**Quantitative Methods**

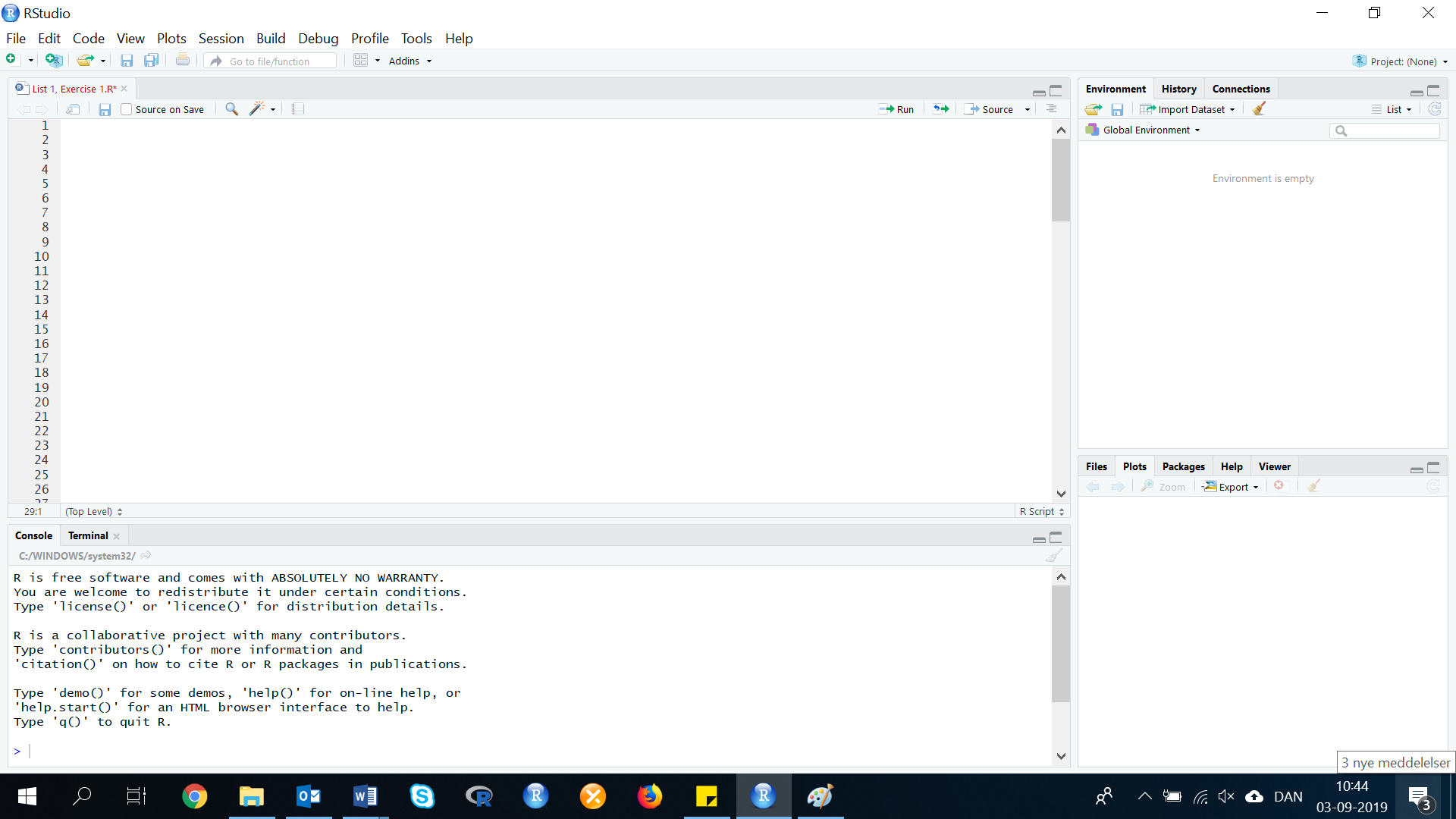
**List of Exercises N. 1**

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

**2.2 Graphical Methods to describe Quantitative Data**

**Exercise 1: (24, SANIT). *Sanitation inspection of cruise ships*.** **To minimize the potential for gastrointestinal disease outbreaks, all passenger cruise ships arriving at U.S. ports are subject to unannounced sanitation inspections. Ships are rated on a 100-point scale by the Centers for Disease Control and Prevention. A score of 86 or higher indicates that the ships is providing an accepted standard of sanitation. The latest (as of Jan. 2010) sanitation scores for 186 cruise ships are saved in the accompanying file. The first 5 and last 5 observations in the data set are listed in the accompanying table.**

Open a new R Script, and your screen should look like the picture below.



Before you start making the exercise you need to clean your screen, open the library and import the data:

*Clean your screen:*

Start with cleaning your screen by removing all data or codes that could be running. Use the following code:

rm(list=ls())

*Install or call the library:*

In this exercise, you need to use the library: MOSAIC

If the package is not install in your R Studio, then install it and afterwards call the library.

Code:

install.packages("mosaic")

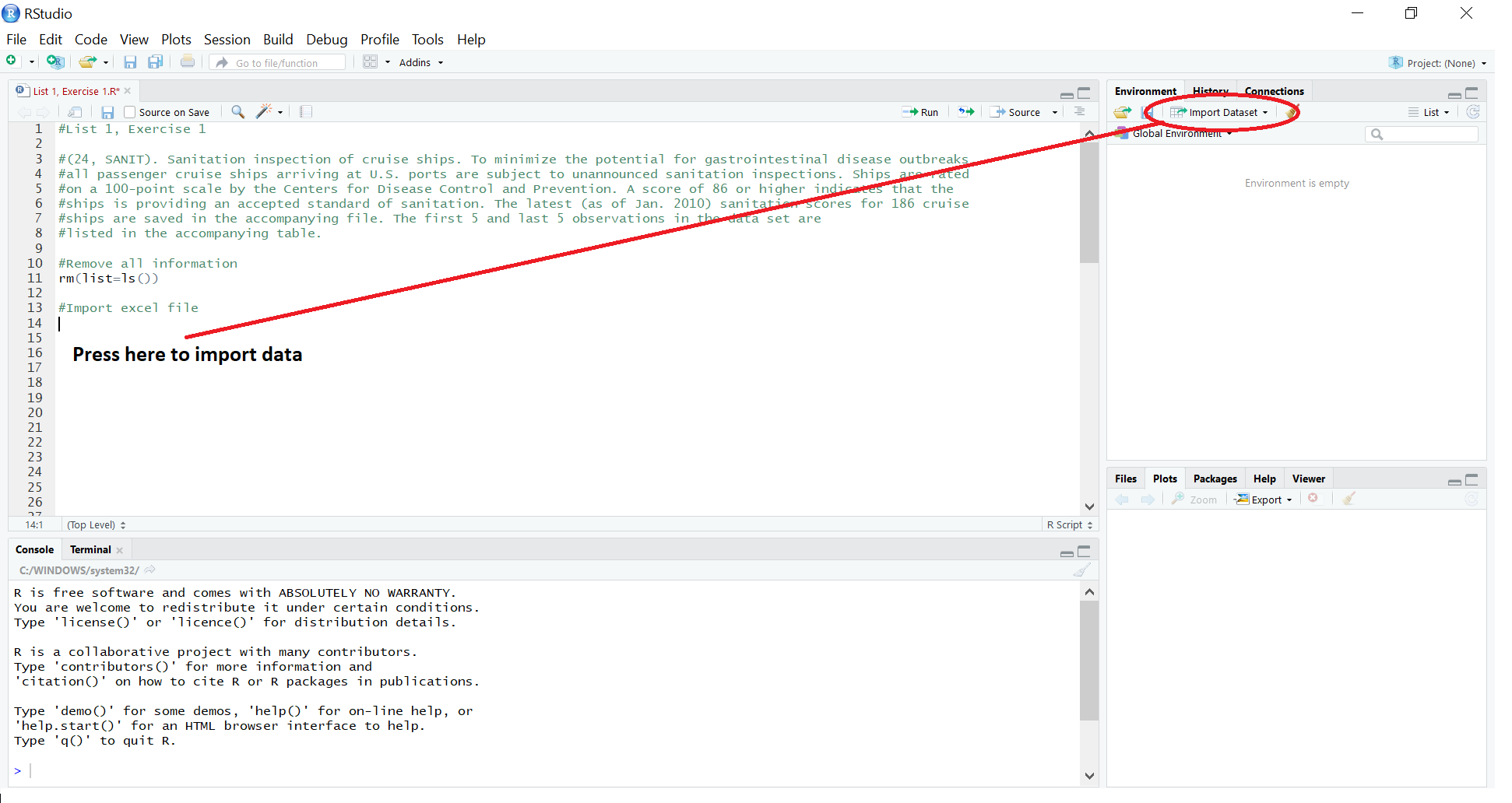
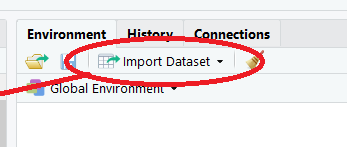
library(mosaic)

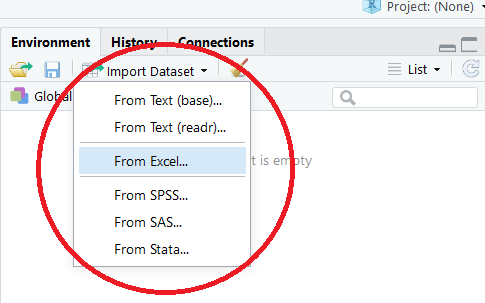
Remember to press ‘Run’ after each line.

*Import data:*

This section has five steps, follow those before you start the exercise.

