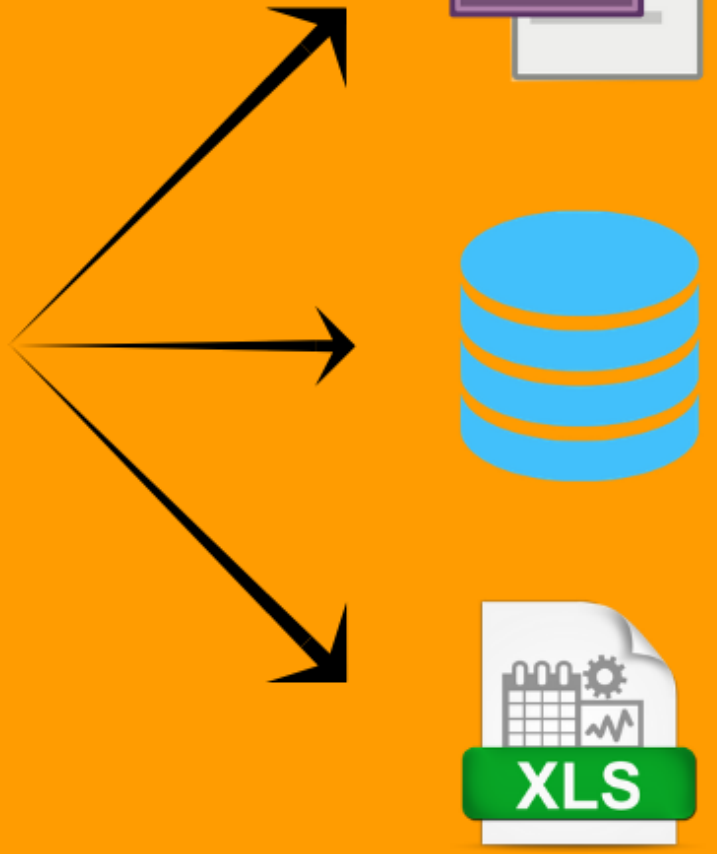
Rsquared Academy



Practical Introduction to Web Scraping in R
(Includes 4 Case Studies)

Connect With Us

- Website (https://www.rsquaredacademy.com/)
- Free Online R Courses (https://rsquared-academy.thinkific.com/)
- R Packages (https://pkgs.rsquaredacademy.com)
- Shiny Apps (https://apps.rsquaredacademy.com)
- Blog (https://blog.rsquaredacademy.com)

- GitHub (https://github.com/rsquaredacademy)
- YouTube (https://www.youtube.com/user/rsquaredin/)
- Twitter (https://twitter.com/rsquaredacademy)
- Facebook (https://www.facebook.com/rsquaredacademy/)
- Linkedin (https://in.linkedin.com/company/rsquared-academy)

Agenda

- what?
- why?
- how?
- use cases
- HTML basics
- case studies

intro

What?

Human Consumption

Data Analysis

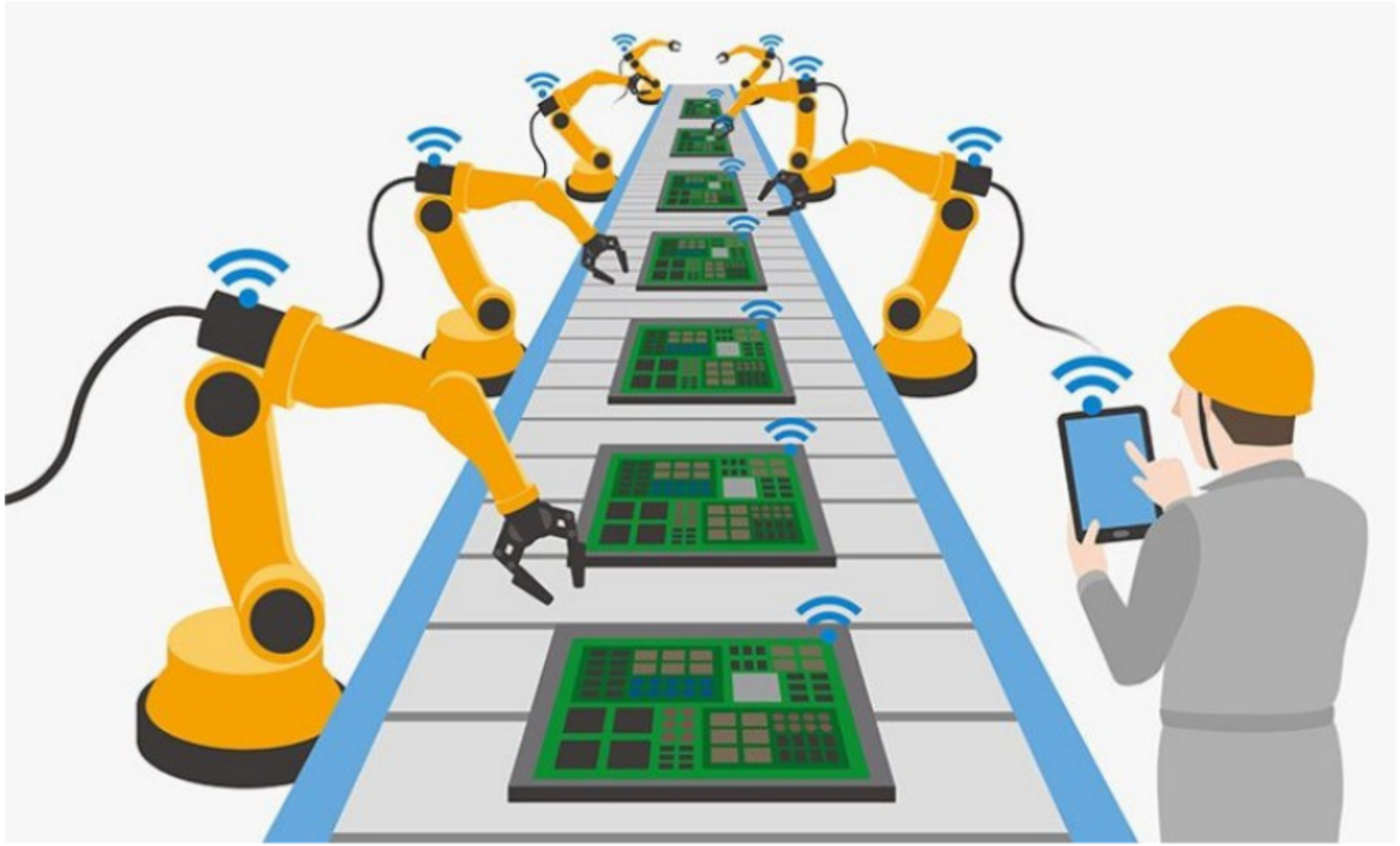
How?

