



DIGITAL
TALENT
SCHOLARSHIP



thematic Academy

Tema Pelatihan: Big Data for Social Science
Pertemuan #5: Data Scraping & Crawling With R

Pemateri: Erika Siregar



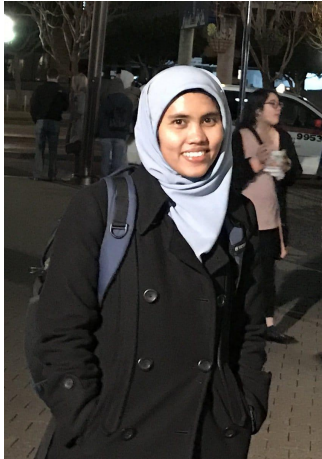
KOMINFO



#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)

Connect with Me:

Email : erika.mukhlisina@gmail.com
GitHub : <https://github.com/erikaris>
Twitter : @erikaris
Instagram : @erikaris15

Learning Objective

In this course you will:

- A. Learn basic understanding of web scraping
- B. Learn how to scrape holiday dates from a website
- C. Learn how to 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.

Self Check

- Ada yang belum pernah pakai R sebelum join acara ini?
- On scale of 1 - 10, how confident are you in using R?
- Experience with web scraping?
- Web programming?
- Inspect Element?
- Sudah install R dan RStudio?
- Siapa yang merasa sudah siap untuk belajar malam ini?

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 scrape?
 - a. Basically any page
 - b. Could be tricky when the page is complex.
 - c. Complex == rich of javascript.



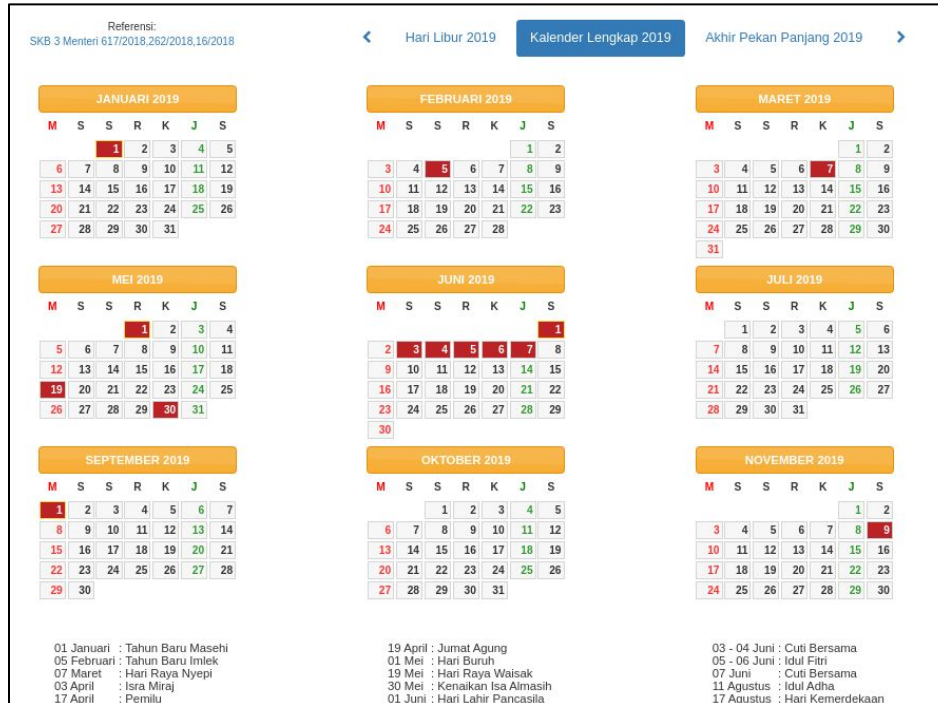
Web Scraping 101

6. What tools to use?
 - a. Python → scrapy, beautiful soup
 - b. R → rvest.
 - i. check the documentation:
<https://rvest.tidyverse.org/>
 - ii. dependent to other libraries such as: xml2, etc
 - iii. important functions:
 1. - read_html() --> convert a website into an XML object.
 2. - html_nodes() --> extract the relevant nodes from the XML object
 3. - html_text() --> extract the tagged data from the wanted nodes.
 4. - html_attrs() --> return a list of the attributes.

