

Intro to R and Installation Guide

Simona Simona

04/06/2019

0.1 What is R and why is it important?

R is a statistical programming language and computing environment which has become a *lingua franca* for the statistical community worldwide and it is increasingly becoming popular in the social sciences too. There are many reasons why this is the case. Probably most importantly, R is free. Licenses for commercial software are sometimes beyond what many students and academic staff can afford. A free software, especially one which is also fancy and powerful can come in handy for this demographic. R comes already loaded with enormous capabilities for data manipulation, analysis and graphics. On top of that, R users are free to extend it's source code and create their own add-on packages to perform unique tasks which may not be supported by existing functionality. This extensibility is one of the most important features of R. It means we can almost literally do anything in R. We can choose to use existing facilities, use packages developed by other users and available for free or create our own packages if what is already available does not satisfy our needs. Of course, we would need to brush up our programming skills if we are to do that. At the time of writing this guide, there were 14,346 available packages on the [cran](#) package repository.

There are several other reasons that make R very attractive. Many software packages have limits in terms of the number of variables in the dataset you import into the software, R is only limited by the memory capacity of your machine. This quality makes it a go-to statistical environment in the era of speedy, big volume and high variety data, otherwise known as **big data**. While most statistical packages would only handle one dataset at a time, RStudio is very happy to hold many datasets in it's environment and the user can interact with all of them at the same time. This can be very handy during data manipulation. It allows you to duplicate datasets and easily replace data points or variables in case of a mistake in one dataset.

R is also very good with graphics. The **ggplot2** package makes very attractive graphs in R. There are many people who have switched to using R from other statistical packages for that reason alone. The learning curve for **ggplot** is relatively steep especially for social science students but it is worth your time investment. It produces publication-ready graphs which may just boost the chance of your paper being accepted.

If you understand how R works, it will be much easier for you to learn other object-oriented programming languages like Python, Java, Ruby and C++ among others. R is a full object-oriented programming language and it comes with it's own functionality. It will therefore, introduce you to object-oriented programming used by the majority of programming languages. As such, if you choose to throw yourself in the programming world, you would have a huge advantage with R in your back pack and hopefully live happily thereafter.

The R community is expansive. It attracts many users who share ideas through many platforms including online courses, R users groups, conferences and Q&A websites. Having access to the internet is almost the sole prerequisite of being an expert in R. If you are stuck with anything, there is almost a 100% chance that someone else faced a similar problem before you. If you make Google your friend and search for solutions to any problem you are having, you will be surprised how much

information there is on the web. You will find many people have already provided useful answers to your problem. If for some reason you don't find what you are looking for, you are also free to ask your own questions. [Stack overflow](#) is the most popular Q&A site for this purpose. Even if you may feel terrible in the beginning, just should know that you are not alone, you have a large community of like-minded people who are eager to help you probably even more than you are willing to ask.

R Markdown has been a game changer for me. It provides an environment which makes R useful beyond data manipulation, analysis and visualisation. It allows researchers to combine the statistical and text components of their projects to create beautifully looking documents in the form of research reports, teaching documents, journal articles and presentations. Additional packages like bookdown and blogdown allow you to create books and websites respectively within the R environment. I don't think it can get any better than that. This very document, my entire PhD thesis and my personal [website](#) were created using R Markdown. I have never written any HTML, CSS or JavaScript code in my life¹. This means you don't only have to use R for data analysis or graphics but for a host of other non-statistical purposes too. Even if your work doesn't involve any computing, you may desire to distinguish yourself by creating cool presentations, journal articles or blog posts. In which case, R is your good friend and learning it will take you a few steps ahead of others.

In terms of jobs, data science has been described as the sexiest job of the 21st century and R is one of the prominent tools in the data science community. Learning R puts you in the bracket of people with highly valuable skills especially now when we literally have an explosion of data infrastructure worldwide. I know that for social scientists like me, this may sound far-fetched and intimidating. It is not. You do not have to know everything, no one does. You can specialise in a little piece of the many variants of data science. There is also a burgeoning field of computational social science which has significantly blurred the qual/quant divide. It largely involves leveraging novel technologies to collect and analyse digital data. You can either be a qualitative or quantitative researcher who is interested in using R to expand your sources of information. For example, you may want to learn how to use R and harvest Twitter data and use it to answer different types of research questions. You might also want to do web scrapping and collect data from the web such as blog posts, Google searches, government policy documents etc. That is computational social science and R can help you do that. But I would say that even if you don't want to do all that, you don't want to be a high flying data or computational scientists, adding R skills on your CV puts you at an advantage in comparison with other graduates in the job market.

