

ISA 401: Business Intelligence & Data Visualization

04: Scraping Webpages in

Fadel M. Megahed, PhD

Endres Associate Professor
Farmer School of Business
Miami University

 @FadelMegahed



 fmegahed

 fmegahed@miamioh.edu


 Automated Scheduler for Office Hours

Spring 2024

Quick Refresher from Last Class

- ✓ Subset data in .
- ✓ Read text-files, binary files (e.g., Excel, SAS, SPSS, Stata, etc), json files, etc.
- ✓ Export data from .

Learning Objectives for Today's Class




- Understand when can we scrape data (i.e., `robots.txt`)
- Scrape a webpage using .

Web Technology

World Wide Web (WWW)

WWW (or the **Web**) is the information system where documents (web pages) are identified by Uniform Resource Locators (**URLs**)

A web page consists of:

-  **HTML** provides the basic structure of the web page
-  **CSS** controls the look of the web page (optional)
-  **JS** is a programming language that can modify the behavior of elements of the web page (optional)

Hypertext Markup Language (HTML)

- with the extension `.html`.
- rendered using a web browser via an URL.
- text files that follows a special syntax that alerts web browsers how to render it.

via a web browser

← → ↻ 🏠 🔒 Not secure | plane crash info.com/2021/2021.htm

Apps Teaching Research Misc MU Mail MU Calendar Overleaf Canvas

2021

| Date | Location / Operator | Aircraft Type / Registration | Fatalities |
|-----------------------------|---------------------------------------------------------------|-----------------------------------------|------------|
| 09 Jan 2021 | Near Jakarta, Indonesia Sriwijaya Air | Boeing 737-524 PK-CLC | 62/62(0) |
| 02 Mar 2021 | Pieri, Sudan South Sudan Supreme Airlines | Let L-410UVP-E HK-4274 | 10/10(0) |
| 28 Mar 2021 | Near Butte, Alaska Soloy Helicopters | Eurocopter AS350B3 Ecureuil N351SH | 5/6(0) |
| 21 May 2021 | Near Kaduna, Nigeria Military - Nigerian Air Force | Beechcraft B300 King Air 350i NAF203 | 11/11(0) |
| 10 Jun 2021 | Near Pailin, Myanmar Military - Myanmar Air Force | Beechcraft 1900D 4610 | 12/14(0) |
| 04 Jul 2021 | Patikul, Sulu, Philippines Military - Philippine Air Force | Lockheed C-130H Hercules 5125 | 50/96(3) |
| 06 Jul 2021 | Palana, Russia Kamchatka Aviation Enterprise | Antonov An 26B-100 RA-26085 | 28/28(0) |
| 12 Sep 2021 | Kazachinskoye, Russia Aeroservice/SILA | Let L-410UVP-E20 RA-67042 | 4/16(0) |
| 27 Dec 2021 | El Cajon, California Med Jet | Learjet 35A N880Z | 4/4(0) |

[Return to Home Page](#)

Copyright © Richard Kebabjian / www.plane crash info.com

via a text editor

```
1 <html>
2
3 <head>
4   <meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
5   <meta name="GENERATOR" content="Microsoft FrontPage 4.0">
6   <meta name="description" content="Aviation accidents">
7   <meta name="keywords" content="aircraft accident, plane crash, aviation disaster, safety, aviation safety, aviation accident,
8   aircraft, plane, statistics, airline statistics, airline, airlines, hijack, pilot, probable cause, crash, boeing, cockpit,
9   <meta name="ProgId" content="FrontPage.Editor.Document">
10  <meta name="Title" content="Aviations accidents 2021">
11  <title>2021</title>
12 </head>
13
14 <body>
15   <p align="center"><b><font face="Arial" color="#B086FF" size="5">2021</font></b></p>
16   <div align="center">
17     <center>
18       <table border="1" cellpadding="4" cellspacing="0" width="700">
19         <tr>
20           <td width="75" bgcolor="#B086FF" align="left"><b><font face="Arial" size="2">Date</font></b></td>
21           <td bgcolor="#B086FF" align="left"><b><font face="Arial" size="2">Location / Operator</font></b></td>
22           <td bgcolor="#B086FF" align="left"><b><font face="Arial" size="2">Aircraft Type / Registration</font></b></td>
23           <td align="right"><b><font face="Arial" size="2">Fatalities</font></b></td>
24         </tr>
25         <tr>
26           <td align="center"><b><font face="Arial" size="2">09 Jan 2021</font></b></td>
27           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-1.htm">09 Jan 2021</a></td>
28           <td align="left" valign="top"><font face="Arial" size="2">Near Jakarta, Indonesia<br>Sriwijaya Air</td>
29           &td align="right" valign="top"><font face="Arial" size="2">62/62(0)</td>
30         </tr>
31         <tr>
32           <td align="center"><b><font face="Arial" size="2">02 Mar 2021</font></b></td>
33           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-2.htm">02 Mar 2021</a></td>
34           <td align="left" valign="top"><font face="Arial" size="2">Pieri, Sudan<br>South Sudan Supreme Airlines</td>
35           &td align="right" valign="top"><font face="Arial" size="2">10/10(0)</td>
36         </tr>
37         <tr>
38           <td align="center"><b><font face="Arial" size="2">28 Mar 2021</font></b></td>
39           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-3.htm">28 Mar 2021</a></td>
40           <td align="left" valign="top"><font face="Arial" size="2">Near Butte, Alaska<br>Soloy Helicopters</td>
41           &td align="right" valign="top"><font face="Arial" size="2">5/6(0)</td>
42         </tr>
43         <tr>
44           <td align="center"><b><font face="Arial" size="2">21 May 2021</font></b></td>
45           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-4.htm">21 May 2021</a></td>
46           <td align="left" valign="top"><font face="Arial" size="2">Near Kaduna, Nigeria<br>Military - Nigerian Air Force</td>
47           &td align="right" valign="top"><font face="Arial" size="2">11/11(0)</td>
48         </tr>
49         <tr>
50           <td align="center"><b><font face="Arial" size="2">10 Jun 2021</font></b></td>
51           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-5.htm">10 Jun 2021</a></td>
52           <td align="left" valign="top"><font face="Arial" size="2">Near Pailin, Myanmar<br>Military - Myanmar Air Force</td>
53           &td align="right" valign="top"><font face="Arial" size="2">12/14(0)</td>
54         </tr>
55         <tr>
56           <td align="center"><b><font face="Arial" size="2">04 Jul 2021</font></b></td>
57           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-6.htm">04 Jul 2021</a></td>
58           <td align="left" valign="top"><font face="Arial" size="2">Patikul, Sulu, Philippines<br>Military - Philippine Air Force</td>
59           &td align="right" valign="top"><font face="Arial" size="2">50/96(3)</td>
60         </tr>
61         <tr>
62           <td align="center"><b><font face="Arial" size="2">06 Jul 2021</font></b></td>
63           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-7.htm">06 Jul 2021</a></td>
64           <td align="left" valign="top"><font face="Arial" size="2">Palana, Russia<br>Kamchatka Aviation Enterprise</td>
65           &td align="right" valign="top"><font face="Arial" size="2">28/28(0)</td>
66         </tr>
67         <tr>
68           <td align="center"><b><font face="Arial" size="2">12 Sep 2021</font></b></td>
69           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-8.htm">12 Sep 2021</a></td>
70           <td align="left" valign="top"><font face="Arial" size="2">Kazachinskoye, Russia<br>Aeroservice/SILA</td>
71           &td align="right" valign="top"><font face="Arial" size="2">4/16(0)</td>
72         </tr>
73         <tr>
74           <td align="center"><b><font face="Arial" size="2">27 Dec 2021</font></b></td>
75           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-9.htm">27 Dec 2021</a></td>
76           <td align="left" valign="top"><font face="Arial" size="2">El Cajon, California<br>Med Jet</td>
77           &td align="right" valign="top"><font face="Arial" size="2">4/4(0)</td>
78         </tr>
79       </table>
80     </center>
81   </div>
82 </body>
83 </html>
```