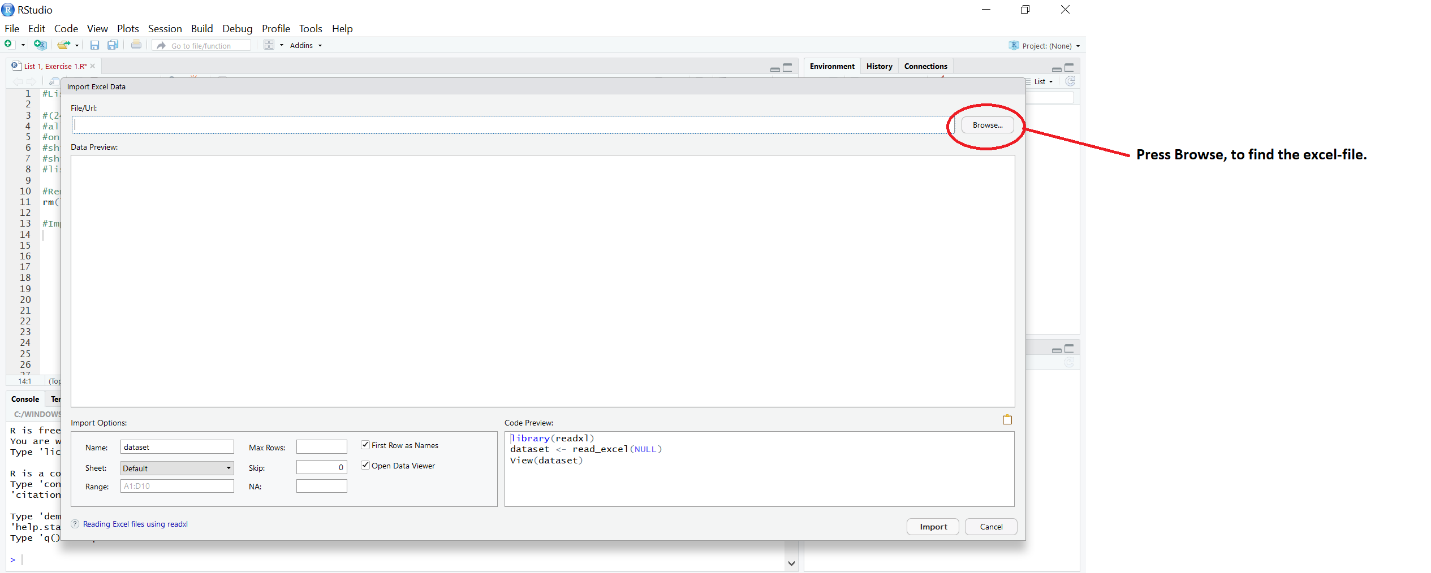
Now you need to import your data from either an Excel or CSV file. In this solution, we will import data from Excel. You can do it in two ways, but in this exercise we will only show the easiest way. See the picture below:

Step 1: Press ‘Import Dataset’.



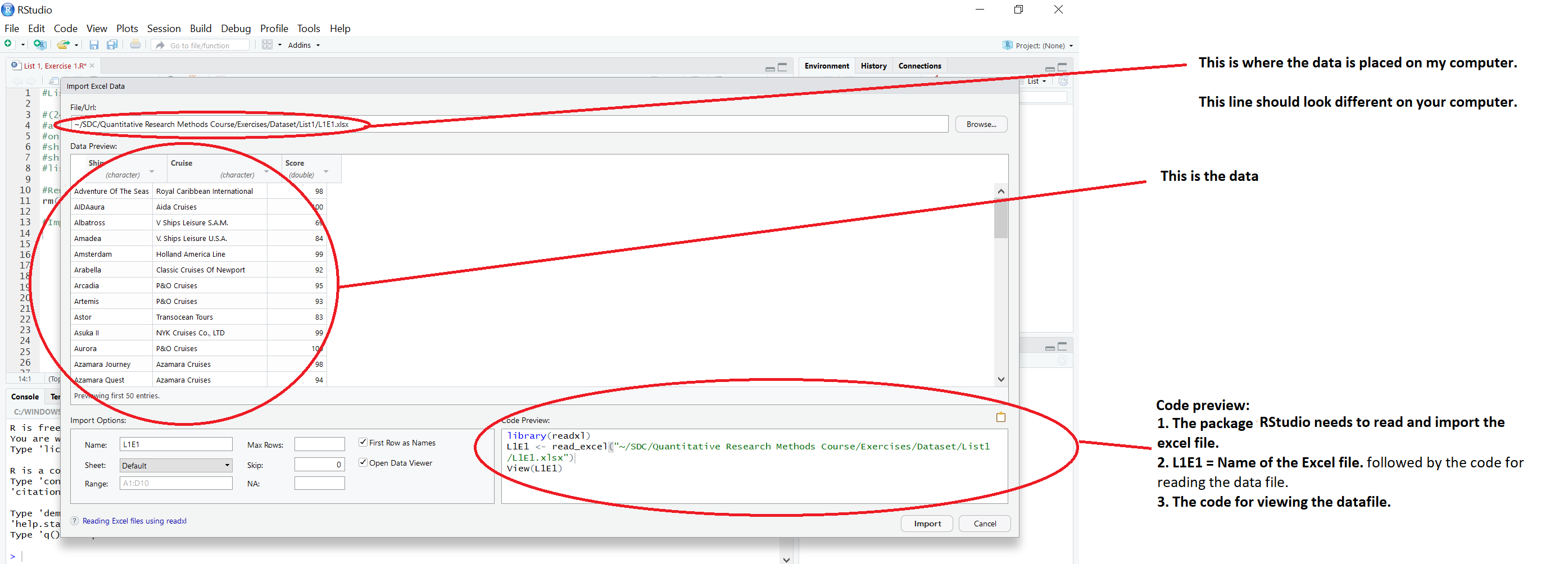
Step 2: Press ‘From …’

Choose which type of data that you would like to import. If you have a CSV file you should use ‘From Text’. However, in this solution we will import data from Excel. Press the bottom ‘From Excel’.



Step 3: Press ‘Browse’ to select your data

When you have browsed and clicked on your data file. Your screen should look almost like the picture below:



To make it really easy for yourself, instead of pressing import, then copy the code preview (press: Ctrl + C), and insert it into your open R Script.

library(readxl)

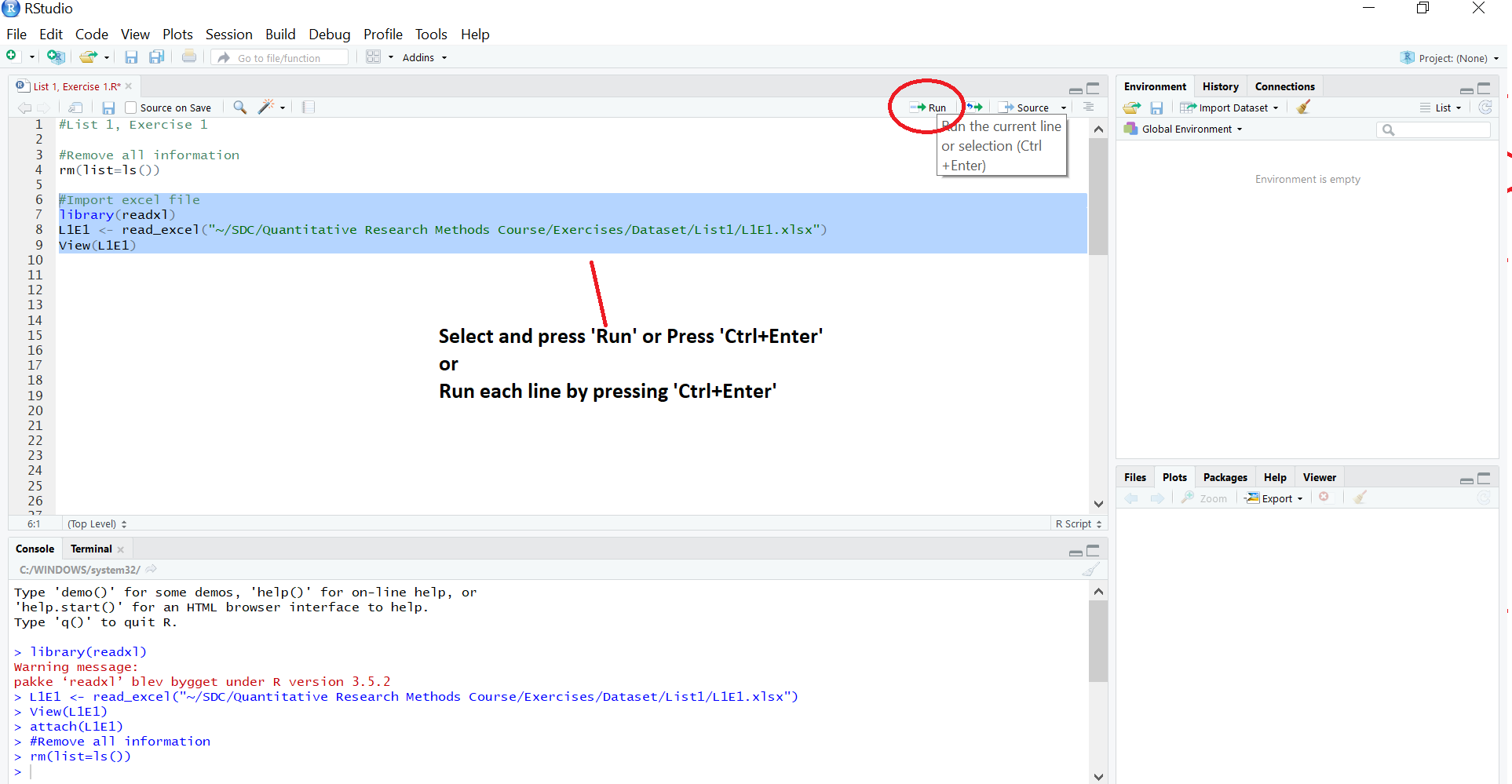
L1E1 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List1/L1E1.xlsx")

View(L1E1)

