Data Science Workflow

Workflow for data analysis is streamlined by using a R project.

Using a R project allows you to find materials, your scripts and data for the work, particularly helpful if you leave the work and don't come back to it for a week or two weeks or months or years. It helps to avoid idiosyncrasies in directory referencing across different operating systems. It's easy to provide your analysis to someone else to run modify and verify.

Getting into the workflow

Storing files in folders, and folders in a 'filing cabinet' helps centralise your work: it keeps it organised so it is easier to find.

Using **RStudio projects** is like providing a filing cabinet for your work! Using them centralises your work, making your life easier as you do not have to manage where files are. For each project, you need to create **one** RStudio project.

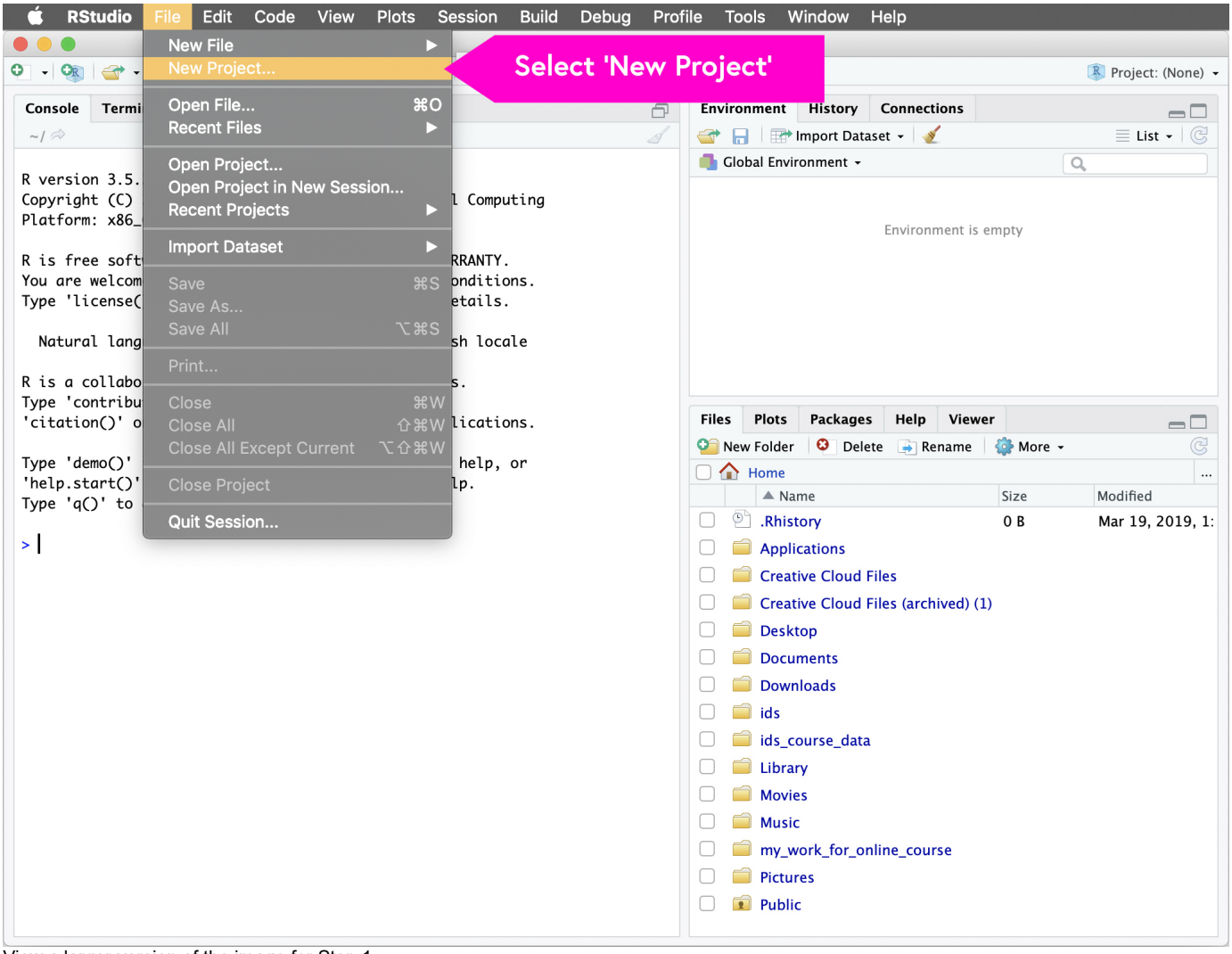
Consider reading Project-oriented workflow (https://www.tidyverse.org/blog/2017/12/workflow-vs-script/) by Jenny Bryan for insights into the importance of workflow.

