DATA VISUALIZATION REPORT   
LAB 01: DATA RELATIONSHIP

GROUP 07

20127449 – Tran Quoc Bao

20127452 – Ho Dang Cao

20127476 – Do Duc Duy

20127567 – Le Nguyen Binh Nam

Instructors:

Dr. Bui Tien Len

Dr. Le Ngoc Thanh

Contents

[EVALUATIONS 3](#_Toc129989296)

[LEVEL OF COMPLETION OF REQUIREMENTS 3](#_Toc129989297)

[LEVEL OF COMPLETION OF MEMBERS 3](#_Toc129989298)

[DETAILED ALGORITHM 4](#_Toc129989299)

[DATA CRAWLING AND REPROCESSING 4](#_Toc129989300)

[DATA CRAWLING 4](#_Toc129989301)

[PRE-PROCESSING 6](#_Toc129989302)

[Store data and reuse 7](#_Toc129989303)

[VISUALIZATION 8](#_Toc129989304)

[I. Scatter Charts 8](#_Toc129989305)

[II. Radar Charts 10](#_Toc129989306)

[III. Percent Bar Charts 10](#_Toc129989307)

[IV. Donut Charts 12](#_Toc129989308)

[References 13](#_Toc129989309)

# EVALUATIONS

## LEVEL OF COMPLETION OF REQUIREMENTS

|  |  |  |  |
| --- | --- | --- | --- |
| ID | TASK | MEMBER | LEVEL |
| 1 | data crawling, pre-processing | 20127452 | 100% |
| 2 | scatter chart - correlation between 2 attributes | 20127452 | 100% |
| 3 | radar chart - attributes comparison between continents | 20127452 | 100% |
| 4 | % bar chart - % of attributes of each continent | 20127452 | 100% |
| 5 | donut chart - % cases contribution of continents | 20127452 | 100% |
|  |  |  | 100% |
|  |  |  | 100% |
|  |  |  | 100% |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## LEVEL OF COMPLETION OF MEMBERS

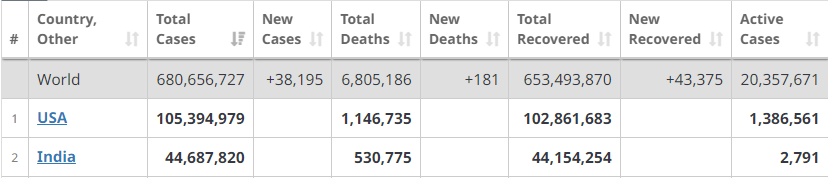
|  |  |  |
| --- | --- | --- |
| ID | FULL NAME | LEVEL |
| 20127449 | TRAN QUOC BAO | 100% |
| 20127452 | HO DANG CAO | 100% |
| 20127476 | DO DUC DUY | 100% |
| 20127567 | LE NGUYEN BINH NAM | 100% |

# DETAILED ALGORITHM

## DATA CRAWLING AND REPROCESSING

### DATA CRAWLING

In this section, we will talk about the process of getting data from the data table in the Worldometer website ([www.worldometers.info/coronavirus/](http://www.worldometers.info/coronavirus/)).



The algorithm will use 3 main libraries/modules, specifically:

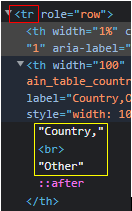
- Use selenium's webdriver function to access and get the HTML code of the website.

- Use BeautifulSoup to extract data by tag from the obtained HTML code.

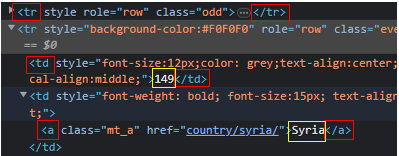
- Use re to extract the names of the data's attributes.

Observe that the data is all in the 'tr' tags. Each 'tr' tag corresponds to 1 object (1 row) in the data table. In there:

- The value of each attribute name cell is in the 'th' tag, but there is a 'br' tag in the middle.



- The value of each data cell is in the 'td' and 'a' tags.



We'll ignore hidden rows - lines with style="display: none" attribute and we take only the first 233 rows, which corresponds to the row number of the data table.



