Web_Scrapping

https://www.youtube.com/watch?v=l37n_HDD1qs/newline

https://slides.rsquaredacademy.com/web-scraping/web-scraping.html#/section-11

what is web scrapping?

–> Web scrapping is the process or technique of extracting data from website and then tidying or reshapping it into format or structure suitable for data Analysis.

How do you do the web scrapping?

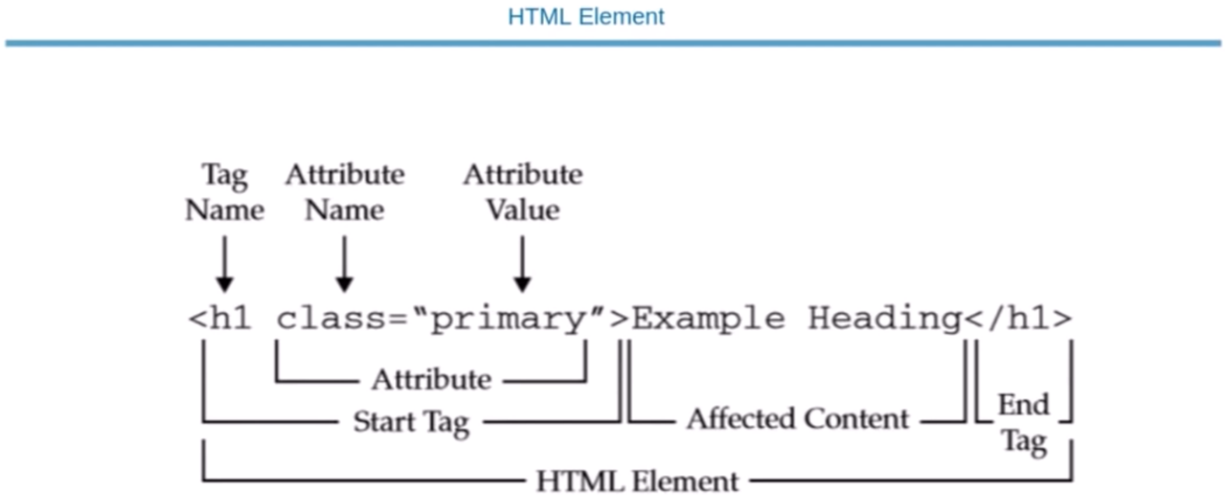
–> Step 1 : fetch the data as a xml document using xml2 package. –> Step 2 : Extract the content using rvest –> Step 3 : store using tibble

Why do you have to scrapp the web?

1. lot of web sites contains useful information we might want to use it for analysis.
2. you cann't copy/ save / download the contents of the website.
3. Web scrapping allows you to automate the data collection from website.

