ISA 401: Business Intelligence & Data Visualization

05: Lab - R Markdown and Web Scraping

Fadel M. Megahed, PhD

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University

Twitter: FadelMegahed
GitHub: fmegahed
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Virtual Office Hours

Spring 2022

Quick Refresher from Last Week

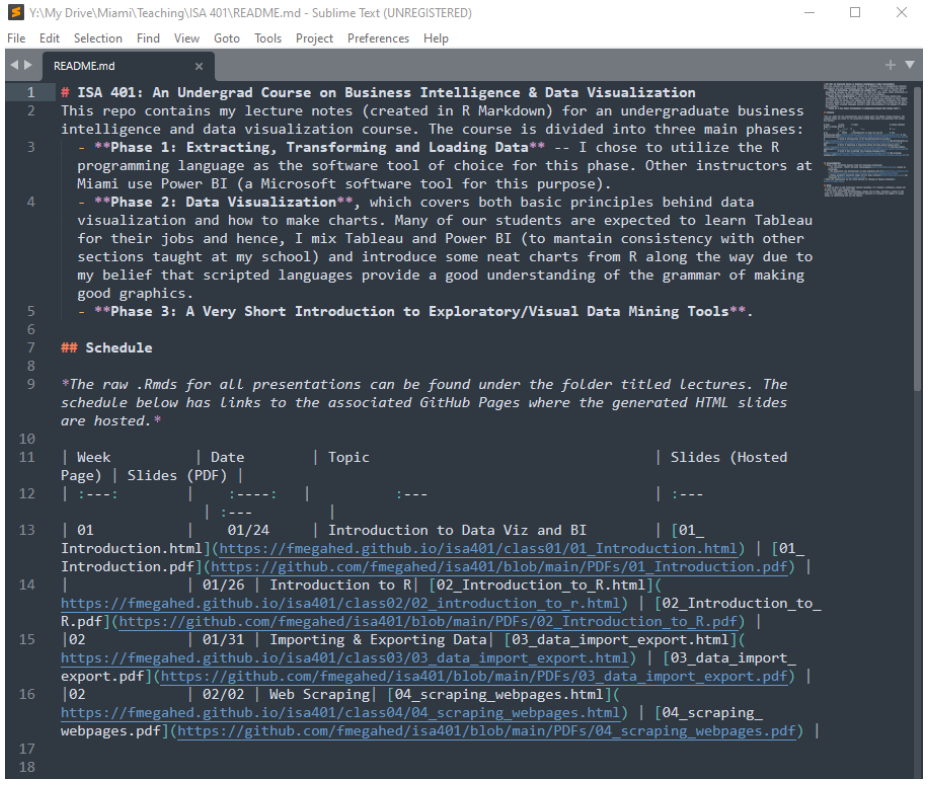
- ✅ Subset data in R.
- ✅ Read text-files, binary files (e.g., Excel, SAS, SPSS, Stata, etc), json files, etc.
- ✅ Export data from R.
- ✅ Understand when can we scrape data (i.e., `robots.txt`)
- ✅ Scrape a webpage Using R
- ✅ Utilize loops or `purr::map` to download data from multiple webpages.

Learning Objectives for Today's Lab

- Perform an extensive web scrape
- Create reproducible reports/analyses using R Markdown

Markdown .md

- A **markup language** to add formatting elements to plain-text documents, contrasting to **WYSIWYG**



.md Basic Syntax

Element	Markdown Syntax

Heading	`# H1`
	`## H2`
	`### H3`

Bold	`**bold text**`
Italic	`*italicized text*`
Blockquote	`> blockquote`

Ordered List	`1. First item`
	`2. Second item`
	`3. Third item`

Unordered List	`- First item`
	`- Second item`
	`- Third item`

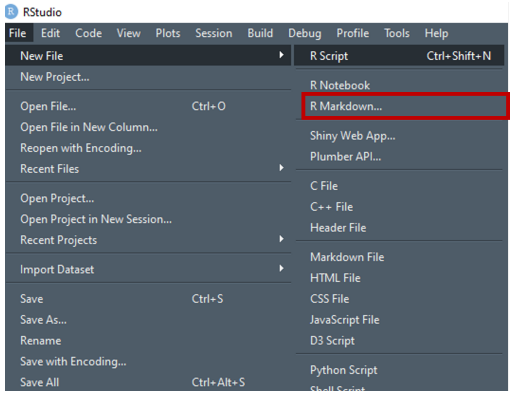
Code	`` `code` ``
Horizontal Rule	`---`
Link	`[title](https://www.example.com)`
Image	`![alt text](image.jpg)`

Note: The table is captured from markdown cheatsheet. More details can be found at markdown syntax.

