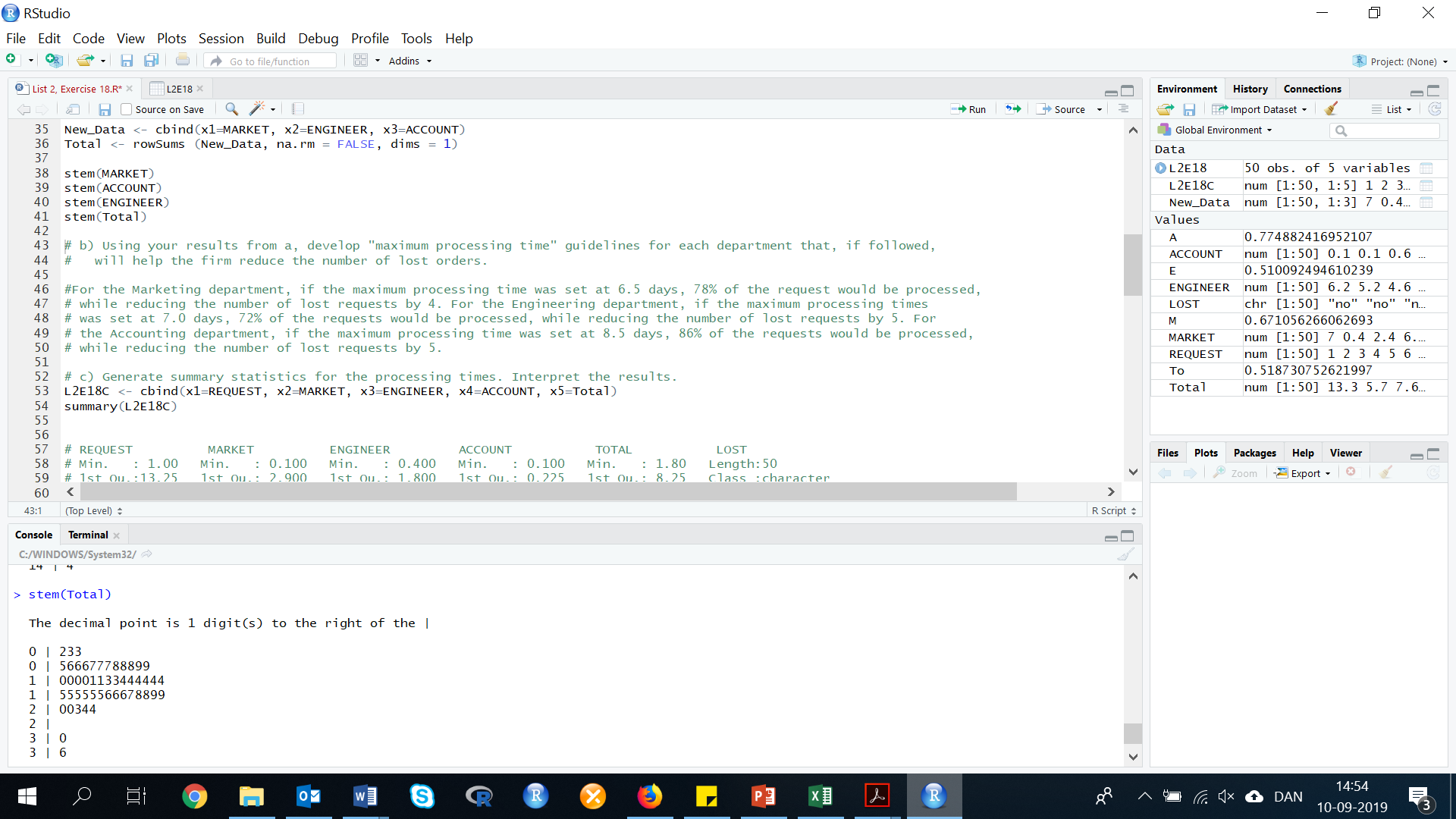
Now we can construct the stem and leaf plot:

stem(VARIABLE)

stem(MARKET)

stem(ACCOUNT)

 stem(ENGINEER)

stem(Total)

Of the 50 requests, 10 were lost. For each of the three departments, the processing times for the lost requests are scattered throughout the distributions. The processing times for the departments do not appear to be related to whether the request was lost or not. However, the total processing times for total processing times could be kept under 17 days, 76% of the data could be maintained, while reducing the number of lost requests to 1.