0.2 Origins

R was created by Ross Ihaka and Robert Gentleman. Every R course you will attend, the instructor is likely to tell you that the R name came from the first letter of the first names of the creators. Well, this is partly true and since I like this reason better, I didn't want to bother finding out what the other part of the reason was. They were at the University of Auckland at the time. R is an implementation of the S language² which was created by John Chambers at the AT&T's Bell Labs in Texas in the 1970s. Now it is maintained and coordinated by the R Development Core Team, which is a committed group of volunteers across the world.

Fig 1. is the R interface and as you can see it is not very appealing and this has been partly why

¹I am told these are the programming languages you need to learn to create a website ground-up

²Don't worry if you don't understand the meaning of that statement. I don't either, so it makes the two of us, probably even more of us

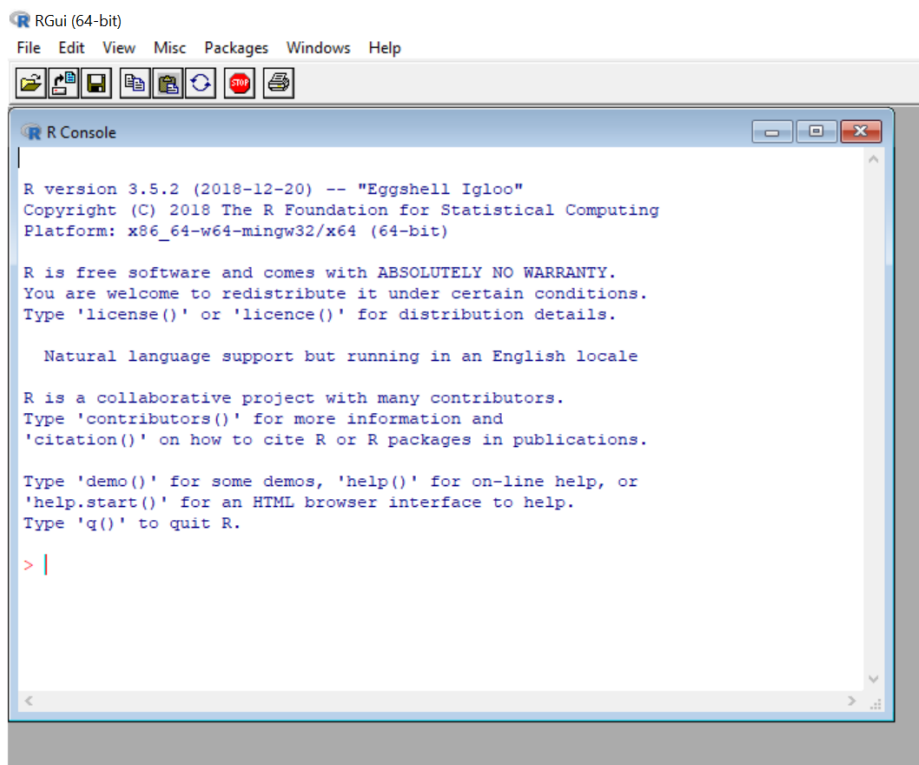


Figure 1: The R interface

not so many social scientists have not been enthusiastic about the software until quite recently. One of the major reasons for the surge in R usage in the social sciences, is the development of RStudio which had its initial release in 2011. RStudio is an open source integrated development environment (IDE). It provides a much more user friendly environment than the original R interface. From now onwards, when we refer to R, we will exchangeably be referring to R and RStudio unless so specified. RStudio continues to provide powerful tools that aid the usability and operationalisation of R not only for data manipulation, data analysis and visualisation but also for non-statistical purposes like creating reports, journal articles and presentations through R Markdown as we have stated above.