Why?


Invest
Work
Live
Play
A World of
Opportunities Await...
Latest News

Why?

Why?



Use Cases

Contacts Scraping

Used Cars Listings

Real Estate Listing

Price Comparison

Reviews Scraping

Price Monitoring

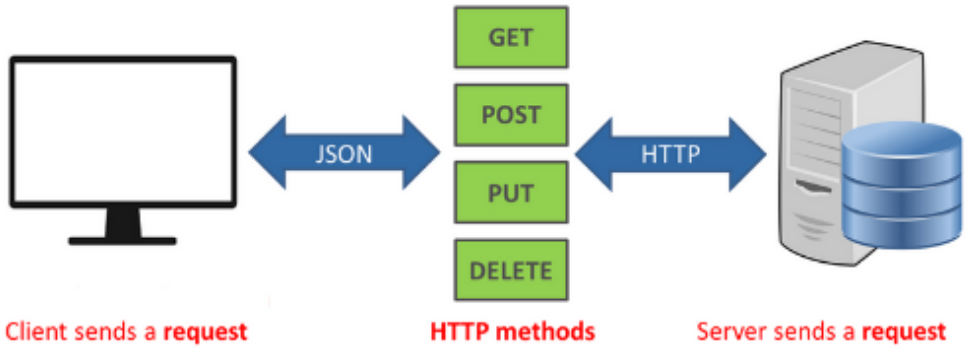
Case Studies

Things to keep in mind...

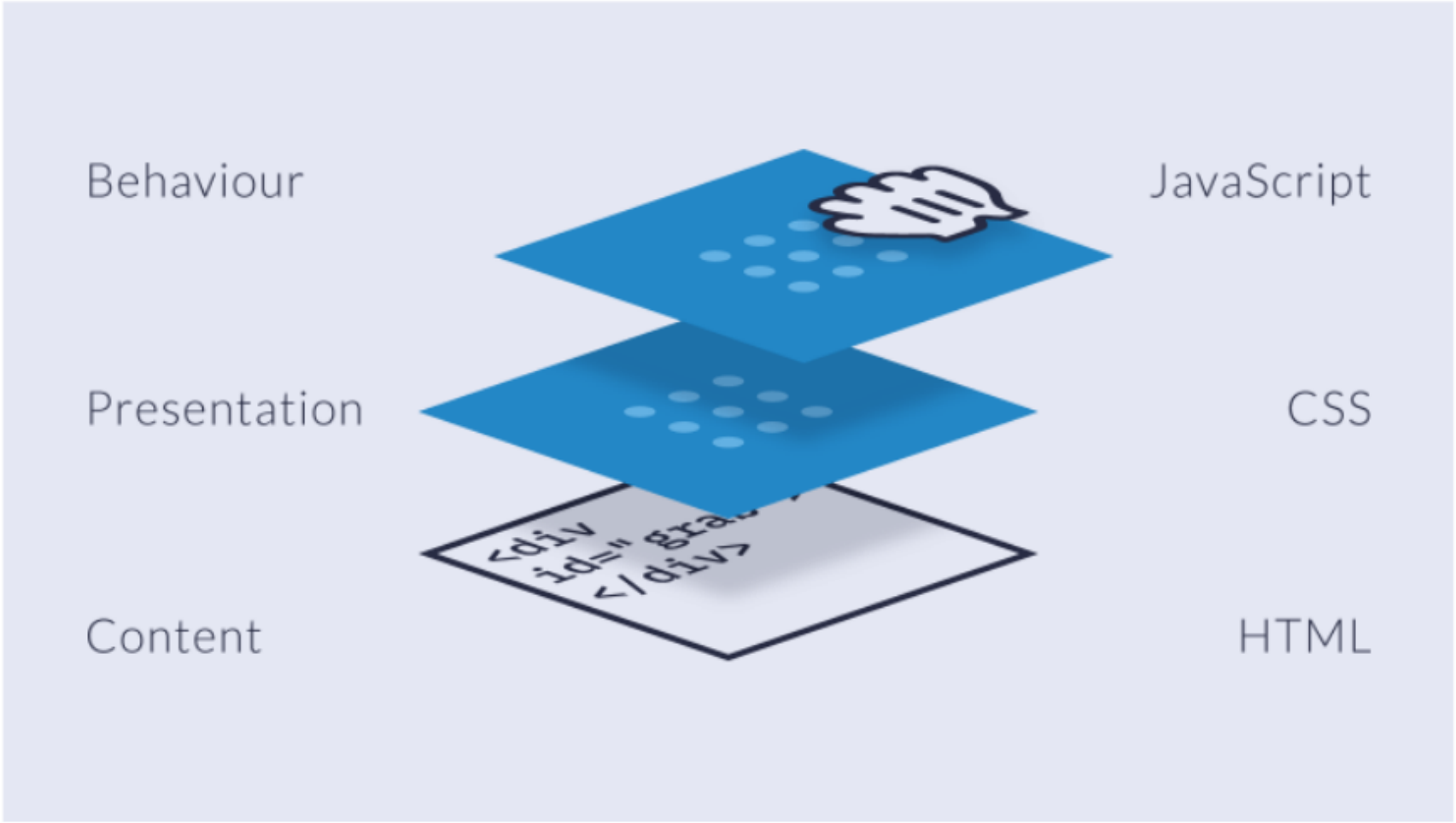
ROBOTS.TXT

```
User-Agent: *
Disallow: /music?
Disallow: /widgets/radio?
Disallow: /show_ads.php

Disallow: /affiliate/
Disallow: /affiliate_redirect.php
Disallow: /affiliate_sendto.php
Disallow: /affiliatelink.php
Disallow: /campaignlink.php
Disallow: /delivery.php

Disallow: /music/+noredirect/

Disallow: /harming/humans
Disallow: /ignoring/human/orders
Disallow: /harm/to/self

Allow: /
```

