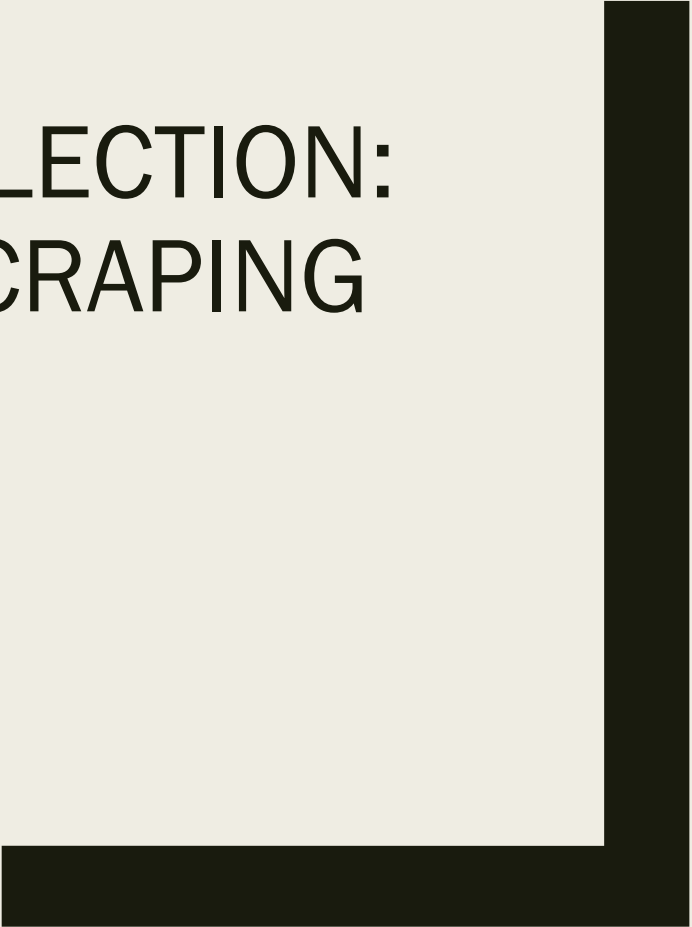




# INTERNET-BASED DATA COLLECTION: FUNDAMENTALS OF WEBSCRAPING

Charlene Mae Celoso  
Assistant Professor  
UP School of Statistics



# Training Outline

## 1. Introduction

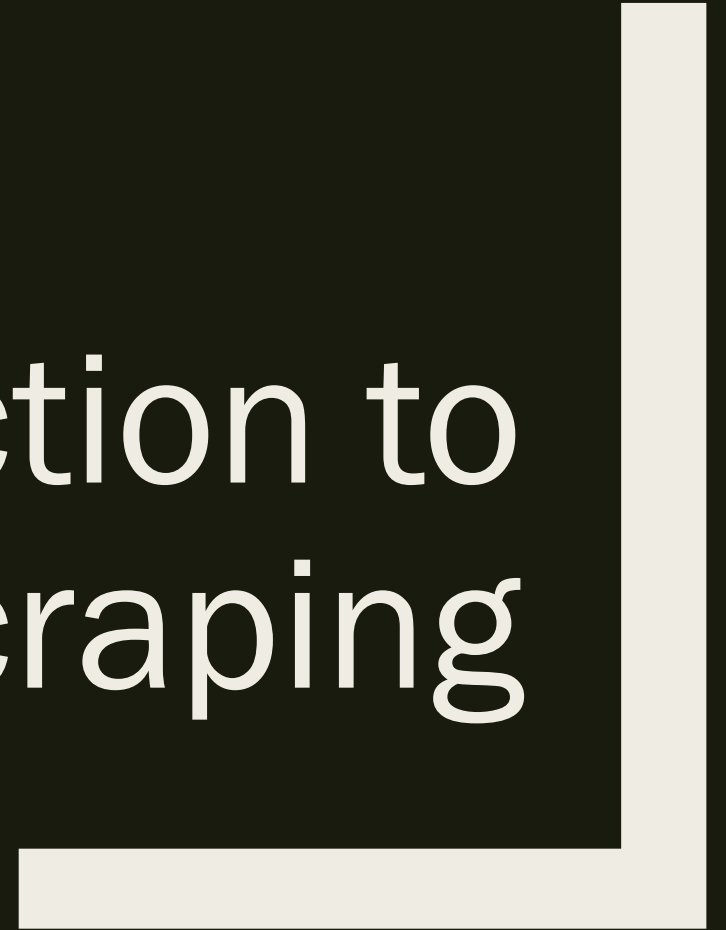
- Why Use Webscraping?
- Possible Sources of Data

