MAKERERE UNIVERSITY

BSE2301 SOFTWARE ENGINEERING MINI PROJECT 2

NAME	STUDENT NUMBER	REGISTRATION NUMBER	SECTION WORKED ON
KIKULE SHAWN J	2000707794	20/U/7794/PS	4
KATO MATHIAS KYESWA LULE	2000707788	20/U/7788/PS	5
NVANNUNGI JULIET	2000707745	20/U/7745/PS	3
KITONSA ELVIS	2000707785	20/U/7785/PS	6
NKINZI KAYLA GOLOOBA S	2000702501	20/U/2501/EVE	1&2

SUPERVISED BY : NDIGEZZA LIVINGSTONE

1. INTRODUCTION

This report entails thorough analysis conducted on the dataset, "covid_19_Dataset.csv".

This dataset is a collection of records collected from the effects covid-19 virus which was declared a global pandemic in 2019. The dataset is describes the presence of this virus in different regions around the world.

2. KEY OBJECTIVES FOR ANALYSIS DATA:

- To determine the number of people who died in each country and region.
- To determine the number of recoveries from the virus in each country and region.
- To determine the confirmed cases in each country and region.
- To determine the number of people who are actively living with the virus in each country and region.
- To determine the number of countries from each region.
- To determine the recovery rate in the most affected countries.
- To determine the number of confirmed cases in the most affected regions.
- To show the relationship between confirmed cases and the recoveries in each region.
- To show the relationship between the new cases and the new recoveries.

3. FEATURES/ COLUMNS ANALYSED

- **Confirmed**
  This column contains the number of cases of covid_19 victims registered in each country or region.
  The main reason for analysis of this column was to determine which country had the highest rise of the covid_19 virus during its initial spread and this involved all the people who received a positive result on a Rapid Antigen Test.

- **Deaths**
  This column contains the number of deaths registered from each country or region during the covid-19 pandemic.
  This column includes all deaths where covid-19 was determined to have been the underlying cause of death.
  This in turn entails the country that had the highest death rate during covid-19 spread.

- **Recovered**
  This column contains the recovery rates from each country.
  With this, we were able to determine a country's development towards controlling the corona virus so as to reduce on the number of deaths registered.

- **Active**
  This column contains all the active cases from each country.
  This enabled us to know how many people were diagnosed with covid_19 based on their exposure to other people with the virus and on their symptoms.
  With this, we were able to know how possible it is to live with the virus without severe effects.

- **WHO region**
  This column contains the different regions that were affected by the covid-19 virus.
  This column contains 6 regions that is Eastern Mediterranean, Europe, Africa, Americas, Western Pacific and South-East Asia.
  With this, we were able to determine which region was highly affected by the covid-19 virus.

4. PROCESSES AND TECHNIQUES USED IN ANALYSIS

4.1 GENERAL DESCRIPTION

The dataset is loaded under the variable name 'data'.

```
data = pd.read_csv("covid_19_Datasets.csv")
```

On Initial visual analysis on the tables, we realised that some of the columns used were vague, so we renamed them using the command:

```
data = data.rename(columns = {"Country/Region": "Country", "Deaths / 100 Cases": "Deaths(%)",
"Recovered / 100 Cases": "Recoveries(%)", "Deaths / 100 Recovered": "Deaths per 100
recoveries"})
```

```
print(data.columns)
```

Columns Before Renaming

```
Index(['Country/Region', 'Confirmed', 'Deaths', 'Recovered', 'Active',
       'New cases', 'New deaths', 'New recovered', 'Deaths / 100 Cases',
       'Recovered / 100 Cases', 'Deaths / 100 Recovered',
       'Confirmed last week', '1 week change', '1 week % increase',
       'WHO Region'],
      dtype='object')
```

Columns After Renaming

```
Index(['Country', 'Confirmed', 'Deaths', 'Recovered', 'Active', 'New cases',
       'New deaths', 'New recovered', 'Deaths(%)', 'Recoveries(%)',
       'Deaths per 100 recoveries', 'Confirmed last week', '1 week change',
       '1 week % increase', 'WHO Region'],
      dtype='object')
```

Using the '.shape' method,

```
print(data.shape)
```

We get to know that the dataset contains 187 rows and 15 columns

```
(187,15)
```

Using the '.info().to_string()' method,

```
print(data.info().to_string())
```

We got detailed description about the columns of the data as seen in the image below.

```
RangeIndex: 187 entries, 0 to 186
Data columns (total 15 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Country                    187 non-null    object
 1   Confirmed                  187 non-null    int64
 2   Deaths                     187 non-null    int64
 3   Recovered                  187 non-null    int64
 4   Active                     187 non-null    int64
 5   New cases                  187 non-null    int64
 6   New deaths                 187 non-null    int64
 7   New recovered              187 non-null    int64
 8   Deaths(%)                  187 non-null    float64
 9   Recoveries(%)              187 non-null    float64
 10  Deaths per 100 recoveries  187 non-null    float64
 11  Confirmed last week        187 non-null    int64
 12  1 week change              187 non-null    int64
 13  1 week % increase          187 non-null    float64
 14  WHO Region                 187 non-null    object
dtypes: float64(4), int64(9), object(2)
memory usage: 22.0+ KB
None
```

This image also tells us that there are 4 columns of data type float, 9 of data type int, and 2 object data types. And that it uses up a memory of 22.0+KB

Using the 'describe().to_string()' method,

```
print(data.describe().to_string())
```

	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths(%)	Recoveries(%)	Deaths per 100 recoveries	Confirmed last week	1 week change	1 week % increase
count	1.870000e+02	187.000000	1.870000e+02	1.870000e+02	187.000000	187.000000	187.000000	187.000000	187.000000	187.00	1.870000e+02	187.000000	187.000000
mean	8.813094e+04	3497.518717	5.063148e+04	3.400194e+04	1222.957219	28.957219	933.812834	3.019519	64.820535	inf	7.868248e+04	9448.459893	13.606203
std	3.833187e+05	14100.002482	1.901882e+05	2.133262e+05	5710.374790	120.037173	4197.719635	3.454302	26.287694	NaN	3.382737e+05	47491.127684	24.509838
min	1.000000e+01	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	1.000000e+01	-47.000000	-3.840000
25%	1.114000e+03	18.500000	6.265000e+02	1.415000e+02	4.000000	0.000000	0.000000	0.945000	48.770000	1.45	1.051500e+03	49.000000	2.775000
50%	5.059000e+03	108.000000	2.815000e+03	1.600000e+03	49.000000	1.000000	22.000000	2.150000	71.320000	3.62	5.020000e+03	432.000000	6.890000
75%	4.046050e+04	734.000000	2.260600e+04	9.149000e+03	419.500000	6.000000	221.000000	3.875000	86.885000	6.44	3.708050e+04	3172.000000	16.855000
max	4.290259e+06	148011.000000	1.846641e+06	2.816444e+06	56336.000000	1076.000000	33728.000000	28.560000	100.000000	inf	3.834677e+06	455582.000000	226.320000

The image above lets us know the count, mean, standard deviation, minimum value, maximum values, 25th percentile, 50th percentile, and 75th percentile of the data in each column.

