DIGITAL TALENT SCHOLARSHIP

Thematic Academy

Tema Pelatihan: Big Data for Social Science

Pertemuan #5: Data Scraping & Crawling With R

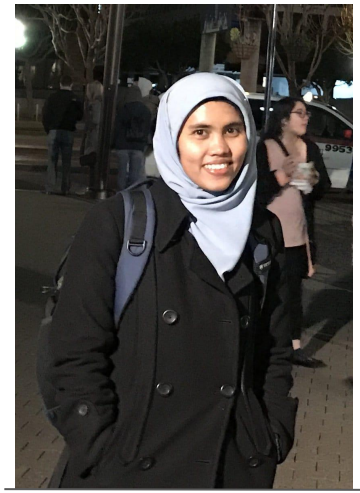
Pemateri: Erika Siregar

KOMINFO

#JADIJAGOANDIGITAL

Badan Penelitian dan Pengembangan Sumber Daya Manusia

Hello, my name is Erika



Erika Siregar

Education

- **Master in Computer Science from Old Dominion University, US**
- **Bachelor of Applied Science from STIS**

What I am doing now:

- **BPS**
- **R-Ladies Jakarta : Cofounder (IG: @rladiesjkt, youtube: R-Ladies Jakarta, GitHub, Whatsapp Group)**
- **Jakarta Machine Learning: Head of Program**

Connect with Me:
Email : erika.mukhlisina@gmail.com
GitHub : https://github.com/erikaris
Twitter : @erikaris
Instagram : @erikaris15
Linkedin : https://www.linkedin.com/in/erika-siregar/

Self Check

1. Ada yang belum pernah pakai R sebelum join acara ini?
2. On scale of 1 - 10, how confident are you in using R?
3. Experience with web scraping?
4. Web programming?
5. HTML nodes & Inspect Element?
6. Sudah install R dan RStudio?
7. Sudah download script untuk hari ini?
8. Sudah take a look at the today's script?
9. Ready for today's materials?

Learning Objective

In this course you will:

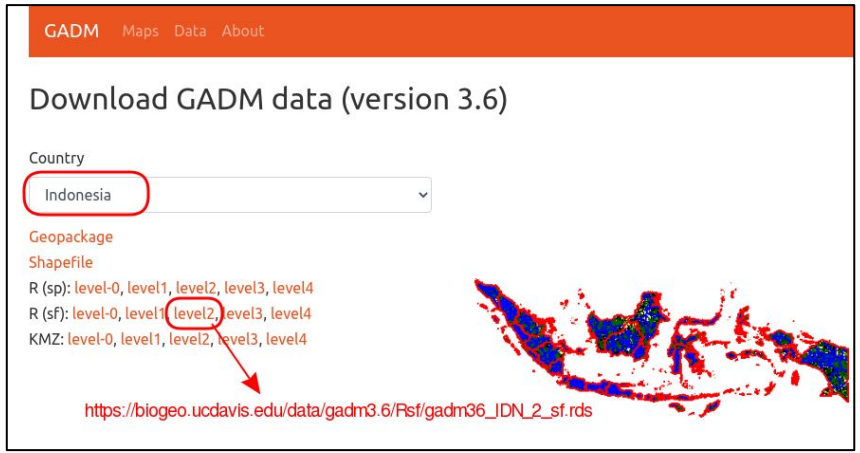
A. Learn basic understanding of web scraping
B. Learn case 1: scrape holiday dates from a website
C. Learn case 2: get insight of **stay-at-home behaviour during pandemic** through web scraping.
    a. scrape and download administrative and movement data from the internet
    b. visualize it as graphs.

Case 1: Scraping Holiday Dates

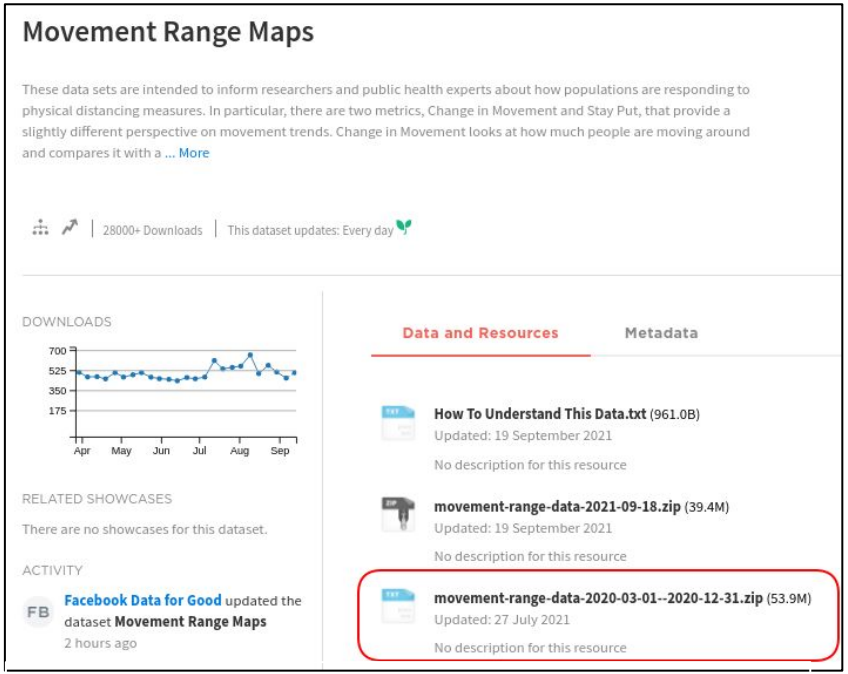
https://www.liburnasional.com/kalender-lengkap-2019/


```
# A tibble: 21 × 2
   date       event
   <date>     <chr>
 1 2019-01-01 Tahun Baru Masehi
 2 2019-02-05 Tahun Baru Imlek
 3 2019-03-07 Hari Raya Nyepi
 4 2019-04-03 Isra Miraj
 5 2019-04-17 Pemilu
 6 2019-04-19 Jumat Agung
 7 2019-05-01 Hari Buruh
 8 2019-05-19 Hari Raya Waisak
 9 2019-05-30 Kenaikan Isa Almasih
10 2019-06-01 Hari Lahir Pancasila
11 2019-06-03 Cuti Bersama
12 2019-06-04 Cuti Bersama
13 2019-06-05 Idul Fitri
14 2019-06-06 Idul Fitri
15 2019-06-07 Cuti Bersama
16 2019-08-11 Idul Adha
17 2019-08-17 Hari Kemerdekaan
18 2019-09-01 Tahun Baru Hijriyah
19 2019-11-09 Maulid Nabi
20 2019-12-24 Cuti Bersama
21 2019-12-25 Hari Natal
```

Case 2: stay-at-home behaviour during pandemic

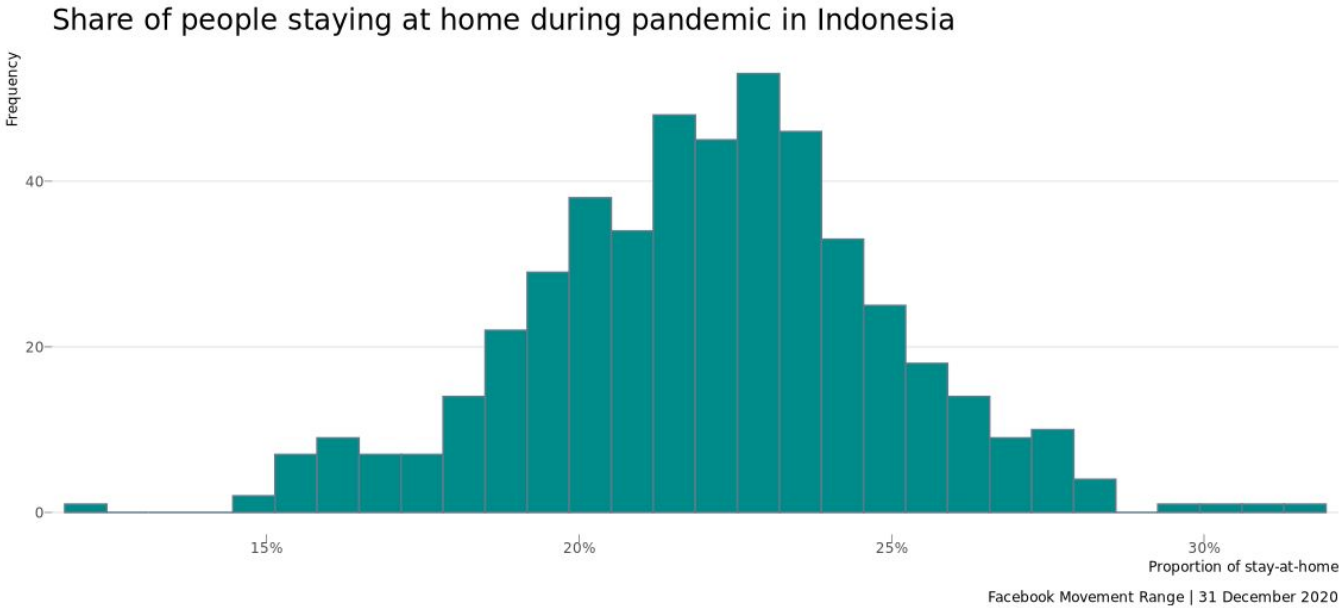


GADM data (spatial)

