IT497 Week 3 Lab

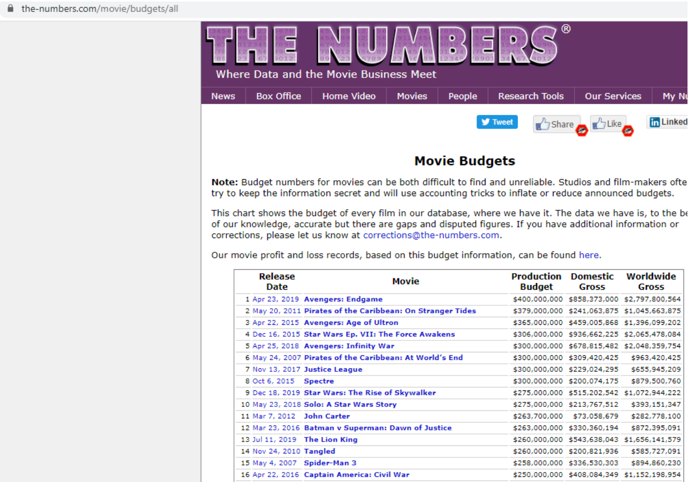
Web Scraping using R and `rvest`

About the Data

Nash Information Services' (https://www.the-numbers.com/) publishes revenue estimates for past movies, analyses for future releases, and many other reports that can be critical to planning an investment or creating a compelling business plan for a movie.

For our lab, we will be using Nash's Movie Budgets (https://www.the-numbers.com/movie/budgets/all)

This chart shows the budget of every film in Nash's database.



Movie Budgets

Note: Budget numbers for movies can be both difficult to find and unreliable. Studios and film-makers ofte try to keep the information secret and will use accounting tricks to inflate or reduce announced budgets.

This chart shows the budget of every film in our database, where we have it. The data we have is, to the be of our knowledge, accurate but there are gaps and disputed figures. If you have additional information or corrections, please let us know at corrections@the-numbers.com.

Our movie profit and loss records, based on this budget information, can be found here.

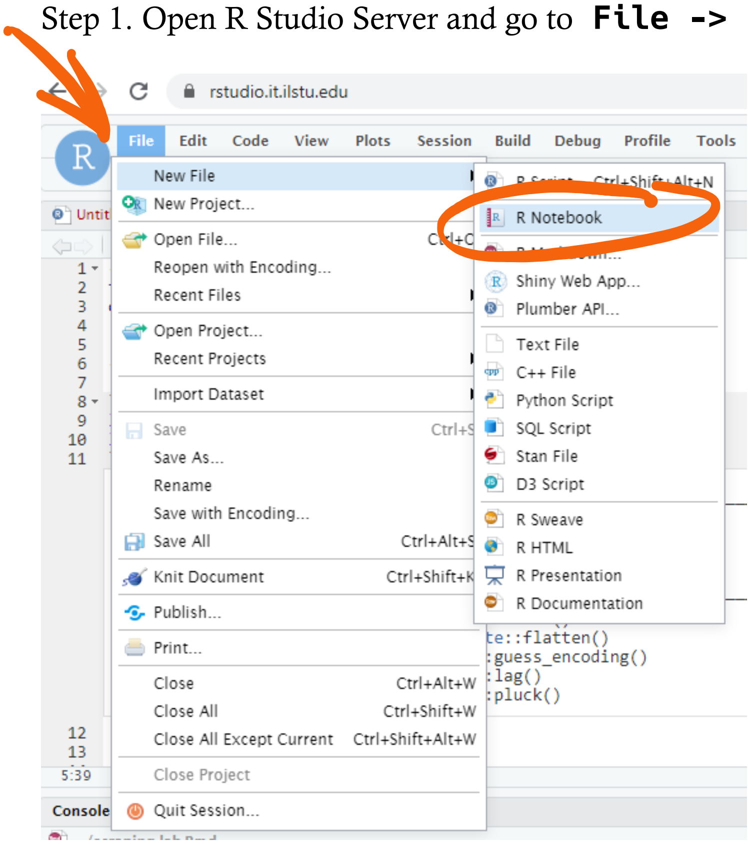
	Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1	Apr 23, 2019	Avengers: Endgame	$400,000,000	$858,373,000	$2,797,800,564
2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$379,000,000	$241,063,875	$1,045,663,875
3	Apr 22, 2015	Avengers: Age of Ultron	$365,000,000	$459,005,868	$1,396,099,202
4	Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,065,478,084
5	Apr 25, 2018	Avengers: Infinity War	$300,000,000	$678,815,482	$2,048,359,754
6	May 24, 2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$963,420,425
7	Nov 13, 2017	Justice League	$300,000,000	$229,024,295	$655,945,209
8	Oct 6, 2015	Spectre	$300,000,000	$200,074,175	$879,500,760
9	Dec 18, 2019	Star Wars: The Rise of Skywalker	$275,000,000	$515,202,542	$1,072,944,222
10	May 23, 2018	Solo: A Star Wars Story	$275,000,000	$213,767,512	$393,151,347
11	Mar 7, 2012	John Carter	$263,700,000	$73,058,679	$282,778,100
12	Mar 23, 2016	Batman v Superman: Dawn of Justice	$263,000,000	$330,360,194	$872,395,091
13	Jul 11, 2019	The Lion King	$260,000,000	$543,638,043	$1,656,141,579
14	Nov 24, 2010	Tangled	$260,000,000	$200,821,936	$585,727,091
15	May 4, 2007	Spider-Man 3	$258,000,000	$336,530,303	$894,860,230
16	Apr 22, 2016	Captain America: Civil War	$250,000,000	$408,084,349	$1,152,198,954