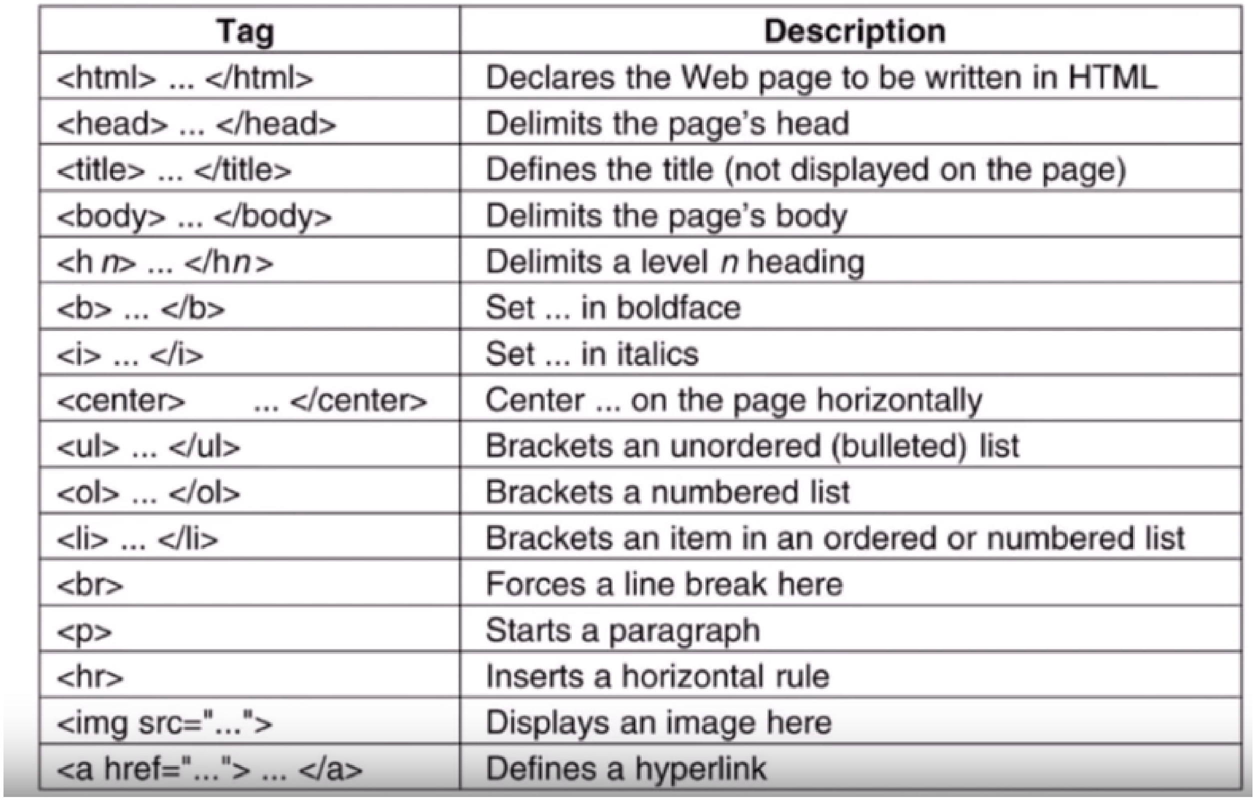
Use cases

1. Contact scrapping
2. Used cars listing
3. Real Estate Listing
4. Price comparison
5. Reviews Scraping
6. Price Monitoring