Steps in Web Scraping

1. Open the webpage
2. Explore it → inspect
3. scrape the page using scripts → obtain the data
4. preprocessed the data

Dissecting HTML Page



what we see

DTS 2021

```
<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="//libfont-awsone.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 keputusan pemerintah">
    <script>
        (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=a.src=g+"&pageid=1";a.async=1;a.src=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["u0870u0806fu0870"]=["site/u0849d",958255],["minB/u0869fu0864",0],["\u0870puu0866fu0869"]
    })/* */
    </script>
</head>
<body>
    <div id="feedback-ribbon">
    <a href="#kontak"><Hubungi Kami></a>
    </div>
    <div class="container-fluid">
    <div class="row row-logo">
    <!-- For Logo -->
    <div class="col-md-4 col-xs-12 text-center"><div id="libnas-linkid-logo" href="/">Libur Nasional</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>
    </div>

```

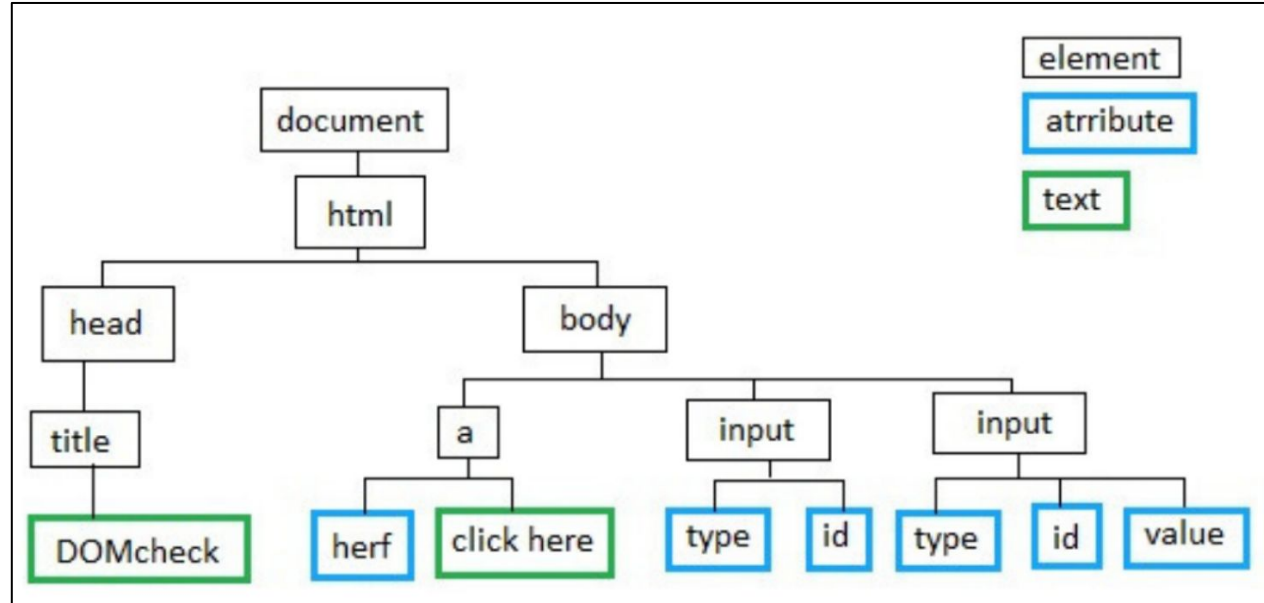
what actually behinds it

#Jadijagoandi

Dissecting HTML Page

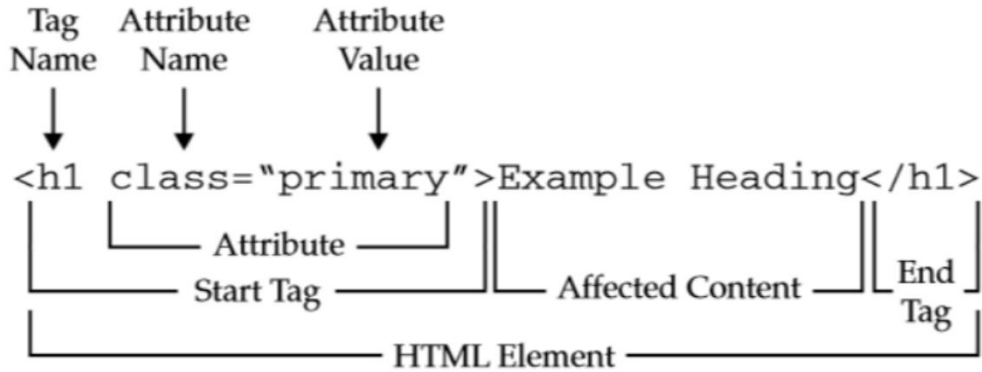
```
<html>
<head>
  <title> Page title </title>
</head>
<body>
  <h1> Page title </h1>
  <p> This is a paragraph. </p>
  <p> This is another paragraph </p>
</body>
</html>
```