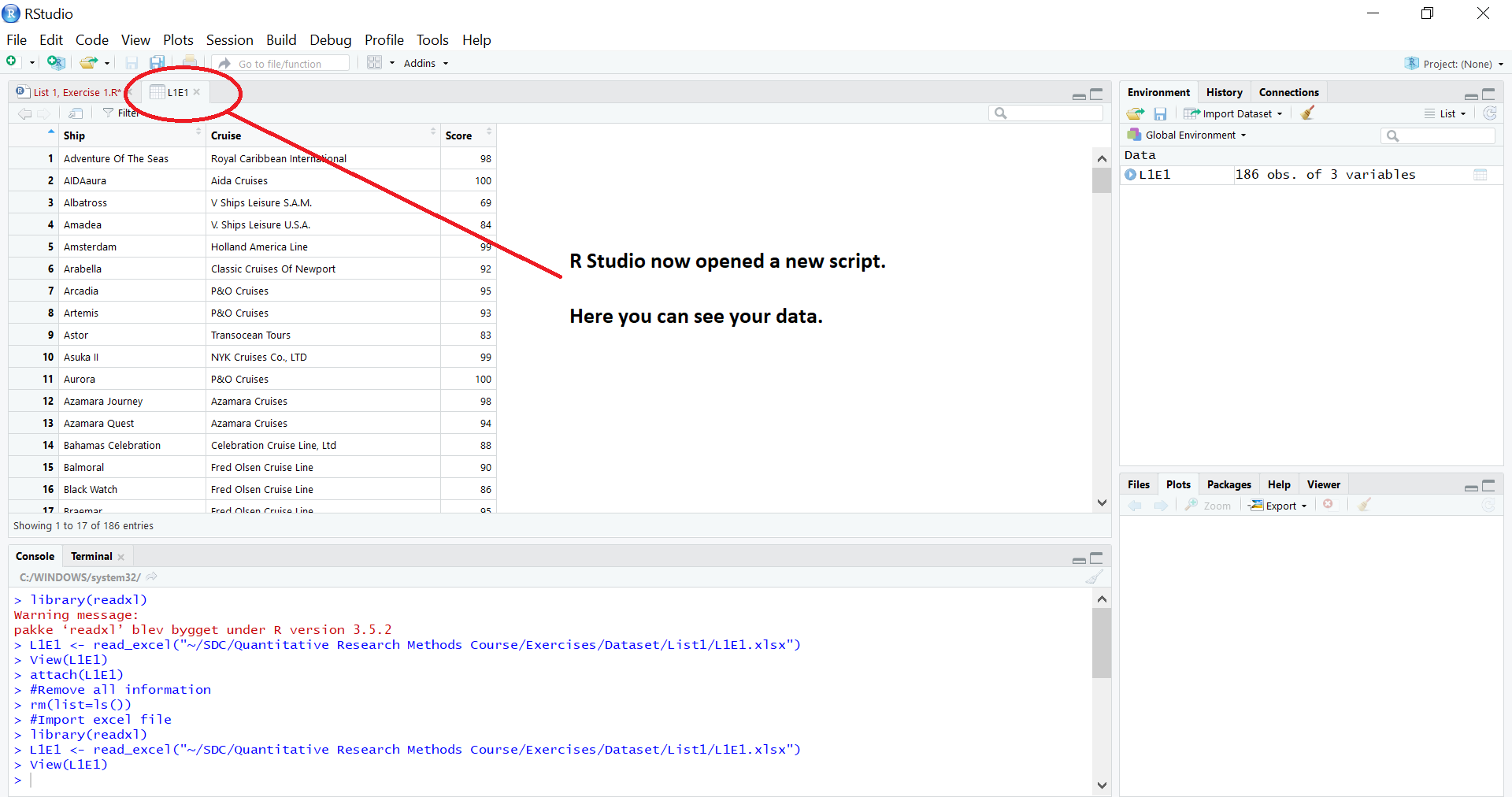
attach(L1E1)

Step 4: Run the code

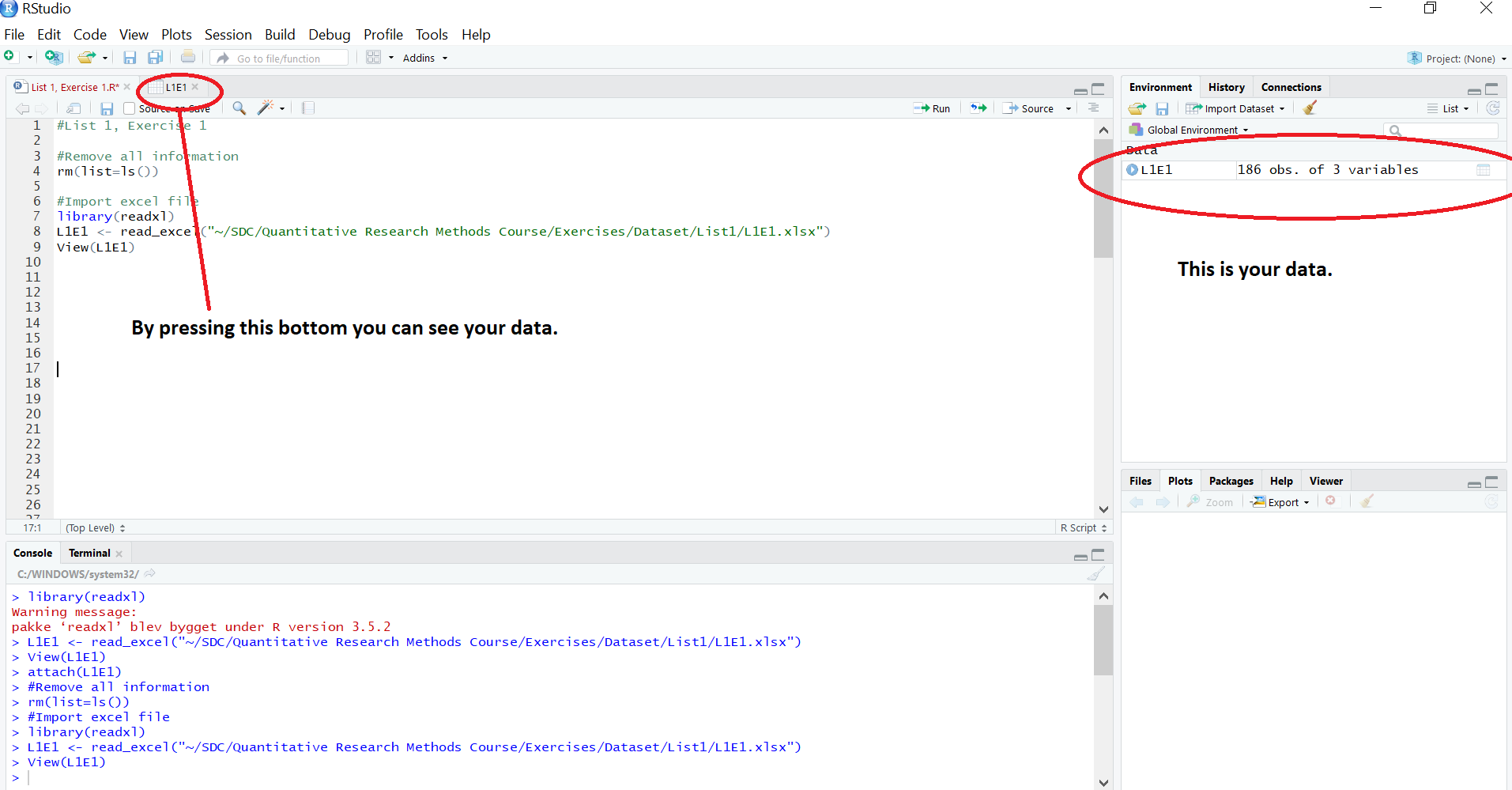
Now the code is in your R Script, and you need to run the code.



The screen should now look like this:



The picture, that appears, explains the different things:



Step 5: Identify and create the different variables:

When you have imported your data, it is important that you attach the dataset. If you do not attach the dataset, R Studio can’t identify the variables.

Write the following code and press run:

attach(L1E1)

or if your data is named differently, then:

attach(DATA)

Now R Studio have identified your variables. Then you need to create the variables in R Script. If you do not create these variables, it is not possible for you to work with the data that you have imported. To create the variables in your R Script, you need to write the names of each variable, so R Studio knows how to identify them in your excel file.

Write the following code and press run:

Ship <- Ship

Cruise <- Cruise

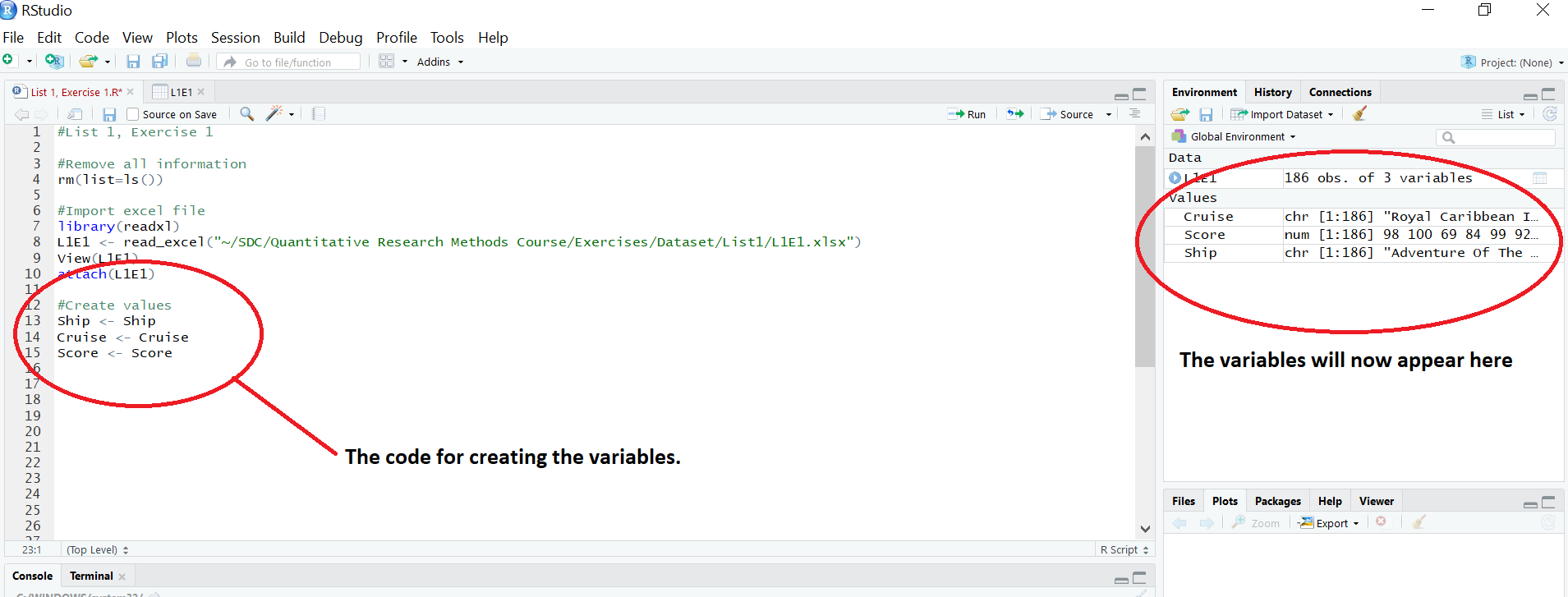
Score <- Score

If your variables are named differently, then you need to make sure that you write the exact name, so pay attention if the name is written with CAPS LOCK, small letters, numbers or a combination. Example:

NAME <- NAME

NAME (IN FRONT OF <- ) is the Name you give the variable. NAME (BEHIND OF <- ) is the name of the variable in the dataset you have imported.

The variables also called ‘Values’ in R Script are now created, and should look like this:



Now you are ready to calculate exercise a, b, and c.

