

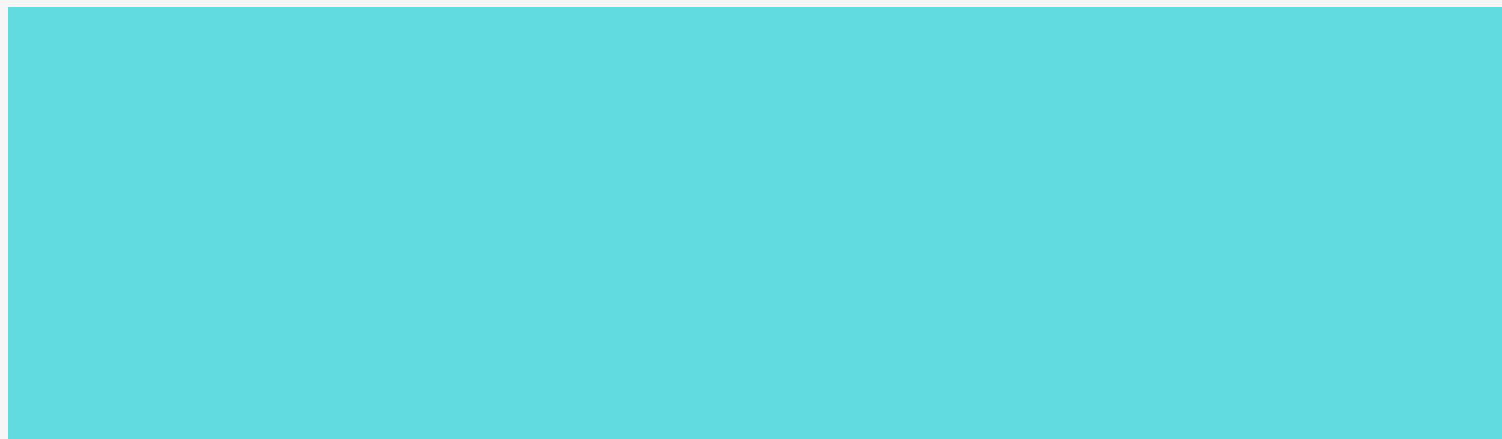


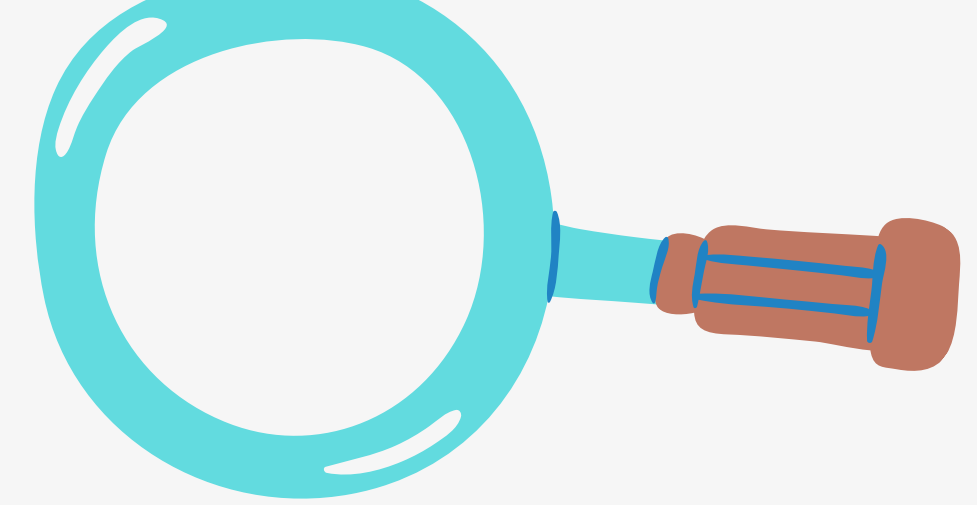
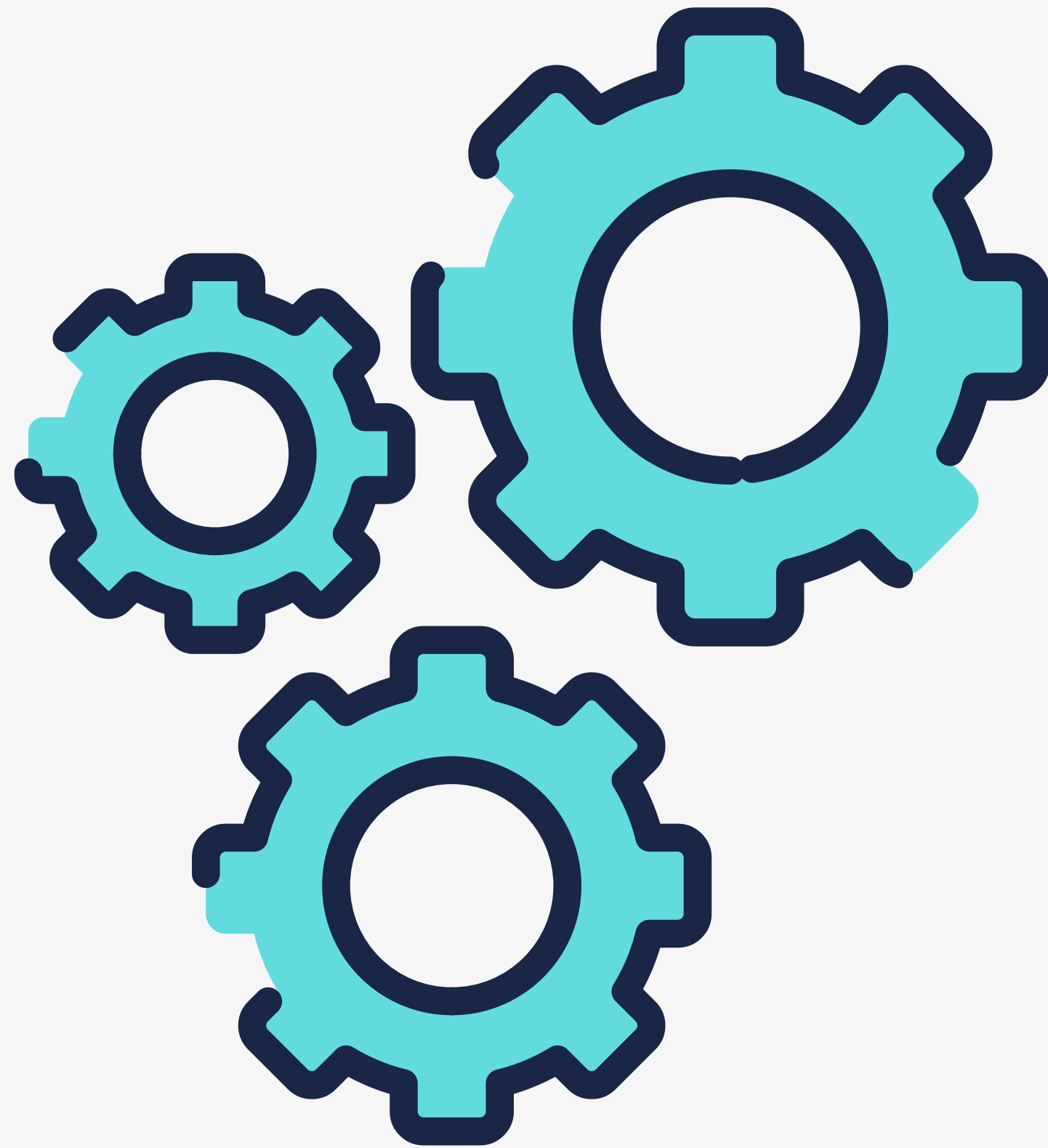
WEB SCRAPING PROJECT

POPOOLA Habib Olaitan
Data Scientist & Analyst | Tutor

PROJECT INSIGHT

Web scraping is the process of collecting and parsing raw data from the Web.
The goal of this project was to extract data from twitter using Python.



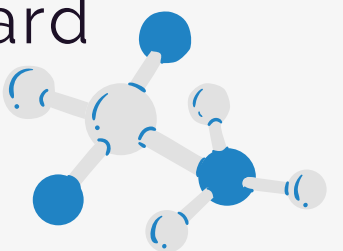


PROJECT BRIEF

In this Project, I made use of Selenium library in python because of the flexibility it offers. I was able to automate the process of logging in to my twitter account using the EdgeOption from selenium tools and extracted the data.

The next stage of the project was data cleaning, this process was required in order to get the extracted data in shape for easier visualizations and more accurate information.

To wrap it up, I created a simple dashboard to visualize the data.



DATA EXTRACTION



Importing the necessary libraries and setting up EdgeOption

```
In [1]: import selenium
        from selenium import webdriver
        from selenium.webdriver.common.by import By
        from selenium.webdriver.common.keys import Keys
        from time import sleep
        from getpass import getpass
        from selenium.common.exceptions import NoSuchElementException
        from msedge.selenium_tools import Edge, EdgeOptions
```

```
In [2]: options = EdgeOptions()
        options.use_chromium = True
        driver = Edge(options=options)
```

<ipython-input-2-e3ed45f20b59>:3: DeprecationWarning: Selenium Tools for Microsoft Edge is deprecated. Please upgrade to Selenium 4 which has built-in support for Microsoft Edge (Chromium): <https://docs.microsoft.com/en-us/microsoft-edge/webdriver-chromium/#upgrading-from-selenium-3> (<https://docs.microsoft.com/en-us/microsoft-edge/webdriver-chromium/#upgrading-from-selenium-3>)

```
    driver = Edge(options=options)
```

```
In [3]: driver.get('https://twitter.com/login')
```

```
In [47]: sleep(3)
```

