HCM City University of Science
Faculty of Information Technology
Ho Chi Minh City, 12/9/2021

INTRODUCTION TO DATA SCIENCE

Report Exercise 2

Author: Nguyen Dang Khoa
ID: 19127177
Class: 19KHDL

Purpose:

The report provides insights about covid data, including current trend, abnormalities in data, relationship between data attributes,.... From the studied knowledge, we can have the first basic view of the data so we can figure out the difficulty of the problems we will be about to face before building our own models or making decisions.

Data collection:

The collected data are provided by Worldometers (a reference website that provides counters and real-time statistics for diverse topics).
Since Worldometers only provides data for the most 3 recent days so the data should be collected every day.

Limitations:

Since Worldometers only provides data for the most 3 recent days, there should be some limitations while collecting data. The script for gathering data must be manually run by the owner, so there could be days when the owner forgot to execute the script.
Moreover, according to Worldometer methods of collecting data, their sources of information also come from Official Websites of Ministries of Health or other Government Institutions and Government authorities' social media accounts. As a result, the data must be translated and processed by hand before displaying on their website.
For those reasons, there could be flaws in collected data and might affect the insight to be introduced in this report.

Completeness:

No	Requirement	Completed
1	Source code	100%
2	Data (up to 2021-12-09)	100%
3	Report	100%

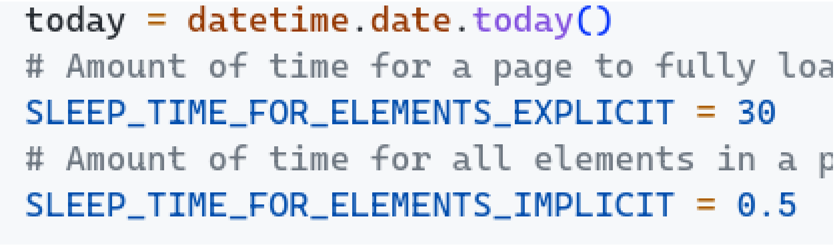
Table of contents

Methods explanation:

crawler.ipynb:

Selenium and HTML parsing are used for crawling data.

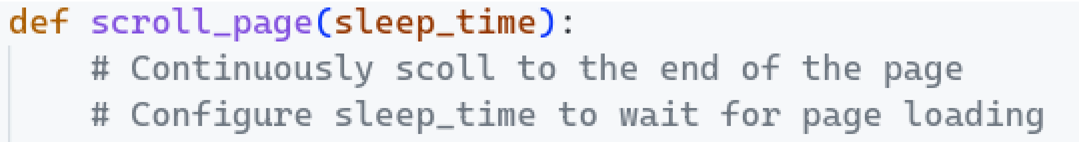
Firstly, the script gets today's date and creates some variables that act as timeouts for web and web elements.



```
today = datetime.date.today()
# Amount of time for a page to fully loa
SLEEP_TIME_FOR_ELEMENTS_EXPLICIT = 30
# Amount of time for all elements in a p
SLEEP_TIME_FOR_ELEMENTS_IMPLICIT = 0.5
```

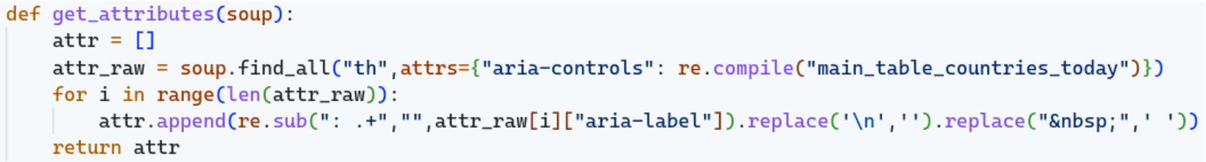
Then, there are some defined functions:

- scroll_page: this function is used to scroll a page in the browser via selenium. It will scroll to the end of the page and use the timeout variable to wait for the page to load. Then it compares the new height of the page with the previous height to decide to scroll.



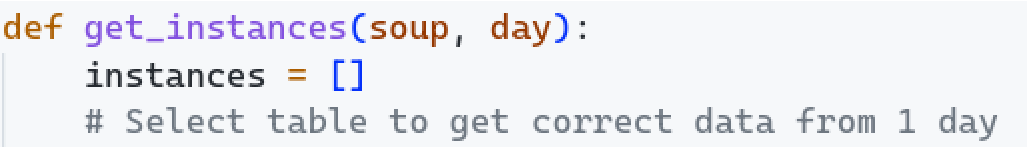
```
def scroll_page(sleep_time):
    # Continuously scoll to the end of the page
    # Configure sleep_time to wait for page loading
```

- get_attributes: this function is used to collect the attributes (columns) which are specified by the website.



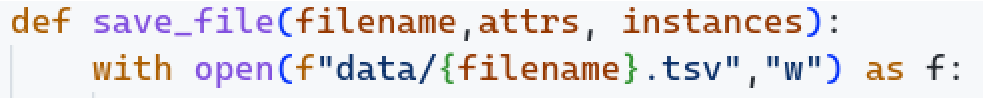
```
def get_attributes(soup):
    attr = []
    attr_raw = soup.find_all("th",attrs={"aria-controls": re.compile("main_table_countries_today")})
    for i in range(len(attr_raw)):
        attr.append(re.sub(": .+","",attr_raw[i]["aria-label"]).replace('\n','').replace("&nbsp;",' '))
    return attr
```

- get_instances: the function is designed to crawl main data on the website. While collecting, the data is transformed to a suitable format so as to work well with explore.ipynb.



```
def get_instances(soup, day):
    instances = []
    # Select table to get correct data from 1 day
```

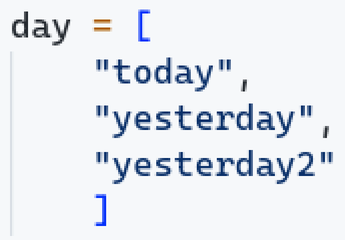
- save_file: this function will plug attributes to the data and save them into tsv files.



```
def save_file(filename,attrs, instances):
    with open(f"data/{filename}.tsv","w") as f:
```

The collecting procedures:

- Firstly, create a list that contains the tag of data table which should be collected



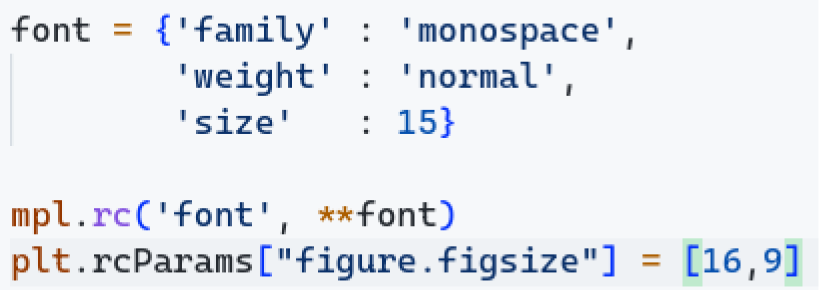
```
day = [
    "today",
    "yesterday",
    "yesterday2"
    ]
```

- Then, the script accesses the Worldometer website and scroll to the end of the page
- Get the attributes
- The next for loop is used to iterate the list “day” while collecting and saving data.

Finally, the data are saved in “data” folder that is located in the same folder as the script

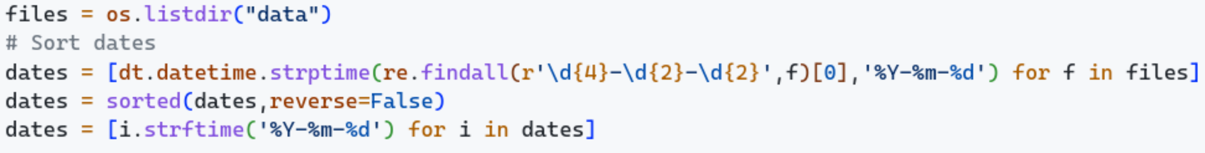
explore.ipynb:

Firstly, the script setup matplotlib for plotting figures by changing font, font size, weight of the font, size of the figure



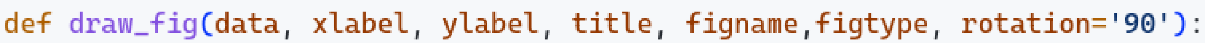
```
font = {'family' : 'monospace',
        'weight' : 'normal',
        'size'   : 15}

mpl.rc('font', **font)
plt.rcParams["figure.figsize"] = [16,9]
```

The load the data from the “data” folder, the data’s titles also have dates on them so the script extracts the date of the data and saves them in a list. Then, the script sorts the data by date for easy accessing.



```
files = os.listdir("data")
# Sort dates
dates = [dt.datetime.strptime(re.findall(r'\d{4}-\d{2}-\d{2}',f)[0],'%Y-%m-%d') for f in files]
dates = sorted(dates,reverse=False)
dates = [i.strftime('%Y-%m-%d') for i in dates]
```

There is a function that is designed for plotting various type of graphs name “draw_fig”:



```
def draw_fig(data, xlabel, ylabel, title, figname,figtype, rotation='90'):
```

The function parameters are:

- data: list of data that should be arranged based on type of graphs to be plotted
- xlabel, ylabel,title: basic arguments for matplotlib