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 <i>n</i> > ... </h <i>n</i> >	Delimits a level <i>n</i> heading
 ... 	Set ... in boldface
<i> ... </i>	Set ... in italics
<center> ... </center>	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
	Displays an image here
 ... 	Defines a hyperlink

Case 1:

Scraping The Holidays

Steps

1. read the webpage
2. identify the element that contains information about holiday dates.
3. select the element and obtain the data
4. preprocess it.

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

Libur Nasional

Kalender Libur Nasional Kalender Libur Sekolah Akhir Pekan Panjang Lain Lain ▾

Hari libur berikutnya adalah **Tahun Baru Hijriyah** yang jatuh di hari **Selasa, 10 Agustus 2021** (4 hari lagi) ✕

Referensi:
SKB 3 Menteri 617/2018, 262/2018, 16/2018

< Hari Libur 2019 **Kalender Lengkap 2019** Akhir Pekan Panjang 2019 >

Twitter Facebook Google+

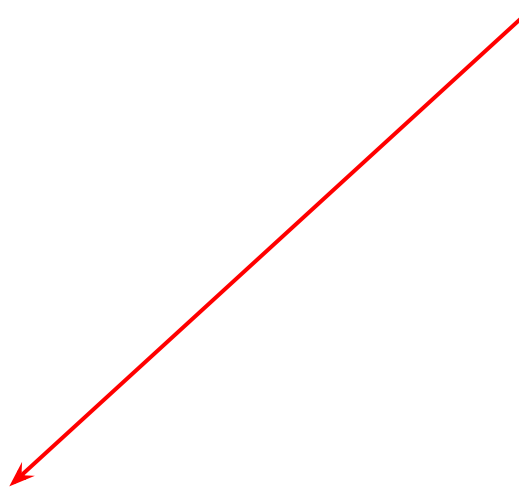
JANUARI 2019							FEBRUARI 2019							MARET 2019							APRIL 2019						
M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S
6	7	8	9	10	11	12	3	4	5	6	7	8	9	3	4	5	6	7	8	9	7	8	9	10	11	12	13
13	14	15	16	17	18	19	10	11	12	13	14	15	16	10	11	12	13	14	15	16	14	15	16	17	18	19	20
20	21	22	23	24	25	26	17	18	19	20	21	22	23	17	18	19	20	21	22	23	21	22	23	24	25	26	27
27	28	29	30	31			24	25	26	27	28	29	30	24	25	26	27	28	29	30	28	29	30				

MEI 2019							JUNI 2019							JULI 2019							AGUSTUS 2019						
M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S
5	6	7	8	9	10	11	2	3	4	5	6	7	8	7	8	9	10	11	12	13	4	5	6	7	8	9	10
12	13	14	15	16	17	18	9	10	11	12	13	14	15	14	15	16	17	18	19	20	11	12	13	14	15	16	17
19	20	21	22	23	24	25	16	17	18	19	20	21	22	21	22	23	24	25	26	27	18	19	20	21	22	23	24
26	27	28	29	30	31		23	24	25	26	27	28	29	28	29	30	31				25	26	27	28	29	30	31

SEPTEMBER 2019							OKTOBER 2019							NOVEMBER 2019							DESEMBER 2019						
M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S
1	2	3	4	5	6	7	6	7	8	9	10	11	12	3	4	5	6	7	8	9	1	2	3	4	5	6	7
8	9	10	11	12	13	14	13	14	15	16	17	18	19	10	11	12	13	14	15	16	8	9	10	11	12	13	14
15	16	17	18	19	20	21	20	21	22	23	24	25	26	17	18	19	20	21	22	23	15	16	17	18	19	20	21
22	23	24	25	26	27	28	27	28	29	30	31			24	25	26	27	28	29	30	22	23	24	25	26	27	28
29	30													29	30	31					29	30	31				

