Web Scraping in the Statistics Curricula: Challenges and Opportunities

Mine Dogucu

Denison University

2018-11-14

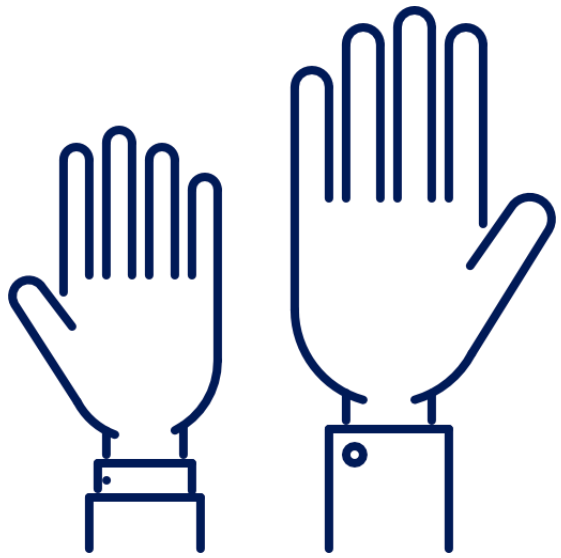
@MineDogucu

bit.ly/scrape_duke

METHODOLOGICAL

APPLIED - COLLOBARATIVE

PEDAGOGICAL



Have you ever scraped data from the web?

WHAT?

WHY?

HOW?

WHEN?

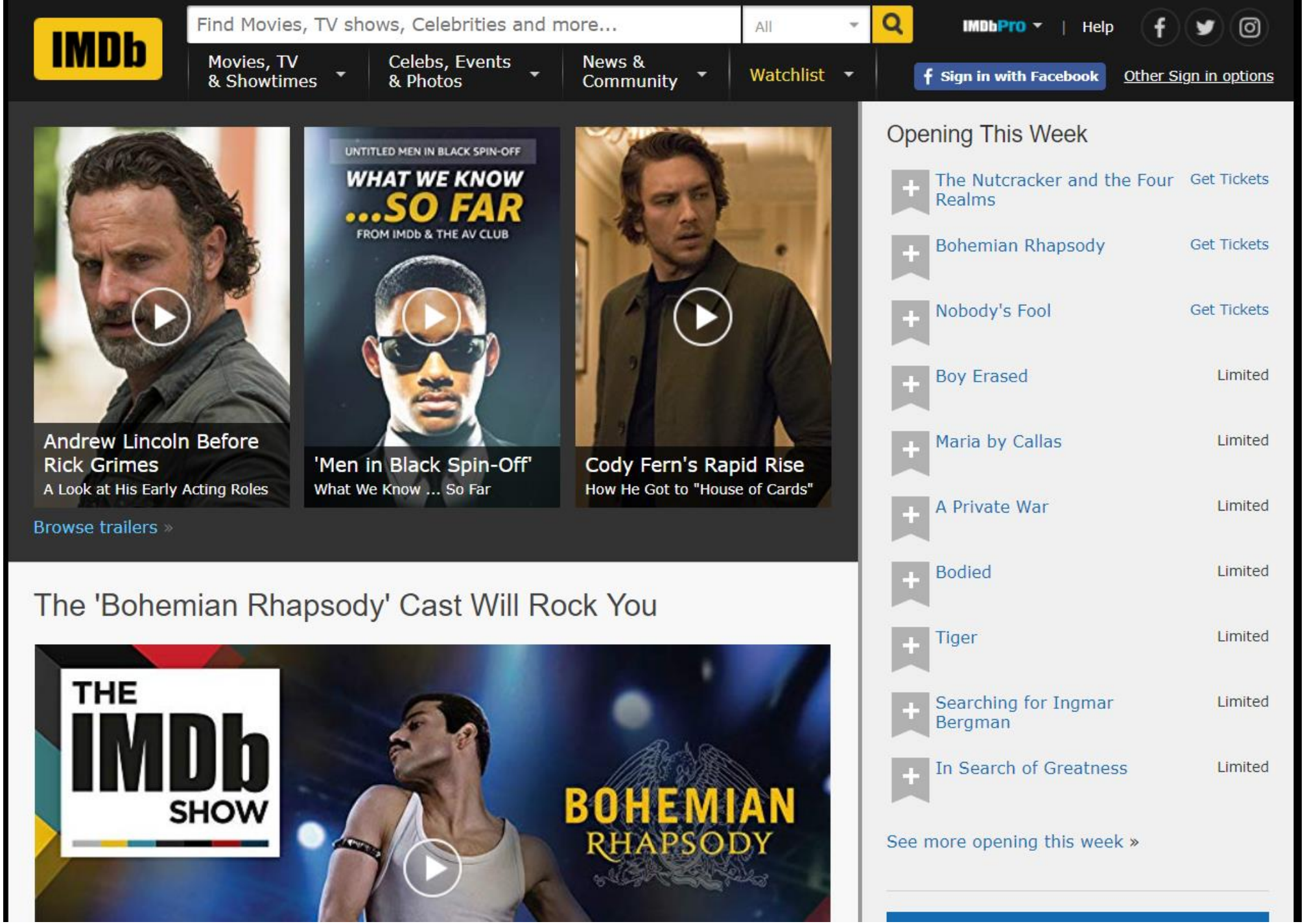
WHAT?

WHY?

HOW?

WHEN?


Web Scraping
Protocol
CSV



Link - Retrieved on 2018-11-04

Support the IMDb by visiting our sponsors.

Search The Database

Take Our User Survey

What's New at IMDb?

Site Index

IMDb feature of the day: [an error occurred while processing this directive]

New to the IMDb? Check out the site introduction or take our tour.
The most comprehensive free source of movie information on the Internet.
Sponsorship and advertising details.

Bandwidth provided by Exec-PC's FilePile

Search - Help - Index - Tour - Linking

Copyright © 1990-1997 The Internet Movie Database Ltd

has provided the graphics for this site.

This site is powered by FreeBSD and Apache

Link – Retrieved on 2018-10-25

Movie credits

As of November 1996 there were over **85,000** titles in the Internet Movie Database, ranging from some of the earliest moving pictures such as Express Train on a Railway Cutting (1898) from last century to several currently in production (we can't show you those, because it'll mean updating this text on a regular basis.. trust us.) and many recent releases

Each week the number of titles in the database grows to cover almost all the new releases, as well as to fill some holes from earlier days.

All types of movie are covered, for example

- blockbusters - Sleepless in Seattle (1993),
- classics - Citizen Kane (1941),
- cult following - Rocky Horror Picture Show, The (1975),
- silent - Kid, The (1921),
- bad - Plan 9 from Outer Space (1958),
- animated - Wrong Trousers, The (1993),
- Hollywood - Fugitive, The (1993),
- world-wide - Belle Epoque (1992),
- and much, much more.

Link – Retrieved on 2018-10-25

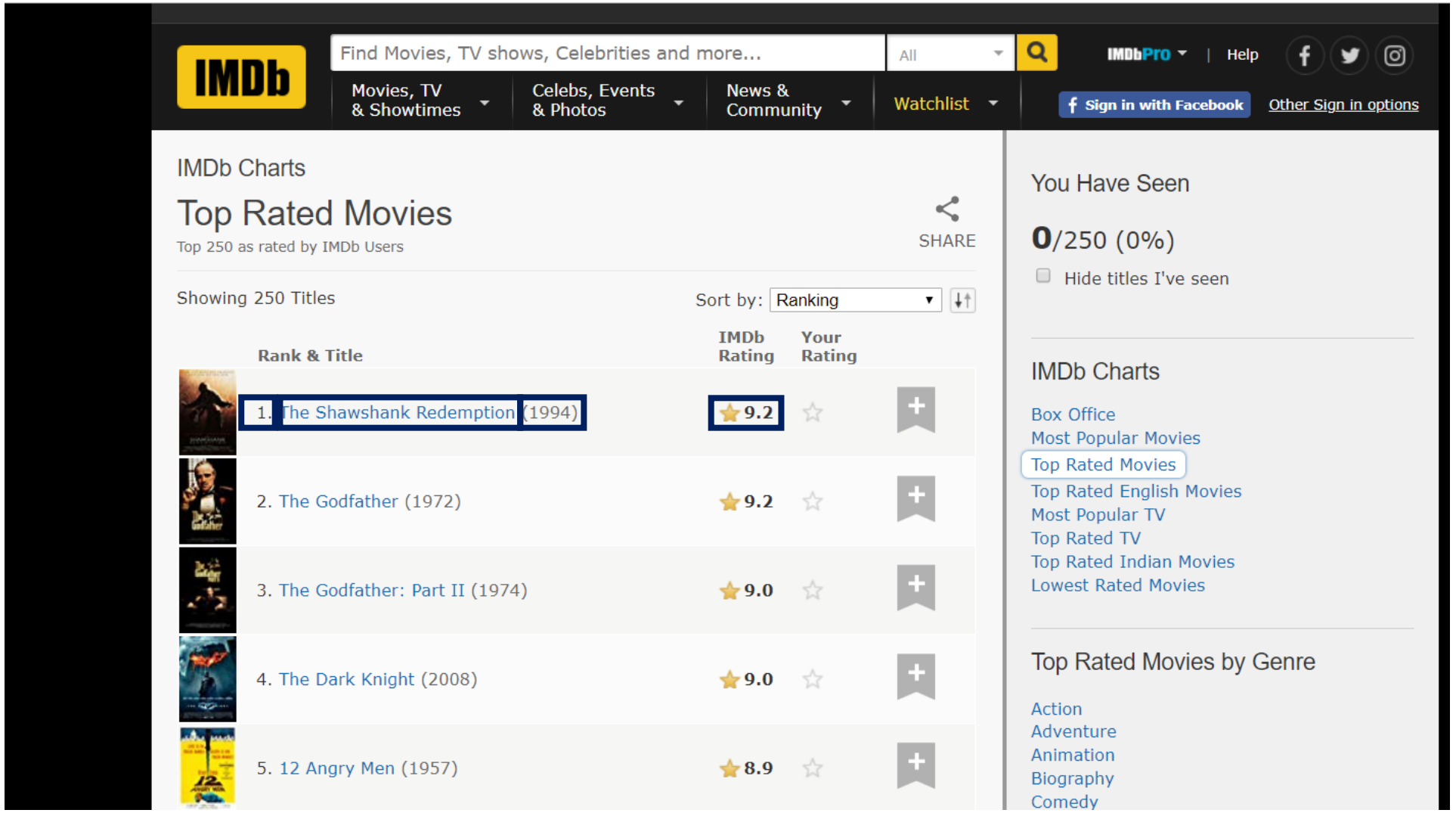
Titles: 5,310,913 (Year Range: 1874 - 2025)

- Titles w/ primary image 1,058,871
- Reviews 3,793,406
- Plots 1,807,393
- Trivia 964,144
- Quotes 907,178
- Parental guides 416,387
- Release dates 6,026,114
- Certificates 1,103,092
- Genres 2,863,475
- Keywords 8,150,624
- Running times 1,687,265
- Soundtrack 916,062

Link – Retrieved on 2018-10-25

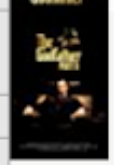
Top 250 movies

https://www.imdb.com/chart/top



Retrieved on 2018-10-25

Hand Scraping

	A	B	C	D	E
1		Rank & Title	IMDb Rating	Your Rating	
2		1. The Shawshank Redemption (1994)	9.2		
3					
4					
5		2. The Godfather (1972)	9.2		
6					
7					
8		3. The Godfather: Part II (1974)	9		
9					
10					
11		4. The Dark Knight (2008)	9		
12					
13					
14		5. 12 Angry Men (1957)	8.9		
15					
16					

.xls

	A	B	C	D
1		Rank & Tit	IMDb Ratii	Your Rating
2		1. The Sha	9.2	
3				
4				
5		2. The Goc	9.2	
6				
7				
8		3. The Goc	9	
9				
10				
11		4. The Dar	9	
12				
13				
14		5. 12 Angr	8.9	
15				

.csv

Web Scraping

	title	year	rating
1	The Shawshank Redemption	1994	9.2
2	The Godfather	1972	9.2
3	The Godfather: Part II	1974	9.0
4	The Dark Knight	2008	9.0
5	12 Angry Men	1957	8.9
6	Schindler's List	1993	8.9
7	The Lord of the Rings: The Return of the King	2003	8.9
8	Pulp Fiction	1994	8.9
9	The Good, the Bad and the Ugly	1966	8.8
10	Fight Club	1999	8.8
11	The Lord of the Rings: The Fellowship of the Ring	2001	8.8

Feature Films Released in 2018

1-50 of 14,485 titles. | Next »

View Mode: Compact | **Detailed**

Sort by: **Popularity▲** | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

1. **Bohemian Rhapsody** (2018)

PG-13 | 134 min | Biography, Drama, Music

8.4 Rate this 49 Metascore

A chronicle of the years leading up to Queen's legendary appearance at the Live Aid (1985) concert.

Director: Bryan Singer | Stars: Rami Malek, Lucy Boynton, Gwilym Lee, Ben Hardy

Votes: 72,662 | Gross: $69.16M

2. **Halloween** (I) (2018)

R | 106 min | Horror, Thriller

7.1 Rate this 67 Metascore

Laurie Strode confronts her long-time foe Michael Myers, the masked figure who has haunted her since she narrowly escaped his killing spree on Halloween night four decades ago.

Director: David Gordon Green | Stars: Jamie Lee Curtis, Judy Greer, Andi Matichak, James Jude Courtney

Votes: 44,006 | Gross: $152.97M

3. **A Star Is Born** (2018)

R | 136 min | Drama, Music, Romance

Retrieved on 2018-11-11

WHAT?

WHY?

HOW?

WHEN?

Six recommendations from GAISE

1. Teach statistical thinking.
   - Teach statistics as an investigative process of problem-solving and decision-making.
   - Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

Desirable Characteristics of Class Activities

Choosing data…

- Relevance – The activity should involve data about topics that interest students. Using real data makes data relevant to a wide variety of student majors. If real data are not used, then the activity should mimic a real-world situation. It should not seem like "busywork" to students. For example, if you use coins or cards to conduct a binomial experiment, explain real-world binomial experiments they could represent.

Statistics Education Literature

- Nature of statistics is changing. Statistics curricula must adapt to this change. One change needed is inclusion of computational skills (Nolan & Lang, 2010).
- Instructors and institutions that have computational skills will be on demand and students who have these skills will get the jobs (Horton, 2015).
- Some schools have already adopted “Data Science in Statistics Curricula” and they included web scraping (Hardin et. al, 2015).