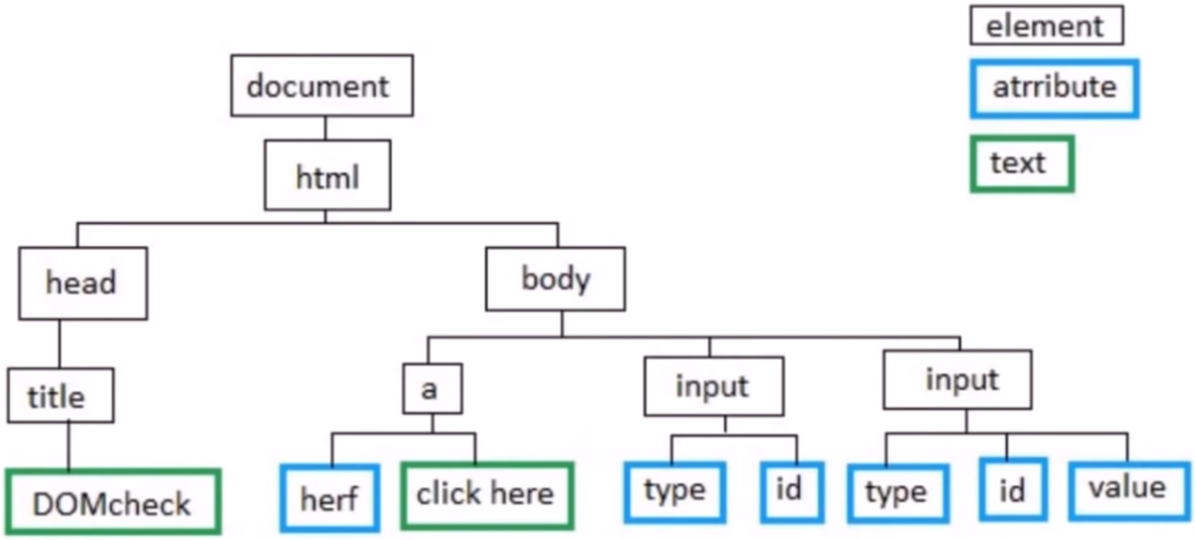
HTML Element

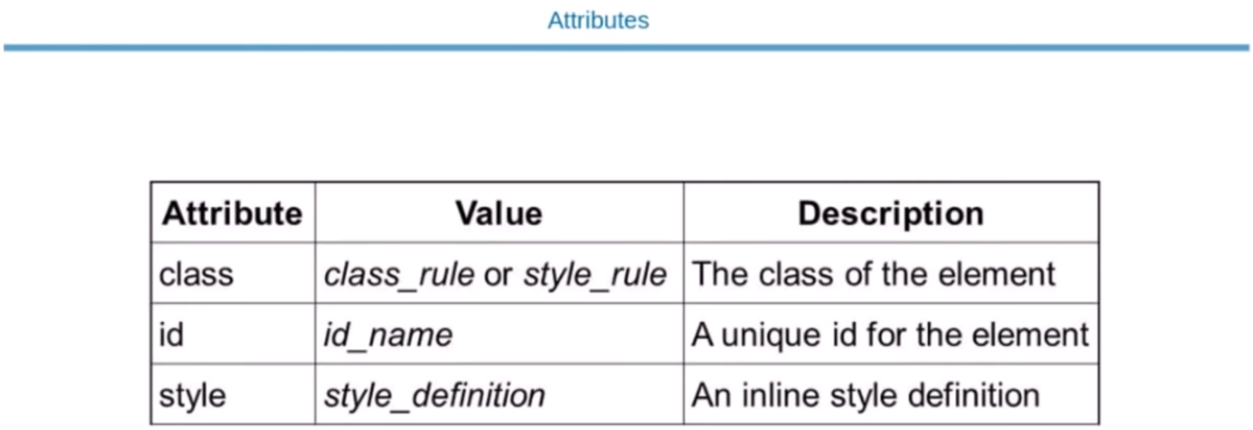
HTML Tags



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
<b> ... </b>	Set ... in boldface
<i> ... </i>	Set ... in italics
<center> ... </center>	Center ... on the page horizontally
<ul> ... </ul>	Brackets an unordered (bulleted) list
<ol> ... </ol>	Brackets a numbered list
<li> ... </li>	Brackets an item in an ordered or numbered list
<br>	Forces a line break here
<p>	Starts a paragraph
<hr>	Inserts a horizontal rule
<img src="...">	Displays an image here
<a href="..."> ... </a>	Defines a hyperlink

DOM





Attributes

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

```
library(robotstxt) # figure out whether or not we can scrape data
library(rvest)     # Extract the data
```

```
## Loading required package: xml2

## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
library(selectr)   # Query selecter
library(xml2)      # fetch data as xml documnet
library(dplyr)     # manipulate data
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)   # pattern matching
library(forcats)   # working with categorical variables
library(magrittr)  # pipe operator
library(tidyr)     # manipulate data
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(ggplot2)   # visualize data
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```r
library(lubridate) # working with dates
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(tibble)    # storing data
library(purrr)     # split a data frame into pieces, fit a model to each piece, compute the summary
```

```
##
## Attaching package: 'purrr'

## The following object is masked from 'package:magrittr':
##
##     set_names

## The following object is masked from 'package:rvest':
##
##     pluck
```

```r
library(backports)
library(future)
```

Case Study 1 : Best Selling Mobile Phones on Amazon website

STEP 1 : Check if we have permission to extract data using robotstxt package

```r
paths_allowed( paths = ("https://www.amazon.in/gp/bestsellers/electronics/1389432031"))
```

```
##
 www.amazon.in                      No encoding supplied: defaulting to UTF-8.

## [1] TRUE
```

True mean allow to scrape the data

False means not allow to scrape data.