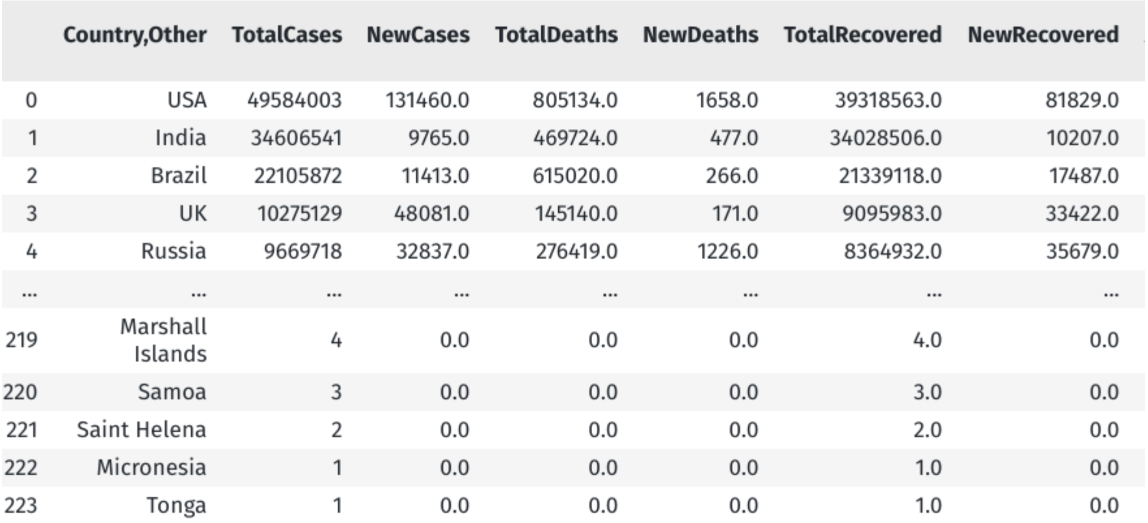
- figname: names of the figures to be saved
- figtype: type of graph to be plotted
- rotation: how items on the x axis should be rotated

This function supports 4 arguments for figtype: “plot”, “hist”, “bar” and “scatter”
Figures are saved in the folder named “figures” which is located in the same folder with the script.

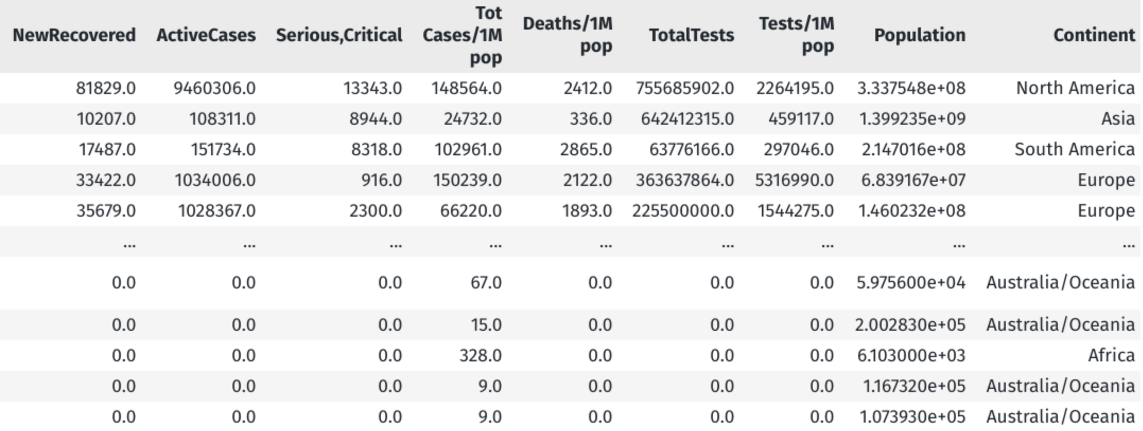
Data cleaning:

There are some instances that have missing values, and the number of instances is 224 which is rather small. So to clean up the data, instead of removing instances with missing values, filling missing values with 0 should be a suitable solution. Exceptionally, for the Continent attribute, ‘Other’ is used to replace missing values becauses this attribute is a nominal attribute.
The purpose is for graph drawing, since the library is not designed to handle nan values.

Data explanation:



	Country,Other	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered
0	USA	49584003	131460.0	805134.0	1658.0	39318563.0	81829.0
1	India	34606541	9765.0	469724.0	477.0	34028506.0	10207.0
2	Brazil	22105872	11413.0	615020.0	266.0	21339118.0	17487.0
3	UK	10275129	48081.0	145140.0	171.0	9095983.0	33422.0
4	Russia	9669718	32837.0	276419.0	1226.0	8364932.0	35679.0
...	...	...	...	...	...	...	...
219	Marshall Islands	4	0.0	0.0	0.0	4.0	0.0
220	Samoa	3	0.0	0.0	0.0	3.0	0.0
221	Saint Helena	2	0.0	0.0	0.0	2.0	0.0
222	Micronesia	1	0.0	0.0	0.0	1.0	0.0
223	Tonga	1	0.0	0.0	0.0	1.0	0.0



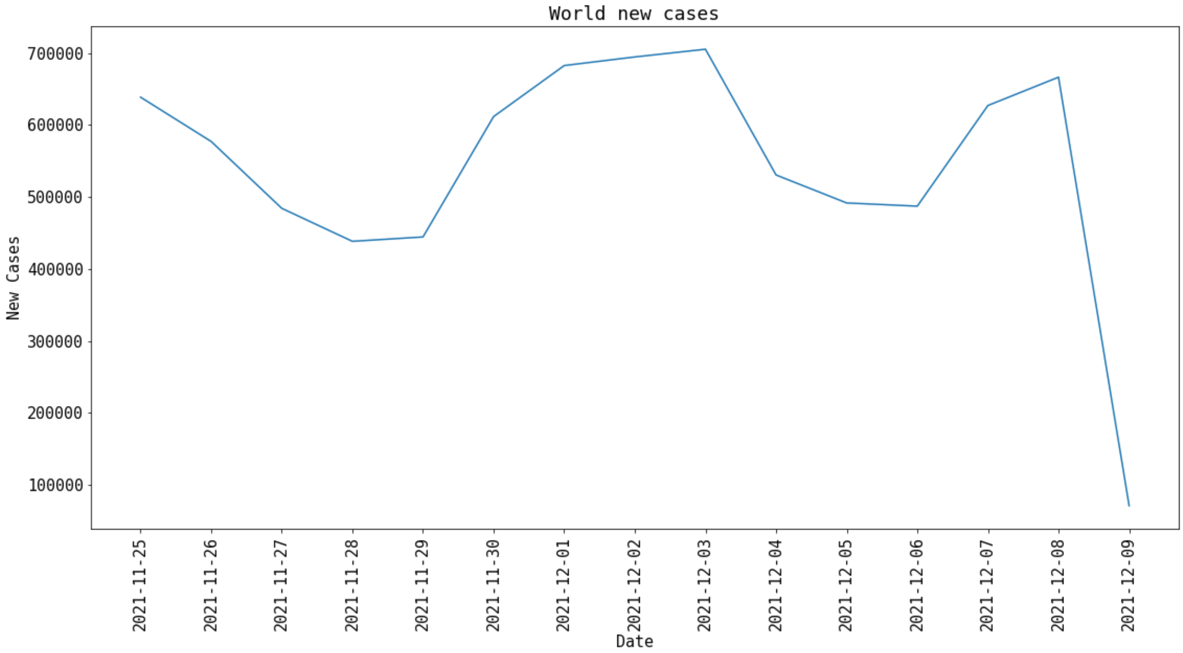
NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	TotalTests	Tests/1M pop	Population	Continent
81829.0	9460306.0	13343.0	148564.0	2412.0	755685902.0	2264195.0	3.337548e+08	North America
10207.0	108311.0	8944.0	24732.0	336.0	642412315.0	459117.0	1.399235e+09	Asia
17487.0	151734.0	8318.0	102961.0	2865.0	63776166.0	297046.0	2.147016e+08	South America
33422.0	1034006.0	916.0	150239.0	2122.0	363637864.0	5316990.0	6.839167e+07	Europe
35679.0	1028367.0	2300.0	66220.0	1893.0	225500000.0	1544275.0	1.460232e+08	Europe
...	...	...	...	...	...	...	...	...
0.0	0.0	0.0	67.0	0.0	0.0	0.0	5.975600e+04	Australia/Oceania
0.0	0.0	0.0	15.0	0.0	0.0	0.0	2.002830e+05	Australia/Oceania
0.0	0.0	0.0	328.0	0.0	0.0	0.0	6.103000e+03	Africa
0.0	0.0	0.0	9.0	0.0	0.0	0.0	1.167320e+05	Australia/Oceania
0.0	0.0	0.0	9.0	0.0	0.0	0.0	1.073930e+05	Australia/Oceania

The data includes 224 instances and 15 attributes. Each instance represents data of a country. So there are 224 countries whose data are collected with the attributes:

- Country,Other: the names of official and non-official countries.
- TotalCases: number of cases in each country including deaths, recovered and active cases.
- NewCases: number of new cases in that current day
- TotalDeaths: number of total death cases
- NewDeaths: number of new death cases
- TotalRecovered: number of total recovered cases. However, according to Worldometer about the **recovery** attributes: *this statistic is highly imperfect, https://www.worldometers.info/coronavirus/about/because reporting can be missing, incomplete, incorrect, based on different definitions, or dated (or a combination of all of these) for many governments, both at the local and national level, sometimes with differences between states within the same country or counties within the same state.*
- ActiveCases: number of cases are currently under treatment
- Serious,Critical: number of active cases which are considered “severe”. According to Worldometers about **this** attribute: the concept of serious cases are defined by each country. Mostly, serious cases are cases are under ICU treatments
- Tot Cases/1M pop: number of cases per 1 million people
- Deaths/1M pop: number of deaths cases per 1 million people
- TotalTests: number of tests conducted
- Tests/1M pop: number of tests conducted per 1 million people
- Population: population of each country
- Continent: which continent the country belongs to