Creating a new RStudio project

Make your way through the following steps to create an RStudio project on your computer for all the work that you'll do in this course.

Step 1: Start a new project

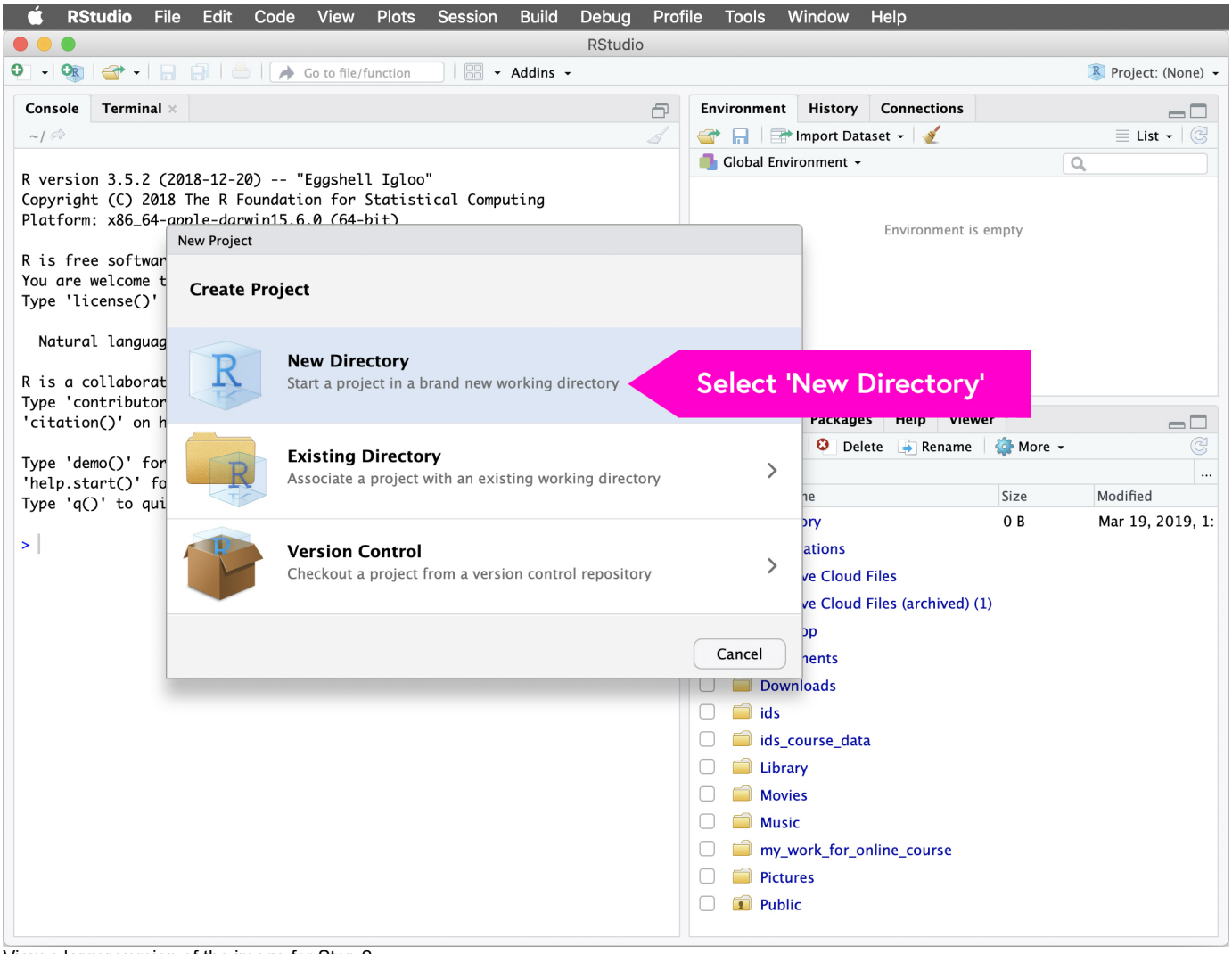
On your computer, open RStudio. Then, select **'File'** > **'New Project'**.



