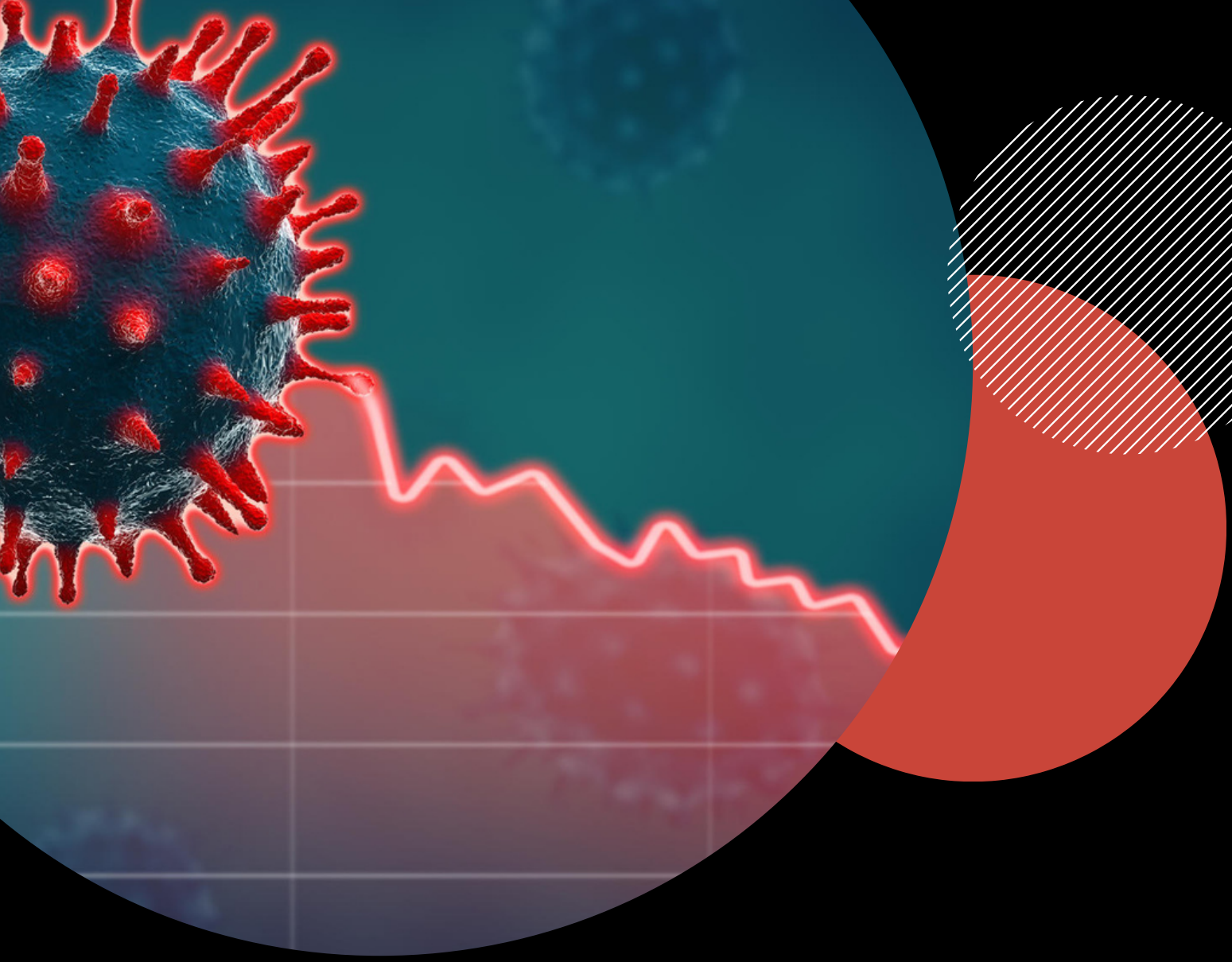
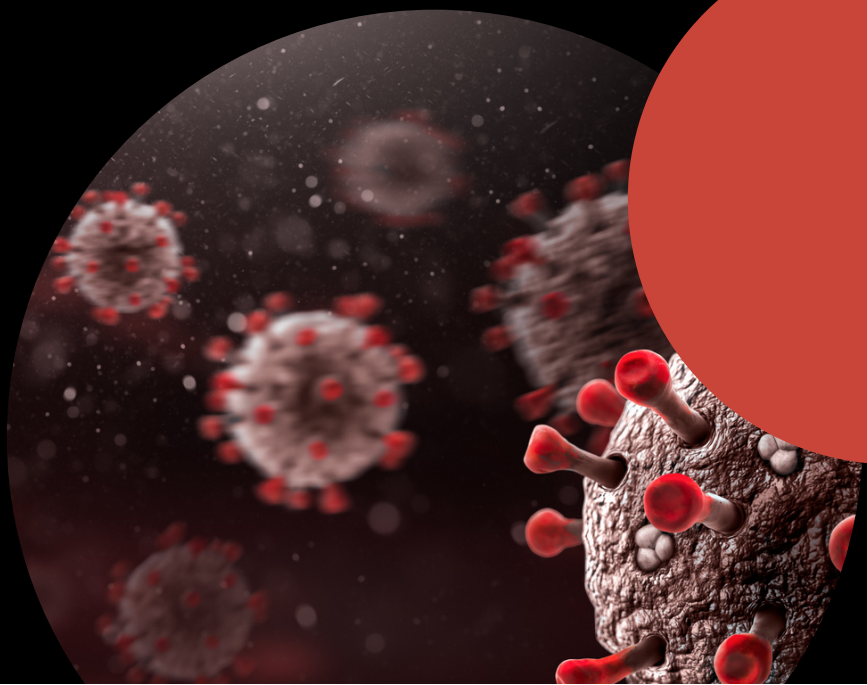