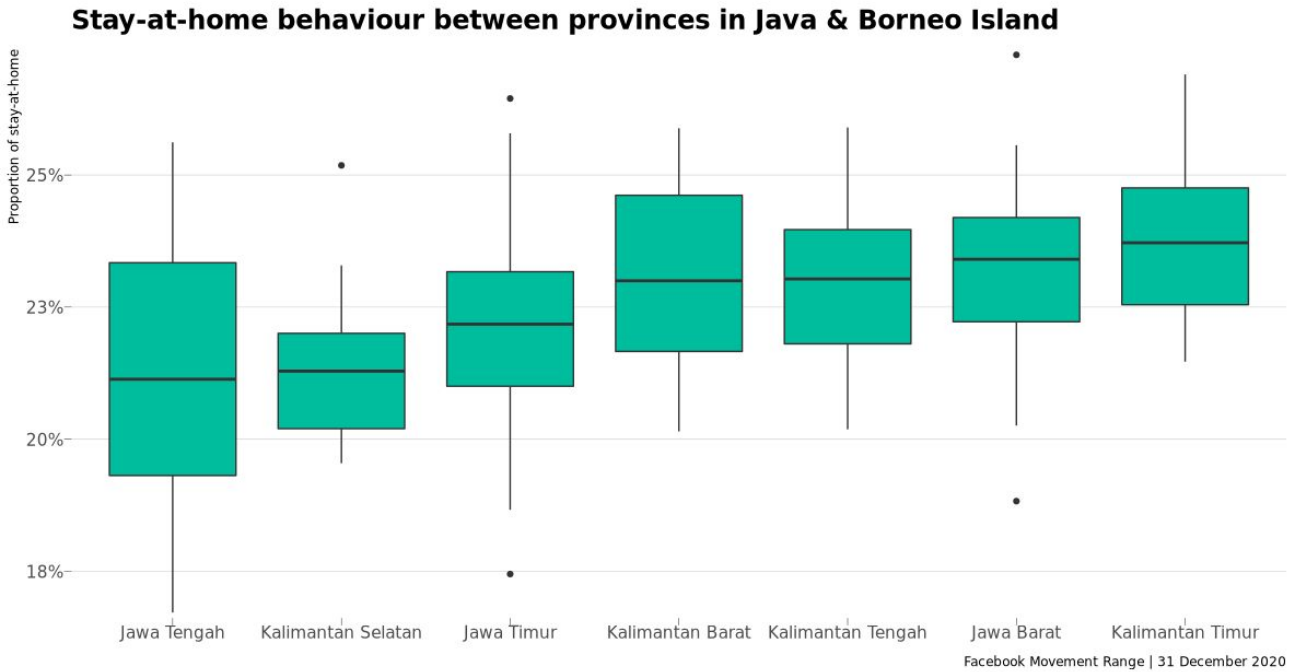


facebook movement data

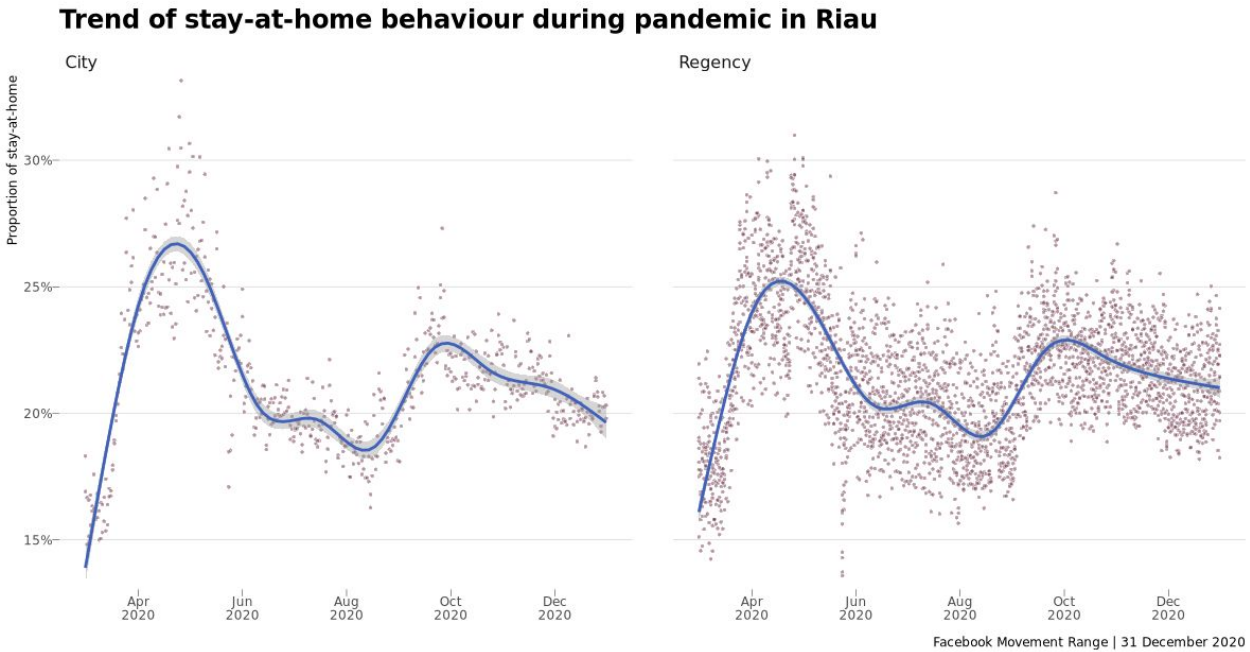
Visualisasi



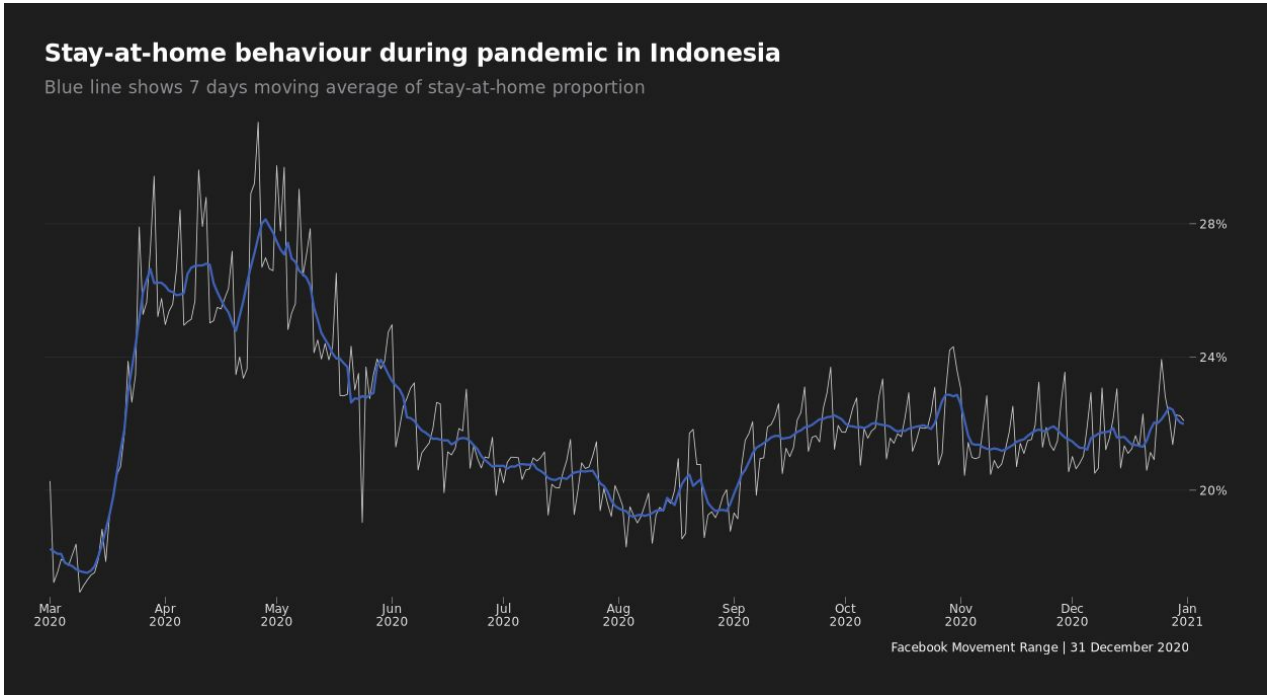
Visualisasi



Visualisasi



Visualisasi



Web Scraping 101

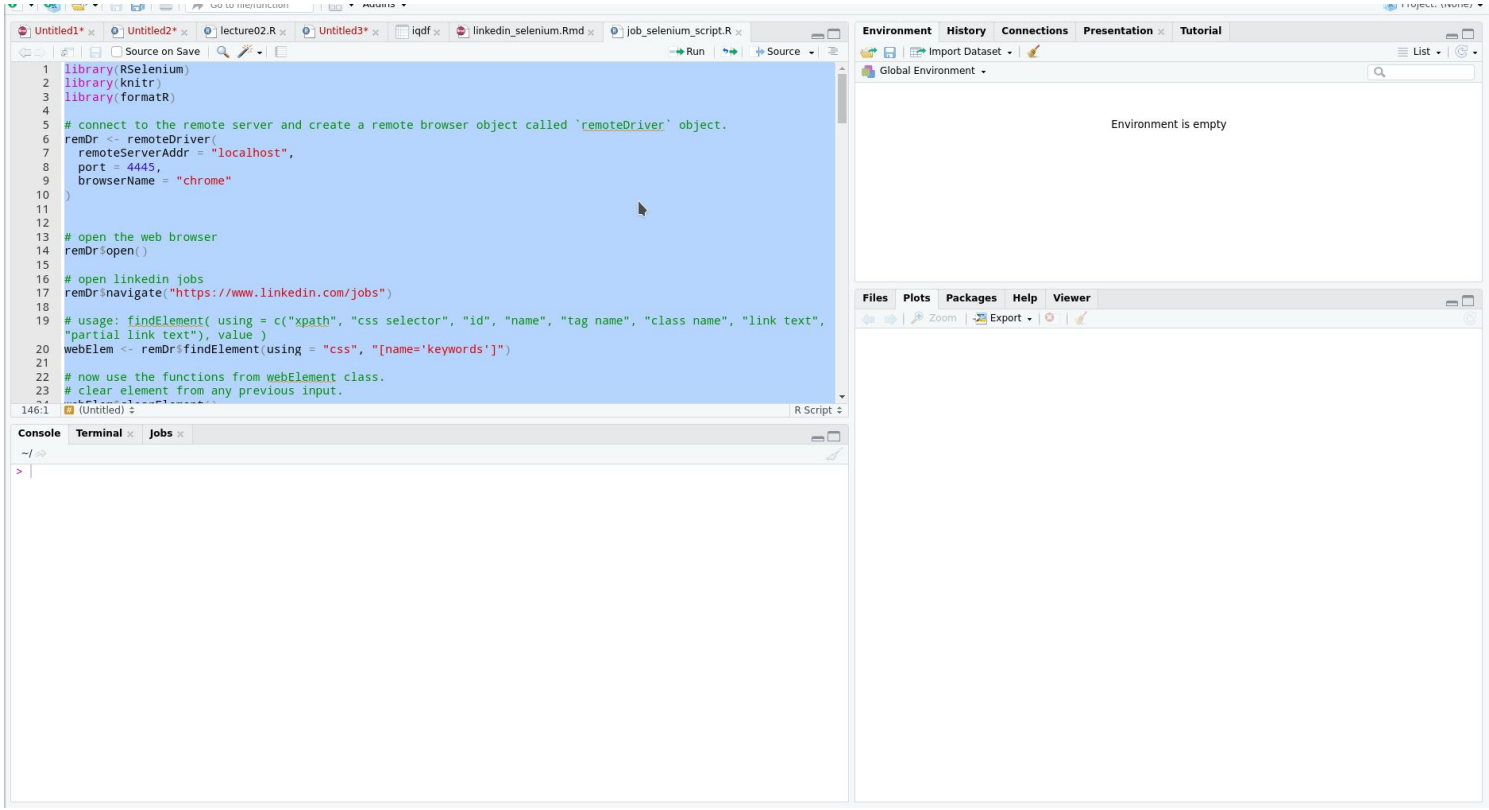
Web Scraping 101

1. Extracting data/information from a website and converting it into a format of your choice (HTML, JSON, CSV, etc.)
2. It's basically copy and paste certain part of the page, but instead of doing it yourself, you ask the computer to do it for you.
3. When do you scrape? → When there is too much to do manually.
4. What can be scraped?
    a. Basically any page
    b. Could be tricky when the page is complex.
    c. Complex == rich of javascript.

3 Types of Web Scraping

1. Simple static page
    a. e.g. Wikipedia, liburnasional.com
    b. rvest (R), scrapy (Python)
2. Dynamic, javascript-heavy page
    a. e.g: E-Commerce, Linkedin