For our lab, we are only going to scrape the top 100.

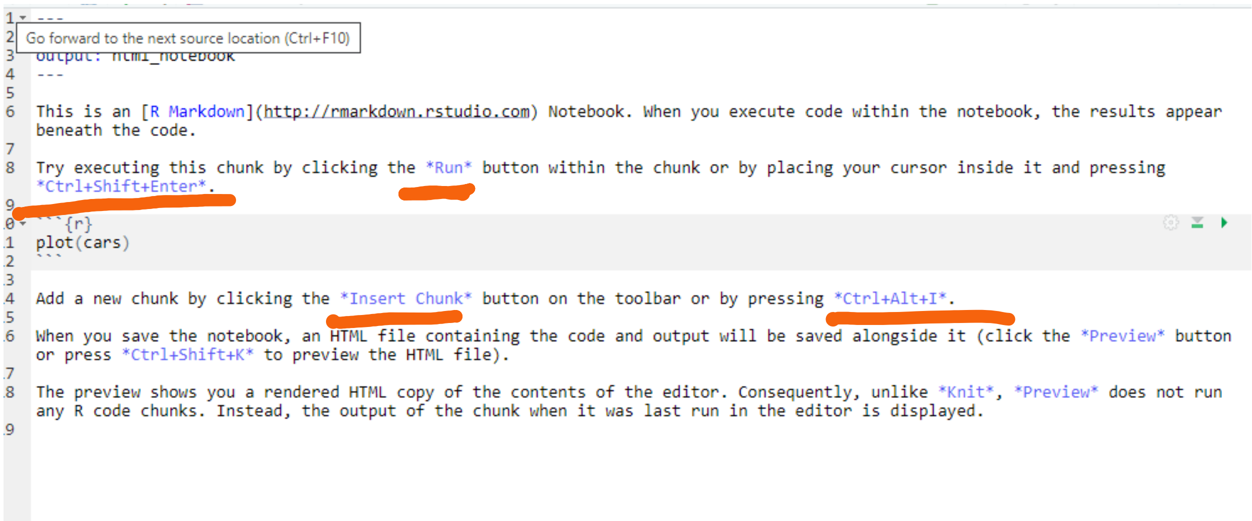
Note: for extra credit, you may scrape additional movies. Instructions for the extra credit follow the lab instructions.