HTML Structure

```
<!DOCTYPE html>

<html>
  <!--This is a comment and ignored by web client.-->
  <head>
    <!--This section contains web page metadata.-->
    <title>ISA 401: Business Intelligence and Data Viz</title>
    <meta name="author" content="Fadel Megahed">
    <link rel="stylesheet" href="css/styles.css">
  </head>

  <body>
    <!--This section contains what you want to display on your web page.-->
    <h1>I'm a first level header</h1>
    <p>This is a <b>paragraph</b>.</p>
  </body>
</html>
```

HTML Syntax

`Author content` Author content

start tag: `Author content`

end tag: `Author content`

content: `Author content`

element name: `Author content`

attribute: `Author content`

attribute name: `Author content`

attribute value: `Author content`

Not all HTML tags have an end tag, for example:

`` → 

HTML Elements

| | |
|--------------------|---------------------------------------------------------------------------------------------------------|
| block element: | <code><div>content</div></code> |
| inline element: | <code>content</code> |
| paragraph: | <code><p>content</p></code> |
| header level 1: | <code><h1>content</h1></code> |
| header level 2: | <code><h2>content</h2></code> |
| italic: | <code><i>content</i></code> |
| emphasised text: | <code>content</code> |
| strong importance: | <code>content</code> |
| link: | <code>content</code> |
| unordered list: | <code> item 1 item 2 </code> |

Cascading Style Sheet (CSS)

- with the extension `.css`
- 3 ways to style elements in HTML:
 - **inline** by using the `style` attribute inside HTML start tag:
`<h1 style="color:blue;">Blue Header</h1>`
 - **externally** by using the `<link>` element:
`<link rel="stylesheet" href="styles.css">`
 - **internally** by defining within `<style>` element:

```
<style type="text/css">  
h1 { color: blue; }  
</style>
```

By convention, the `<style>` and `<link>` elements tend to go into the `<head>` section of the HTML document.

CSS Syntax

```
<style type="text/css">
h1 { color: blue; }
</style>
<h1>This is a header</h1>
```

This is a header

| | |
|-----------------|----------------------------------|
| selector: | <code>h1 { color: blue; }</code> |
| property: | <code>h1 { color: blue; }</code> |
| property name: | <code>h1 { color: blue; }</code> |
| property value: | <code>h1 { color: blue; }</code> |

You may have multiple properties for a single selector. ➡

```
h1 {
  color: blue;
  font-size: 16pt;
}
```

CSS Properties

```
<div>Sample text</div>
```

| | | |
|-------------------|---------------------------------------------------------------|-------------|
| background color: | <code>div { background-color: yellow; }</code> | Sample text |
| text color: | <code>div { color: purple; }</code> | Sample text |
| border: | <code>div { border: 1px dashed brown; }</code> | Sample text |
| left border only: | <code>div { border-left: 10px solid pink; }</code> | Sample text |
| text size: | <code>div { font-size: 10pt; }</code> | Sample text |
| padding: | <code>div { background-color: yellow; padding: 10px; }</code> | Sample text |
| margin: | <code>div { background-color: yellow; margin: 10px; }</code> | Sample text |

CSS Properties

```
<div>Sample text</div>
```

center align text: `div { background-color: yellow;
padding-top: 20px;
text-align: center; }`