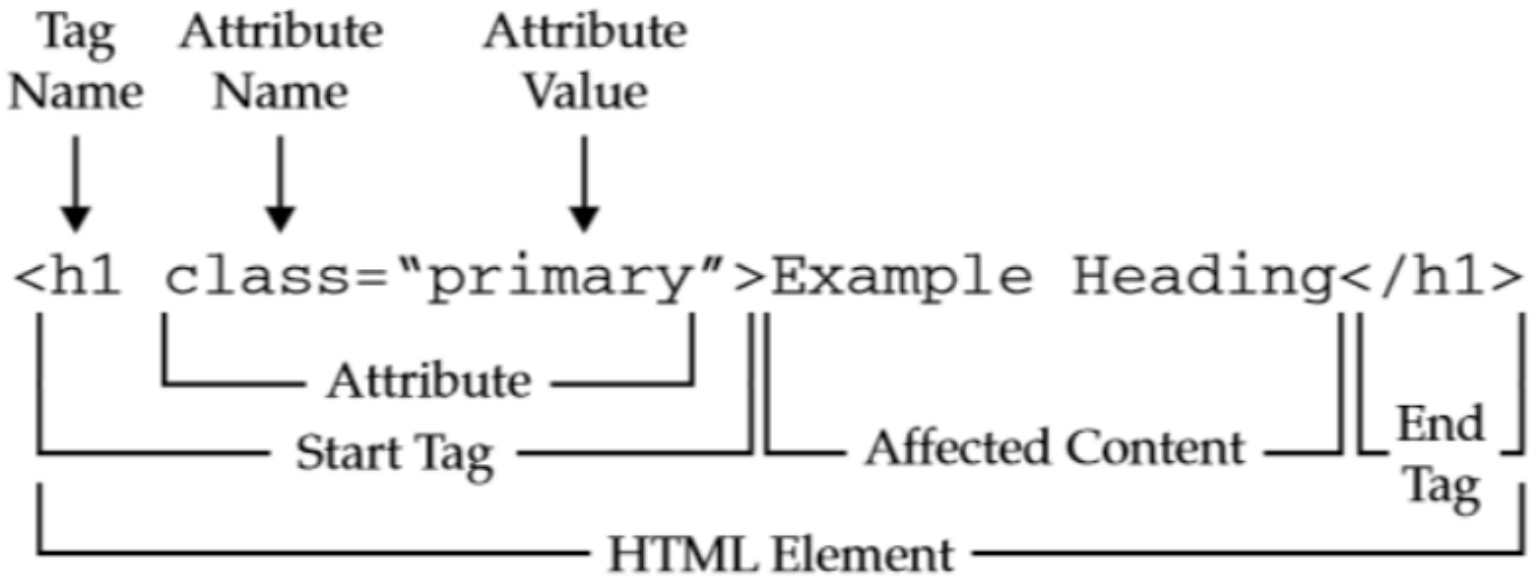
The code Has changed!



