

Applied Missing data analysis with SPSS and R(Studio)

Martijn Heymans and Iris Eekhout

2018-09-04

Contents

Preface	5
0.1 The goal of this Manual	5
0.2 Multiple Imputation in SPSS and R	6
0.3 Notation and annotation in this manual	6
1 Software applications	7
1.1 SPSS, Data and Variable View windows	7
1.2 Analyzing data in SPSS	10
1.3 Data Transformations in SPSS	12
1.4 The Output window in SPSS	12
1.5 The Syntax Editor in SPSS	12
1.6 Reading and saving data in SPSS	13
1.7 R and RStudio	13

Preface

The attention for missing data is growing and so will be the application of methods to solve the missing data problem. From our experience, researchers with missing data still find it difficult to reserve time to evaluate the missing data and from that to find a reasonable solution to handle their missing data for their main data analysis. This manual is developed for researchers that are looking for a solution of their missing data problem or want to learn more about missing data. The manual is developed as a result of a missing data course that we give. Further, we are also active in providing statistical advice in general and more specific about missing data. Because our time to give advice is mostly limited we wanted to give researchers a practical guide to help them get started with their missing data problem. Leading methodologists and statisticians and leading journals have published papers about the problems of missing data and warned researchers to take missing data seriously (Sterne et al., BMJ 2009, Little et al. NEJM 2012, Peng et al. 2015, JAMA). Hopefully this manual will help researchers to find the best solution for their missing data problem. We hope you will enjoy this manual and that you learn from it, at least to take missing data seriously and that you will use recommended methods to solve your missing data problem.

0.1 The goal of this Manual

In this manual the software packages SPSS and R play a central role. The combination of these two software packages may seem a coincidence, but it is not. For a long time, SPSS was the most popular software package worldwide to do statistical data analysis. Currently, R is growing in popularity fast and will probably become one of the most popular Software packages to do data analysis. Also for applied researchers. Both SPSS and R have their advantages and disadvantages. An advantage of SPSS is that it is a user-friendly software package compared to R and works with windows where you can for example drag your variables to. Subsequently, you can click the OK button and the statistical analysis procedure you prespecified gives you the output results. A disadvantage of SPSS may be that you are overloaded with statistical output that may not all needed to answer your research question. Compared to SPSS you could say that R is a more user-unfriendly software package where you need to use R code to activate statistical procedures and to get statistical results. R output will show more specific results, without extra information. Furthermore, R works much faster when it comes to running statistical procedures by using 1 or 2 lines of R code, compared to visiting a couple of windows in SPSS to activate the same statistical test. There is one other advantage of R and that is, that it is open source. This makes it possible for applied researchers to follow the calculations of complex procedures as the estimation of missing values closely along the line. You could say that R brings you to the heart of the matter. With R it is possible to turn complex data analysis functions and formula's into computer code that can be used by everybody and vice versa. Because it is open source, you are able to read the code that is used for the analysis and to relate that code or pieces of code to the statistical output. This makes it possible to evaluate step by step the code and thus the statistical procedures and relate them to the subsequent results. You can copy specific parts of code from functions that others have written and evaluate what happens. This is one of the major advantages of R if you compare it to the closed source statistical package SPSS. R brings you a big learning environment when it comes to the understanding of all kind of statistical procedures as missing data analysis.

0.2 Multiple Imputation in SPSS and R

Multiple Imputation (MI) is a procedure that is developed in the 1970's by Donald Rubin. Later, around the 1990's Multiple imputation was further developed and became more popular. For a long time, MI was only available for S-Plus and R software (S-plus is the commercial alternative of R), where it was further developed by Stef van Buuren, a statistician from TNO, Leiden, The Netherlands. For a long time, it was not possible to do MI analysis in SPSS because it was not available in SPSS. So, it was far out of reach for applied researchers for a long time. It became available from SPSS version 17. From that time MI is now used more by applied researchers. In this manual the handling of missing data is the main topic. We will also show how to apply these methods in both software packages SPSS and R. To apply the imputation methods that are discussed both software packages make use of random starting procedures. SPSS and R use for that intern random number generators. Because these are different, result might slightly differ. Our intention is not to compare the software packages SPSS and R and their output resultys. Both are trustful packages, it is more the estimation procedures that might lead to the differences. The imputation methods, will be applied in SPSS version 24 and with R software version 3.4.3. The R examples will be presented by using the output from RStudio version (version 1.1.383 – © 2009-2017 RStudio, Inc.). RStudio is an integrated development environment (IDE) for R. RStudio includes a wide range of productivity enhancing features and runs on all major platforms. As already stated, R allows you to program the statistical formula's yourself. We have therefore chosen to explain the formula's in more detail in combination with the application in R. The more applied researchers will be satisfied with the explanation and application of methods in SPSS.

0.3 Notation and annotation in this manual

The name of R packages, libraries and functions can be recognized by using Courier new lettertype, for example the package mice will be written as mice.

R code of the procedures used in the manual is marked grey and the explanation in these grey parts can be found in the grey parts itself annotated by the # symbol. The lines that start with the symbol > are R Code lines that have been running in the R Console in RStudio. Example:

R code XX

```
# Activate the foreign package and read in the SPSS dataset

library(foreign)
dataset <- read.spss(file="Backpain 50 missing.sav", to.data.frame=T)
```

```
## re-encoding from UTF-8
```

Chapter 1

Software applications

Statistical software programs can help us to analyze our data. SPSS and R are such programs. Although SPSS and R are among the most popular programs to do statistical data analyses nowadays, they do not have much in common. One of the greatest differences is that SPSS works with menu options that make windows appear and you can click buttons to select options, whereas R works with lines of code that you have to type in to run analyses. This makes SPSS more user-friendly than R for applied researchers. In SPSS you are overloaded with output tables, and in R you only get output on demand. In this Chapter we will explore the different possibilities of the SPSS (IBM 2016) and the R software language (Matloff, 2011, Dalgaard, 2008). We will run R via RStudio, the integrated development environment (IDE) for R. RStudio includes a wide range of productivity enhancing features, which makes it easier to work with than with the R console on its own.

1.1 SPSS, Data and Variable View windows

In this manual we work with SPSS version 24 (IBM, 2016). When you start SPSS Version 24 a start-up window appears. In this window, you can directly open the files that were active during your previous use of SPSS. These files can be found and easily opened in the “Recent files” window (Figure 1.1). If you do not want to see this window the next time that you open SPSS, select “Don’t show this dialog in the future”.

When you click on Close on the right side below, the window will close and you will see an empty Data View window. Now you are in the SPSS Data Editor window. This window is always open when you start SPSS. The name “SPSS Data Editor” is also visible at the top of the screen and is called “IBM SPSS Statistics Data Editor” (Figure 1.2).

In the SPSS Data Editor, you have the possibility to go to the Data View and Variable View windows. In the Data View window, you can enter data yourself or read in data by using the options in the file menu. In Figure 1.2 you see an example of a dataset in the Data View window. Each row in the Data View window represents a case and in the columns you will find the variable names. In the Data View window, you can do all kind of data manipulations by using the different menu’s above in the window. From here you can click on the tab Variable View, in the lower left corner of the window. Then the Variable view window will appear (Figure 1.3).

In the Variable View window, you can add new variables, by entering the name in the name column. Further, you can change the columns by using the following options: Type: Here you can change the type of variables in your dataset. Mostly you work with numeric variables, i.e. a variable whose values are numbers. Other possibilities are Date variables which is a numeric variable whose values are displayed in one of several calendar-date or clock-time formats or String variables, a character (text) variable that can contain any characters up to the defined length. String values are not numeric and therefore are not used in calculations. Width: By default SPSS defines a numeric variable with 8 digits for each new variable.