Sample text

font family: `div { font-family: Marker Felt, times; }`

Sample text

strike: `div { text-decoration: line-through; }`

~~Sample text~~

underline: `div { text-decoration: underline; }`

Sample text

opacity: `div { opacity: 0.3 }`

Sample text

CSS Selector

| | |
|---------|-------------------------------------------|
| * | selects all elements |
| div | selects all <div> elements |
| div, p | selects all <div> and <p> elements |
| div p | selects all <p> within <div> |
| div > p | selects all <p> one level deep in <div> |
| div + p | selects all <p> immediately after a <div> |
| div ~ p | selects all <p> preceded by a <div> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
  <p>Hi!</p>
```

```
  How are you?
```

```
  <div class="child nice">
```

```
    <p>Hello!</p>
```

```
  </div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
<p>Hi!</p>
```

```
<blockquote class="child rebel">
```

```
  <p>Don't talk to me!</p>
```

```
</blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|---------|-------------------------------------------|
| * | selects all elements |
| div | selects all <div> elements |
| div, p | selects all <div> and <p> elements |
| div p | selects all <p> within <div> |
| div > p | selects all <p> one level deep in <div> |
| div + p | selects all <p> immediately after a <div> |
| div ~ p | selects all <p> preceded by a <div> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
  <p>Hi!</p>
```

```
  How are you?
```

```
  <div class="child nice">
```

```
    <p>Hello!</p>
```

```
  </div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
  <p>Hi!</p>
```

```
  <blockquote class="child rebel">
```

```
    <p>Don't talk to me!</p>
```

```
  </blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|------------|---------------------------------------------------------------------------------|
| * | selects all elements |
| blockquote | selects all <code><blockquote></code> elements |
| div, p | selects all <code><div></code> and <code><p></code> elements |
| div p | selects all <code><p></code> within <code><div></code> |
| div > p | selects all <code><p></code> one level deep in <code><div></code> |
| div + p | selects all <code><p></code> immediately after a <code><div></code> |
| div ~ p | selects all <code><p></code> preceded by a <code><div></code> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
<p>Maybe stories are just data with a soul.</p>
<footer>-Brene Brown</footer>
</blockquote>
```

```
<div id="p1" class="parent">
  Hmm
  <p>Hi!</p>
  How are you?
  <div class="child nice">
    <p>Hello!</p>
  </div>
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
  <p>Hi!</p>
  <blockquote class="child rebel">
    <p>Don't talk to me!</p>
  </blockquote>
</div>
```

```
<span class="child">
  <span class="parent child rebel">
    <p>Clean your room!</p>
  </span>
</span>
```

```
<p>End of households</p>
```


CSS Selector

| | |
|---------|-------------------------------------------|
| * | selects all elements |
| div | selects all <div> elements |
| div, p | selects all <div> and <p> elements |
| div p | selects all <p> within <div> |
| div > p | selects all <p> one level deep in <div> |
| div + p | selects all <p> immediately after a <div> |
| div ~ p | selects all <p> preceded by a <div> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
  <p>Hi!</p>
```

```
  How are you?
```

```
  <div class="child nice">
```

```
    <p>Hello!</p>
```

```
  </div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
  <p>Hi!</p>
```

```
  <blockquote class="child rebel">
```

```
    <p>Don't talk to me!</p>
```

```
  </span>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|---------|---------------------------------------------------------------------------------|
| * | selects all elements |
| div | selects all <code><div></code> elements |
| div, p | selects all <code><div></code> and <code><p></code> elements |
| div p | selects all <code><p></code> within <code><div></code> |
| div > p | selects all <code><p></code> one level deep in <code><div></code> |
| div + p | selects all <code><p></code> immediately after a <code><div></code> |
| div ~ p | selects all <code><p></code> preceded by a <code><div></code> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
  <p>Hi!</p>
```

```
  How are you?
```

```
  <div class="child nice">
```

```
    <p>Hello!</p>
```

```
  </div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
  <p>Hi!</p>
```

```
  <blockquote class="child rebel">
```

```
    <p>Don't talk to me!</p>
```

```
  </blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|------------|-------------------------------------------|
| * | selects all elements |
| div | selects all <div> elements |
| div, p | selects all <div> and <p> elements |
| p div | selects all <div> within <p> |
| div > p | selects all <p> one level deep in <div> |
| div + p | selects all <p> immediately after a <div> |
| div ~ p | selects all <p> preceded by a <div> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
  <p>Hi!</p>
```

```
  How are you?
```

```
  <div class="child nice">
```

```
    <p>Hello!</p>
```

```
  </div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
<p>Hi!</p>
```

```
<blockquote class="child rebel">
```

```
  <p>Don't talk to me!</p>
```

```
</blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|---------|---------------------------------------------------------------------------------|
| * | selects all elements |
| div | selects all <code><div></code> elements |
| div, p | selects all <code><div></code> and <code><p></code> elements |
| div p | selects all <code><p></code> within <code><div></code> |
| div > p | selects all <code><p></code> one level deep in <code><div></code> |
| div + p | selects all <code><p></code> immediately after a <code><div></code> |
| div ~ p | selects all <code><p></code> preceded by a <code><div></code> |

Ignores inline elements like

```
<h1>This is a sample h1</h1>  
span, i, b,...
```

```
<blockquote>  
<p>Maybe stories are just data with a soul.</p>  
<footer>-Brene Brown</footer>  
</blockquote>
```

```
<div id="p1" class="parent">  
  Hmm  
  <p>Hi!</p>  
  How are you?  
  <div class="child nice">  
    <p>Hello!</p>  
  </div>  
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">  
  <p>Hi!</p>  
  <blockquote class="child rebel">  
    <p>Don't talk to me!</p>  
  </blockquote>  
</div>
```