    b. selenium
3. Social Media (Twitter)
    a. official API → rtweet, tweepy
    b. others, e.g: twint (python)

Brief Demo of Selenium



Please Do Keep in Mind

1. The web code and design can change anytime.
2. Be mindful in maintaining the number of requests
3. If there is an API, use it

Rvest

1. check the documentation: https://rvest.tidyverse.org/
2. dependent to other libraries such as: xml2, etc
3. important functions:
   a. **read_html()** --> convert a website into an XML object.
   b. **html_nodes()** --> extract the relevant nodes from the XML object
   c. **html_text()** --> extract the tagged data from the wanted nodes.
   d. **html_attrs()** --> return a list of the attributes.

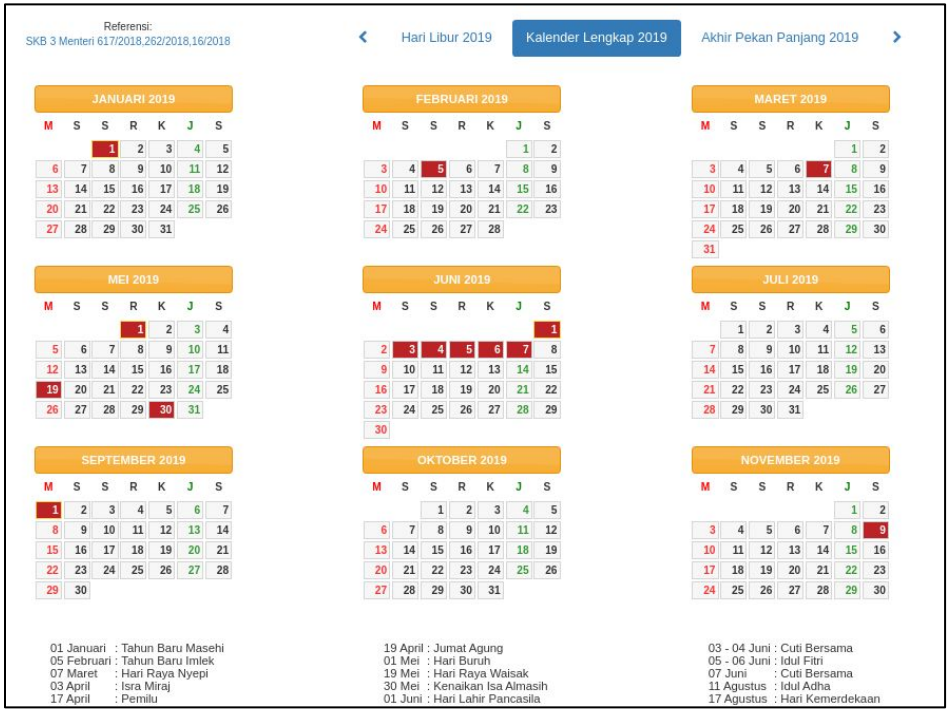
Steps in Web Scraping

1. Open the webpage
2. Explore it → **inspect** → **THE MOST IMPORTANT STEP**
3. scrape the page using scripts → obtain the data
4. preprocessed the data