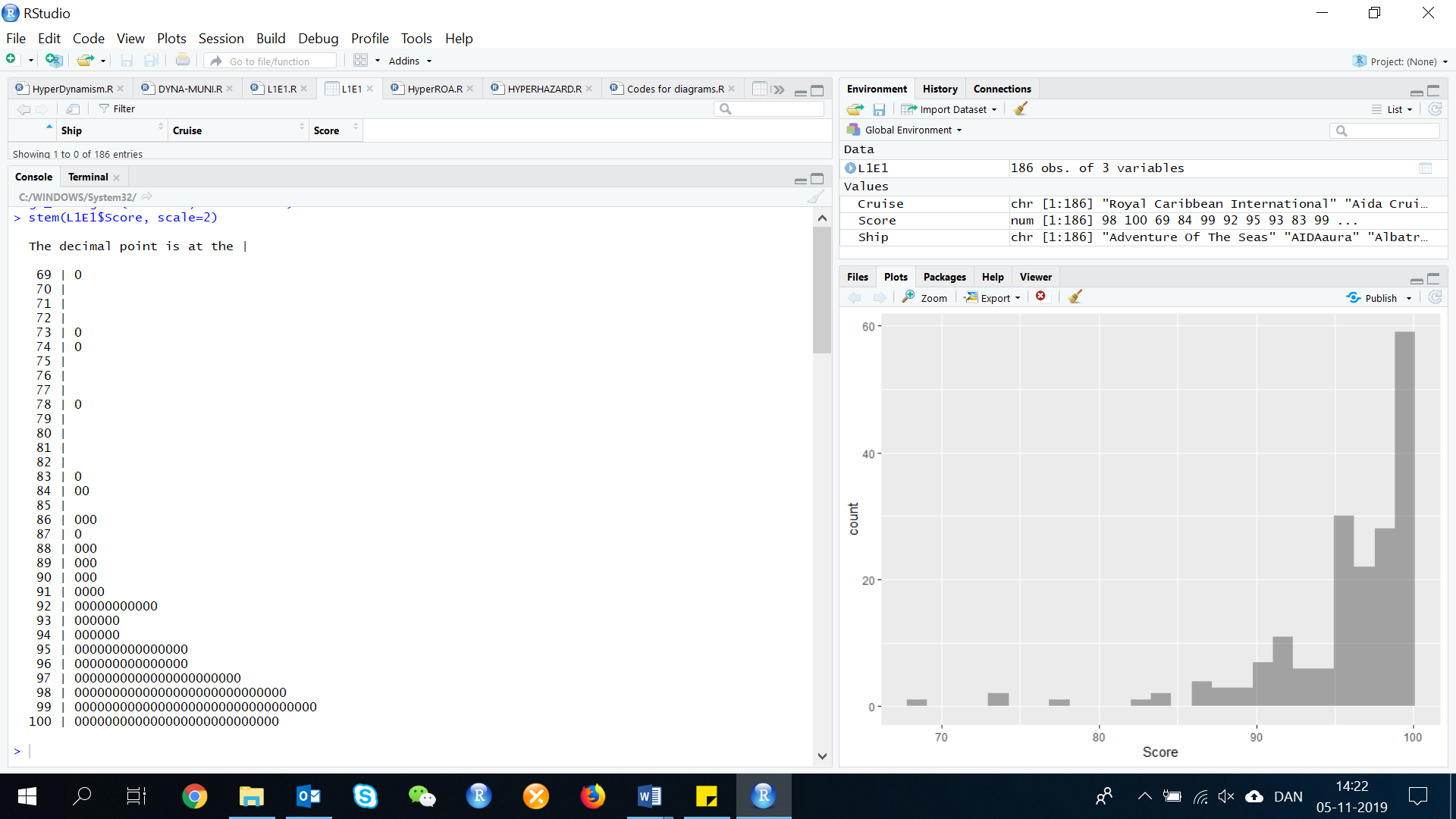
1. **Generate both a stem-and-leaf display and histogram of the data.**

In the first exercise, you need to run both a Stem and Leaf diagram and a Histogram.

Stem-and-leaf:

stem(VARIABLE)

stem(Score, scale=2)



The bigger the scale, the bigger the size of the stem and leaf display.

Histogram:

To make a histogram write the one of the following codes:

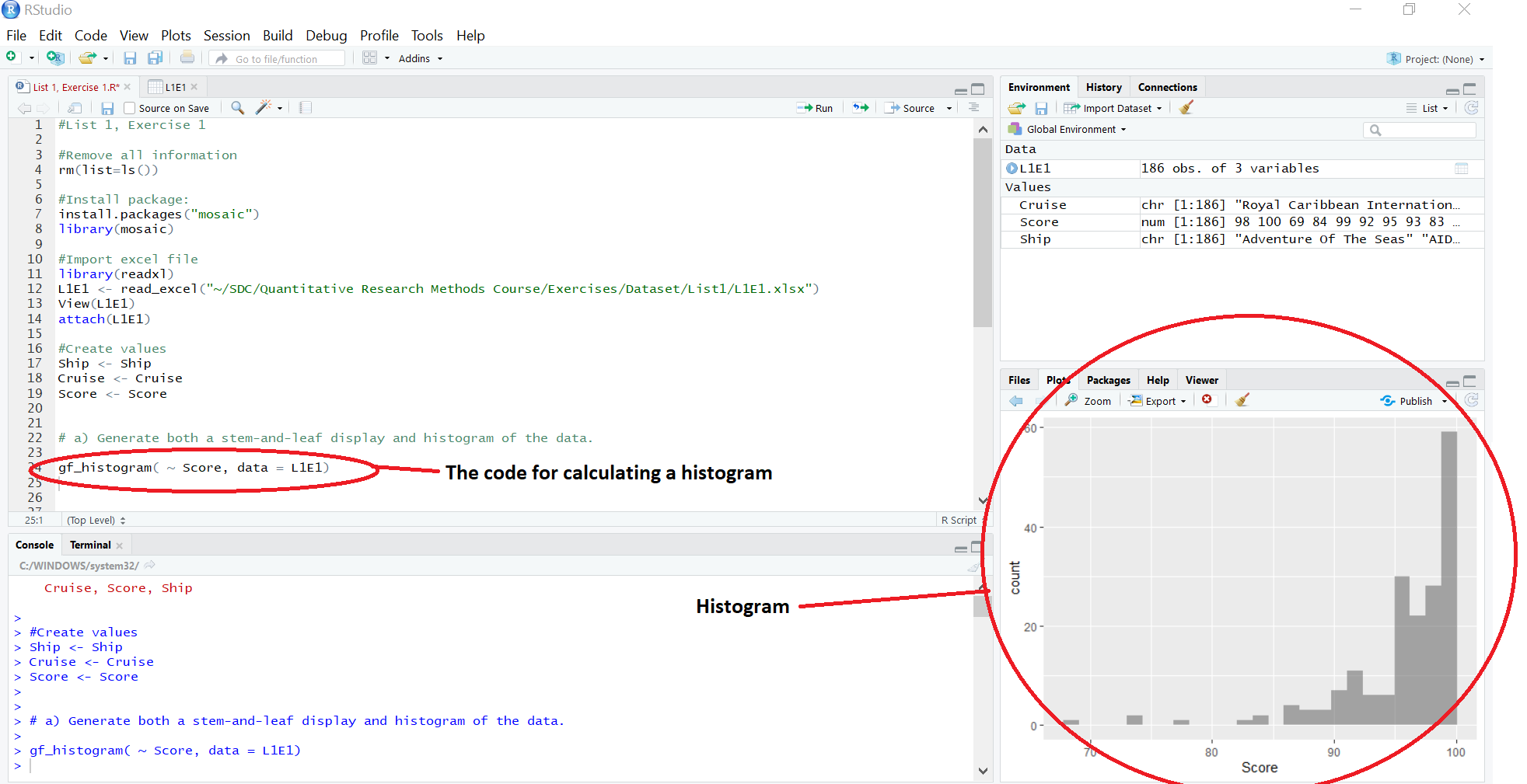
gf\_histogram( ~ VARIABLE, data = DATA)

hist(VARIABLE)

Insert the variable and data in the parentes

gf\_histogram( ~ Score, data = L1E1)

Now your R Script should show the histogram under the section plots, as seen below.



1. **Use the graphs to estimate the proportion of ships that have an accepted sanitation standard. Which graph did you use?**

From stem and leaf plot you can count 7 scores below 86. We can read of the raw data that we have 186 scores. (100/186)\*(186-7)= 96.24 % of the ships have an accepted sanitation standard.

1. **Locate the inspection score of 69 (Albatross) on the graph. Which graph did you use?**

We used the Stem and leaf diagram, which shows the score.

**Exercise 2: (26, NFL). *Most valuable NFL teams*. Each year, Forbes reports on the value of all teams in the National Football League. Although England’s soccer team, Manchester United, is the most valuable team in the world (1.8 billion USD), the NFL now has 15 teams worth at least 1 billion USD. For 2011, Forbes reports that the Dallas Cowboys are the most valuable team in the NFL, worth 1.85 billion USD. The current values in millions of USD of all 32 NFL teams, as well as the percentage changes in the values from 2010 to 2011, debt-to-value ratios, annual revenues, and operating incomes are listed in the table of the file NFL.**

Before you start answering the questions in the exercise, you have to clean the script, load/call the library ‘Mosaic’, import data and create the variables/values. Use the following codes:

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

Library ‘mosaic’: library(mosaic)

Import data: library(readxl)

L1E2 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List1/L1E2.xlsx")

View(L1E2)

attach(L1E2)