Signed in to my twitter account

```
In [48]: username = driver.find_element_by_xpath("//input[@name='text']")
        username.send_keys('popsyynn')
```

```
In [49]: next_button = driver.find_element_by_xpath("//span[contains(text(),'Next')]")
        next_button.click()
```

```
In [50]: my_password = getpass()
```

```
.....
```

```
In [51]: password = driver.find_element_by_xpath("//input[@name='password']")  
password.send_keys(my_password)
```

```
In [52]: password.send_keys(Keys.RETURN)
```

```
In [53]: search_input = driver.find_element_by_xpath("//input[@aria-label='Search query']")  
search_input.send_keys('Ronaldo')
```

```
In [54]: search_input.send_keys(Keys.RETURN)
```

```
In [55]: people = driver.find_element_by_xpath("//span[contains(text(),'People')]")  
people.click()
```

```
In [56]: profile = driver.find_element_by_xpath("//*[@id='react-root']/div/div/div[2]/main/div/div/div/div/div/div/div[3]/div/section")  
profile.click()
```

Scrapping of data

```
In [57]: UserTag = driver.find_element_by_xpath('//div[@data-testid="User-Names"]').text
```

```
In [58]: print(UserTag)
```

```
Cristiano Ronaldo  
@Cristiano  
,  
24 Sep
```

```
In [59]: Time_stamp = driver.find_element_by_xpath('//time').get_attribute('datetime')
```

```
In [60]: print(Time_stamp)
```

```
2022-09-24T21:38:35.000Z
```

```
In [61]: Tweet = driver.find_element_by_xpath('//div[@data-testid="tweetText"]').text
```

```
In [62]: Tweet
```

```
Out[62]: 'Grande jogo, importante vitória equipa!Continuamos focados no nosso objetivo \nObrigado ao público português pelo apoio fantástico.'
```

```
In [63]: Reply = driver.find_element_by_xpath('//div[@data-testid="reply"]').text  
Retweet = driver.find_element_by_xpath('//div[@data-testid="retweet"]').text  
Like = driver.find_element_by_xpath('//div[@data-testid="like"]').text
```

```
In [64]: print(UserTag)  
print(Time_stamp)  
print(Reply)  
print(Retweet)  
print(Like)
```

```
Cristiano Ronaldo  
@Cristiano
```

```
.
```

```
24 Sep
```

```
2022-09-24T21:38:35.000Z
```

```
13.2K
```

```
11.3K
```

```
182.6K
```

```
In [73]: articles = driver.find_elements_by_xpath('//article[@data-testid="tweet"]')
```

```
In [74]: len(articles)
```

```
Out[74]: 4
```



```
In [80]: print(Usertags)
          print(Timestamps)
          print(Tweets)
          print(Replies)
          print(Retweets)
          print(Likes)
```



```
'My 2nd Premier League Player Of The Month Award, the 6th in my career. I'm as happy to win today as I was in my early
days, the hunger for victory and achievements never fades away. Thanks to everyone that made this possible. ', 'My 2nd
Premier League Player Of The Month Award, the 6th in my career. I'm as happy to win today as I was in my early days, th
e hunger for victory and achievements never fades away. Thanks to everyone that made this possible. ', 'My 2nd Premier
League Player Of The Month Award, the 6th in my career. I'm as happy to win today as I was in my early days, the hunger
for victory and achievements never fades away. Thanks to everyone that made this possible. ', 'Always good to be with m
y Bro's ', 'Always good to be with my Bro's ', 'Always good to be with my Bro's ', 'Always good to be with my Bro's ',
'Always good to be with my Bro's ', 'Always good to be with my Bro's ', 'Always good to be with my Bro's ', 'Always goo
d to be with my Bro's ', 'Always good to be with my Bro's ', 'Never stop dreaming.\nWatch my chat with Sir Alex now \n#
MUFC #BringingYouCloser', 'Never stop dreaming.\nWatch my chat with Sir Alex now \n#MUFC #BringingYouCloser', 'Never st
op dreaming.\nWatch my chat with Sir Alex now \n#MUFC #BringingYouCloser', 'Never stop dreaming.\nWatch my chat with Si
r Alex now \n#MUFC #BringingYouCloser', 'Never stop dreaming.\nWatch my chat with Sir Alex now \n#MUFC #BringingYouClos
er', 'Never stop dreaming.\nWatch my chat with Sir Alex now \n#MUFC #BringingYouCloser', 'Never stop dreaming.\nWatch m
y chat with Sir Alex now \n#MUFC #BringingYouCloser', '', '', '', '', '', 'É sempre um orgulho voltar à seleção e repre
sentar o nosso país!', 'É sempre um orgulho voltar à seleção e representar o nosso país!', 'É sempre um orgulho voltar
à seleção e representar o nosso país!']
['13.2K', '13.2K', '13.2K', '13.2K', '13.2K', '13.2K', '13.2K', '13.2K', '6,501', '6,501', '6,501', '6,501', '5,277',
'5,277', '5,277', '5,277', '5,277', '5,277', '5,277', '5,277', '1,225', '1,225', '1,225', '1,225', '1,225',
'1,225', '1,225', '2,965', '2,965', '2,965', '2,965', '2,965', '2,965', '4,356', '4,356', '4,356']
['11.3K', '11.3K', '11.3K', '11.3K', '11.3K', '11.3K', '11.3K', '11.3K', '24.4K', '24.4K', '24.4K', '24.4K', '14.2K',
'14.2K', '14.2K', '14.2K', '14.2K', '14.2K', '14.2K', '14.2K', '4,585', '4,585', '4,585', '4,585', '4,585',
'4,585', '4,585', '4,619', '4,619', '4,619', '4,619', '4,619', '17.3K', '17.3K', '17.3K']
['182.6K', '182.6K', '182.6K', '182.6K', '182.7K', '182.7K', '182.7K', '182.7K', '337.1K', '337.1K', '337.1K', '428.9
K', '428.9K', '428.9K', '428.9K', '428.9K', '428.9K', '428.9K', '428.9K', '428.9K', '59.8K', '59.8K', '59.8K', '59.8K',
'59.8K', '59.8K', '59.8K', '81.8K', '81.8K', '81.8K', '81.8K', '81.8K', '81.8K', '324.8K', '324.8K', '324.8K']
```

```
In [82]: len(set(Tweets))
```

```
Out[82]: 6
```

```
In [83]: import pandas as pd
```

```
In [84]: data = pd.DataFrame(zip(Usertags, Timestamps, Tweets, Replies, Retweets, Likes), columns=[ 'Usertags', 'Timestamps', 'Tweets', 'Re
plies', 'Retweets', 'Likes'])
```

```
In [90]: data = data.drop_duplicates()
```