The Key to Web Scraping is Selecting the Right Element (Node)

1. Know about css selector → most commonly used class (.) and id (#)
   a. **Spend time to learn about it!**
2. How:
   a. chrome extension: SelectorGadget
   b. inspect element

Dissecting HTML Page



what we see

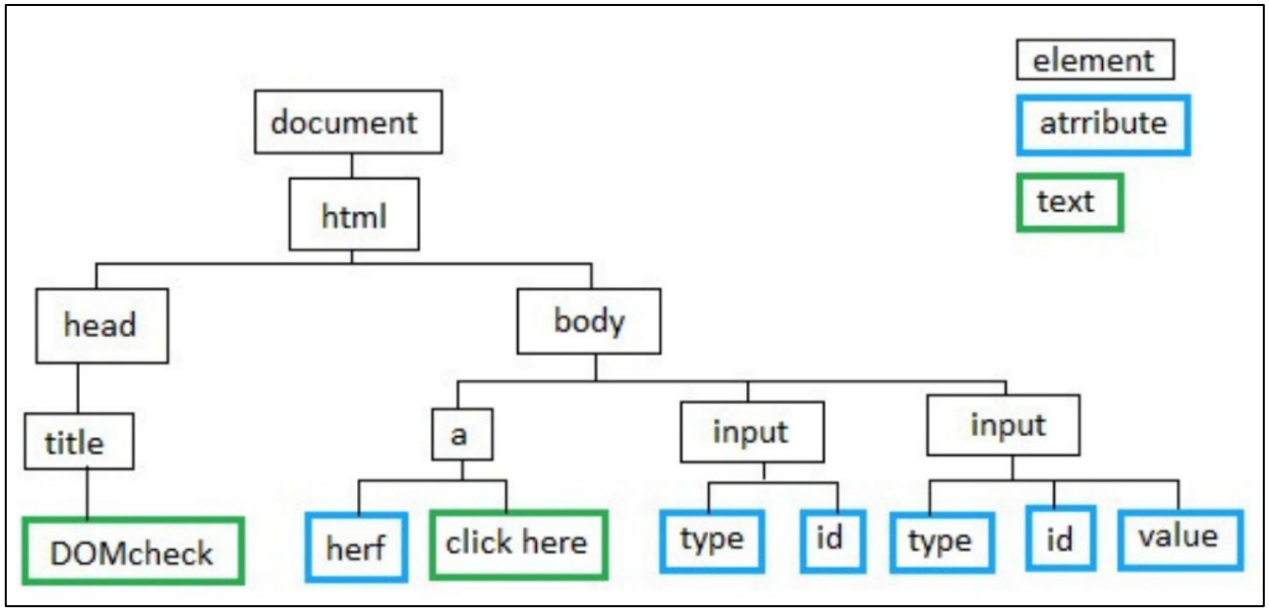
```
<html>
    <head>
            <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
        <meta name="viewport" content="width=device-width, initial-scale=1.0">
        <meta property="og:image" content="http://www.liburnasional.com/img/share_liburnasional.png" />
        <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.4/css/bootstrap.min.css">
        <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.4/css/bootstrap-theme.min.css">
        <link rel="stylesheet" href="/lib/font-awesome-4.6.1/css/font-awesome.min.css">
        <link rel="stylesheet" href="/lib/jquery-ui-1.11.4/jquery-ui.min.css">
        <link href='https://fonts.googleapis.com/css?family=Yanone+Kaffeesatz:700' rel='stylesheet' type='text/css'>
        <link rel="stylesheet" href="/lib/magnific-popup-1.1.0/dist/magnific-popup.css">
        <link rel="stylesheet" href="/css/libnas.css">
        <link rel="shortcut icon" href="/favicon.ico" />
            <title>Kalender Lengkap Indonesia Tahun 2019</title>
    <meta name="description" content="Kalender lengkap tahun 2019  beserta hari libur nasional dan cuti bersama berdasarkan kepu
            (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
            (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
            m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
            })(window,document,'script','//www.google-analytics.com/analytics.js','ga');

            ga('create', 'UA-63859452-5', 'auto');
            ga('require', 'linkid', 'linkid.js');
            ga('send', 'pageview');

            /**
            * Function that tracks a click on an outbound link in Google Analytics.
            * This function takes a valid URL string as an argument, and uses that URL string
            * as the event label.
            */
            var trackOutboundLink = function(url) {
               ga('send', 'event', 'outbound', 'click', url, {'hitCallback':
                 function () {
                 document.location = url;
                 }
               });
            }
        </script>

        <!-- PopAds.net Popunder Code for www.liburnasional.com | 2017-09-25,958255,0,0 -->
        <script type="text/javascript" data-cfasync="false">
        /*<![CDATA[/* */
        /* Privet darkv. Each domain is 2h fox dead */
         (function(){ var u=window;u["_\u0070\u006f\u0070"]=[["site\u0049d",958255],["minB\u0069\u0064",0],["\u0070opu\u006e\u00
        /*]]>/* */
        </script>
    </head>
    <body>
            <div id="feedback-ribbon">
            <a href="/kontak/">Hubungi Kami  <i class="fa fa-envelope-o"></i></a>
        </div>          <div class="container-fluid">
                <div class="row row-logo">
            <!-- For Logo -->
            <div class="col-md-4 col-xs-12 text-center"><a id="libnas-linkid-logo" href="/">Libur Nasional</a></div>
            <!-- For Banner -->
            <div class="col-md-8 col-xs-12 text-center">
                        <script type="text/javascript">
            agoda_ad_client = "1722771_44280";
            agoda_ad_width = 468;
            agoda_ad_height = 90;
            width = document.documentElement.clientWidth;
        </script>
        <script type="text/javascript" src="//banner.agoda.com/js/show_ads.js"></script>
```

what actually behinds it

Dissecting HTML Page

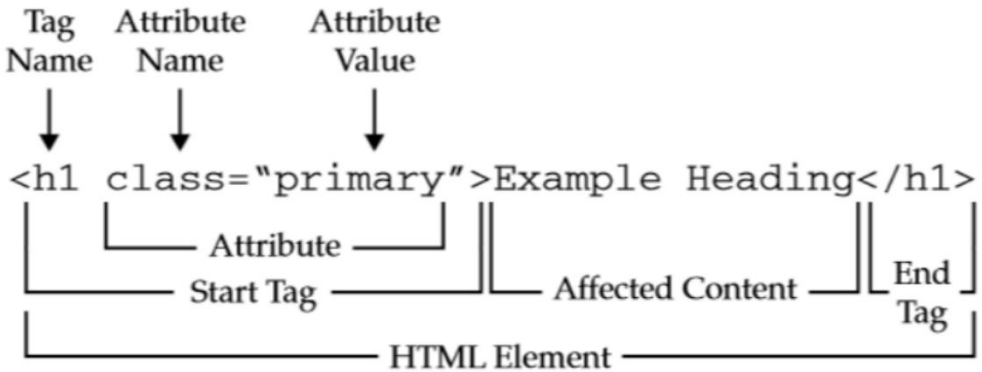
```
<html>
  <head>
    <title> Page title </title>
  </head>
  <body>
    <h1> Page title </h1>
    <p> This is a paragraph. </p>
    <p> This is another paragraph </p>
  </body>
</html>
```

html structure



html tree → Document Object Model (DOM)

HTML Elements



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

Case 1:
Scraping The Holidays

Getting Started: Go to the intended page → Decide what you want to scrape.