Create variables: Rank <- Rank

Team <- Team

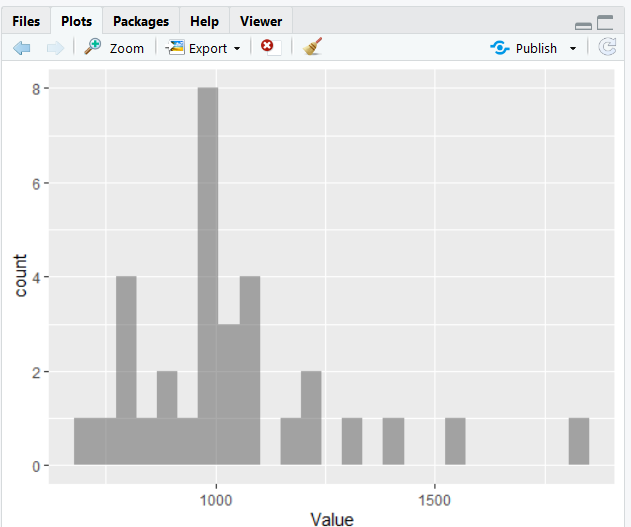
Value <- Value

Chang1Yr <-Chang1Yr

Debt\_Value <- Debt\_Value

Revenue <- Revenue

Income <- Income

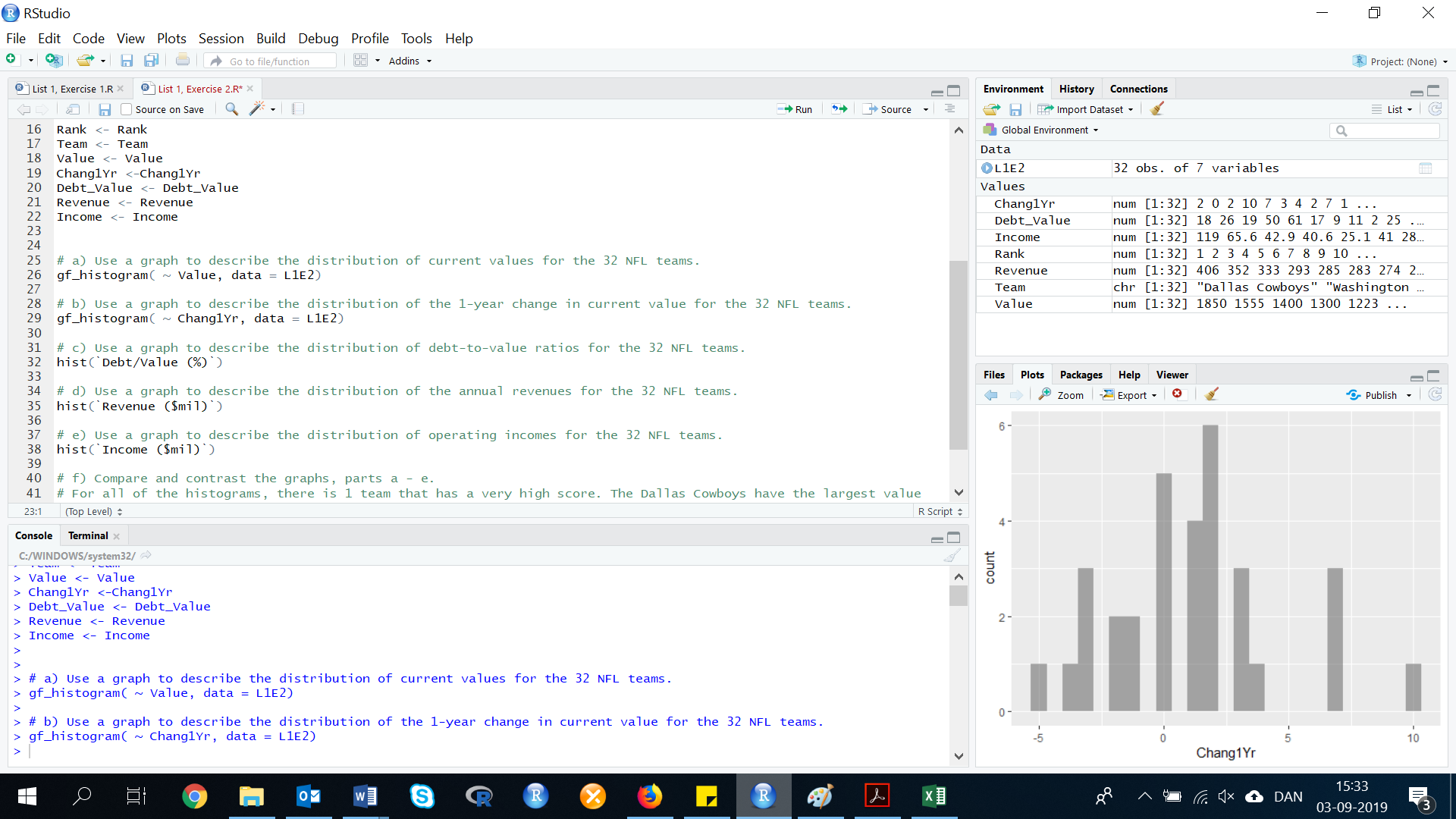
1. **Use a graph to describe the distribution of current values for the 32 NFL teams.**

We use the same code, as we did in List 1, Exercise 1a.

Code: gf\_histogram( ~ Variable, data = DATA)

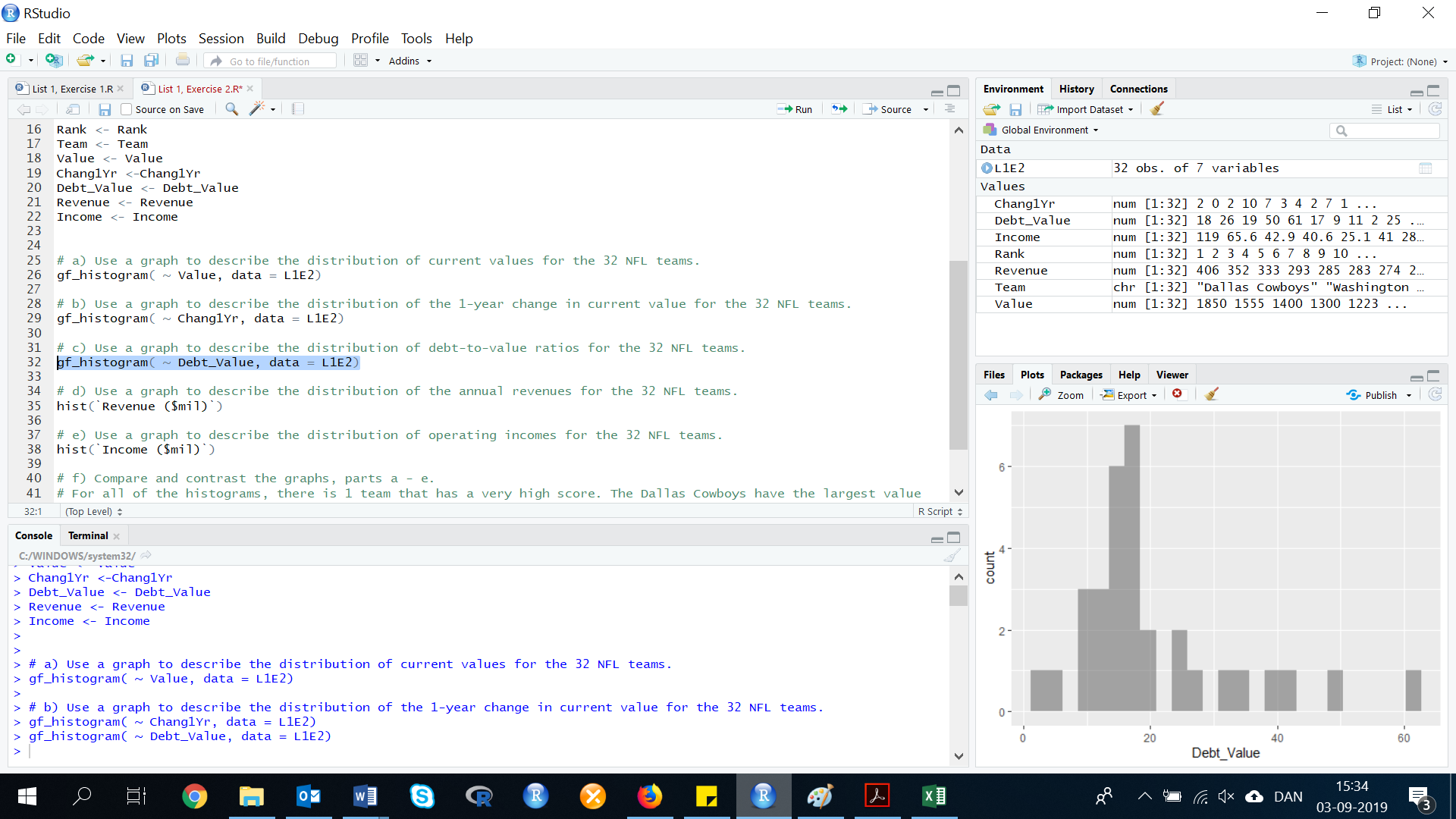
Insert the data and variable:

gf\_histogram( ~ Value, data = L1E2)

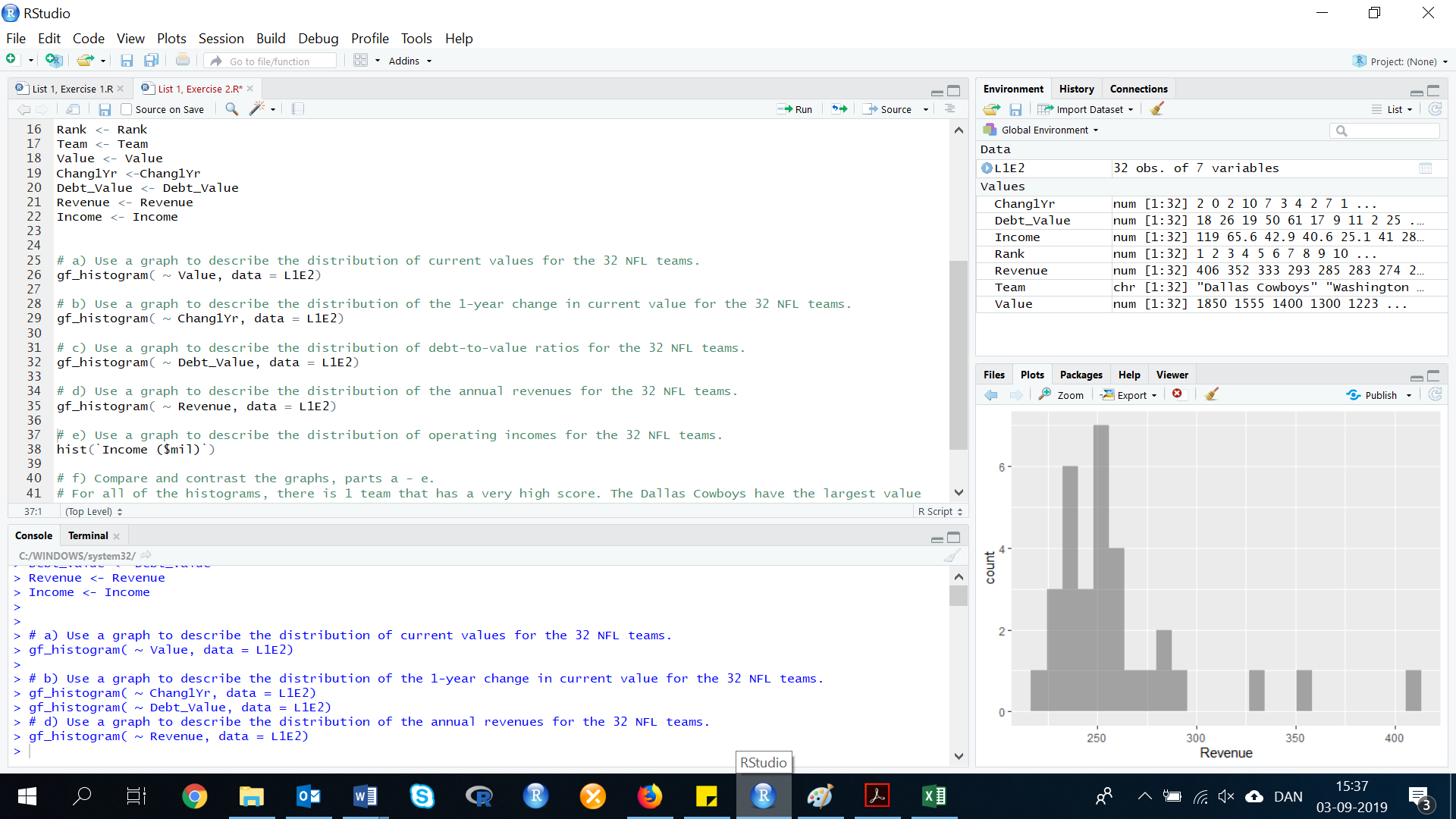


1. **Use a graph to describe the distribution of the 1-year change in current value for the 32 NFL teams.**

gf\_histogram( ~ Chang1Yr, data = L1E2)

