

# ISA 401: Business Intelligence & Data Visualization

## 04: Scraping Webpages in and

Fadel M. Megahed, PhD

Endres Associate Professor  
Farmer School of Business  
Miami University

 @FadelMegahed





 fmegahed

 fmegahed@miamioh.edu



 Automated Scheduler for Office Hours

Fall 2023

# Quick Refresher from Last Class

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

# Learning Objectives for Today's Class




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

# 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
<a href="#">09 Jan 2021</a>	Near Jakarta, Indonesia Sriwijaya Air	Boeing 737-524 PK-CLC	62/62(0)
<a href="#">02 Mar 2021</a>	Pieri, Sudan South Sudan Supreme Airlines	Let L-410UVP-E HK-4274	10/10(0)
<a href="#">28 Mar 2021</a>	Near Butte, Alaska Soloy Helicopters	Eurocopter AS350B3 Ecureuil N351SH	5/6(0)
<a href="#">21 May 2021</a>	Near Kaduna, Nigeria Military - Nigerian Air Force	Beechcraft B300 King Air 350i NAF203	11/11(0)
<a href="#">10 Jun 2021</a>	Near Pyn Oo Lwin, Myanmar Military - Myanmar Air Force	Beechcraft 1900D 4610	12/14(0)
<a href="#">04 Jul 2021</a>	Patikul, Sulu, Philippines Military - Philippine Air Force	Lockheed C-130H Hercules 5125	50/96(3)
<a href="#">06 Jul 2021</a>	Palana, Russia Kamchatka Aviation Enterprise	Antonov An 26B-100 RA-26085	28/28(0)
<a href="#">12 Sep 2021</a>	Kazachinskoye, Russia Aeroservice/SILA	Let L-410UVP-E20 RA-67042	4/16(0)
<a href="#">27 Dec 2021</a>	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  </head>
12
13 <body>
14   <p align="center"><b><font face="Arial" color="#B086FF" size="5">2021</font></b></p>
15   <div align="center">
16     <center>
17       <table border="1" cellpadding="4" cellspacing="0" width="700">
18         <tr>
19           <td width="75" bgcolor="#B086FF" align="left"><b><font face="Arial" size="2">Date</font></b></td>
20           <td bgcolor="#B086FF" align="left"><b><font face="Arial" size="2">Location / Operator</b></td>
21           <td bgcolor="#B086FF" align="left"><b><font face="Arial" size="2">Aircraft Type / Registration</b></td>
22           <td align="right" valign="top"><b><font face="Arial" size="2">Fatalities</b></td>
23         </tr>
24         <tr>
25           <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-1.htm">09 Jan 2021</a></td>
26           <td align="left" valign="top"><font face="Arial" size="2">Near Jakarta, Indonesia<br>Sriwijaya Air
27         </td>
28         <td align="left" valign="top"><font face="Arial" size="2">Boeing 737-524<br>PK-CLC</td>
29         &td align="right" valign="top"><font face="Arial" size="2">62/62(0)</td>
30       </tr>
31       <tr>
32         <td align="left" valign="top"><font face="Arial" size="2"><a href="2021-2.htm">02 Mar 2021</a></td>
33         <td align="left" valign="top"><font face="Arial" size="2">Pieri, Sudan<br>South Sudan Supreme Airlines
34       </td>
35         <td align="left" valign="top"><font face="Arial" size="2">Let L-410UVP-E<br>HK-4274 </td>
36         &td align="right" valign="top"><font face="Arial" size="2">10/10(0)</td>
37       </tr>
38     </center>
39   </div>
40 </body>
41 </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

`<span style="color:blue;">Author content</span>` Author content

---

start tag:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
------------	--------------------------------------------------------------------------

end tag:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
----------	--------------------------------------------------------------------------

content:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
----------	--------------------------------------------------------------------------

element name:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
---------------	--------------------------------------------------------------------------

attribute:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
------------	--------------------------------------------------------------------------

attribute name:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
-----------------	--------------------------------------------------------------------------

attribute value:	<code>&lt;span style="color:blue;"&gt;Author content&lt;/span&gt;</code>
------------------	--------------------------------------------------------------------------