In [91]: data

Out[91]:

	User	tags	Timestamps	Tweets	Replies	Retweets	Likes
0	Cristiano Ronaldo	@Cristiano	24 Sep 2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...	13.2K	11.3K	182.6K
4	Cristiano Ronaldo	@Cristiano	24 Sep 2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...	13.2K	11.3K	182.7K
8	Cristiano Ronaldo	@Cristiano	30 Apr 2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	6,501	24.4K	337.1K
9	Cristiano Ronaldo	@Cristiano	12 May 2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	6,501	24.4K	337.1K
11	Cristiano Ronaldo	@Cristiano	12 May 2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	6,501	24.4K	428.9K
12	Cristiano Ronaldo	@Cristiano	9 May 2022-05-09T10:32:07.000Z	Always good to be with my Bro's	5,277	14.2K	428.0K
20	Cristiano Ronaldo	@Cristiano	9 May 2022-05-09T10:32:07.000Z	Always good to be with my Bro's	5,277	14.2K	59.8K
21	Cristiano Ronaldo	@Cristiano	18 Dec 2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...	1,225	4,585	59.8K
27	Cristiano Ronaldo	@Cristiano	18 Dec 2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...	1,225	4,585	81.8K
28	Cristiano Ronaldo	@Cristiano	13 Dec 2021-12-13T12:04:26.000Z		2,965	4,619	81.8K
32	Cristiano Ronaldo	@Cristiano	13 Dec 2021-12-13T12:04:26.000Z		2,965	4,619	324.8K
33	Cristiano Ronaldo	@Cristiano	30 Aug 2021-08-30T19:47:56.000Z	É sempre um orgulho voltar à seleção e represe...	4,355	17.3K	324.8K

In [92]: data.to_csv('C:\\Users\\user\\Documents\\datasets\\tweet.csv', index = False)

In []:

DATA CLEANING



```
In [61]: data = pd.read_csv('C:\\Users\\user\\Documents\\datasets\\tweet.csv')
```

```
In [62]: data['Replies'] = data['Replies'].str[:-1]
data['Retweets'] = data['Retweets'].str[:-1]
data['Likes'] = data['Likes'].str[:-1]
```

```
In [63]: data
```

Out[63]:

	Usertags	Timestamps	Tweets	Replies	Retweets	Likes
0	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...	13.2	11.3	182.6
1	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...	13.2	11.3	182.7
2	Cristiano Ronaldo\n@Cristiano\n\n30 Apr	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	6,50	24.4	337.1
3	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	6,50	24.4	337.1
4	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	6,50	24.4	428.9
5	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09T10:32:07.000Z	Always good to be with my Bro's	5,27	14.2	428.9
6	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09T10:32:07.000Z	Always good to be with my Bro's	5,27	14.2	59.8
7	Cristiano Ronaldo\n@Cristiano\n\n18 Dec 2021	2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...	1,22	4,58	59.8
8	Cristiano Ronaldo\n@Cristiano\n\n18 Dec 2021	2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...	1,22	4,58	81.8
9	Cristiano Ronaldo\n@Cristiano\n\n13 Dec 2021	2021-12-13T12:04:26.000Z	NaN	2,96	4,61	81.8
10	Cristiano Ronaldo\n@Cristiano\n\n13 Dec 2021	2021-12-13T12:04:26.000Z	NaN	2,96	4,61	324.8
11	Cristiano Ronaldo\n@Cristiano\n\n30 Aug 2021	2021-08-30T19:47:56.000Z	É sempre um orgulho voltar à seleção e represe...	4,35	17.3	324.8

```
In [64]: data['Replies'] = (pd.to_numeric(data['Replies'].str.replace(',', ''), errors='coerce'))
data['Retweets'] = (pd.to_numeric(data['Retweets'].str.replace(',', ''), errors='coerce'))
data['Likes'] = pd.to_numeric(data['Likes'])
```

```
data['Retweets'] = (data['Retweets'])*1000
data['Likes'] = (data['Likes'])*1000
```

10/9/22, 1:26 PM

In [67]:

data

Out[67]:

	Usertags	Timestamps	Tweets	Replies	Retweets	Likes
0	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...	13200.0	11300.0	182600.0
1	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...	13200.0	11300.0	182700.0
2	Cristiano Ronaldo\n@Cristiano\n\n30 Apr	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	650000.0	24400.0	337100.0
3	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	650000.0	24400.0	337100.0
4	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...	650000.0	24400.0	428900.0
5	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09T10:32:07.000Z	Always good to be with my Bro's	527000.0	14200.0	428900.0
6	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09T10:32:07.000Z	Always good to be with my Bro's	527000.0	14200.0	59800.0
7	Cristiano Ronaldo\n@Cristiano\n\n18 Dec 2021	2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...	122000.0	458000.0	59800.0
8	Cristiano Ronaldo\n@Cristiano\n\n18 Dec	2021-12-				

9	Cristiano Ronaldo\n@Cristiano\n\n13 Dec 2021	2021-12-13T12:04:26.000Z		NaN	296.0	461.0	81.8
10	Cristiano Ronaldo\n@Cristiano\n\n13 Dec 2021	2021-12-13T12:04:26.000Z		NaN	296.0	461.0	324.8
11	Cristiano Ronaldo\n@Cristiano\n\n30 Aug 2021	2021-08-30T19:47:56.000Z	É sempre um orgulho voltar à seleção e represe...		435.0	17.3	324.8

```
In [66]: data['Replies'] = (data['Replies'])*1000
data['Retweets'] = (data['Retweets'])*1000
data['Likes'] = (data['Likes'])*1000
```

```
In [67]: data
```

```
Out[67]:
```

	Usertags	Timestamps		Tweets	Replies	Retweets	Likes
0	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...		13200.0	11300.0	182600.0
1	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24T21:38:35.000Z	Grande jogo, importante vitória equipa!Continu...		13200.0	11300.0	182700.0
2	Cristiano Ronaldo\n@Cristiano\n\n30 Apr	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...		650000.0	24400.0	337100.0
3	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...		650000.0	24400.0	337100.0
4	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12T17:54:32.000Z	My 2nd Premier League Player Of The Month Awar...		650000.0	24400.0	428900.0
5	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09T10:32:07.000Z		Always good to be with my Bro's	527000.0	14200.0	428900.0
6	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09T10:32:07.000Z		Always good to be with my Bro's	527000.0	14200.0	59800.0
7	Cristiano Ronaldo\n@Cristiano\n\n18 Dec 2021	2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...		122000.0	458000.0	59800.0
8	Cristiano Ronaldo\n@Cristiano\n\n18 Dec 2021	2021-12-18T10:39:17.000Z	Never stop dreaming.\nWatch my chat with Sir A...		122000.0	458000.0	81800.0
9	Cristiano Ronaldo\n@Cristiano\n\n13 Dec 2021	2021-12-13T12:04:26.000Z		NaN	296000.0	461000.0	81800.0
10	Cristiano Ronaldo\n@Cristiano\n\n13 Dec 2021	2021-12-13T12:04:26.000Z		NaN	296000.0	461000.0	324800.0
11	Cristiano Ronaldo\n@Cristiano\n\n30 Aug 2021	2021-08-30T19:47:56.000Z	É sempre um orgulho voltar à seleção e represe...		435000.0	17300.0	324800.0

```
In [68]: data['Timestamps'] = pd.to_datetime(data['Timestamps'])
```