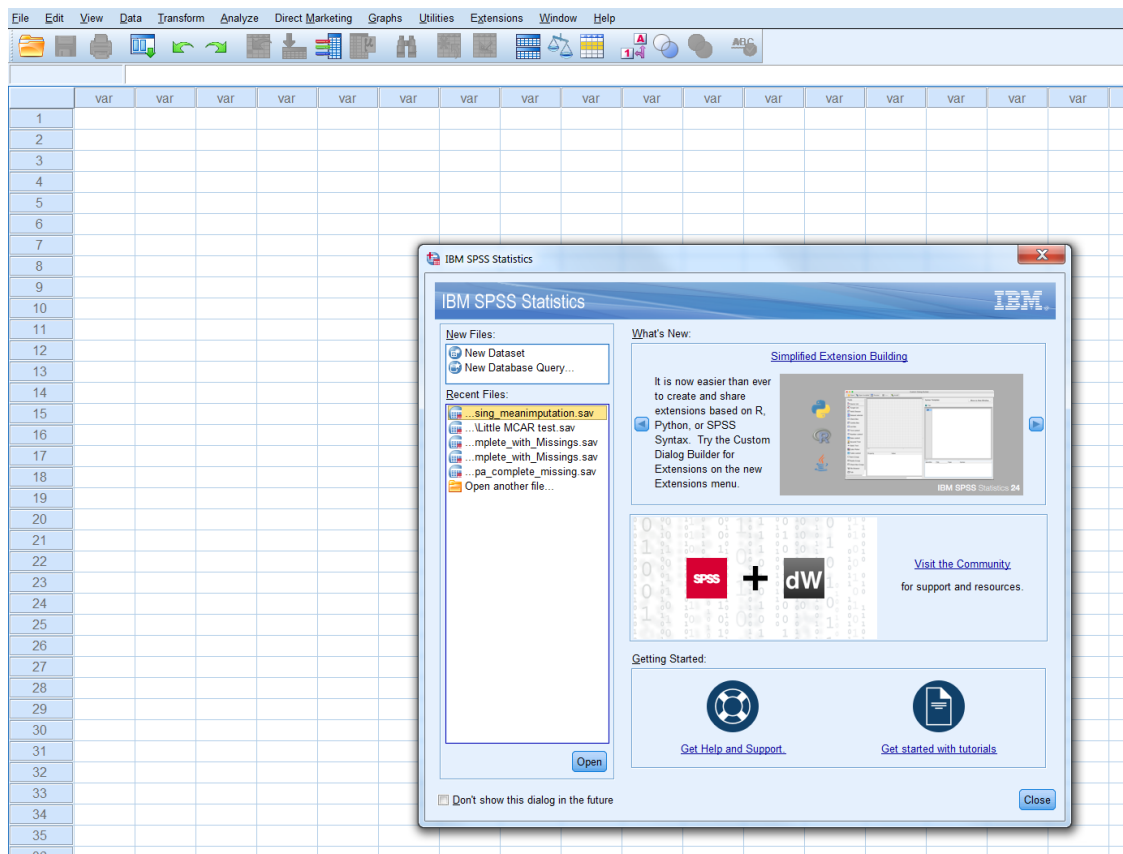
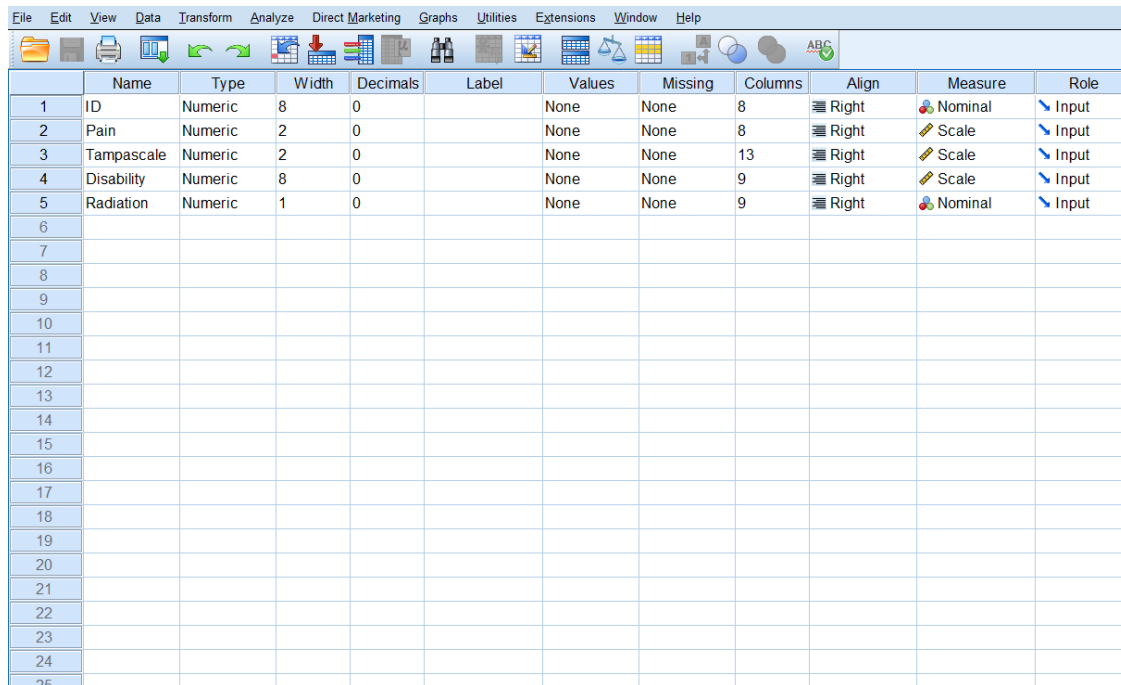


Figure 1.1: First window after you have started SPSS

	ID	Pain	Tampascale	Disability	Radiation	var
1	1	9	45	20	1	
2	2	6	.	10	0	
3	3	1	36	1	0	
4	4	5	38	14	0	
5	5	6	44	14	1	
6	6	7	.	11	1	
7	7	8	43	18	0	
8	8	6	43	11	1	
9	9	2	.	11	1	
10	10	4	36	3	0	
11	11	5	38	16	1	
12	12	9	47	14	0	
13	13	0	32	3	1	
14	14	6	.	12	0	
15	15	3	34	13	0	
16	16	6	42	8	1	
17	17	3	35	11	0	
18	18	1	31	1	0	
19	19	2	31	7	0	
20	20	4	32	9	1	
21	21	5	.	13	0	
22	22	5	39	12	0	
23	23	4	34	8	1	
24	24	8	47	13	1	
25	25	5	.	6	0	

Figure 1.2: Data View window in SPSS



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	Numeric	8	0		None	None	8	Right	Nominal	Input
2	Pain	Numeric	2	0		None	None	8	Right	Scale	Input
3	Tampascale	Numeric	2	0		None	None	13	Right	Scale	Input
4	Disability	Numeric	8	0		None	None	9	Right	Scale	Input
5	Radiation	Numeric	1	0		None	None	9	Right	Nominal	Input
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											

Figure 1.3: Variable View window in SPSS

Decimals: the number of decimal places displayed.

Label: The variable name.

Values: To assign numbers to the categories of a variable. To define Variable values do the following: 1. Click the button in the Values cell for the variable that you want to define. 2. For each value, enter the value and a label. 3. Click Add to enter the value label. 4. Click OK.

Missing: Here you can define specified data values as user-missing. You can enter up to three discrete (individual) missing values, a range of missing values, or a range plus one discrete value.

Columns: To change the number of characters displayed in the Data View window.

Align: Here you can specify the alignment of your data.

Measure: Here you can specify the level of each variable, scale (continuous), ordinal or nominal.

Role: Here you can define the role of the variable during your analysis. Examples are, Input for independent variable, Target for dependent or outcome variable, Both, independent and dependent variable. There are more possibilities, but most of the times you use the default Input setting.

1.2 Analyzing data in SPSS

All statistical procedures in SPSS can be found under the Analyze button (Figure 1.4). Here you also will find the option “Multiple Imputation” which plays an important role in this manual. We will use this menu later on in Chapter 4.

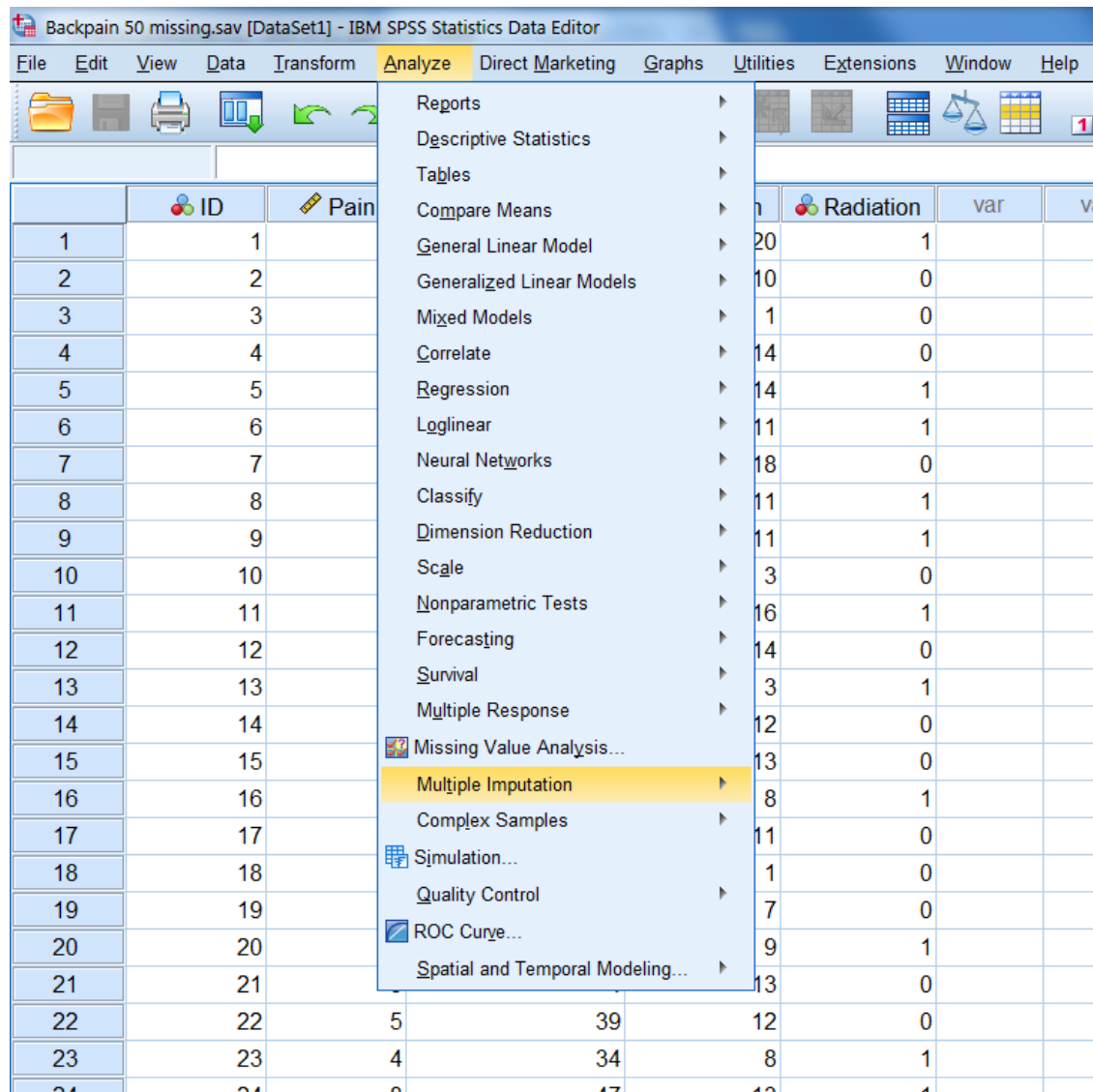


Figure 1.4: Statistical procedures that can be found under the Analyze menu in SPSS