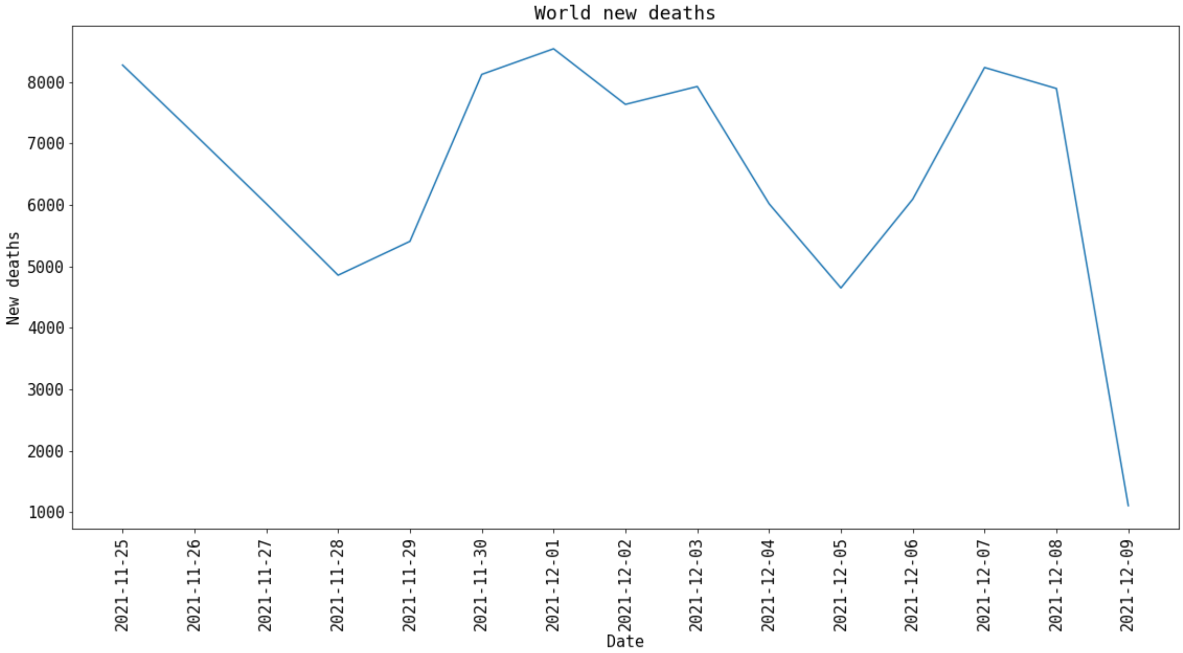
Observation on all collected days:

1. New cases:



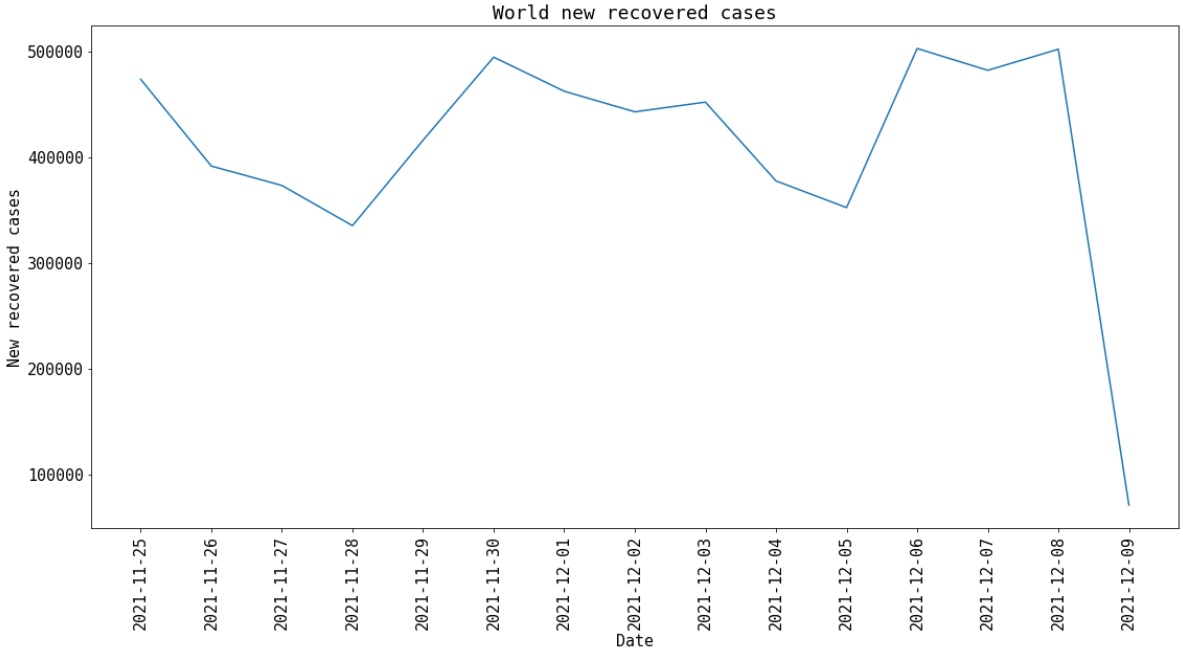
On 2021-12-09, the number of new cases seems to be the lowest since this day is the day the report is written and the data is not fully updated for the rest of the day.
From the graph, the number of new cases decreased from around 64000 to approximately 42000 from 2021-11-25 to 2021-11-28.
The number of cases thrive significantly in the first week of December (2021-11-29 to 2021 12-05) and reach a maximum value of 70000 on 2021-11-03.
More interesting, the number of new cases on the weekends seems to be notably low. Especially on 27 - 28 November and 04 - 05 December, the number of new cases is around 41000 and 47000, respectively.
One of the theories can be used to describe the abnormalities in the shape of the graphs is weekend breaks. As mentioned before, the data are also collected by hand by Worldometer staff and the staff usually have weekend breaks as well as the Governments' media team.
In general, the number of new cases is slowly rising in December, and seems to reach over 72000 cases per day in the next few days.

2. New deaths



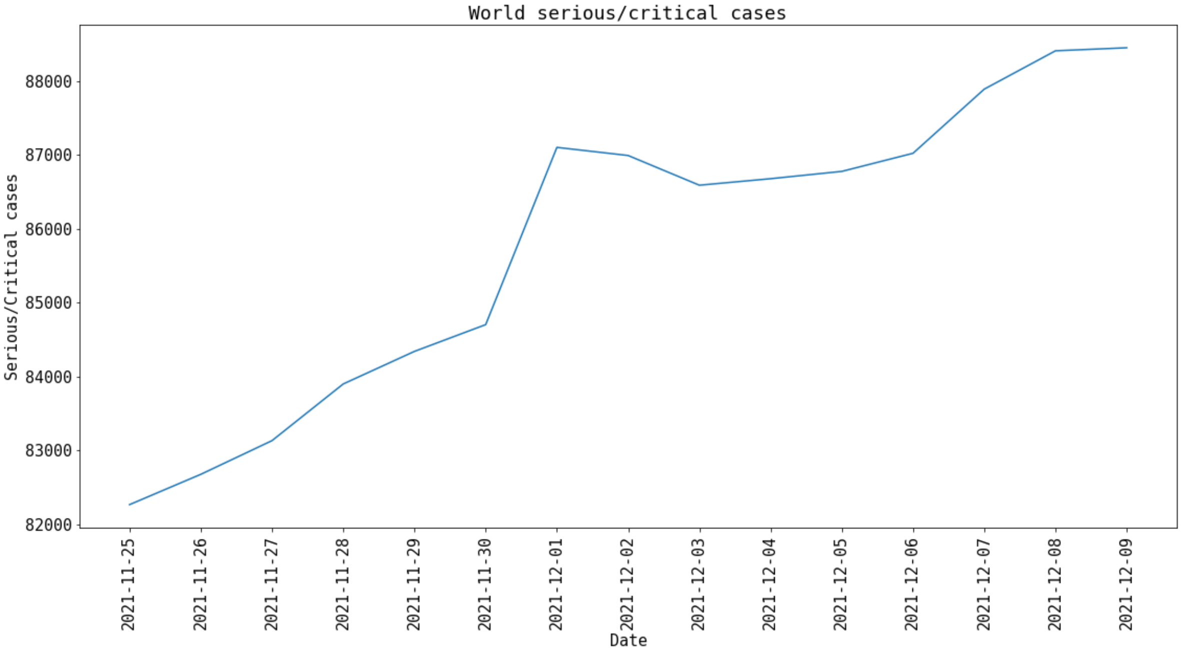
The NewDeaths graphs and NewCases graphs seem to have a similar shape. By the way, the NewDeaths graph cannot tell us much about the trend of new deaths. The abnormalities in the graph's shape can be also explained like the NewCases graph.

3. New recovered cases:



Again, the graph shows no trend of the data. From the data explanation part of the report, we can assume that the data is unreliable, and thus they are mostly useless for analyzing. However, it has a similar shape comparing to NewCases, the knowledge in NewCases graph can also be applied to this graph

4. Serious/Critical cases:

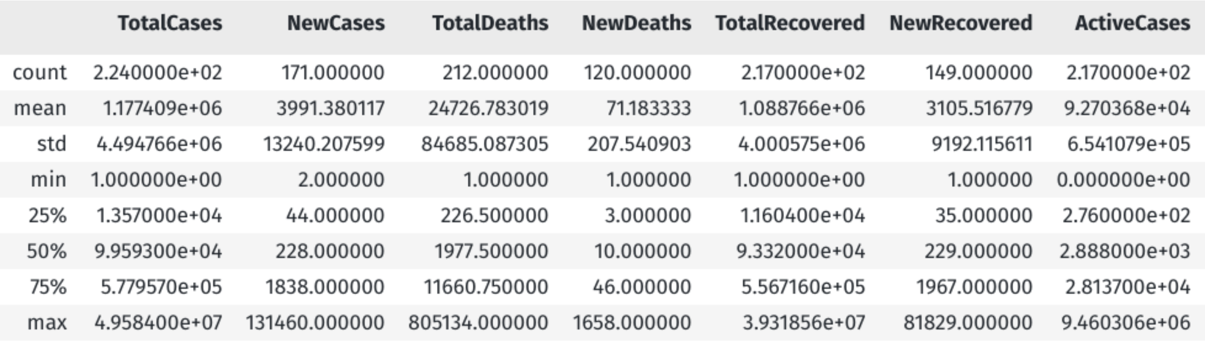


The number of serious cases follows an increasing trend. The number of serious cases rises dramatically from 82000 to 90000 within 15 days. Especially, from 30 November to 1 December, the number of reported serious cases increases from about 85000 to 87000 within a day.