We shall learn how to download RStudio in this chapter and inspect its environment. But we must note before installing RStudio, we need to install the original R first RStudio will not work without R. I always compare the relationship between them to that of an engine and car body. The body of a car provides a user friendly environment for you to comfortably drive a car. You don't even have to know what is going on under the hood for you to enjoy the ride but of course, the car would never run without the engine unless you are witnessing an apocalyptic miracle. We will thus start by finding R and downloading it and then move on to RStudio.

0.3 Finding and Installing R

Downloading and installing R is not very different from any software. R can be downloaded from the Comprehensive R Archive Network (CRAN). You can Google the **R project** or use the this website: <https://www.r-project.org/>. The first 2 lines under the getting started subtitle, click on the



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Developer Pages](#)

[R Blog](#)

R Foundation

[Foundation](#)

[Board](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- [R version 3.5.3 \(Great Truth\) prerelease versions](#) will appear starting Friday 2019-03-01. Final release is scheduled for Monday 2019-03-11.
- [R version 3.5.2 \(Eggshell Igloo\)](#) has been released on 2018-12-20.
- The R Foundation Conference Committee has released a [call for proposals](#) to host useR! 2020 in North America.
- You can now support the R Foundation with a renewable subscription as a [supporting member](#)
- The R Foundation has been awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers.

Figure 2: The R project website

highlighted **download R** or the highlighted **CRAN mirror** at the end of the paragraph. See Fig. 2 for this. When you open the cran mirror page, you should see that it resembles Fig. 3 and browse through the countries on left hand side of the screen and click on the city nearest to you. You don't have to be completely accurate here. Essentially, whatever city you choose should work just fine. After this process you should now see Fig 4. Choose your operating system accordingly. If you are using a Windows, please don't choose Mac. It is that simple! It should be pretty straightforward from here. On the next page click on **install R for the first time** if you are using a Windows machine. Then you should see the latest version of R on the next page **Download R 3.6.0 for Windows**(this is true at the time of writing the chapter. The latest version is expected to change with time).

If you are using Mac, after fig.4 and you have selected appropriately you should see a window resembling figure 5. The latest version of R should appear on your left hand side marked **R-3.6.0pkg**. If you click on that, you should get your installer and can launch it accordingly as you would with any other software.


I don't have instructions for Linux. I have never seen the Linux operating system before, so I can not possibly give instructions on something I don't know. If you are using Linux system and you can't figure out how to do it yourself, you can bring your machine over to the workshop, I am sure we will manage to figure it out together.

0.4 Installing RStudio

It is worthy repeating that RStudio is the platform we shall be using in this book but it should be installed only after you have followed the instructions above and have installed R accordingly. RStudio is found on this website: <https://www.rstudio.com/>. Or you can also Google it, which

CRAN Mirrors	
The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page , windows release , windows old release .	
If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO .	
0-Cloud	Automatic redirection to servers worldwide, currently sponsored by Rstudio
https://cloud.r-project.org/ http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Algeria	
https://cran.usthb.dz/ http://cran.usthb.dz/	University of Science and Technology Houari Boumediene University of Science and Technology Houari Boumediene
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
https://cran.csiro.au/ http://cran.csiro.au/ https://mirror.aarnet.edu.au/pub/CRAN/ https://cran.ms.unimelb.edu.au/ https://cran.curtin.edu.au/	CSIRO CSIRO AARNET School of Mathematics and Statistics, University of Melbourne Curtin University of Technology
Austria	
https://cran.wu.ac.at/ http://cran.wu.ac.at/	Wirtschaftsuniversität Wien Wirtschaftsuniversität Wien

Figure 3: The R CRAN Mirror Page



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)


R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2018-12-20, Eggshell Igloo) [R-3.5.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.

Figure 4: The R Platform distributions



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

Documentation

[Manuals](#)

[FAQs](#)

R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS X 10.1) are no longer supported but you can find the last supported release of R for these systems (which is R 1.7.1) [here](#). Releases for old Mac OS X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the [old](#) directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries, please use the normal precautions with downloaded executables.

As of 2016/03/01 package binaries for R versions older than 2.12.0 are only available from the [CRAN archive](#) so users of such versions should adjust the CRAN mirror setting accordingly.

R 3.5.2 "Eggshell Igloo" released on 2018/12/20

Important: since R 3.4.0 release we are now providing binaries for OS X 10.11 (El Capitan) and higher using non-Apple toolkit to provide support for OpenMP and C++17 standard features. To compile packages you may have to download tools from the [tools](#) directory and read the corresponding note below.