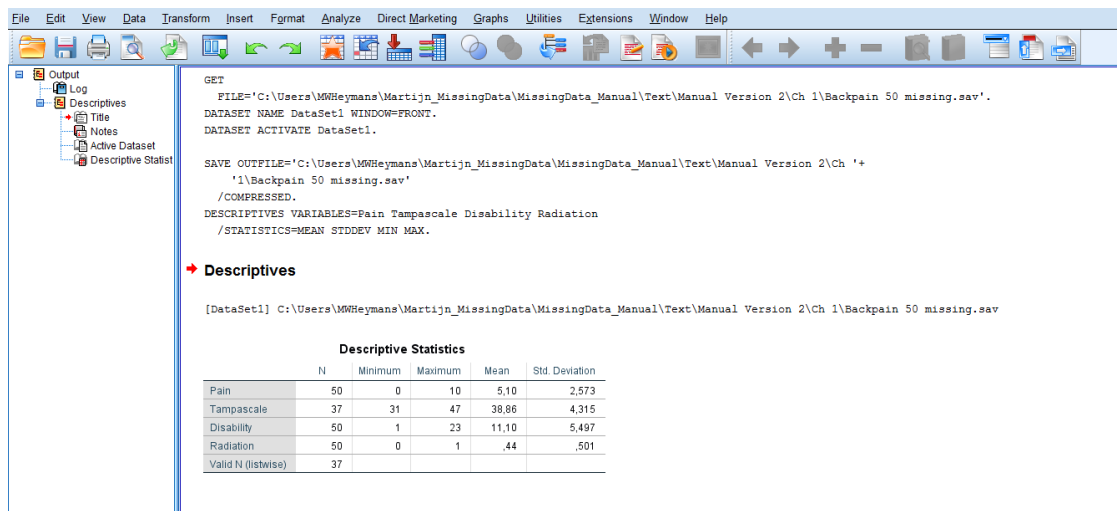


Figure 1.5: Part of the Output or Viewer window in SPSS after making use of Descriptive Statistics under the Analyze menu

1.3 Data Transformations in SPSS

Two other interesting buttons are Data and Transform. The Data menu allows you to make changes to the data editor. Here you can add new variables or cases. You can also use the Split File option, to get analyses results separately for categories of a variable. The Transform menu allows you to manipulate your variables by for example dichotomizing a numeric variable.

1.4 The Output window in SPSS

If you have run your analyses in SPSS, an SPSS Output (or viewer) Window will pop-up. The main body of the Output Window consists of two panes (left and right panes). In the left pane you will find an outline of the output. In the right pane you will find the actual output of your statistical procedure (Figure 1.5).

1.5 The Syntax Editor in SPSS

In the syntax editor of SPSS, you use the SPSS syntax programming language. You can run all SPSS procedures by typing in commands in this syntax editor window, instead of using the graphical user interface, i.e. by using your mouse and clicking on the menu's. You can get access to the syntax window in two ways. The first is just by opening a new syntax file by navigating to File -> New -> Syntax. This will open a new syntax window (Figure 1.6).

Now you can start writing your syntax directly in this window. You can also generate syntax by accessing statistical procedures through the dropdown menus and clicking the Paste button instead of clicking the OK button after you have specified the options. When you have clicked the Paste button, a new Syntax Editor window will pop up or the new syntax will automatically be added to the open Syntax Editor window. This is a very useful way to keep track of the analysis that you have performed. An example can be found in Figure 1.7, where the syntax is shown for the Descriptive Statistics procedure of Figure 1.5.

By using the SPSS Syntax it is possible for users to perform the same analyses over and over again or to adapt the analysis via the syntax code for complex calculations in the data. In this manual we will not use SPSS syntax code to access statistical procedures, however we recommend to use the SPSS syntax to keep

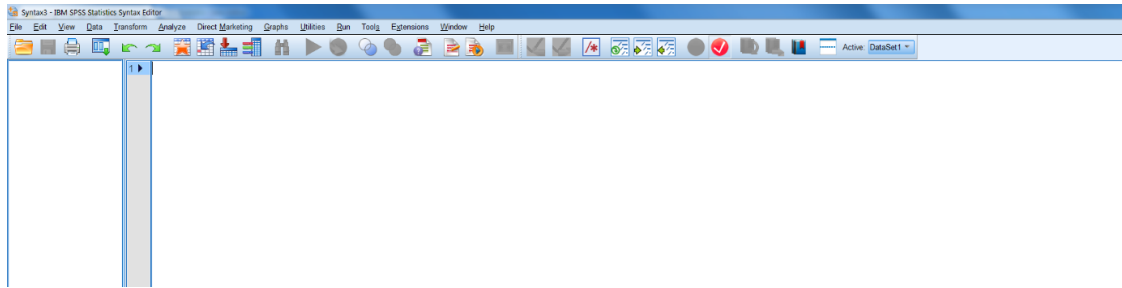


Figure 1.6: Screenshot of new syntax file

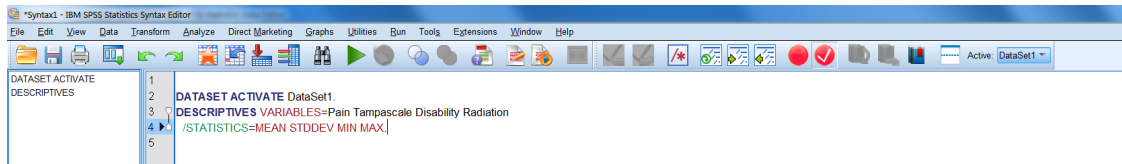


Figure 1.7: Screenshot of Syntax editor of SPSS including the Syntax code for descriptive statistics

track of the analysis that you have performed. SPSS is most frequently used via the graphical user interface, and we will use that method also in this manual.

1.6 Reading and saving data in SPSS

Reading in data in SPSS is very easy; via the menu File choose for File -> Open -> Data. All kind of file types can be selected. Of course the SPSS .sav files, but also .por, .xlsx, .csv, SAS, Stata, etc. (Figure 1.8). After you have selected a specific file type you may have to go through several steps before you see the data in the Data View window. These steps are not necessary for SPSS files, they open directly in the data editor.

Saving files in SPSS is possible via the Save Data As option under the menu File. You can choose the same kind of file types (Figure 1.9).

1.7 R and RStudio

RStudio is an integrated environment to work with the software program R. Consequently, to work with RStudio, R has to be installed. RStudio uses the R language and is also freely available. In this manual we will only show some possibilities and options in RStudio that are needed to run the R code and the programs that are discussed in this manual. For more information about RStudio and its possibilities visit the RStudio website at www.rstudio.com. When you open RStudio the following screen will appear.

There are three windows opened:

1. On the left is the Console window

This is the main window to run R code (see below for more information about the Console window).

2. Right above is the window where you can choose between the Environment and History tabs (e.g. history tracks the code you typed in the Console window).
3. At the right side below is the window where you can choose between Files, Plots, Packages, Help and Viewer tabs.