View a larger version of the image for Step 1. (https://ugc.futurelearn.com/uploads/assets/43/d0/43d05b5f-0bf3-4944-9738-b69c90c15478.jpg)

Step 2: Set a directory

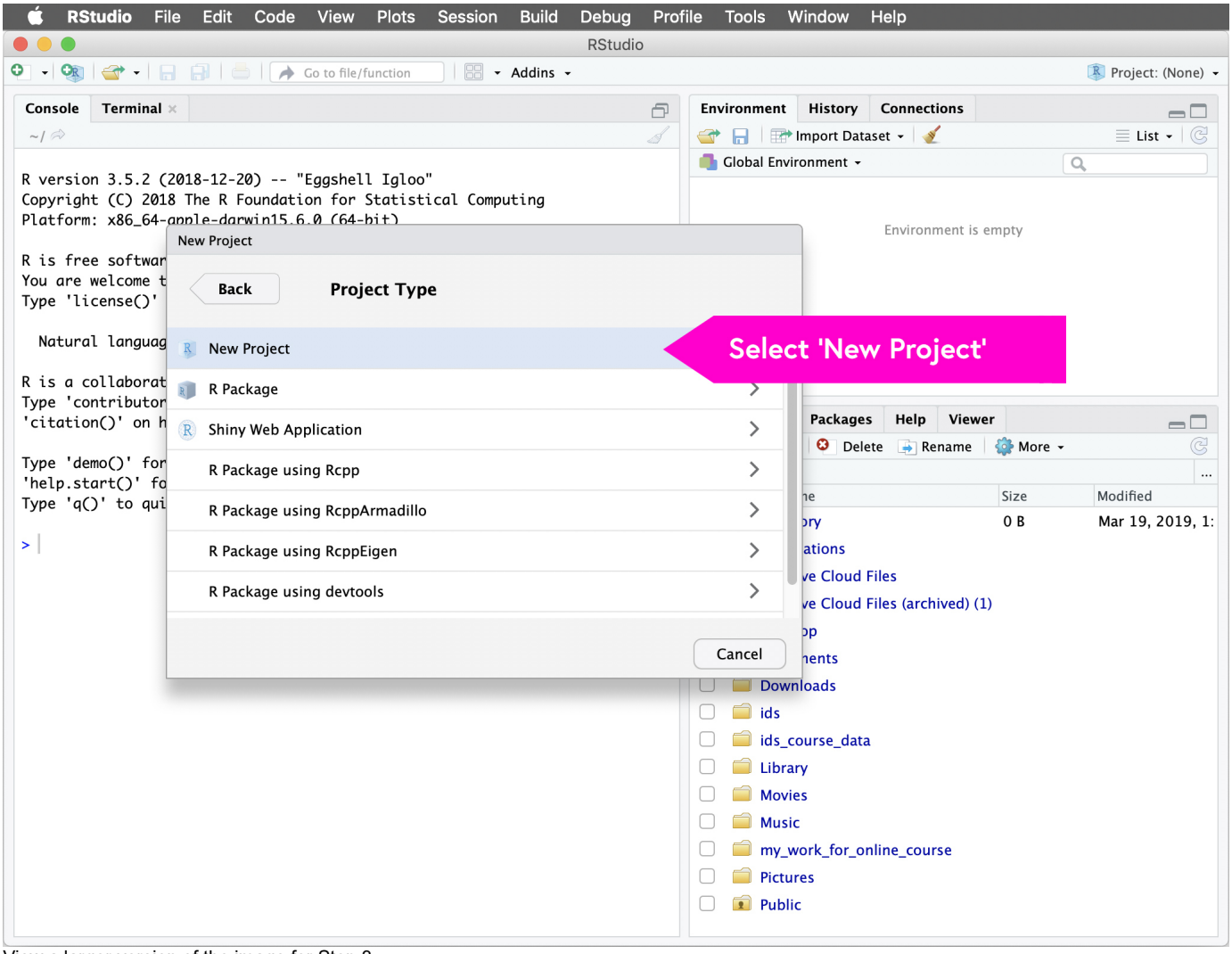
A pop-up window labelled **'New Project'** should be displayed in RStudio. From the pop-up window, select **'New directory'**.



