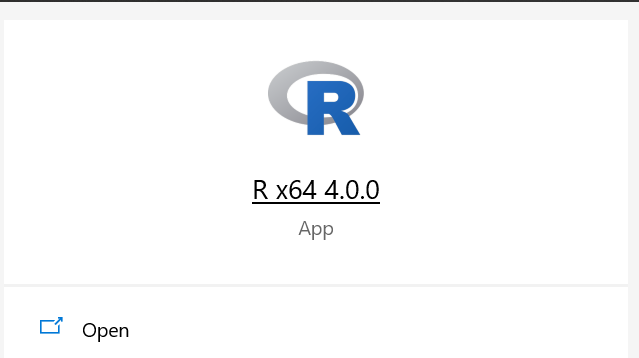
# If you don’t have R in your computer, install it following these instructions.

If you do not have R installed in your computer, please follow the procedures below.

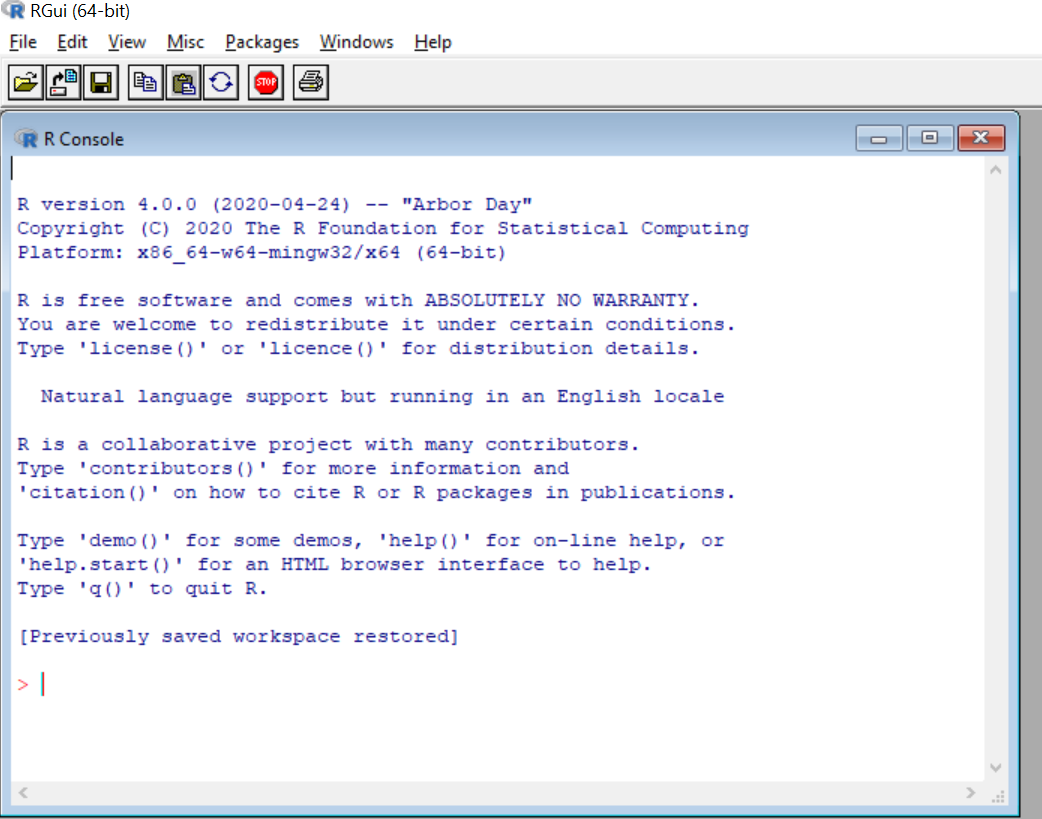
* Execute the exe installR.exe included in the distribution folder and follow the instructions from the installation wizard. This process shouldn’t take more than 2 minutes.
* Once it is done, just run R.