## 2. Webscraping in MS Excel Using "Web Query"

## 3. Webscraping in R

- Introduction to CSS Selectors
- Using the "rvest" Package for Webscraping

# Introduction to Webscrapping



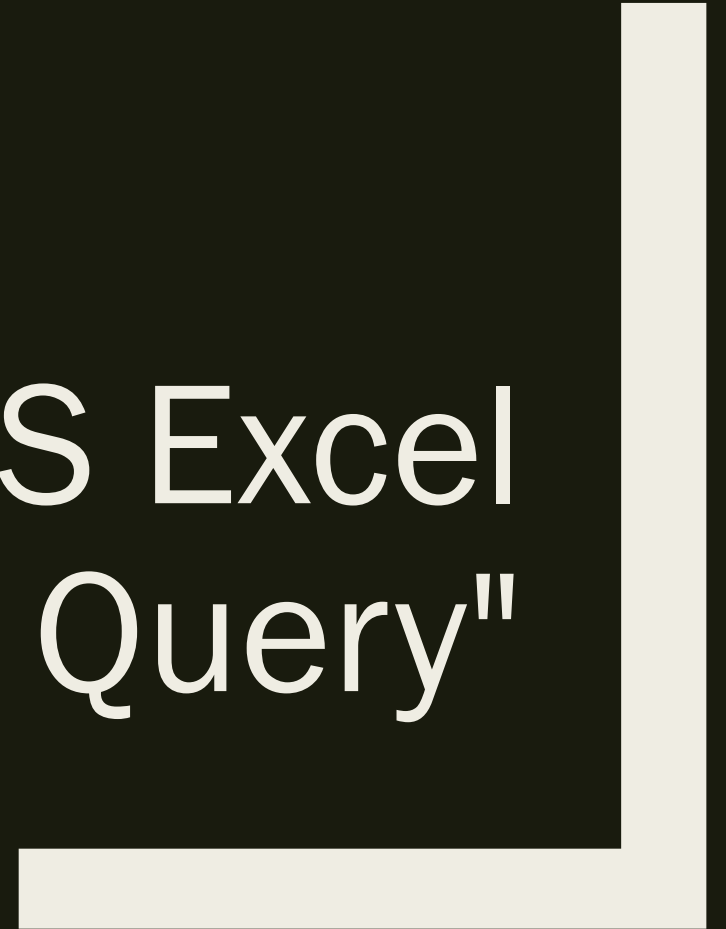
# Webscraping

- a technique employed to extract data from websites whereby the data is extracted and saved to a local file in your computer (typically, in a spreadsheet)
- helpful in cases where manually copying and pasting data is tedious

# Possible Sources of Data

- basically, almost any website!
- The task becomes easier when data is organized as a table.