html basics


Behaviour
JavaScript
Presentation
CSS
Content
HTML

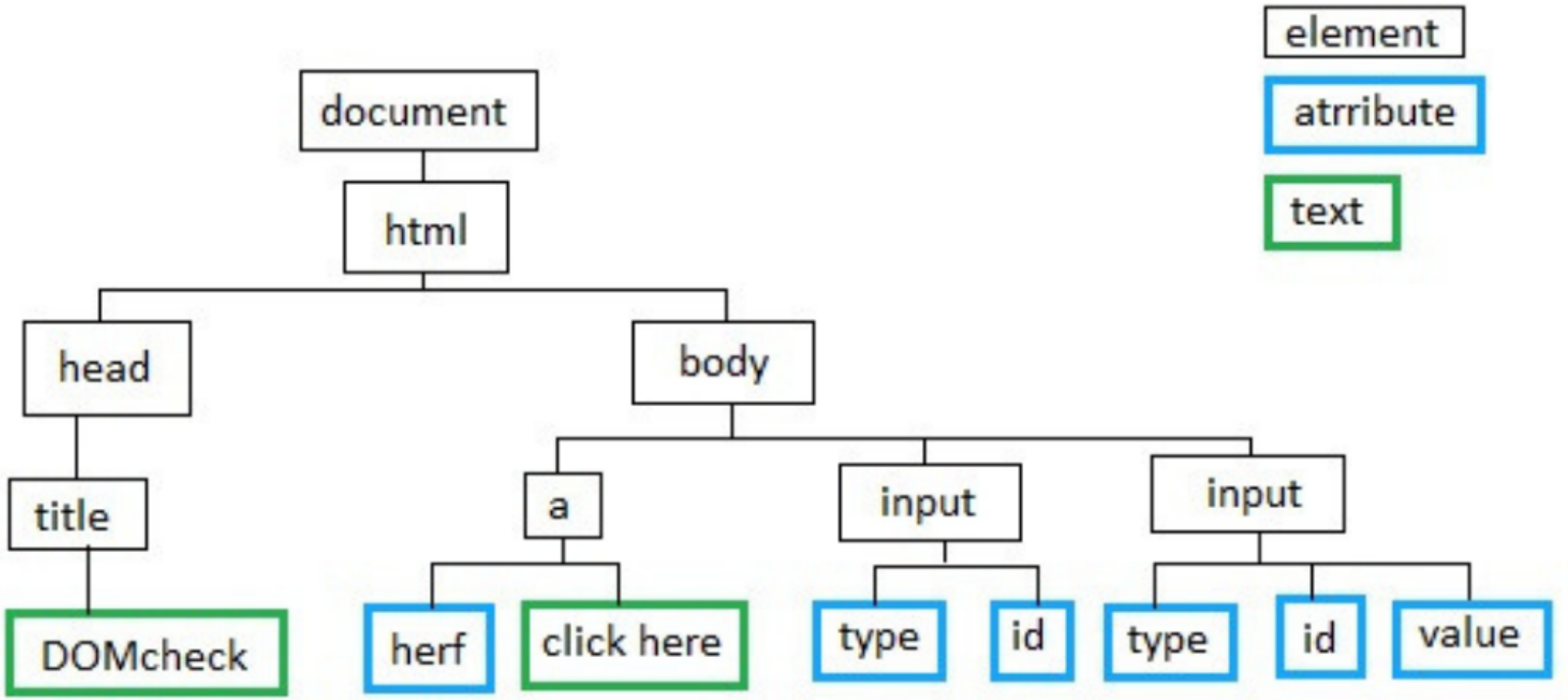
HTML Element



HTML Tags

Tag	Description
<html> ... </html>	Declares the Web page to be written in HTML
<head> ... </head>	Delimits the page's head
<title> ... </title>	Defines the title (not displayed on the page)
<body> ... </body>	Delimits the page's body
<h *n*> ... </h*n*>	Delimits a level *n* heading
...	Set ... in boldface
...	Set ... in italics
...	Center ... on the page horizontally
...	Brackets an unordered (bulleted) list
...	Brackets a numbered list
...	Brackets an item in an ordered or numbered list
 	Forces a line break here
<p>	Starts a paragraph
<hr>	Inserts a horizontal rule
<img src="...">	Displays an image here
 ... 	Defines a hyperlink

DOM


element
atrribute
text
document
html
head
body
title
a
input
input
DOMcheck
herf
click here
type
id
type
id
value

Attribute	Value	Description
class	*class_rule* or *style_rule*	The class of the element
id	*id_name*	A unique id for the element
style	*style_definition*	An inline style definition

case studies

Libraries

```
library(robotstxt)
library(rvest)
library(selectr)
library(xml2)
library(dplyr)
library(stringr)
library(forcats)
library(magrittr)
library(tidyr)
library(ggplot2)

library(lubridate)
library(tibble)
library(purrr)
```

IMDB Top 50


PULP FICTION
INCEPTION
CITY OF GOD
Casablanca
ONCE UPON A TIME IN THE WEST
THE PROFESSIONAL
CITY LIGHTS
CITIZEN KANE
DOUBLE INDEMNITY

```
paths_allowed(
  paths = c("https://www.imdb.com/search/title?groups=top_250&sort=user_
)
```

```
##
 www.imdb.com                          No encoding supplied: defaulting to U
```

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

Read Web Page

```
imdb <- read_html("https://www.imdb.com/search/title?groups=top_250&sort
imdb
```

```
## {xml_document}
## <html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; chars
## [2] <body id="styleguide-v2" class="fixed">\n\n            <img heigh
```