An Overview Of the Data
At the top

```
print(data.head(4)) # which returns a summarised view of the first 4 rows
```

	Country	Confirmed	...	1 week % increase	WHO Region
0	Afghanistan	36263	...	2.07	Eastern Mediterranean
1	Albania	4880	...	17.00	Europe
2	Algeria	27973	...	18.07	Africa
3	Andorra	907	...	2.60	Europe

At the Bottom

```
print(data.tail(4)) # which returns a summarised view of the bottom 4 rows
```

```
            Country  Confirmed  ...  1 week % increase            WHO Region
183  Western Sahara         10  ...               0.00                Africa
184           Yemen       1691  ...               4.45  Eastern Mediterranean
185          Zambia       4552  ...              36.86                Africa
186        Zimbabwe       2704  ...              57.85                Africa
```

4.2 DATA CLEANING

From the image above we can see that there are no null values, no duplicate values, no empty cells and no data written in the wrong format in any of the columns. This helps us to know that we do not need the methods needed for dropping and filling null values.

```
print(data.isnull().sum())
```

Proves this by returning 0 null values against all columns

```
Country                    0
Confirmed                  0
Deaths                     0
Recovered                  0
Active                     0
New cases                  0
New deaths                 0
New recovered              0
Deaths(%)                  0
Recoveries(%)              0
Deaths per 100 recoveries  0
Confirmed last week        0
1 week change              0
1 week % increase          0
WHO Region                 0
```

Using the .duplicated() method, we also get to know that there aren't any duplicated rows in the dataset.

4.3 EXPLORATORY ANALYSIS

The total number of;

-Confirmed Cases

```
print(data["Confirmed"].sum())
```

Returns

```
16480485
```

-Deaths

```
print(data["Deaths"].sum())
```

Returns

```
654036
```

-Recovered

```
print(data["Recovered"].sum())
```

Returns

```
9468087
```

-Active Cases

```
print(data["Active"].sum())
```

Returns

```
6358362
```

Regions Where Data was collected from

```
regions = print(data["WHO Region”].unique())
print(regions)
```

Returns

```
['Eastern Mediterranean' 'Europe' 'Africa' 'Americas' 'Western Pacific'
 'South-East Asia']
```

-Countries From each region where data was collected

```
print(data["WHO Region”].value_counts())
```

Returns

```
Europe                   56
Africa                   48
Americas                 35
Eastern Mediterranean    22
Western Pacific          16
South-East Asia          10
Name: WHO Region, dtype: int64
```

-new cases

```
print(data["New cases”].sum())
```

returns

```
228693
```

-new deaths

```
print(data["New deaths”].sum())
```

returns

```
5415
```

-new recovered cases

```
print(data["New recovered”].sum())
```

returns

```
174623
```

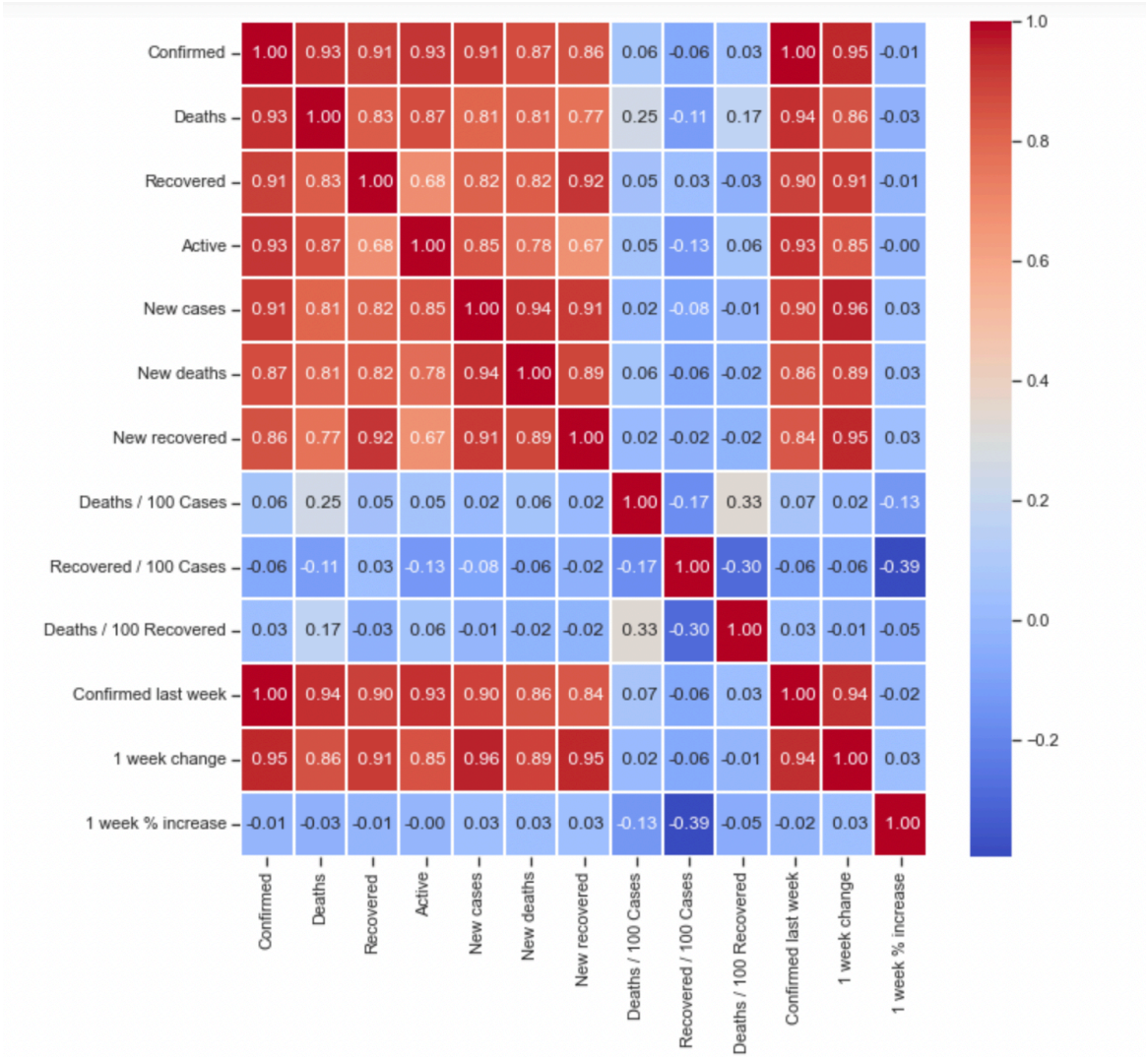
SORTING DATA BY REGIONS

```
import numpy as np

#Grouping data based on the WHO Region........
group_by=covid_dataset.groupby('WHO Region')
x= group_by['Confirmed'].agg(np.mean)        #returns the average number of confirmed cases in each WHO region.....
q= group_by['Confirmed'].agg(np.sum)         #returns the total number of confirmed cases in each WHO Region......
y= group_by['Deaths'].agg(np.sum)            #returns the total number of deaths......
z= group_by['Active'].agg(np.sum)            #returns the total number of active cases........
p= group_by['Country'].agg(np.size)   #returns the total number of countries in each WHO Region.....
r =group_by['Recovered'].agg(np.sum)         #total number of recovered cases in each WHO Region......

print('')
print('The average number of cases confirmed in each WHO Region:')
print(x)
print('')
print('The total number of cases confirmed in each WHO Region:')
print(q)
print('')
print('The total number of deaths in each WHO Region:')
print(y)
print('')
print('The total number of active cases in each WHO Region:')
print(z)
print('')
print('The total number of countries in each WHO Region:')
print(p)
print('')
print('The total number of recovered cases in each WHO Region:')
print(r)
```

