What is R?

- Open-source programming language and software environment
- Focuses on statistical computing and graphics
- Large and active user community with extensive libraries
- Free to use and readily available on various platforms

Why R for Web Scraping?

- Rich ecosystem of web scraping libraries: `rvest`, `RSelenium`
- Data manipulation and analysis strength of R
- Integration with popular data visualization libraries like `ggplot2`
- Open-source and free to use

R Capabilities for Web Scraping

- R goes beyond just harvesting data
- Key functionalities for web scraping tasks:
  - Sending HTTP requests to websites (obtaining the content)
  - Parsing HTML structure (understanding the website's layout)

- Extracting specific data using selectors (targeting desired information)
- Cleaning and transforming extracted data (preparing for analysis)
- Navigating dynamic website using `RSelenium` package

Direct data import from the web

In [1]:
```
df <- read.csv("http://s.anilz.net/wb_energy")
head(df)

dx <- read.csv("https://data.ny.gov/api/views/d6yy-54nr/rows.csv")
head(dx)
```

A data.frame: 6 × 11

	year	country	ccode	ele_rural	ele_total	ele_urban	en_int	ren_ele	ren_con	tot_ele	tfec
	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1990	Afghanistan	AFG	NA	0.01000	52.03698	1.884113	764	6312.3920	1128	39639.420
2	1990	Albania	ALB	100.000000	100.00000	100.00000	7.912243	2848	20429.1800	3296	80057.645
3	1990	Algeria	DZA	96.392315	98.27138	100.00000	3.500935	135	811.7773	16104	458040.442
4	1990	American Samoa	ASM	NA	NA	NA	NA	0	0.0000	100	306.000
5	1990	Andorra	AND	100.000000	100.00000	100.00000	NA	120	952.1450	120	6670.695
6	1990	Angola	AGO	7.518615	11.39781	22.68237	4.605300	725	135443.7000	841	187451.703

A data.frame: 6 × 3

	Draw.Date	Winning.Numbers	Multiplier
	<chr>	<chr>	<int>
1	09/26/2020	11 21 27 36 62 24	3
2	09/30/2020	14 18 36 49 67 18	2
3	10/03/2020	18 31 36 43 47 20	2
4	10/07/2020	06 24 30 53 56 19	2
5	10/10/2020	05 18 23 40 50 18	3
6	10/14/2020	21 37 52 53 58 05	2

Using rvest package for static website scraping

Example 1. share price scraping

In [ ]:
```
#loading necessary packages
library(rvest) #see https://rvest.tidyverse.org/articles/harvesting-the-web.html for details
library(dplyr)

#loading webpage content
webpage <- read_html("https://www.sharesansar.com/today-share-price")

#extracting table from the webpage
tables <- html_table(webpage)

#checking the number of tables available in the webpage
length(tables)
```

In [3]:
```
#storing the table in a dataframe
df1 <- tables[[1]]
head(df1)
```

A tibble: 6 × 21

S.No	Symbol	Conf.	Open	High	Low	Close	VWAP	Vol	Prev. Close	⋯	Trans.	Diff	Range	Diff %	Range %	VW
<int>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	⋯	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<db
1	ACLBSL	39.42	983.10	998.00	970.00	986.00	979.31	5,844.00	1,000.00	⋯	97	-14.0	28.00	-1.40	2.89	0.
2	ADBL	49.67	267.90	267.90	261.50	261.50	263.74	19,809.00	268.00	⋯	152	-6.5	6.40	-2.43	2.45	-0.
3	ADBLD83	61.57	1,061.00	1,101.50	1,061.00	1,101.50	1,070.29	350.00	1,080.00	⋯	8	21.5	40.50	1.99	3.82	2.
4	AHL	43.71	505.00	508.90	492.00	500.00	499.00	16,996.00	497.00	⋯	139	3.0	16.90	0.60	3.43	0.
5	AHPC	44.72	161.00	161.00	156.00	156.00	157.69	118,322.00	157.90	⋯	375	-1.9	5.00	-1.20	3.21	-1.
6	AKJCL	53.18	216.00	220.10	213.60	213.60	216.51	42,372.00	216.90	⋯	188	-3.3	6.50	-1.52	3.04	-1.