Keynote by Chief Statistician of the United States, Nancy Potok

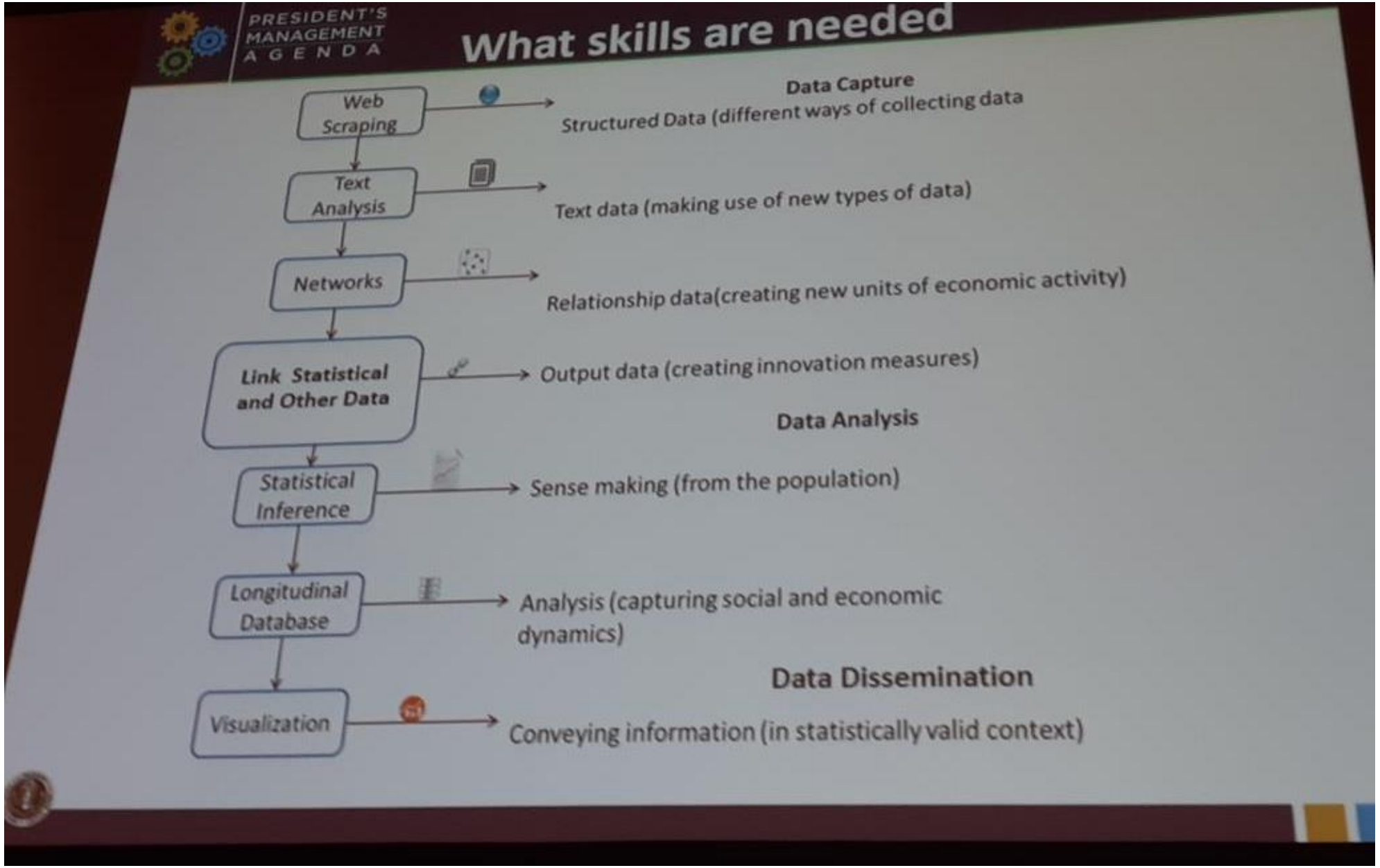


Photo:
Kim Flagg Sellers

indeed

Data Science Internship, Spring 2019

Weber Shandwick - Washington, DC 20005

General Responsibilities:

- Understand client background and needs, including general business strategy, industry issues, products and services, key customers and competitors in the marketplace
- Participate in strategic brainstorming sessions when invited by account leads or supervisors
- Create and maintain optimal data pipelines and data integration solutions
- Assemble large, complex data sets that meet business requirements through web scraping or API's
- Identify, design, and implement internal process improvements: automating manual processes, optimizing data delivery, re-designing infrastructure for greater scalability, etc.
- Mine and analyze data to drive optimization and improvement of campaign management, marketing techniques and business strategies

indeed

Digital Product Data Analyst

Galco Industrial Electronics - Madison Heights, MI 48071

TASKS, DUTIES, AND RESPONSIBILITIES

- Write & optimize database queries to transform raw, disparate datasets into our standard
- Process data from a range of formats, including CSV, JSON, XML
- Write basic scripts for data manipulation, data fetching, web scraping, and simple automation
- Drive continuous improvement of our processes and internal tools
- Drive automation capabilities with data entry and data extraction
- Working knowledge of SQL or Progress is a plus
- Ability to write and optimize queries
- Joins, aggregation, indexes, and transactions * Ability to write and automate basic scripts for data manipulation and retrieval including CSV, JSON, XML data formats
- General understanding of APIs, command line, client server architecture, networks, and (S)FTP
- Use Excel to standardize and structure product data into our format
- Familiar with load sheets, appropriate keywords, effective copy, and product classifications
- Research products to ensure data accuracy
- Work with internal and external resources (ex. Manufacturers) to ensure products are accurately described
- Maintain departmental records of progress and completed work

indeed

Research Analyst

Decision Resources Group - Parsippany, NJ

RESPONSIBILITIES

- Research, review, and enter pharmaceutical product management data efficiently and accurately for all therapeutic drug classes and for all payer segments: Commercial, PBM, Commercial Medicaid, State Medicaid and Medicare plans
- Work with the Analyst team to create best practices for capturing plan changes and management updates
- Leverage a set technology tools, including proprietary web scraping and natural language processing software to drive efficiency
- Collaborate with other DRG teams to facilitate a variety of data sets
- Support the timely and appropriate response to client inquiries regarding the data
- Serve as key support for the delivery of data, reports, and ad hoc research assignments

indeed

Data Analysis Librarian

UNC-Chapel Hill - Chapel Hill, NC

Minimum Qualifications:

Required ALA-accredited master's degree in Library or Information Science, or related advanced degree. Coursework in statistics, and knowledge of statistical methods. Strong customer service orientation and excellent communication and interpersonal skills. Demonstrated advanced data skills, including data cleaning/wrangling/normalization, using regular expressions, and web scraping. Demonstrated experience with data analysis tools such as R, STATA, SPSS, and SAS. Proficiency with at least one programming language (such as Python, Java, or R). Demonstrated aptitude for quickly learning new tools and technologies. Experience working effectively with a team to plan and complete projects. Demonstrated ability to work with diverse populations as well as demonstrated commitment to diversity, inclusion, and accessibility. Preferred Advanced degree in statistics. Experience providing data-related services in a library or research setting. Experience teaching technology, either one-on-one or in a classroom setting. Proficiency using tools and programming libraries to support text analysis. Geospatial technology skills. Data visualization skills.ALA-accredited master's degree in Library or Information Science, or related advanced degree.

WHAT?

WHY?

HOW?

WHEN?


Client
Server
1. Request
2. HTML document
3. Render

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb Users

SHARE

Showing 250 Titles

Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating
 1. The Shawshank Redemption (1994)		
 2. The Godfather (1972)		
 3. The Godfather: Part II (1974)		
 4. The Dark Knight (2008)	9.0	
 5. 12 Angry Men (1957)	8.9	

Back Alt+Left Arrow
Forward Alt+Right Arrow
Reload Ctrl+R
Save as... Ctrl+S
Print... Ctrl+P
Cast...
Translate to English
AdBlock
View page source Ctrl+U
Inspect Ctrl+Shift+I

```
<!DOCTYPE html>
<html
    xmlns:og="http://ogp.me/ns#"
    xmlns:fb="http://www.facebook.com/2008/fbml">
    <head>

<script type='text/javascript'>var ue_t0=ue_t0||+new Date();</script>
<script type='text/javascript'>
window.ue_ihb = (window.ue_ihb || window.ueinit || 0) + 1;
if (window.ue_ihb === 1) {

var ue_csm = window,
    ue_hob = +new Date();
(function(d){var e=d.ue=d.ue||{},f=Date.now||function(){return+new Date};e.d=function(b){return f()-(b?
0:d.ue_t0)};e.stub=function(b,a){if(!b[a]){var c=[];b[a]=function()
{c.push([c.slice.call(arguments),e.d(),d.ue_id])};b[a].replay=function(b){for(var
a;a=c.shift();)b(a[0],a[1],a[2])};b[a].isStub=1}};e.exec=function(b,a){return function(){if(1==window.ueinit)try{return
b.apply(this,arguments)}catch(c){ueLogError(c,{attribution:a||"undefined",logLevel:"WARN"})}}}})(ue_csm);


    var ue_err_chan = 'jserr';
(function(d,e){function h(f,b){if(!(a.ec>a.mxe)&&f){a.ter.push(f);b=b||{};var
c=f.logLevel||b.logLevel;c&&c!==k&&c!==m&&c!==n&&c!==p||a.ec++;c&&c!=k||a.ecf++;b.pageURL=""+(e.location?
e.location.href:"");b.logLevel=c;b.attribution=f.attribution||b.attribution;a.erl.push({ex:f,info:b})}}function
l(a,b,c,e,g){d.ueLogError({m:a,f:b,l:c,c:""+e,err:g,fromOnError:1,args:arguments},g?
{attribution:g.attribution,logLevel:g.logLevel}:void 0);return!1}var k="FATAL",m="ERROR",n="WARN",p="DOWNGRADED",a=
{ec:0,ecf:0,
pec:0,ts:0,erl:[],ter:[],mxe:50,startTimer:function(){a.ts++;setInterval(function()
{d.ue&&a.pec<a.ec&&d.uex("at");a.pec=a.ec},1E4)}};l.skipTrace=1;h.skipTrace=1;h.isStub=1;d.ueLogError=h;d.ue_err=a;e.oner
ror=l})(ue_csm,window);


var ue_id = 'SDKEQZE363X3YQV65TSF',
    ue_url,
    ue_navtiming = 1,
    ue_mid = 'A1EVAM02EL8SFB',
    ue_sid = '130-8783331-0922553',
    ue_sn = 'www.imdb.com',
```

```
["sitbreaderrightpageturner","sitbreaderleftpageturner","sitbreaderpagecontainer"].indexOf(a))return!0}function m()
{n=!1;h=0;b.clearTimeout(s)}function D(){b.ue.onSushiUnload(function(){ue.event({violationType:"unresponsive-
clicks",violationCount:g,totalScanned:k},"csm","csm.ArmoredCXGuardrailsViolation.3")});b.ue.onunload(function()
{ue.count("armored-cxguardrails.unresponsive-clicks.violations",
g);ue.count("armored-cxguardrails.unresponsive-
clicks.violationRate",g/k*100||0)})}if(b.MutationObserver&&b.addEventListener&&Object.keys&&d&&d.ue&&d.ue.log&&d.ue_unrt&
&d.ue_utils){var
z=d.ue_unrt,r="cel",w="unr_mcm",C="res_mcm",p=b.performance,f=d.ue_utils,n=!1,h=0,s=0,q=!1,l=0,g=0,k=0;b.addEventListener
&&
(b.addEventListener("mousedown",x,!0),b.addEventListener("beforeunload",m,!0),b.addEventListener("visibilitychange",m,!0)
,b.addEventListener("pagehide",m,!0));b.ue&&b.ue.event&&b.ue.onSushiUnload&&
b.ue.onunload&&D();(new MutationObserver(A)).observe(t,{childList:!0,attributes:!0,characterData:!0,subtree:!0})})}})
(ue_csm,window,document);


ue_csm.ue.exec(function(g,e){if(e.ue_err){var f="";e.ue_err.errorHandlers||(e.ue_err.errorHandlers=
[]);e.ue_err.errorHandlers.push({name:"fctx",handler:function(a)
{if(!a.logLevel||"FATAL"===a.logLevel)if(f=g.getElementsByTagName("html")[0].innerHTML){var b=f.indexOf("var
ue_t0=ue_t0||+new Date();");if(-1!=b){var b=f.substr(0,b).split("\n"),d=Math.max(b.length-5-1,0),b=b.slice(d,b.length-
1);a.fcsmln=b.length+1;a.cinfo=a.cinfo||{};for(var c=0;c<b.length;c++)a.cinfo[d+c+1+""]=b[c]}b=f.split("\n");a.cinfo=
a.cinfo||{};if(!(a.f||void 0===a.l||a.l in a.cinfo))for(c=+a.l-1,d=Math.max(c-2,0),c=Math.min(c+2,b.length-
1);d<=c;d++)a.cinfo[d+1+""]=b[d]}}})}},"fatals-context")(document,window);




}
/* ⚠ */
</script>

</div>

<noscript>
    <img height="1" width="1" style='display:none;visibility:hidden;' src='//fls-
na.amazon.com/1/batch/1/OP/A1EVAM02EL8SFB:130-8783331-
0922553:SDKEQZE363X3YQV65TSF$uedata=s:%2Fgp%2Fuedata%3Fnoscript%26id%3DSDKEQZE363X3YQV65TSF:0' alt=""/>
</noscript>

    </body>
</html>
```

Duke University is a private research university located in Durham, NC. *I am currently visiting Duke University.*

Denison University is a private liberal arts college located in Granville, OH.