1:26 PM

Data Cleaning - Jupyter Notebook

In [69]: data

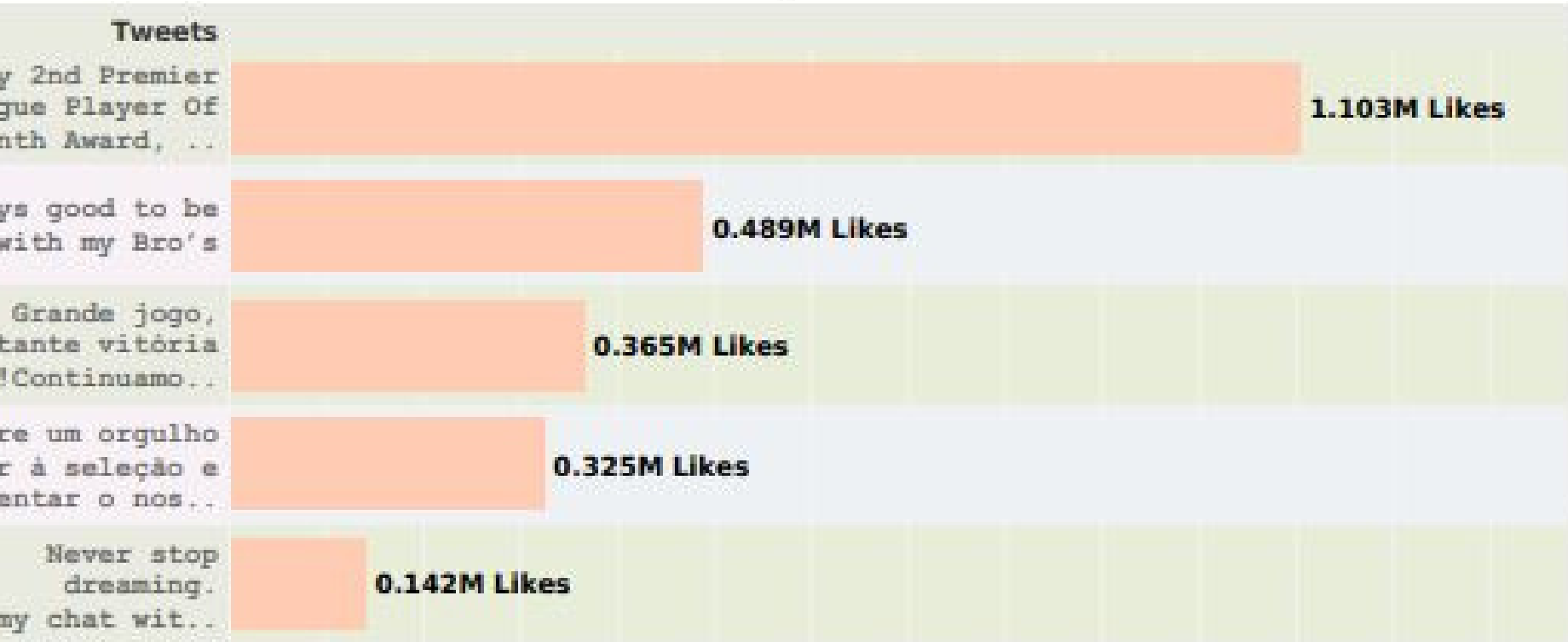
Out[69]:

	Usertags	Timestamps	Tweets	Replies	Retweets	Likes
0	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24 21:38:35+00:00	Grande jogo, importante vitória equipa!Continu...	13200.0	11300.0	182600.0
1	Cristiano Ronaldo\n@Cristiano\n\n24 Sep	2022-09-24 21:38:35+00:00	Grande jogo, importante vitória equipa!Continu...	13200.0	11300.0	182700.0
2	Cristiano Ronaldo\n@Cristiano\n\n30 Apr	2022-05-12 17:54:32+00:00	My 2nd Premier League Player Of The Month Awar...	650000.0	24400.0	337100.0
3	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12 17:54:32+00:00	My 2nd Premier League Player Of The Month Awar...	650000.0	24400.0	337100.0
4	Cristiano Ronaldo\n@Cristiano\n\n12 May	2022-05-12 17:54:32+00:00	My 2nd Premier League Player Of The Month Awar...	650000.0	24400.0	428900.0
5	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09 10:32:07+00:00	Always good to be with my Bro's	527000.0	14200.0	428900.0
6	Cristiano Ronaldo\n@Cristiano\n\n9 May	2022-05-09 10:32:07+00:00	Always good to be with my Bro's	527000.0	14200.0	59800.0
7	Cristiano Ronaldo\n@Cristiano\n\n18 Dec 2021	2021-12-18 10:39:17+00:00	Never stop dreaming.\nWatch my chat with Sir A...	122000.0	458000.0	59800.0
8	Cristiano Ronaldo\n@Cristiano\n\n18 Dec	2021-12-18				