- intended page: https://www.liburnasional.com/kalender-lengkap-2019/
- Part to scrape: daftar hari libur nasional

```
# A tibble: 21 × 2
   date       event
   <date>     <chr>
 1 2019-01-01 Tahun Baru Masehi
 2 2019-02-05 Tahun Baru Imlek
 3 2019-03-07 Hari Raya Nyepi
 4 2019-04-03 Isra Miraj
 5 2019-04-17 Pemilu
 6 2019-04-19 Jumat Agung
 7 2019-05-01 Hari Buruh
 8 2019-05-19 Hari Raya Waisak
 9 2019-05-30 Kenaikan Isa Almasih
10 2019-06-01 Hari Lahir Pancasila
11 2019-06-03 Cuti Bersama
12 2019-06-04 Cuti Bersama
13 2019-06-05 Idul Fitri
14 2019-06-06 Idul Fitri
15 2019-06-07 Cuti Bersama
16 2019-08-11 Idul Adha
17 2019-08-17 Hari Kemerdekaan
18 2019-09-01 Tahun Baru Hijriyah
19 2019-11-09 Maulid Nabi
20 2019-12-24 Cuti Bersama
21 2019-12-25 Hari Natal
```

Scraping holidays the human way

- open the page: https://www.liburnasional.com/kalender-lengkap-2019/
- select, copy, and paste the holiday dates.

Doing step 1: The rvest way

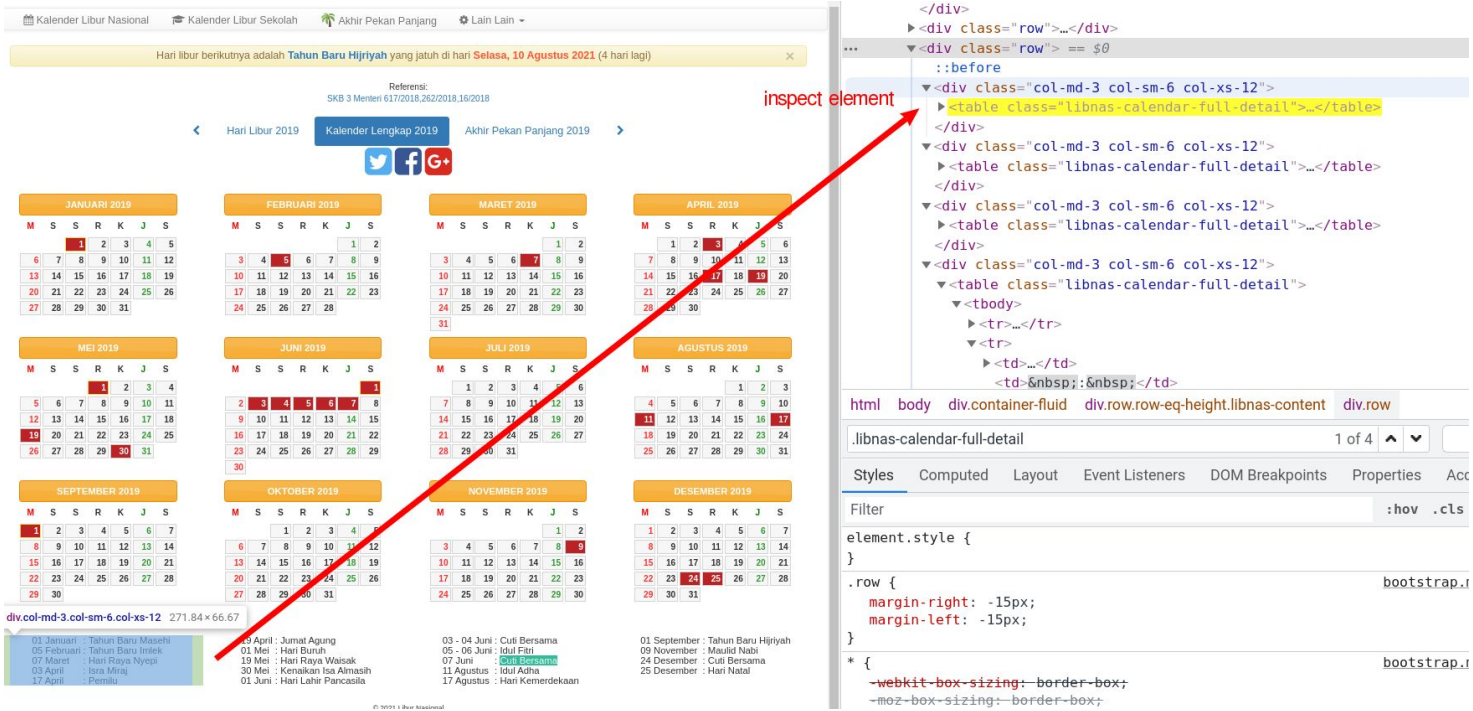
1. open the page
2. function: read_html()

```
> libur2019_html <- read_html(libur2019_url)
> libur2019_html
{html_document}
<html>
[1] <head>\n<meta http-equiv="Content-Type" content
="text/html; charset=utf-8">\n<meta name="viewport" c
o ...
[2] <body>\r\n\t\t\t\t<div id="feedback-ribbon">\r\n
\t\t\t<a href="/kontak/">Hubungi Kami  <i class="fa f
 ...
```

read_html() is equal to a human open a page in a browser

Doing step 2: Explore, Inspect, and Select Elements (The rvest way)

right click on the browser → inspect



Now that you know the element of your interest, select it using html_nodes()

```
>   read_html(libur2019_url) %>%
+   html_nodes(".libnas-calendar-full-detail") %>%
```

why is this?

```
{xml_nodeset (4)}
[1] <table class="libnas-calendar-full-detail">\n<tr>\n<td ...
[2] <table class="libnas-calendar-full-detail">\n<tr>\n<td ...
[3] <table class="libnas-calendar-full-detail">\n<tr>\n<td ...
[4] <table class="libnas-calendar-full-detail">\n<tr>\n<td ...
```

doesn't look really good → html_table()

```
>   read_html(libur2019_url) %>%
+   html_nodes(".libnas-calendar-full-detail") %>%
+   html_table()
```

```
[[1]]
# A tibble: 5 × 3
  X1          X2    X3
  <chr>       <chr> <chr>
1 01 Januari  :     Tahun Baru Masehi
2 05 Februari :     Tahun Baru Imlek
3 07 Maret    :     Hari Raya Nyepi
4 03 April    :     Isra Miraj
5 17 April    :     Pemilu

[[2]]
# A tibble: 5 × 3
  X1       X2    X3
  <chr>    <chr> <chr>
1 19 April :     Jumat Agung
2 01 Mei   :     Hari Buruh
3 19 Mei   :     Hari Raya Waisak
4 30 Mei   :     Kenaikan Isa Almasih
5 01 Juni  :     Hari Lahir Pancasila

[[3]]
# A tibble: 5 × 3
  X1           X2    X3
  <chr>        <chr> <chr>
1 03 - 04 Juni :     Cuti Bersama
2 05 - 06 Juni :     Idul Fitri
3 07 Juni      :     Cuti Bersama
4 11 Agustus   :     Idul Adha
5 17 Agustus   :     Hari Kemerdekaan

[[4]]
# A tibble: 4 × 3
  X1           X2    X3
  <chr>        <chr> <chr>
1 01 September :     Tahun Baru Hijriyah
2 09 November  :     Maulid Nabi
3 24 Desember  :     Cuti Bersama
4 25 Desember  :     Hari Natal
```

Preprocess the data

```
[[1]]
# A tibble: 5 × 3
  X1          X2    X3
  <chr>       <chr> <chr>
1 01 Januari  :     Tahun Baru Masehi
2 05 Februari :     Tahun Baru Imlek
3 07 Maret    :     Hari Raya Nyepi
4 03 April    :     Isra Miraj
5 17 April    :     Pemilu

[[2]]
# A tibble: 5 × 3
  X1       X2    X3
  <chr>    <chr> <chr>
1 19 April :     Jumat Agung
2 01 Mei   :     Hari Buruh
3 19 Mei   :     Hari Raya Waisak
4 30 Mei   :     Kenaikan Isa Almasih
5 01 Juni  :     Hari Lahir Pancasila

[[3]]
# A tibble: 5 × 3
  X1           X2    X3
  <chr>        <chr> <chr>
1 03 - 04 Juni :     Cuti Bersama
2 05 - 06 Juni :     Idul Fitri
3 07 Juni      :     Cuti Bersama
4 11 Agustus   :     Idul Adha
5 17 Agustus   :     Hari Kemerdekaan

[[4]]
# A tibble: 4 × 3
  X1           X2    X3
  <chr>        <chr> <chr>
1 01 September :     Tahun Baru Hijriyah
2 09 November  :     Maulid Nabi
3 24 Desember  :     Cuti Bersama
4 25 Desember  :     Hari Natal
```

libur2019_list

bind_rows()