Hyper Text Markup Language tags

```
<html>

</html>
```

HTML tags

```
<html>
<head>

</head>
<body>

</body>

</html>
```

```
<html>
<head>
</head>
<body>
        <p>Duke University is a private research university
            located in Durham, NC. I am currently
            visiting Duke University.
        </p>
        <p> Denison University is a private liberal arts
             college located in Granville, OH.
        </p>
</body>
</html>
```

What we have

Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH.

```
<html>
<head>
</head>
<body>
        <p>Duke University is a private research university
            located in Durham, NC. I am currently
            visiting Duke University.
        </p>
        <p> Denison University is a private liberal arts
             college located in Granville, OH.
        </p>
</body>
</html>
```

What we would like

Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH

```
<a href = "https://www.duke.edu">Duke
    University </a> is a private research university
    located in Durham, NC. I am currently visiting
    Duke University.
```

Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Duke University

<a > HTML tag

href attribute (name)

https://www.duke.edu attribute(value)

Duke University content

```
<html>
<head>
</head>
<body>
        <p><a href = "https://www.duke.edu">Duke
            University</a> is a private research university
            located in Durham, NC. I am currently visiting
            Duke University.
        </p>
        <p><a href ="https://denison.edu">  Denison
            University</a> is a private liberal arts college
            located in Granville, OH.
        </p>
</body>
</html>
```

Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH

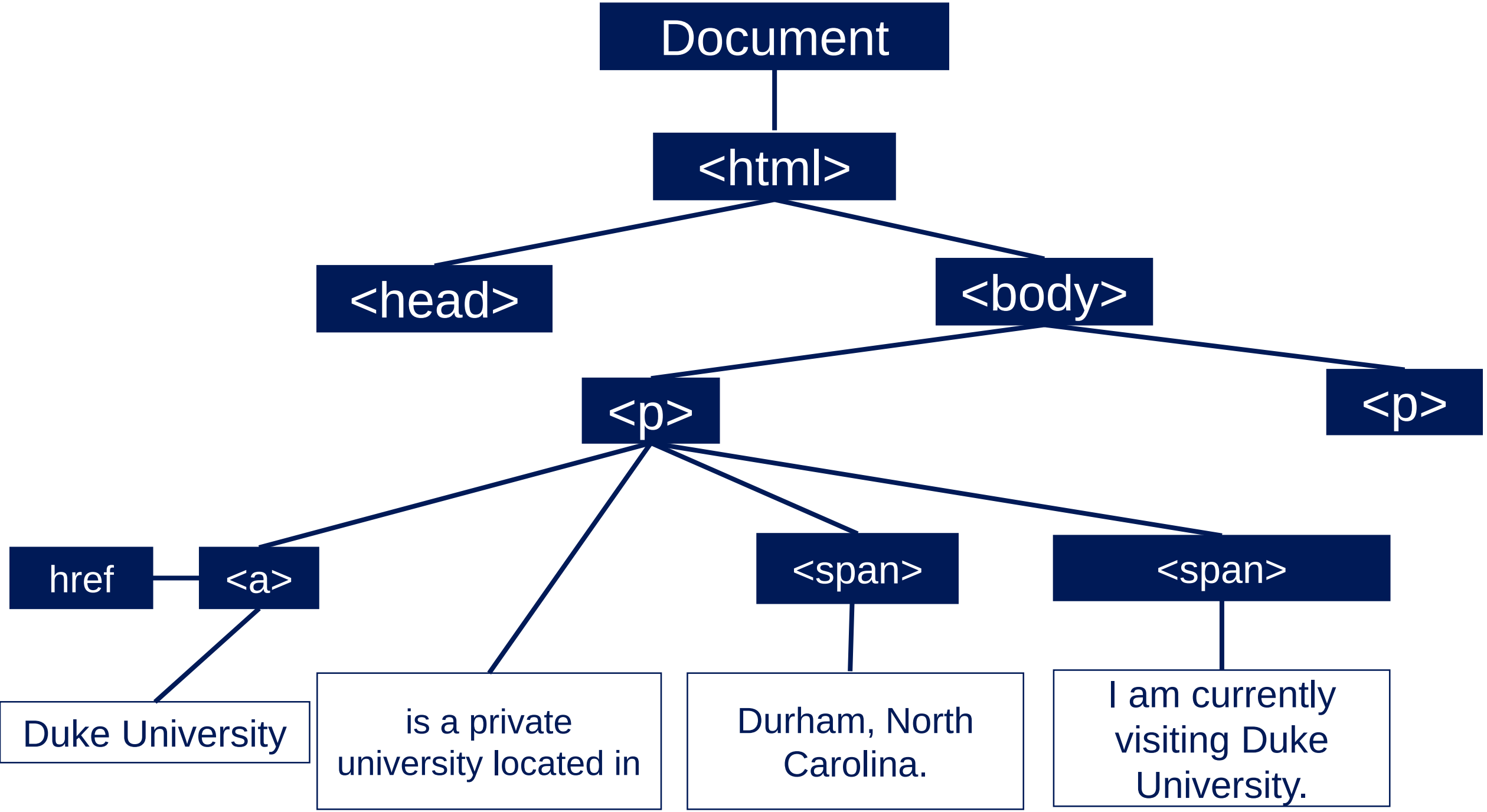
Duke University is a private research university located in Durham, NC. *I am currently visiting Duke University.*

Denison University is a private liberal arts college located in Granville, OH.

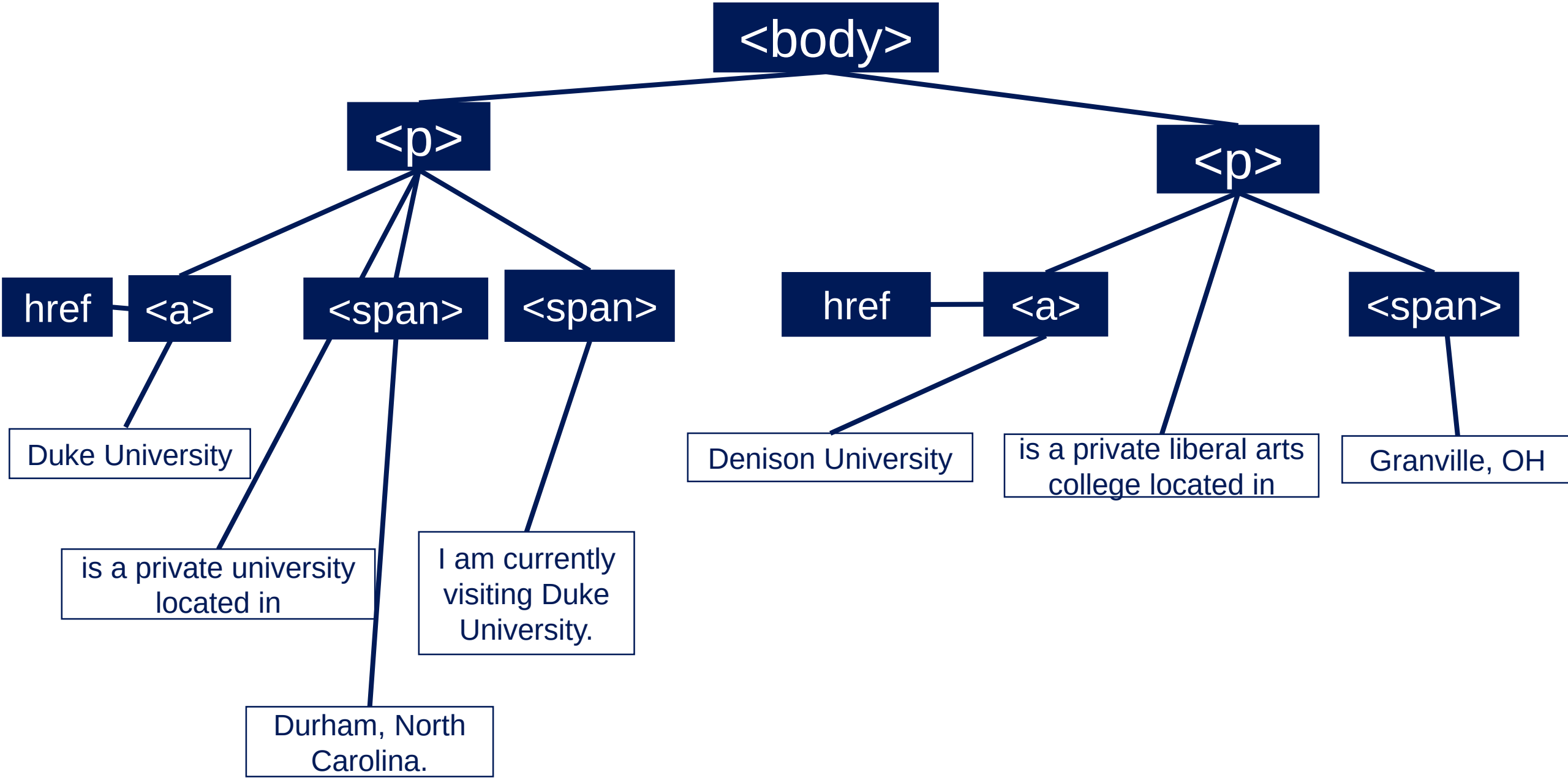
```
<html>
<head>
</head>
<body>
        <p><a href = "https://www.duke.edu">Duke
            University</a> is a private research university
            located in <span> Durham, NC </span>.
            <span> I am currently visiting Duke University.
            </span>
        </p>
        <p><a href ="https://denison.edu">  Denison
            University</a> is a private liberal arts college
            located in <span> Granville, OH</span>.
        </p>
</body>
</html>
```

Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH


Document
<html>
<head>
<body>
<p>
<p>
href
<a>

Duke University
is a private university located in
Durham, North Carolina.
I am currently visiting Duke University.


<body>
<p>
<p>
href
<a>

href
<a>

Duke University
is a private university located in
Durham, North Carolina.
I am currently visiting Duke University.
Denison University
is a private liberal arts college located in
Granville, OH

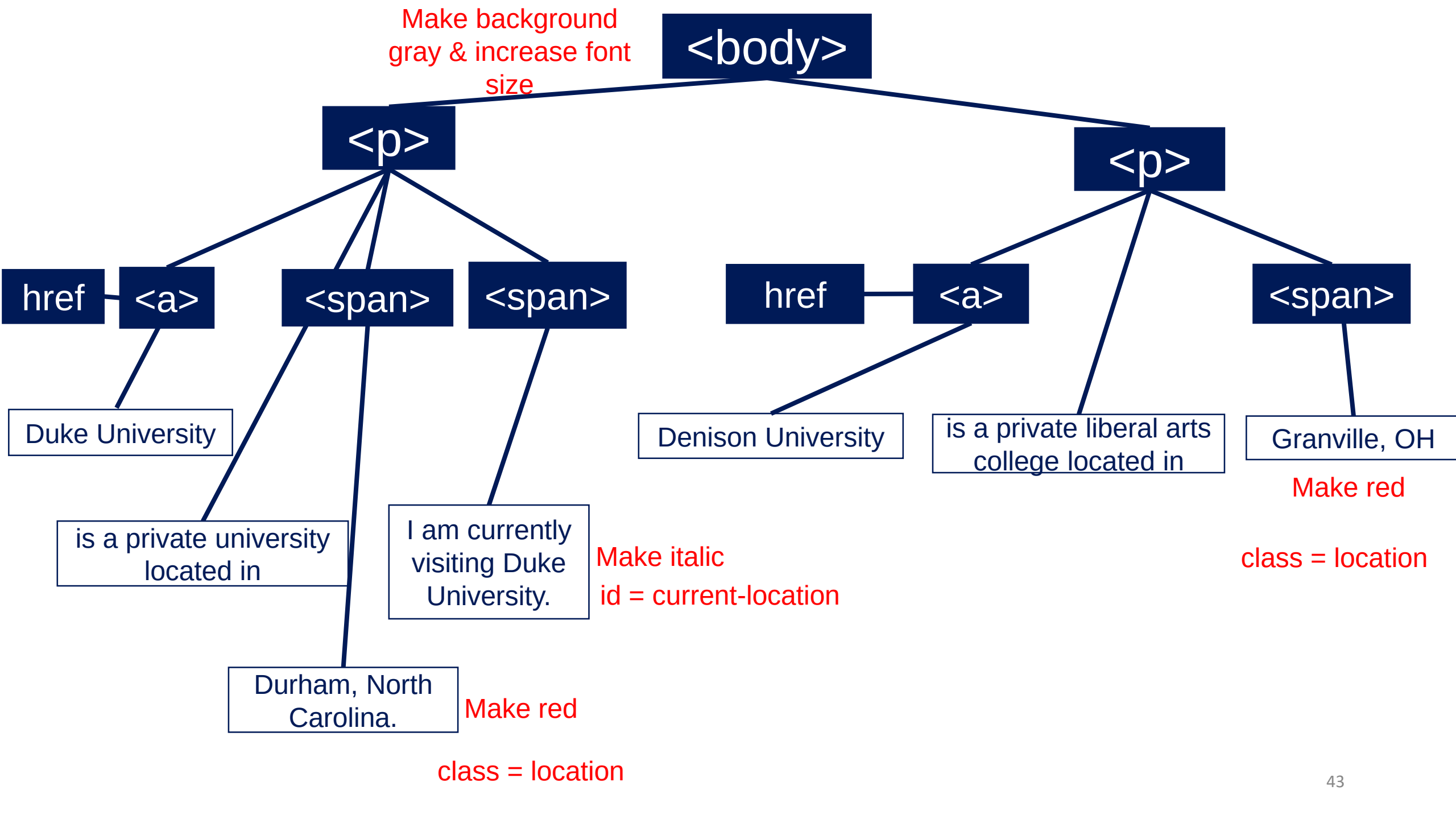
Duke University is a private research university located in Durham, NC. I am currently visiting Duke University.

Denison University is a private liberal arts college located in Granville, OH

Duke University is a private research university located in Durham, NC. *I am currently visiting Duke University.*

Denison University is a private liberal arts college located in Granville, OH.