01 Januari : Tahun Baru Masehi
05 Februari : Tahun Baru Imlek
07 Maret : Hari Raya Nyepi
03 April : Isra Miraj
17 April : Pemilu
19 April : Jumat Agung
01 Mei : Hari Buruh
19 Mei : Hari Raya Waisak
30 Mei : Kenaikan Isa Almasih
01 Juni : Hari Lahir Pancasila
03 - 04 Juni : Cuti Bersama
05 - 06 Juni : Idul Fitri
07 Juni : Cuti Bersama
11 Agustus : Idul Adha
17 Agustus : Hari Kemerdekaan
01 September : Tahun Baru Hijriyah
09 November : Maulid Nabi
24 Desember : Cuti Bersama
25 Desember : Hari Natal

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



Scraping holidays the human way

Libur Nasional

Kalender Libur Nasional Kalender Libur Sekolah Akhir Pekan Panjang Lain Lain ▾

Hari libur berikutnya adalah **Tahun Baru Hijriyah** yang jatuh di hari **Selasa, 10 Agustus 2021** (4 hari lagi) ✕

Referensi:
SKB 3 Menteri 617/2018, 262/2018, 16/2018

< Hari Libur 2019 **Kalender Lengkap 2019** Akhir Pekan Panjang 2019 >

🐦 📘 🌐

JANUARI 2019							FEBRUARI 2019							MARET 2019							APRIL 2019						
M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S
			1	2	3	4						1	2						1	2			1	2	3	4	5
6	7	8	9	10	11	12	3	4	5	6	7	8	9	3	4	5	6	7	8	9	7	8	9	10	11	12	13
13	14	15	16	17	18	19	10	11	12	13	14	15	16	10	11	12	13	14	15	16	14	15	16	17	18	19	20
20	21	22	23	24	25	26	17	18	19	20	21	22	23	17	18	19	20	21	22	23	21	22	23	24	25	26	27
27	28	29	30	31			24	25	26	27	28	29	30	24	25	26	27	28	29	30	28	29	30				

MEI 2019							JUNI 2019							JULI 2019							AGUSTUS 2019						
M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S
			1	2	3	4						1			1	2	3	4	5	6				1	2	3	
5	6	7	8	9	10	11	2	3	4	5	6	7	8	7	8	9	10	11	12	13	4	5	6	7	8	9	10
12	13	14	15	16	17	18	9	10	11	12	13	14	15	14	15	16	17	18	19	20	11	12	13	14	15	16	17
19	20	21	22	23	24	25	16	17	18	19	20	21	22	21	22	23	24	25	26	27	18	19	20	21	22	23	24
26	27	28	29	30	31		23	24	25	26	27	28	29	28	29	30	31				25	26	27	28	29	30	31
							30																				

SEPTEMBER 2019							OKTOBER 2019							NOVEMBER 2019							DESEMBER 2019							
M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	M	S	S	R	K	J	S	
1	2	3	4	5	6	7				1	2	3	4	5						1	2	1	2	3	4	5	6	7
8	9	10	11	12	13	14	6	7	8	9	10	11	12	3	4	5	6	7	8	9	8	9	10	11	12	13	14	
15	16	17	18	19	20	21	13	14	15	16	17	18	19	10	11	12	13	14	15	16	15	16	17	18	19	20	21	
22	23	24	25	26	27	28	20	21	22	23	24	25	26	17	18	19	20	21	22	23	22	23	24	25	26	27	28	
29	30						27	28	29	30	31			24	25	26	27	28	29	30	29	30	31					

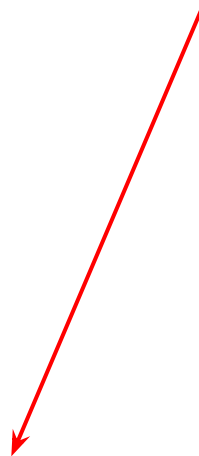
01 Januari : Tahun Baru Masehi
05 Februari : Tahun Baru Imlek
07 Maret : Hari Raya Nyepi
03 April : Isra Miraj
17 April : Pemilu