```
<span class="child">  
  <span class="parent child rebel">  
    <p>Clean your room!</p>  
  </span>  
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|---------|---------------------------------------------------------------------------------|
| * | selects all elements |
| div | selects all <code><div></code> elements |
| div, p | selects all <code><div></code> and <code><p></code> elements |
| div p | selects all <code><p></code> within <code><div></code> |
| div > p | selects all <code><p></code> one level deep in <code><div></code> |
| div + p | selects all <code><p></code> immediately after a <code><div></code> |
| div ~ p | selects all <code><p></code> preceded by a <code><div></code> |

Ignores inline elements like

```
<h1>This is a sample h1</h1>
```

```
<blockquote>
<p>Maybe stories are just data with a soul.</p>
<footer>-Brene Brown</footer>
</blockquote>
```

```
<div id="p1" class="parent">
  Hmm
  <p>Hi!</p>
  How are you?
  <div class="child nice">
    <p>Hello!</p>
  </div>
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
  <p>Hi!</p>
  <blockquote class="child rebel">
    <p>Don't talk to me!</p>
  </blockquote>
</div>
```

```
<span class="child">
  <span class="parent child rebel">
    <p>Clean your room!</p>
  </span>
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|-------------------------|---------------------------------------------------------------------------------|
| <code>*</code> | selects all elements |
| <code>div</code> | selects all <code><div></code> elements |
| <code>div, p</code> | selects all <code><div></code> and <code><p></code> elements |
| <code>div p</code> | selects all <code><p></code> within <code><div></code> |
| <code>div > p</code> | selects all <code><p></code> one level deep in <code><div></code> |
| <code>div + p</code> | selects all <code><p></code> immediately after a <code><div></code> |
| <code>div ~ p</code> | selects all <code><p></code> preceded by a <code><div></code> |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
  <p>Hi!</p>
```

```
  How are you?
```

```
  <div class="child nice">
```

```
    <p>Hello!</p>
```

```
  </div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
  <p>Hi!</p>
```

```
  <blockquote class="child rebel">
```

```
    <p>Don't talk to me!</p>
```

```
  </blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|-------------------------|-----------------------------------------------------------------------------------------------------------------|
| <code>.classname</code> | selects all elements with the attribute <code>class="classname"</code> . |
| <code>.c1.c2</code> | selects all elements with <i>both</i> <code>c1</code> and <code>c2</code> within its class attribute. |
| <code>.c1 .c2</code> | selects all elements with class <code>c2</code> that is a descendant of an element with class <code>c1</code> . |
| <code>#idname</code> | selects all elements with the attribute <code>id="idname"</code> . |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
<p>Hi!</p>
```

```
  How are you?
```

```
<div class="child nice">
```

```
  <p>Hello!</p>
```

```
</div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
<p>Hi!</p>
```

```
<blockquote class="child rebel">
```

```
  <p>Don't talk to me!</p>
```

```
</blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|----------------------|-----------------------------------------------------------------------------------------------------------------|
| <code>.parent</code> | selects all elements with the attribute <code>class="parent"</code> . |
| <code>.c1.c2</code> | selects all elements with <i>both</i> <code>c1</code> and <code>c2</code> within its class attribute. |
| <code>.c1 .c2</code> | selects all elements with class <code>c2</code> that is a descendant of an element with class <code>c1</code> . |
| <code>#idname</code> | selects all elements with the attribute <code>id="idname"</code> . |

`<h1>This is a sample h1`

Note some offspring do not inherit class from their parents.

`</h1>`

`<p>Maybe stories are just data with a soul.</p>`

`<footer>-Brene Brown</footer>`

`</div>`

`<div id="p1" class="parent">`

 Hmm

Hi!

 How are you?

Hello!

</div>

</div>

`<p>Household 1</p>`

`<div class="parent">`

Hi!

Don't talk to me!

</div>

</div>

``

Clean your room!

`<p>End of households</p>`

13 / 30

CSS Selector

| | |
|-------------------------|-------------------------------------------------------------------------|
| <code>.classname</code> | selects all elements with the attribute <code>class="classname".</code> |
|-------------------------|-------------------------------------------------------------------------|

| | |
|---------------------------|-------------------------------------------------------------------------------------------------------------|
| <code>.child.rebel</code> | selects all elements with <i>both</i> <code>child</code> and <code>rebel</code> within its class attribute. |
|---------------------------|-------------------------------------------------------------------------------------------------------------|

| | |
|----------------------|-----------------------------------------------------------------------------------------------------------------|
| <code>.c1 .c2</code> | selects all elements with class <code>c2</code> that is a descendant of an element with class <code>c1</code> . |
|----------------------|-----------------------------------------------------------------------------------------------------------------|

| | |
|----------------------|-------------------------------------------------------------------|
| <code>#idname</code> | selects all elements with the attribute <code>id="idname".</code> |
|----------------------|-------------------------------------------------------------------|

```
<h1>This is a sample html</h1>
```

```
<blockquote>
<p>Maybe stories are just data with a soul.</p>
<footer>-Brene Brown</footer>
</blockquote>
```

```
<div id="p1" class="parent">
  Hmm
  <p>Hi!</p>
  How are you?
  <div class="child nice">
    <p>Hello!</p>
  </div>
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
  <p>Hi!</p>
  <blockquote class="child rebel">
    <p>Don't talk to me!</p>
  </blockquote>
</div>
```