---

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

`` → 



# HTML Elements

block element:	<code>&lt;div&gt;content&lt;/div&gt;</code>
inline element:	<code>&lt;span&gt;content&lt;/span&gt;</code>
paragraph:	<code>&lt;p&gt;content&lt;/p&gt;</code>
header level 1:	<code>&lt;h1&gt;content&lt;/h1&gt;</code>
header level 2:	<code>&lt;h2&gt;content&lt;/h2&gt;</code>
italic:	<code>&lt;i&gt;content&lt;/i&gt;</code>
emphasised text:	<code>&lt;em&gt;content&lt;/em&gt;</code>
strong importance:	<code>&lt;strong&gt;content&lt;/strong&gt;</code>
link:	<code>&lt;a href="https://github.com/fmegahed/isa401"&gt;content&lt;/a&gt;</code>
unordered list:	<code>&lt;ul&gt; &lt;li&gt;item 1&lt;/li&gt; &lt;li&gt;item 2&lt;/li&gt; &lt;/ul&gt;</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 <blockquote> 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

<code>*</code>	selects all elements
<code>div</code>	selects all <code>&lt;div&gt;</code> elements
<code>div, p</code>	selects all <code>&lt;div&gt;</code> and <code>&lt;p&gt;</code> elements
<code>div p</code>	selects all <code>&lt;p&gt;</code> within <code>&lt;div&gt;</code>
<code>div &gt; p</code>	selects all <code>&lt;p&gt;</code> one level deep in <code>&lt;div&gt;</code>
<code>div + p</code>	selects all <code>&lt;p&gt;</code> immediately after a <code>&lt;div&gt;</code>
<code>div ~ p</code>	selects all <code>&lt;p&gt;</code> preceded by a <code>&lt;div&gt;</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>
```

```
  </span>
```

```
</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>&lt;div&gt;</code> elements
<code>div, p</code>	selects all <code>&lt;div&gt;</code> and <code>&lt;p&gt;</code> elements
<code>div p</code>	selects all <code>&lt;p&gt;</code> within <code>&lt;div&gt;</code>
<code>div &gt; p</code>	selects all <code>&lt;p&gt;</code> one level deep in <code>&lt;div&gt;</code>
<code>div + p</code>	selects all <code>&lt;p&gt;</code> immediately after a <code>&lt;div&gt;</code>
<code>div ~ p</code>	selects all <code>&lt;p&gt;</code> preceded by a <code>&lt;div&gt;</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>&lt;div&gt;</code> elements
div, p	selects all <code>&lt;div&gt;</code> and <code>&lt;p&gt;</code> elements
div p	selects all <code>&lt;p&gt;</code> within <code>&lt;div&gt;</code>
div > p	selects all <code>&lt;p&gt;</code> one level deep in <code>&lt;div&gt;</code>
div + p	selects all <code>&lt;p&gt;</code> immediately after a <code>&lt;div&gt;</code>
div ~ p	selects all <code>&lt;p&gt;</code> preceded by a <code>&lt;div&gt;</code>

Ignores inline elements like `span, i, b, ...`

```
<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

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

Ignores inline elements like `span, i, b, ...`

```
<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

*	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

<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

`.parent` selects all elements with the attribute `class="parent"`.

`.c1.c2` selects all elements with *both* `c1` and `c2` within its class attribute.

`.c1 .c2` selects all elements with class `c2` that is a descendant of an element with class `c1`.

`#idname` selects all elements with the attribute `id="idname"`.

Note some offspring do not inherit class from their parents.

```
<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>.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.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
```

```
<blockquote>
```

```
<p>Maybe stories are jus
```

```
<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>
```

# JS 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

# rvest: Step 1 - Reading Static HTML Pages in R

[bulletin.miamioh.edu/courses-instruction/isa/](#)

Apps Teaching Research Misc MU Mail MU Calendar Overleaf Canvas Reading list

MIAMI UNIVERSITY Support Miami myMiami

# GENERAL BULLETIN 2021–2022

Search Topics and People

Print Options

Search Bulletin

- 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 Awards and Other Financial Assistance for Graduate Students
- Special Opportunities and Programs
- The College of Arts and Science
- College of Creative Arts
- College of Education, Health and Society
- College of Engineering and Computing
- Farmer School of Business
- College of Liberal Arts and Applied Science
- The Graduate School
- Admission for Graduate Students

Home • Courses of Instruction • Information Systems & Analytics (ISA)

## Information Systems & Analytics (ISA)

### ISA 125. Introduction to Business Statistics. (3)

This course provides an introduction to data, probability, sampling and its importance to analytical decision-making in business. Upon successful completion of this course, students will have the foundational skills necessary to summarize data, describe relationships among variables, and conduct one-sample and two-sample statistical inference. Note: Credit for graduation will not be given for more than one of [STA 125](#), [ISA 125](#), [STA 261](#), [STA 301](#), or STA 368.