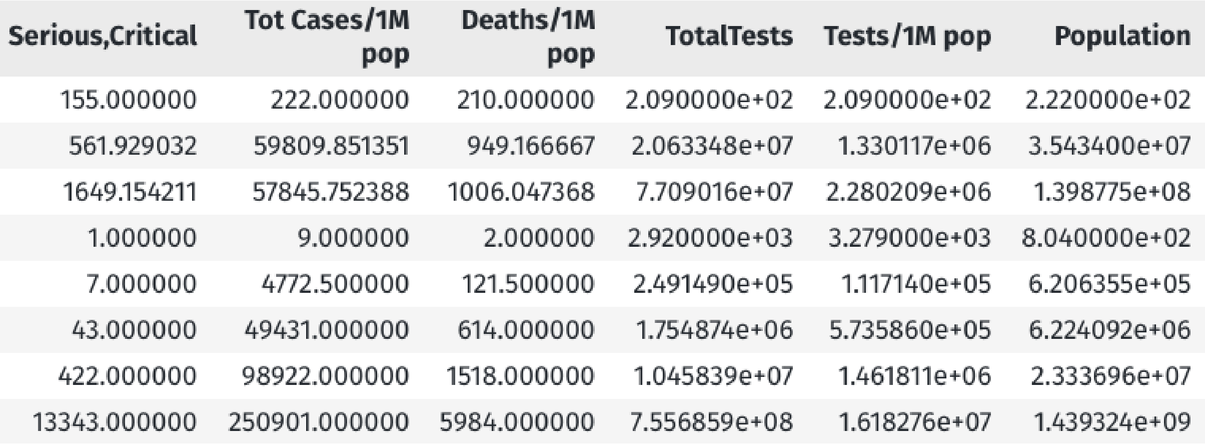
The reason why the number of serious cases increases could be caused by the weather. The cold weather also plays a significant role in worsening the covid 19 symptoms. Moreover, this number will increase even more since more events, occasions, and festivals will be held by the end of the year. Therefore, more and more people gather in one place, and thus spread the disease easily.

Observation on December 1:

Description



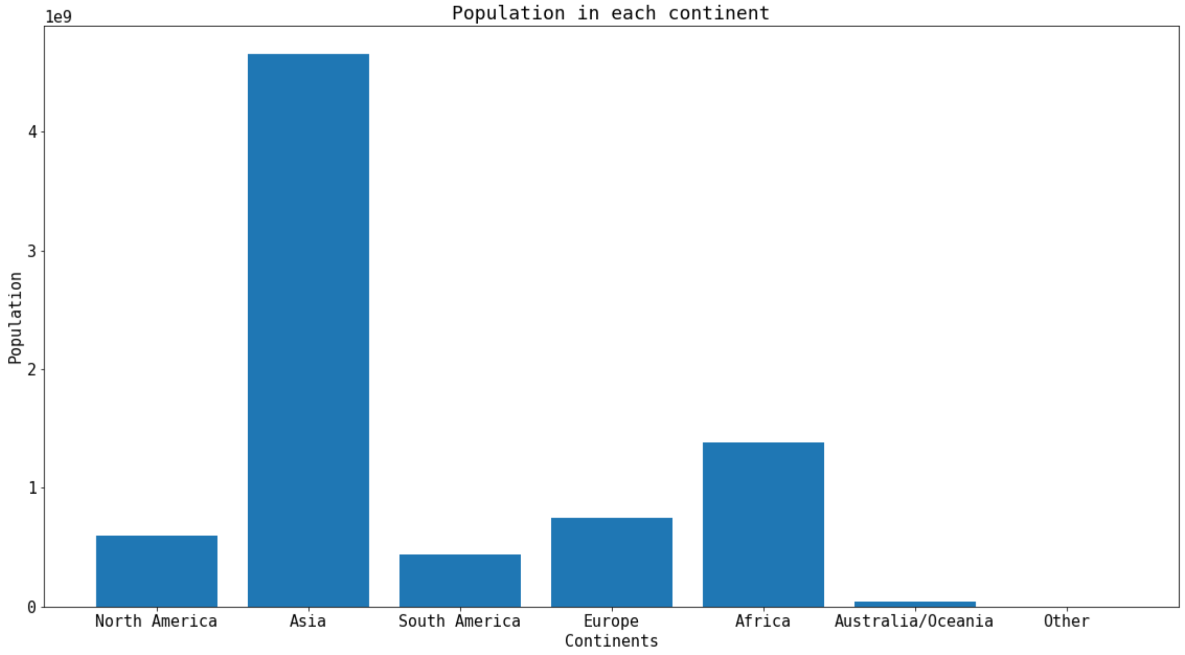
	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases
count	2.240000e+02	171.000000	212.000000	120.000000	2.170000e+02	149.000000	2.170000e+02
mean	1.177409e+06	3991.380117	24726.783019	71.183333	1.088766e+06	3105.516779	9.270368e+04
std	4.494766e+06	13240.207599	84685.087305	207.540903	4.000575e+06	9192.115611	6.541079e+05
min	1.000000e+00	2.000000	1.000000	1.000000	1.000000e+00	1.000000	0.000000e+00
25%	1.357000e+04	44.000000	226.500000	3.000000	1.160400e+04	35.000000	2.760000e+02
50%	9.959300e+04	228.000000	1977.500000	10.000000	9.332000e+04	229.000000	2.888000e+03
75%	5.779570e+05	1838.000000	11660.750000	46.000000	5.567160e+05	1967.000000	2.813700e+04
max	4.958400e+07	131460.000000	805134.000000	1658.000000	3.931856e+07	81829.000000	9.460306e+06

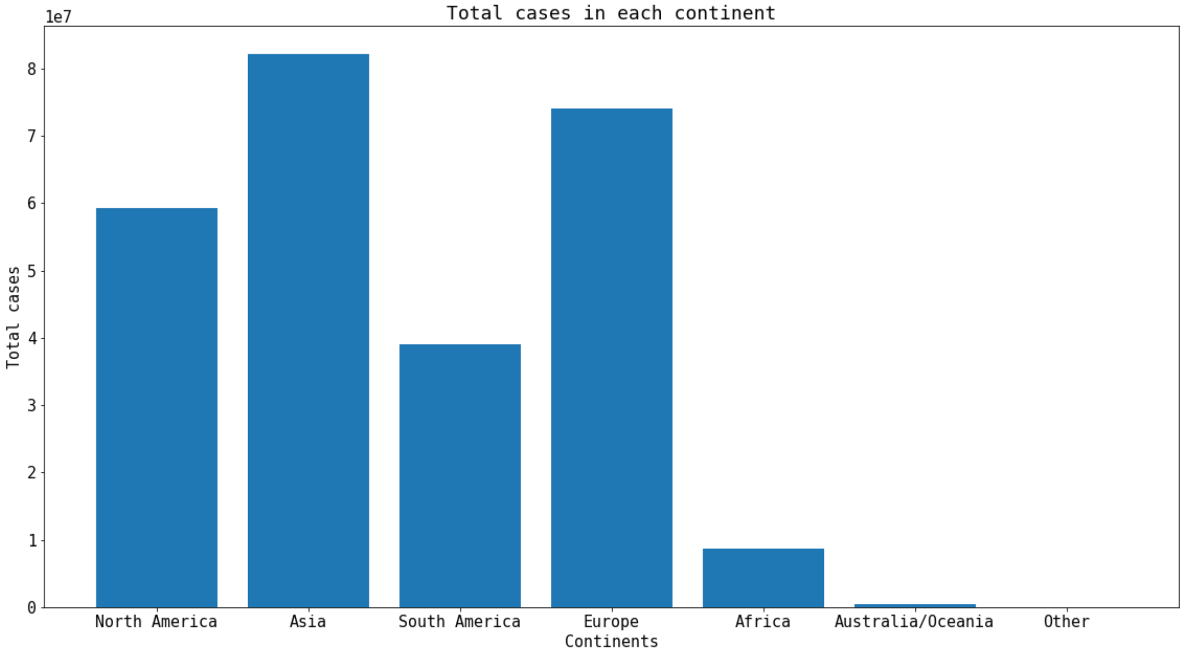


Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	TotalTests	Tests/1M pop	Population
155.000000	222.000000	210.000000	2.090000e+02	2.090000e+02	2.220000e+02
561.929032	59809.851351	949.166667	2.063348e+07	1.330117e+06	3.543400e+07
1649.154211	57845.752388	1006.047368	7.709016e+07	2.280209e+06	1.398775e+08
1.000000	9.000000	2.000000	2.920000e+03	3.279000e+03	8.040000e+02
7.000000	4772.500000	121.500000	2.491490e+05	1.117140e+05	6.206355e+05
43.000000	49431.000000	614.000000	1.754874e+06	5.735860e+05	6.224092e+06
422.000000	98922.000000	1518.000000	1.045839e+07	1.461811e+06	2.333696e+07
13343.000000	250901.000000	5984.000000	7.556859e+08	1.618276e+07	1.439324e+09

The instances in the data frame are sorted by TotalCases descendingly and there seems to be no problems with the data.