View a larger version of the image for Step 2. (https://ugc.futurelearn.com/uploads/assets/a4/58/a458f9a9-b8cf-4b67-be37-e79de37e6bbe.jpg)

Step 3: Select the project type

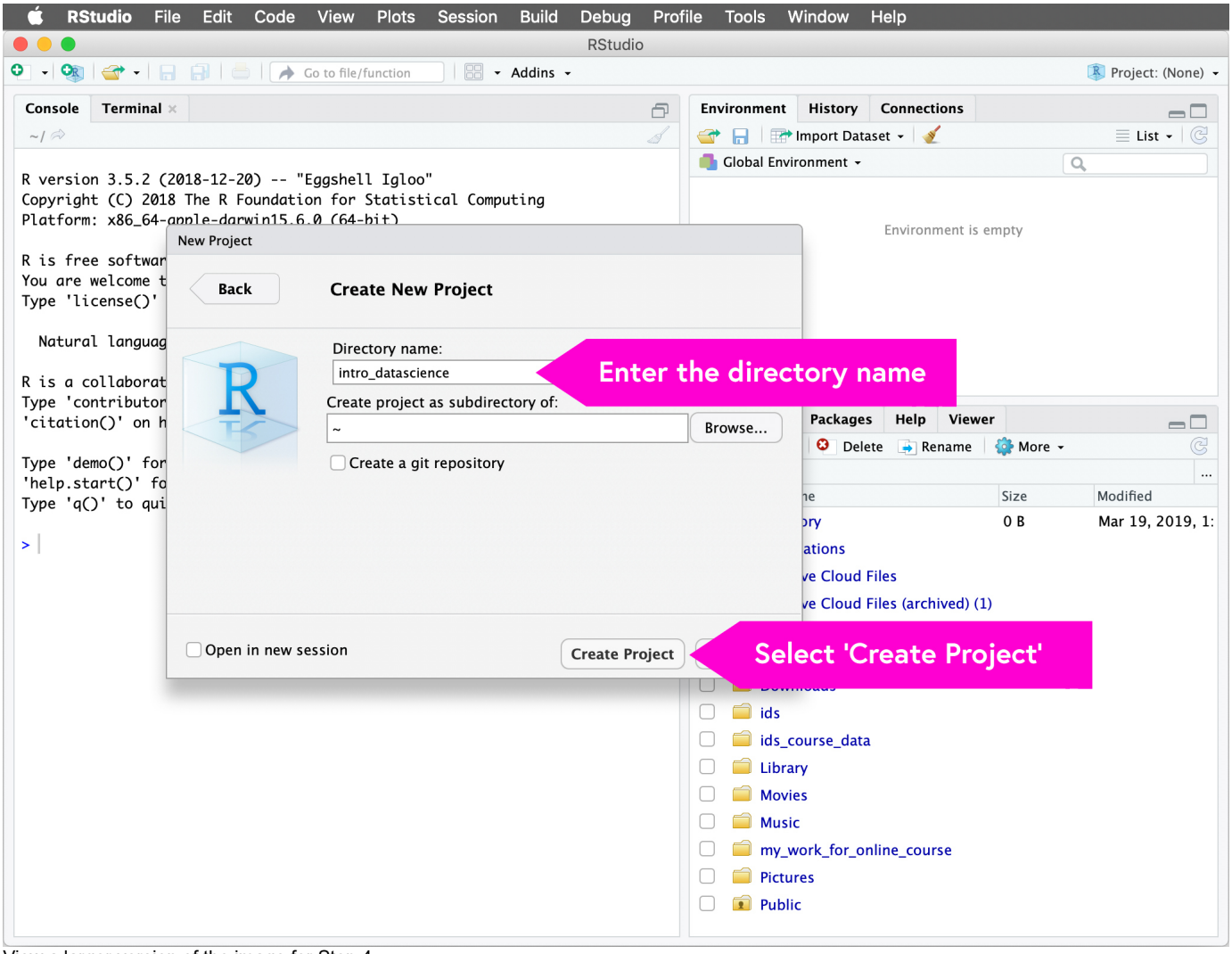
You're creating a new project, so select **'New project'**.