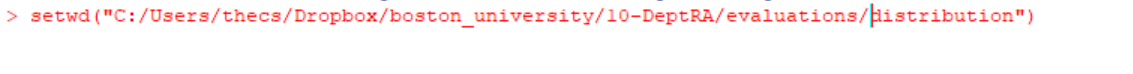
Now let’s get into the scanning

# Preliminary steps

1. Place the pdf to scan in the folder **distribution/to\_scan**
2. Open R for windows

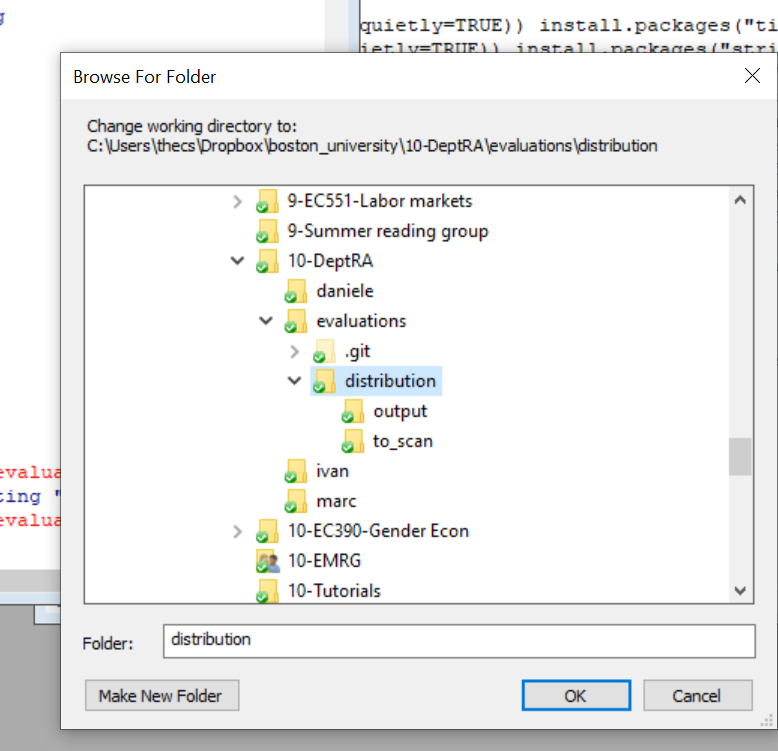


1. Set the working directory to the folder distribution folder. There are two ways of doing this:
   1. Running: **setwd(“[path\_to\_folder]/distribution”).** To indicate de path, use “/” or “\\” instead of “\”.

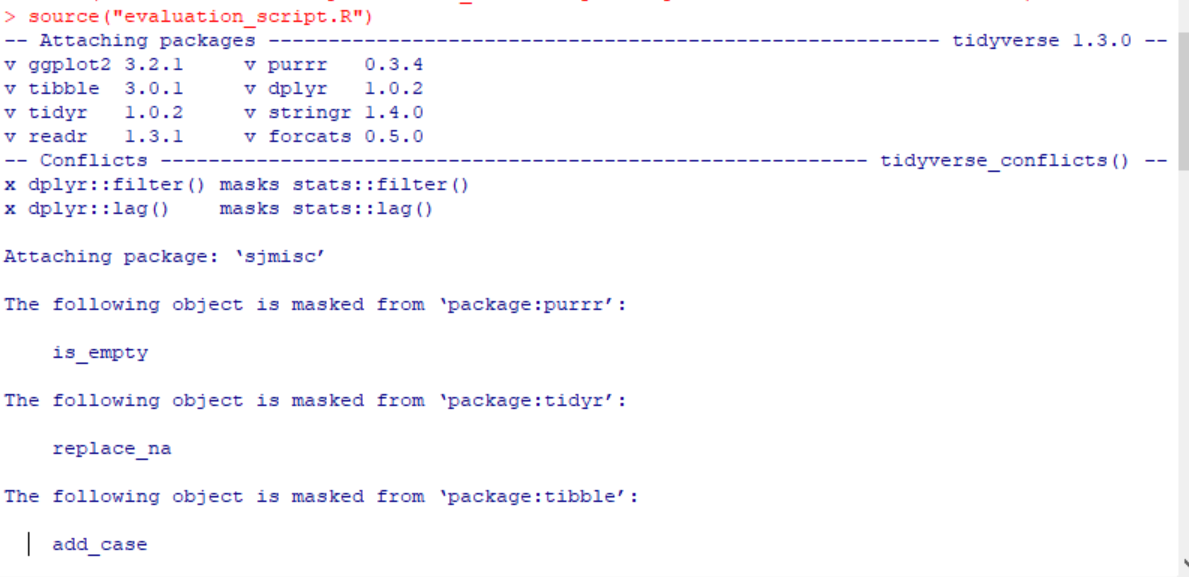
****

Make sure to write the path in quotes.