1. Total cases in each continent:



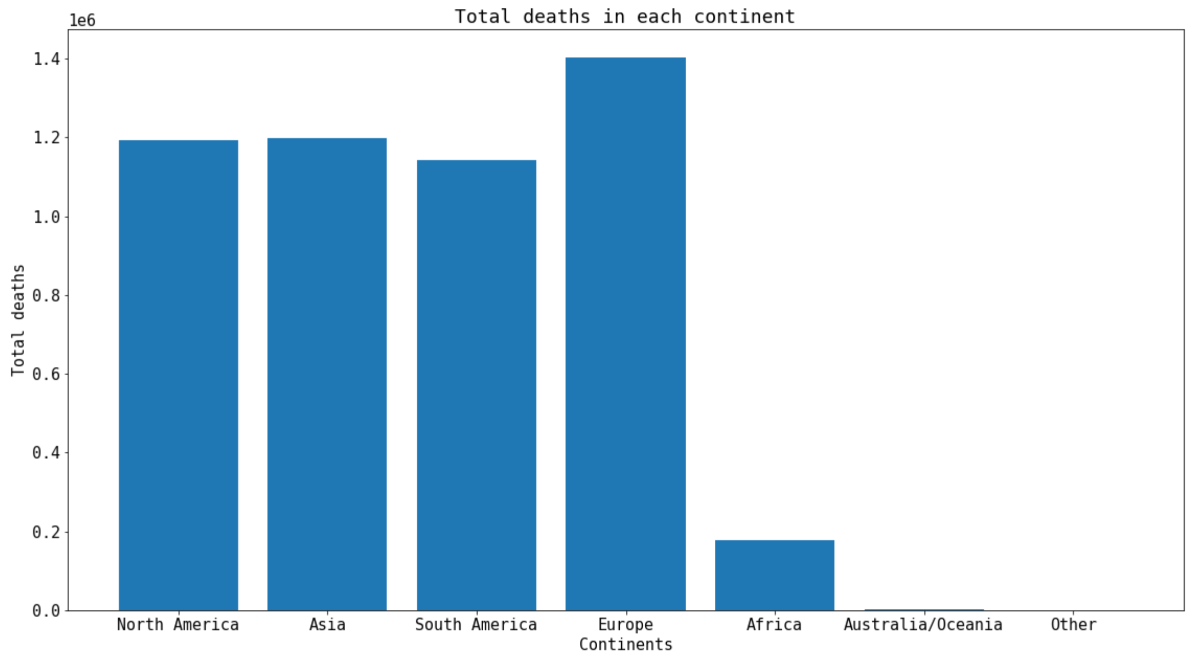


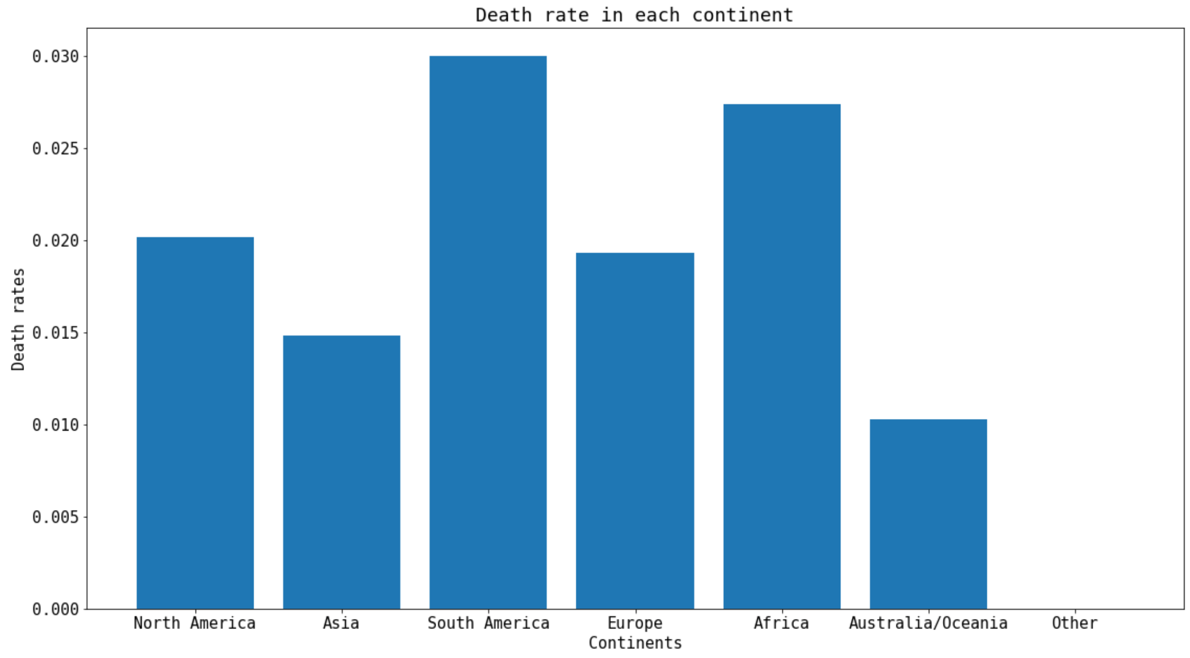
North America, South America, Europe are continents which have the percentage of total covid cases to their population that are much higher than Asia whose population is the highest in the world.

Australia/Oceania has the lowest number of cases since Oceania has only one country which is Australia.

2. Deaths, Active cases, Serious cases in each continent:

Deaths



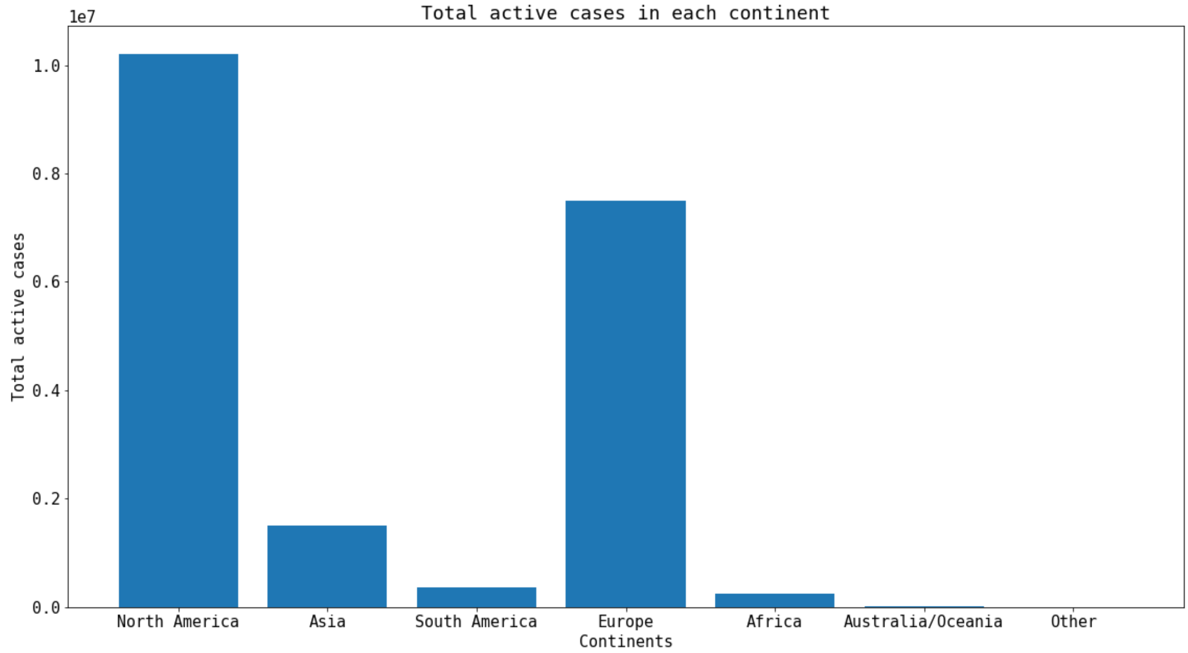


The number of reported deaths in Europe reaches 1.4 million cases by the date of December 1. Meanwhile, despite the huge gap between North America, South Africa and Asia population, the number of deaths in these continents is approximately the same.

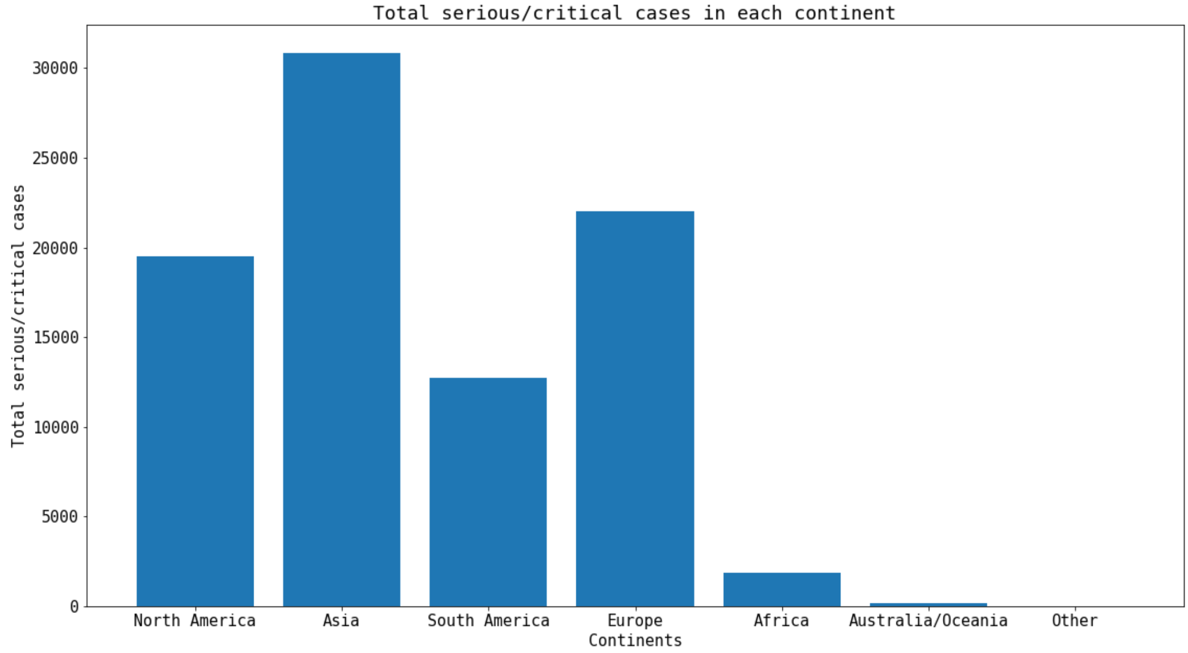
Furthermore, the death rate in South America is nearly 0.030 (30 deaths/1000 people) which made South Africa become the continent with the highest death rate.

