

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

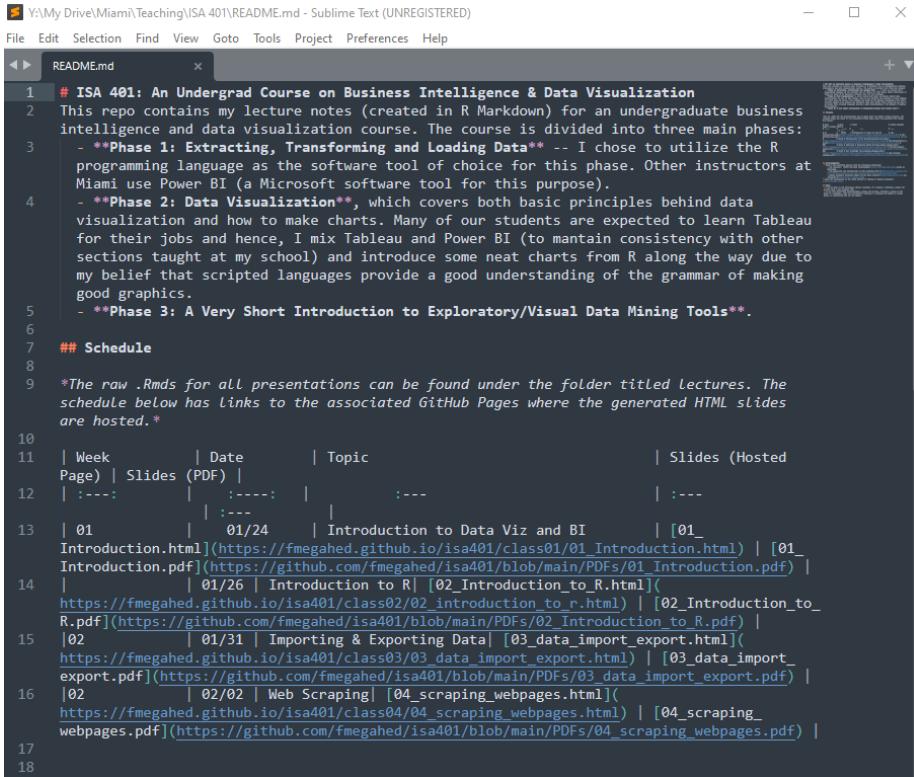
Learning Objectives for Today's Lab

- 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**



The screenshot shows a Sublime Text window with the file "README.md" open. The main pane displays the following R Markdown code:

```
1 # ISA 401: An Undergrad Course on Business Intelligence & Data Visualization
2 This repo contains my lecture notes (created in R Markdown) for an undergraduate business
3 intelligence and data visualization course. The course is divided into three main phases:
4 - **Phase 1: Extracting, Transforming and Loading Data** -- I chose to utilize the R
5 programming language as the software tool of choice for this phase. Other instructors at
6 Miami use Power BI (a Microsoft software tool for this purpose).
7 - **Phase 2: Data Visualization**, which covers both basic principles behind data
8 visualization and how to make charts. Many of our students are expected to learn Tableau
9 for their jobs and hence, I mix Tableau and Power BI (to maintain consistency with other
10 sections taught at my school) and introduce some neat charts from R along the way due to
11 my belief that scripted languages provide a good understanding of the grammar of making
12 good graphics.
13 - **Phase 3: A Very Short Introduction to Exploratory/Visual Data Mining Tools**.
14
15 ## Schedule
16
17 *The raw .Rmds for all presentations can be found under the folder titled Lectures. The
18 schedule below has links to the associated GitHub Pages where the generated HTML slides
19 are hosted.*
```

Below the code, there is a preview pane showing a table of contents for the course schedule:

Week	Date	Topic	Slides (Hosted Page)
01	01/24	Introduction to Data Viz and BI	[01_Introduction.html](https://fmegahed.github.io/isa401/class01/01_Introduction.html) [01_Introduction.pdf](https://github.com/fmegahed/isa401/blob/main/PDFs/01_Introduction.pdf)
02	01/26	Introduction to R [02_Introduction_to_R.html](https://fmegahed.github.io/isa401/class02/02_introduction_to_r.html) [02_Introduction_to_R.pdf](https://github.com/fmegahed/isa401/blob/main/PDFs/02_Introduction_to_R.pdf)	
03	01/31	Importing & Exporting Data [03_data_import_export.html](https://fmegahed.github.io/isa401/class03/03_data_import_export.html) [03_data_import_export.pdf](https://github.com/fmegahed/isa401/blob/main/PDFs/03_data_import_export.pdf)	
04	02/02	Web Scraping [04_scraping_webpages.html](https://fmegahed.github.io/isa401/class04/04_scraping_webpages.html) [04_scraping_webpages.pdf](https://github.com/fmegahed/isa401/blob/main/PDFs/04_scraping_webpages.pdf)	

QR .md Basic Syntax

Element	Markdown Syntax
Heading	# H1 ## H2 ### H3
Bold	bold text
Italic	<i>italicized text</i>
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)

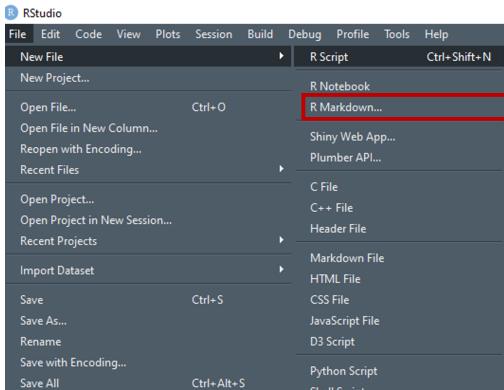
Note: The table is captured from [markdown cheatsheet](#). More details can be found at [markdown syntax](#).

Rmarkdown



@allison_horst

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

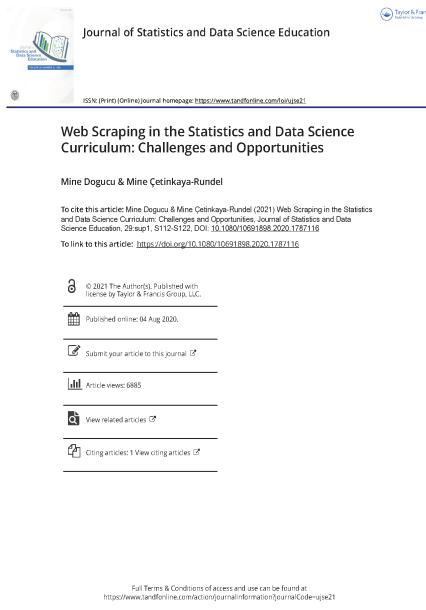
Summary of Main Points

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



- PDF of Published Paper
- ePub of Published Paper



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