# Webscraping in MS Excel Using "Web Query"

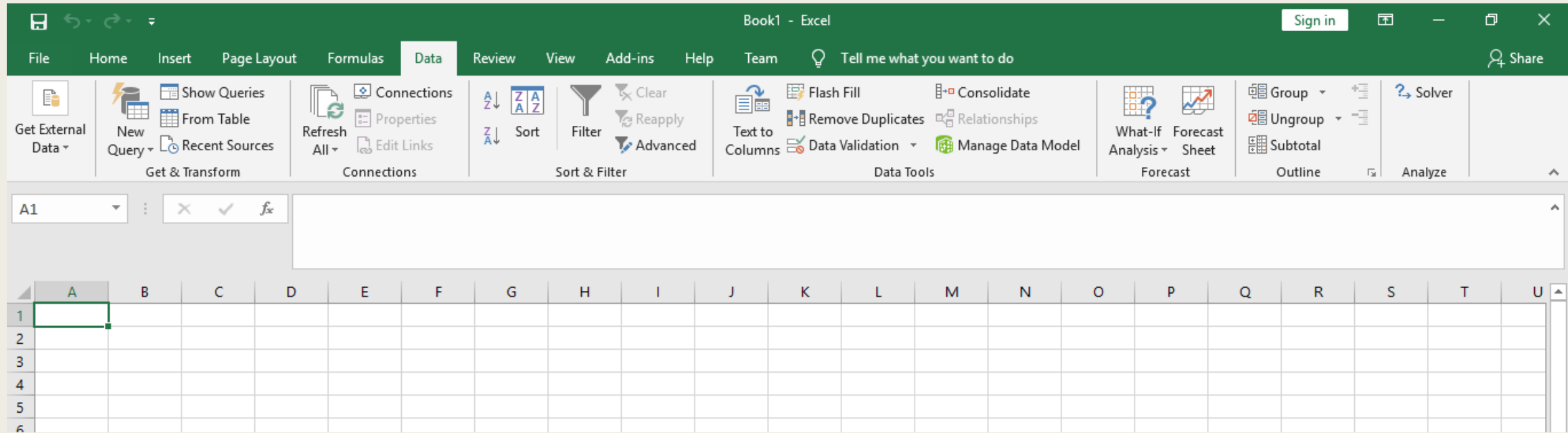


# Web Query

- a tool in MS Excel which can fetch data from web pages
- it automatically finds all the tables from a particular web page

# Steps

Go to the “Data” pane.

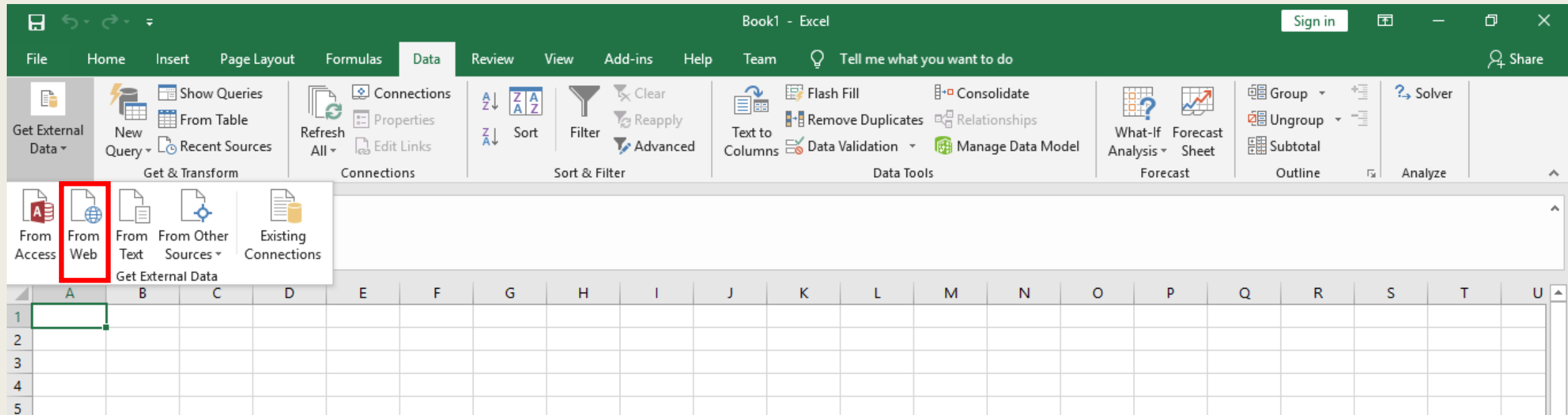






# Steps

Select “From Web.” A dialog box will appear.



# Steps

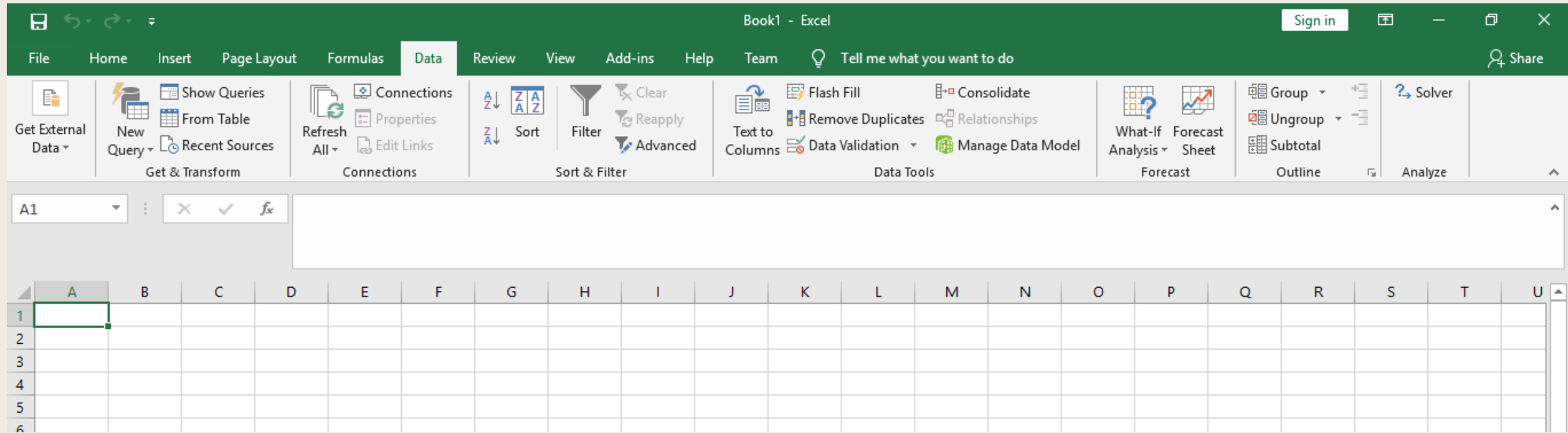
- In the dialog box, enter the URL of the web page where you want to fetch data.
- Once the page loads, select all tables that you want to get.