The second continent with the high death rate is Africa whose population is around 600 million people.

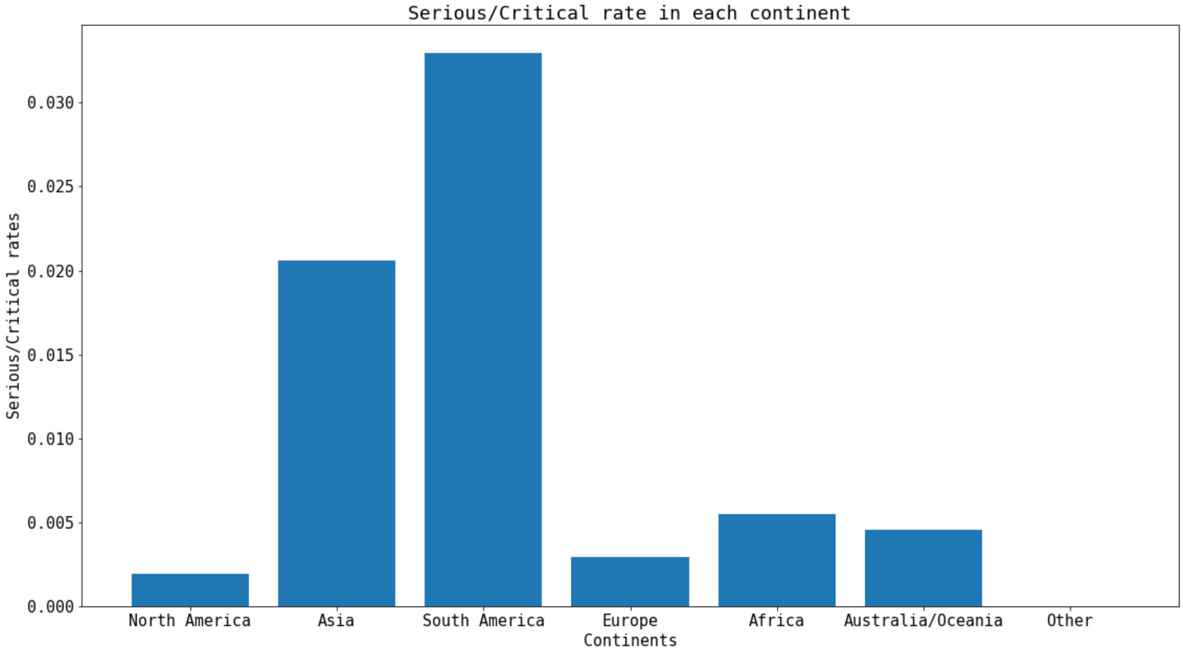
Active cases



Number of active cases in North America is 10 million cases which is surprisingly high according to reports on December 1. Secondly, the number of cases in Europe is approximately 7 million cases on December 1. Therefore, there is a huge gap in the number of active cases between North America, Europe and other continents.



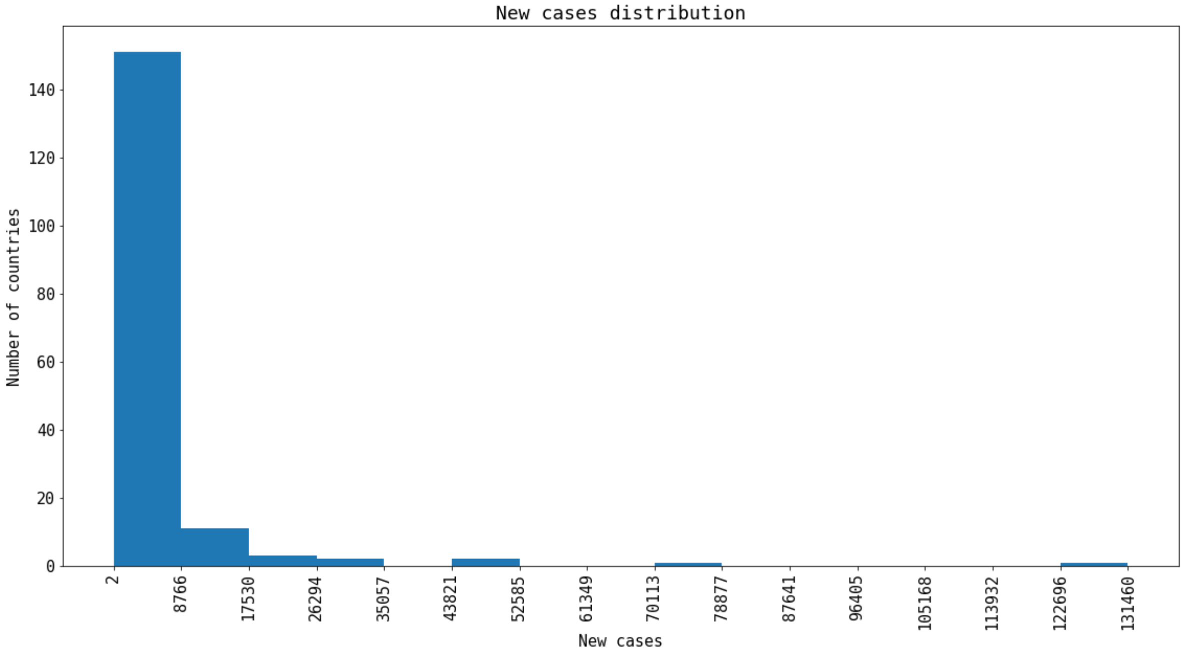
Despite having a large number of active cases, the number of serious cases is pretty low which is roughly 20000 cases. However, the number of active cases in Asia is pretty low but the number of serious cases in this continent is much higher than the rest of the world, which is approximately 30000 cases.

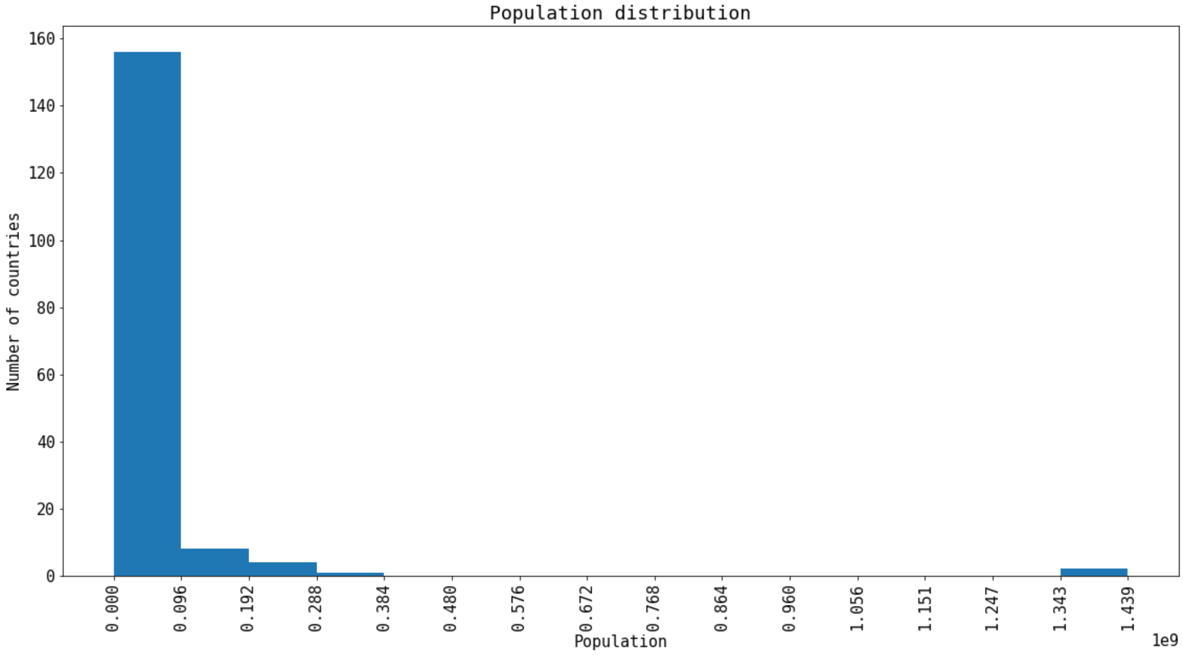


The serious case rate in South America is almost 0.050 (50 serious cases/1000 cases) which tells that South America has the highest serious case rate compared to other continents even though the number of active cases is rather small . In contrast, in North America, the serious case rate is surprisingly small in spite of having the highest number of active cases.