View a larger version of the image for Step 3. (https://ugc.futurelearn.com/uploads/assets/f3/62/f362f08f-8cbb-4e4e-8baf-e7a9f180fe7e.jpg)

Step 4: Give your directory a name

Enter the name of the directory you want to create. You can call it anything you like, for example: **“intro_datascience”** or **“data_science_decision_making”** or something else that’s meaningful to you. Select **‘Create project’** once you’ve named your directory.

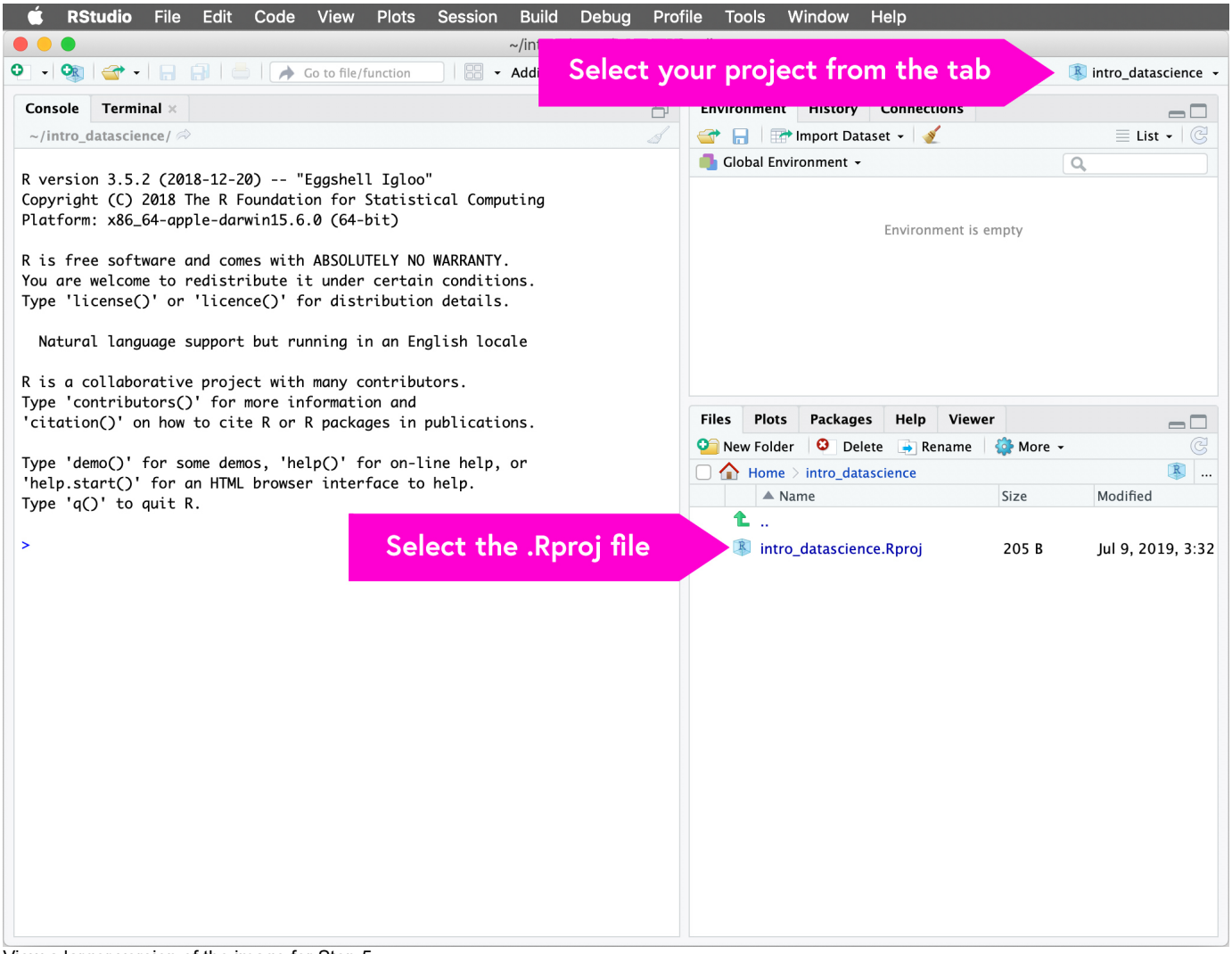


View a larger version of the image for Step 4. (https://ugc.futurelearn.com/uploads/assets/8a/bb/8abb2ac8-62d9-4b6e-b9ee-972c8c034a11.jpg)

Step 5: Well done, you've now created a RStudio project!

Your RStudio will have a projects tab on the upper right hand corner. Remember, every time you start to work on something for this course, be sure to open this project!

Whenever you are ready to write code for the course go to your project folder and then select the **'.Rproj'** file. This will automatically open RStudio and take you to the right directory!



View a larger version of the image for Step 5. (https://ugc.futurelearn.com/uploads/assets/d4/6c/d46cb5de-320d-4fc0-80e8-0eeb61f16049.jpg)

Installing packages

Packages are the way R users share useful code. You can think of each R package as a book. Once you've installed a package, you can load the code contained in it using **`library`** - which is like checking out a book from the library!

There are more than 14,000 packages available on CRAN (https://cran.r-project.org/) contributed by a range of R users, from experts to those relatively new to R. There are another several hundred on the Bioconductor (https://www.bioconductor.org/) archive, which focuses primarily on bioinformatics applications. There are many more on people's Github pages that may be in development. For example: Earo Wang's `mists` package (https://github.com/earowang/mists)