HTML
CSS


Make background gray & increase font size
<body>
<p>
<p>
href
<a>

href
<a>

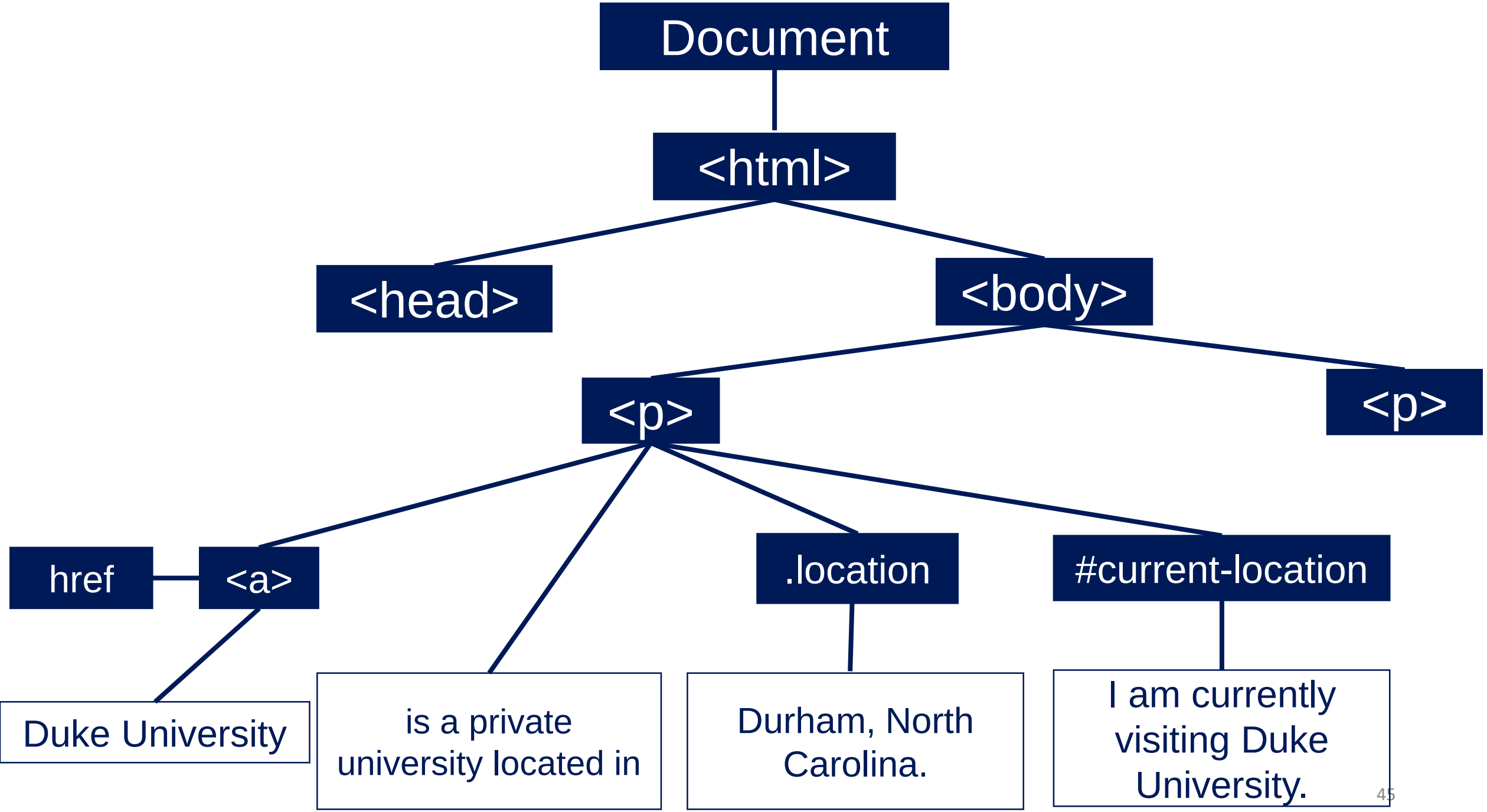
Duke University
is a private university located in
I am currently visiting Duke University.
Durham, North Carolina.
Make red
class = location
Make italic
id = current-location
Denison University
is a private liberal arts college located in
Granville, OH
Make red
class = location

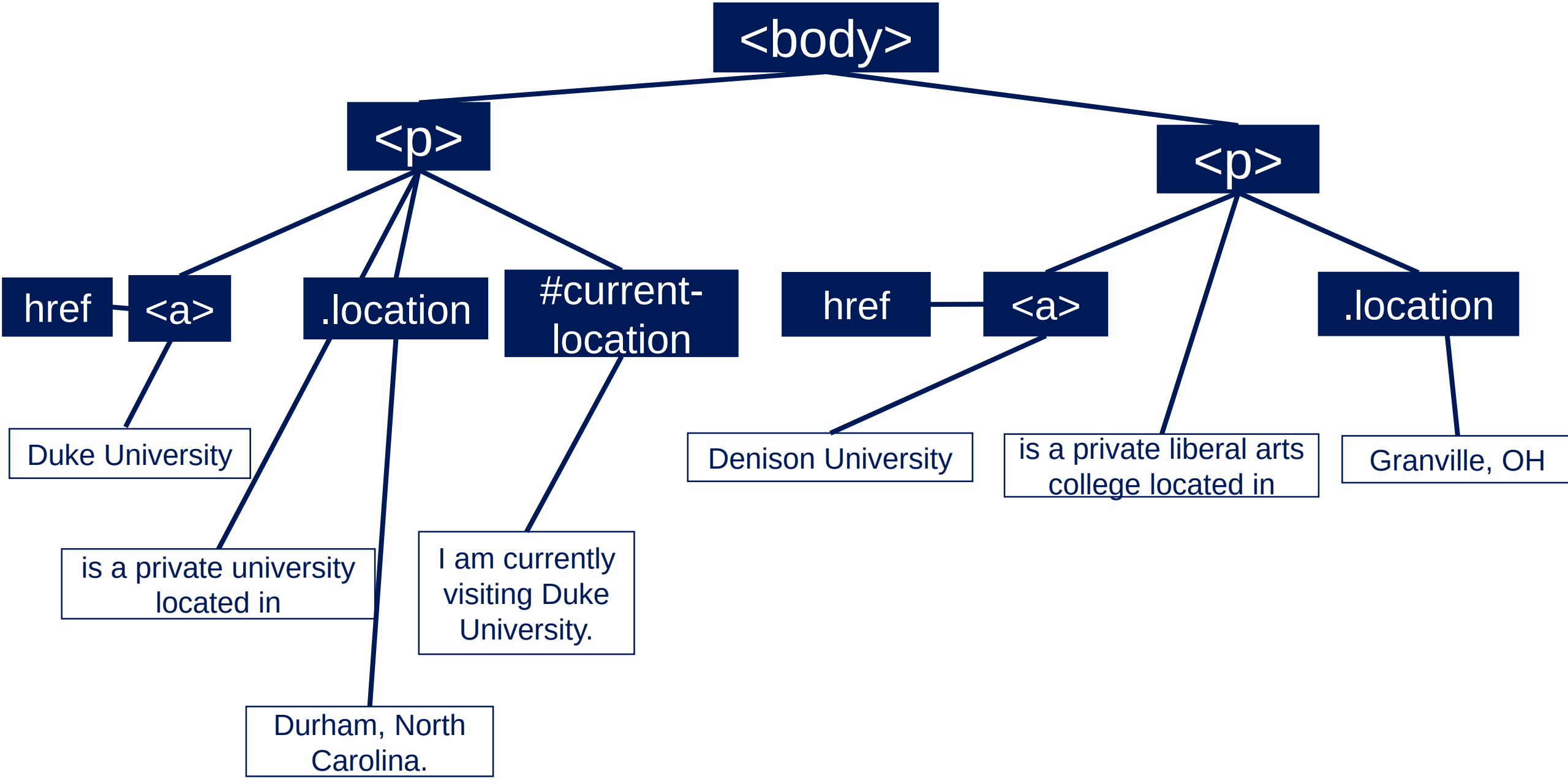
```
<html>
<head>
        <style>
        body{
                background-color: lightgray;
                font-size: 20px;
        }
        .location{
                color: red;
        }
        #current-location{
        font-style: italic;
        }
        </style>
</head>
```

```
<body>
        <p><a href = "https://www.duke.edu">Duke
            University</a> is a private research
            university located in <span class =
            “location”> Durham, NC </span>. <span id =
            “current-location”>I am currently visiting
            Duke University </span>.
        </p>
        <p><a href ="https://denison.edu">  Denison
            University</a> is a private liberal arts college
            located in <span class= “location”> Granville,
            OH </span>.
        </p>
</body>
</html>
```

Duke University is a private research university located in Durham, NC. *I am currently visiting Duke University.*

Denison University is a private liberal arts college located in Granville, OH.


Document
<html>
<head>
<body>
<p>
<p>
href
<a>
.location
#current-location
Duke University
is a private university located in
Durham, North Carolina.
I am currently visiting Duke University.

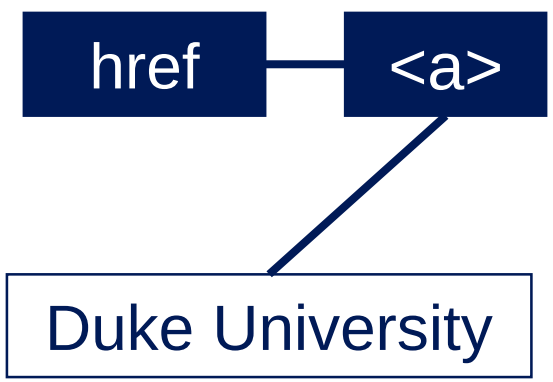

<body>
<p>
<p>
href
<a>
.location
#current-location
href
<a>
.location
Duke University
is a private university located in
Durham, North Carolina.
I am currently visiting Duke University.
Denison University
is a private liberal arts college located in
Granville, OH

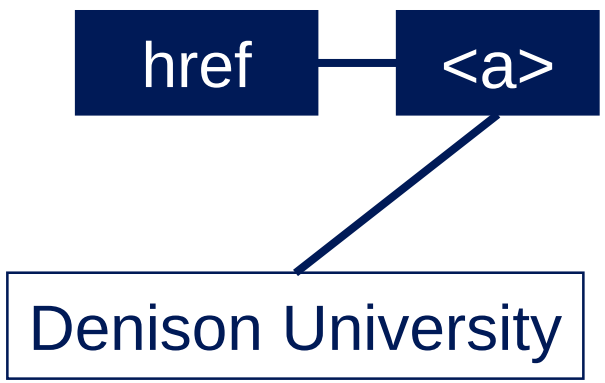

rvest
www.rstudio.com

rvest::read_html()

<html>

rvest::html_nodes("a")



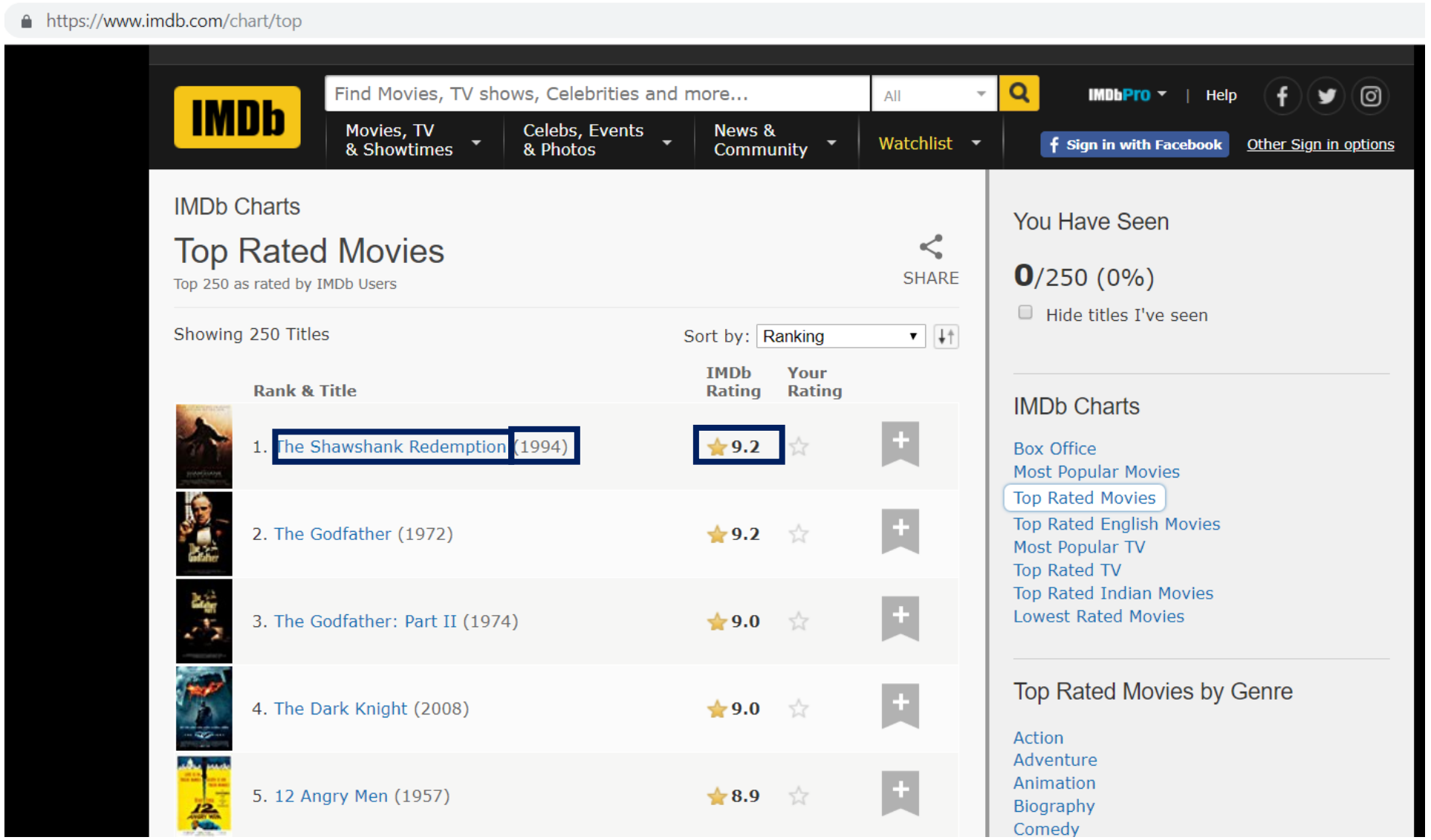


rvest::html_text()

Duke University

Denison University

Top 250 movies



Retrieved on 2018-10-25

chrome web store

SelectorGadget

Offered by: selectorgadget.com

★★★★½ 70 | Developer Tools | 86,860 users

Add to Chrome

Overview Reviews Support Related

bit.ly/SelectorGadget

https://www.imdb.com/chart/top

IMDb | Find Movies, TV shows, Celebrities and more... | All | IMDbPro | Help

Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist | Sign in with Facebook | Other Sign in options

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb Users

SHARE

Showing 250 Titles

Sort by: Ranking

	Rank & Title	IMDb Rating	Your Rating	
	1. The Shawshank Redemption (1994)	9.2		+
	2. The Godfather (1972)	9.2		+
	3. The Godfather: Part II (1974)	9.0		+
	4. The Dark Knight (2008)	9.0		+
	5. 12 Angry Men (1957)	8.9		+
	6. Schindler's List (1993)	8.9		+
	7. The Lord of the Rings: The Return of the King (2003)	8.9		+

You Have Seen

0/250 (0%)

Hide titles I've seen

IMDb Charts

Box Office
Most Popular Movies
Top Rated Movies
Top Rated English Movies
Most Popular TV
Top Rated TV
Top Rated Indian Movies
Lowest Rated Movies

Top Rated Movies by Genre

Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror

No valid path found. | Clear | Toggle Position | XPath | ? | X

HTML Nodes

Our Interest	Node
Title	.titleColumn a
Year	.secondaryInfo
Rating	.ratingColumn.imdbRating

Title: .titleColumn a

Year: .secondaryInfo

```
<td class="titleColumn">
    1.
    <a href="/title/tt0111161/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=e31d89dd-322d-4646-8962-
327b42fe94b1&pf_rd_r=3BTCBAC4HJN0GK95J818&pf_rd_s=center-
1&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_tt_1"
title="Frank Darabont (dir.), Tim Robbins, Morgan Freeman" >The Shawshank Redemption</a>
    <span class="secondaryInfo">(1994)</span>
</td>
<td class="ratingColumn imdbRating">
        <strong title="9.2 based on 2,011,548 user ratings">9.2</strong>
</td>
```

Year: .ratingColumn.imdbRating

Step 1 - rvest::read_html()

Download the HTML file

```
> read_html("http://www.imdb.com/chart/top")
{xml_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.fa
cebook.com/2008/fbml">
[1] <head>\n<meta http-equiv="Content-Type" content=" ...
[2] <body id="styleguide-v2" class="fixed">\n\n       ...
>
```