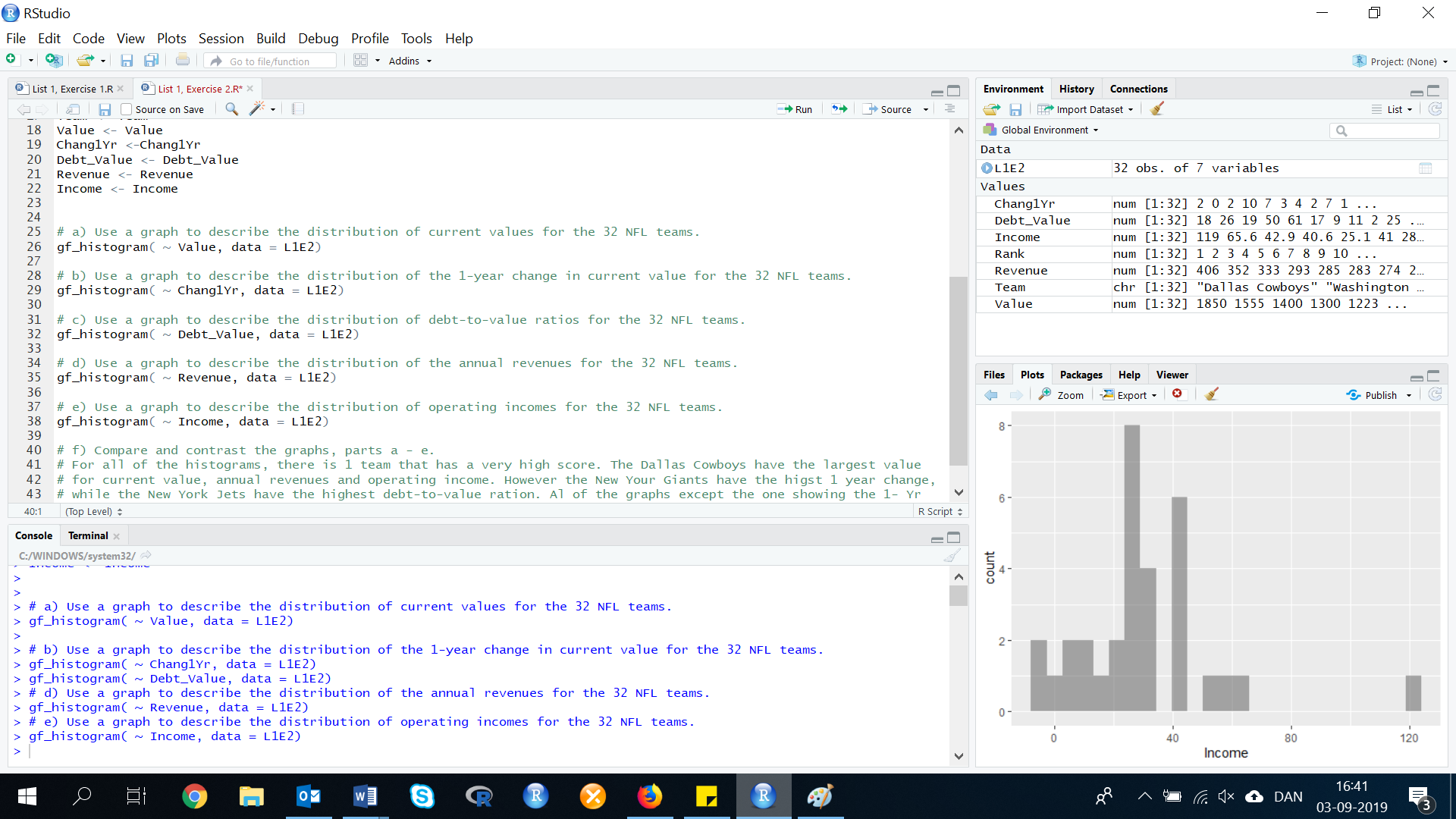
1. **Use a graph to describe the distribution of debt-to-value ratios for the 32 NFL teams.**

gf\_histogram( ~ Debt\_Value, data = L1E2)

1. **Use a graph to describe the distribution of the annual revenues for the 32 NFL teams.**

gf\_histogram( ~ Revenue, data = L1E2)

1. **Use a graph to describe the distribution of operating incomes for the 32 NFL teams.**

gf\_histogram( ~ Income, data = L1E2)

1. **Compare and contrast the graphs, parts a – e.**

For all of the histograms, there is one team that has a very high score. The Dallas Cowboys have the largest value for current value, annual revenues and operating income. However the New Your Giants have the highest 1 year change, while the New York Jets have the highest debt-to-value ration. Al of the graphs except the one showing the 1- Year value changes are skewed to the right.

**Exercise 3. (28, WRKCTR). *Items arriving and depending at work center?* In a manufacturing plant, a work center is a specific production facility that consists of one or more people and/or machines and is treated as one unit for the purpose of capacity requirements for planning and job scheduling. If jobs arrive at a particular work center at a faster rate than they depart, the work center impedes the overall production process and is referred to as a bottleneck. The data in the table were collected by an operations manager for use in investing a potential bottleneck work center. Construct dot plots for two sets of data. Do the dot plots suggest that the work center may be a bottleneck? Explain.**

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 R Script: rm(list=ls())

Library: library(mosaic)

install.packages(“dplyr”)

library(dplyr)

Import data: library(readxl)

L1E3A <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List1/L1E3A.xlsx")

View(L1E3A)

attach(L1E3A)

Create values: NUMBER <- NUMBER

WHEN <- WHEN

In this exercise, you have to make two sets of dotplots/Stripcharts. To do that you have to split the data into two. So you have one dataset with arrivals and one with the departures.

Step 1: Split the dataset into two new dataframes

To split the data into two new datasets, you need to use the R code ‘Filter’ from the dplyr package. As there is to many filter functions/codes in Rstudio, we need to tell it which package to use. We do this by writing:

**dplyr::** in front of the function, as below.

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

The filter() function can be used to generate a new data frame containing just the Arrivals or just the Departures.

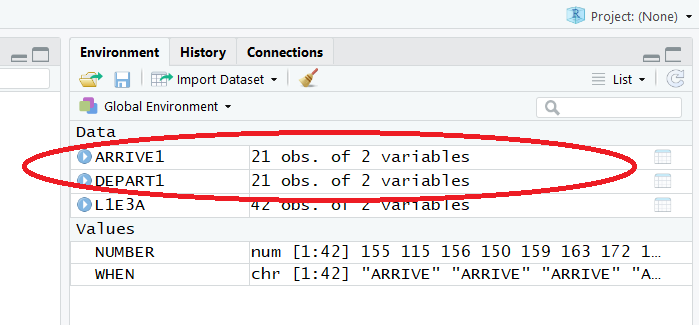
dplyr::filter(L1E3A, WHEN == “ARRIVE”)

Now the R Console gives us the number of arrivals. However, if we do not make this code into a dataset, it is not possible to calculate or make a plot with the new data frame. Therefore, we need to create our new data frame into a new dataset.

ARRIVE1 <- dplyr::filter(L1E3A, WHEN == “ARRIVE”)

We do the same with Departures:

DEPART1 <- dplyr::filter(L1E3A, WHEN == “DEPART”)

We now have two new dataframes:

Step 2: Create dotplots

The code for making a stripchart:

stripchart(VARIABLE)

If you want it to be in a specific color with text and so forth, you can always add extra codes:

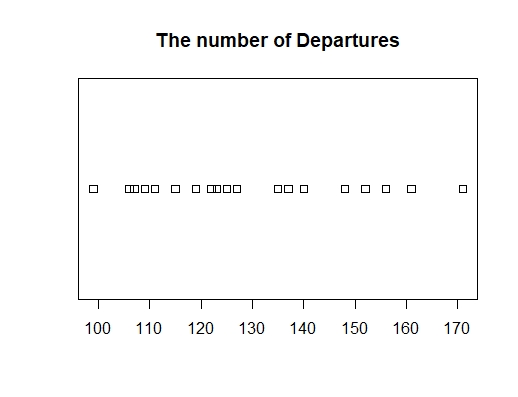
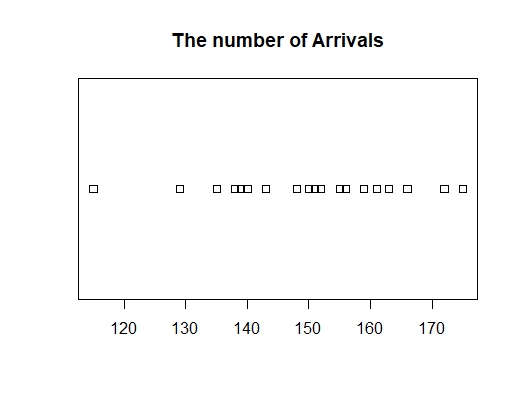
stripchart(VARIABLE, col="black", main="The number of Arrivals and Departures", border="black")

A stripchart produces a one dimensional scatterplot (also called dot plot) of the given data. The stripchart is a good alternative to making a boxplot, when the sample size is small.

Now we need to chose our data and variable. If we just write ‘NUMBER’ as our variable, it will give us the one from L1E3, and not from the new dataframe. Therefore, specify from which data you need the variable ‘NUMBER’, by writing e.g.: ARRIVE1$NUMBER

stripchart(ARRIVE1$NUMBER, col="black", main="The number of Arrivals", border="black")

stripchart(DEPART1$NUMBER, col="black", main="The number of Departures", border="black")



Most of the numbers of items arriving at the work center per hour are in the 135-165 area. Most of the numbers of items departing the work center per hour are in the 110-140 area. Because the number of items arriving is larger than the number of items departing, there will probably be a bottleneck.