1. **Using your results from a, develop “maximum processing time” guidelines for each department that, if followed, will help the firm reduce the number of lost orders.**

For the Marketing department, if the maximum processing time was set at 6.5 days, 78% of the request would be processed, while reducing the number of lost requests by 4.

For the Engineering department, if the maximum processing times was set at 7.0 days, 72% of the requests would be processed, while reducing the number of lost requests by 5.

For the Accounting department, if the maximum processing time was set at 8.5 days, 86% of the requests would be processed, while reducing the number of lost requests by 5.

1. **Generate summary statistics for the processing times. Interpret the results.**

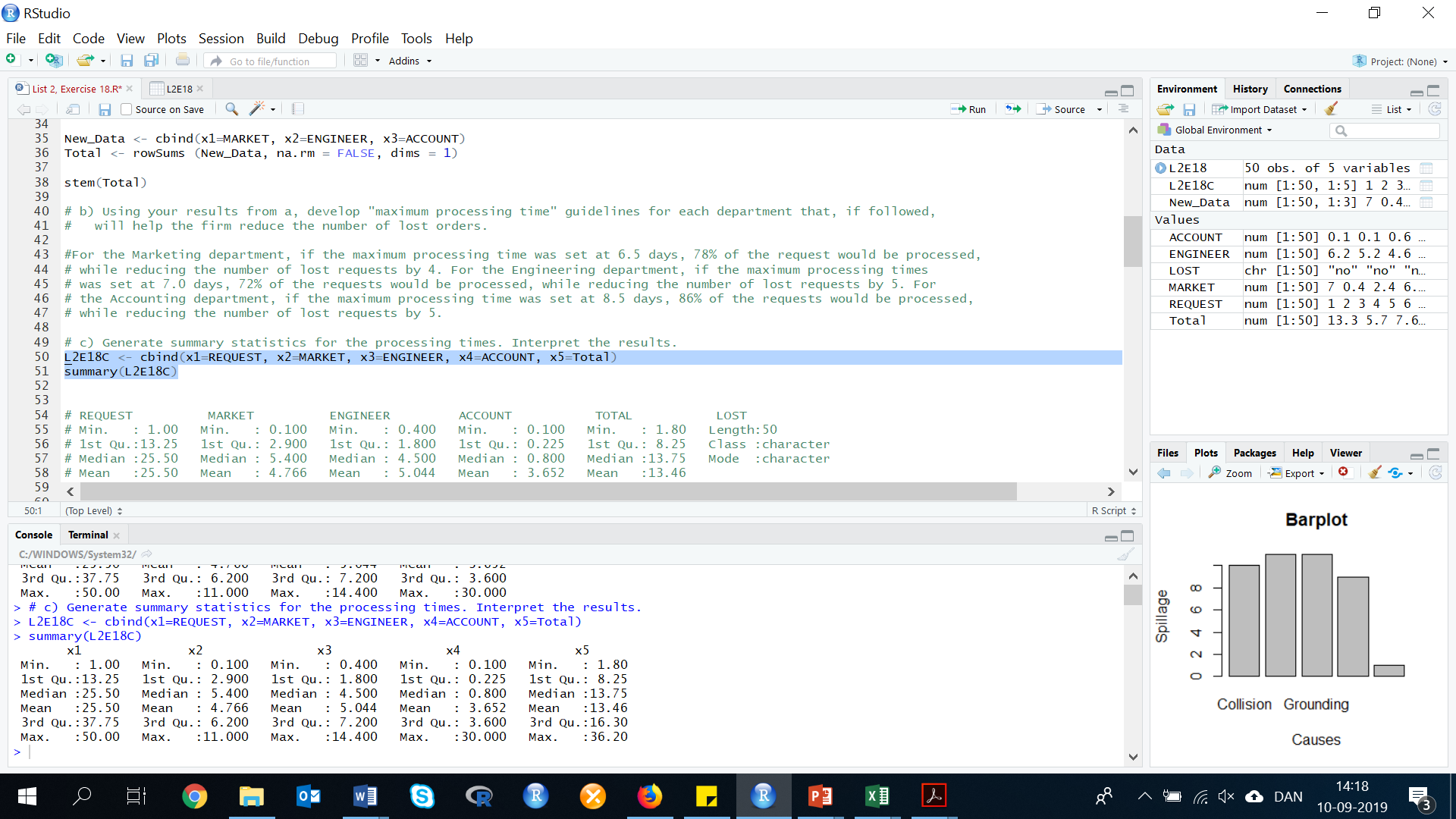
In this exercise, we need to bind the variables Request, Market, Engineer, Account and Total together. To do this, we will use the following code:

cbind(VARIABLE1, VARIABLE2, VARIABLE3, VARIABLEn)

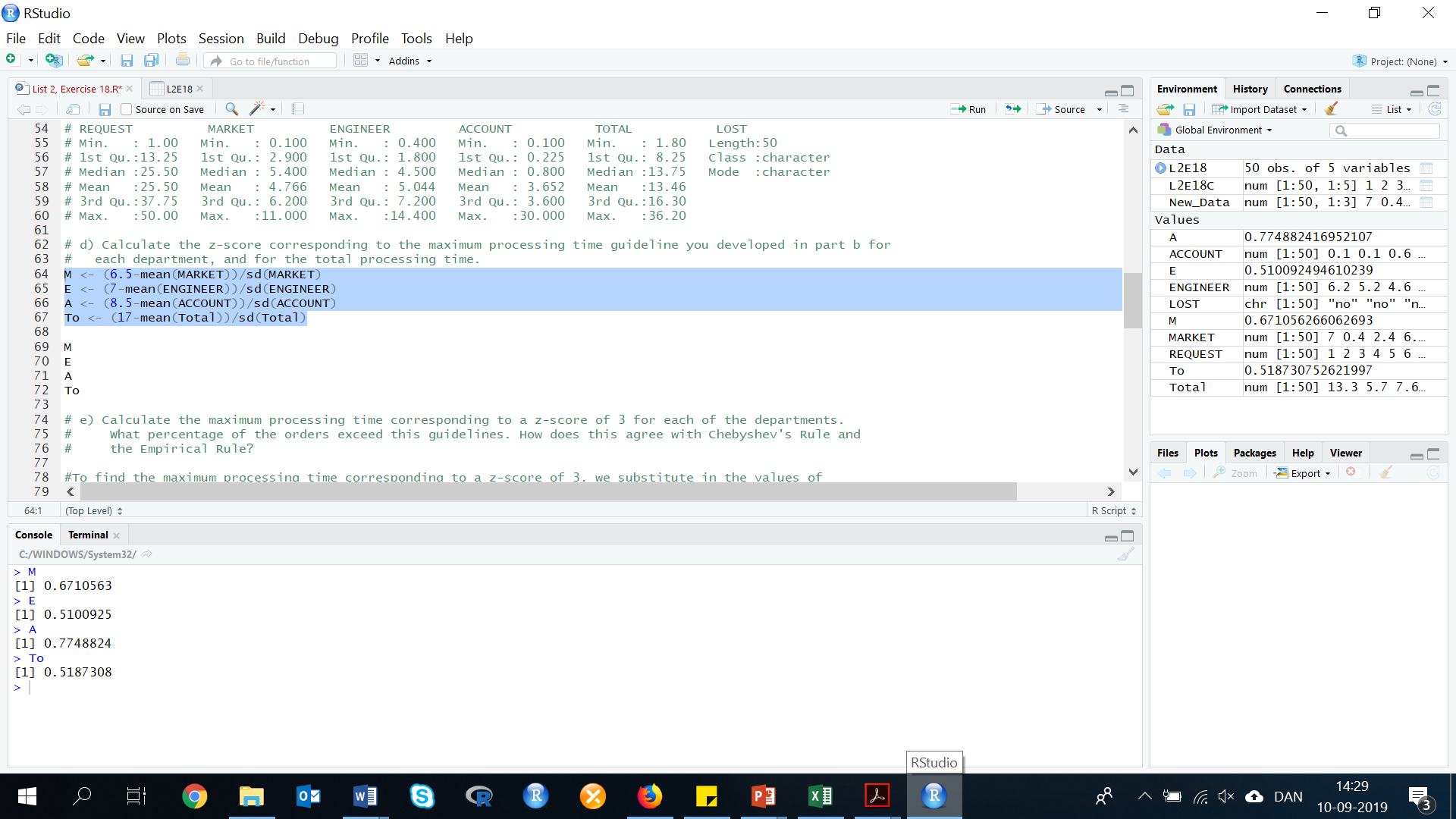
L2E18C <- cbind(x1=REQUEST, x2=MARKET, x3=ENGINEER, x4=ACCOUNT, x5=Total)