# Another Option: Power Query

- also known as “Get and Transform”
- can fetch data from various sources (not just limited to webpages)

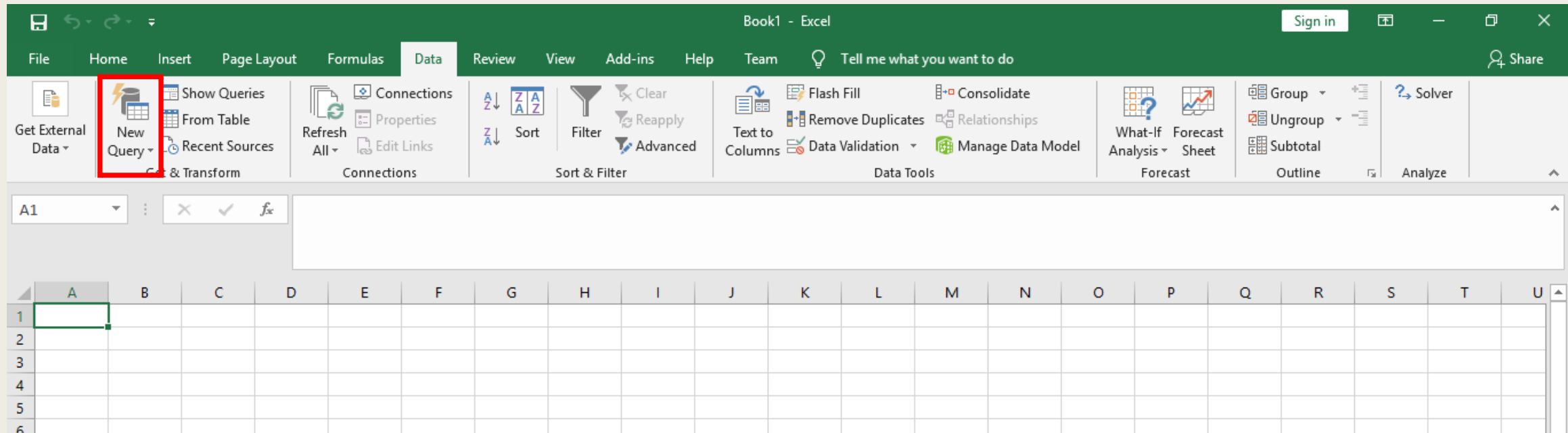
# Steps

Go to the “Data” pane.