Step 2 - rvest::html_nodes()

Extract specific nodes

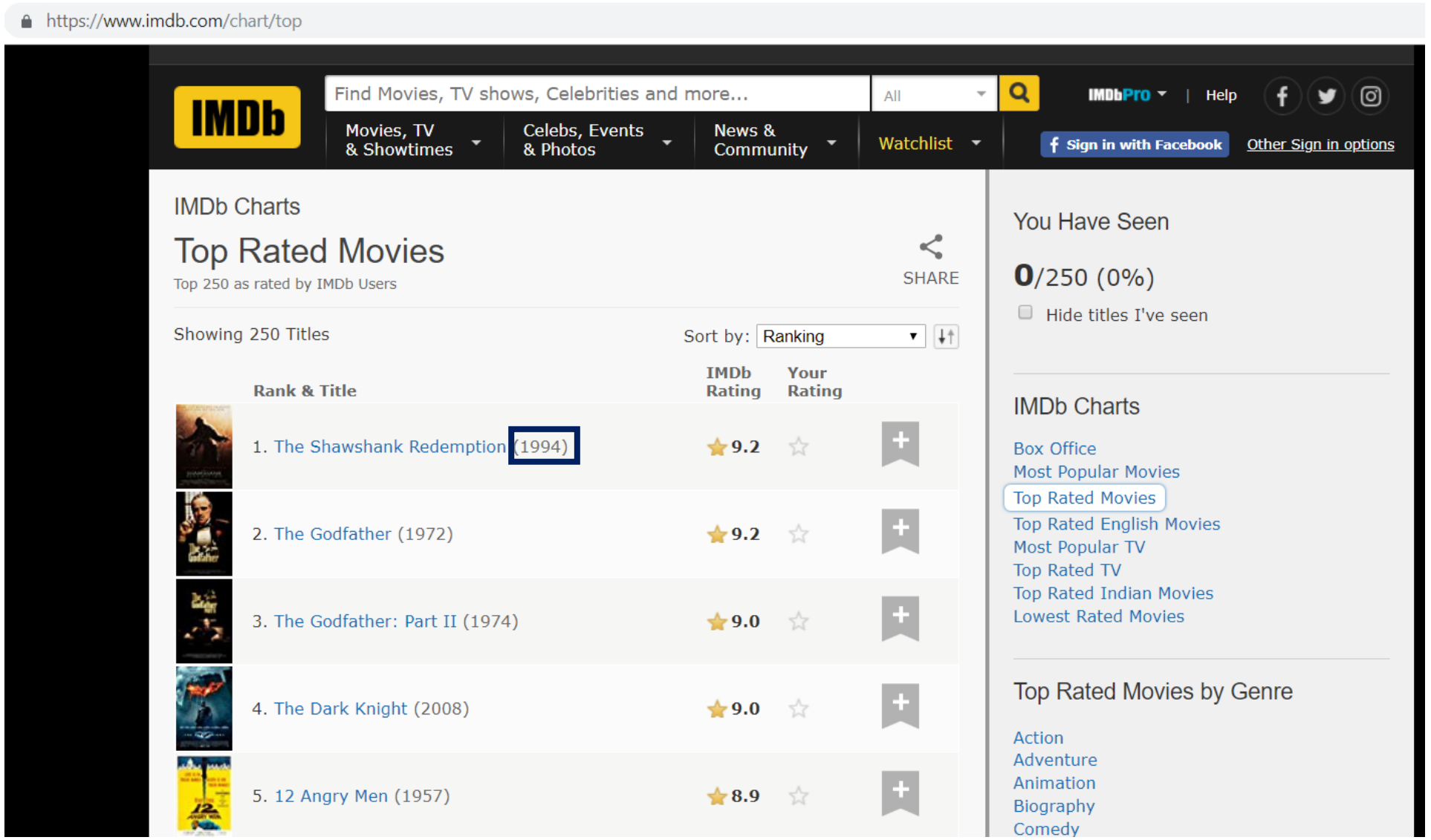
```
> read_html("http://www.imdb.com/chart/top") %>%
+   html_nodes(".titleColumn a")
{xml_nodeset (250)}
 [1] <a href="/title/tt0111161/?pf_rd_m=A2FGELUUNOQJN ...
 [2] <a href="/title/tt0068646/?pf_rd_m=A2FGELUUNOQJN ...
 [3] <a href="/title/tt0071562/?pf_rd_m=A2FGELUUNOQJN ...
 [4] <a href="/title/tt0468569/?pf_rd_m=A2FGELUUNOQJN ...
 [5] <a href="/title/tt0050083/?pf_rd_m=A2FGELUUNOQJN ...
 [6] <a href="/title/tt0108052/?pf_rd_m=A2FGELUUNOQJN ...
 [7] <a href="/title/tt0167260/?pf_rd_m=A2FGELUUNOQJN ...
```

Step 3 - rvest::html_text()

Extract text

```
> read_html("http://www.imdb.com/chart/top") %>%
+   html_nodes(".titleColumn a") %>%
+   html_text()
  [1] "The Shawshank Redemption"

  [2] "The Godfather"

  [3] "The Godfather: Part II"

  [4] "The Dark Knight"

  [5] "12 Angry Men"
```

Top 250 movies



Retrieved on 2018-10-25

HTML Nodes

Our Interest	Node
Title	.titleColumn a
Year	.secondaryInfo
Rating	.ratingColumn.imdbRating

```
> read_html("http://www.imdb.com/chart/top") %>%
+   html_nodes(".secondaryInfo") %>%
+   html_text()
  [1] "(1994)" "(1972)" "(1974)" "(2008)" "(1957)"
  [6] "(1993)" "(2003)" "(1994)" "(1966)" "(1999)"
 [11] "(2001)" "(1994)" "(1980)" "(2010)" "(2002)"
 [16] "(1975)" "(1990)" "(1999)" "(1954)" "(2002)"
 [21] "(1995)" "(1977)" "(1991)" "(1946)" "(1997)"
 [26] "(1995)" "(2001)" "(1998)" "(1994)" "(1999)"
 [31] "(2014)" "(1960)" "(1998)" "(1931)" "(1968)"
 [36] "(1942)" "(1936)" "(2011)" "(2002)" "(2006)"
 [41] "(1991)" "(1985)" "(2014)" "(1954)" "(1981)"
 [46] "(2000)" "(1994)" "(2006)" "(1979)" "(2000)"
 [51] "(2018)" "(1979)" "(1940)" "(1988)" "(1988)"
 [56] "(1950)" "(2006)" "(1964)" "(1957)" "(1980)"
```

```
> read_html("http://www.imdb.com/chart/top") %>%
+   html_nodes(".secondaryInfo") %>%
+   html_text() %>%
+   str_remove("\\(") %>%                    # remove (
+   str_remove("\\)") %>%                    # remove )
+   as.numeric()
  [1] 1994 1972 1974 2008 1957 1993 2003 1994 1966 1999
 [11] 2001 1994 1980 2010 2002 1975 1990 1999 1954 2002
 [21] 1995 1977 1991 1946 1997 1995 2001 1998 1994 1999
 [31] 2014 1960 1998 1931 1968 1942 1936 2011 2002 2006
 [41] 1991 1985 2014 1954 1981 2000 1994 2006 1979 2000
 [51] 2018 1979 1940 1988 1988 1950 2006 1964 1957 1980
 [61] 2012 2008 1997 1957 1999 2012 2003 2017 1986 1984
 [71] 1981 1941 1995 1958 1959 1992 1983 2016 1931 2016
```

Web scraping

```
> read_html("http://www.imdb.com/chart/top") %>%
+   html_nodes(".secondaryInfo") %>%
+   html_text() %>%
```

String manipulation

```
+   str_remove("\\(") %>%                    # remove (
+   str_remove("\\)") %>%                    # remove )
+   as.numeric()
```

```r
# Read the entire page ----
page <- read_html("http://www.imdb.com/chart/top")

# Scrape titles ----
titles <- page %>%
  html_nodes(".titleColumn a") %>%
  html_text()

# Scrape years ----
years <- page %>%
  html_nodes(".secondaryInfo") %>%
  html_text() %>%
  str_remove("\\(") %>%                 # remove (
  str_remove("\\)") %>%                 # remove )
  as.numeric()

# Scrape ratings ----
ratings <- page %>%
  html_nodes("#main strong") %>%
  html_text() %>%
  as.numeric()

# Save titles, years, and rating together in a tibble ----
imdb_top_250 <- tibble(
  title = titles,
  year = years,
  rating = ratings
)
```

	title	year	rating
1	The Shawshank Redemption	1994	9.2
2	The Godfather	1972	9.2
3	The Godfather: Part II	1974	9.0
4	The Dark Knight	2008	9.0
5	12 Angry Men	1957	8.9
6	Schindler's List	1993	8.9
7	The Lord of the Rings: The Return of the King	2003	8.9
8	Pulp Fiction	1994	8.9
9	The Good, the Bad and the Ugly	1966	8.8
10	Fight Club	1999	8.8
11	The Lord of the Rings: The Fellowship of the Ring	2001	8.8

WHAT?

WHY?

HOW?

WHEN?

Prior Knowledge (Before Learning Web Scraping)

Data types – character, factor, integer, double

Experience working with rectangular data

Familiarity with R packages, functions, arguments

Knowledge Building (While Learning Web Scraping)