summary(L2E18C)

The R Console gave us the following result:



1. **Calculate the z-score corresponding to the maximum processing time guideline you developed in part b for each department, and for the total processing time.**

The z-scores corresponding to the maximum times guidelines developed for each department and the total are as follows. We use the following equation to measure the z-scores:

M <-(6.5-mean(MARKET))/sd(MARKET)

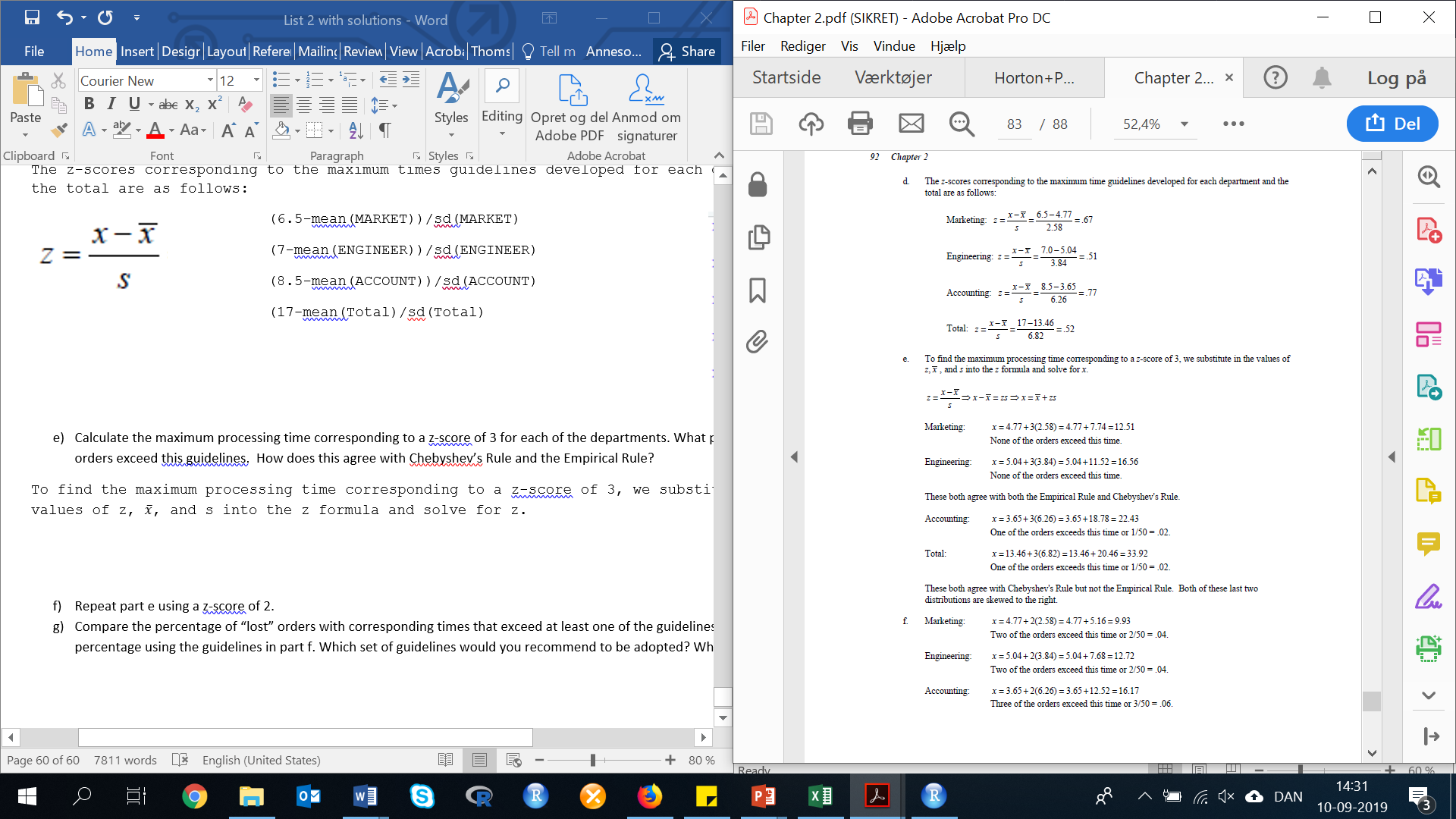
E <- (7-mean(ENGINEER))/sd(ENGINEER)