Title

```
<a href="/title/tt0111161/?ref_=adv_li_i"
   > <img alt="The Shawshank Redemption"
   ...</a>       </div>
<div class="lister-item-content">
<h3 class="lister-item-header">
   <span class="lister-item-index unbold text-primary">1.</span>
   <a href="/title/tt0111161/?ref_=adv_li_tt"
      >The Shawshank Redemption</a>
   <span class="lister-item-year text-muted unbold">(1994)</span>
</h3>
```

```
imdb %>%
  html_nodes(".lister-item-content h3 a") %>%
  html_text() -> movie_title

movie_title
```

```
##  [1] "The Shawshank Redemption"
##  [2] "The Godfather"

##  [3] "The Dark Knight"
##  [4] "The Godfather: Part II"
##  [5] "The Lord of the Rings: The Return of the King"
##  [6] "Pulp Fiction"
##  [7] "Schindler's List"
##  [8] "Il buono, il brutto, il cattivo"
##  [9] "12 Angry Men"
## [10] "Inception"
## [11] "Fight Club"
## [12] "The Lord of the Rings: The Fellowship of the Ring"
## [13] "Forrest Gump"
## [14] "The Lord of the Rings: The Two Towers"
## [15] "The Matrix"
## [16] "Goodfellas"
## [17] "Star Wars: Episode V - The Empire Strikes Back"
```

Year

```
<div class="lister-item-content">
<h3 class="lister-item-header">
    <span class="lister-item-index unbold text-primary">1.</span>
    <a href="/title/tt0111161/?ref_=adv_li_tt"
        >The Shawshank Redemption</a>
    <span class="lister-item-year text-muted unbold">(1994)</span>
</h3>
```

Year of Release

```
imdb %>%
  html_nodes(".lister-item-content h3 .lister-item-year") %>%
  html_text() %>%
  str_sub(start = 2, end = 5) %>%
  as.Date(format = "%Y") %>%
  year() -> movie_year

movie_year
```

```
##  [1] 1994 1972 2008 1974 2003 1994 1993 1966 1957 2010 1999 2001 1994
## [15] 1999 1990 1980 1975 1954 2014 2002 2001 1998 1999 1997 1995 1995
## [29] 1991 1977 1946 2018 2016 2018 2018 2014 2011 2006 2006 2002 2000
## [43] 1998 1994 1991 1988 1988 1985 1981 1979
```

Certificate

```
<a href="/title/tt0111161/?ref_=adv_li_i"
   > <img alt="The Shawshank Redemption"
   ...
</a>       </div>
<div class="lister-item-content">
<h3 class="lister-item-header">
   <span class="lister-item..."
      >The Shawshank Redemption</a>
   <span class="lister-item-year text-muted unbold">(1994)</span>
</h3>
<p class="text-muted ">
   <span class="certificate">R</span>
   <span class="ghost">|</span>
   <span class="runtime">142 min</span>
   <span class="ghost">|</span>
   <span class="genre">
   Drama          </span>
</p>
```

```
imdb %>%
  html_nodes(".lister-item-content p .certificate") %>%
  html_text() -> movie_certificate

movie_certificate
```

```
##  [1] "A"     "A"     "UA"    "PG-13" "A"     "A"     "UA"    "A"
##  [9] "PG-13" "PG-13" "PG-13" "A"     "A"     "PG"    "UA"    "R"

## [17] "PG"    "A"     "A"     "PG-13" "A"     "R"     "A"     "A"
## [25] "U"     "PG"    "UA"    "U"     "U"     "UA"    "A"     "UA"
## [33] "PG-13" "A"     "R"     "R"     "R"     "A"     "U"     "U"
## [41] "R"     "U"     "PG"    "R"
```

Runtime

```
<a href="/title/tt0111161/?ref_=adv_li_i"
   > <img alt="The Shawshank Redemption"
   ...
</a>       </div>
<div class="lister-item-content">
<h3 class="lister-item-header">
   <span class="lister-item..."
      >The Shawshank Redemption</a>
   <span class="lister-item-year text-muted unbold">(1994)</span>
</h3>
<p class="text-muted ">
   <span class="certificate">R</span>
   <span class="ghost">|</span>
   <span class="runtime">142 min</span>
   <span class="ghost">|</span>
   <span class="genre">
   Drama          </span>
</p>
```

Runtime

```
imdb %>%
  html_nodes(".lister-item-content p .runtime") %>%
  html_text() %>%
  str_split(" ") %>%
  map_chr(1) %>%
  as.numeric() -> movie_runtime

movie_runtime
```

```
##  [1] 142 175 152 202 201 154 195 161  96 148 139 178 142 179 136 146
## [18] 133 207 169 130 125 169 189 116 106 127 110 118 121 130 139 161
## [35] 149 106 112 130 151 150 113 155 119  88 137 155  89 116 115 147
```