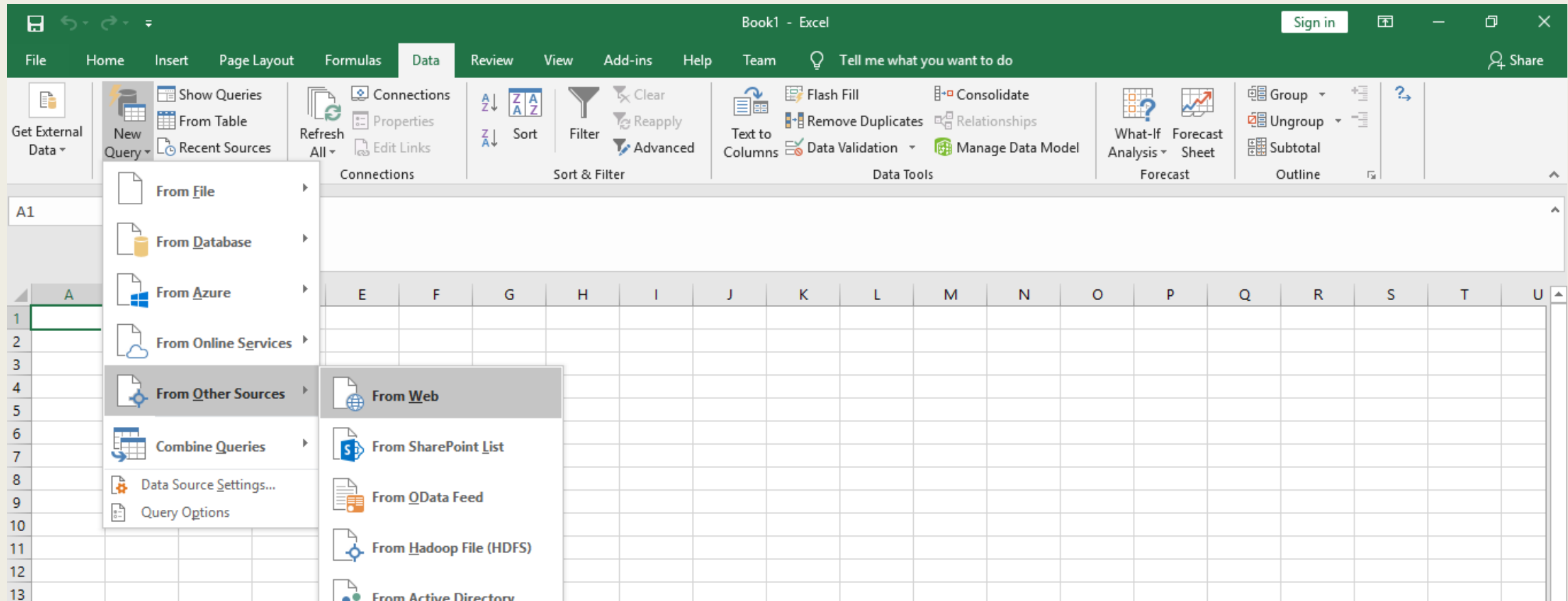
# Steps

Select “New Query.”



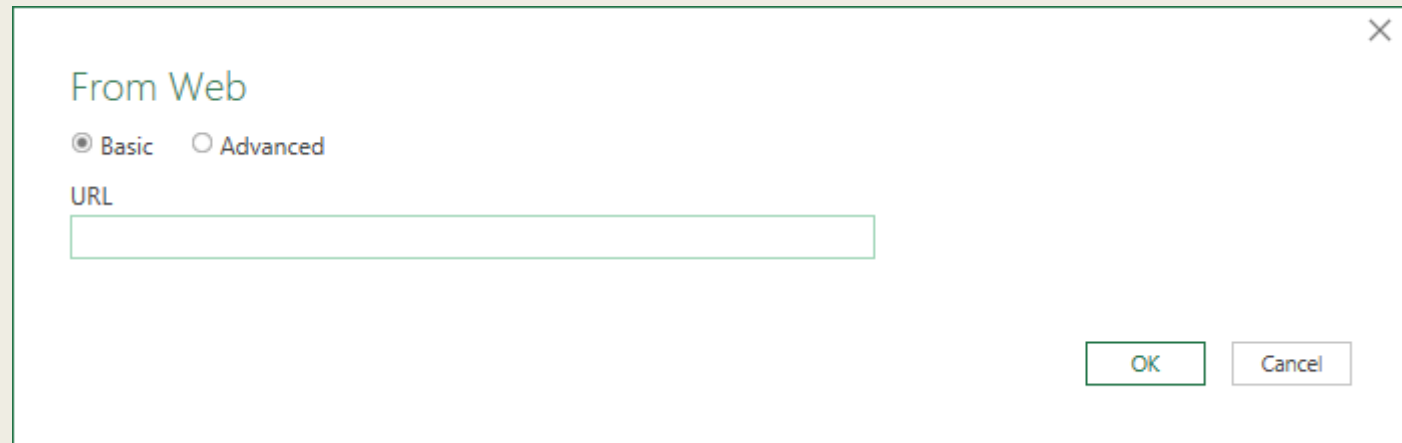
# Steps

Select “From Other Sources” > “From Web”



# Steps

A dialog box will appear. Enter the URL of the web page where you want to fetch data.



The image shows a dialog box titled "From Web" with a close button (X) in the top right corner. Below the title, there are two radio buttons: "Basic" (selected) and "Advanced". Below the radio buttons, there is a label "URL" followed by a text input field. At the bottom right of the dialog box, there are two buttons: "OK" and "Cancel".



# Steps

Select the table/s that you want to load in your spreadsheet. You may modify the entries (if necessary).

# Main Limitation of Using MS Excel

- You may only get data from web pages that contain tables.
- Hence, another tool is necessary.