Prerequisites: MTH 102 or [MTH 121](#) or [MTH 125](#) or [MTH 151](#); ACT Math Score of 22 or higher; SAT Math Score of 540 or higher; or Miami International Math Placement Test score of 8 or higher; or successful completion of [MTH 025](#); or permission of department chair.

Cross-listed with STA.

### ISA 177. Independent Studies. (0-6)

### ISA 203. Supplementary Business Statistics. (1)

Review of elementary statistics. Regression analysis and statistical process control. For students needing additional coursework to complete the topics in ISA 205.

Prerequisite: [MTH 151](#), [STA 261](#) or equivalents.

### ISA 211. Information Technology and Data Driven Decision Making in Business. (3)

Introduction to the concepts of information systems and analytics used to support organizations for the non-business major. Focus is on the critical information technology and systems impacting the operations of organizations in the digital world. Additionally, how organizations use business analytics to make data-driven decisions will be covered.

### ISA 225. Principles of Business Analytics. (3)

Provides a continuation of the study of data and its importance to analytical decision-making in business. Topics include: probability and classification, data visualization, two or more population inference, predictive modeling with simple and multiple regression analysis, business forecasting, data-mining. Emphasis on computer implementation, analysis of real data, and communication of results.

Prerequisite: (MTH 141 or [MTH 151](#)) and ISA/[STA 125](#).

### ISA 235. Information Technology and the Intelligent Enterprise. (3) (MPT)


Focuses on the strategic role of information technology and systems. Topics include: Challenges faced by managers in firms, understanding key technologies and how they help meet these challenges, and the processes, policies and procedures needed to manage technical and digital assets.

Prerequisite: [CSE 148](#).

### ISA 241. Database for Analytics. (1.5)

This course is designed to help students develop knowledge and skills related to collection, manipulation, and