Genre

```
<a href="/title/tt0111161/?ref_=adv_li_i"
   > <img alt="The Shawshank Redemption"
   ...
</a>       </div>
<div class="lister-item-content">
<h3 class="lister-item-header">
   <span class="lister-item..."
      >The Shawshank Redemption</a>
   <span class="lister-item-year text-muted unbold">(1994)</span>
</h3>
<p class="text-muted ">
   <span class="certificate">R</span>
   <span class="ghost">|</span>
   <span class="runtime">142 min</span>
   <span class="ghost">|</span>
   <span class="genre">
   Drama          </span>
</p>
```

```
imdb %>%
  html_nodes(".lister-item-content p .genre") %>%
  html_text() %>%
  str_trim() -> movie_genre

movie_genre
```

```
##  [1] "Drama"                          "Crime, Drama"
##  [3] "Action, Crime, Drama"           "Crime, Drama"
##  [5] "Adventure, Drama, Fantasy"      "Crime, Drama"
##  [7] "Biography, Drama, History"      "Western"
##  [9] "Drama"                          "Action, Adventure, Sci-Fi"
## [11] "Drama"                          "Adventure, Drama, Fantasy"
## [13] "Drama, Romance"                 "Adventure, Drama, Fantasy"
## [15] "Action, Sci-Fi"                 "Biography, Crime, Drama"
## [17] "Action, Adventure, Fantasy"     "Drama"
## [19] "Adventure, Drama"               "Adventure, Drama, Sci-Fi"
## [21] "Crime, Drama"                   "Animation, Adventure, Family"
## [23] "Drama, War"                     "Crime, Drama, Fantasy"
## [25] "Comedy, Drama, Romance"         "Crime, Mystery, Thriller"
## [27] "Crime, Drama, Mystery"          "Action, Crime, Drama"
## [29] "Crime, Drama, Thriller"         "Action, Adventure, Fantasy"
## [31] "Drama, Family, Fantasy"         "Crime, Thriller"
```

Rating

```
<div class="ratings-bar">
    <div class="inline-block ratings-imdb-rating" name="ir" data-value="9.3">
        <span class="global-sprite rating-star imdb-rating"></span>
        <strong>9.3</strong>
    </div>
```

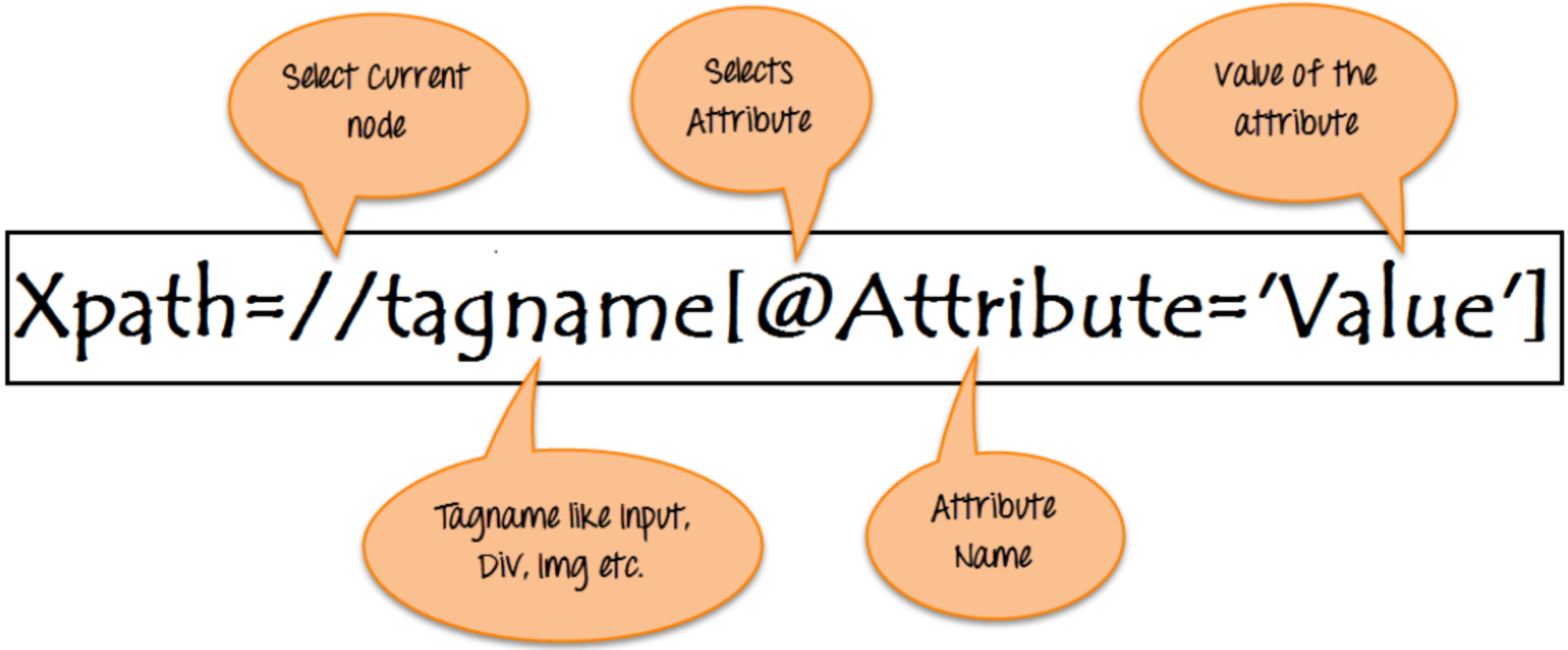
Rating

```
imdb %>%
  html_nodes(".ratings-bar .ratings-imdb-rating") %>%
  html_attr("data-value") %>%
  as.numeric() -> movie_rating

movie_rating
```

```
##  [1] 9.3 9.2 9.0 9.0 8.9 8.9 8.9 8.9 8.9 8.8 8.8 8.8 8.8 8.7 8.7 8.7
## [18] 8.7 8.7 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.6 8.5 8.5
## [35] 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5 8.5
```