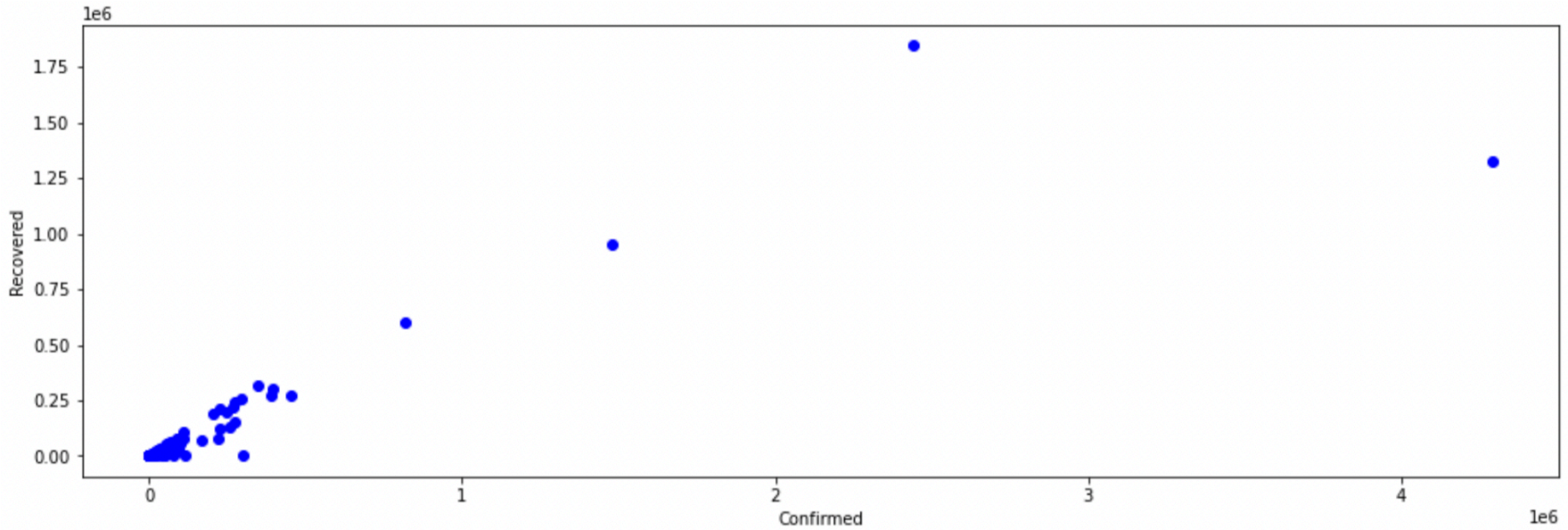
```
The average number of cases confirmed in each WHO Region:
WHO Region
Africa                   15066.812500
Americas                252551.028571
Eastern Mediterranean    67761.090909
Europe                   58920.053571
South-East Asia         183529.700000
Western Pacific          18276.750000
Name: Confirmed, dtype: float64

The total number of cases confirmed in each WHO Region:
WHO Region
Africa                   723207
Americas                8839286
Eastern Mediterranean   1490744
Europe                  3299523
South-East Asia         1835297
Western Pacific          292428
Name: Confirmed, dtype: int64

The total number of deaths in each WHO Region:
WHO Region
Africa                   12223
Americas                342732
Eastern Mediterranean    38339
Europe                  211144
South-East Asia          41349
Western Pacific           8249
Name: Deaths, dtype: int64

The total number of active cases in each WHO Region:
WHO Region
Africa                   270339
Americas                4027938
Eastern Mediterranean    251005
Europe                  1094656
South-East Asia          637015
Western Pacific           77409
Name: Active, dtype: int64

The total number of countries in each WHO Region:
WHO Region
Africa                  48
Americas                35
Eastern Mediterranean   22
Europe                  56
South-East Asia         10
Western Pacific         16
Name: Country, dtype: int64

The total number of recovered cases in each WHO Region:
WHO Region
Africa                   440645
Americas                4468616
Eastern Mediterranean   1201400
Europe                  1993723
South-East Asia         1156933
Western Pacific          206770
Name: Recovered, dtype: int64
```