# Exercise

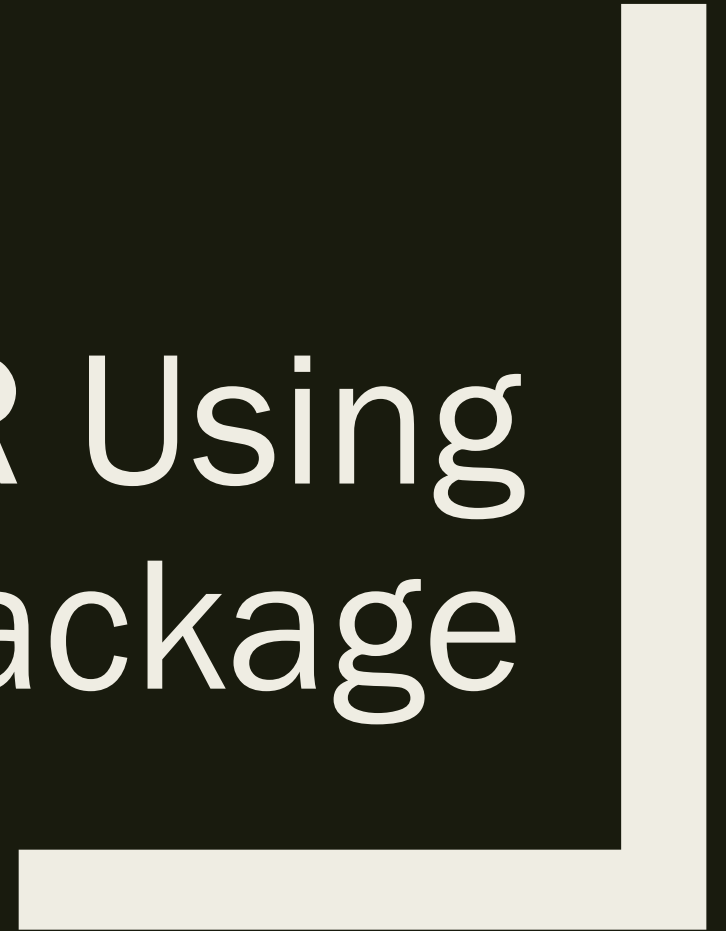
- Get the Dam Water Level Update from the PAGASA website:

<http://bagong.pagasa.dost.gov.ph/flood/>

# Quick Example: Transforming Data

- [https://en.wikipedia.org/wiki/List\\_of\\_cities\\_in\\_the\\_Philippines](https://en.wikipedia.org/wiki/List_of_cities_in_the_Philippines)

# Webscraping in R Using the *rvest* Package



# R and RStudio: An Introduction

- R: a free software environment for statistical computing and graphics
- RStudio: an integrated development environment (IDE) for R (basically, it provides an interface to code more easily using R)

In order for RStudio to work, R must be installed first.

# Installing R and RStudio

## R

- Go to <https://cran.r-project.org/>. Download R according to your operating system (Windows/Linux/Mac)
- During installation, just keep the default settings.

## RStudio

- Go to <https://www.rstudio.com/products/rstudio/download/>. Download the free version of “RStudio Desktop” that suits your operating system.
- During installation, just keep the default settings.

# If you can't install *rvest*...

- Go to **rstudio.cloud** then sign up for an account.
- Go to

<https://rstudio.cloud/project/485493>



# CSS Selectors: An Introduction

- CSS is a language that describes how HTML elements should be displayed. (HTML is the standard markup language for web pages.)
- CSS selectors define the elements to which a set of CSS rules apply.
- These CSS selectors are used in *rvest* to identify what we want to scrape. It enables us to obtain data beyond those that appear in tables.
- The *SelectorGadget* extension in Google Chrome may be used to easily identify CSS selectors.

# CSS Selectors: An Introduction

- The SelectorGadget extension in Google Chrome may be used to easily identify CSS selectors.