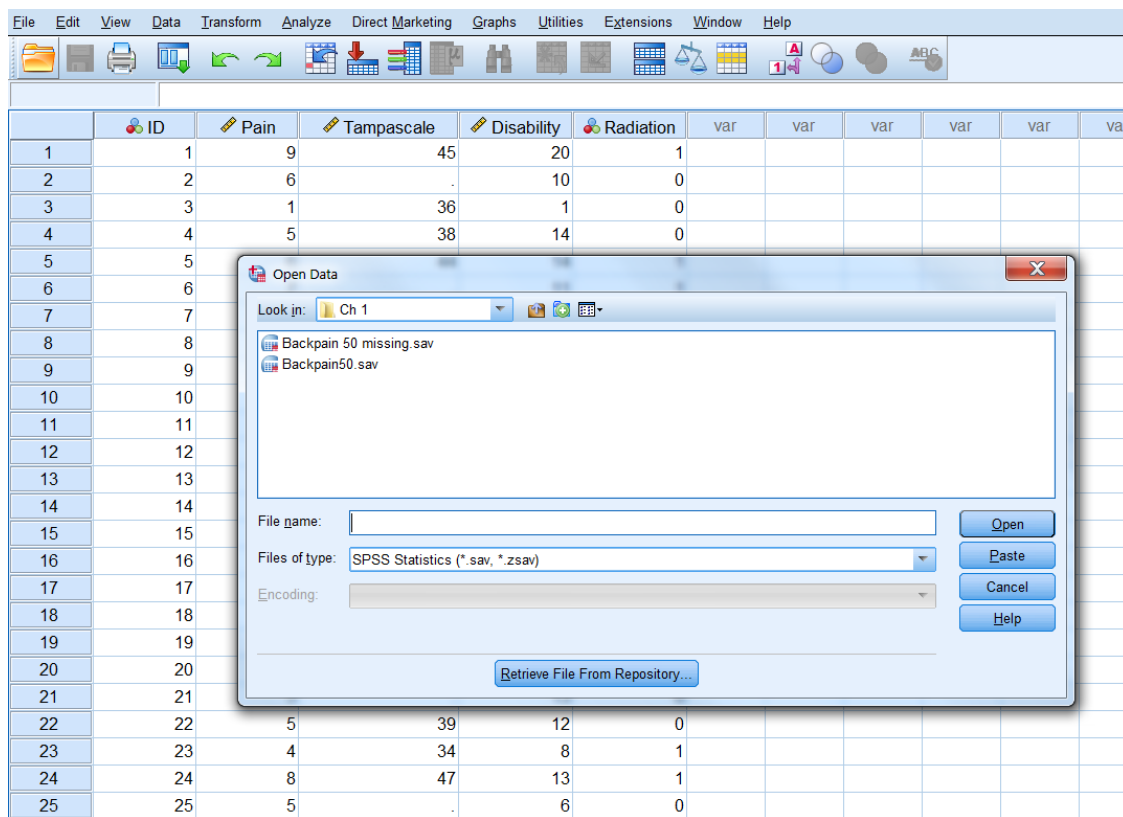


Figure 1.8: Window to read in different file types in SPSS

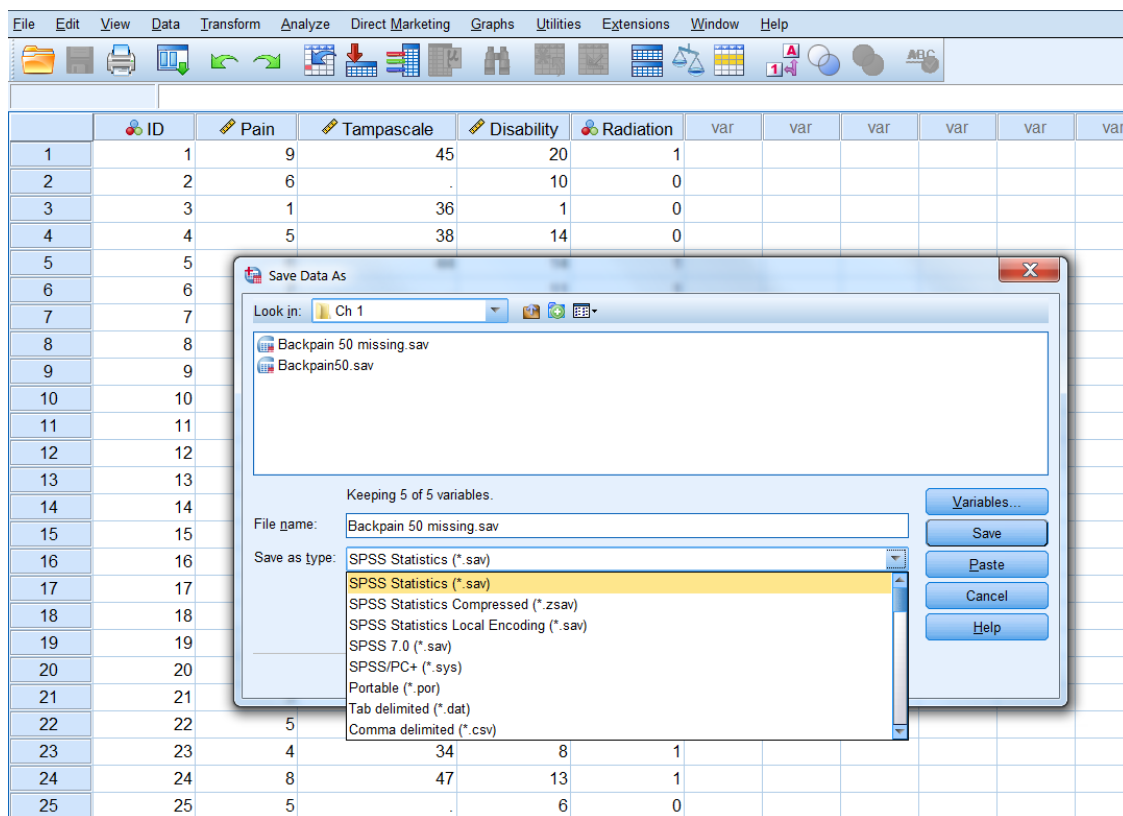


Figure 1.9: Option Save Data As under the menu File

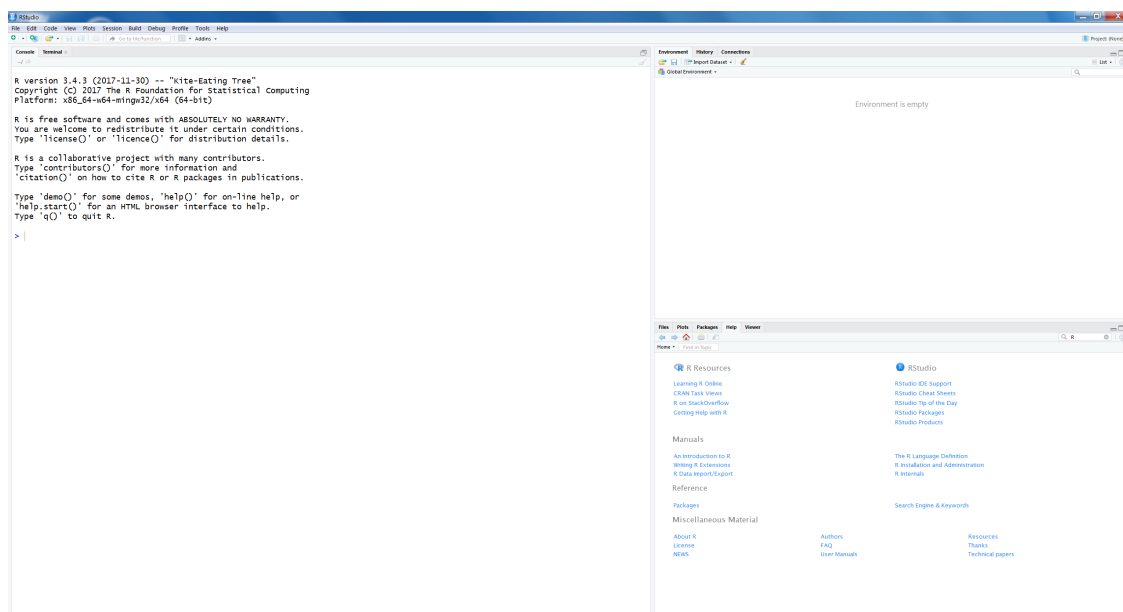


Figure 1.10: First screen that appears after you have started RStudio

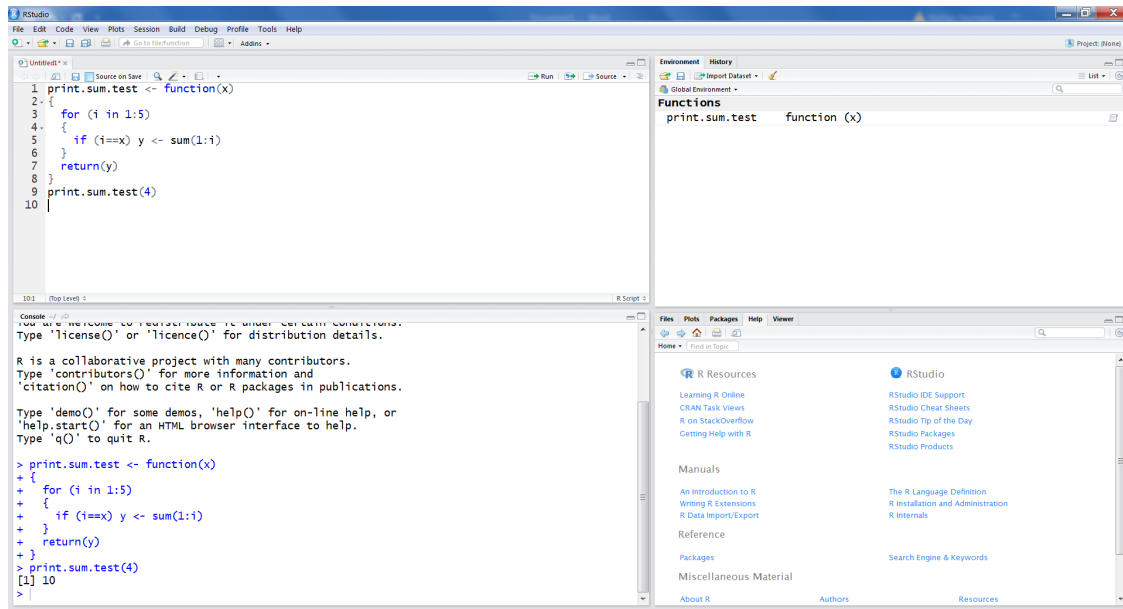


Figure 1.11: Script file example in RStudio

When you enter code in the Console window you will directly receive a result. For example, you can type the following code and the result will appear directly in the Console window.