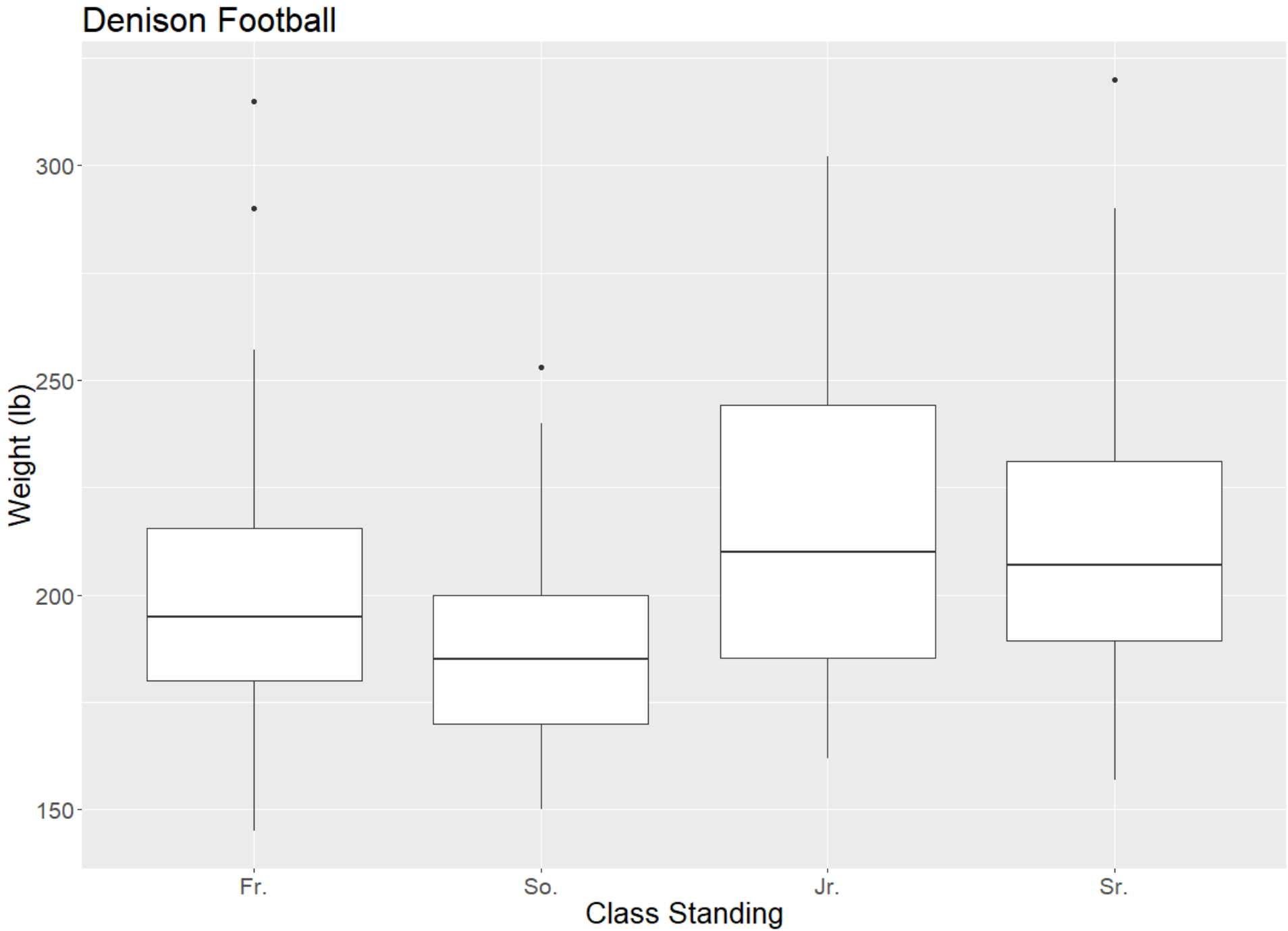
Optional

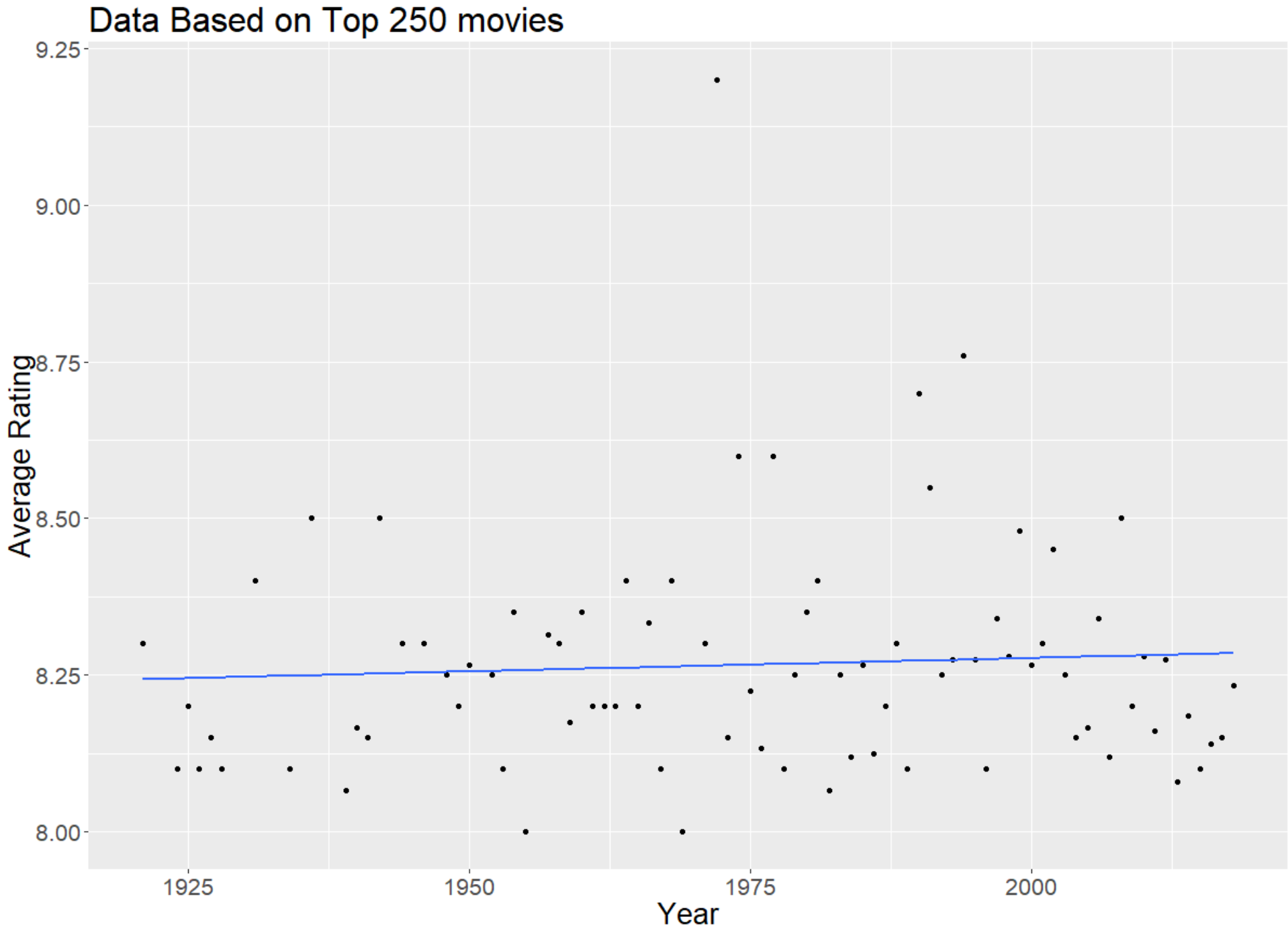
Optional

Optional

Subsequent Activities (After Having Learned Web Scraping)

STATISTICS


Denison Football
300
250
200
150
Weight (lb)
Fr.
So.
Jr.
Sr.
Class Standing


Data Based on Top 250 movies
9.25
9.00
8.75
8.50
8.25
8.00
Average Rating
1925
1950
1975
2000
Year

Predicting Movie Ratings

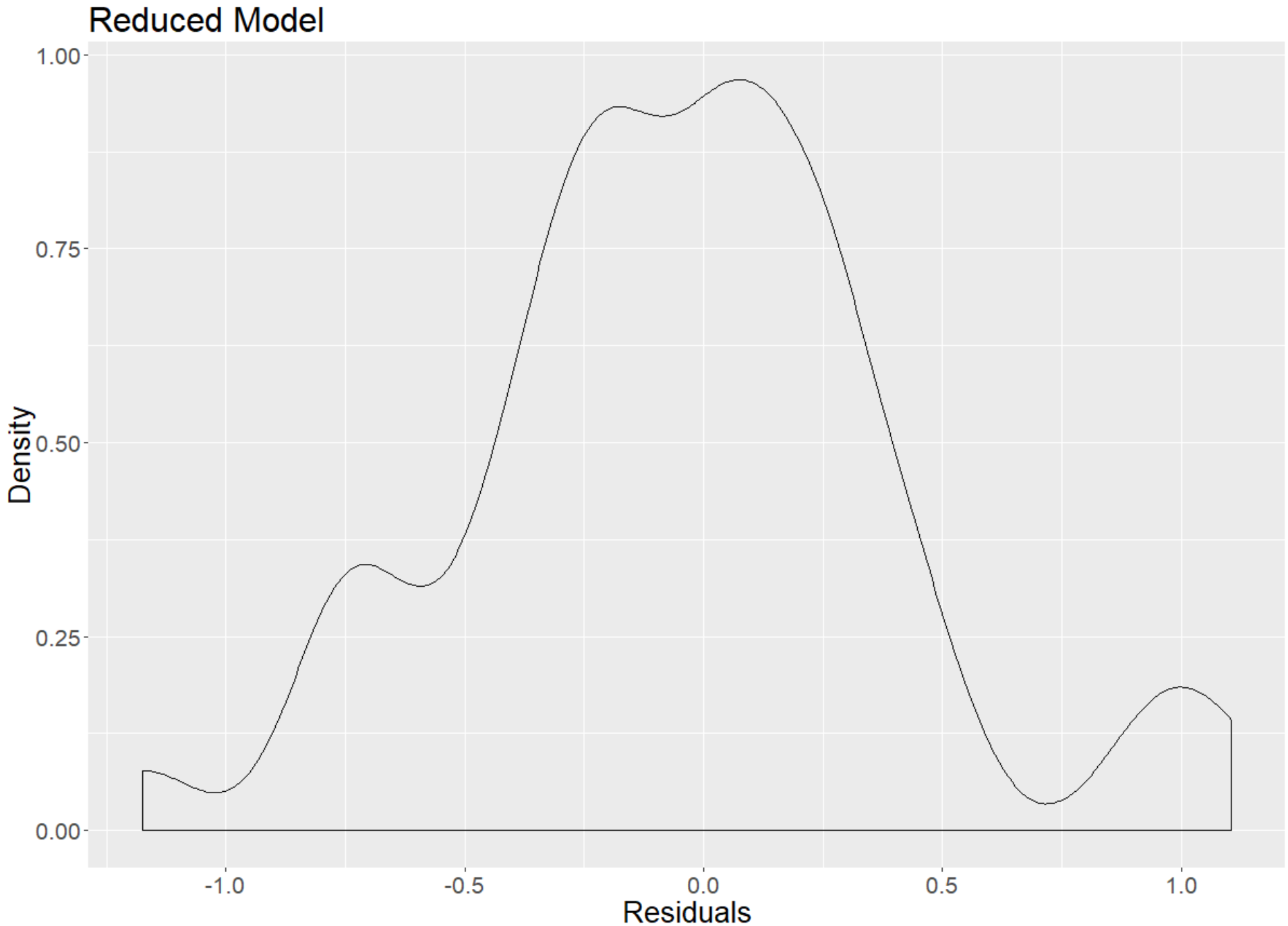
```
> tidy(full_model)
# A tibble: 5 x 5
  term           estimate   std.error statistic  p.value
  <chr>             <dbl>       <dbl>     <dbl>    <dbl>
1 (Intercept)  4.74        0.178           26.7   1.60e-93
2 meta_score   0.0243      0.00130         18.7   1.51e-57
3 run_time     0.00492     0.00142          3.46  5.85e- 4
4 vote         0.00000115  0.000000231      4.99  8.86e- 7
5 gross       -0.000151    0.000201        -0.751 4.53e- 1
```

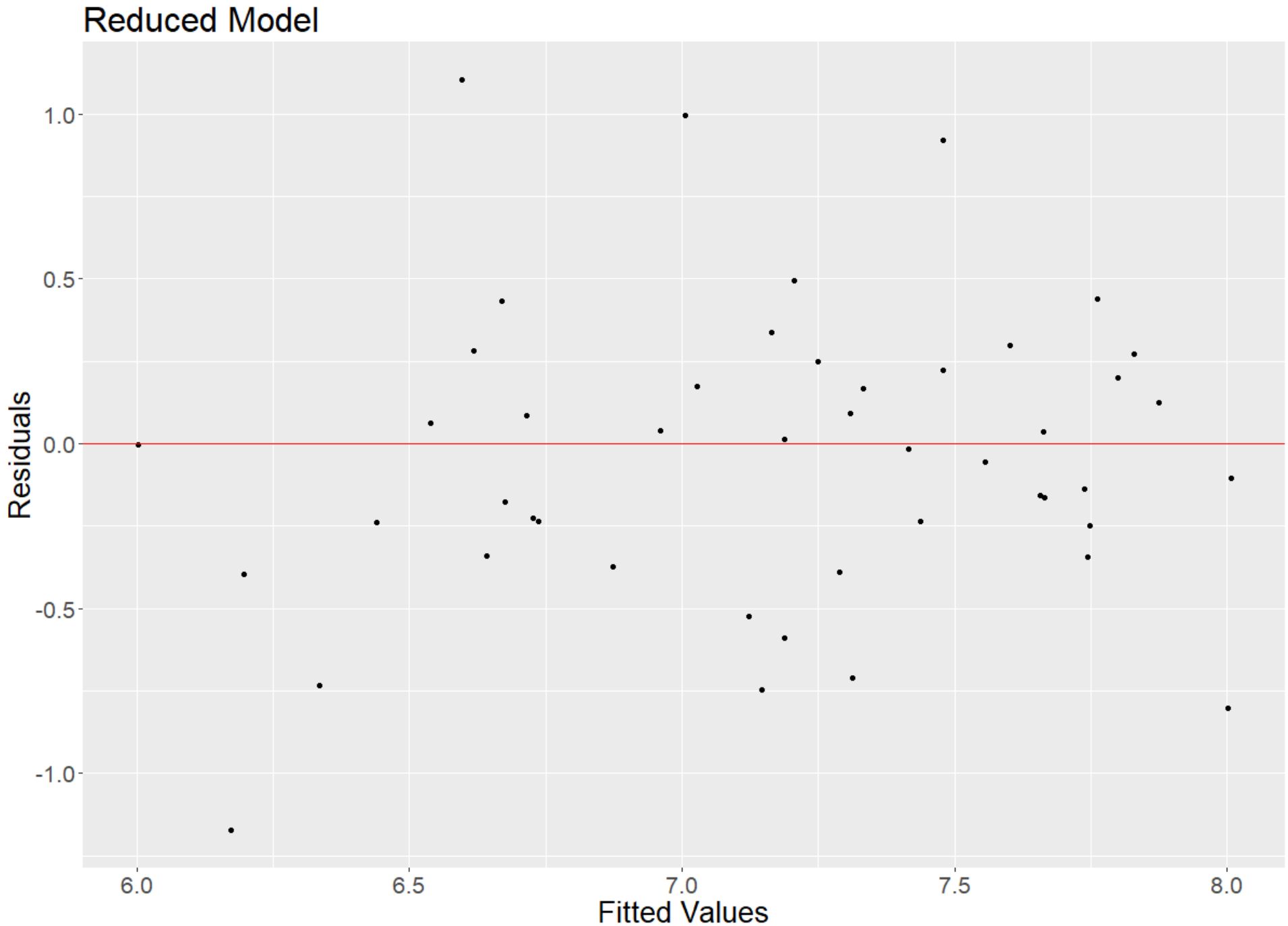
```
> reduced_model <- step(full_model, direction = "backward"
)
```

```
> tidy(reduced_model)
# A tibble: 4 x 5
  term           estimate   std.error statistic  p.value
  <chr>             <dbl>       <dbl>     <dbl>    <dbl>
1 (Intercept) 4.74        0.178             26.7  1.14e-93
2 meta_score  0.0246      0.00125           19.7  2.96e-62
3 run_time    0.00477     0.00141            3.39 7.52e- 4
4 vote        0.00000103  0.000000170        6.07 2.80e- 9
```

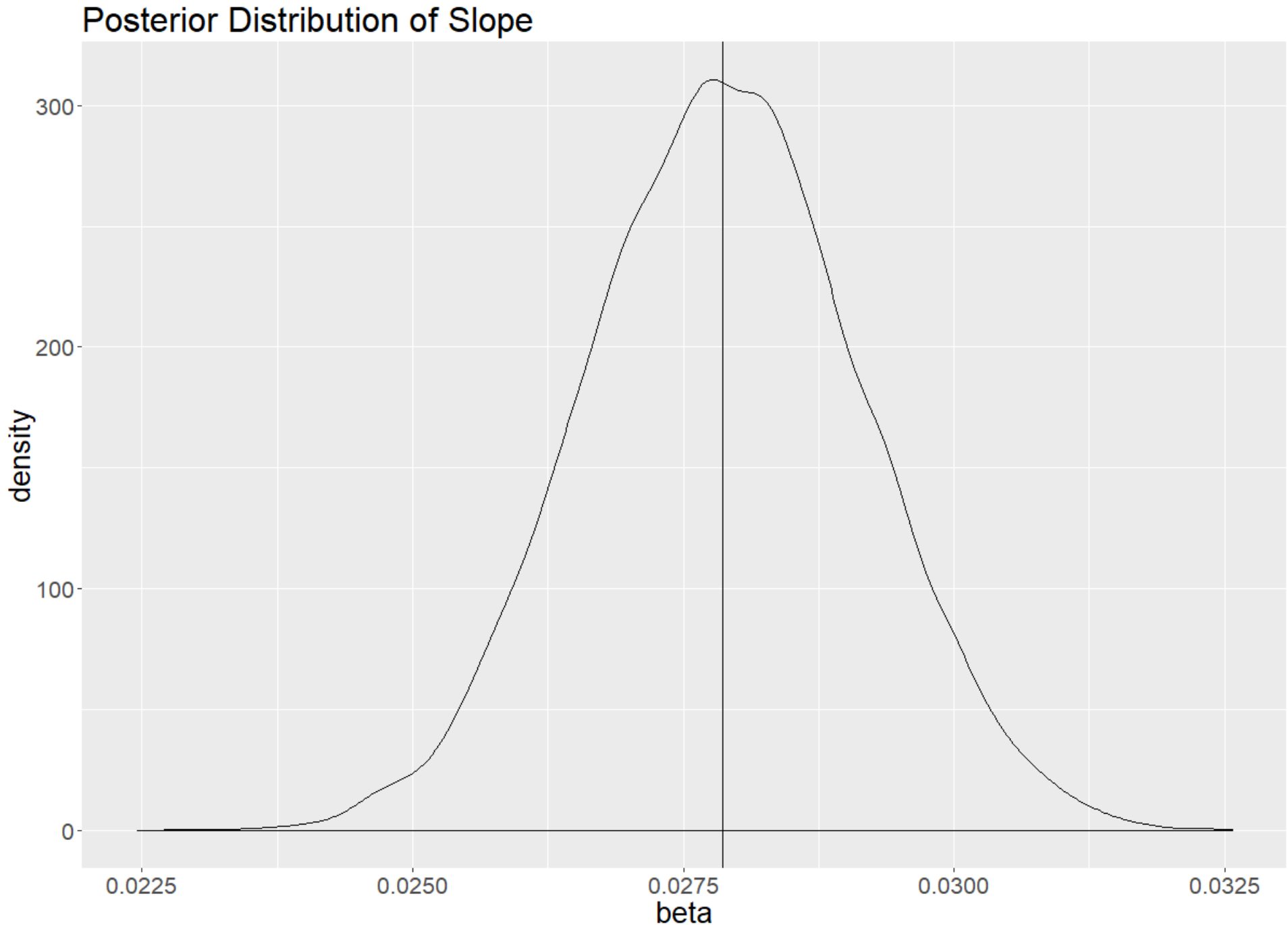
```
> glance(full_model)$AIC
[1] 496.9882
```