5. DATA VISUALISATION

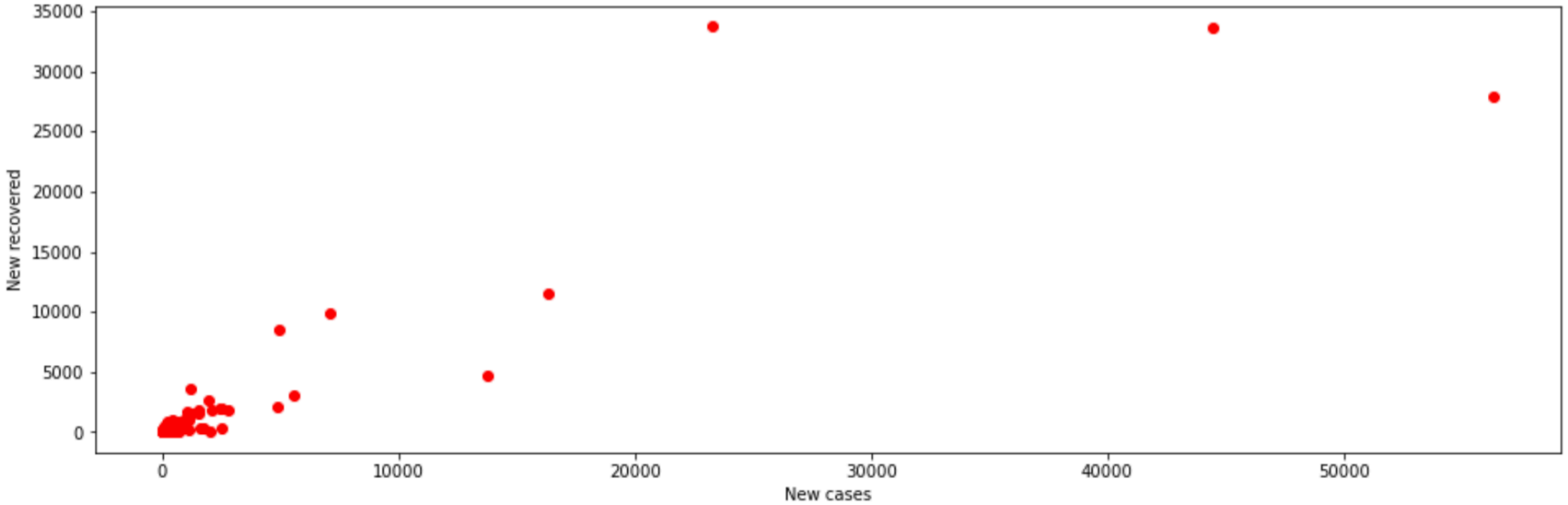
- **Correlations**

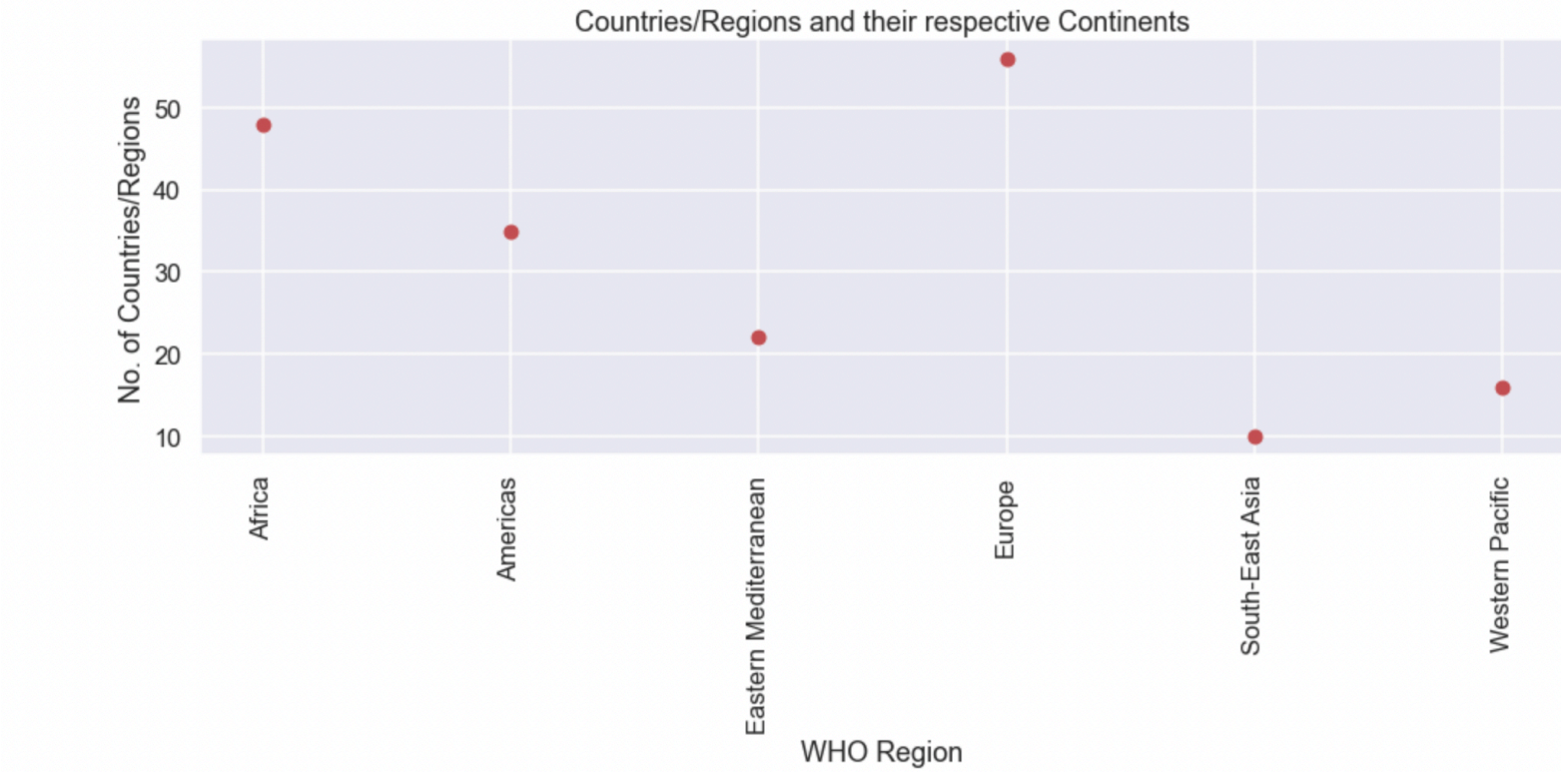


SCATTER GRAPH FOR RECOVERED CASES AGAINST CONFIRMED CASES

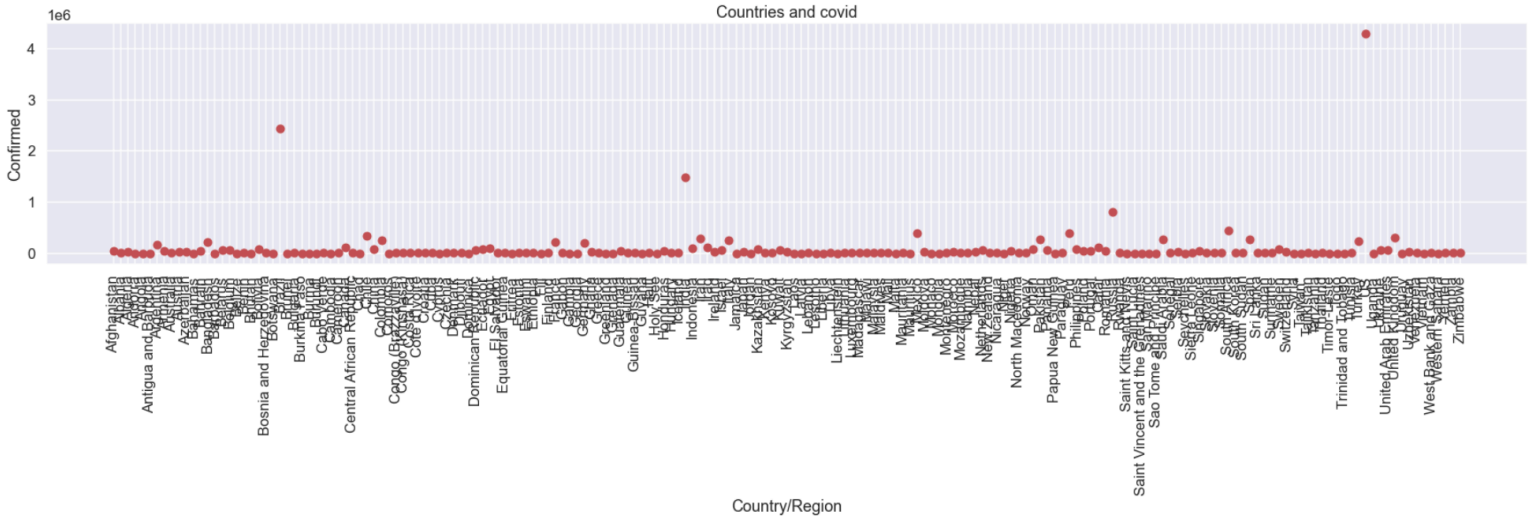