19 April : Jumat Agung
01 Mei : Hari Buruh
19 Mei : Hari Raya Waisak
30 Mei : Kenaikan Isa Almasih
01 Juni : Hari Lahir Pancasila

03 - 04 Juni : Cuti Bersama
05 - 06 Juni : Idul Fitri
07 Juni : Cuti Bersama
11 Agustus : Idul Adha
17 Agustus : Hari Kemerdekaan

01 September : Tahun Baru Hijriyah
09 November : Maulid Nabi
24 Desember : Cuti Bersama
25 Desember : Hari Natal

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



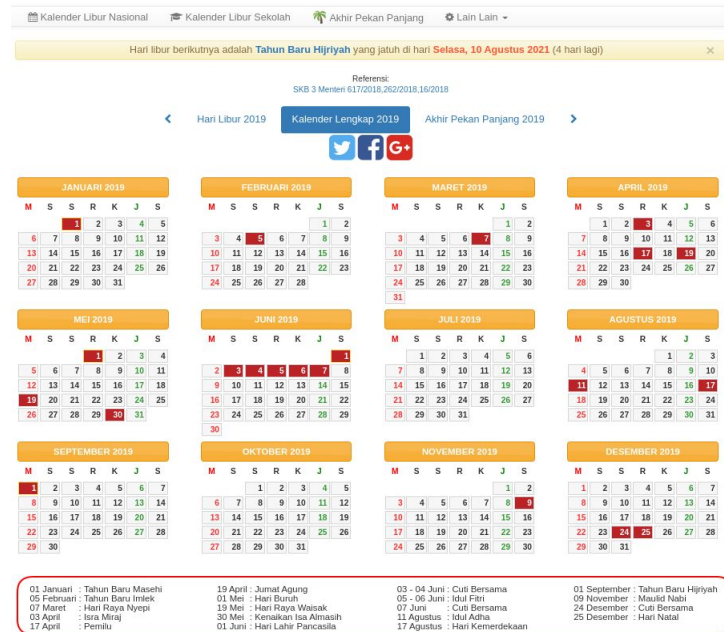
Doing step 1: The rvest way

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\t<div id="feedback-ribbon">\r\n
\t\t\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

Libur Nasional



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

right click on the browser → inspect

The screenshot shows a web browser displaying a calendar application. The calendar is for the year 2019, showing months from January to December. The date 10 August 2021 is highlighted in yellow. The 'Inspect' tool is open on the right side of the browser, showing the DOM tree. A red arrow points from the text 'inspect element' to the 'libnas-calendar-full-detail' table in the DOM tree. The DOM tree shows the following structure:

```
<div class="row">...</div>
<div class="row"> == $0
  <div class="col-md-3 col-sm-6 col-xs-12">
    <table class="libnas-calendar-full-detail">...</table>
  </div>
  <div class="col-md-3 col-sm-6 col-xs-12">
    <table class="libnas-calendar-full-detail">...</table>
  </div>
  <div class="col-md-3 col-sm-6 col-xs-12">
    <table class="libnas-calendar-full-detail">...</table>
  </div>
  <div class="col-md-3 col-sm-6 col-xs-12">
    <table class="libnas-calendar-full-detail">
      <tbody>
        <tr>...</tr>
        <tr>...</tr>
      </tbody>
    </table>
  </div>
```


The 'Styles' pane shows the following styles for the selected element:

```
element.style {
}
.row {
  margin-right: -15px;
  margin-left: -15px;
}
* {
  -webkit-box-sizing: border-box;
  -moz-box-sizing: border-box;
}
```

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. What's data preprocessing?
2. Example of data preprocessing that you know?