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process. For example type

```
md5 R-3.5.2.pkg
```

in the *Terminal* application to print the MD5 checksum for the R-3.5.2.pkg image. On Mac OS X 10.7 and later you can also validate the signature using

```
pkgutil --check-signature R-3.5.2.pkg
```

Latest release:

R-3.5.2.pkg
MD5-sha1: 0c4ee980c9a1799329c0d524454563e
SHA1-bash: c71d8060d6b6c345ac968892a8f26b466a2c5b7
(ca. 74MB)

R 3.5.2 binary for OS X 10.11 (El Capitan) and higher, signed package. Contains R 3.5.2 framework, Rapp GUI 1.70 in 64-bit for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 5.2. The latter two components are optional and can be omitted when choosing "custom install", they are only needed if

Figure 5: The R Download Window for MacOS

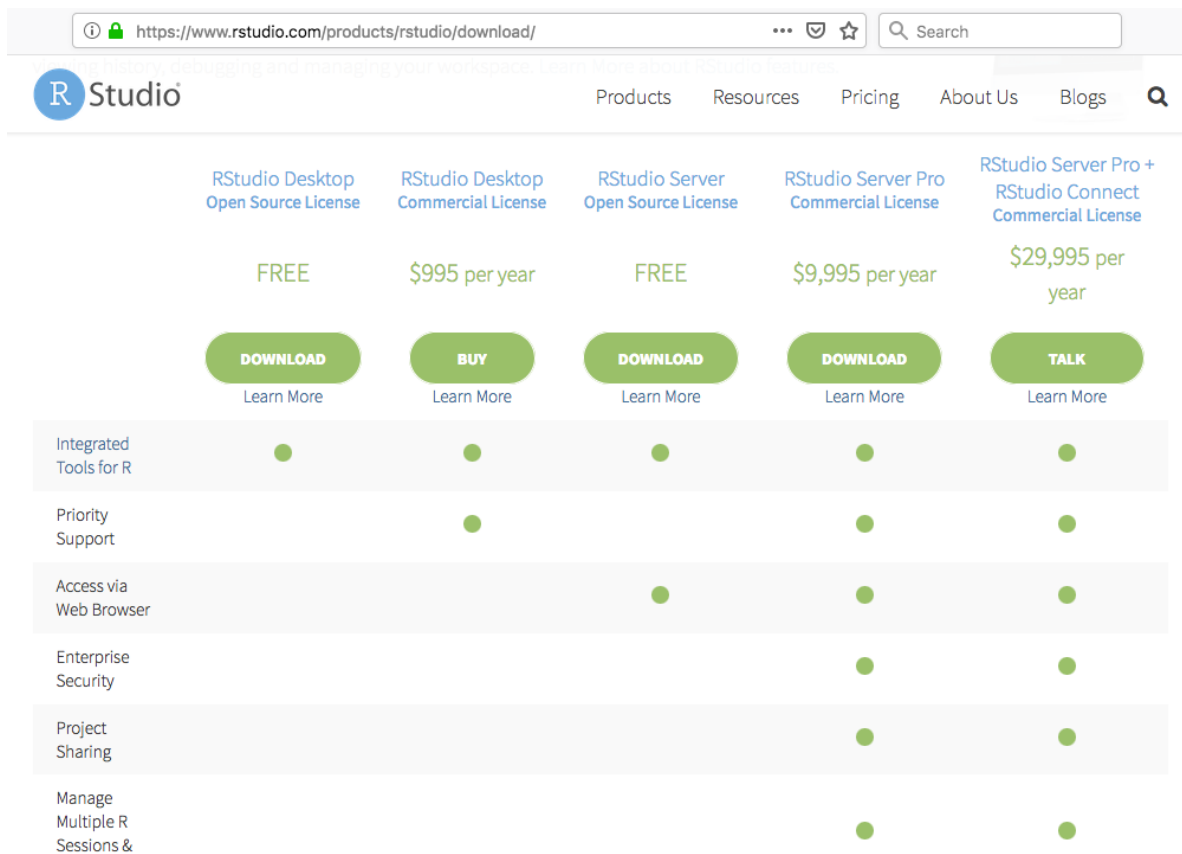


Figure 6: The RStudio Download Window

should eventually take you to the same website. When it opens, you will see a few RStudio products listed, mainly RStudio, Shiny and R Packages. Under R Studio, on the left hand side, click the **Download** button which should take you to the RStudio download page as shown in fig. 6. There are different licences that RStudio offers, some of which are commercial. The open source versions of RStudio are more than enough for us and since we have marketed R so well for being an open source software, we should equally emphasise that the free version of RStudio offers you literally everything you need. So, unless you are a business and you think you are likely to benefit more from the commercial licences, or you feel the urge to contribute something to RStudio for the wonderful job they are doing, you probably want to stay on the free side of things.

Under the free licence, you can either click on the **download** button or scrawl down to the **installers for supported platforms** section. Select the installers according to your operating system and follow the straightforward instructions to install the software. After this process we are basically good to go. R and RStudio should be appearing on the Start menu for Windows or the Applications folder for Mac. As I said, we are going to be using RStudio and so, at this point you can double-click on it's icon to open the environment, which takes us to the next section.