A <- (8.5-mean(ACCOUNT))/sd(ACCOUNT)

To <- (17-mean(Total)/sd(Total)

1. **Calculate the maximum processing time corresponding to a z-score of 3 for each of the departments. What percentage of the orders exceed this guidelines. How does this agree with Chebyshev’s Rule and the Empirical Rule?**

To find the maximum processing time corresponding to a z-score of 3, we substitute in the values of z, , and s into the z formula and solve for z.



Marketing: mean(MARKET)+3\*sd(MARKET)

Result: 12.51796

Engineering: mean(ENGINEER)+3\*sd(ENGINEER)

Result: 16.5478

These both agree with both the Empirical Rule and Chebyshev’s Rule.

Accounting: mean(ACCOUNT)+3\*sd(ACCOUNT)

Result: 22.4213

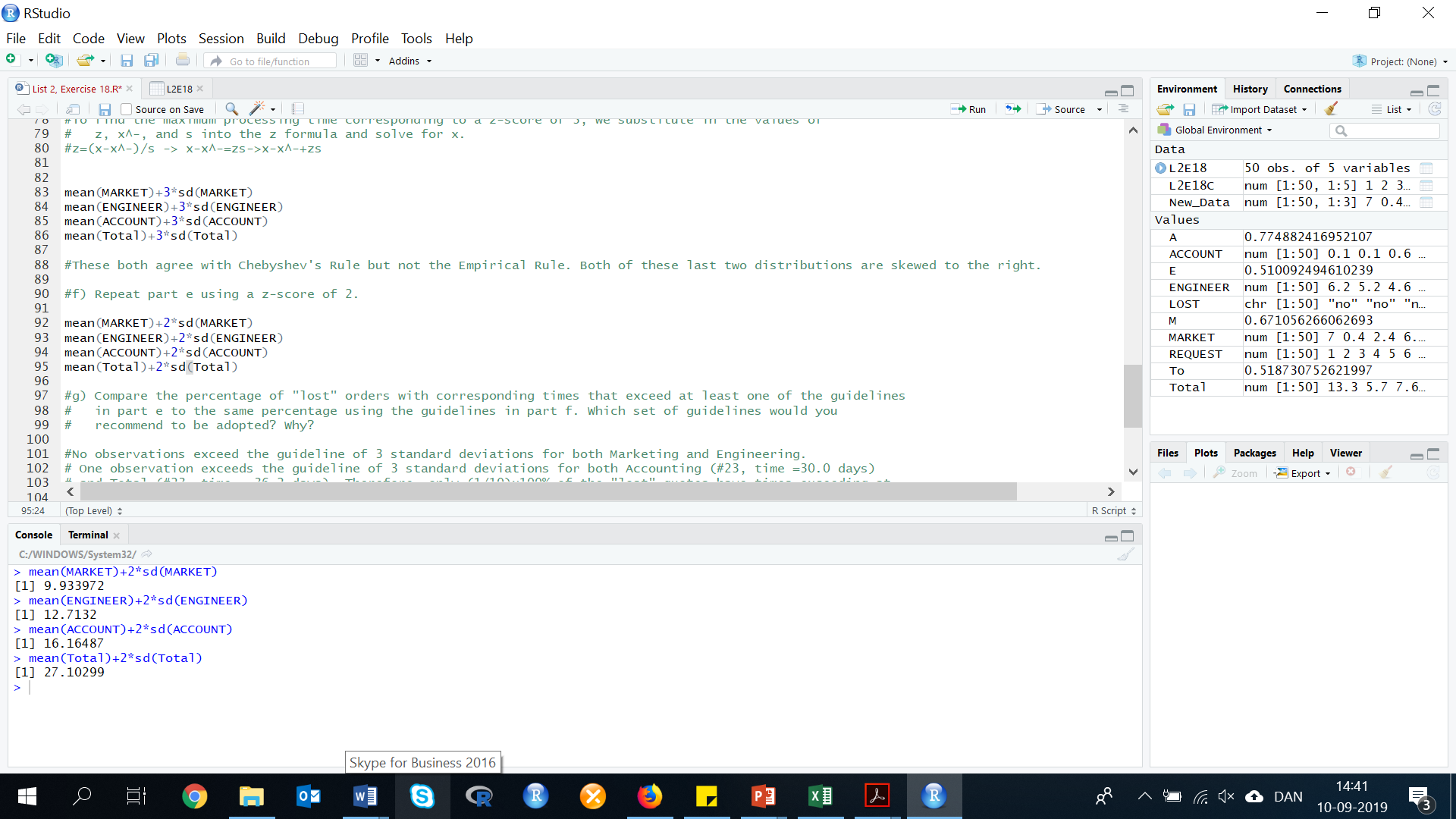
Total: mean(Total)+3\*sd(Total)

Result: 33.92348

These both agree with Chesbyshev’s Rule but not the Empirical Rule. Both of these last two distributions are skewed to the right.

1. **Repeat part e using a z-score of 2.**

Lets now calculate with z = 2.

Marketing: mean(MARKET)+2\*sd(MARKET)

Result: 9.933972

Engineering: mean(ENGINEER)+2\*sd(ENGINEER)

Result: 12.7132

These both agree with both the Empirical Rule and Chebyshev’s Rule.

Accounting: mean(ACCOUNT)+2\*sd(ACCOUNT)

Result: 16.16487

Total: mean(Total)+2\*sd(Total)

Result: 27.10299

1. **Compare the percentage of “lost” orders with corresponding times that exceed at least one of the guidelines in part e to the same percentage using the guidelines in part f. Which set of guidelines would you recommend to be adopted? Why?**

No observations exceed the guideline of 3 standard deviations for both Marketing and Engineering. One observation exceeds the guideline of 3 standard deviations for both Accounting (#23, time =30.0 days) and Total (#23, time = 36.2 days). Therefore, only (1/10)x100% of the "lost" quotes have times exceeding at least one of the 3 standard deviation guidelines.

Two observations exceed the guideline of 2 standard deviations for both Marketing (#31, time =11.0 days and #48, time=10.0 days) and Engineering (#4, time =13.0 days and #49, time = 14.4 days). Three observations exceed the guideline of 2 standard deviations for Accounting (#20, time = 22.0 days; #23, time = 30.0 days; and #36, time = 18.2 days). Two observations exceed the guideline of 2 standard deviations for Total (#20, time = 30.2 days and #23, time = 36.2 days). Therefore, (7/10)\*100%=70% of the "lost" quotes have times exceeding at least one the standard deviation guidelines.

We would recommend the standard deviation guideline since it covers 70% of the lost quotes, while having very few other quotes exceed the guidelines.