DATA RELATIONSHIP

DATA VISUALIZATION

SUBMITTED TO: LECTURER LE NGOC THANH

TABLE OF CONTENTS



About Us

Võ Văn Hoàng

20127028

Nguyễn Đức Minh

20127049

Ngô Văn Trung Nguyên

20127054

Nguyễn Minh Tuấn

20127092

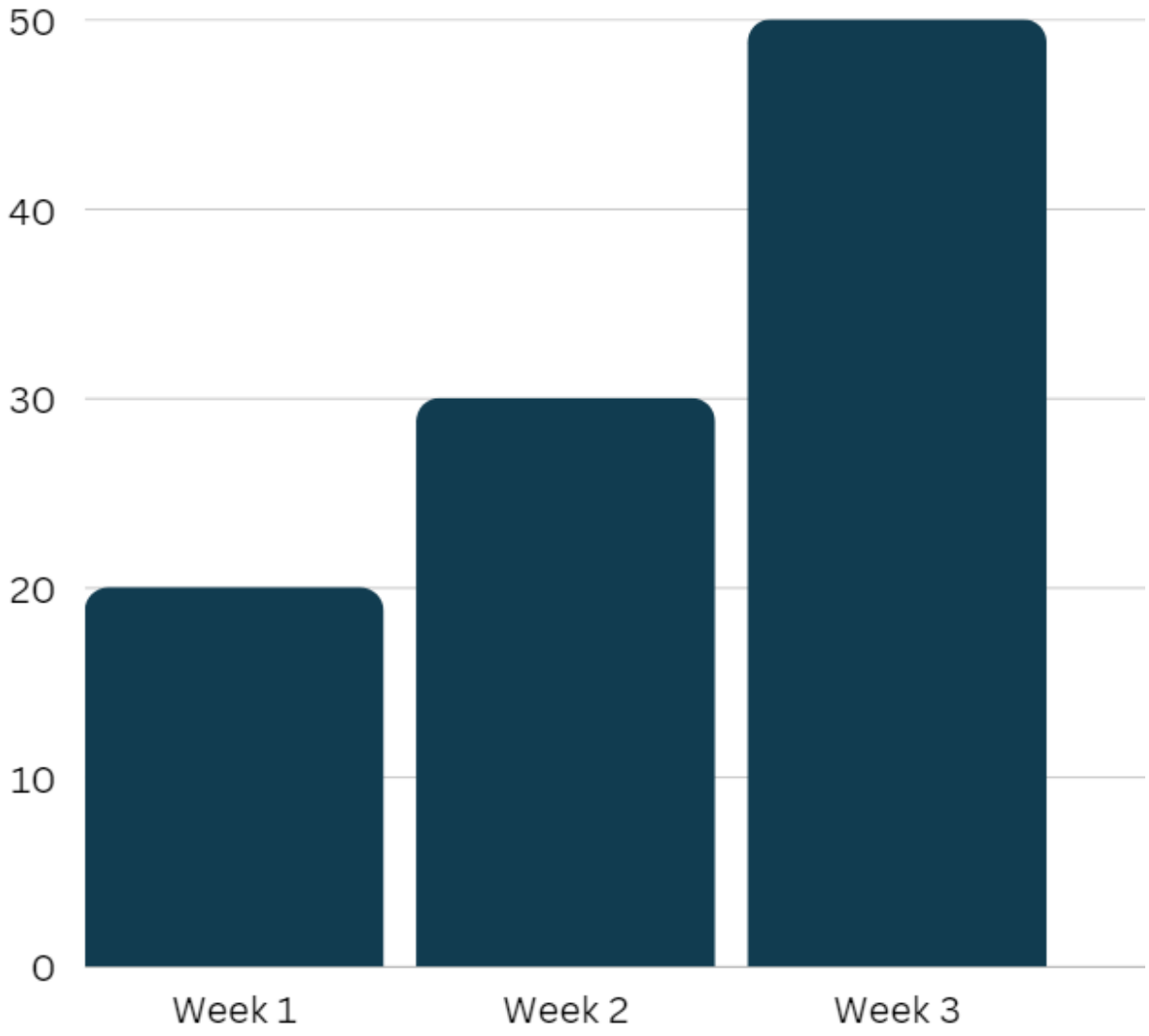
Nguyễn Trương Minh Khôi

20127214

Task Allocation

Task Description	Person in charge
Data Visualization (Continent part) Lab 01 Report	Vo Van Hoang 20127028
Data Collection Data Visualization (World part)	Nguyen Duc Minh 20127049
Data Visualization (Country part) Notebook presentation	Ngo Van Trung Nguyen 20127054
Data Visualization (Country part) Notebook support	Nguyen Minh Tuan 20127092
Data Collection Data Visualization (Continent part)	Nguyen Truong Minh Khoi 20127214

Lab Completion Rate



This is a 3-week lab of Data Visualization. We share work for each team member equally.
Week 1: Generally, we decided how to represent the content and crawl data.
Week 2: Try to visualize data by world and continents, find out some special countries.
Week 3: Explore more about the relationship of each column, try to visualize data as meaningful as possible and make a check again before submitting.

Approaching Methods

In order to understand the data deeply, first of all, we carried out the procedures of data, such as: reading data, explore data, etc.
After making sure data is ready for analysis, we do the visualization of relations between different data columns. After that, we selected the useful data fields to add into the report.

Working Flow

Step 1

Crawl data, try to create some new columns, delete some meaningless rows,...

Step 2

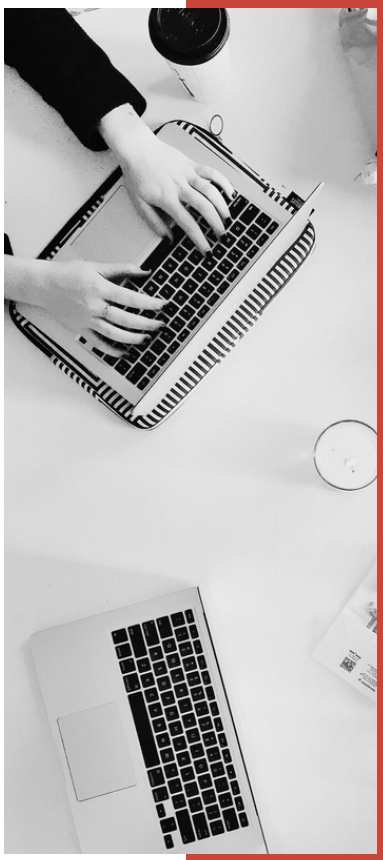
Select columns: Through visualization on correlation to select typical fields for the analysis.

Step 3

Visualize the data, check and give explantion for each kind of chart.

Step 4

Try to find out relationship between different columns with is related to machine learning.



Preparation

Library we used