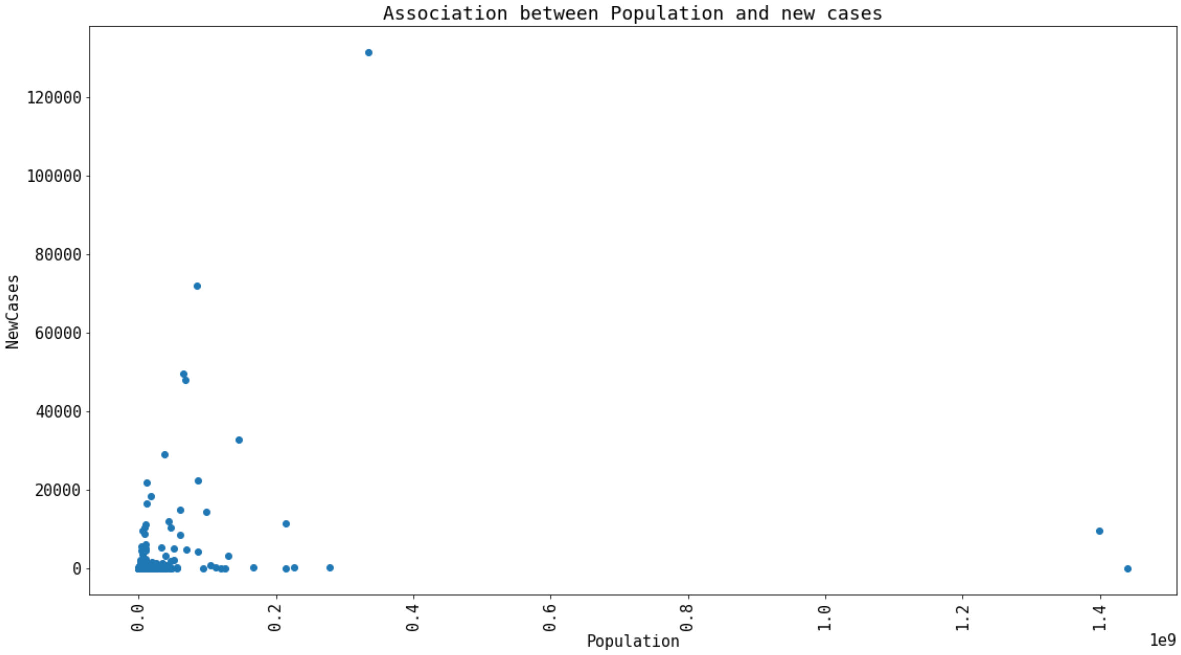
3. Association between attributes and their distribution:

Population and NewCases

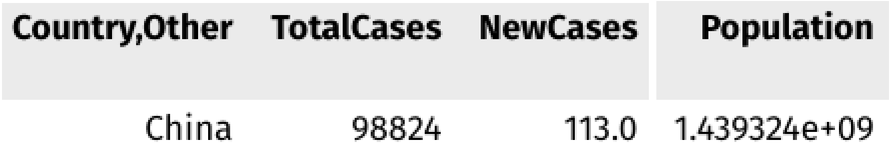




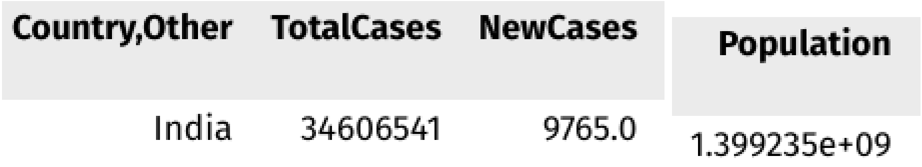
We can see that the number of countries with a population between 0 and 96 million is over 150. That explains why the number of countries, whose new cases are low, is also high.



In our expectation, the number of new cases should be proportional to the population. However, China and India are special cases.



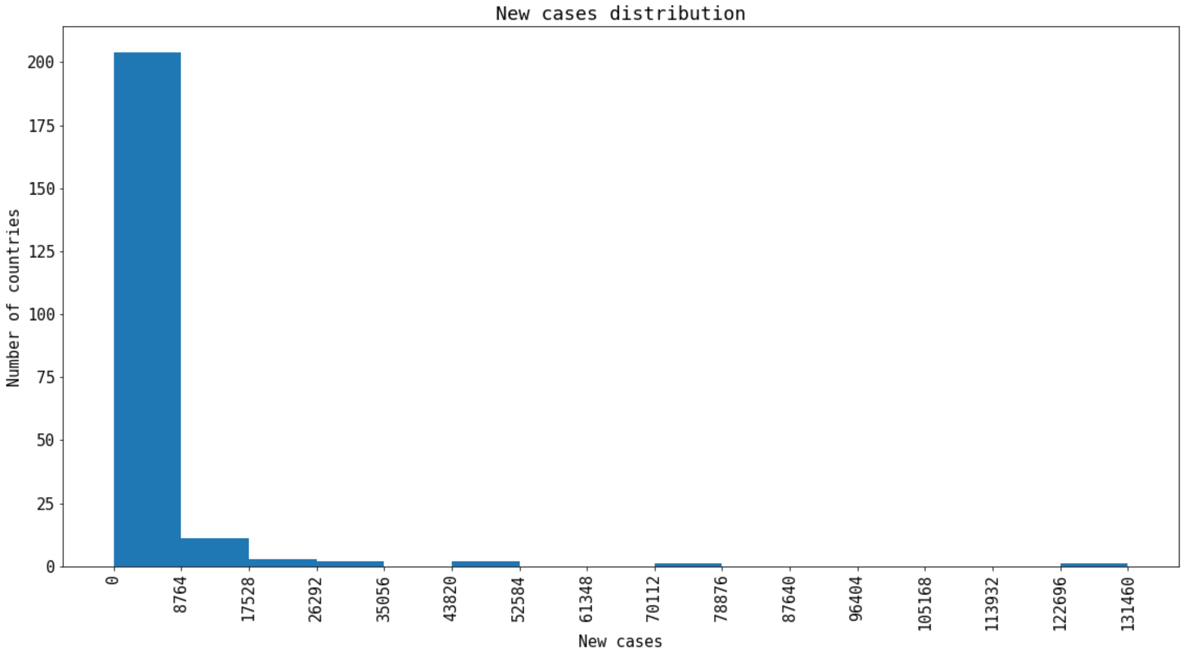
Country,Other	TotalCases	NewCases	Population
China	98824	113.0	1.439324e+09

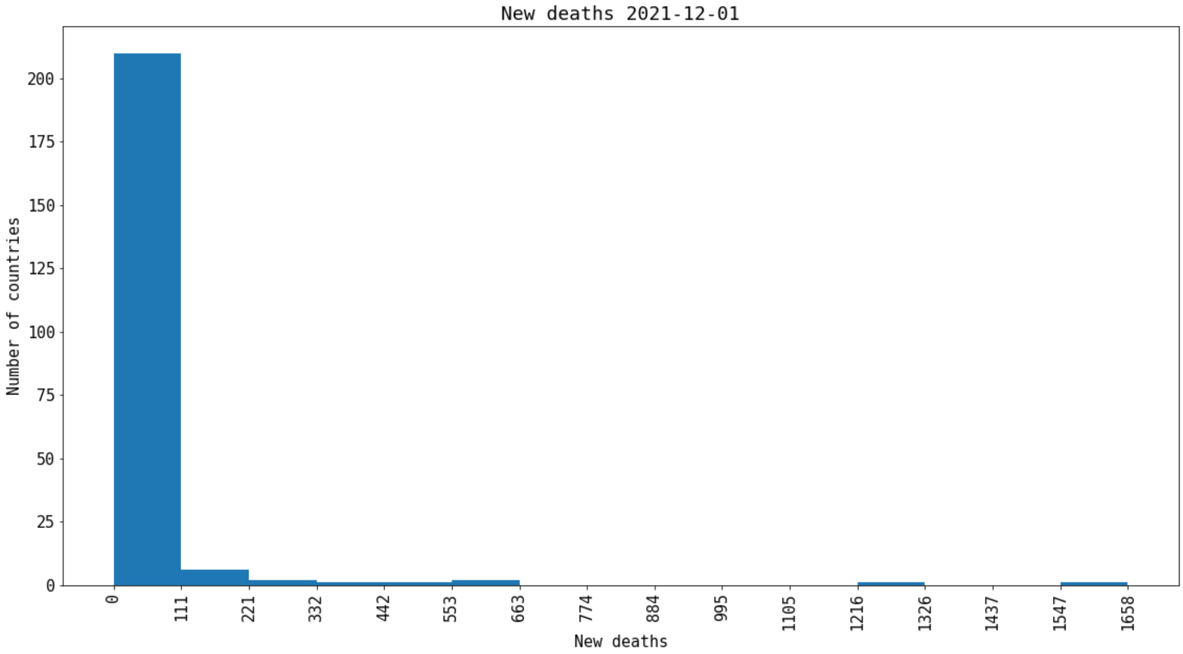


Country,Other	TotalCases	NewCases	Population
India	34606541	9765.0	1.399235e+09

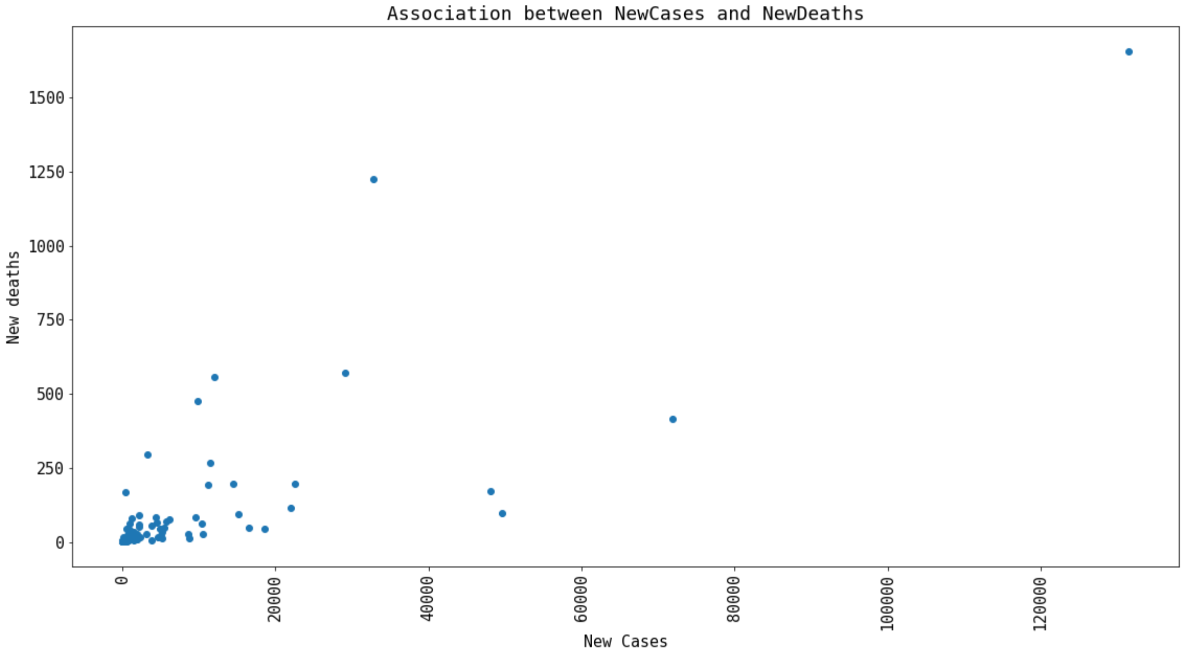
Even though the population of China is over 1.4 billion, the number of new cases is only 113, the numbers for India are also nearly 1.4 billion and 9765 new cases, respectively.