After getting the html code containing the rows in the data table, we divide it into two parts for different processing: the attribute name row and the value table.

* Attribute row:
  + Use re.findall with the syntax: >([\w\d,/ ]\*)< to get a list containing the property names as follows:



* + Due to the 'br' tag, attribute names are split into contiguous elements. Then concatenate adjacent (other than ' ') elements to get the final list:

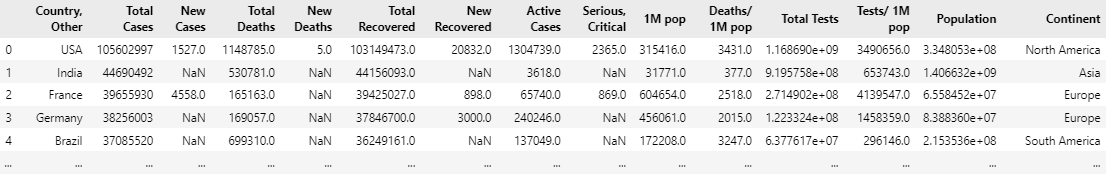


* Value table:
  + We filter out the cell data on each row by finding all the 'td' and 'a' tags.
  + Get the last data on each row using the get\_string function.
  + Illustrated problems to be handled on each row:



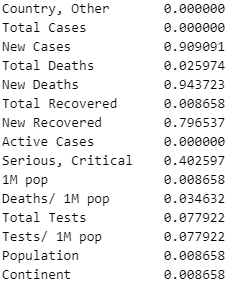
* + The get\_string function includes the following steps:
    - remove id value
    - remove redundant empty value
    - remove duplicated country name
    - get pure data by string attribute of Selenium
    - Replace ' ' (space characters) by nan values.
    - Astype numeric values from string to integer.

The crawled data is put into a data frame:

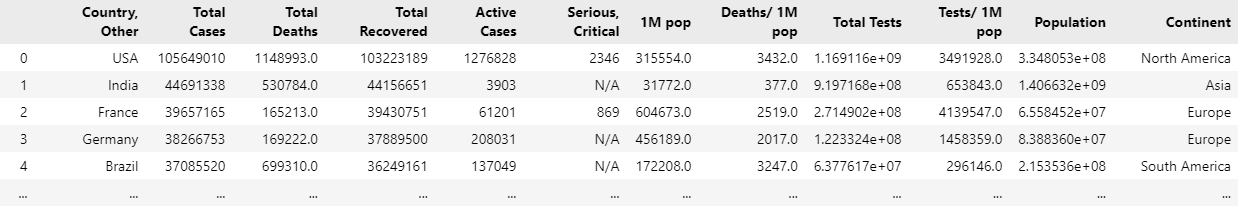
****

### PRE-PROCESSING

Observe missing ratio of attributes:



**Comment:** Because the covid situation is no longer serious, the attributes such as New Cases, New Deaths, New Recovered are not updated regularly (have high missing ratio). They wil be dropped.



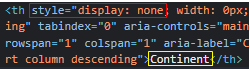
### Store data and reuse

The data frame is saved as .csv file daily.

Read the file to visualize later.

**Note that**: our data frame has 231 rows × (14 + 1) columns.

The extra column comes from the Continent attribute, which is hidden in the website.

