Web scrape a Global Bike-Sharing Systems Wiki Page

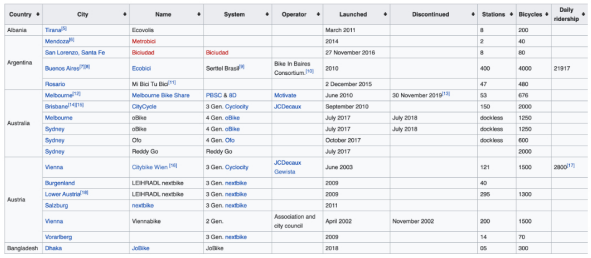
Estimated time needed: **20** minutes

Lab Overview:

Before getting your hands dirty on the actual data analysis tasks, you first need to obtain some background and context information about well-known bike sharing systems worldwide, such as their location, launch date, rental bike size, and so on.

You can get such information from this Wiki page:

https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems



Country	City	Name	System	Operator	Launched	Discontinued	Stations	Bicycles	Daily ridership
Albania	Tirana[6]	Ecovolis			March 2011		8	200	
Argentina	Mendoza[6]	Metrobici			2014		2	40	
	San Lorenzo, Santa Fe	Biciudad	Biciudad		27 November 2016		8	80	
	Buenos Aires[7][8]	Ecobici	Sertel Brasil[9]	Bike In Baires Consortium.[10]	2010		400	4000	21917
	Rosario	Mi Bici Tu Bici[11]			2 December 2015		47	480	
Australia	Melbourne[12]	Melbourne Bike Share	PBSC & BO	Motivate	June 2010	30 November 2019[13]	53	676	
	Brisbane[14][15]	CityCycle	3 Gen. Cyclocity	JCDecaux	September 2010		150	2000	
	Melbourne	oBike	4 Gen. oBike		July 2017	July 2018	dockless	1250	
	Sydney	oBike	4 Gen. oBike		July 2017	July 2018	dockless	1250	
	Sydney	Ofo	4 Gen. Ofo		October 2017		dockless	600	
	Sydney	Reddy Go	Reddy Go		July 2017			2000	
Austria	Vienna	Citybike Wien.[16]	3 Gen. Cyclocity	JCDecaux Gewista	June 2003		121	1500	2800[17]
	Burgenland	LEIHRADL nextbike	3 Gen. nextbike		2009		40		
	Lower Austria[18]	LEIHRADL nextbike	3 Gen. nextbike		2009		295	1300	
	Salzburg	nextbike	3 Gen. nextbike		2011				
	Vienna	Viennabike	2 Gen.	Association and city council	April 2002	November 2002	200	1500	
	Vorarlberg		3 Gen. nextbike		2009		14	70	
Bangladesh	Dhaka	Jolibike	Jolibike		2018		05	300	

First import necessary libraries for the webscraping task.

In this lab, you need to use the `rvest` library to obtain the bike sharing systems table from the above web page, convert the table into a data frame, and write the data frame to a csv file for future data wrangling and analysis tasks.

```
[1]  # Check if need to install rvest` library
     require("rvest")

     library(rvest)
```

```
     Loading required package: rvest
```

TASK: Extract bike sharing systems HTML table from a Wiki page and convert it into a data frame

TODO: Get the root HTML node

```
[2]  url <- "https://en.wikipedia.org/wiki/List_of_bicycle-sharing_systems"
     # Get the root HTML node by calling the `read_html()` method with URL
```

```
[4]  # Get the root HTML node by calling the `read_html()` method with URL
     root_node <- read_html(url)
```

Note that this HTML page at least contains three child `<table>` nodes under the root HTML node. So, you will need to use `html_nodes(root_node, "table")` function to get all its child `<table>` nodes:

```
<html>
  <table>(table1)</table>
  <table>(table2)</table>
  <table>(table3)</table>
  ...
</html>
```

table_nodes <- html_nodes(root_node, "table")

You can use a `for` loop to print each table, and then you will see that the actual the bike sharing table is the second element `table_nodes[[2]]`.

Next, you need to convert this HTML table into a data frame using the `html_table()` function. You may choose to include `fill = TRUE` argument to fill any empty table rows/columns.

```
[5] table_nodes <- html_nodes(root_node, "table")
    table_nodes
```

```
{xml_nodeset (6)}
[1] <table class="box-Split plainlinks metadata ambox ambox-move" role="prese ...
[2] <table class="box-Overlinked plainlinks metadata ambox ambox-style" role= ...
[3] <table class="wikitable sortable" style="text-align:left"><tbody>\n<tr>\n ...
[4] <table class="nowraplinks mw-collapsible autocollapse navbox-inner" style ...
[5] <table class="nowraplinks navbox-subgroup" style="border-spacing:0"><tbod ...
[6] <table class="nowraplinks navbox-subgroup" style="border-spacing:0"><tbod ...
```

```
[6] # Convert the bike-sharing system table into a dataframe
    bike_share <- html_table(table_nodes, fill=TRUE)
```

Summarize the bike sharing system data frame

```
[7] # Summarize the dataframe
    summary(bike_share)
```

```
     Length Class  Mode
[1,]  2     tbl_df list
[2,]  2     tbl_df list
[3,] 10     tbl_df list
[4,] 10     tbl_df list
[5,]  2     tbl_df list
[6,]  2     tbl_df list
```

```
[8] df=bike_share[[3]]
```

```
[9] summary(df)
```

```
   Country              City              Name              System
 Length:515         Length:515         Length:515         Length:515
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
   Operator           Launched         Discontinued        Stations
 Length:515         Length:515         Length:515         Length:515
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
   Bicycles         Daily ridership
 Length:515         Length:515
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Export the data frame as a csv file called `raw_bike_sharing_systems.csv`

```
# Export the dataframe into a csv file
write.csv(df,"raw_bike_sharing_systems.csv")
```

For more details about webscraping with `rvest`, please refer to the previous webscraping notebook here:

Webscraping in R

Authors

Yan Luo

Other Contributors

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2021-04-05	0.1	Yan	Initial version created