STEP 2 : Read Web page

```r
top_phones <- read_html("https://www.amazon.in/gp/bestsellers/electronics/1389432031")
top_phones
```

```
## {xml_document}
## <html class="a-no-js" data-19ax5a9jf="dingo">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset= ...
## [2] <body class="a-aui_149818-c a-aui_152852-c a-aui_157141-c a-aui_1586 ...
```

Data scrapping from imdb

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

```
##
 www.imdb.com                      No encoding supplied: defaulting to UTF-8.
```

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

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

```
## {xml_document}
## <html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset= ...
## [2] <body id="styleguide-v2" class="fixed">\n\n            <img height=" ...
```

Title

```
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] "The Good, the Bad and the Ugly"
##  [9] "12 Angry Men"
## [10] "Avengers: Endgame"
## [11] "Inception"
## [12] "Fight Club"
## [13] "The Lord of the Rings: The Fellowship of the Ring"
## [14] "Forrest Gump"
## [15] "The Lord of the Rings: The Two Towers"
## [16] "The Matrix"
## [17] "Goodfellas"
## [18] "Star Wars: Episode V - The Empire Strikes Back"
## [19] "One Flew Over the Cuckoo's Nest"
## [20] "Seven Samurai"
## [21] "Interstellar"
## [22] "City of God"
## [23] "Spirited Away"
## [24] "Saving Private Ryan"
## [25] "The Green Mile"
## [26] "Life Is Beautiful"
## [27] "The Usual Suspects"
## [28] "Se7en"
## [29] "Léon: The Professional"
## [30] "The Silence of the Lambs"
## [31] "Star Wars: Episode IV - A New Hope"
## [32] "It's a Wonderful Life"
## [33] "Andhadhun"
## [34] "Dangal"
## [35] "Spider-Man: Into the Spider-Verse"
## [36] "Avengers: Infinity War"
## [37] "Whiplash"
## [38] "The Intouchables"
## [39] "The Prestige"
## [40] "The Departed"
## [41] "The Pianist"
## [42] "Memento"
## [43] "Gladiator"
## [44] "American History X"
## [45] "The Lion King"
## [46] "Terminator 2: Judgment Day"
## [47] "Cinema Paradiso"
## [48] "Grave of the Fireflies"
## [49] "Back to the Future"
## [50] "Raiders of the Lost Ark"
```

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 2019 2010 1999 2001 1994
## [15] 2002 1999 1990 1980 1975 1954 2014 2002 2001 1998 1999 1997 1995 1995
## [29] 1994 1991 1977 1946 2018 2016 2018 2018 2014 2011 2006 2006 2002 2000
## [43] 2000 1998 1994 1991 1988 1988 1985 1981
```

Certificate

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

movie_certificate
```

```
##  [1] "R"         "R"         "PG-13"     "R"         "PG-13"
##  [6] "R"         "R"         "R"         "Not Rated" "PG-13"
## [11] "PG-13"     "R"         "PG-13"     "PG-13"     "PG"
## [16] "R"         "R"         "PG"        "R"         "Not Rated"
## [21] "PG-13"     "R"         "PG"        "R"         "R"        
## [26] "PG-13"     "R"         "R"         "R"         "R"        
## [31] "PG"        "PG"        "Not Rated" "Not Rated" "PG"       
## [36] "PG-13"     "R"         "R"         "PG-13"     "R"        
## [41] "R"         "R"         "R"         "R"         "G"        
## [46] "R"         "R"         "Not Rated" "PG"        "PG"
```

Run Time

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

#movie_runtime
```

Genre

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

#movie_genre
```

Rating

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

#movie_rating
```

Votes

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

#movie_votes
```