A SCATTER GRAPH FOR NEW RECOVERED CASES AND NEW CASES

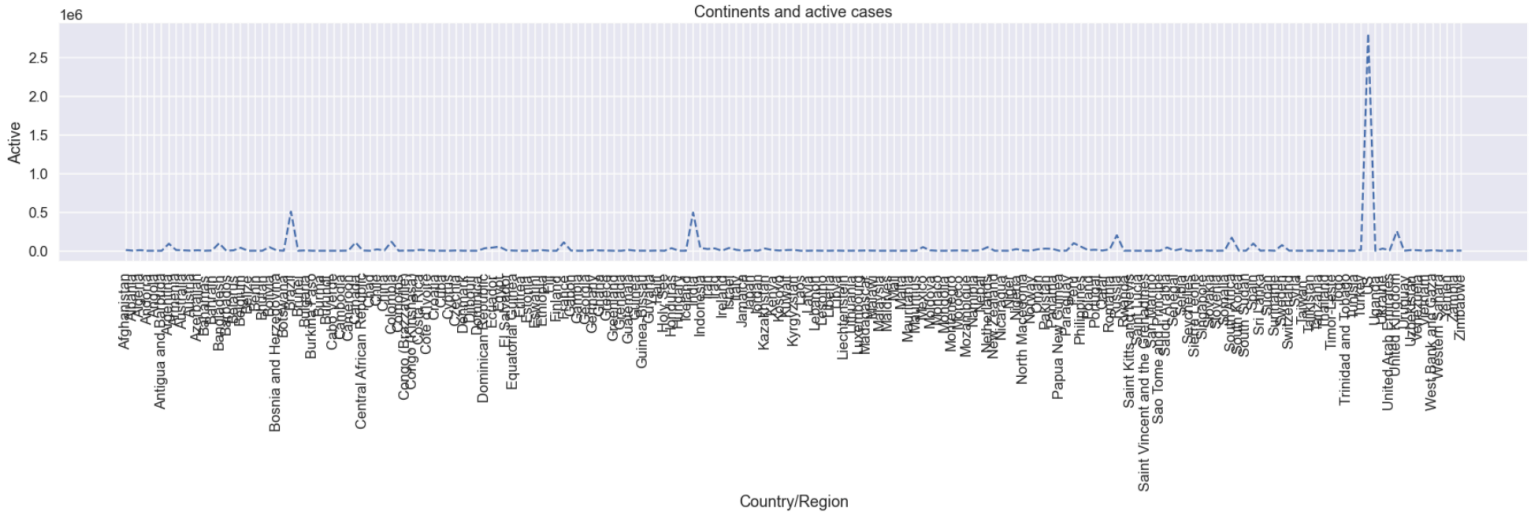



Countries/Regions and their respective Continents
No. of Countries/Regions
50
40
30
20
10
Africa
Americas
Eastern Mediterranean
Europe
South-East Asia
Western Pacific
WHO Region