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



The screenshot shows the RStudio interface with the 'Packages' tab selected. The 'rvest' package is listed with a green circular update icon next to its name. The version '1.0.2' is highlighted in a red box. The description 'Easily Harvest (Scrape) Web Pages' is also visible.

```
if(require(pacman)==FALSE) install.packages(pacman::p_load(rvest))
isa_courses = read_html("http://bulletin.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

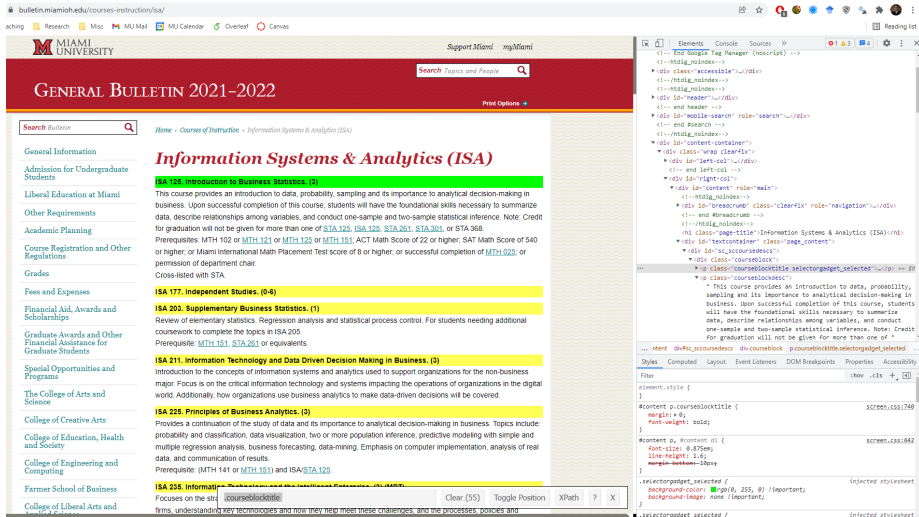
## Inspector

The screenshot shows a web browser displaying the 'GENERAL BULLETIN 2021-2022' page. The page content includes sections for 'Information Systems & Analytics (ISA)', 'ISA 125, Introduction to Business Statistics', 'ISA 177, Independent Studies', 'ISA 203, Supplementary Business Statistics', 'ISA 211, Information Technology and Data Driven Decision Making in Business', and 'ISA 225, Principles of Business Analytics'. The Inspector shows the DOM tree on the left and the selected element's HTML structure on the right.

## Selector Gadget

The screenshot shows the same web browser page with the Selector Gadget tool active. The tool highlights the 'Information Systems & Analytics (ISA)' section. The Selector Gadget interface shows the selected element and its corresponding CSS selector.

# rvest: Step 2 - Selecting HTML Elements

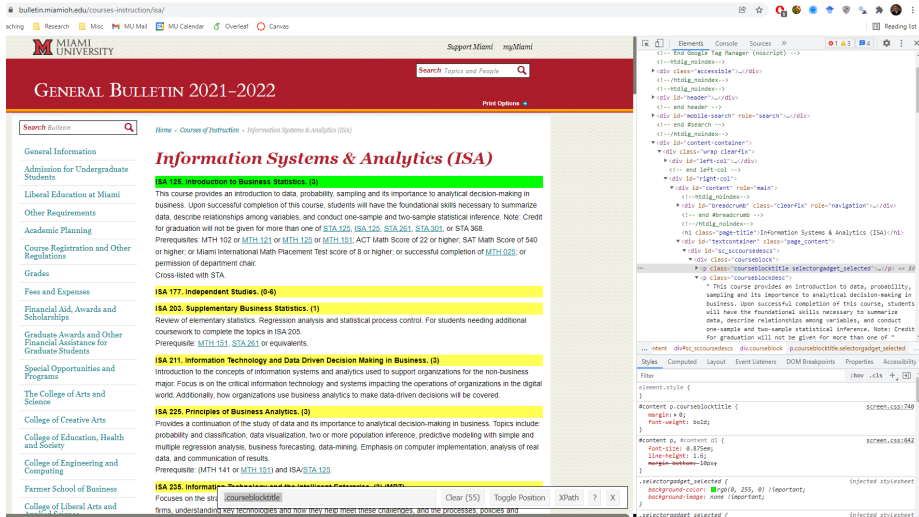


```
isa_course_titles = isa_courses |>
  html_elements(css = "p.courseblocktitle")
isa_course_titles
```