Votes

```
<meta itemprop="ratingValue" content="9.3" />
 <meta itemprop="bestRating" content="10" />
 <meta itemprop="ratingCount" content="2070946" />
```


Select Current node
Selects Attribute
Value of the attribute
Xpath=//tagname[@Attribute='Value']
Tagname like Input, Div, Img etc.
Attribute Name

Votes

```
imdb %>%
  html_nodes(xpath = '//meta[@itemprop="ratingCount"]') %>%
  html_attr('content') %>%
  as.numeric() -> movie_votes

movie_votes
```

```
##  [1] 2072893 1422292 2038787  987020 1475650 1621033 1074273  615219
##  [9]  585562 1817393 1658750 1492209 1589127 1334563 1489071  895033
## [17] 1040130  822277  280024 1276946  637716  549410 1096231 1000909
## [25]  545280  897576 1271530  913352 1118817 1109777  352837   39132
## [33]  118413  174125  617621  605417  666327 1052901 1064050  633675
## [41] 1021511 1198326  941917  823238  897607  198398  192715  923178
## [49]  803033  542311
```

Revenue

```
<p class="sort-num_votes-visible">
    <span class="text-muted">Votes:</span>
    <span name="nv" data-value="2070946">2,070,946</span>
    <span class="ghost">|</span>              <span class="text-muted">Gross:</span>
    <span name="nv" data-value="28,341,469">$28.34M</span>
</p>
```

```
imdb %>%
  html_nodes(xpath = '//span[@name="nv"]') %>%
  html_text() %>%
  str_extract(pattern = "^\\$.*") %>%
  na.omit() %>%
  as.character() %>%
  append(values = NA, after = 30) %>%
  append(values = NA, after = 46) %>%
  str_sub(start = 2, end = nchar(.) - 1) %>%


  as.numeric() -> movie_revenue

movie_revenue
```

```
##  [1]  28.34 134.97 534.86  57.30 377.85 107.93  96.07   6.10   4.36 2
## [11]  37.03 315.54 330.25 342.55 171.48  46.84 290.48 112.00   0.27 1
## [21]   7.56  10.06 216.54 136.80  57.60  23.34 100.13  19.50 130.74 3
## [31]     NA   1.19  12.39 190.24 678.82  13.09  13.18  53.09 132.38
## [41]  25.54 187.71   6.72 312.90 204.84  11.99     NA 210.61 248.16
```

Putting it all together…

```
top_50 <- tibble(title = movie_title, release = movie_year,
    `runtime (mins)` = movie_runtime, genre = movie_genre, rating = movi
    votes = movie_votes, `revenue ($ millions)` = movie_revenue)

top_50
```

```
## # A tibble: 50 x 7
##    title    release `runtime (mins)` genre   rating  votes `revenue (
##    <chr>      <dbl>            <dbl> <chr>    <dbl>  <dbl>
##  1 The Sha~    1994              142 Drama      9.3 2.07e6
##  2 The God~    1972              175 Crime,~    9.2 1.42e6
##  3 The Dar~    2008              152 Action~    9   2.04e6
##  4 The God~    1974              202 Crime,~    9   9.87e5
##  5 The Lor~    2003              201 Advent~    8.9 1.48e6
##  6 Pulp Fi~    1994              154 Crime,~    8.9 1.62e6
##  7 Schindl~    1993              195 Biogra~    8.9 1.07e6
##  8 Il buon~    1966              161 Western    8.9 6.15e5
##  9 12 Angr~    1957               96 Drama      8.9 5.86e5
## 10 Incepti~    2010              148 Action~    8.8 1.82e6
## # ... with 40 more rows
```

RBI Governors

```
paths_allowed(
  paths = c("https://en.wikipedia.org/wiki/List_of_Governors_of_Reserve_
)
```

```
##
 en.wikipedia.org
```