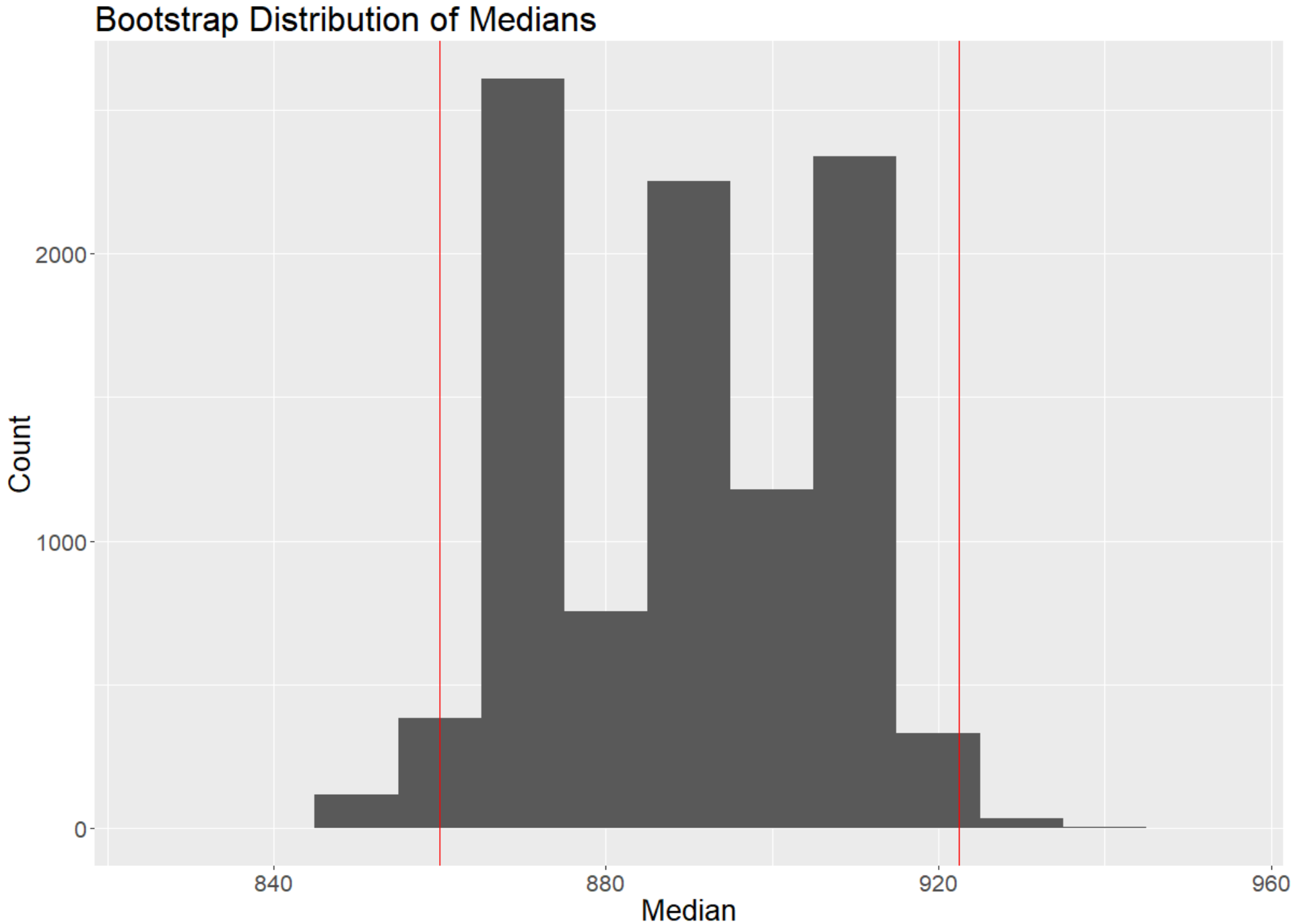
```
> glance(reduced_model)$AIC
[1] 495.5577
```


Reduced Model
1.00
0.75
0.50
0.25
0.00
Density
-1.0
-0.5
0.0
0.5
1.0
Residuals


Reduced Model
1.0
0.5
0.0
-0.5
-1.0
Residuals
6.0
6.5
7.0
7.5
8.0
Fitted Values

```
> model.script <- "
+ data {
+ int N;
+ real sigma;
+ real rating[N];
+ real meta_score[N];
+ }
+ transformed data {
+ real m_meta_score;
+ m_meta_score = mean(meta_score);
+ }
+ parameters {
+ real alpha;
+ real beta;
+ }
+ model {
+ alpha ~ normal(0,1/.001);
+ beta ~ normal(0,1/.001);
+ for (i in 1:N)
+ rating[i] ~ normal(alpha +
+                    beta*(meta_score[i]- m_meta_score),
+                   sigma);
+ }"
```


Posterior Distribution of Slope
300
200
100
0
density
0.0225
0.0250
0.0275
0.0300
0.0325
beta


Bootstrap Distribution of Medians
Count
2000
1000
0
840
880
920
960
Median

When?

- After prior topics have been mastered
- After some data maturity has been gained
- Before the desired subsequent activity

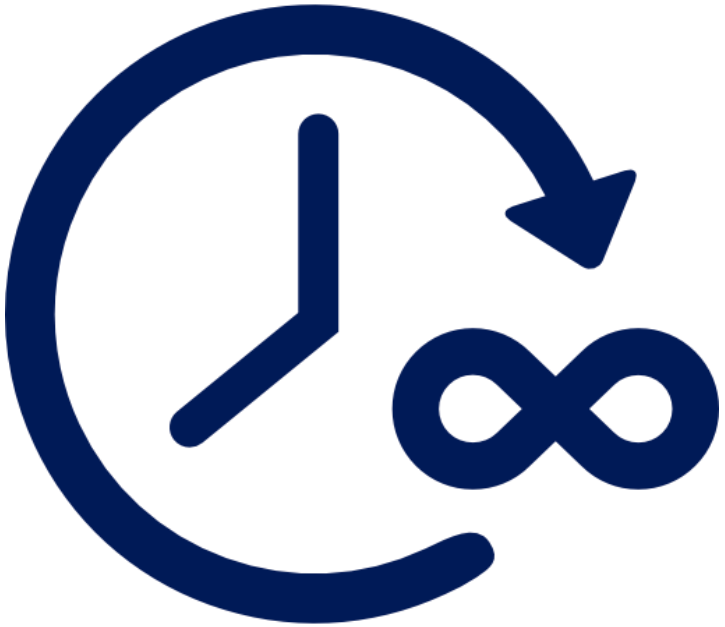
WHAT?

WHY?

HOW?

WHEN?

Making use of the written code



Duke
TRINITY COLLEGE OF ARTS & SCIENCES

STATISTICAL SCIENCE

About | Events | News | Resources

SEARCH

Undergraduate | M.S. Program | Ph.D. Program | Research | Courses | People | Alumni

StatSci @ Duke

The Department of Statistical Science is nationally ranked in the top 5 research departments and as a top 10 graduate program. Recognized as the world's leading center for Bayesian statistics and its interdisciplinary applications, the Department is a Duke campus hub for statistical and computational research. The Department administers and teaches a broad range of undergraduate statistics courses, including introductory courses and more advanced courses for the Statistical Science major, as well as many courses in statistics and related areas at the M.S. and Ph.D. level.

Learn More

Link – Retrieved 2018-10-25

ISDS Home (no frames)

Conferences

Courses

Directory

Discussion papers

FTP Server

Seminars

ISDS info

C. I. S.

Stats Sites: Bayesian

Stats Sites: General

Duke Home

Duke Library

Duke Phone Directory

Institute of Statistics & Decision Sciences

Duke University

WHAT'S NEW?

- Jim Berger is new *Annals* editor
- Statistics Week 1997
- *Award Winners:*
  - Dalene Stangl wins Teaching Award
  - Mike West wins Mitchell Prize

DEPARTMENTAL INFO

- ISDS FAQs, ftp, CIS server, etc.

RESEARCH

- Faculty research areas
- Seminar series
- Discussion papers
- Conferences

GRADUATE STUDIES

- Graduate program descriptions

TEACHING

- Course homepages
- Graduate course listings
- Undergraduate course listings

PEOPLE

- Faculty
- Staff
- Students
- Visitors
- Alumni

OTHER SITES

- Bayesian statistics sites
- Other Statistics sites

CONSULTING

- Statistical consulting center

webmaster@stat.duke.edu

Last Revised: *January 26, 1998*

Date of archive – 1998-02-13

Link – Retrieved on 2018-10-25

https://columbus.craigslist.org/d/apts-housing-for-rent/sea

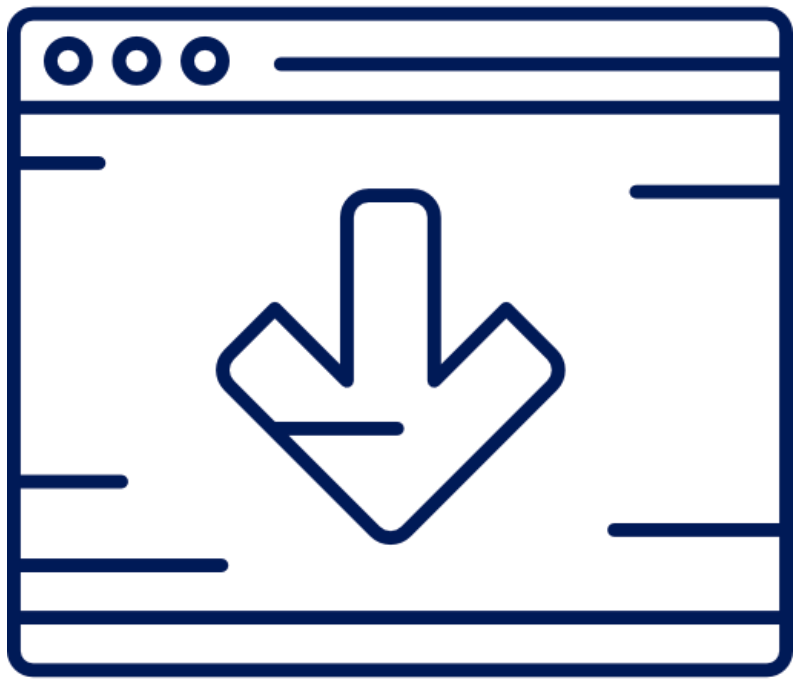
Latest

Show All

Hrm.

The Wayback Machine has not archived that URL.

Link – Retrieved on 2018-10-25



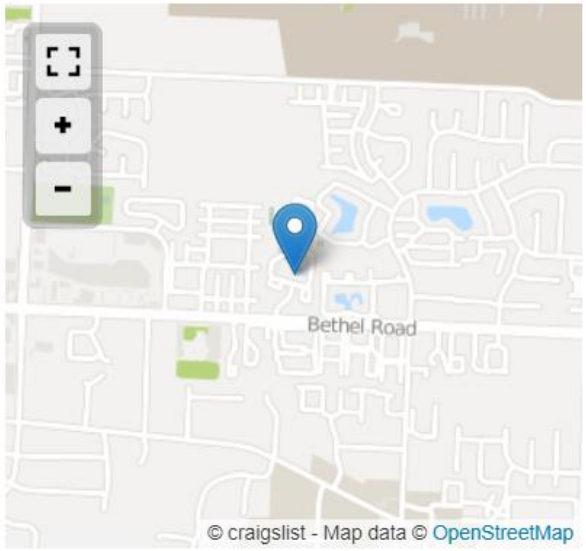
Website server can be down

Data entry

☆ **$00 / 2br - Won't Last! Most Requested 2 Bedroom with Loft in Columbus! 1 Left (Columbus)**

image 1 of 4

Luxury lakeside community just 9 minutes from The Ohio State University featuring 1, 2 and 3 bedroom floor plans!



5399 Coachman Rd

(google map)

2BR / 2Ba available now

cats are OK - purrr

dogs are OK - wooof

apartment

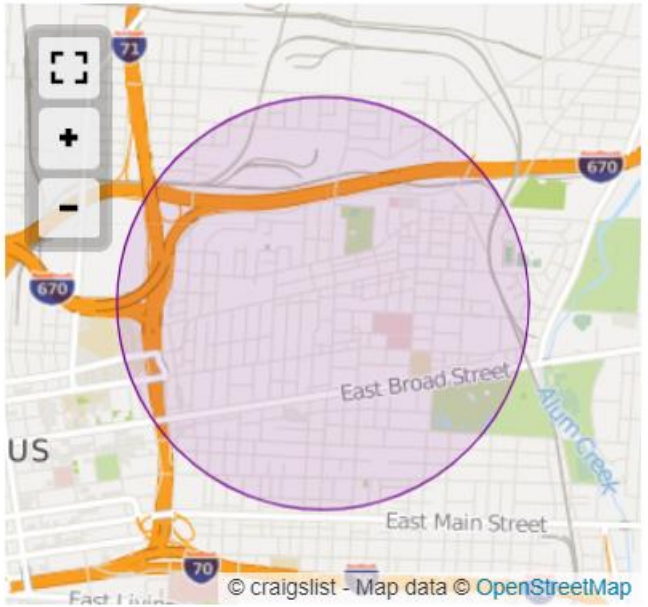
w/d in unit

off-street parking

Link – Retrieved 2018-11-11

☆ $500 / 1250ft^2 - 2 Bed, 2 B@th, 2 Parking St@ll

image 1 of 2



(google map)

1250ft^2

apartment

Come see this Portland Tower 2BR, 2BA with 2 indoor heated parking stalls and huge walkout patio with southern exposure and gas grill hook up! Located just one block from Skyway, 3 blocks from Downtown East Commons, Light Rail Station and US Bank Stadium, this downtown condo is an amazing location!

Link – Retrieved 2018-11-11

Feature Films Released in 2018– Missing Nodes

29. **The Grinch** (2018)

PG | 90 min | Animation, Comedy, Family

★ **6.1** ☆ Rate this 50 Metascore

A grumpy Grinch plots to ruin Christmas for the village of Whoville.

Directors: Yarrow Cheney, Scott Mosier | Stars: Benedict Cumberbatch, Cameron Seely, Rashida Jones, Pharrell Williams

Votes: 1,484

30. **Creed II** (2018)

PG-13 | Drama, Sport | **Completed**

Under the tutelage of Rocky Balboa, newly crowned light heavyweight champion Adonis Creed faces off against Viktor Drago, the son of Ivan Drago.