R code 1.1

```
3 + 3
```

```
## [1] 6
```

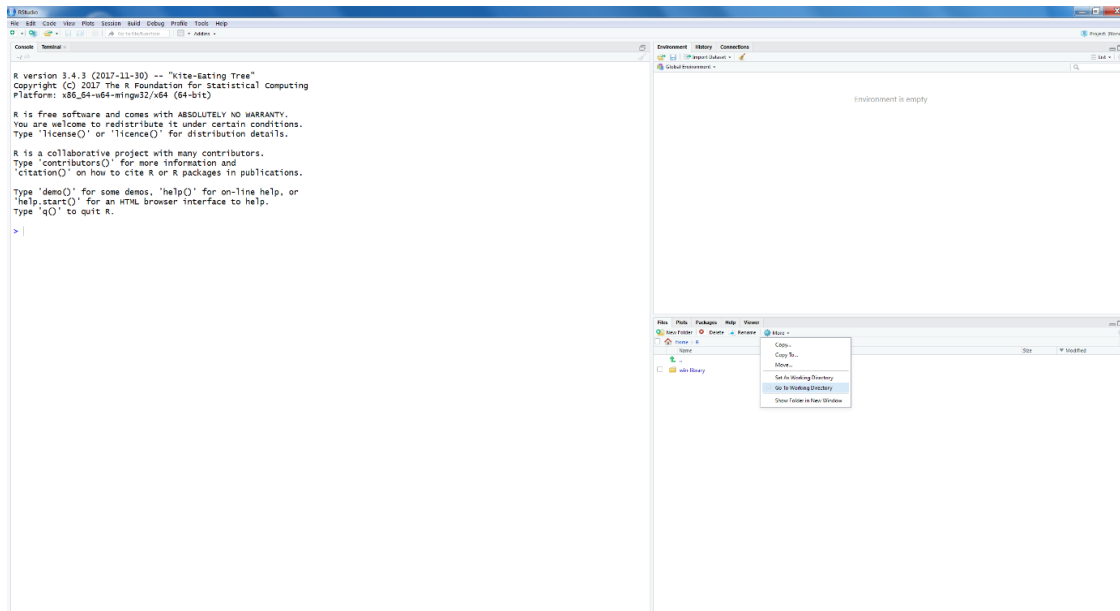



Figure 1.12: Working directory selection in RStudio

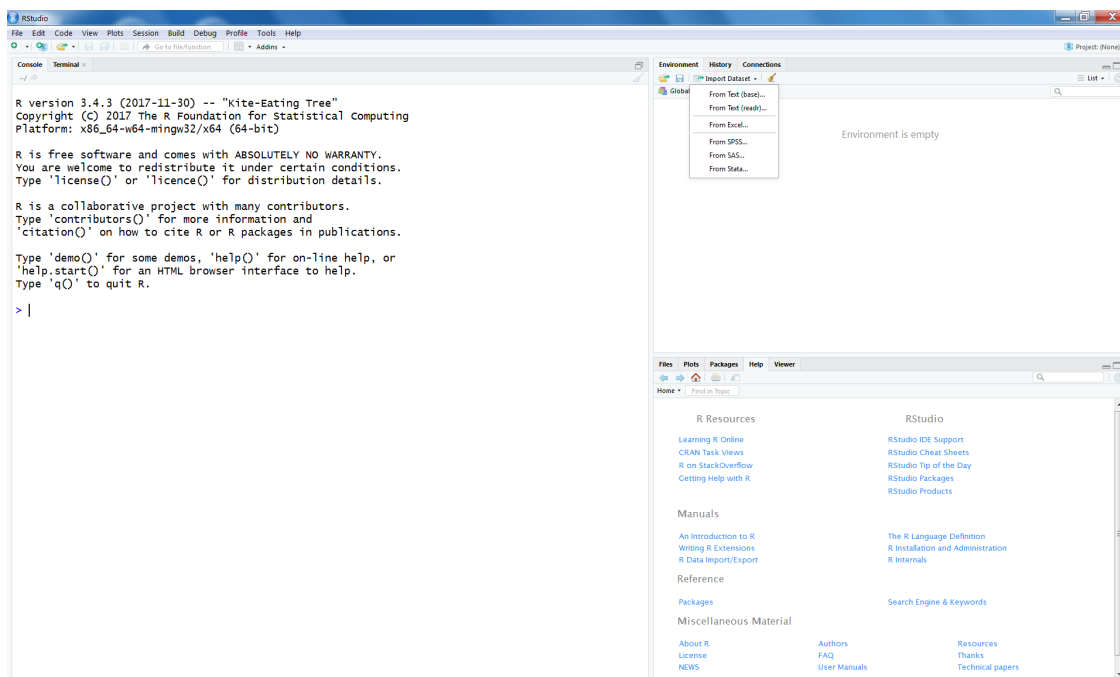


Figure 1.13: Screen to import datasets in RStudio

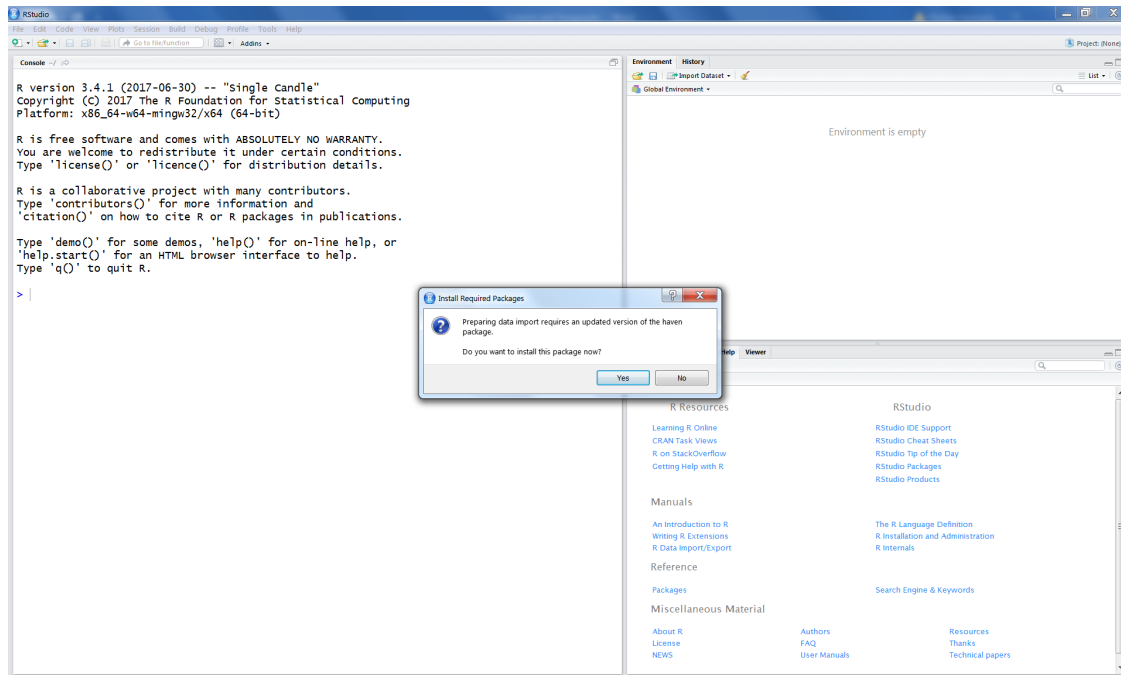


Figure 1.14: Window to install the haven package

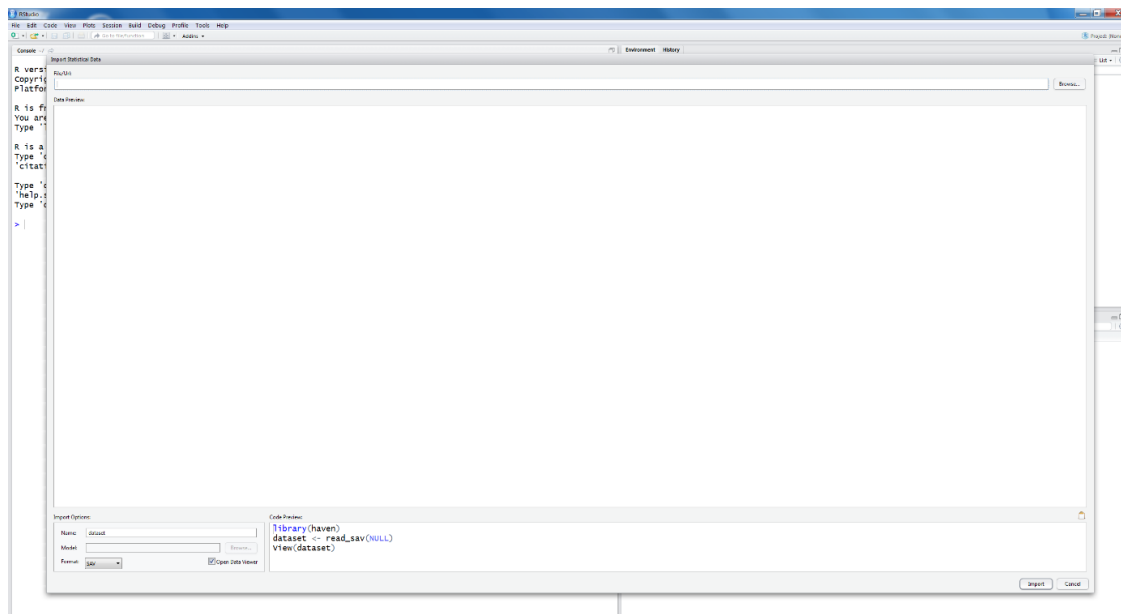


Figure 1.15: The Import Statistical Data window in RStudio