- *To install this, go to:*

- <https://chrome.google.com/webstore/category/extensions>*

# CSS Selectors: An Introduction

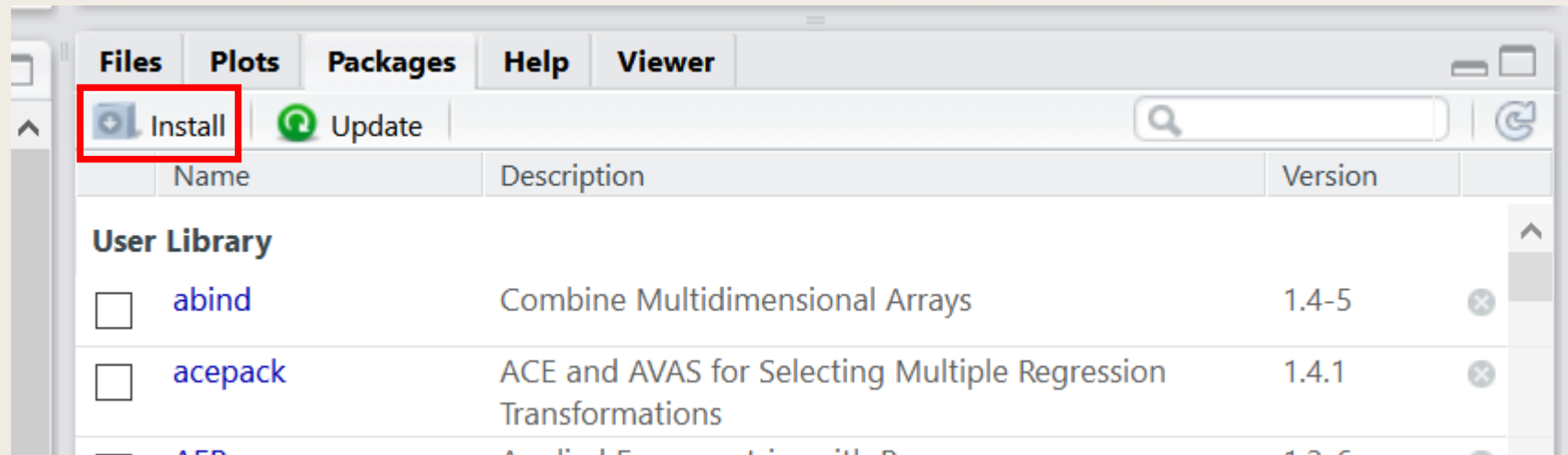
- The SelectorGadget extension in Google Chrome may be used to easily identify CSS selectors.

- *To install this, go to:*

- [\*https://chrome.google.com/webstore/category/extensions\*](https://chrome.google.com/webstore/category/extensions)

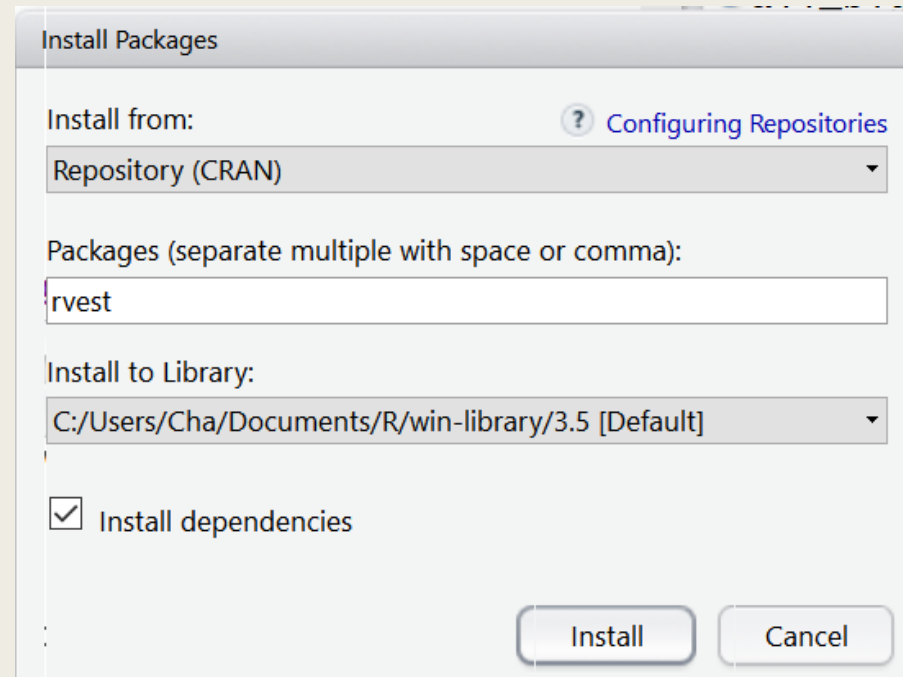
# Using *rvest*

Install the package: look for the “Install” button at the “Packages” tab on the lower right portion of the screen.