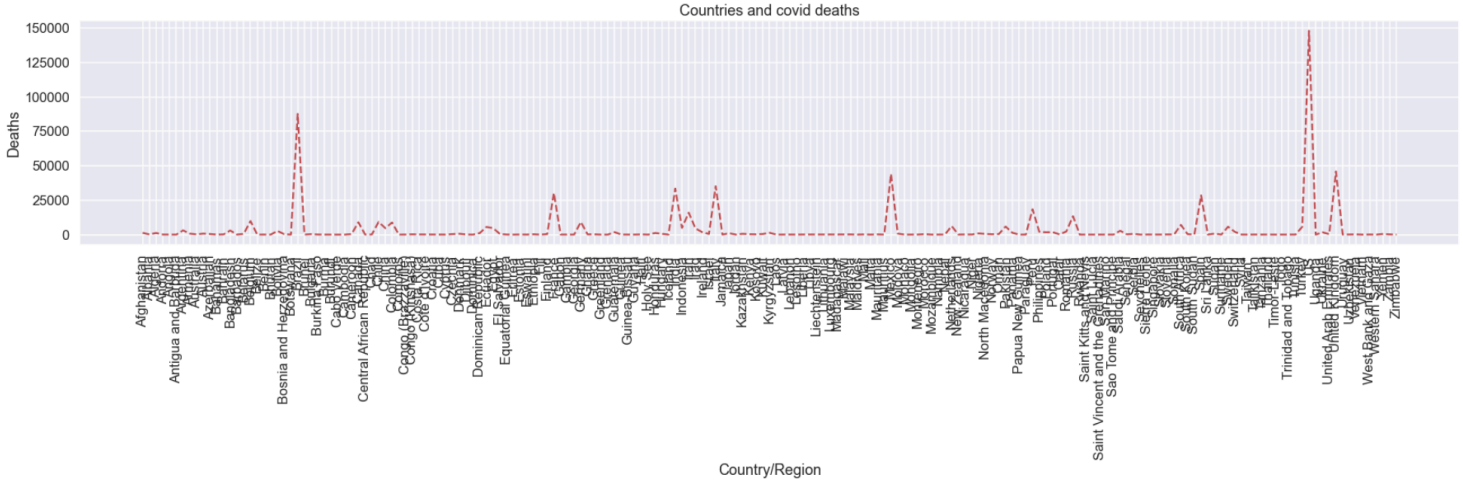
- Confirmed cases in per country



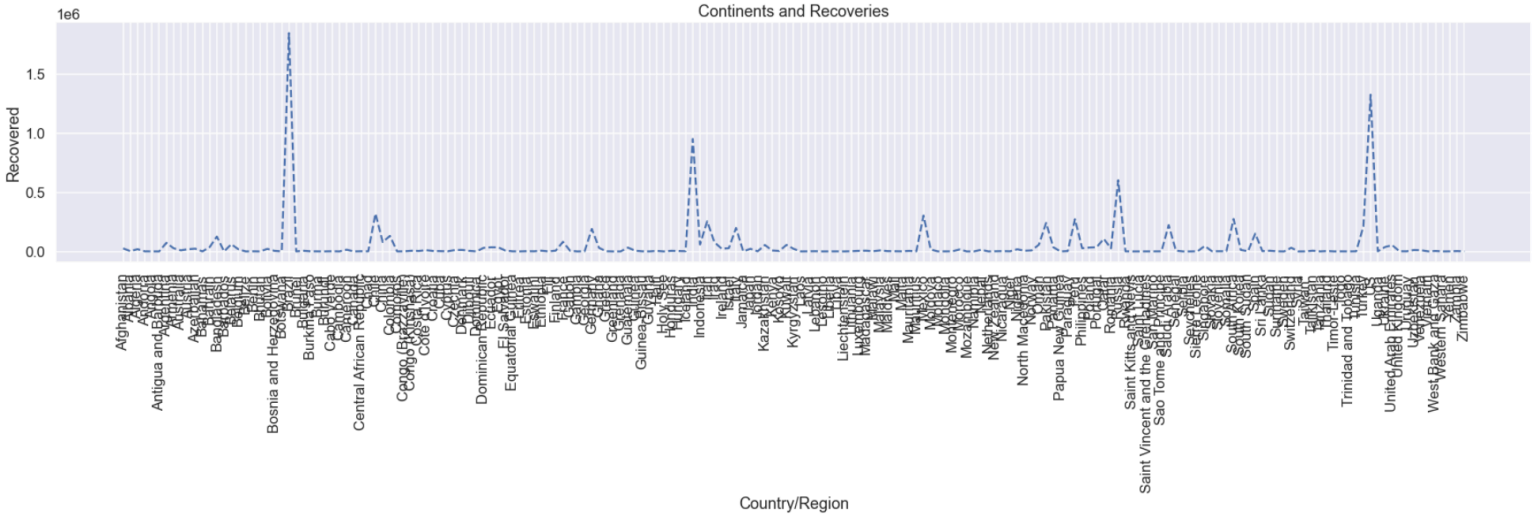
- Active cases per country



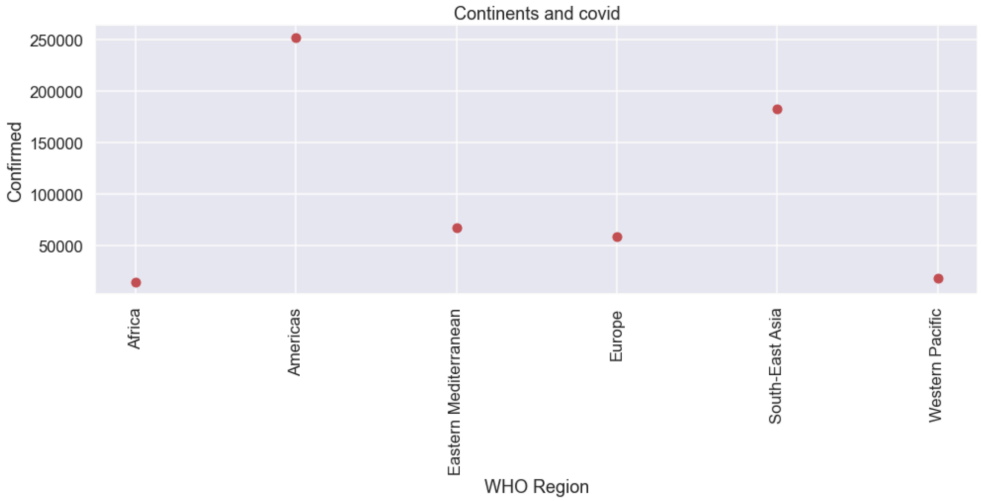
- **Deaths per Country**



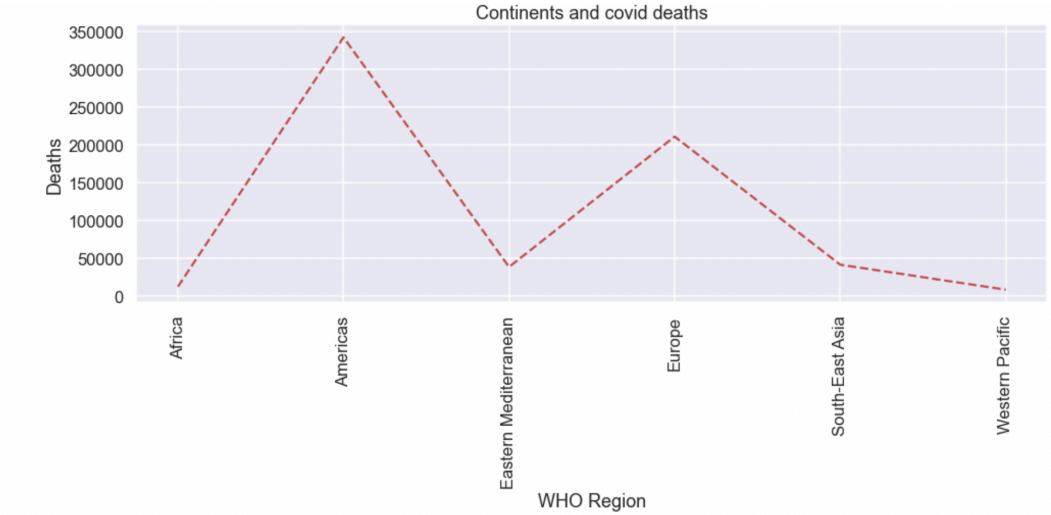
- **Recovered Cases per Country**



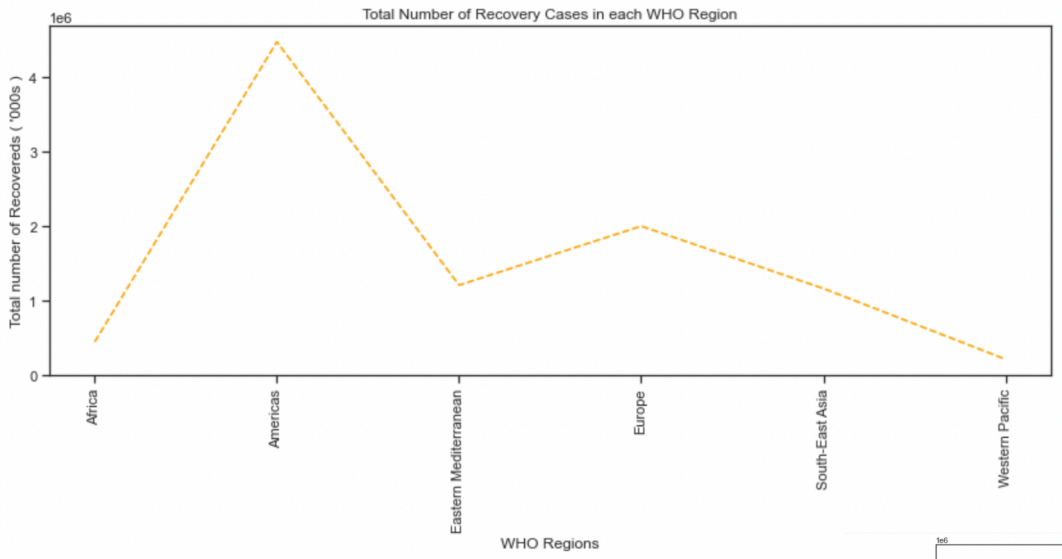
Confirmed Cases For each Region



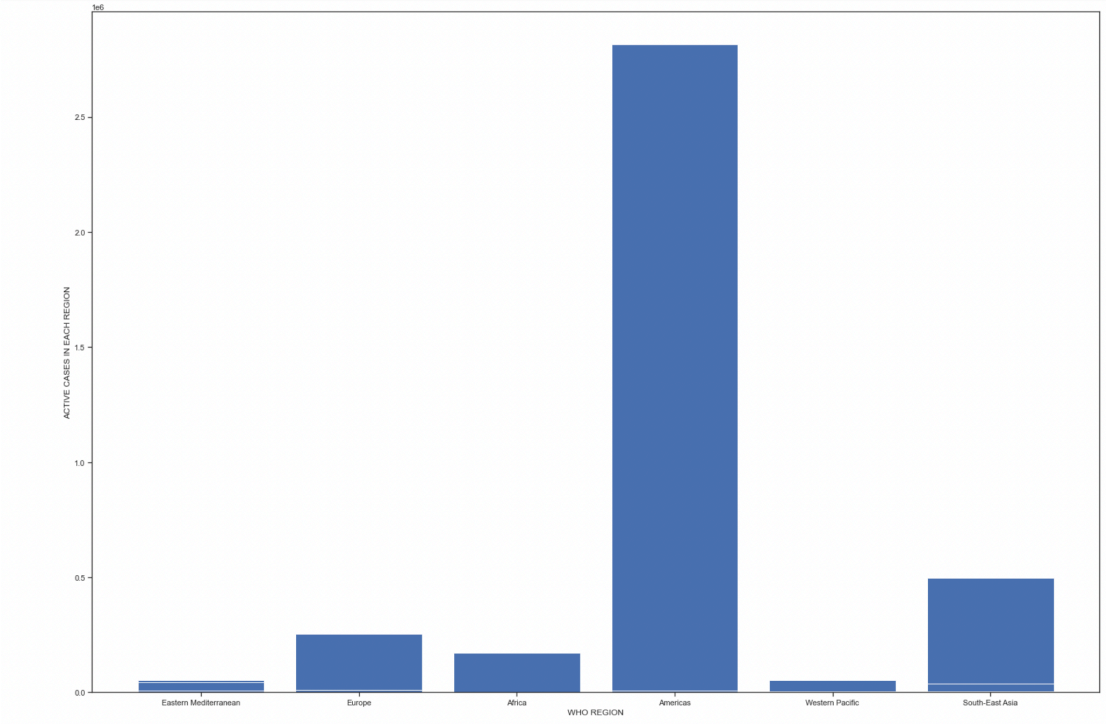
Deaths For Each Region



Recovered Cases For Each Region



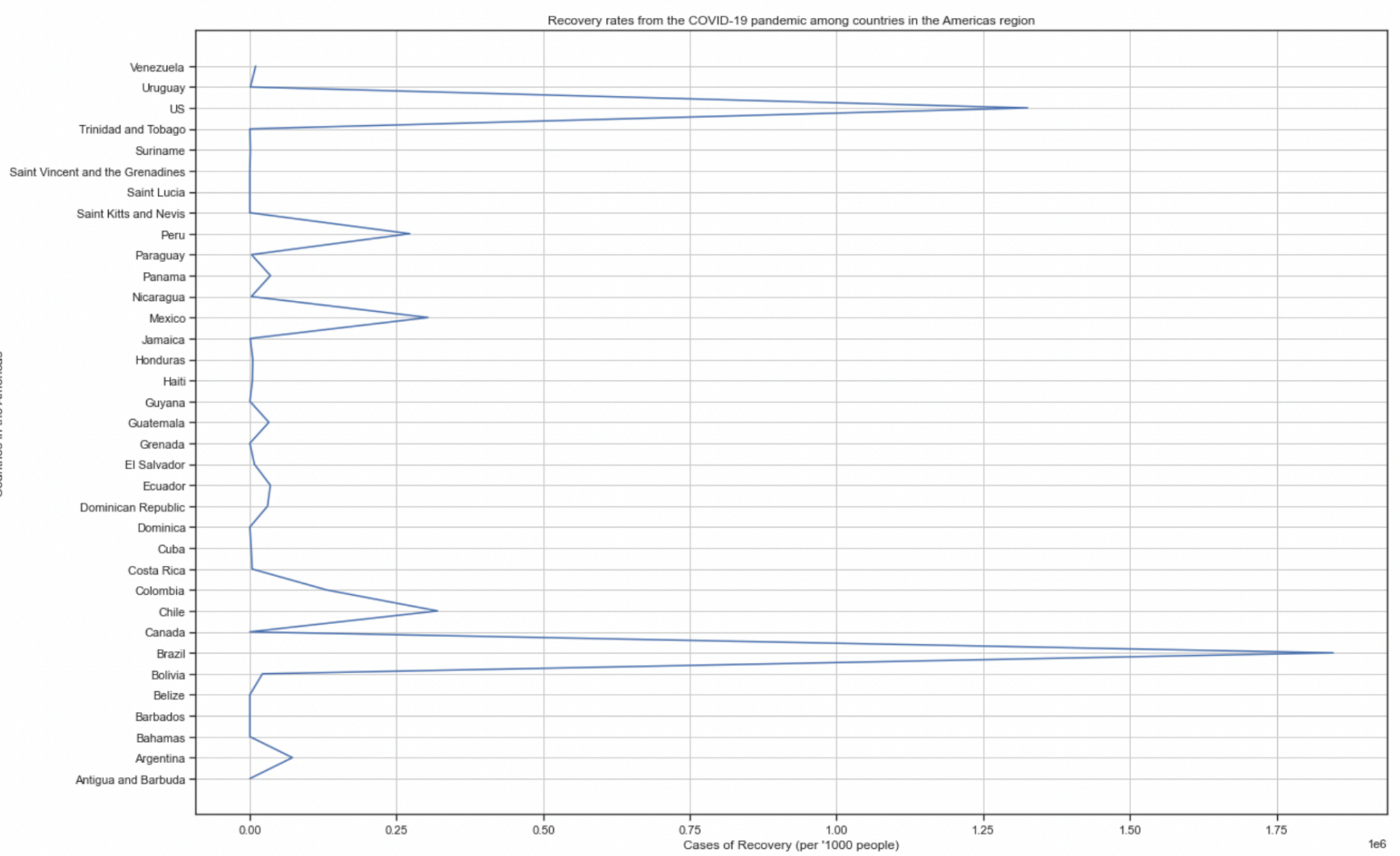
Active Cases for Each Region