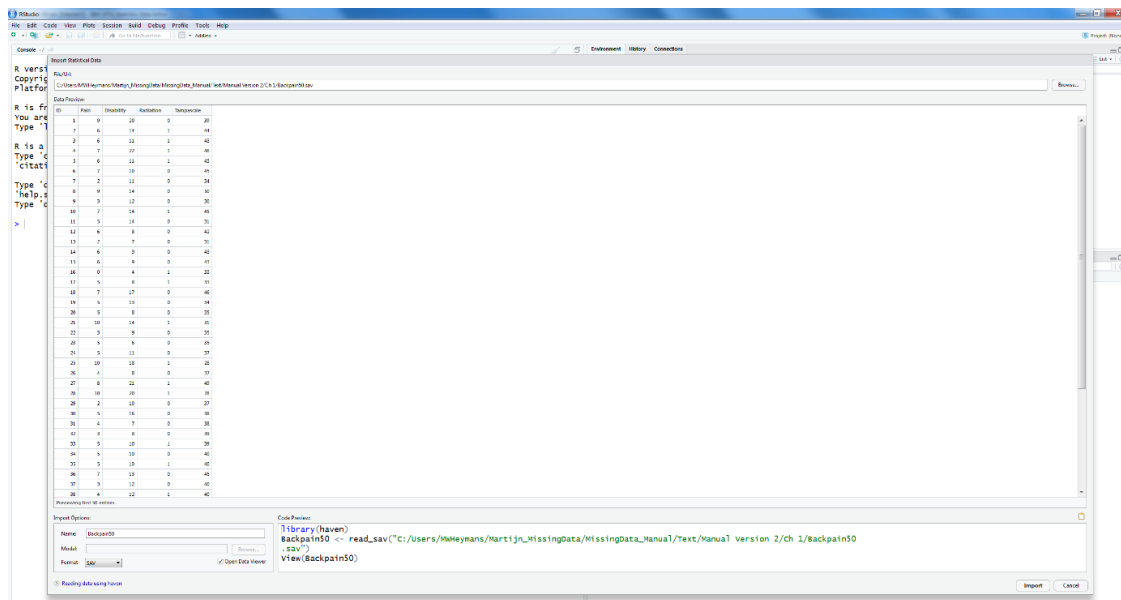


Figure 1.16: A preview of the dataset in RStudio

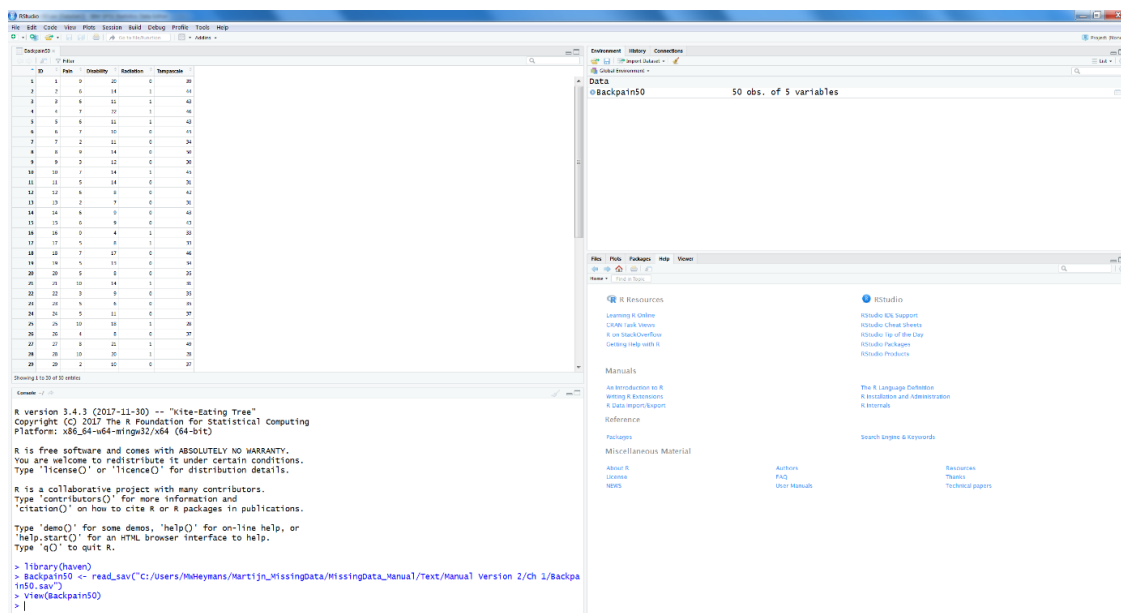


Figure 1.17: Imported dataset in RStudio

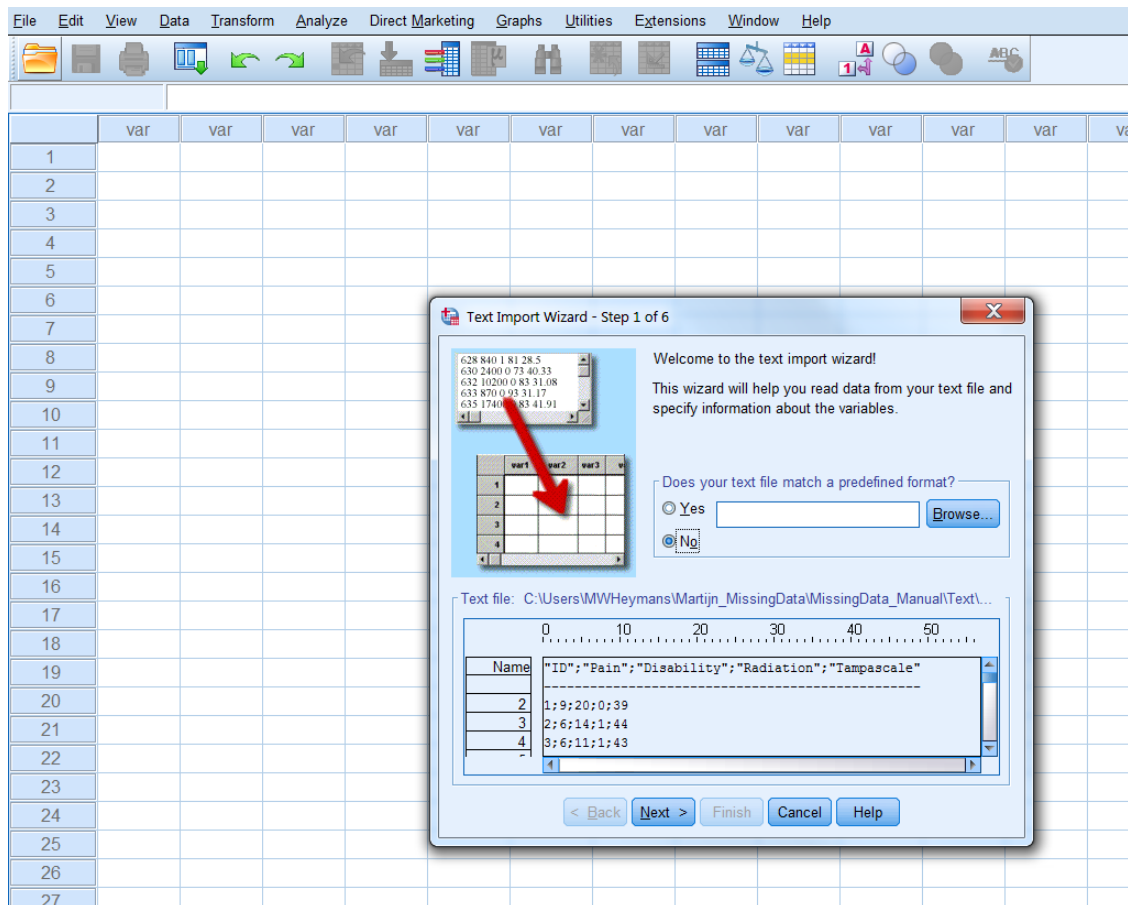


Figure 1.18: Step 1 of the Text Import Wizard

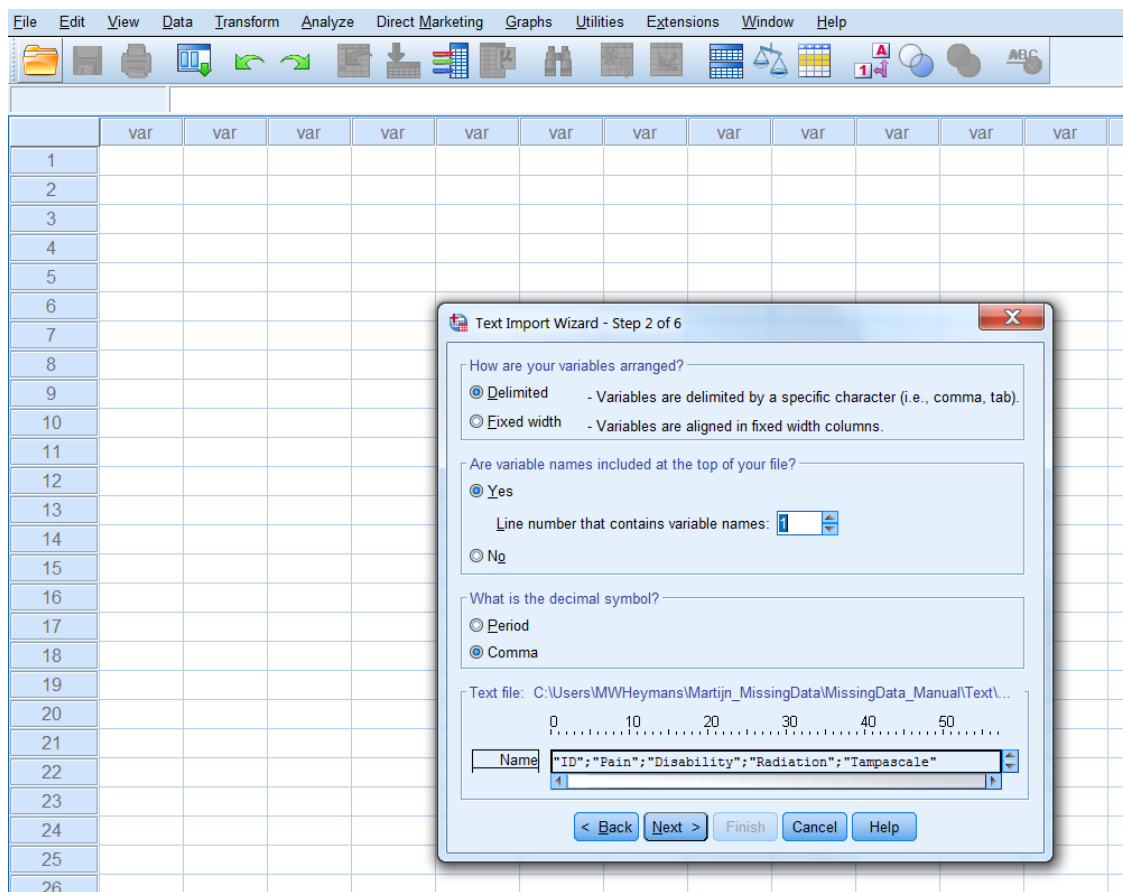


Figure 1.19: Step 2 of the Text Import Wizard

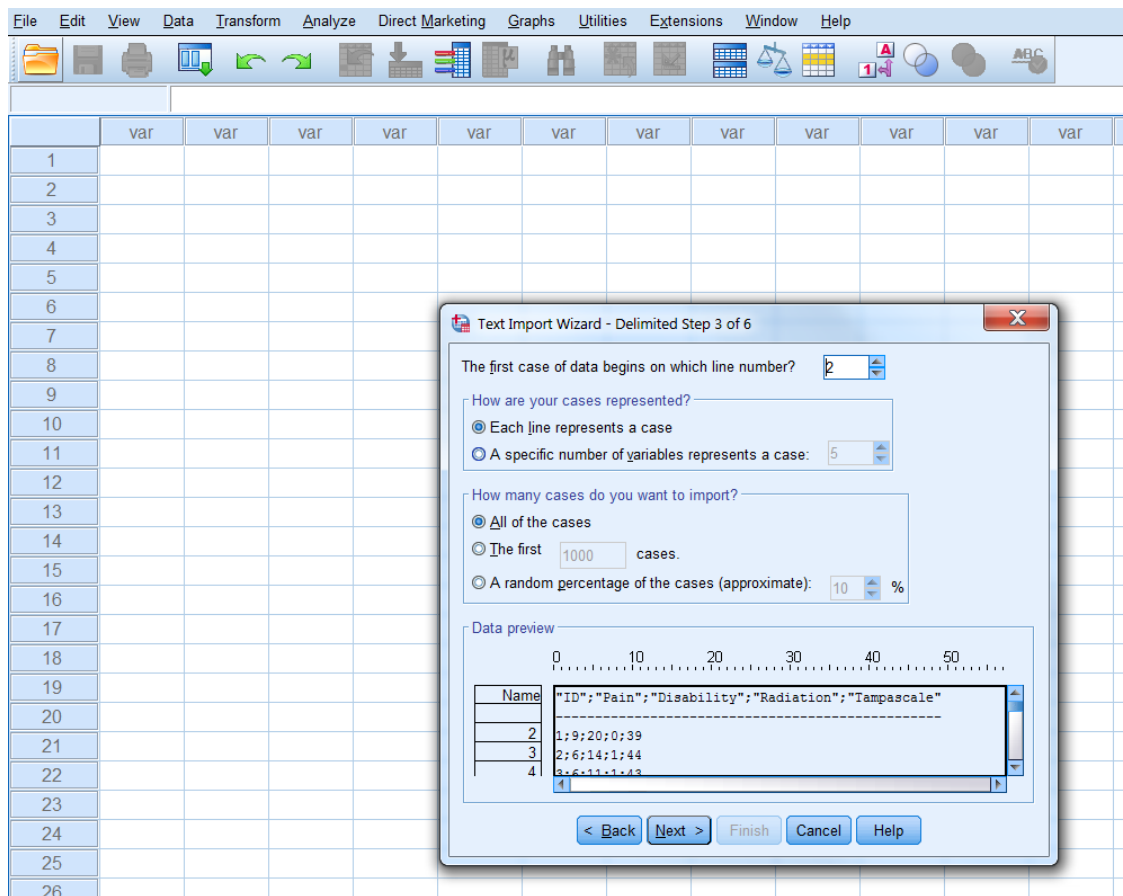