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
```



`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

libur2019_list

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 <chr>	month <chr>	year <chr>	event <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  
0 01-June-2019 Hari Lahir Pancasila  
1 03-June-2019 Cuti Bersama  
2 04-June-2019 Cuti Bersama  
3 05-June-2019 Idul Fitri  
4 06-June-2019 Idul Fitri  
5 07-June-2019 Cuti Bersama  
6 11-August-2019 Idul Adha  
7 17-August-2019 Hari Kemerdekaan  
8 01-September-2019 Tahun Baru Hijriyah  
9 09-November-2019 Maulid Nabi  
0 24-December-2019 Cuti Bersama  
1 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
```

Step 4: 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
```

```
libur2019 <-  
  libur2019_semiclean %>%  
  mutate(  
    # https://www.rdocumentation.org  
    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"  
    )  
  ) %>%  
  unite(  
    col = "date",  
    date,  
    month,  
    year,  
    sep = "-"  
  ) %>%  
  mutate(  
    # https://www.rdocumentation.org  
    date = as.Date(date, format = "%e-%B-%Y")  
  ) %>%  
  arrange(date)
```

This is where `Sys.setlocale()` comes in handy

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

```
libur2019 <-  
  libur2019_semiclean %>%  
  unite(  
    col = "date",  
    date,  
    month,  
    year,  
    sep = "-"  
  ) %>%  
  mutate(  
    date = as.Date(date, format = "%e-%B-%Y")  
  ) %>%  
  arrange(date) %>%  
  print(n = Inf)
```

```
Sys.setlocale("LC_TIME", "en_US.UTF-8")
```

without Sys.locale()

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

Notice the
difference?

Why?

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

```
# 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!

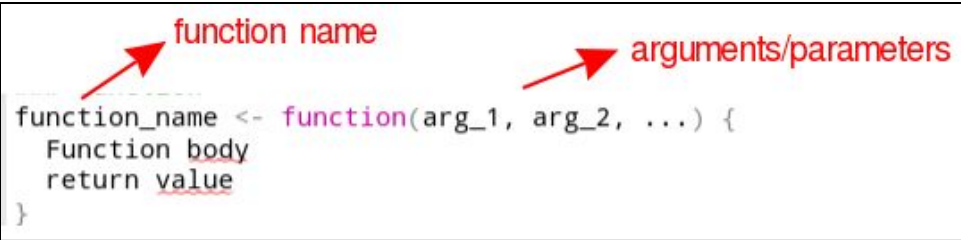
But how if we wanted to perform similar procedure but for a different year?

Of course we can re-write our previous code.

But is there any better ways?

INTRODUCTION: function creation in R.

Function in R



The diagram shows the syntax of an R function definition within a light gray box. Above the box, two red arrows point to specific parts of the code. The first arrow, labeled "function name", points to the variable `function_name` on the left. The second arrow, labeled "arguments/parameters", points to the list `(arg_1, arg_2, ...)` inside the `function` call. The code inside the box is as follows:


```
function_name <- function(arg_1, arg_2, ...) {  
  Function body  
  return value  
}
```

In the code, `function` is highlighted in pink, and `value` is underlined in red.

Create an empty function

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

```
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)  
}
```



fill it with crawling &
preprocessing script

Add several conditions

```
get_holidays <- function(year) {  
  if (length(year) != 1) {  
    stop("please specify one year only", call. = FALSE)  
    res <- NULL  
  }  
  
  if (!inherits(year, c("integer", "character"))) {  
    stop("year must be a positive integer of character string", call. = FALSE)  
    res <- NULL  
  }  
  
  if (nchar(year) != 4) {  
    stop("year must be a four digits number", call. = FALSE)  
    res <- NULL  
  }  
  
  if (as.numeric(year) < 2013 | as.numeric(year) > 2023) {  
    stop("only data since 2013 up to 2023 are available", call. = FALSE)  
    res <- NULL  
  }  
  
  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
 - 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.
4. Visualize:
 - a. histogram
 - b. boxplot
 - c. dll

Facebook Movement Range data

```
[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 "FIPS" for U.S. data or "GADM" 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/>

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 collected earlier

GADM Maps Data About

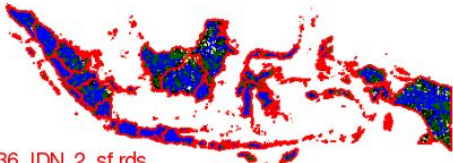
Download GADM data (version 3.6)

Country
Indonesia

Geopackage
Shapefile

R (sp): level-0, level1, level2, level3, level4
R (sf): level-0, level1, level2, level3, level4
KMZ: level-0, level1, level2, level3, level4

https://biogeo.ucdavis.edu/data/gadm3.6/Rsf/gadm36_IDN_2_sf.rds

A map of Indonesia with its administrative boundaries highlighted in red and blue. The map is located on the right side of the GADM website interface shown in the screenshot.