As an example, you can install the **`rmarkdown`** package for making reproducible reports using the following code chunk:

```
install.packages("rmarkdown")
```

> The **`tidyverse`** of packages contains many useful functions for data analytics, but to use these functions you have to load **`tidyverse`** into your R session. A good rule of thumb is to load **`tidyverse`** before you begin any analysis. To do this, place the code: **`library(tidyverse)`** , inside the **first** code chunk of your R Markdown file.

Give it a go!

Continue to develop your skills in RStudio by making your way through this exercise. Follow the instructions specified in **'Creating a new RStudio project'** to create another RStudio project that will contain all your materials as you work through this course.

Then, use the console to install the `tidyverse` (https://www.tidyverse.org/) and `visdat` (http://visdat.njtierney.com/) packages. You'll use these packages throughout the course to read, visualise and analyse data.

After you've installed the packages, run `library(tidyverse)` at the console, then, run the following code chunk:

```
glimpse(diamonds)
filter(diamonds, carat <= 2.5)
```

- **What happened when you ran `library(tidyverse)` at the console?**
- **Do you think that you'll always need to run that code to use the `tidyverse` ?**
- **What happened when you ran the code chunk?**

Once you've set up your workflow, within the **Comments**, share with other learners what you've learned from creating a project, installing packages and running code chunks, for example, what was your experience of working in the console, and were you able 'to do' data science?

Also consider reading and commenting on or **'Liking'** contributions made by other learners or following learners with similar interests as you.