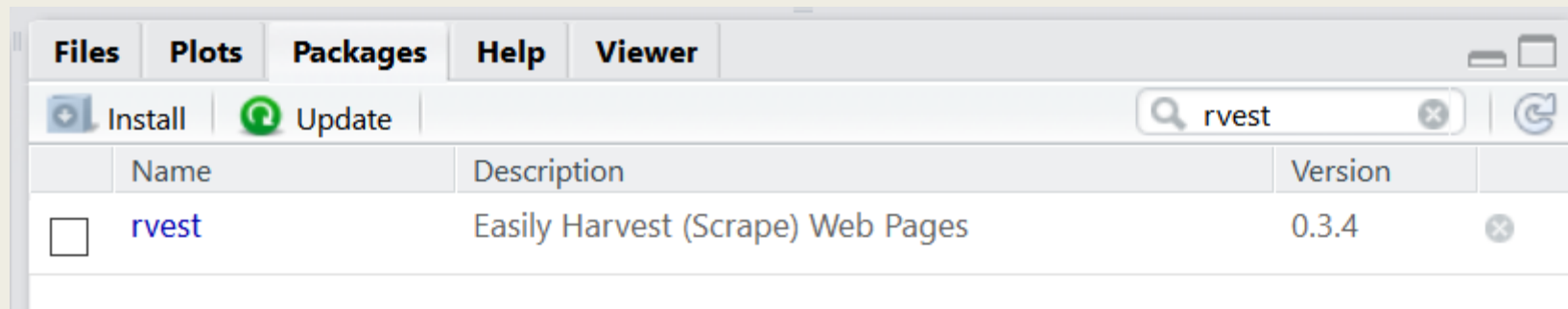
# Using *rvest*

A dialog box will appear. Input “rvest” as the package to be installed.



# Using *rvest*

Wait for the installation to be finished. Once *rvest* is successfully installed, you should be able to find it under the “Packages” tab.

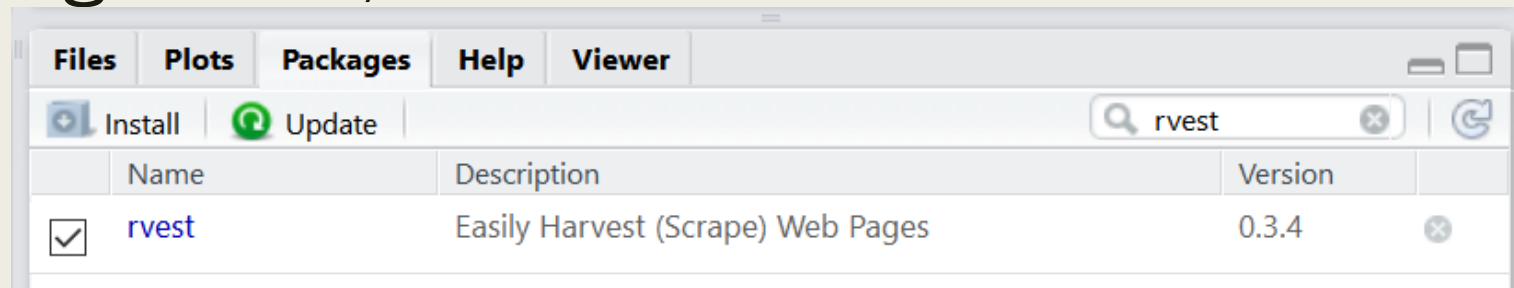


# Using *rvest*

Installing a package does not mean that it is ready for use! We still have to load it.

To load the *rvest* package, either

- Click the checkbox beside its name under the “Packages” tab,



- or run “library(rvest)”

# Using *rvest*

## Typical Workflow:

1. `read_html()`: to read all HTML codes from the webpage
2. `html_nodes()`: to scrape all elements that correspond to a particular CSS selector
3. `html_text()`: to get the actual contents

Further data processing may be necessary after step 3.



# Demonstration

[https://www.imdb.com/search/title/?groups=top\\_100  
&count=100](https://www.imdb.com/search/title/?groups=top_100&count=100)

The website above shows the Top 100 movies in IMDb (Internet Movie Database).

# Demonstration: Scraping Tables

[https://en.wikipedia.org/wiki/List\\_of\\_cities\\_in\\_the\\_Philippines](https://en.wikipedia.org/wiki/List_of_cities_in_the_Philippines)

# EXTRA: Scraping Text from Images

- “Optical Character Recognition”

- Website:

**ocr.space**

# Exercise

- Obtain the titles of the news articles in The Philippine Star that result from searching “Philippine Statistics Authority.”

<https://www.philstar.com/search/philippine%20statistics%20authority>