	X1	X2	X3
	<chr>	<chr>	<chr>
1	01 Januari	:	Tahun Baru Masehi
2	05 Februari	:	Tahun Baru Imlek
3	07 Maret	:	Hari Raya Nyepi
4	03 April	:	Isra Miraj
5	17 April	:	Pemilu
6	19 April	:	Jumat Agung
7	01 Mei	:	Hari Buruh
8	19 Mei	:	Hari Raya Waisak
9	30 Mei	:	Kenaikan Isa Almasih
10	01 Juni	:	Hari Lahir Pancasila
11	03 - 04 Juni	:	Cuti Bersama
12	05 - 06 Juni	:	Idul Fitri
13	07 Juni	:	Cuti Bersama
14	11 Agustus	:	Idul Adha
15	17 Agustus	:	Hari Kemerdekaan
16	01 September	:	Tahun Baru Hijriyah
17	09 November	:	Maulid Nabi
18	24 Desember	:	Cuti Bersama
19	25 Desember	:	Hari Natal

Preprocess the data

```
transmute(
   date = paste(X1, "2019"),
   event = X3
  )
```

```
# A tibble: 19 × 2
   date              event
   <chr>             <chr>
 1 01 Januari 2019   Tahun Baru Masehi
 2 05 Februari 2019  Tahun Baru Imlek
 3 07 Maret 2019     Hari Raya Nyepi
 4 03 April 2019     Isra Miraj
 5 17 April 2019     Pemilu
 6 19 April 2019     Jumat Agung
 7 01 Mei 2019       Hari Buruh
 8 19 Mei 2019       Hari Raya Waisak
 9 30 Mei 2019       Kenaikan Isa Almasih
10 01 Juni 2019      Hari Lahir Pancasila
11 03 - 04 Juni 2019 Cuti Bersama
12 05 - 06 Juni 2019 Idul Fitri
13 07 Juni 2019      Cuti Bersama
14 11 Agustus 2019   Idul Adha
15 17 Agustus 2019   Hari Kemerdekaan
16 01 September 2019 Tahun Baru Hijriyah
17 09 November 2019  Maulid Nabi
18 24 Desember 2019  Cuti Bersama
19 25 Desember 2019  Hari Natal
```

handle this

Preprocess the data

```
separate_rows(
    date,
    sep = " - "
  )
```

```
# A tibble: 21 × 2
   date              event
   <chr>             <chr>
 1 01 Januari 2019   Tahun Baru Masehi
 2 05 Februari 2019  Tahun Baru Imlek
 3 07 Maret 2019     Hari Raya Nyepi
 4 03 April 2019     Isra Miraj
 5 17 April 2019     Pemilu
 6 19 April 2019     Jumat Agung
 7 01 Mei 2019       Hari Buruh
 8 19 Mei 2019       Hari Raya Waisak
 9 30 Mei 2019       Kenaikan Isa Almasih
10 01 Juni 2019      Hari Lahir Pancasila
11 03                Cuti Bersama
12 04 Juni 2019      Cuti Bersama
13 05                Idul Fitri
14 06 Juni 2019      Idul Fitri
15 07 Juni 2019      Cuti Bersama
16 11 Agustus 2019   Idul Adha
17 17 Agustus 2019   Hari Kemerdekaan
18 01 September 2019 Tahun Baru Hijriyah
19 09 November 2019  Maulid Nabi
20 24 Desember 2019  Cuti Bersama
21 25 Desember 2019  Hari Natal
```

```
separate(
    date,
    into = c("date", "month", "year"),
    sep = "\\s",
    fill = "right"
  )
```

```
# A tibble: 21 × 4
   date  month     year  event
   <chr> <chr>     <chr> <chr>
 1 01    Januari   2019  Tahun Baru Masehi
 2 05    Februari  2019  Tahun Baru Imlek
 3 07    Maret     2019  Hari Raya Nyepi
 4 03    April     2019  Isra Miraj
 5 17    April     2019  Pemilu
 6 19    April     2019  Jumat Agung
 7 01    Mei       2019  Hari Buruh
 8 19    Mei       2019  Hari Raya Waisak
 9 30    Mei       2019  Kenaikan Isa Almasih
10 01    Juni      2019  Hari Lahir Pancasila
11 03    NA        NA    Cuti Bersama
12 04    Juni      2019  Cuti Bersama
13 05    NA        NA    Idul Fitri
14 06    Juni      2019  Idul Fitri
15 07    Juni      2019  Cuti Bersama
16 11    Agustus   2019  Idul Adha
17 17    Agustus   2019  Hari Kemerdekaan
18 01    September 2019  Tahun Baru Hijriyah
19 09    November  2019  Maulid Nabi
20 24    Desember  2019  Cuti Bersama
21 25    Desember  2019  Hari Natal
```

Preprocess the data

```
separate(
    date,
    into = c("date", "month", "year"),
    sep = "\\s",
    fill = "right"
  )
```

```
# A tibble: 21 × 4
   date  month     year  event
   <chr> <chr>     <chr> <chr>
 1 01    Januari   2019  Tahun Baru Masehi
 2 05    Februari  2019  Tahun Baru Imlek
 3 07    Maret     2019  Hari Raya Nyepi
 4 03    April     2019  Isra Miraj
 5 17    April     2019  Pemilu
 6 19    April     2019  Jumat Agung
 7 01    Mei       2019  Hari Buruh
 8 19    Mei       2019  Hari Raya Waisak
 9 30    Mei       2019  Kenaikan Isa Almasih
10 01    Juni      2019  Hari Lahir Pancasila
11 03    NA        NA    Cuti Bersama
12 04    Juni      2019  Cuti Bersama
13 05    NA        NA    Idul Fitri
14 06    Juni      2019  Idul Fitri
15 07    Juni      2019  Cuti Bersama
16 11    Agustus   2019  Idul Adha
17 17    Agustus   2019  Hari Kemerdekaan
18 01    September 2019  Tahun Baru Hijriyah
19 09    November  2019  Maulid Nabi
20 24    Desember  2019  Cuti Bersama
21 25    Desember  2019  Hari Natal
```

deal with this

Preprocess the data

fill(month, year, .direction = "up")

A tibble: 21 × 4

	date	month	year	event
	<chr>	<chr>	<chr>	<chr>
1	01	Januari	2019	Tahun Baru Masehi
2	05	Februari	2019	Tahun Baru Imlek
3	07	Maret	2019	Hari Raya Nyepi
4	03	April	2019	Isra Miraj
5	17	April	2019	Pemilu
6	19	April	2019	Jumat Agung
7	01	Mei	2019	Hari Buruh
8	19	Mei	2019	Hari Raya Waisak
9	30	Mei	2019	Kenaikan Isa Almasih
10	01	Juni	2019	Hari Lahir Pancasila
11	03	Juni	2019	Cuti Bersama
12	04	Juni	2019	Cuti Bersama
13	05	Juni	2019	Idul Fitri
14	06	Juni	2019	Idul Fitri
15	07	Juni	2019	Cuti Bersama
16	11	Agustus	2019	Idul Adha
17	17	Agustus	2019	Hari Kemerdekaan
18	01	September	2019	Tahun Baru Hijriyah
19	09	November	2019	Maulid Nabi
20	24	Desember	2019	Cuti Bersama
21	25	Desember	2019	Hari Natal

Preprocess the data

```
mutate(
   month = recode(
     month,
     "Januari" = "January",
     "Februari" = "February",
     "Maret" = "March",
     "April" = "April",
     "Mei" = "May",
     "Juni" = "June",
     "Juli" = "July",
     "Agustus" = "August",
     "September" = "September",
     "Oktober" = "October",
     "November" = "November",
     "Desember" = "December"
   )
 )
```

```
# A tibble: 21 × 4
   date  month     year  event
   <chr> <chr>     <chr> <chr>
 1 01    January   2019  Tahun Baru Masehi
 2 05    February  2019  Tahun Baru Imlek
 3 07    March     2019  Hari Raya Nyepi
 4 03    April     2019  Isra Miraj
 5 17    April     2019  Pemilu
 6 19    April     2019  Jumat Agung
 7 01    May       2019  Hari Buruh
 8 19    May       2019  Hari Raya Waisak
 9 30    May       2019  Kenaikan Isa Almasih
10 01    June      2019  Hari Lahir Pancasila
11 03    June      2019  Cuti Bersama
12 04    June      2019  Cuti Bersama
13 05    June      2019  Idul Fitri
14 06    June      2019  Idul Fitri
15 07    June      2019  Cuti Bersama
16 11    August    2019  Idul Adha
17 17    August    2019  Hari Kemerdekaan
18 01    September 2019  Tahun Baru Hijriyah
19 09    November  2019  Maulid Nabi
20 24    December  2019  Cuti Bersama
21 25    December  2019  Hari Natal
```