```
<span class="child">
  <span class="parent child rebel">
    <p>Clean your room!</p>
  </span>
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|---------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| <code>.classname</code> | selects all elements with the attribute <code>class="classname"</code> . |
| <code>.c1.c2</code> | selects all elements with <i>both</i> <code>c1</code> and <code>c2</code> within its class attribute. |
| <code>.parent</code> <code>.rebel</code> | selects all elements with class <code>rebel</code> that is a descendant of an element with class <code>parent</code> . |
| <code>#idname</code> | selects all elements with the attribute <code>id="idname"</code> . |

```
<h1>This is a sample html</h1>
```

```
<blockquote>
```

```
<p>Maybe stories are just data with a soul.</p>
```

```
<footer>-Brene Brown</footer>
```

```
</blockquote>
```

```
<div id="p1" class="parent">
```

```
  Hmm
```

```
<p>Hi!</p>
```

```
  How are you?
```

```
<div class="child nice">
```

```
  <p>Hello!</p>
```

```
</div>
```

```
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
```

```
<p>Hi!</p>
```

```
<blockquote class="child rebel">
```

```
  <p>Don't talk to me!</p>
```

```
</blockquote>
```

```
</div>
```

```
<span class="child">
```

```
<span class="parent child rebel">
```

```
  <p>Clean your room!</p>
```

```
</span>
```

```
</span>
```

```
<p>End of households</p>
```

CSS Selector

| | |
|-------------------------|-----------------------------------------------------------------------------------------------------------------|
| <code>.classname</code> | selects all elements with the attribute <code>class="classname"</code> . |
| <code>.c1.c2</code> | selects all elements with <i>both</i> <code>c1</code> and <code>c2</code> within its class attribute. |
| <code>.c1 .c2</code> | selects all elements with class <code>c2</code> that is a descendant of an element with class <code>c1</code> . |
| <code>#p1</code> | selects all elements with the attribute <code>id="p1"</code> . |

Unlike `class`, you can only have one `id` value and must be unique in the whole HTML document.

```
<h1>This is a sample htr
</h1>
<blockquote>
  <p>Maybe stories are jus
  </p>
</blockquote>
<footer>-Brene Brown</footer>
```

```
<div id="p1" class="parent">
  Hmm
  <p>Hi!</p>
  How are you?
  <div class="child nice">
    <p>Hello!</p>
  </div>
</div>
```

```
<p>Household 1</p>
```

```
<div class="parent">
  <p>Hi!</p>
  <blockquote class="child rebel">
    <p>Don't talk to me!</p>
  </blockquote>
</div>
```

```
<span class="child">
  <span class="parent child rebel">
    <p>Clean your room!</p>
  </span>
</span>
```

```
<p>End of households</p>
```

JavaScript (JS)*

- JS is a programming language and enable interactive components in HTML documents.
- 2 ways to insert JS into a HTML document:
 - **internally** by defining within `<script>` element:

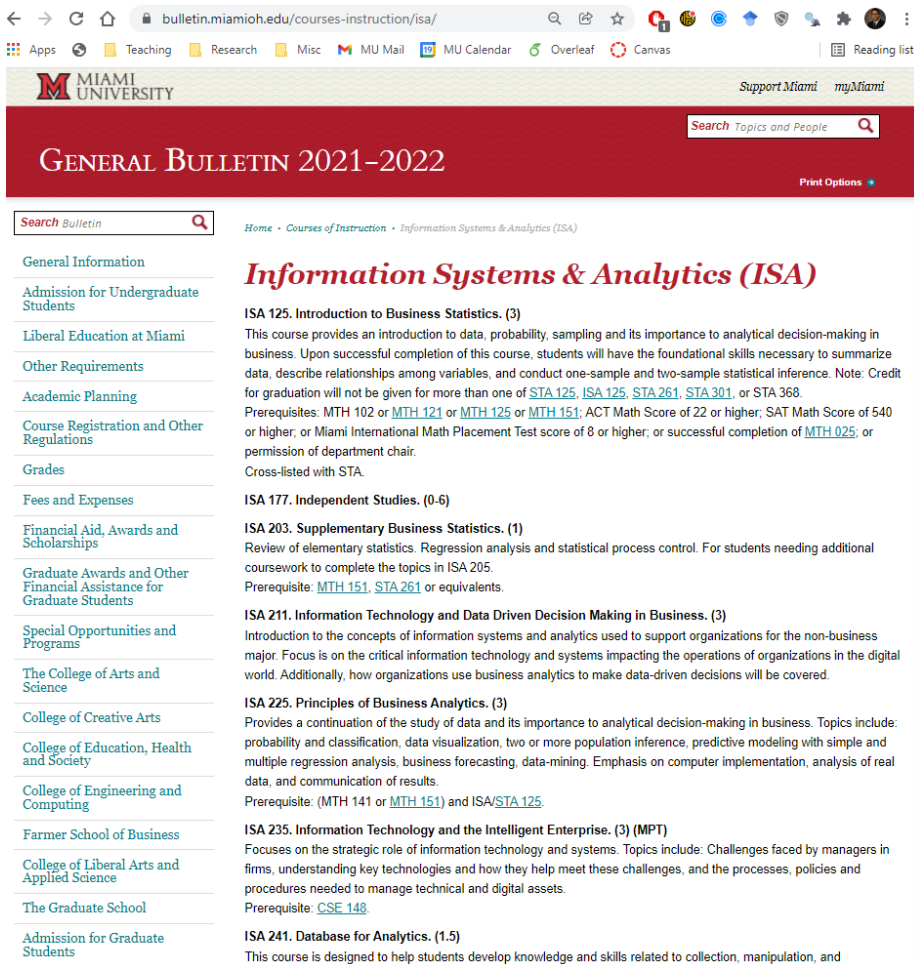
```
<script>  
document.getElementById("p1").innerHTML = "content";  
</script>
```