Following are the steps you need to follow:

Step 1. Open R Studio Server and go to **`File -> New File -> R Notebook`**

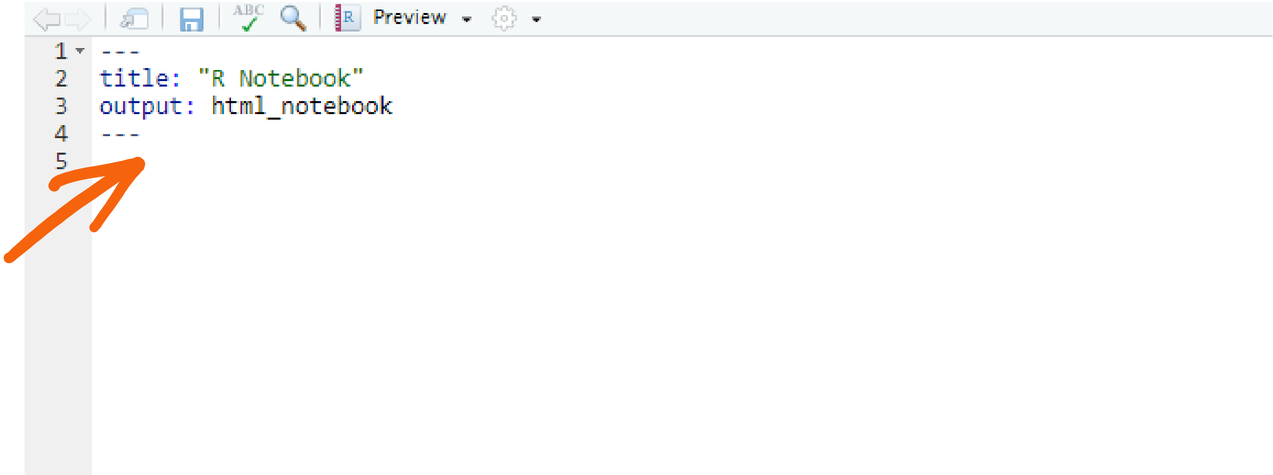


Step 2. R Studio will create a template file like this. Read through the text and pay special attention to everything in blue. That is how to "Run" code chunks and how to "Insert Chunk".



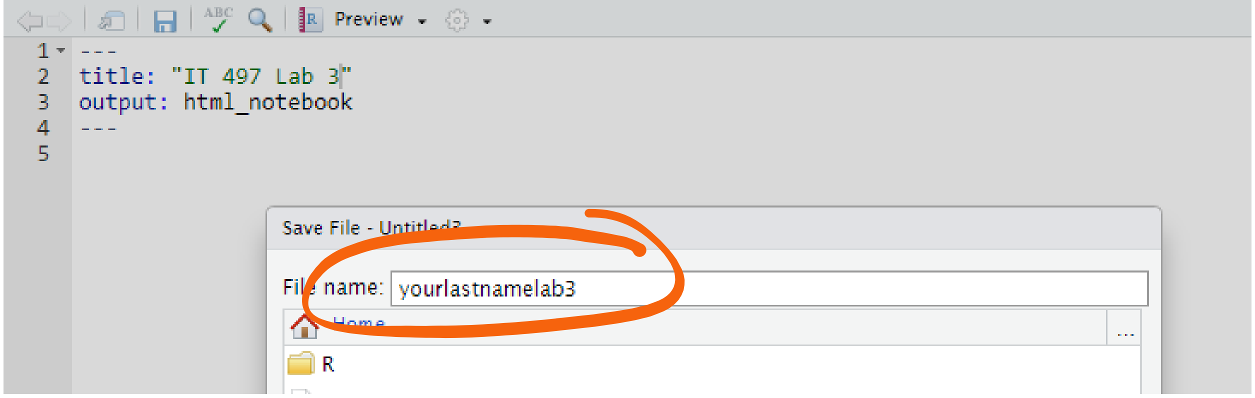
~~~
---

output: html_notebook
---

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```{r}
plot(cars)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.
~~~