In [4]:
```
#filtering upper and lower circuit stock
#See https://github.com/tempgita/training/blob/master/archieved/R%20Training%20(old)/Day%203-Session%203/dplyr%20-%20A%20Gramm
filtered_df1 <- df1 %>% filter(`Diff %` > 9 | `Diff %`< -9) %>% arrange(`Diff %`)
filtered_df1
```

A tibble: 3 × 21

S.No	Symbol	Conf.	Open	High	Low	Close	VWAP	Vol	Prev. Close	⋯	Trans.	Diff	Range	Diff %	Range %	VWAP %
<int>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	⋯	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>
102	KBSH	51.72	1,695.40	1,695.40	1,557.00	1,557.00	1,562.78	12,879.00	1,730.00	⋯	203	-173	138.40	-10	8.89	-0.37
140	MKLB	51.89	1,836.00	1,836.00	1,620.00	1,620.00	1,627.02	1,405.00	1,800.00	⋯	36	-180	216.00	-10	13.33	-0.43
242	SAMAJ	65.63	2,151.00	2,409.00	2,151.00	2,409.00	2,301.38	10,537.00	2,190.00	⋯	160	219	258.00	10	11.99	4.47

In [5]:
```
write.csv(filtered_df1, file = "example1.csv", row.names = FALSE)
```

Example 2. Forex from NRB

In [6]:
```
#loading webpage content
webpage <- read_html("https://www.nrb.org.np")

#extracting table from the webpage
tables <- html_table(webpage)

#checking the number of tables available in the webpage
length(tables)
```

2

In [7]:
```
df1 <- tables[[1]]
df2 <- tables[[2]]

df1
df2
```

A tibble: 6 × 3

Currency	Buy	Sell
<chr>	<dbl>	<dbl>
USD	133.40	134.00
EUR	142.95	143.59
GBP	169.03	169.79
AUD	88.08	88.47
SGD	98.61	99.06
JPY	8.46	8.50

A tibble: 8 × 3

Last Updated	13/06/2024	12/06/2024
<chr>	<chr>	<chr>
Total Deposits (in NPR Billion)	6,242	6,235
Commercial Banks Total Deposits (in NPR Billion)	5,525	5,519
Other BFIs Total Deposits (in NPR Billion)	717	716
Total Lending (in NPR Billion)	5,133	5,131
Commercial Banks Total Lending (in NPR Billion)	4,542	4,541
Other BFIs Total Lending (in NPR Billion)	591	591
CD Ratio (in %)	80.08	80.14
Interbank Interest Rate LCY - Weighted Avg. (in %)	2.95	2.97

In [8]:
```
#keeping USD and JPY only
filtered_df1 <- df1 %>% filter(Currency=='USD' | Currency =='JPY')
filtered_df1
```

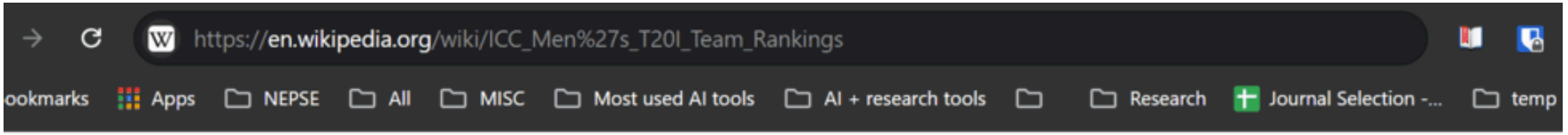
```
write.csv(filtered_df1, file = "example2.csv", row.names = FALSE)
```

A tibble: 2 × 3

Currency	Buy	Sell
<chr>	<dbl>	<dbl>
USD	133.40	134.0
JPY	8.46	8.5

Practice 1. Web-scrape the Historical ranking table from https://en.wikipedia.org/wiki/ICC_Men%27s_T20I_Team_Rankings and save it as practice1.csv

---



Contents hide

Historical rankings [edit]

This table lists the teams that have historically held the highest rating since the T20I rankings was introduced.[*citation needed*] In April 2018, the ICC decided to grant full T20I status to all its members. As a result, ratings of leading teams since 2018 have been considerably higher, and cannot be directly compared to those before that date.

Country	Start	End	Duration	Cumulative	Highest Rating
England	24 October 2011[4]	7 August 2012[5]	289 days	289 days	140
South Africa	8 August 2012	11 September 2012	35 days	35 days	137
England	12 September 2012	21 September 2012	10 days	299 days	130
South Africa	22 September 2012	28 September 2012	7 days	42 days	134
	29 September				