```
## {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  
## [2] "ISA 177. Independent Studies. (0  
## [3] "ISA 203. Supplementary Business  
## [4] "ISA 211. Information Technology  
## [5] "ISA 225. Principles of Business  
## [6] "ISA 235. Information Technology  
## [7] "ISA 241. Database for Analytics.  
## [8] "ISA 242. Programming for Analyti  
## [9] "ISA 245. Database Systems and Da  
## [10] "ISA 250. Basic Math for Analytic  
## [11] "ISA 277. Independent Studies. (0  
## [12] "ISA 281. Concepts in Business Pr  
## [13] "ISA 291. Applied Regression Anal  
## [14] "ISA 301. Business Data Communica  
## [15] "ISA 303. Enterprise Systems. (3)  
## [16] "ISA 305. Information Technology
```

# 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 an  or 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 = "2022/", domain = "planecrashinfo.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 ⓘ

<b>The Appliance Repair Do..</b> 4.6 ★★★★★ · <a href="#">See reviews</a> ✓ Google guaranteed Alameda (510) 871-3938 Open now	<b>Healthy Duct Cleaning S...</b> 4.9 ★★★★★ · <a href="#">See reviews</a> ✓ Google guaranteed Daly City (415) 993-1965 Open now	<b>Atlas Trillo Heating &amp; Air</b> 4.5 ★★★★★ · <a href="#">See reviews</a> ✓ 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, ...

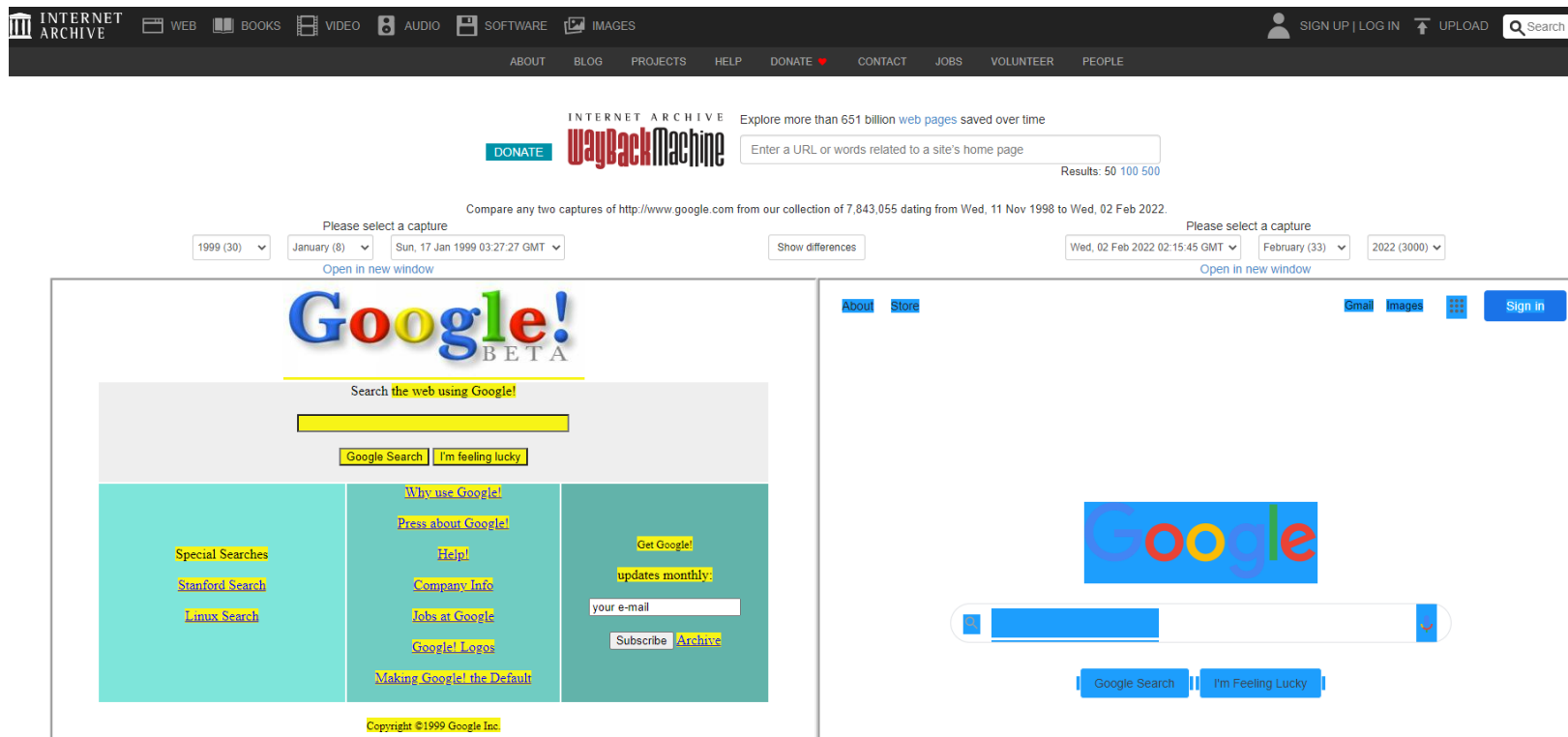
**Best Hvac contractors in San Francisco, CA - Yelp**  
[https://www.yelp.com/search?find\\_desc=hvac+contractors&find\\_loc=San+Francisco+CA](https://www.yelp.com/search?find_desc=hvac+contractors&find_loc=San+Francisco+CA)

# 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

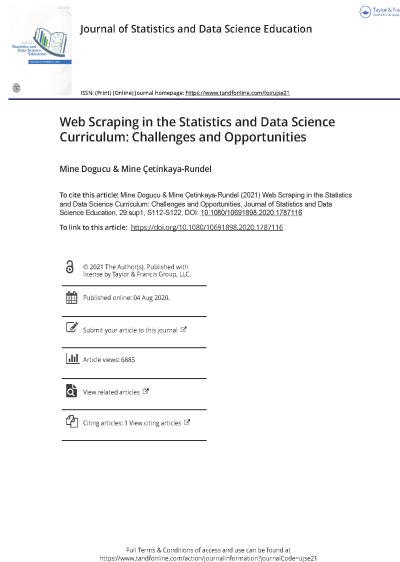
# Summary of Main Points

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

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

# 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