Introduction to programming for data science

STAT 201

Arvind Krishna

2022-09-20

Table of contents

Preface

This book is currently being written for the course STAT201.

1 R: Introduction

1.1 Installing R

Go to the The Comprehensive R Archive Network (CRAN): https://cran.r-project.org/

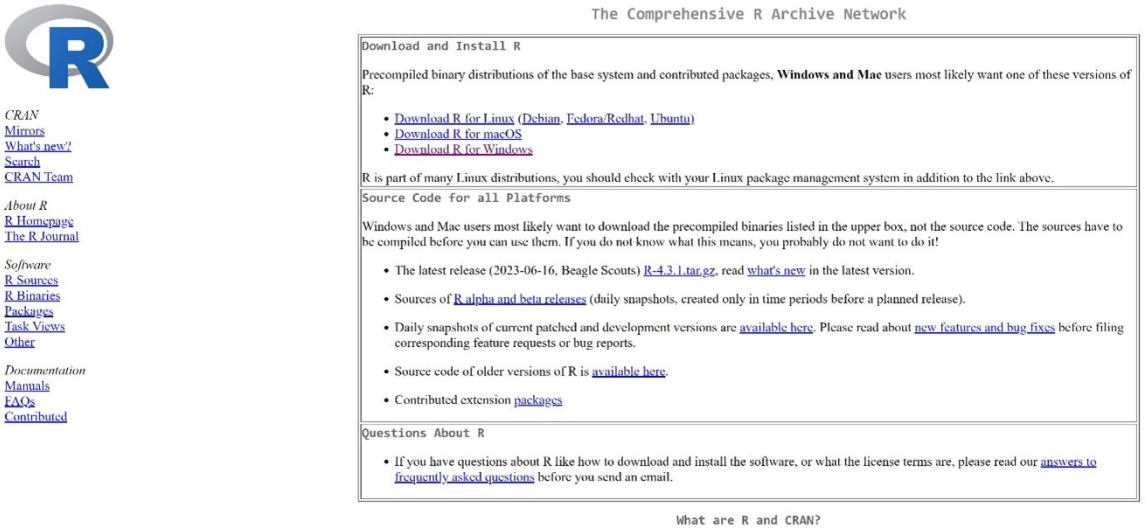

Figure 1.1: CRAN

Under "Download and Install R," choose "Linux," "MacOS X" or "Windows." If you choose Windows, on the next page choose "base," and on the following page choose "Download R 4.3.1 for Windows" to download the setup program.

If you choose MacOS X or Linux you will need to read through the instructions to find the downloads you need for your machine.

Once you have downloaded the setup program, execute it and follow the instructions for installing R on your system. If you have an earlier version of R already installed, you may continue to use it, or you can uninstall it and then install the most recent version, which is R 4.3.1.

1.2 Installing RStudio

https://rstudio.com/products/rstudio/download/

Choose your version: RStudio Desktop, Open Source License, Free. After you install RStudio, you can double click on it and open:

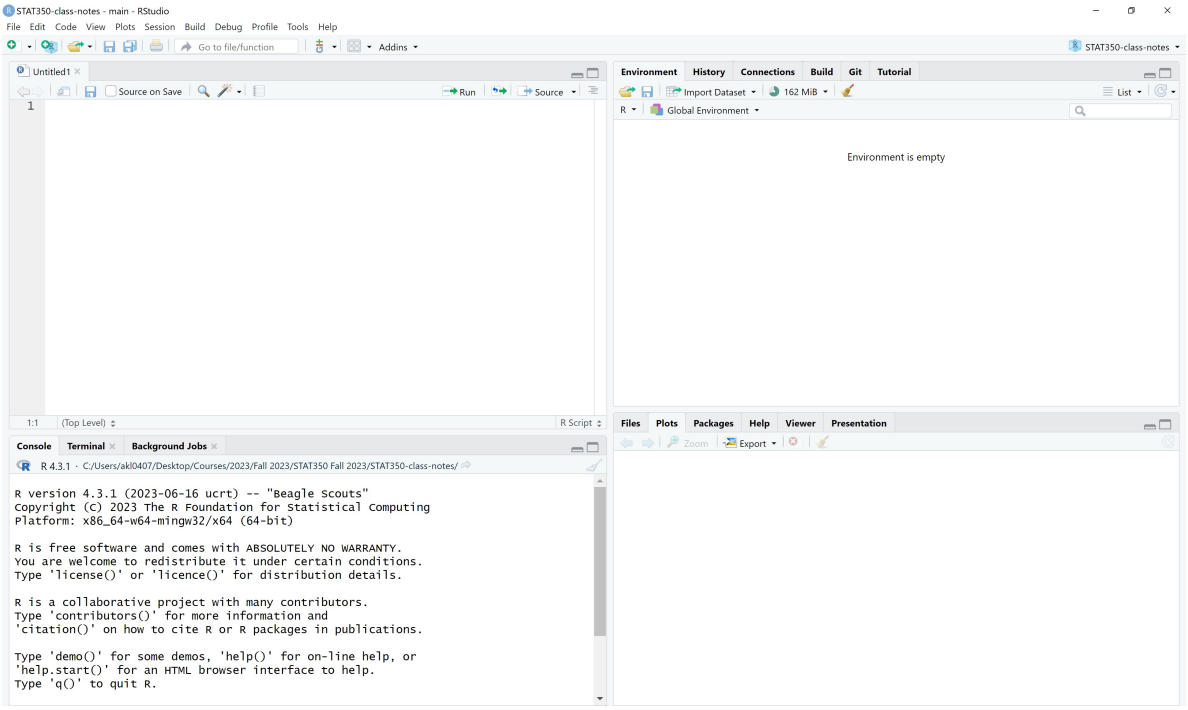

Figure 1.2: R Studio

Usually you will want to import data from a file corresponding to data associated with a homework problem. Such a file will usually end with the extensions **.txt* or **.dat*. The data files for this course will always be available on the CD that comes with the text and/or on the course web page. A data file will consist of columns of numbers, with nothing separating the columns but "white space." If each column has a title on top describing what the data in the column represents (e.g., `age`, `weight`, `income`, etc.), we will say that the file has a *header*.

1.3 Working directory

The easiest way to import the data into R and have it readily available for the current and future sessions is to first save the data file into your working directory. For example mine is `C:\stat350`.

To set up the working directory, select the project option by choosing `File` menu, then `New Project`, and then `Create Project from Existing Directory`.

To start writing a new R script, navigate to the `New File` option in the `File` menu, and select `Quarto Document`. This will create a **.qmd* file. You can write both code and formatted-text in this document. When working on assignment / exam problems, you will work on the **.qmd* file, render it as HTML and then submit. You can view some examples on how to write R code and text in a **.qmd* file and render it as HTML here.

For rough work, i.e., work that won't be graded, you may use the `R script` option to write code.

1.4 Getting started with code

1.4.1 Reading data

Suppose you want to work with the data from Problem 19 of Chapter 1, which is in a file named `CH01PR19.txt` which you have saved from the CD or the course web page in the `Datasets` folder within your R working directory. Assume the file has no header. You will want to create a Table object in R containing this data. First choose an appropriate name for the table. Assume you choose to name it `Data`. Then, you can execute the following code :

```
Data <- read.table("./Datasets/CH01PR19.txt")
```

Then there will be a Table object in R named `Data` containing the data in rows and columns. To view it, you would type

```
Data
```

However, if it is a large file, you might not be able to view the whole table at once. In that case, you may use the `head()` function, which will display only the first 6 rows of `Data`:

```
head(Data)
```

```
     V1 V2
1 3.897 21
2 3.885 14
3 3.778 28
4 2.540 22
5 3.028 21
6 3.865 31
```

A Assignment templates and Datasets

Assignment templates and datasets used in the book can be found here