Revenue

```
#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
```

Putting it all togather…

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

#top_50
```

Case study 2 : RBI Governors

STEP 1 : robotstxt

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

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

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

STEP 2 : Read Web Page

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

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

STEP 3 : List of Governors

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

profile
```

```
##    No.         Officeholder Portrait        Term start          Term end
## 1    1        Osborne Smith       NA      1 April 1935      30 June 1937
## 2    2   James Braid Taylor       NA       1 July 1937  17 February 1943
## 3    3       C. D. Deshmukh       NA  11 August 1943ii       30 May 1949
## 4    4     Benegal Rama Rau       NA       1 July 1949   14 January 1957
## 5    5    K. G. Ambegaonkar       NA   14 January 1957  28 February 1957
## 6    6     H. V. R. Iyengar       NA      1 March 1957  28 February 1962
## 7    7   P. C. Bhattacharya       NA      1 March 1962      30 June 1967
## 8    8     Lakshmi Kant Jha       NA       1 July 1967        3 May 1970
## 9    9        B. N. Adarkar       NA        4 May 1970      15 June 1970
## 10  10 Sarukkai Jagannathan       NA      16 June 1970       19 May 1975
## 11  11      N. C. Sen Gupta       NA       19 May 1975    19 August 1975
## 12  12           K. R. Puri       NA    20 August 1975        2 May 1977
## 13  13        M. Narasimham       NA        3 May 1977  30 November 1977
## 14  14          I. G. Patel       NA   1 December 1977 15 September 1982
## 15  15       Manmohan Singh       NA 16 September 1982   14 January 1985
## 16  16         Amitav Ghosh       NA   15 January 1985   4 February 1985
## 17  17       R. N. Malhotra       NA   4 February 1985  22 December 1990
## 18  18    S. Venkitaramanan       NA  22 December 1990  21 December 1992
## 19  19        C. Rangarajan       NA  22 December 1992  21 November 1997
## 20  20          Bimal Jalan       NA  22 November 1997  6 September 2003
## 21  21   Y. Venugopal Reddy       NA  6 September 2003  5 September 2008
## 22  22          D. Subbarao       NA  5 September 2008  4 September 2013
## 23  23       Raghuram Rajan       NA  4 September 2013  4 September 2016
## 24  24          Urjit Patel       NA  4 September 2016  11 December 2018
## 25  25      Shaktikanta Das       NA  12 December 2018         Incumbent
##    Term in office                                  Background
## 1        821 days                                      Banker
## 2       2057 days          Indian Civil Service (ICS) officer
## 3       2150 days                                 ICS officer
## 4       2754 days                                 ICS officer
## 5         45 days                                 ICS officer
## 6       1825 days                                 ICS officer
## 7       1947 days   Indian Audit and Accounts Service officer
## 8       1037 days                                 ICS officer
## 9         42 days                                   Economist
## 10      1798 days                                 ICS officer
## 11        92 days                                 ICS officer
## 12       621 days
## 13       211 days        Career Reserve Bank of India officer
## 14      1749 days                                   Economist
## 15       851 days                                   Economist
## 16        20 days                                      Banker
## 17      2147 days Indian Administrative Service (IAS) officer
## 18       730 days                                 IAS officer
## 19      1795 days                                   Economist
## 20      2114 days                                   Economist
## 21      1826 days                                 IAS officer
## 22      1825 days                                 IAS officer
## 23      1096 days                                   Economist
## 24       972 days                                   Economist
## 25       143 days                                 IAS officer
##
## 1
## 2                                                                                             Deputy
## 3                                                                                          Deputy Gov
## 4                                                          Ambassador of India to the United States\n
## 5
## 6
## 7                                                                                          Chairman o
## 8
## 9
## 10
## 11
## 12                                                                                                  C
## 13
## 14 Director of the London School of Economics\n\nDeputy Administrator of the United Nations Developme
## 15                                                                         Secretary in the Ministry
## 16                                                                                    Deputy Governor
## 17                                                                                        Finance Sec
## 18
## 19
## 20                                                                       Finance Secretary\n\nBanking
## 21                                                             Executive Director at the Internationa
## 22                                                                           Finance Secretary\n\nMem
## 23
## 24
## 25                                             Member of the Fifteenth Finance Commission\nSherpa of
##    Reference(s)
## 1           [1]
## 2           [2]
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25    [3][4][5]
```