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	<i>NA</i>	IDN.1.19_1	Pidie Jaya	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1118	ID.AC.PJ	<i>list(list(c(96.34819031, 96.34815971</i>
14	IDN	Indonesia	IDN.1_1	Aceh	<i>NA</i>	IDN.1.21_1	Sabang	<i>NA</i>	<i>NA</i>	Kota	City	1172	ID.AC.SA	<i>list(list(c(95.11791992, 95.11804191</i>
15	IDN	Indonesia	IDN.1_1	Aceh	<i>NA</i>	IDN.1.22_1	Simeulue	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1101	ID.AC.SI	<i>list(list(c(96.66509247, 96.66764831</i>
16	IDN	Indonesia	IDN.1_1	Aceh	<i>NA</i>	IDN.1.23_1	Subulussalam	<i>NA</i>	<i>NA</i>	Kota	City	1175	ID.AC.SU	<i>list(list(c(97.85132599, 97.85079191</i>
194	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.1_1	Badung	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5103	ID.BA.BD	<i>list(list(c(115.21576691, 115.215691</i>
195	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.2_1	Bangli	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5106	ID.BA.BN	<i>list(list(c(115.39585876, 115.395811</i>
196	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.3_1	Buleleng	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5108	ID.BA.BL	<i>list(list(c(114.59265137, 114.592721</i>
197	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.4_1	Denpasar	<i>NA</i>	<i>NA</i>	Kota	City	5171	ID.BA.DE	<i>list(list(c(115.23284149, 115.232671</i>
198	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.5_1	Gianyar	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5104	ID.BA.GI	<i>list(list(c(115.3017807, 115.3018721</i>
199	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.6_1	Jembrana	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5101	ID.BA.JE	<i>list(list(c(114.64027405, 114.639761</i>
200	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.7_1	Karangasem	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5107	ID.BA.KA	<i>list(list(c(115.55763245, 115.556841</i>
201	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.8_1	Klungkung	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5105	ID.BA.KL	<i>list(list(c(115.58135223, 115.581161</i>
202	IDN	Indonesia	IDN.2_1	Bali	<i>NA</i>	IDN.2.9_1	Tabanan	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	5102	ID.BA.TA	<i>list(list(c(115.15141296, 115.151591</i>
356	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.4_1	Bangka	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1901	ID.BB.BA	<i>list(list(c(105.75632477, 105.756171</i>
353	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.1_1	Bangka Barat	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1903	ID.BB.BB	<i>list(list(c(105.45385742, 105.452271</i>
354	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.2_1	Bangka Selatan	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1905	ID.BB.BS	<i>list(list(c(106.51670074, 106.516751</i>
355	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.3_1	Bangka Tengah	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1904	ID.BB.BG	<i>list(list(c(105.75260162, 105.752691</i>
358	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.6_1	Belitung	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1902	ID.BB.BE	<i>list(list(c(107.73009491, 107.731301</i>
357	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.5_1	Belitung Timur	<i>NA</i>	<i>NA</i>	Kabupaten	Regency	1906	ID.BB.BT	<i>list(list(c(107.96173096, 107.962191</i>
359	IDN	Indonesia	IDN.3_1	Bangka Belitung	<i>NA</i>	IDN.3.7_1	Pangkalpinang	<i>NA</i>	<i>NA</i>	Kota	City	1971	ID.BB.PP	<i>list(list(c(106.11721039, 106.110981</i>
434	IDN	Indonesia	IDN.4_1	Banten	<i>NA</i>	IDN.4.1_1	Cilegon	<i>NA</i>	<i>NA</i>	Kota	City	3672	ID.BT.CL	<i>list(list(c(105.99121094, 105.991771</i>
435	IDN	Indonesia	IDN.4_1	Banten	<i>NA</i>	IDN.4.2_1	Kota Serang	<i>NA</i>	<i>NA</i>	Kota	City	3673	ID.BT.SK	<i>list(list(c(106.15544891, 106.155251</i>
436	IDN	Indonesia	IDN.4_1	Banten	<i>NA</i>	IDN.4.3_1	Kota Tangerang	<i>NA</i>	<i>NA</i>	Kota	City	3671	ID.BT.TM	<i>list(list(c(106.56490326, 106.565361</i>

Showing 22 to 45 of 502 entries, 14 total columns

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