- **externally** by using the `src` attribute to refer to the external file:

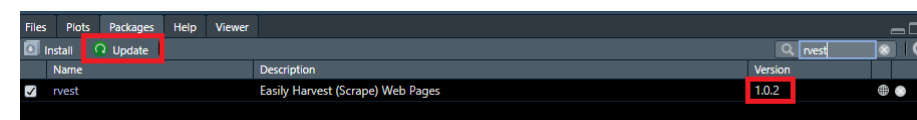
```
<script src="js/myjs.js"></script>
```

Web Scrapping

Invest: Step 1 - Reading Static HTML Pages



Use `{rvest}` \geq v1.0.2 (if not, update)



```
if(require(pacman)==FALSE) install.packages("pacman")
pacman::p_load(rvest)
isa_courses = read_html("http://bulletin.brown.edu/undergraduate/academic-requirements/requirements-for-the-degree-in-science/")
isa_courses
```

```
## {html_document}
## <html xml:lang="en" lang="en" dir="ltr">
## [1] <head>\n<title>Information System
## [2] <body>\n\n\n\n\n\n\n\n<!-- Google Tag
```

rvest: Step 2 - Selecting HTML Elements

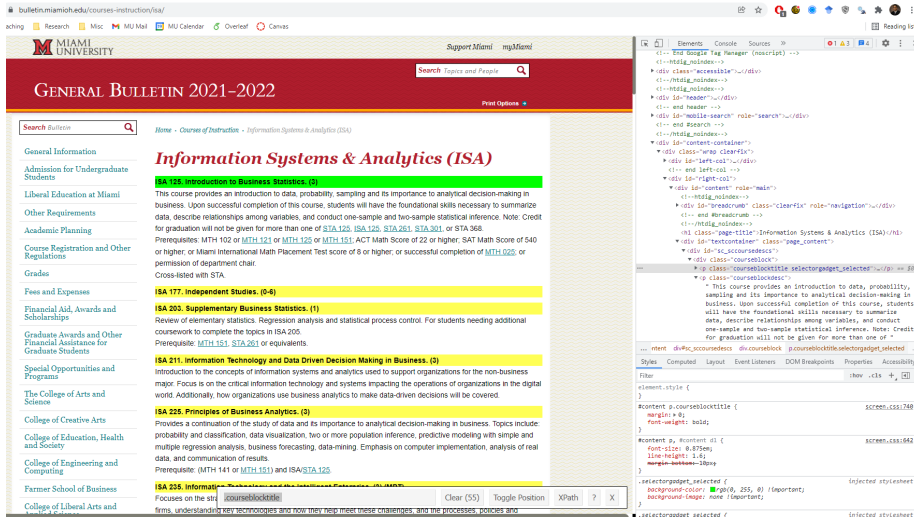
Inspector

The screenshot shows the web browser's developer tools for the Miami University General Bulletin 2021-2022 page. The left pane displays the page structure, including a sidebar menu with links like 'General Information', 'Admission for Undergraduate Students', 'Liberal Education at Miami', 'Other Requirements', 'Academic Planning', 'Course Registration and Other Regulations', 'Grades', 'Fees and Expenses', 'Financial Aid, Awards and Scholarships', 'Graduate Appeals and Other Financial Assistance for Graduate Students', 'Special Opportunities and Programs', 'The College of Arts and Sciences', 'College of Creative Arts', 'College of Education, Health and Society', and 'College of Engineering and Computing'. The middle pane shows the HTML code, and the right pane displays the CSS styles. The 'Information Systems & Analytics (ISA)' section is highlighted in the middle pane.

Selector Gadget

The screenshot shows the web browser's developer tools for the Miami University General Bulletin 2021-2022 page. The left pane displays the page structure, including a sidebar menu with links like 'General Information', 'Admission for Undergraduate Students', 'Liberal Education at Miami', 'Other Requirements', 'Academic Planning', 'Course Registration and Other Regulations', 'Grades', 'Fees and Expenses', 'Financial Aid, Awards and Scholarships', 'Graduate Appeals and Other Financial Assistance for Graduate Students', 'Special Opportunities and Programs', 'The College of Arts and Sciences', 'College of Creative Arts', 'College of Education, Health and Society', 'College of Engineering and Computing', 'Farmer School of Business', and 'College of Liberal Arts and Sciences'. The middle pane shows the HTML code, and the right pane displays the CSS styles. The 'Information Systems & Analytics (ISA)' section is highlighted in the middle pane.