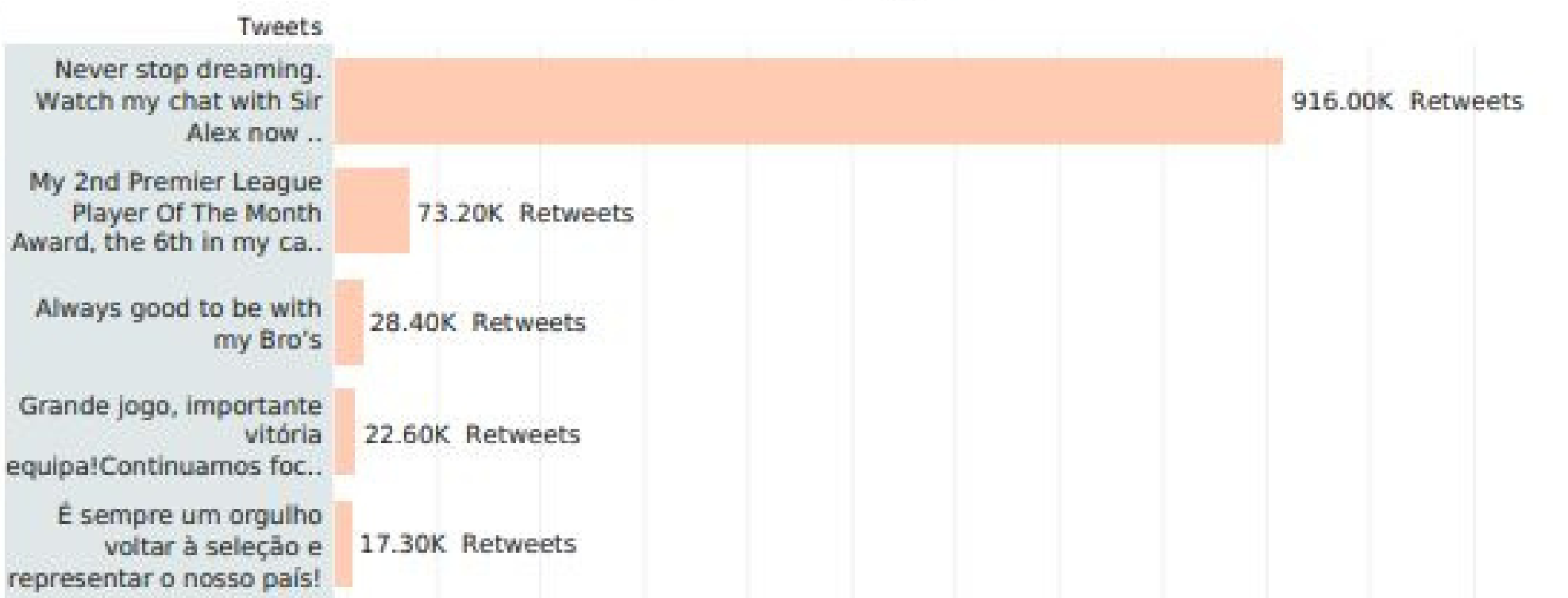
DATA VISUALIZATION



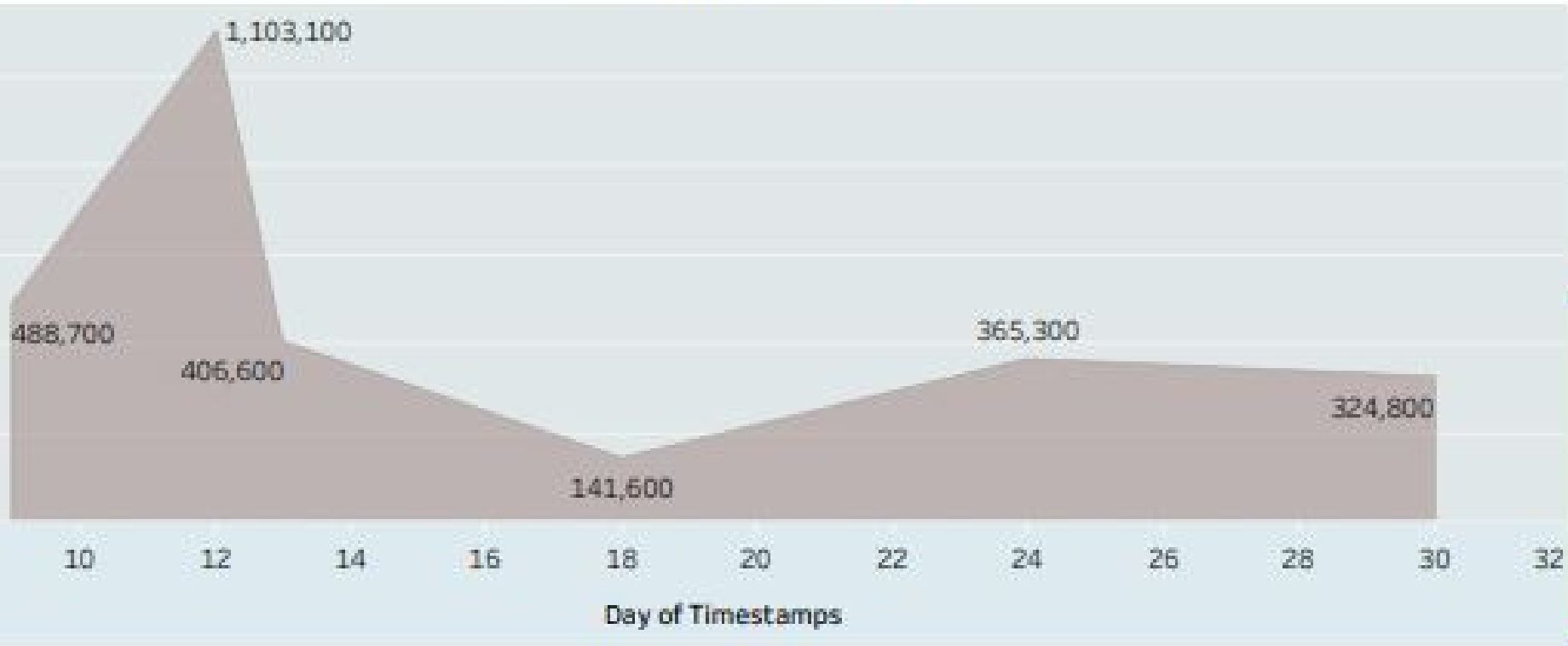
Number of Likes per Tweets



Number of Retweets per Tweet



Likes by Timestamps



Likes

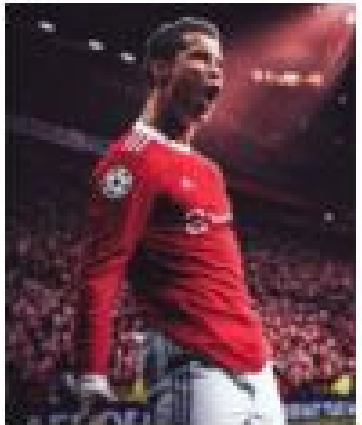