**2.3 Numerical Measures of Central Tendency**

**Exercise 4. (47, 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. Import data
3. Create variables
4. Run Library(mosaic)

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

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

Library: library(mosaic)

Import data from Excel: library(readxl)

L1E4 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List1/L1E4.xlsx")

View(L1E4)

attach(L1E4)

Create variables: Honey <- Honey

DM <- DM

Control <- Control

1. **Find the median improvement score for the honey dosage group.**

The median is a measure of central tendency in which the data is sorted so that the numerical value of the median separates the higher half of the data from the lower half.

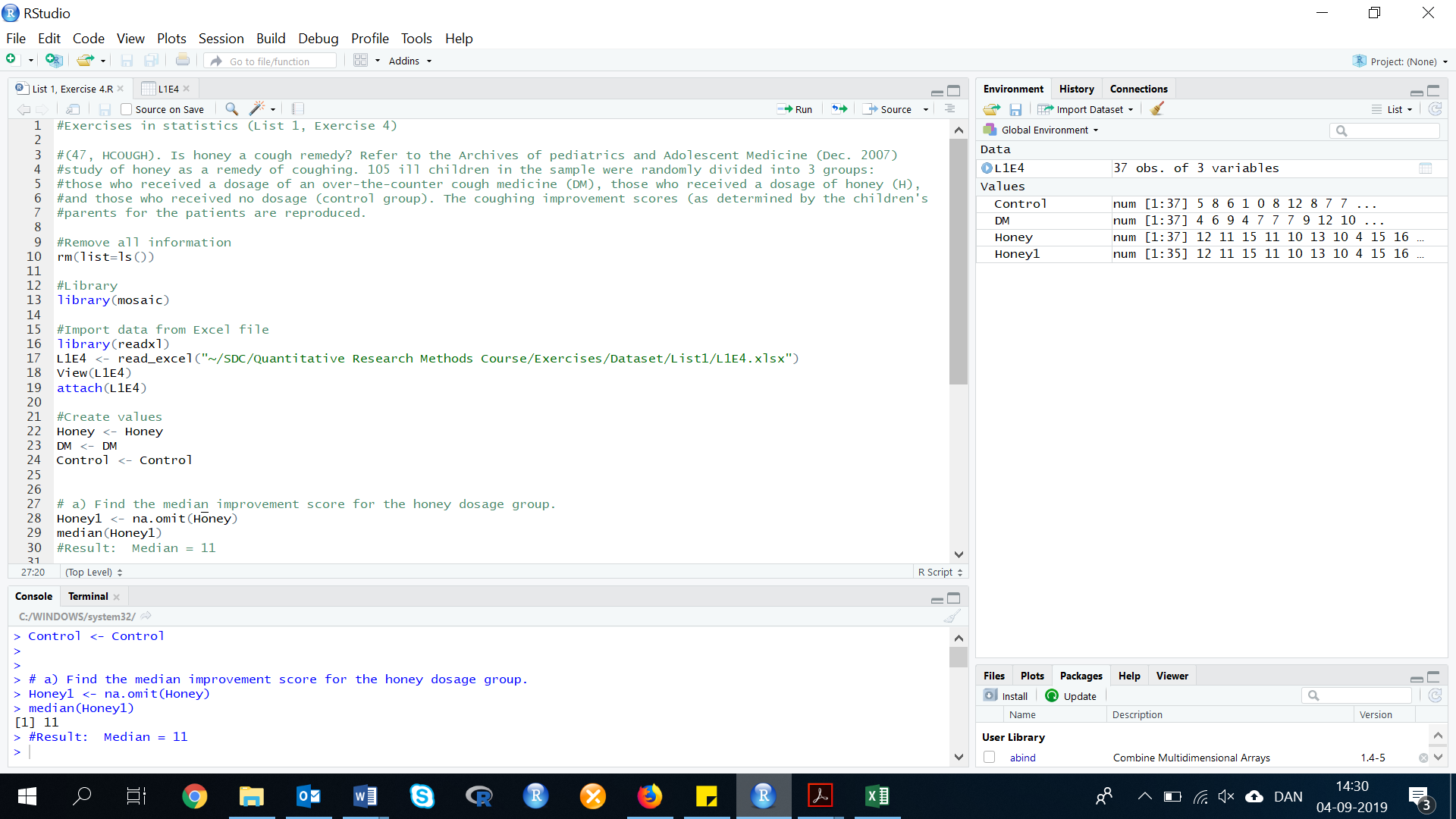
To find the median of Honey, there is two ways to do it, but before calculating Honey, you should look at the data.

If you look at the data you have imported from Excel, it says that Honey have 37 observations. However, if you count the cells for Honey, you will see that only 35 of the cells contain a value, while the others are empty or saying NA. This means that no value is available in these cells and should be removed before you calculate the median. Therefore, the first step in calculating the median for Honey, would be to remove all the cells with missing values.

To do this, you will have to use the following code:

na.omit(Variable)

na.omit(Honey)

Now we have removed the value from Honey (but it is not removed from L1E4), so we need to make our ‘na.omit(Honey)’ into a variable. Otherwise, we cant work with the variable Honey without the missing values.

We do this by creating a new ‘Honey’ Variable (We call it Honey1):

Honey1 <- na.omit(Honey)

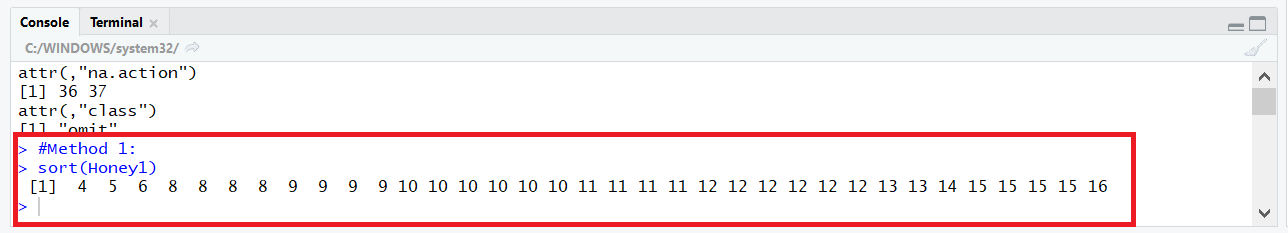
Method 1: Manual sorting the numerical values from smallest to largest.

With this method, you to have to count the numerical values. To see the values of ‘Honey1’ without the missing values, we need to sort the variable.

Therefore, use the code ‘sort()’:

sort(VARIABLE)

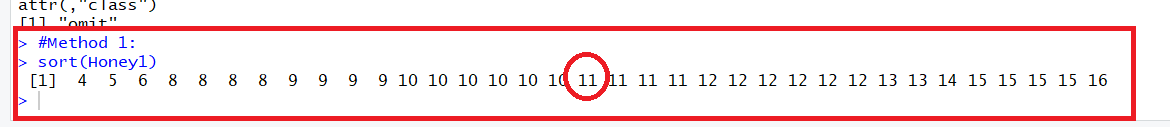
sort(Honey1)



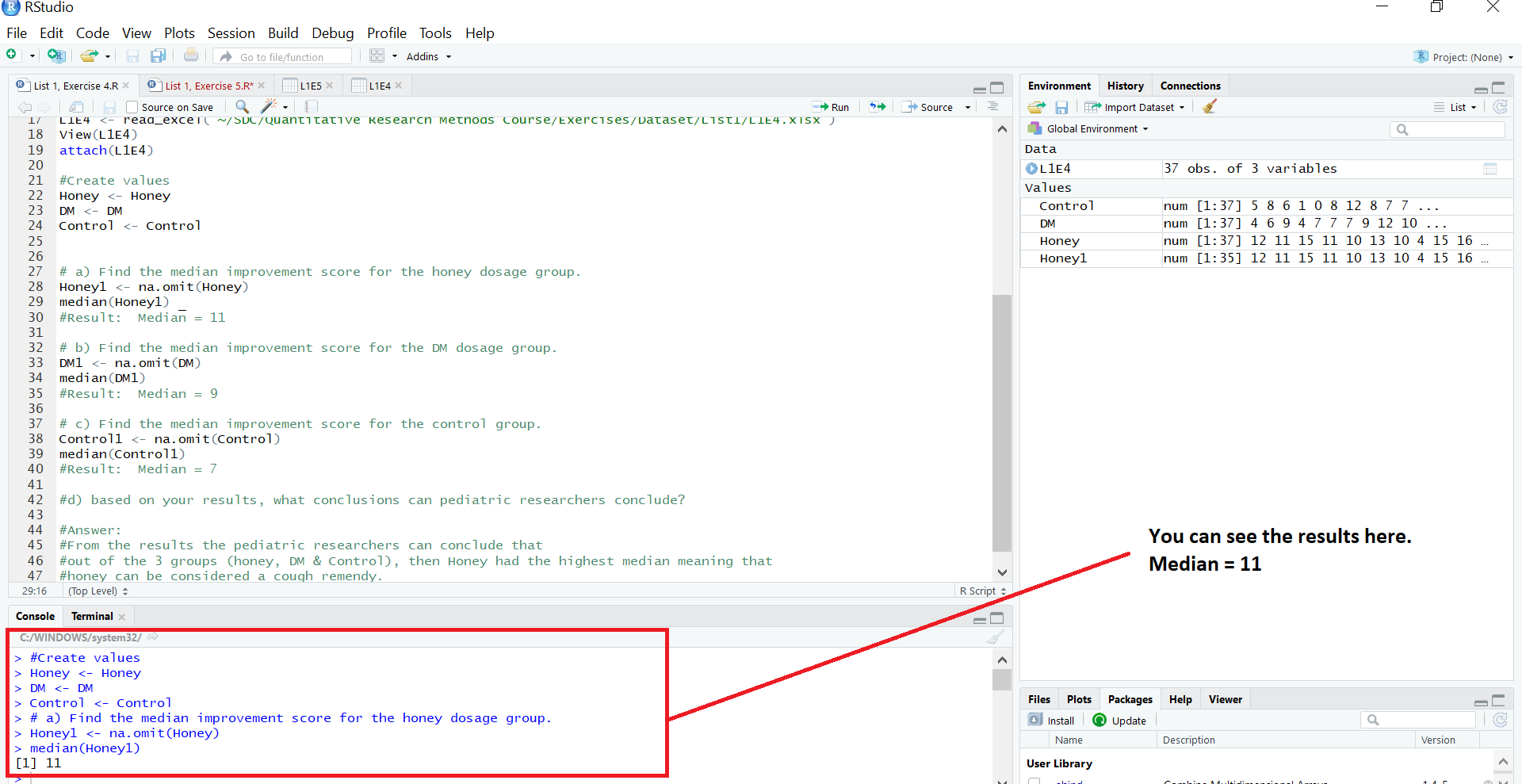
The position of the median can be found by:

* If n is odd it’s the middle value of the sequence.
* If n is even, average of 2 middle values.