Director: Steven Caple Jr. | Stars: Tessa Thompson, Sylvester Stallone, Michael B. Jordan, Dolph Lundgren

31. **Mandy** (I) (2018)

Not Rated | 121 min | Action, Horror, Thriller

★ **6.7** ☆ Rate this 81 Metascore

The enchanted lives of a couple in a secluded forest are brutally shattered by a nightmarish hippie cult and their demon-biker henchmen, propelling a man into a spiraling, surreal rampage of vengeance.

Director: Panos Cosmatos | Stars: Nicolas Cage, Andrea Riseborough, Linus Roache, Ned Dennehy

Votes: 22,867 | Gross: $1.21M

29. **The Grinch** (2018)

PG | 90 min | Animation, Comedy, Family

6.1 Rate this 50 Metascore

A grumpy Grinch plots to ruin Christmas for the village of Whoville.

Directors: Yarrow Cheney, Scott Mosier | Stars: Benedict Cumberbatch, Cameron Seely, Rashida Jones, Pharrell Williams

Votes: 1,484

30. **Creed II** (2018)

PG-13 | Drama, Sport | **Completed**

Under the tutelage of Rocky Balboa, newly crowned light heavyweight champion Adonis Creed faces off against Viktor Drago, the son of Ivan Drago.

Director: Steven Caple Jr. | Stars: Tessa Thompson, Sylvester Stallone, Michael B. Jordan, Dolph Lundgren

31. **Mandy** (I) (2018)

Not Rated | 121 min | Action, Horror, Thriller

6.7 Rate this 81 Metascore

The enchanted lives of a couple in a secluded forest are brutally shattered by a nightmarish hippie cult and their demon-biker henchmen, propelling a man into a spiraling, surreal rampage of vengeance.

Director: Panos Cosmatos | Stars: Nicolas Cage, Andrea Riseborough, Linus Roache, Ned Dennehy

Votes: 22,867 | Gross: $1.21M

32. **Apostle** (2018)

TV-MA | 130 min | Drama, Fantasy, Horror

.sort-num_votes-visible

Cl

29. **The Grinch** (2018)

PG | 90 min | Animation, Comedy, Family

6.1 Rate this 50 Metascore

A grumpy Grinch plots to ruin Christmas for the village of Whoville.

Directors: Yarrow Cheney, Scott Mosier | Stars: Benedict Cumberbatch, Cameron Seely, Rashida Jones, Pharrell Williams

Votes: 1,484

30. **Creed II** (2018)

PG-13 | Drama, Sport | **Completed**

Under the tutelage of Rocky Balboa, newly crowned light heavyweight champion Adonis Creed faces off against Viktor Drago, the son of Ivan Drago.

Director: Steven Caple Jr. | Stars: Tessa Thompson, Sylvester Stallone, Michael B. Jordan, Dolph Lundgren

31. **Mandy** (I) (2018)

Not Rated | 121 min | Action, Horror, Thriller

6.7 Rate this 81 Metascore

The enchanted lives of a couple in a secluded forest are brutally shattered by a nightmarish hippie cult and their demon-biker henchmen, propelling a man into a spiraling, surreal rampage of vengeance.

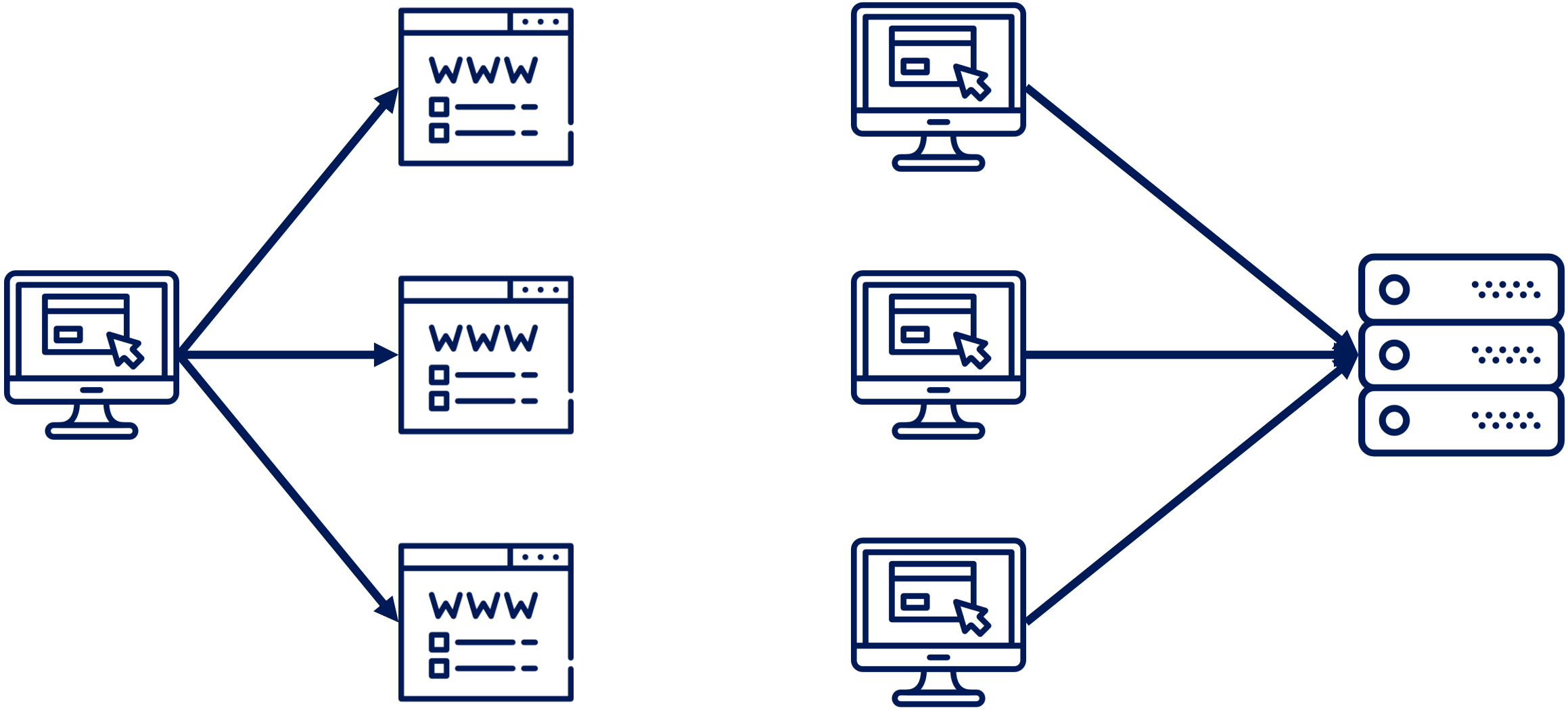
Director: Panos Cosmatos | Stars: Nicolas Cage, Andrea Riseborough, Linus Roache, Ned Dennehy

Votes: 22,867 | Gross: $1.21M

32. **Apostle** (2018)

TV-MA | 130 min | Drama, Fantasy, Horror

.lister-item-content


WWW
WWW
WWW

Challenge	Solution
Reproducibility	Remains unresolved: nature of the work (Partial solutions: Internet Archive + Local downloads)
Server	Make students aware of such issue + Give ample time
Data entry issues	Save these examples for more advanced courses.
Missing Data	Assign websites that are likely to have well structured nodes that are not missing.
Speed + Traffic	Assign a small subset of the website

WHAT?

WHY?

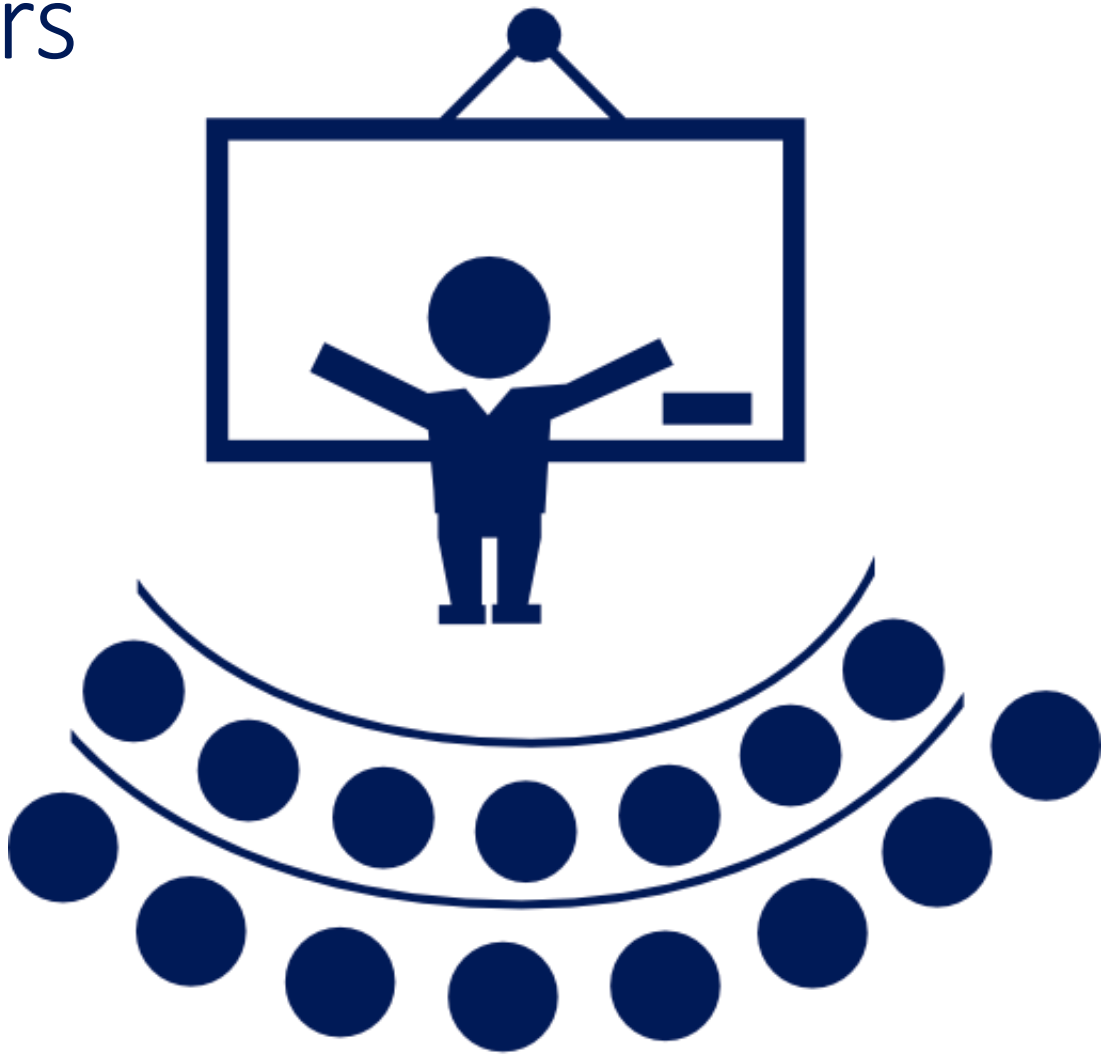
HOW?

WHEN?

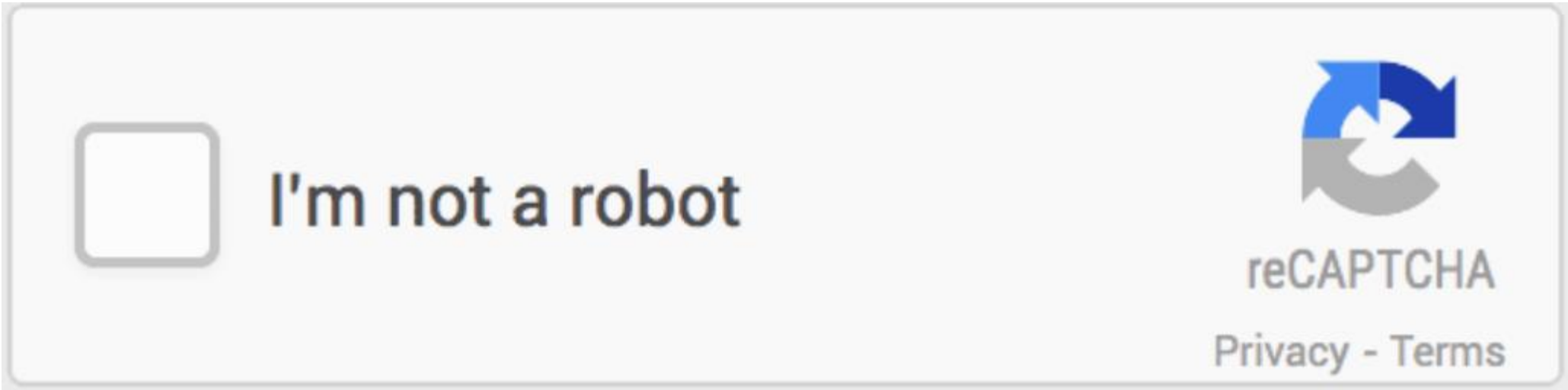
Variety of sources for data

craigslist

Not only for the learners



Ethics


I'm not a robot
reCAPTCHA
Privacy - Terms

```
> library(robotstxt)
> paths_allowed("http://www.imdb.com")
 www.imdb.com                          No encoding supplied: d
efaulting to UTF-8.


[1] TRUE
> paths_allowed("http://www.facebook.com")
 www.facebook.com

[1] FALSE
>
```

TERMS OF USE

“I feel like I can rule the world with web scraping”


meaningful
endless
COMPUTING
STATISTICS

Dr. Mine Cetinkaya-Rundel

eCOTS

Manuscript

QUESTIONS?

 dogucum@denison.edu

 @MineDogucu

 bit.ly/scrape_duke

References

GAISE College Report ASA Revision Committee, “Guidelines for Assessment and Instruction in Statistics Education College Report 2016,” http://www.amstat.org/education/gaise.

J. Hardin, R. Hoerl, Nicholas J. Horton, D. Nolan, B. Baumer, O. Hall-Holt, P. Murrell, R. Peng, P. Roback, D. Temple Lang & M. D. Ward (2015) Data Science in Statistics Curricula: Preparing Students to “Think with Data”, The American Statistician, 69:4, 343-353, DOI: 10.1080/00031305.2015.1077729

Nicholas J. Horton (2015) Challenges and Opportunities for Statistics and Statistical Education: Looking Back, Looking Forward, The American Statistician, 69:2, 138-145, DOI: 10.1080/00031305.2015.1032435

Deborah Nolan & Duncan Temple Lang (2010) Computing in the Statistics Curricula, The American Statistician, 64:2, 97- 107, DOI: 10.1198/tast.2010.09132