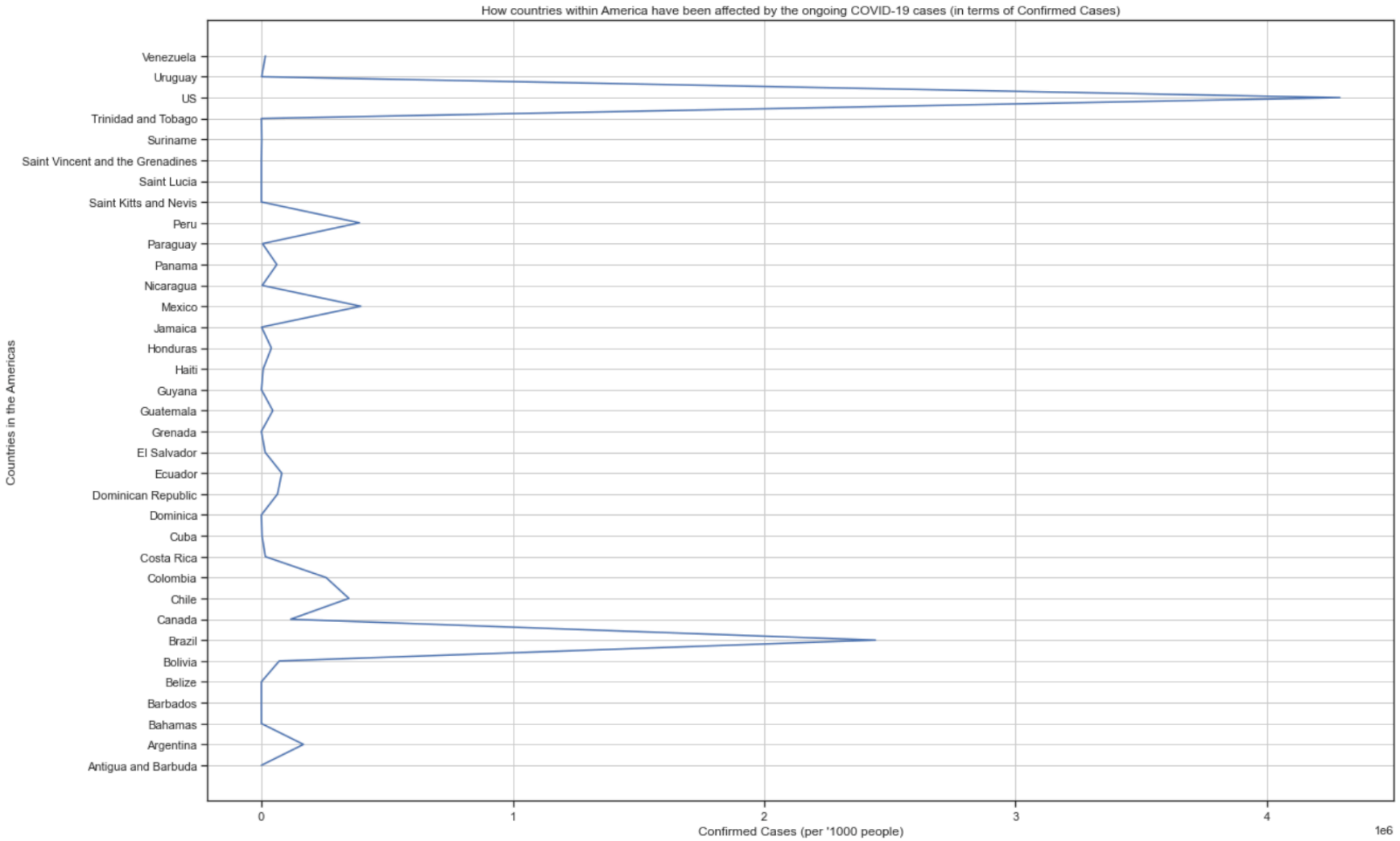
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
\

-

- Recovery rate in most affected



- **Confirmed Cases in the most affected Region**



6. CONCLUSIONS

- The Americas and South-East Asia were the regions with most confirmed cases, with the US and Brazil having the most confirmed cases (4290259 and 2442375 cases respectively) whereas Africa had the lowest number of confirmed cases with the Western Sahara having the lowest cases(10) recorded. This comes to show that the virus is most communicable in the Americas and South-East Asia and least communicable in Africa, maybe due to the different climatic conditions in the countries
- The Americas and South-East Asia were the regions with the most confirmed deaths, with the US and Brazil having 148011 and 87618 deaths respectively, whereas Africa and Western Pacific having the lowest deaths.
- The rate of recovery in the Americas and South-East Asia was the highest recorded, this comes to show that, however high the rate of deaths is in these regions, correct measures are being taken to fight the virus in these countries.