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

Read Web Page

```
rbi_guv <- read_html("https://en.wikipedia.org/wiki/List_of_Governors_of
rbi_guv
```

```
## {xml_document}
## <html class="client-nojs" lang="en" dir="ltr">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; chars
## [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-
```

```
rbi_guv %>%
  html_nodes("table") %>%
  html_table() %>%
  extract2(2) -> profile

profile
```

```
##    No.          Officeholder Portrait         Term start           Term
## 1    1         Osborne Smith       NA       1 April 1935      30 June 1
## 2    2    James Braid Taylor       NA        1 July 1937  17 February 1
## 3    3        C. D. Deshmukh       NA  11 August 1943ii        30 May 1
## 4    4      Benegal Rama Rau       NA        1 July 1949   14 January 1
## 5    5     K. G. Ambegaonkar       NA   14 January 1957  28 February 1
## 6    6      H. V. R. Iyengar       NA      1 March 1957  28 February 1
## 7    7    P. C. Bhattacharya       NA      1 March 1962      30 June 1
## 8    8      Lakshmi Kant Jha       NA        1 July 1967        3 May 1
## 9    9         B. N. Adarkar       NA         4 May 1970       15 June 1
## 10  10 Sarukkai Jagannathan       NA       16 June 1970        19 May 1
## 11  11       N. C. Sen Gupta       NA        19 May 1975     19 August 1
## 12  12            K. R. Puri       NA    20 August 1975        2 May 1
## 13  13          M. Narasimham       NA         3 May 1977  30 November 1
## 14  14            I. G. Patel       NA   1 December 1977 15 September 1
## 15  15        Manmohan Singh       NA 16 September 1982   14 January 1
```

```
profile %>%
  separate(`Term in office`, into = c("term", "days")) %>%
  select(Officeholder, term) %>%
  arrange(desc(as.numeric(term)))
```

```
##            Officeholder term
## 1      Benegal Rama Rau 2754
## 2        C. D. Deshmukh 2150
## 3        R. N. Malhotra 2147
## 4           Bimal Jalan 2114
## 5    James Braid Taylor 2057
## 6    P. C. Bhattacharya 1947
## 7    Y. Venugopal Reddy 1826
## 8      H. V. R. Iyengar 1825
## 9           D. Subbarao 1825
## 10 Sarukkai Jagannathan 1798
## 11        C. Rangarajan 1795
## 12          I. G. Patel 1749
## 13       Raghuram Rajan 1096
## 14     Lakshmi Kant Jha 1037
## 15          Urjit Patel  947
## 16       Manmohan Singh  851
```

```
profile %>%
  count(Background)
```

```
## # A tibble: 9 x 2
##   Background                                        n
##   <chr>                                         <int>
## 1 ""                                                1
## 2 Banker                                            2
## 3 Career Reserve Bank of India officer              1
## 4 Economist                                         7
## 5 IAS officer                                       4
## 6 ICS officer                                       7
## 7 Indian Administrative Service (IAS) officer       1
## 8 Indian Audit and Accounts Service officer         1
## 9 Indian Civil Service (ICS) officer                1
```

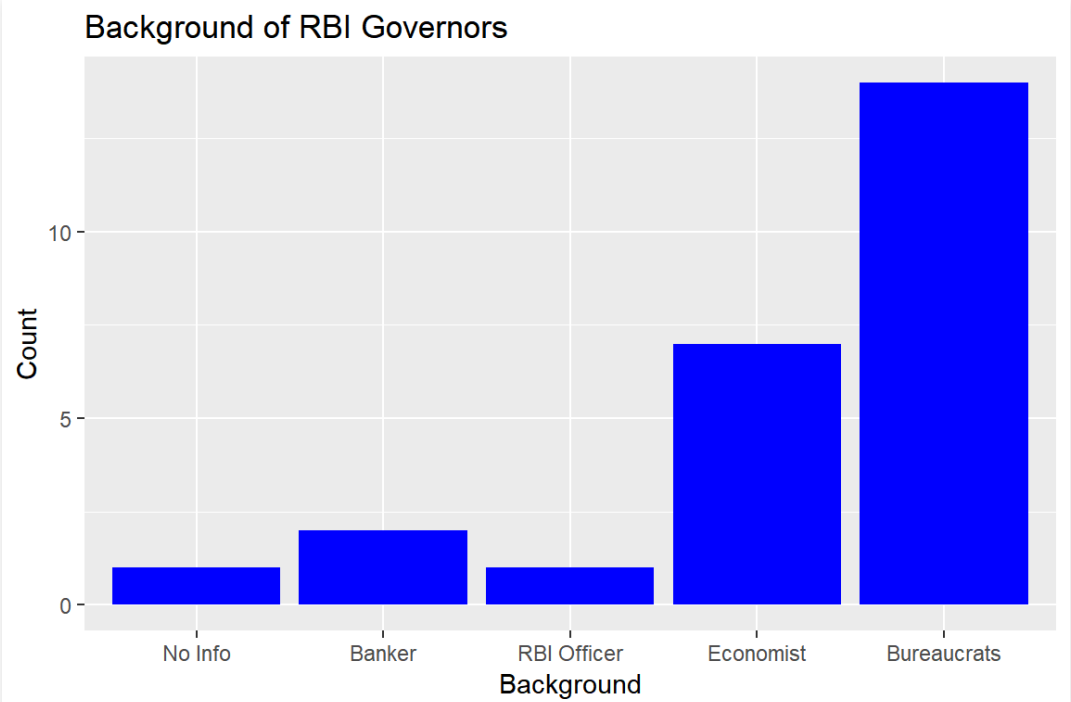
```
profile %>%
  pull(Background) %>%
  fct_collapse(
    Bureaucrats = c("IAS officer", "ICS officer",
    "Indian Administrative Service (IAS) officer",
    "Indian Audit and Accounts Service officer",
    "Indian Civil Service (ICS) officer"),
    `No Info` = c(""),
    `RBI Officer` = c("Career Reserve Bank of India officer")
  ) %>%

  fct_count() %>%
  rename(background = f, count = n) -> backgrounds
```

```
backgrounds
```

```
## # A tibble: 5 x 2
##   background  count
##   <fct>       <int>
## 1 No Info         1
## 2 Banker          2
## 3 RBI Officer     1
##
## 4 Economist       7
## 5 Bureaucrats    14
```

Backgrounds

```
backgrounds %>%
  ggplot() +
  geom_col(aes(background, count), fill = "blue") +
  xlab("Background") + ylab("Count") +
  ggtitle("Background of RBI Governors")
```



summary

Summary

- web scraping is the extraction of data from web sites
- best for static & well structured HTML pages
- review robots.txt file
- HTML code can change any time
- if API is available, please use it
- do not overwhelm websites with requests

Thank You

For more information please visit our website
www.rsquaredacademy.com