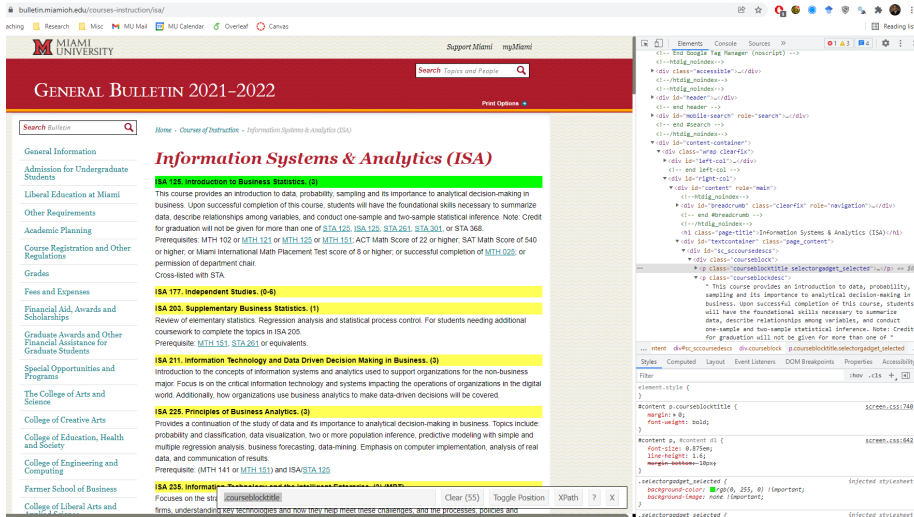
rvest: Step 2 - Selecting HTML Elements



```
isa_course_titles = isa_courses |>
  html_elements(css = "p.courseblocktitle")
isa_course_titles
```

```
## {xml_nodeset (50)}
## [1] <p class="courseblocktitle"><str
## [2] <p class="courseblocktitle"><str
## [3] <p class="courseblocktitle"><str
## [4] <p class="courseblocktitle"><str
## [5] <p class="courseblocktitle"><str
## [6] <p class="courseblocktitle"><str
## [7] <p class="courseblocktitle"><str
## [8] <p class="courseblocktitle"><str
## [9] <p class="courseblocktitle"><str
## [10] <p class="courseblocktitle"><str
## [11] <p class="courseblocktitle"><str
## [12] <p class="courseblocktitle"><str
## [13] <p class="courseblocktitle"><str
## [14] <p class="courseblocktitle"><str
## [15] <p class="courseblocktitle"><str
```


rvest: Step 3 - Getting HTML Text



```
isa_course_titles_en = isa_course_titles_html_text2()
```

```
isa_course_titles_en
```

```
## [1] "ISA 125. Introduction to Business Statistics. (3)"
## [2] "ISA 177. Independent Studies. (0-4)"
## [3] "ISA 203. Supplementary Business Statistics. (1)"
## [4] "ISA 211. Information Technology and Data Driven Decision Making in Business. (2)"
## [5] "ISA 225. Principles of Business Analytics. (2)"
## [6] "ISA 235. Information Technology and Data Driven Decision Making in Business. (2)"
## [7] "ISA 241. Database for Analytics. (3)"
## [8] "ISA 242. Programming for Analytics. (3)"
## [9] "ISA 245. Database Systems and Data Analytics. (3)"
## [10] "ISA 250. Basic Math for Analytics. (3)"
## [11] "ISA 277. Independent Studies. (0-4)"
## [12] "ISA 281. Concepts in Business Analytics. (3)"
## [13] "ISA 291. Applied Regression Analysis. (3)"
## [14] "ISA 301. Business Data Communications. (3)"
## [15] "ISA 303. Enterprise Systems. (3)"
## [16] "ISA 305. Information Technology and Data Driven Decision Making in Business. (2)"
```

Demo: Scraping the Course Descriptions

- We will build on the previous example and we will scrape the **course descriptions** associated with these courses.
- Then, we will create a **data frame** containing **both** the **course titles** and **descriptions**
- Then, we will **export the results to a CSV** so that we can analyze that in a separate program if we wanted to.

Non-Graded Class Activity

| Activity | Your Solution | My Solution |
|----------|---------------|-------------|
|----------|---------------|-------------|

- Go to [this database on plane crashes](#)
- Scrape the HTML table. **Note the difference from text elements:**
 - The CSS selector for `html_elements()` will be different.
 - You will extract a table (in its **entirety**) and hence:
 - we will use `html_table()` instead of `html_text2()`
- Store the scraped data in an appropriate location on your computer (e.g., within the data folder for ISA 401)

04:00

Non-Graded Class Activity

Activity

Your Solution

My Solution

Over the next 4 minutes, use a  script file to perform the tasks outlined in the activity panel.

04:00

Non-Graded Class Activity

Activity

Your Solution

My Solution

Please refer to our discussion in class

Legal and Ethical Issues with Web Scraping

Robots.txt

When scraping/crawling the web you need to be aware of `robots.txt`.

The robots exclusion standard, also known as the robots exclusion protocol or simply robots.txt, is a standard used by websites to communicate with web crawlers and other web robots. The standard specifies how to inform the web robot about which areas of the website should not be processed or scanned. --- [Wikipedia](#)

Using the excellent `robotstxt` 📦 to check if scraping/crawling a specific directory is allowed.

```
if(require(robotstxt)==FALSE) install.packages("robotstxt")
robotstxt::paths_allowed(paths = "2024/", domain = "plane crashinfo.com", bot = "*")
```

```
## [1] TRUE
```


Terms of Service


Most large companies have **terms of service** that supplement what is permitted and/or disallowed on their `robots.txt` file. Examples include:

- [Yelp's US Terms of Service](#)
- [LinkedIn Terms of Service](#)

Ethical/Legal Considerations

- **Use of publicly available reviews as a part of your service:** Would you classify the [Yelp vs Google Feud](#) as such an example?

**Jeremy Stoppelman**
@jeremys · [Follow](#)



Wow Google, congrats on a new low. Consumer searches for Yelp gets "reviews" which are Google Ads.

About 7,020,000 results (0.84 seconds)

HVAC pros serving San Francisco

Sponsored ⓘ

| | | |
|---------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| The Appliance Repair Do.. 4.6 ★★★★★ · See reviews ✓ Google guaranteed Alameda (510) 871-3938 Open now | Healthy Duct Cleaning S... 4.9 ★★★★★ · See reviews ✓ Google guaranteed Daly City (415) 993-1965 Open now | Atlas Trillo Heating & Air 4.5 ★★★★★ · See reviews ✓ Google guaranteed San Jose (408) 915-7800 Opens Tue at 8 AM |
|---------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|

→ [More HVAC pros in San Francisco](#)

Heating & Air Conditioning/HVAC in San Francisco - Yelp
<https://www.yelp.com/c/sf/hvac> ▼
The Best Heating & Air Conditioning/HVAC in San Francisco on Yelp. Read about places like: Air Flow Pros Heating And Air Conditioning, Kohler Heating, ...