NewCases and NewDeaths:



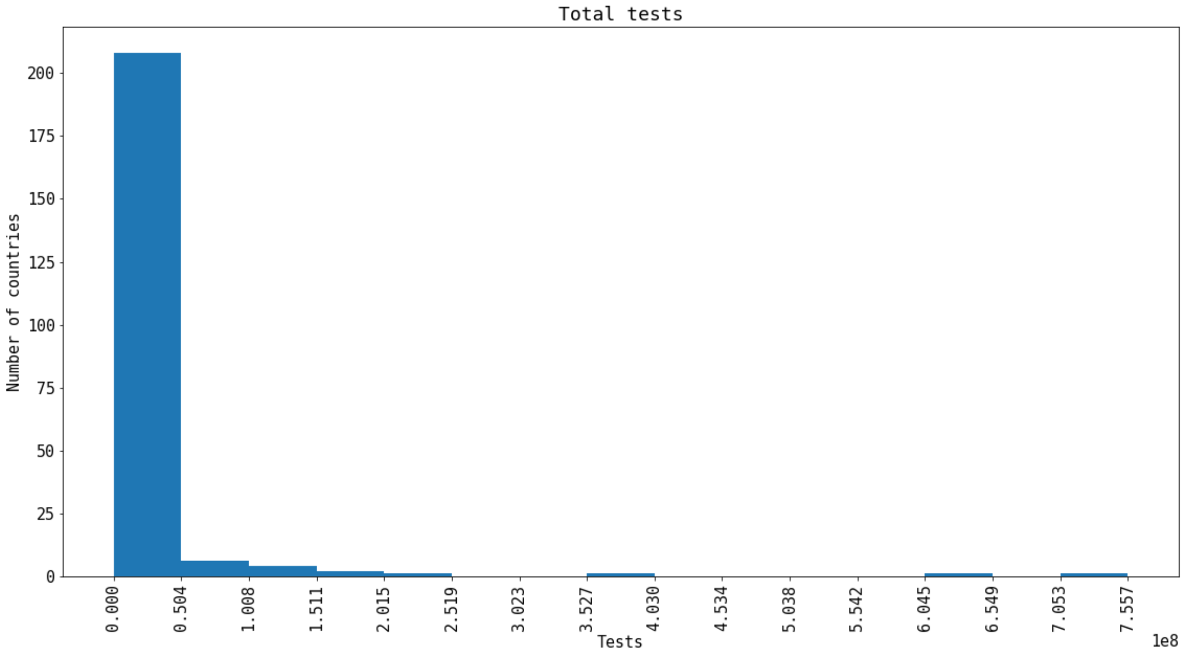


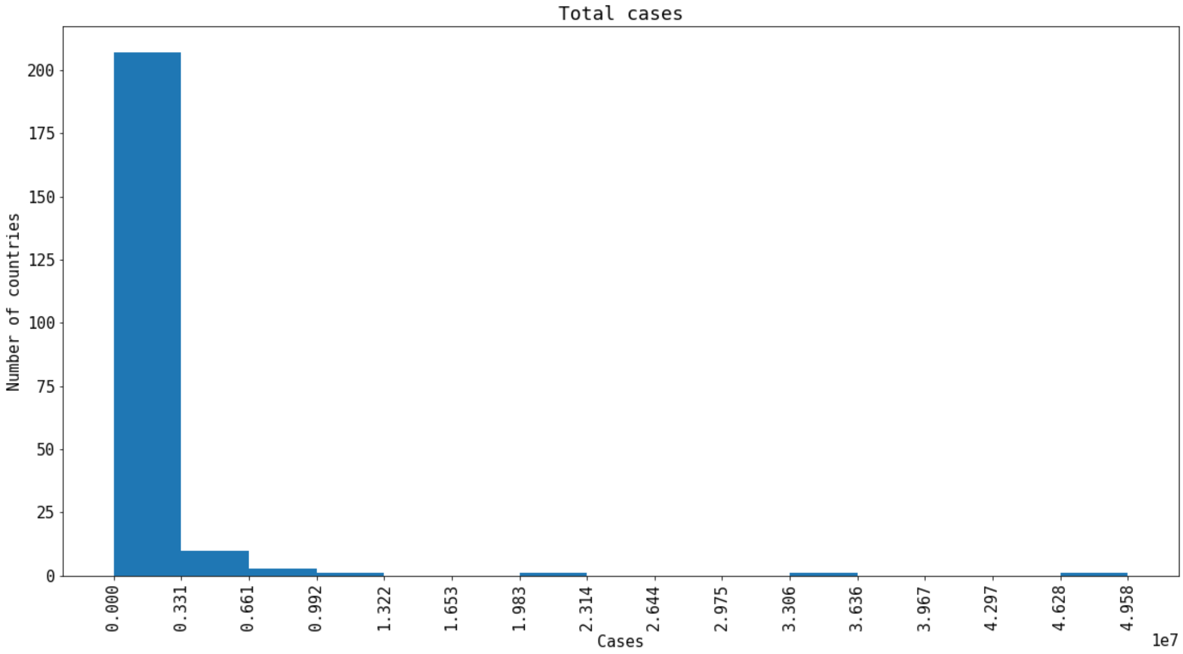
Most of the countries in the world have the number of new cases between 0 - 8764 as well as the number of new deaths which is between 0 - 111.



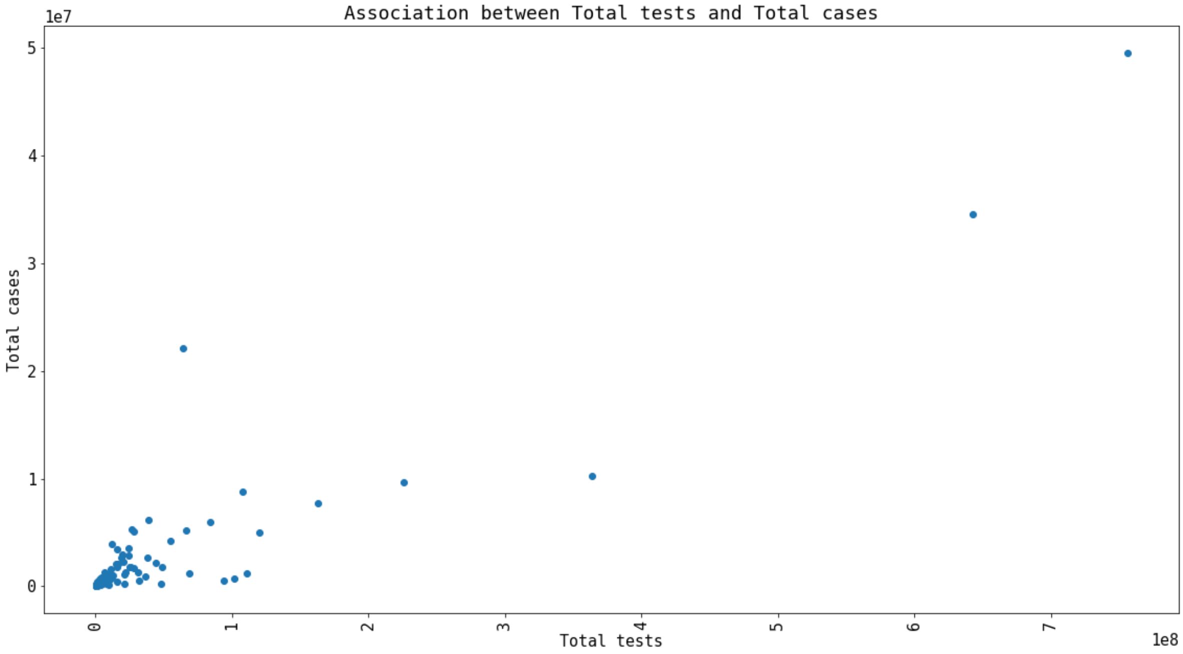
From the scatter plot, we can see that there is a relationship between NewCases and NewDeaths. It looks like the number of new cases is proportional to the number of new deaths. As a result, we can conclude that the higher number of new cases will result in the higher number of deaths.

TotalTest and TotalCases:





There are over 200 countries whose number of tests conducted and total cases are between 0 and 50,400,000, and between 0 and 3,310,000, respectively.



There is obviously a rising trend between TotalCases and TotalTests. More tests conducted will result in more cases to be found. This number could be the result of either low accuracy tests or great efforts of testing for covid 19 cases by governments.

REFERENCES:

<https://www.worldometers.info/coronavirus/about/> - About Worldometer

<https://machinelearningcoban.com/tabml_book/ch_data_processing/eda_purpose.html> - EDA purpose

<https://drive.google.com/drive/folders/1Q4SiCkFViU_ODI__bqHzzHVTvmZ__8QV> - Introduction to data science - Thang Le Ngoc

<https://drive.google.com/drive/folders/18cMq6zT-XA_XJd4SH8wMmxexAnG5eukb> - Data science Techniques - Kien Tran Trung