Rmarkdown


%>%
ggplo
readr
for
@allison_horst
YAML
rmarkdown
.Rmd
knitr
.md
pandoc

Create R Markdown .Rmd




````
---
title: "ISA 401: R Markdown Demo"
author: "Fadel Megahed"
date: "February 06, 2022"
output: html_document
---

```{r setup, include = FALSE}
library(knitr)
opts_chunk$set(echo = TRUE)
```

## R Markdown

You can embed an R code chunk like this:

```{r yelp_reviews, message = FALSE}
library(tidyverse)
patterson_reviews <- read_rds("https://ti
patterson_reviews
```

See <http://rmarkdown.rstudio.com>.
````

R Markdown Cheatsheet [1]

R Markdown Cheatsheet [2]

Demo

In class, we will do a live demo of some of the basic functionalities and settings of R Markdown.

Lab 01

Lab 01 (Group Assignment 03 Questions)

> **In your pre-assigned group**, please submit the **HTML knitted** from the Markdown template titled 05_web_scraping_lab_template.Rmd on Canvas
>
> **This submission is due Friday February 11, 2022 at 11:59 PM.**

Q1 Q2 Q3 Q4 Q5

- Click on Found and Impounded Property Listing on the Property - Lost, Found, Impounded Page
- Go to the Google Doc containing a table of lost and found items.
- **Please scrape the table and print it out.**

Lab 01 (Group Assignment 03 Questions)

> **In your pre-assigned group**, please submit the **HTML knitted** from the Markdown template titled 05_web_scraping_lab_template.Rmd on Canvas
>
> **This submission is due Friday February 11, 2022 at 11:59 PM.**

Q1 | Q2 | Q3 | Q4 | Q5

- Currently, the Farmer School of Business has seven academic departments (i.e., Accountancy, Economics, Entrepreneurship, Finance, Information Systems & Analytics, Marketing, and Management.
- Using the code chunk below, please write code that will produce and print a **single tibble containing information on ALL departments and the following variables:** (a) department name, (b) faculty/staff's name, (c) faculty/staff's position, and (d) faculty/staff's website.

Lab 01 (Group Assignment 03 Questions)

> **In your pre-assigned group**, please submit the **HTML knitted** from the Markdown template titled 05_web_scraping_lab_template.Rmd on Canvas
>
> **This submission is due Friday February 11, 2022 at 11:59 PM.**

Q1 Q2 Q3 Q4 Q5

- The most popular listings on Netflix are rated and reviews on ImDb. Based on this webpage and its following pages, please create a **tibble** that contains the following: *title*, *years*, *age classification*, *duration*, *genres*, *IMDb Rating*, *1-2 sentence summary*, *stars*, and *votes*.
- **Your tibble should contain a variable for the 9 items above for each of the 50 titles found on the page.**

Lab 01 (Group Assignment 03 Questions)

> **In your pre-assigned group**, please submit the **HTML knitted** from the Markdown template titled 05_web_scraping_lab_template.Rmd on Canvas
>
> **This submission is due Friday February 11, 2022 at 11:59 PM.**

Q1 Q2 Q3 **Q4** Q5

- Expand on the previous example to capture the top **300** titles on Netflix (i.e., the information across six pages).

Lab 01 (Group Assignment 03 Questions)

> **In your pre-assigned group**, please submit the **HTML knitted** from the Markdown template titled 05_web_scraping_lab_template.Rmd on Canvas
>
> **This submission is due Friday February 11, 2022 at 11:59 PM.**

Q1 Q2 Q3 Q4 Q5

- In assignment 02, I shared with you an RDS file containing four variables and all the reviews that were performed on Patterson Cafe on Yelp. Use what you have learned in class to potentially recreate the same results.

Recap

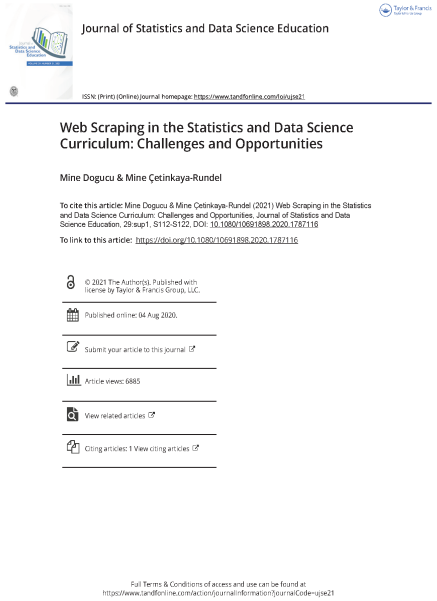
Summary of Main Points

By now, you should be able to do the following:

- Perform an extensive web scrape
- Create reproducible reports/analyses using R Markdown

Supplementary Reading on R Markdown

Supplementary Reading on Web Scraping



Journal of Statistics and Data Science Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ujse21

Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities

Mine Dogucu & Mine Çetinkaya-Rundel

To cite this article: Mine Dogucu & Mine Çetinkaya-Rundel (2021) Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities, Journal of Statistics and Data Science Education, 29:sup1, S112-S122, DOI: 10.1080/10691898.2020.1787116

To link to this article: https://doi.org/10.1080/10691898.2020.1787116

Published online: 04 Aug 2020.

- PDF of Published Paper
- ePub of Published Paper

- Selector Gadget
- Getting Started with rvest
- Practical Web Scraping in R