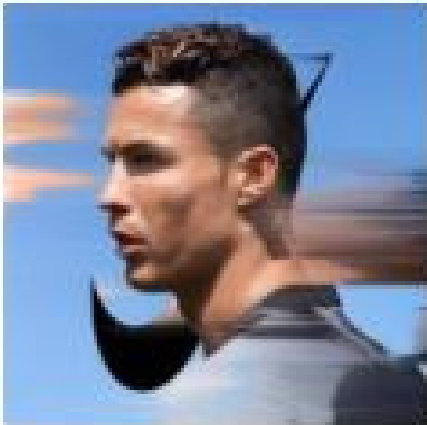
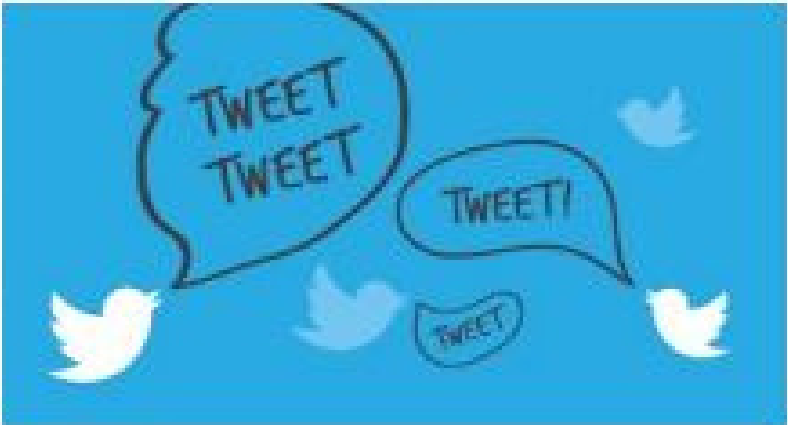
2.8M

Retweets

1.98M

Replies

4.3M



**THANK
YOU FOR
VIEWING**





Oshodi, Lagos



+2348174336178



<https://popseyynn.github.io/>



popseyynn@gmail.com