Figure 1.20: Step 3 of the Text Import Wizard

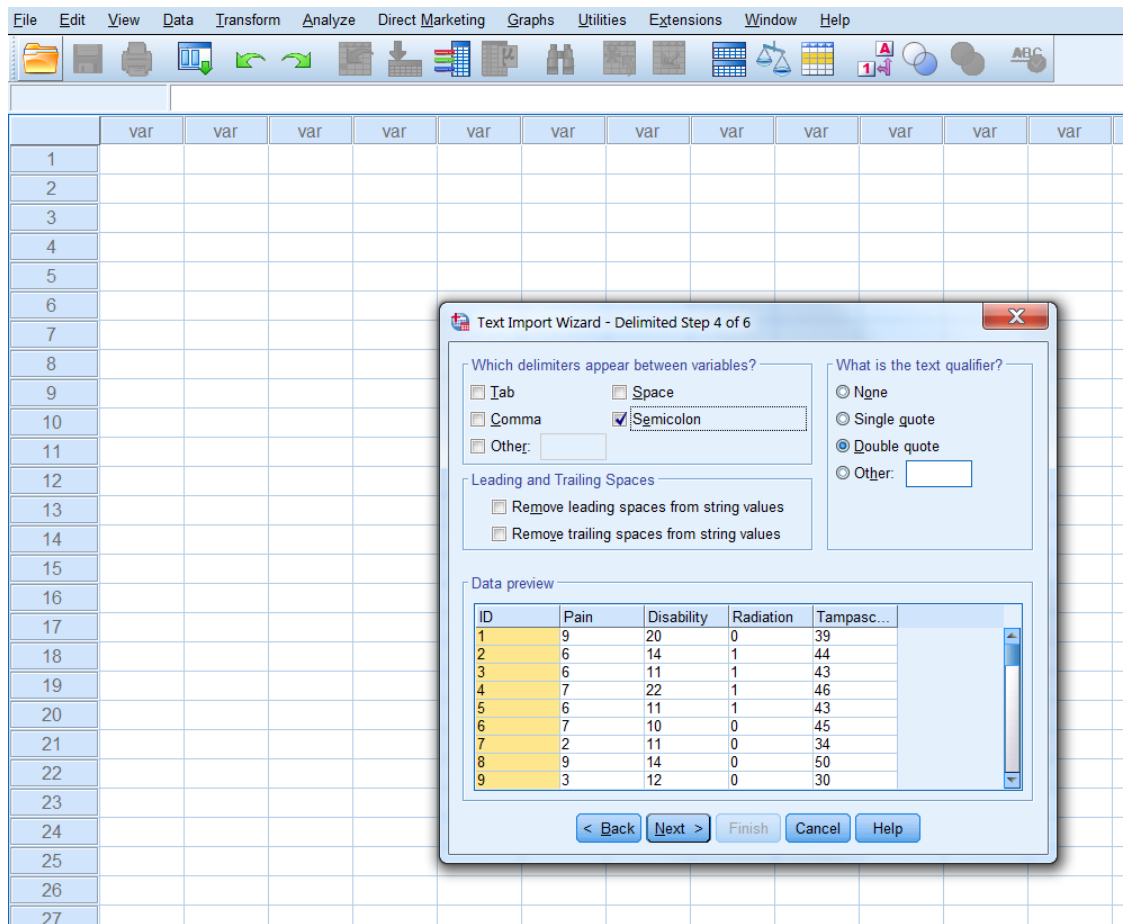


Figure 1.21: Step 4 of the Text Import Wizard

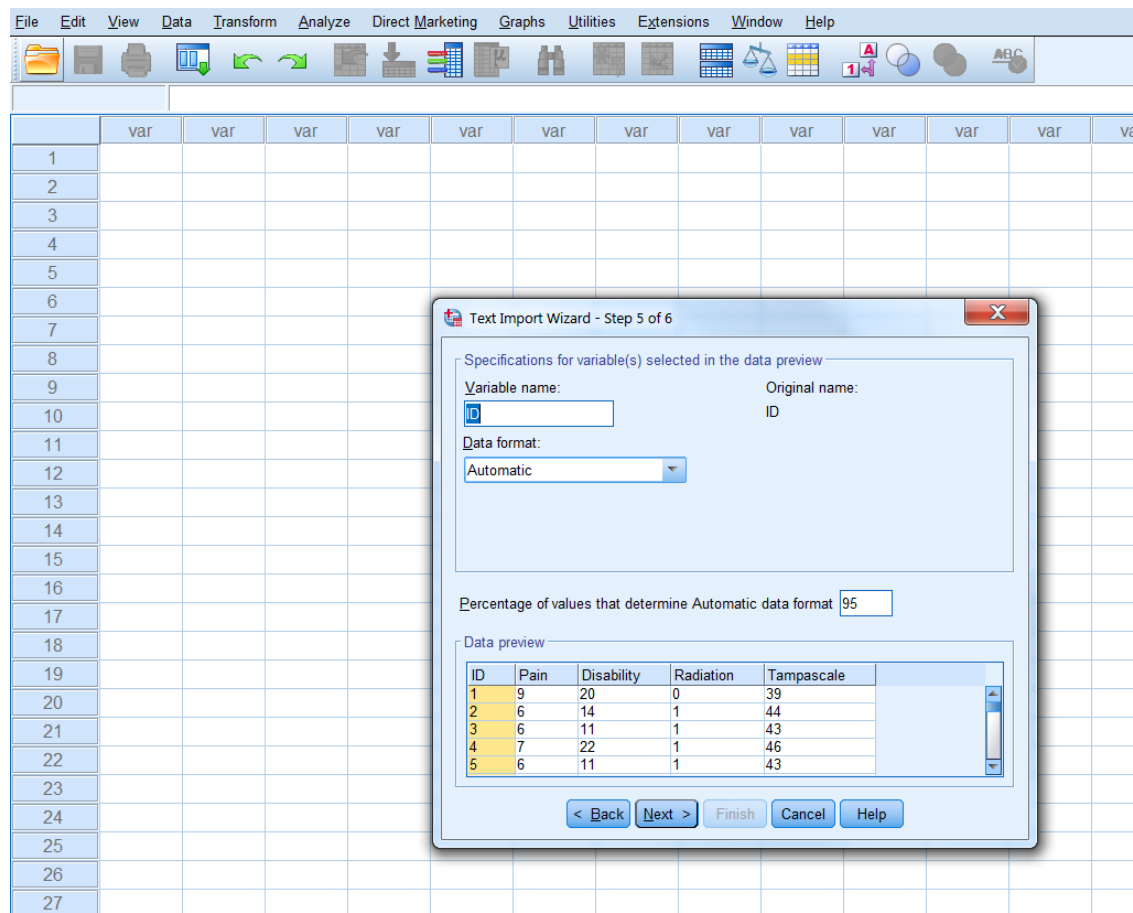


Figure 1.22: Step 5 of the Text Import Wizard

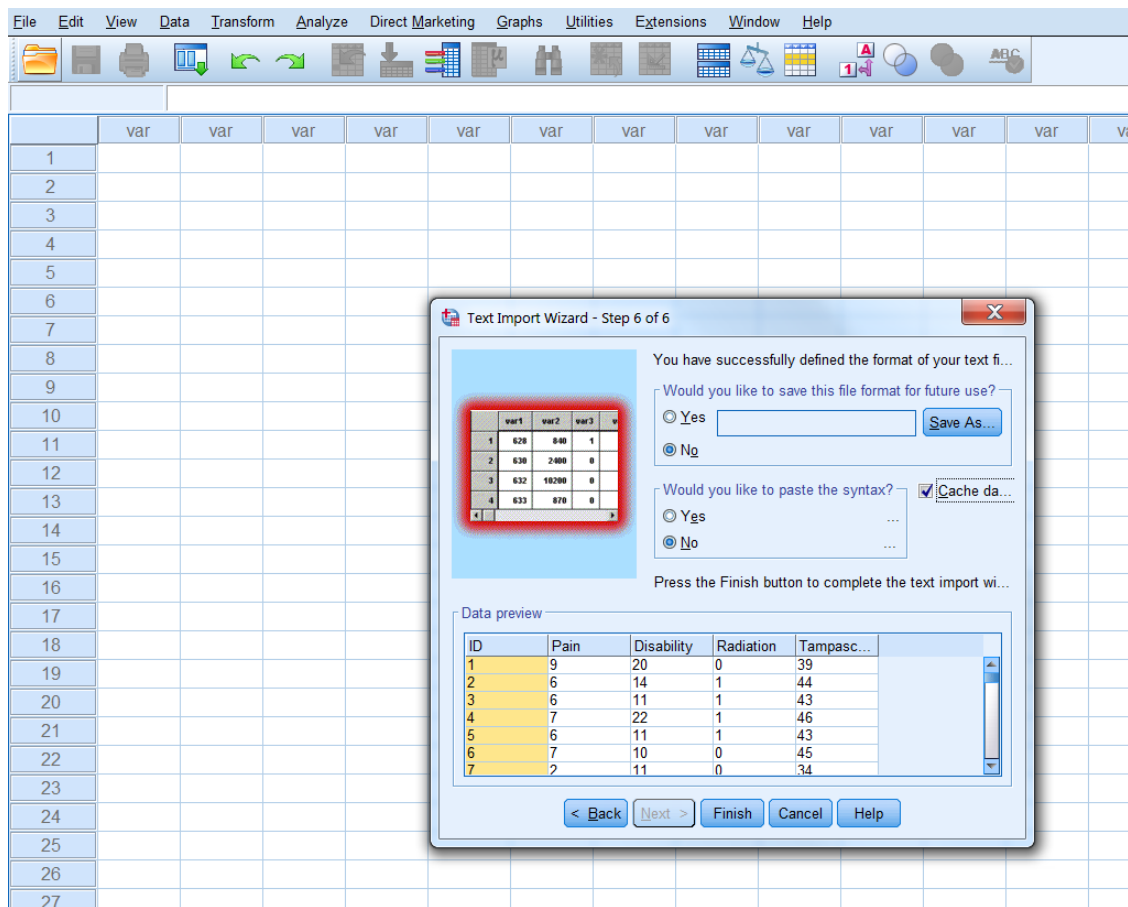


Figure 1.23: Step 6 of the Text Import Wizard

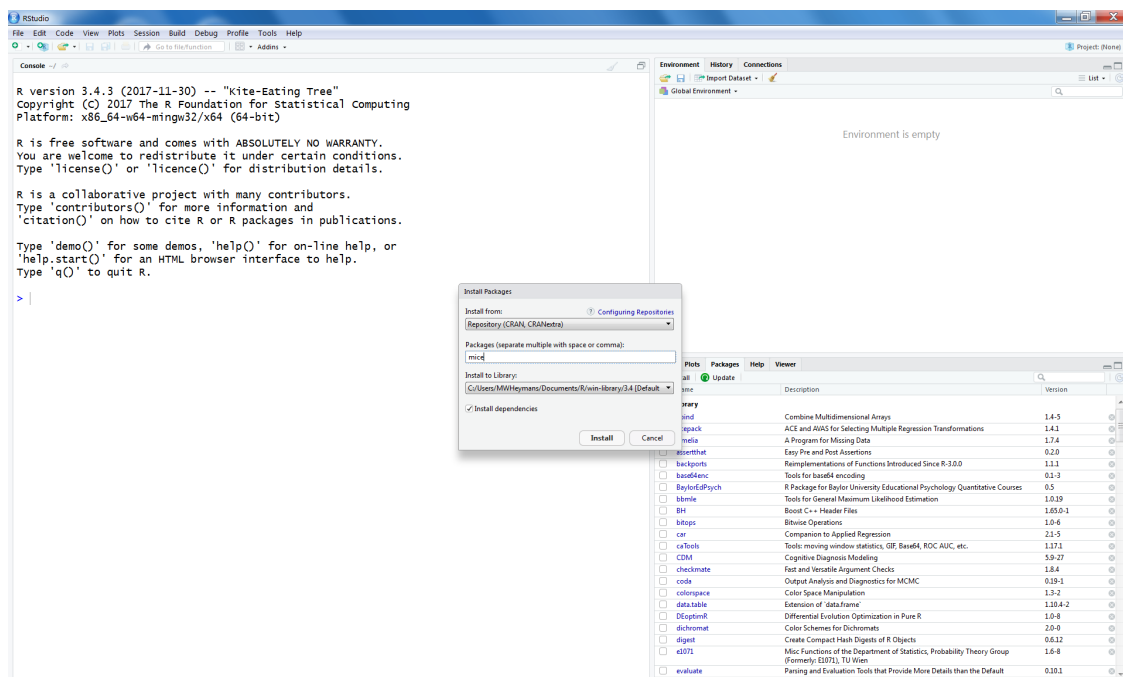


Figure 1.24: Install packages Window in RStudio to install packages from the CRAN website