Step 3. Delete everything below the 3 dashes:



```
---
title: "R Notebook"
output: html_notebook
---

```

Step 4. Change the title to `IT497 Lab 3` and save the file as `yourlastnamelab3.Rmd`



```
---
title: "IT 497 Lab 3"
output: html_notebook
---

```

File name: yourlastnamelab3

Step 5. Insert a new R Code Chunk

If you have not already installed the `tidyverse`, do it in the 1st code chunk like this:

```
# Install from CRAN

install.packages("tidyverse")
```

The tidyverse packages are the packages at heart of data science/research using R. We will learn much more about many of these packages throughout the semester. However, if you want to do a little reading on your own, a great place to start is here: https://www.tidyverse.org/

If you have not already, you will also need to install `rvest`.

```
install.packages("rvest")
```

rvest helps you scrape information from web pages. It is designed to make it easy to express common web scraping tasks, inspired by libraries like beautiful soup. For more information about rvest, please visit: https://github.com/tidyverse/rvest

A nice rvest tutorial can be found here: https://www.datacamp.com/community/tutorials/r-web-scraping-rvest

Please note that you only need to do this ONE TIME THIS SEMESTER.
After this time, you will just load the tidyverse and revest packages using the library command. Insert the following R code into your code chunk. In code chunk #1, you should have the following R code:

```
library(tidyverse)
library(rvest)
```

Step 6. Insert a new R Code Chunk. In code chunk #2, you should have the following R code:

```
base_url <- "https://www.the-numbers.com/movie/budgets/all"
base_webpage <- read_html(base_url)
```

Step 7. Insert a new R Code Chunk. In code chunk #3, you should have the following R code:

```
new_urls<- "https://www.the-numbers.com/movie/budgets/all/%s"

table_base <- rvest::html_table(base_webpage)[[1]] %>%
  tibble::as_tibble(.name_repair = "unique") # repair the repeated columns
```

Step 8. Insert a new R Code Chunk. In code chunk #4, you should list the first 10 movies (you know how to do this).

Step 9. Insert a new R Code Chunk. In code chunk #5, write your data to a *.csv file. You will submit this file via ReggieNet.

In code chunk #5, you should have the following R code:

```
write.csv(table_base,"moviesData.csv")
```

YOU SHOULD BE ABLE TO KNIT YOUR RMD FILE TO EITHER HTML OR WORD.

Extra Credit:

There are 2 extra credit opportunities this week

You may do both to earn up to 5 points extra credit,

Up to 3 additional extra credit points
In this week's lab, we scraped 100 movie records. Our lab was based off of this R Blogger's Tutorial: https://www.r-bloggers.com/tutorial-web-scraping-of-multiple-pages-using-r/

The bottom of the tutorial explains how to read 100's of additional movie records.

For up to 3 additional extra credit points, scrape an additional 600 movie records. Submit the 600 records following the lab's instructions (there should just be more records).

Up to 2 additional extra credit points
This video https://vimeo.com/110804387 explains how to use a Word Template to format your knitted output.

For up to two points, create a word template that makes the R code in your knitted word document **9-point Consolas.** Using the template when creating your output and submit your template via ReggieNet along with the other lab materials.

Deliverables: Upload the following to ReggieNet before the due date/time.

1. Your `moviesData.csv file`
2. `Your *.Rmd file`
3. `Either a Word or HTML file knitted from your *.Rmd file`
4. `Any extra credit files you want to include,`