* 1. Go to File > Change dir and look for the distribution folder. Once you find it, click on OK

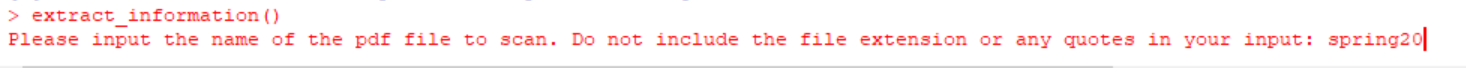
****

1. In the command line write **source("evaluation\_script.R")** and press enter. The console will likely print some warnings in blue. Just ignore them.

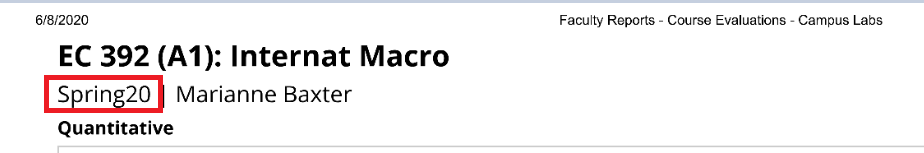


# Scanning

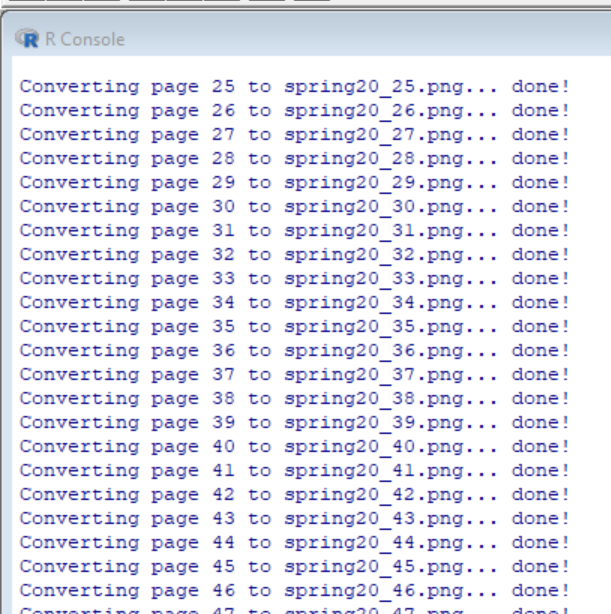
1. To start the scanning process write **extract\_information()** and press enter.
   1. Write the name of the file to scan. In my case, the pdf file is named spring20. Note that the program assumes that pdf is searchable. If this is not the case, you need to run the pdf though Acrobats’s OCR first.



* 1. Write the semester to scan. The “semester name” is indicated in the evaluation pdf in the snapshot below.

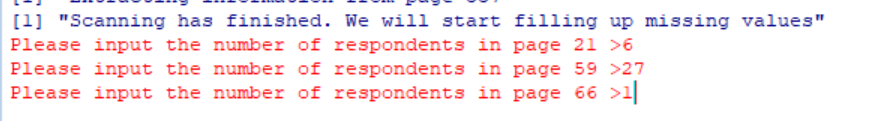


1. After doing this, the scanning process will start. This can take up to 30 minutes.
   1. The first time you scan a pdf file, the program will the program will create a bunch of png files in the folder to\_scan. There’s one png file for each page of the pdf. These png files are used to extract the course section. The program will inform on the progress of the pdf to png file conversion.



* 1. If the png files already exist (because for whatever reason this is the second time you are scanning a file), you can speed up the process by running **extract\_information(convert=FALSE)**

1. Once the extracting of the information is done. The program checks whether there is any information that was erroneous, or unable to extract. If there are errors, it will prompt you to input manually the missing data.
   1. The prompt indicates both the page number from the **pdf file** and the missing data point that you must input.
   2. If you made a mistake in the manual inputting, **do not stop the program.** It is more efficient to continue until the end and correct the mistake in the resulting csv file. The csv file includes a column called “complete”. All the lines with a FALSE in this column were manually inputted.



1. Once the scanning + manual input is done, the program will create the file **output/final\_database.csv**. Ignore the “NA introduced by coercion” warnings.