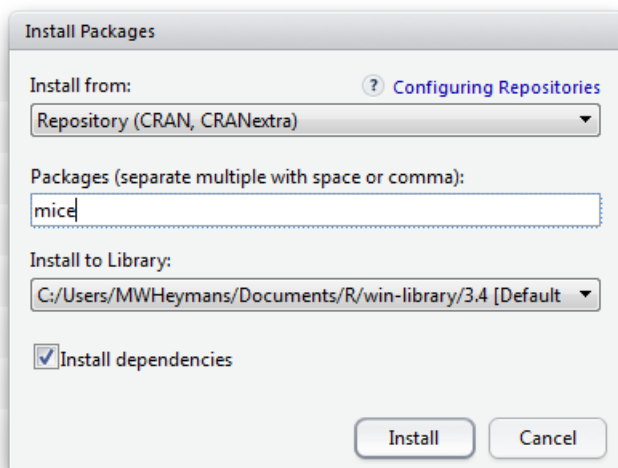


Figure 1.25: Enlarged Install packages Window in RStudio to install packages from the CRAN website

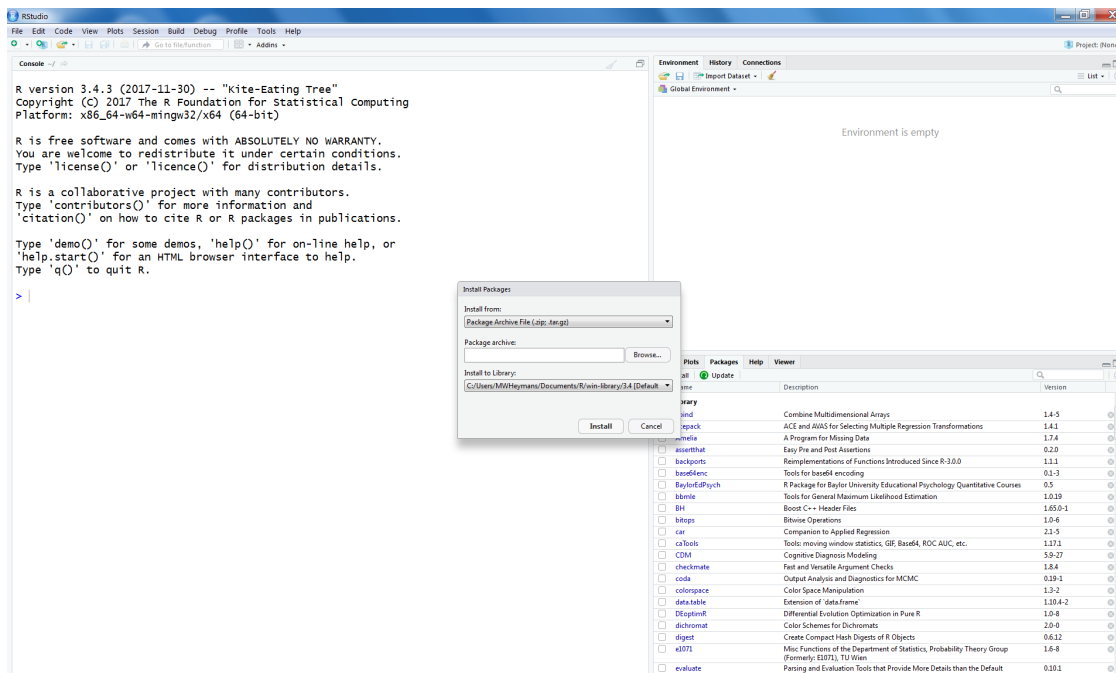


Figure 1.26: Install packages Window in RStudio to install packages from zip files

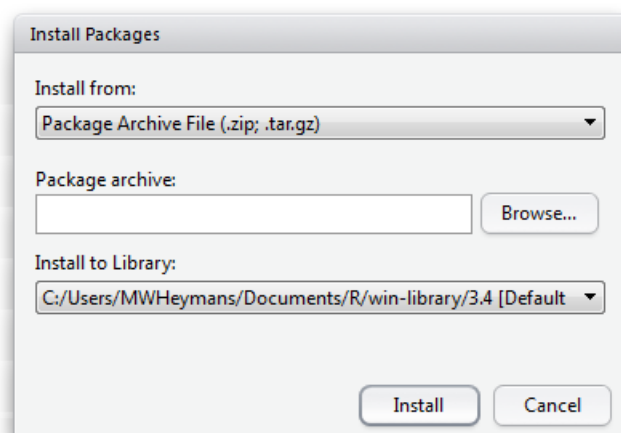


Figure 1.27: Enlarged Install packages Window in RStudio to install packages from zip files