There is 35 values in the variable Honey1, then the middle number is 18, and the median is 11.



Method 2: Calculate the median in R

To calculate the median, we need the following code:

median(VARIABLE)

Now insert the variable, and it will calculate the mean to you.

median(Honey1)

See the picture below, to understand where R Studio are giving you the result of your calculation. It is giving you the result in the so called ‘Console’.

Result: median = 11

1. **Find the median improvement score for the DM dosage group.**

You now know the two types of methods. In the following exercises, we will only focus on Method 2, where we use R to calculate the median.

As with Honey, we will now do the same thing to the variable ‘DM’.

DM1 <- na.omit(DM)

median(DM1)

Result: Median = 9

1. **Find the median improvement score for the control group.**

Control1 <- na.omit(Control)

median(Control1)

Result: Median = 7

1. **Based on your results, what conclusions can pediatric researches draw?**

From the results the pediatric researchers can conclude that out of the three groups (Honey, DM & Control), Honey had the highest median meaning that Honey can be considered a cough remedy.

**Exercise 5. (48, BIODEG). Refer to the Journal of Petroleum Geology (April 2010) study of the environmental factors associated with biodegradation in crude oil reservoirs. The amount of dioxide (milligrams / liter) and the presence / absence of crude oil was determined for each of 16 water specimens collected from mine reservoir. The data are repeated in the accompanying table:**



First, we need to do as followed:

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

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

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

Library: library(mosaic)

Library(dplyr)

Import data from Excel: library(readxl)

L1E5 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List1/L1E5.xlsx")

View(L1E5)

attach(L1E5)

Create values: Dioxide <- Dioxide

Oil <- Oil

1. **Find the mean dioxide level of the 16 water specimens. Interpret this value.**

To calculate the mean of the dioxide level, you need to use the following code:

mean(VARIABLE)

If you insert the variable, that you need to calculate the mean of, then R studio will give you the result in console (just as with the median in List 1, Exercise 4).

mean(Dioxide)

Result: Mean dioxide level is 1.8125

1. **Find the median dioxide level of the 16 water specimens. Interpret this value.**

median(Dioxide)

Result: Median dioxide level is 1.35

1. **Find the mode dioxide level of the 16 water specimens. Interpret this value.**

To calculate the mode of the dioxide level you need to use a code for mode. In R there is not specified a code to calculate mode. However, we recommend that you use the following code:

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

mode <- getmode(VARIABLE)

print(mode)

Insert the variable, and R will calculate the mode for the dioxide level:

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

mode <- getmode(Dioxide)

print(mode)

Result: Mode = 4

1. **Find the median dioxide level of the 10 water specimens with no crude oil present.**

As in List 1, Exercise 2, we need to filter our variable:

dplyr::filter(Data, Variable ==”VALUE”)

Remember to make the filter into a variable. When the data are filtered and a new variable is created, we can calculate the mean:

mean(DATA$Variable)

Insert the data and variables:

Dioxide10 <- dplyr::filter(L1E5, Oil == ”No”)

median(Dioxide10$Dioxide)

Result: Median = 2.85

The median level of the 10 water specimens with no crude oil are 2.85.

1. **Find the median dioxide level of the 6 water specimens with crude oil present.**

In this exercise, the calculation/coding is quite similar to exercise d.

First, we need to filter our data and create a new data frame. Then we can calculate the median dioxide level of the 6 water specimens with crude oil.

Dioxide6 <- dplyr::filter(L1E5, Oil == “Yes”)

median(Dioxide6$Dioxide)

Result: Median = 0.45

The median level of the 6 water specimens with crude oil are 0.45.

1. **Compare results d and e. Make a statement about the association between dioxide level and presence / absence of crude oil.**

The results show that, when there is a higher level of dioxide in the water, there is no crude oil present.

**Exercise 6. (50, PGA). *Ranking driving performance of professional golfers*. A group of Northeastern University researchers developed a new method for ranking the total driving performance of golfers on the professional Golf Association (PGA) tour (The Sport Journal, Winter 2007.)The method requires knowing a golfer’s average driving distance (yards) and driving accuracy (percent of drives that land in the fairway). The values of these two variables are used to compute a driving performance index. Data for the top 40 PGA golfers (as ranked by the new method) are saved in the accompanying file. The first five and last five observations are listed in the table below.**



First, we need to do as followed:

1. Clean our script
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)

L1E6 <- read\_excel("~/SDC/Quantitative Research Methods Course/Exercises/Dataset/List1/L1E6.xlsx")

View(L1E6)

attach(L1E6)

Create values: ACCURACY <- ACCURACY

DISTANCE <- DISTANCE

INDEX <- INDEX

PLAYER <- PLAYER

1. **Find the mean, median and mode for the 40 driving performance index values.**

In Exercise 5, we have calculated the mean, median, and mode. Therefore, we can use the same code just with new data and variables.

mean(INDEX)

median(INDEX)

getmode <- function(v) {

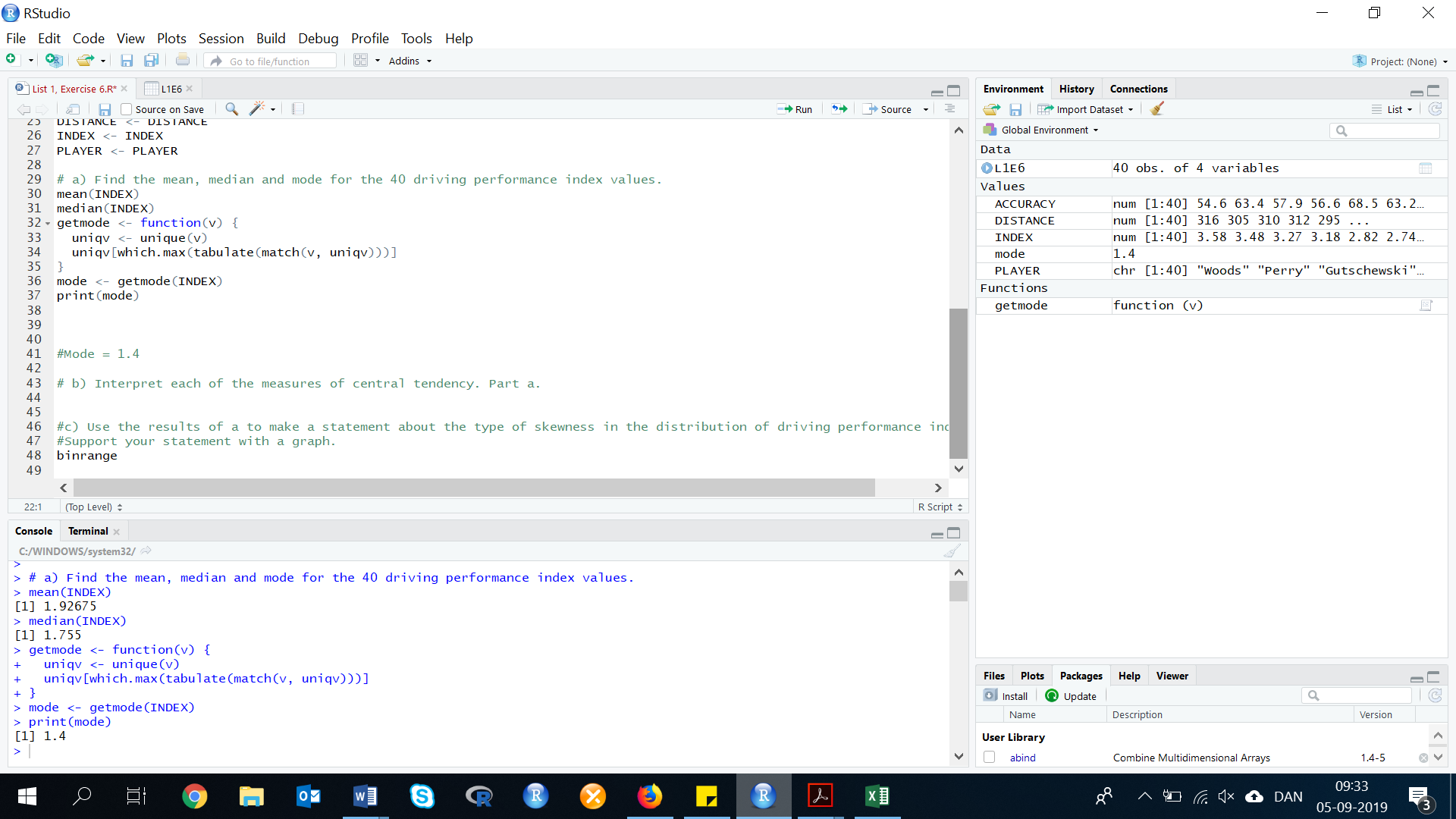
uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

mode <- getmode(INDEX)

print(mode)



Results: Mean = 1.92675, Median = 1.755, and Mode = 1.4

NOTE:

The mean is the average.

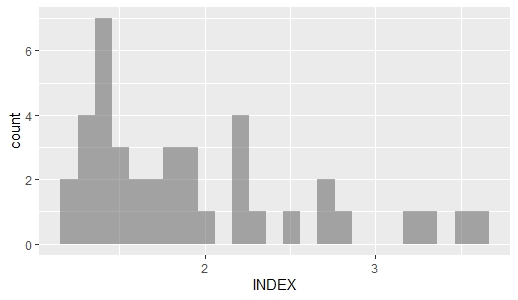
The median is found as the 20th and 21st observations, once the data have been ordered. The 20th and 21st observation are 1.75 and 1.76.

The mode is the number that occurs the most in the variable.

1. **Interpret each of the measures of central tendency. Part a.**

The sample average driving performance index is 1.927. The median driving performance index is 1.755. Half of all driving performance indexes are less than 1.755 and half are higher. The most common driving performance index value is 1.4.

1. **Use the results (found in exercise A) to make a statement about the type of skewness in the distribution of driving performance indexes. Support your statement with a graph.**

In this exercise, you have to make a histogram of the variable index.

The code for a histogram is:

gf\_histogram( ~ VARIABLE, data = DATA)

Insert the Variable and data, and it should look like this:

gf\_histogram( ~ INDEX, data = L1E6)

The histogram should look like this histogram (Look at the picture to the right).

As the mean (1.927) is higher than the median (1.755), the data is skewed to the right, which also can be seen in the histogram.