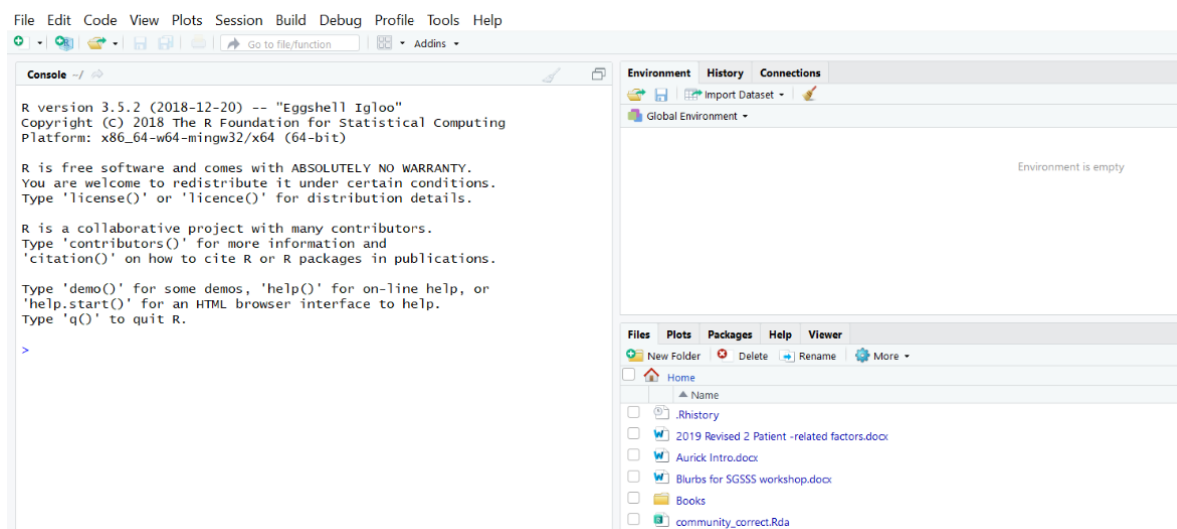


Figure 7: The RStudio Environment

0.5 The RStudio Environment

When you open RStudio, we are essentially in a new R session and the window you get should look like Fig.7. The new session of RStudio has 4 windows, although only 3 will appear initially. We will get to the fourth one shortly. The main window on the left hand side, is the **console** and this is where the main outputs will be shown. The right hand side has two windows and each one of them has a few tabs. The top right window has 3 tabs. The **Environment** is a place where your data frames, values and functions that are imported or created within RStudio are shown. The **History** tab keeps a record of all the commands we have used during our current session. You may not use the **Connections** tab but it allows you to connect 2 databases including relational ones. It is definitely for advanced users. The bottom right window has multiple tabs. It is the area you should see your files that are in your working directory under the **Files** tab. **Plots** is the place where the plots executed in the console or the r script (to be discussed later) will appear. A list of all packages installed on our computer will appear in the **packages** tab. The **Help** tab displays documentations for any R functions or packages that you have sought further information on. RStudio allows you to seek help about a particular function or package by typing the relevant code in the R console or source script and the results will be shown in the help tab. Many people may not even do this as they have Google for friend. They would rather consult Google so that they have options in terms of which website to read the documentations from and also the webpage might have a better resolution compared to the RStudio tab. Finally we have the **Viewer** tab which displays the web content locally. This is useful when you want to use the more advanced facilities in RStudio to create websites using the **blogdown** package or interactive web applications using **RShiny** respectively. You will be able to preview them using this tab before launching them on the internet. When I was creating my website in RStudio, I was able to see it in this tab before deploying it on the web.