Preprocess the data

```
unite(
    col = "date",
    date,
    month,
    year,
    sep = "-"
  )
```

```
# A tibble: 21 × 2
   date                event
   <chr>               <chr>
 1 01-January-2019     Tahun Baru Masehi
 2 05-February-2019    Tahun Baru Imlek
 3 07-March-2019       Hari Raya Nyepi
 4 03-April-2019       Isra Miraj
 5 17-April-2019       Pemilu
 6 19-April-2019       Jumat Agung
 7 01-May-2019         Hari Buruh
 8 19-May-2019         Hari Raya Waisak
 9 30-May-2019         Kenaikan Isa Almasih
10 01-June-2019        Hari Lahir Pancasila
11 03-June-2019        Cuti Bersama
12 04-June-2019        Cuti Bersama
13 05-June-2019        Idul Fitri
14 06-June-2019        Idul Fitri
15 07-June-2019        Cuti Bersama
16 11-August-2019      Idul Adha
17 17-August-2019      Hari Kemerdekaan
18 01-September-2019   Tahun Baru Hijriyah
19 09-November-2019    Maulid Nabi
20 24-December-2019    Cuti Bersama
21 25-December-2019    Hari Natal
```

Preprocess the data

```
mutate(
   date = as.Date(date, format = "%e-%B-%Y")
  ) %>%
arrange(date)
```

learn more about date format in R:
https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/strptime

A tibble: 21 × 2

	date	event
	<date>	<chr>
1	2019-01-01	Tahun Baru Masehi
2	2019-02-05	Tahun Baru Imlek
3	2019-03-07	Hari Raya Nyepi
4	2019-04-03	Isra Miraj
5	2019-04-17	Pemilu
6	2019-04-19	Jumat Agung
7	2019-05-01	Hari Buruh
8	2019-05-19	Hari Raya Waisak
9	2019-05-30	Kenaikan Isa Almasih
10	2019-06-01	Hari Lahir Pancasila
11	2019-06-03	Cuti Bersama
12	2019-06-04	Cuti Bersama
13	2019-06-05	Idul Fitri
14	2019-06-06	Idul Fitri
15	2019-06-07	Cuti Bersama
16	2019-08-11	Idul Adha
17	2019-08-17	Hari Kemerdekaan
18	2019-09-01	Tahun Baru Hijriyah
19	2019-11-09	Maulid Nabi
20	2019-12-24	Cuti Bersama
21	2019-12-25	Hari Natal

Great!

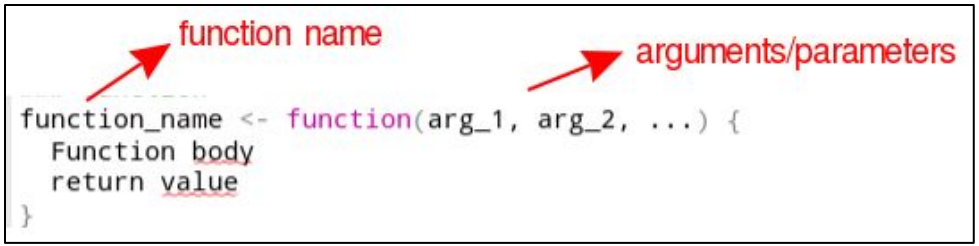
But how if we wanted to perform similar procedure but for a different year?

Of course we can re-write our previous code.

But is there any better ways?

INTRODUCTION: function creation in R.

Function in R



getHoliday(2020)
getHoliday(2021)

```r
get_holidays <- function(year) {

  Sys.setlocale("LC_TIME", "id_ID.UTF-8")

  res <-
    str_glue("https://www.liburnasional.com/kalender-lengkap-{year}/") %>%
    read_html() %>%
    html_nodes(".libnas-calendar-full-detail") %>%
    html_table() %>%
    bind_rows() %>%
    transmute(
      date = str_glue("{X1} {year}"),
      event = X3
    ) %>%
    separate_rows(
      date,
      sep = " - "
    ) %>%
    separate(
      date,
      into = c("date", "month", "year"),
      sep = "\\s",
      fill = "right"
    ) %>%
    fill(month, year, .direction = "up") %>%
    unite(
      col = "date",
      date,
      month,
      year,
      sep = "-"
    ) %>%
    mutate(
      date = as.Date(date, format = "%e-%B-%Y")
    ) %>%
    arrange(date)

  return(res)
}
```

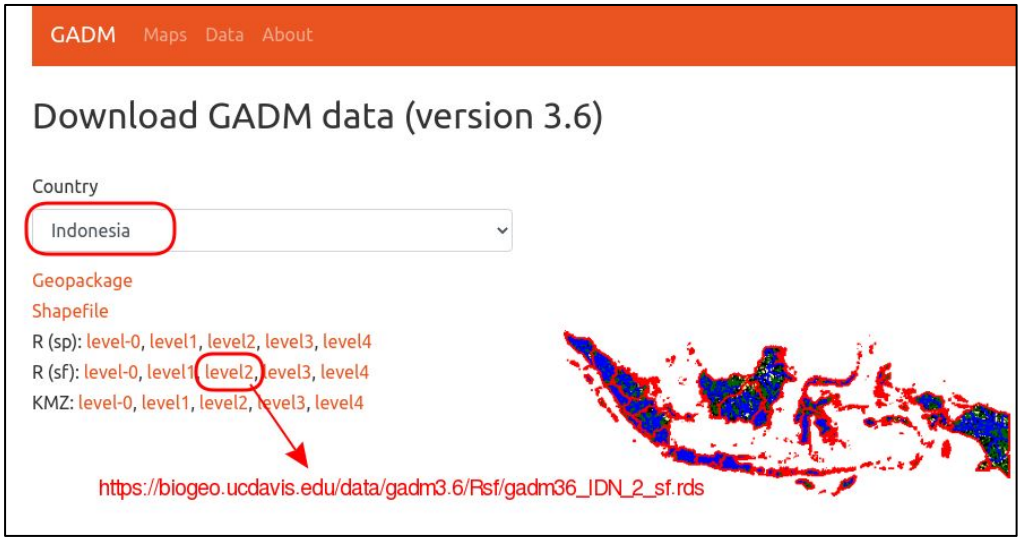
Case 2:
Stay-at-home behaviour during pandemic

Steps:

1. Get the information of administrative areas in Indonesia → Indonesia GADM data
   a. download → curl → https://biogeo.ucdavis.edu/data/gadm3.6/Rsf/gadm36_IDN_2_sf.rds
   b. import it to R
   c. extract the required columns/information
2. Get the information of people's movement → Facebook movement data.
   a. download
   b. import it to R
   c. extract the required columns/information → filter Indonesia only and selected columns only
3. Left join 2 and 1 based on field 'polygon_id'
4. Visualize:
   a. histogram
   b. boxplot
   c. dll

Indonesia GADM data

1. GADM stands for Global Administrative Areas
2. a spatial database of the location of the world's administrative areas/boundaries for use in GIS and similar software.
3. https://gadm.org/
4. We'll use it to give spatial information about the movement data we'll collect later.