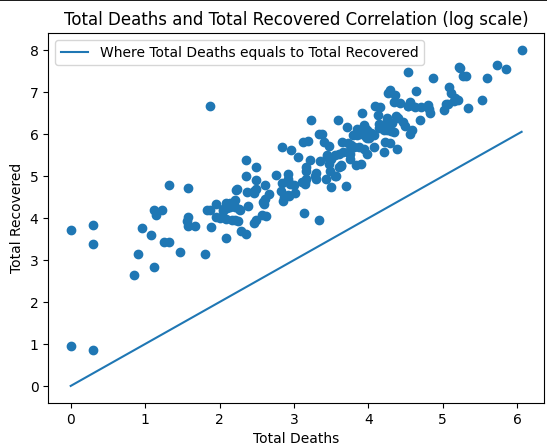


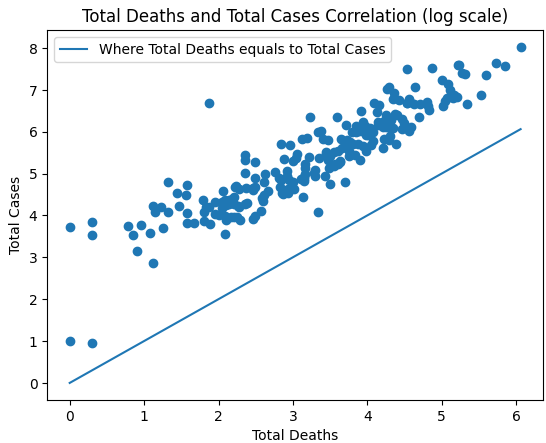
## VISUALIZATION

### Scatter Charts

We built a function to display the correlation between 2 given properties. Also, visualize the mean line where the points have x coordinate equal to y coordinate to compare the magnitude of the 2 attributes.

#### Charts





#### Comments

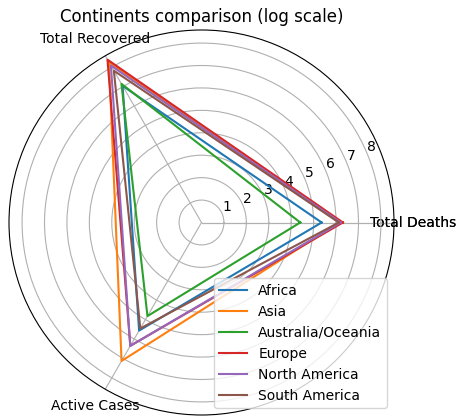
All countries have more recovered cases than death cases.

Recovered cases and death cases have positive linear correlation.

no difference with correlation between Recovered cases and Death cases

### Radar Charts

#### Charts



#### Comments

Recovered rate is high in all continents.

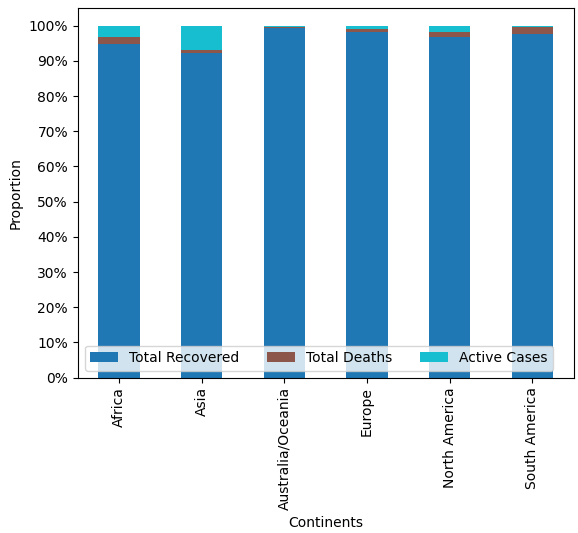
Asia has highest total cases but the number of recovered and deaths is near to America and Europe.

-

### Percent Bar Charts

%

#### Charts



#### Comments

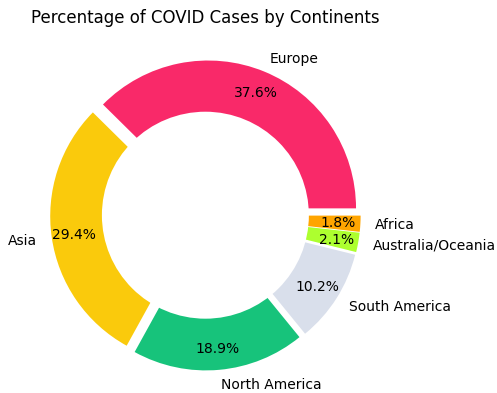
Asia has the highest active cases.

Australia/Oceania seems to be dealing with Covid epidemic most effectively.

Meanwhile, Africa and South America have the highest Deaths ratio.

### Donut Charts

#### Charts



#### Comments

Europe has the highest total cases.

Meanwhile, Africa has weak medical background, but the number of total cases of Africa is the smallest one.

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.