Now it is time to go back to the fourth window and it is called the **source** or **script** editor. The R console can be used to write and execute code as highlighted earlier but the code written there cannot be saved or replicated. That's why we need the script window. It is the top left window in RStudio. It supports many forms of script files but here we are dealing with **.R** files. It does not

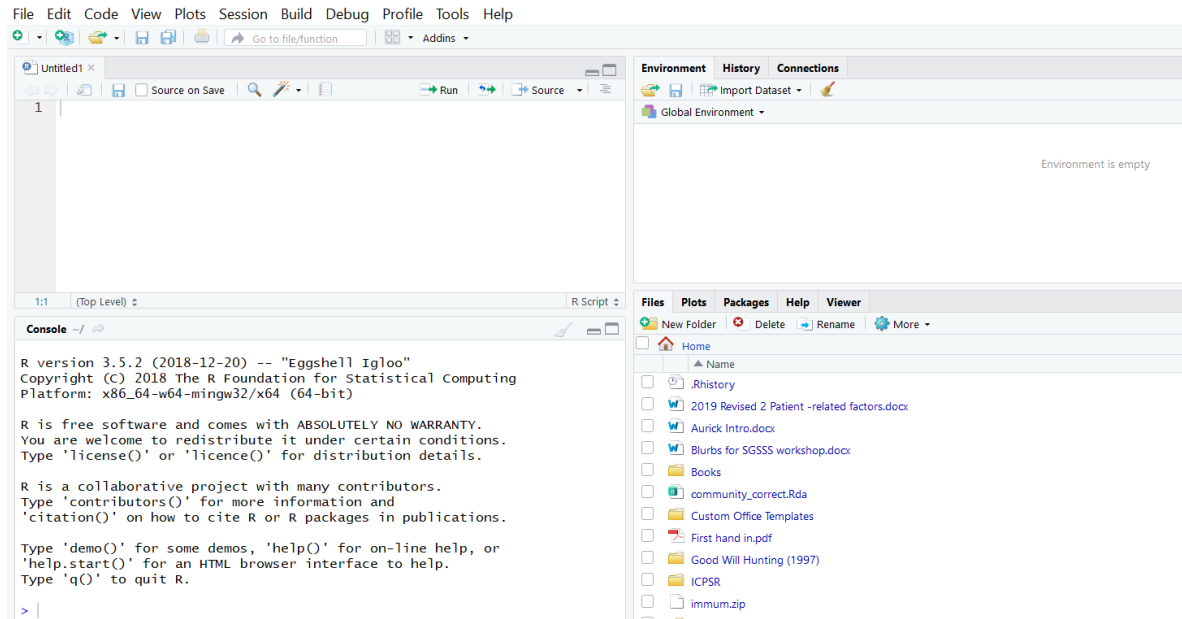


Figure 8: The RStudio Interface

always open when you launch RStudio for the first time. To open it, you use the **File** → **New File** → **R Script** menu. The File menu is located in the left hand corner of RStudio. The source script window is where we type our code to be executed in the console. We can most importantly document our code and save it for future use. For those of you familiar with **Stata**, this is similar to the **do-file**. The full RStudio interface looks like the figure above (Fig.8). This window is going to be home for any serious analysis. We can save the code we write here for our analysis or graphics and come back to it later. Not only that, we can also easily correct errors in our code. The script editor comes with handy productivity enhancing features such as auto completion, syntax highlighting, multiple-file editing and find/replace. To run an R code, you just need to click anywhere on the line of code you want to run and then click the Run button on the top right corner of the **source script** window. Alternatively you can use **Ctrl + Enter** if you are using a windows machine or **Cmd + Enter** on a Mac.

See you in the workshop