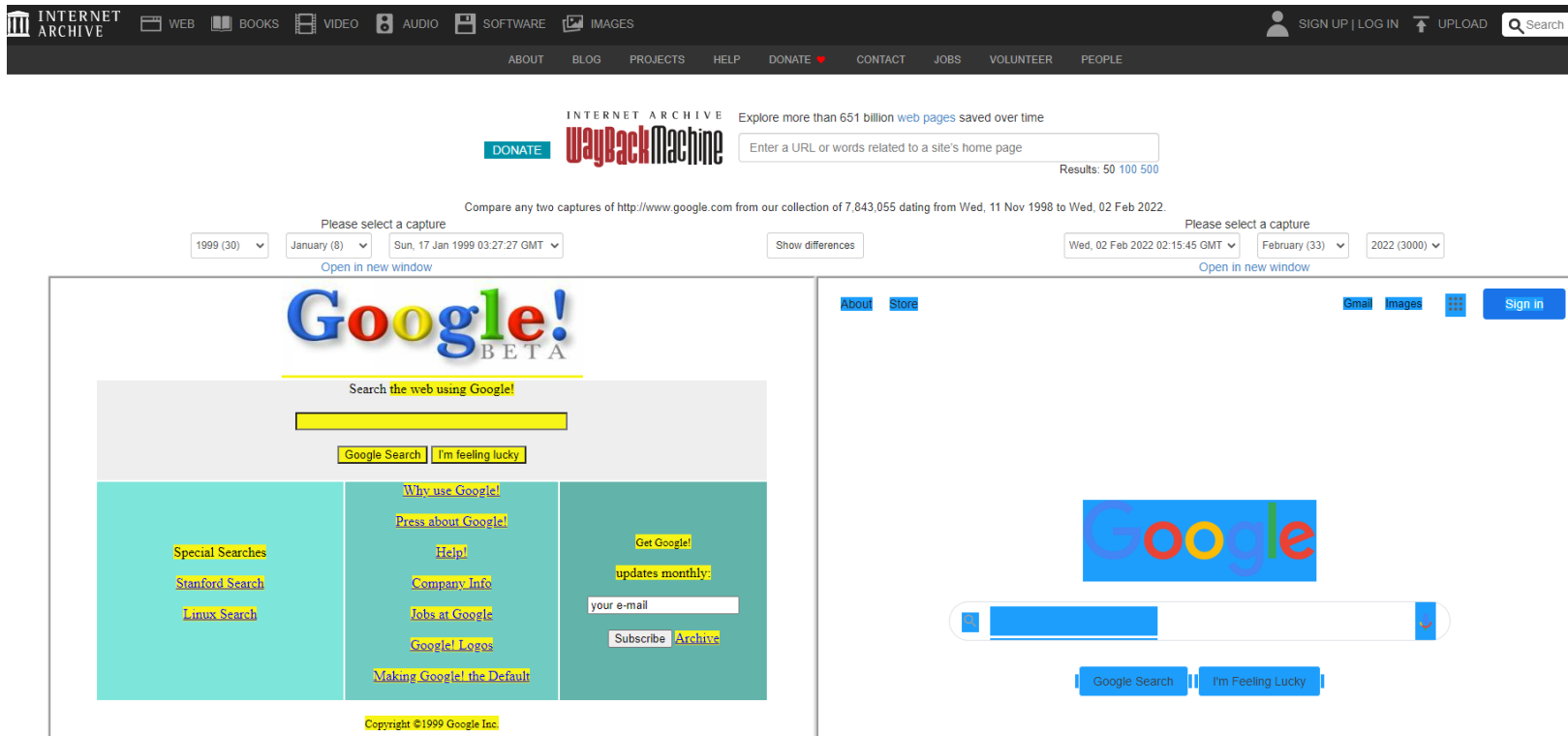
Best Hvac contractors in San Francisco, CA - Yelp
https://www.yelp.com/search?find_desc=hvac+contractors&find_loc=San+Francisco+CA

Ethical/Legal Considerations

- **Use of publicly available profiles as a part of your service:**
 - LinkedIn vs Hiq Labs: Ninth Circuit Decision in 2019
 - Revival of Case in 2021 by Supreme Court

Ethical/Legal Considerations

- What about scraping entire websites/webpages for the purpose of archiving the internet?



The evolution of the home page for Google per the Wayback Machine

Recap

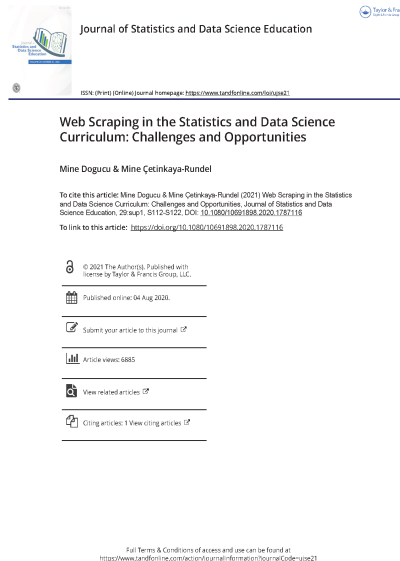
Summary of Main Points

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

- Understand when can we scrape data (i.e., `robots.txt`)
- Scrape a webpage using .

Things to Do to Prepare for Next Class

- Go over your notes, read through the supplementary material (below) and complete **Assignment 04** on Canvas.



- PDF of Published Paper
- ePub of Published Paper

- Selector Gadget
- Getting Started with rvest