STEP 4 : Sort

```
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  972
## 16       Manmohan Singh  851
## 17        Osborne Smith  821
## 18    S. Venkitaramanan  730
## 19           K. R. Puri  621
## 20        M. Narasimham  211
## 21      Shaktikanta Das  143
## 22      N. C. Sen Gupta   92
## 23    K. G. Ambegaonkar   45
## 24        B. N. Adarkar   42
## 25         Amitav Ghosh   20
```

STEP 5 : Backgrounds

```
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
```

STEP 6 : Backgrounds

```
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
```

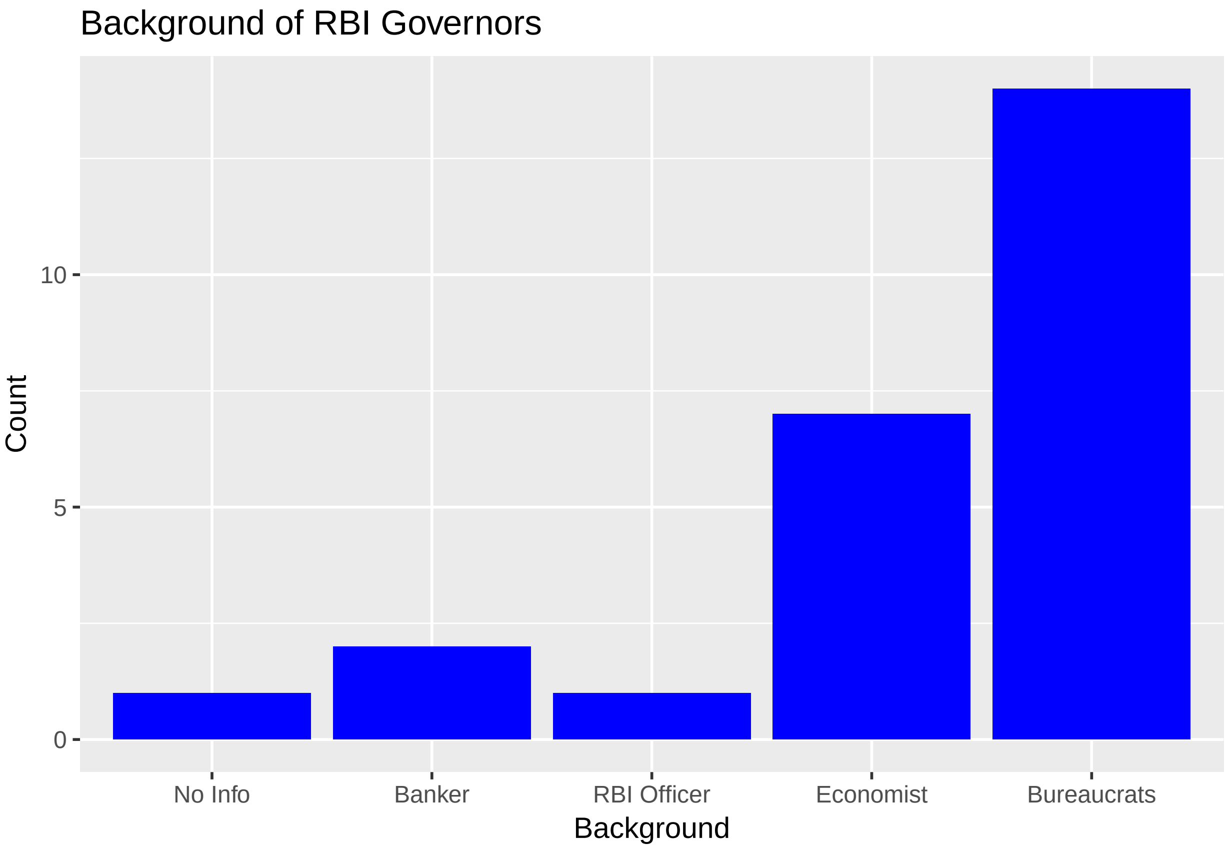
STEP 7 : Backgrounds

```
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
```

STEP 8 : Backgrounds

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


Background of RBI Governors
Count
No Info
Banker
RBI Officer
Economist
Bureaucrats
Background