Download Indonesia GADM data

	GID_0	NAME_0	GID_1	NAME_1	NL_NAME_1	GID_2	NAME_2	VARNAME_2	NL_NAME_2	TYPE_2	ENGTYPE_2	CC_2	HASC_2	geometry
11	IDN	Indonesia	IDN.1_1	Aceh	NA	IDN.1.19_1	Pidie Jaya	NA	NA	Kabupaten	Regency	1118	ID.AC.PJ	list(list(c(96.34819031, 96.3481597
14	IDN	Indonesia	IDN.1_1	Aceh	NA	IDN.1.21_1	Sabang	NA	NA	Kota	City	1172	ID.AC.SA	list(list(c(95.11791992, 95.1180419
15	IDN	Indonesia	IDN.1_1	Aceh	NA	IDN.1.22_1	Simeulue	NA	NA	Kabupaten	Regency	1101	ID.AC.SI	list(list(c(96.66509247, 96.6676483
16	IDN	Indonesia	IDN.1_1	Aceh	NA	IDN.1.23_1	Subulussalam	NA	NA	Kota	City	1175	ID.AC.SU	list(list(c(97.85132599, 97.8507919
194	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.1_1	Badung	NA	NA	Kabupaten	Regency	5103	ID.BA.BD	list(list(c(115.21576691, 115.2156
195	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.2_1	Bangli	NA	NA	Kabupaten	Regency	5106	ID.BA.BN	list(list(c(115.39585876, 115.39581
196	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.3_1	Buleleng	NA	NA	Kabupaten	Regency	5108	ID.BA.BL	list(list(c(114.59265137, 114.59272
197	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.4_1	Denpasar	NA	NA	Kota	City	5171	ID.BA.DE	list(list(c(115.23284149, 115.23267
198	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.5_1	Gianyar	NA	NA	Kabupaten	Regency	5104	ID.BA.GI	list(list(c(115.3017807, 115.301872
199	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.6_1	Jembrana	NA	NA	Kabupaten	Regency	5101	ID.BA.JE	list(list(c(114.64027405, 114.63976
200	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.7_1	Karangasem	NA	NA	Kabupaten	Regency	5107	ID.BA.KA	list(list(c(115.55763245, 115.5568
201	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.8_1	Klungkung	NA	NA	Kabupaten	Regency	5105	ID.BA.KL	list(list(c(115.58135223, 115.5811
202	IDN	Indonesia	IDN.2_1	Bali	NA	IDN.2.9_1	Tabanan	NA	NA	Kabupaten	Regency	5102	ID.BA.TA	list(list(c(115.15141296, 115.1515
356	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.4_1	Bangka	NA	NA	Kabupaten	Regency	1901	ID.BB.BA	list(list(c(105.75632477, 105.7561
353	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.1_1	Bangka Barat	NA	NA	Kabupaten	Regency	1903	ID.BB.BB	list(list(c(105.45385742, 105.4522
354	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.2_1	Bangka Selatan	NA	NA	Kabupaten	Regency	1905	ID.BB.BS	list(list(c(106.51670074, 106.5167
355	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.3_1	Bangka Tengah	NA	NA	Kabupaten	Regency	1904	ID.BB.BG	list(list(c(105.75260162, 105.7526
358	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.6_1	Belitung	NA	NA	Kabupaten	Regency	1902	ID.BB.BE	list(list(c(107.73009491, 107.7313
357	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.5_1	Belitung Timur	NA	NA	Kabupaten	Regency	1906	ID.BB.BT	list(list(c(107.96173096, 107.9621
359	IDN	Indonesia	IDN.3_1	Bangka Belitung	NA	IDN.3.7_1	Pangkalpinang	NA	NA	Kota	City	1971	ID.BB.PP	list(list(c(106.11721039, 106.1109
434	IDN	Indonesia	IDN.4_1	Banten	NA	IDN.4.1_1	Cilegon	NA	NA	Kota	City	3672	ID.BT.CL	list(list(c(105.99121094, 105.9917
435	IDN	Indonesia	IDN.4_1	Banten	NA	IDN.4.2_1	Kota Serang	NA	NA	Kota	City	3673	ID.BT.SK	list(list(c(106.15544891, 106.1552
436	IDN	Indonesia	IDN.4_1	Banten	NA	IDN.4.3_1	Kota Tangerang	NA	NA	Kota	City	3671	ID.BT.TM	list(list(c(106.56490326, 106.5653

Showing 22 to 45 of 502 entries, 14 total columns

GADM Metadata

Variable names for level "i", where "i" can be 1, 2, 3, 4, or 5

Variable	Type	Description
GID_i	String	Preferred unique ID at level i. See discussion below
ID_i	Integer	Alternative unique identifies at level 1. See discussion below
NAME_i	String	Official name in latin script
VARNAME_i	String	Variant name. Alternate names in usage for the place, separated by pipes |
NL_NAME_i	String	Non-Latin name. Official name in a non-latin script (e.g. Arabic, Chinese, Russian, Korean)
HASC_i	String	HASC. A unique ID from Statoids
CC_i	String	Country code. Uniqe ID used within the country
TYPE_i	String	Administrative type in local language
ENGTYPE_i	String	Administrative type in English (following commonly used translations)
VALIDFR_i	String	Valiid From. Date from which data is known to have started. default: Unknown. Format is YYYY-MM-DD or YYYY-MM or YYYY
VALIDTO_i	String	Valid To. Date at which data is no longer valid. default: Present or Current. Format is YYYY-MM-DD or YYYY-MM or YYYY
REMARKS_i	String	Comments about edits, relevant to history. For example "This is a split from Matam region."

https://gadm.org/metadata.html

Facebook Movement Range data

- **What?**
  - How populations are responding to physical distancing measures.
  - Two metrics:
    - Change in Movement: how much people are moving around
    - Stay Put: looks at the proportion of population that appear to stay within a small area during an entire day.
- **Who?**
  - people who opt in to Location History and background location collection

```
[erikaris@erikaris-inspiron7000 movement-range-data-2020-03-01-2020-12-31]$ head movement-range-data-2020-03-01--2020-12-31.txt
ds      country polygon_source  polygon_id      polygon_name    all_day_bing_tiles_visited_relative_change      all_day_ratio_single_tile_users baseline_name   baseline_type
2020-03-01      AGO     GADM    AGO.10.10_1     Lubango -0.02992        0.18751 full_february   DAY_OF_WEEK
2020-03-02      AGO     GADM    AGO.10.10_1     Lubango 0.06746 0.10521 full_february   DAY_OF_WEEK
2020-03-03      AGO     GADM    AGO.10.10_1     Lubango 0.05873 0.11397 full_february   DAY_OF_WEEK
2020-03-04      AGO     GADM    AGO.10.10_1     Lubango 0.01288 0.10492 full_february   DAY_OF_WEEK
2020-03-05      AGO     GADM    AGO.10.10_1     Lubango 0.02753 0.11056 full_february   DAY_OF_WEEK
2020-03-06      AGO     GADM    AGO.10.10_1     Lubango 0.03778 0.11842 full_february   DAY_OF_WEEK
2020-03-07      AGO     GADM    AGO.10.10_1     Lubango 0.05156 0.12881 full_february   DAY_OF_WEEK
2020-03-08      AGO     GADM    AGO.10.10_1     Lubango -0.02778        0.18324 full_february   DAY_OF_WEEK
2020-03-09      AGO     GADM    AGO.10.10_1     Lubango 0.0476  0.11499 full_february   DAY_OF_WEEK
```

Facebook Movement Range data (2)

```
Facebook Movement Range Maps

This data includes movement changes measured by Facebook throughout March,
April, May, and June 2020 starting from a baseline in February. Data is provided in one global tab-delimited
text file.

Columns

- ds: Date stamp for movement range data row in YYYY-MM-DD form
- country: Three-character ISO-3166 country code
- polygon_source: Source of region polygon, either â€œFIPSâ€ for U.S. data or â€œGADMâ€ for global data
- polygon_id: Unique identifier for region polygon, either numeric string for U.S. FIPS codes or alphanumeric
string for GADM regions
- polygon_name: Region name
- all_day_bing_tiles_visited_relative_change: Positive or negative change in movement relative to baseline
- all_day_ratio_single_tile_users: Positive proportion of users staying put within a single location
- baseline_name: When baseline movement was calculated pre-COVID-19
- baseline_type: How baseline movement was calculated pre-COVID-19
```

- https://data.humdata.org/dataset/c3429f0e-651b-4788-bb2f-4adbf222c90e/resource/435ed157-6f7a-4e8f-a63a-2aa177b9bd05/download/readme.txt
- https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/

Tools / Lab Online

Let’s jump back to RStudio

Summary

1. Web scraping:
   a. Identify the element to scrape
   b. scrape it using selector
   c. preprocess it
   d. do something on it (visualization, etc.)
2. Explore more:
   a. scraping for dynamic webpage → involving javascript
   b. scraping for page that need authentication.

#Jadijagoandigital

Terima Kasih

digitalent.kominfo

DTS_kominfo

digitalent.kominfo

digital talent scholarship