```
# Concatenating Files Library
import glob
import pandas as pd
# Visualizing Libraries
import matplotlib.pyplot as plt
import seaborn as sns
import geopandas as gpd
import plotly.express as px
# Converting Code of Countries Libraries
from pycountry_convert import country_alpha2_to_continent_code, country_name_to_country_alpha2
import pycountry
import plotly.graph_objects as go
# Other necessary Libraries
## For time handling
import datetime as dt
## For string handling
import re
#list all csv files only
csv_files = glob.glob('*.{}'.format('csv'))
csv_files
```

We also use preprocess function in order to preprocess something useful

```
# Function to preprocess DataFrame

def preprocess_df(df):
    df.dropna(subset = ['Country, Other'], inplace = True)
```

Web Crawling Libraries

```
# Import Web Crawling Libraries

from bs4 import BeautifulSoup
import requests
import numpy as np
import pandas as pd
import datetime as dt
```

Preprocessing

```python
# Function to preprocess DataFrame

def preprocess_df(df):
    df.dropna(subset = ['Country, Other'], inplace = True)

    for col in df.columns:
        if col == 'New Deaths':
            continue
        df[col] = df[col].str.replace(',', '')
        df[col] = df[col].str.replace('.', '', regex = True)
        df[col] = df[col].str.replace('+', '', regex = True)

    df['Date'] = df['Date'].str.replace('Last updated: ', '')
    df['Date'] =  pd.to_datetime(df['Date'], format = '%B %d %Y %H:%M ')
    numeric_header = list(df.columns)
    numeric_header.remove('Country, Other')
    numeric_header.remove('Date')

    for col in df.columns:
        if col in numeric_header:
            df[col] = df[col].astype(float).astype('Int64')
```

Preprocessing:

- Convert Data Time Update from Object to Date Time
- New case, new deaths, new recovered: eliminating "+", convert to float
- Countries which can not be classified into continents automatically will be filled manually

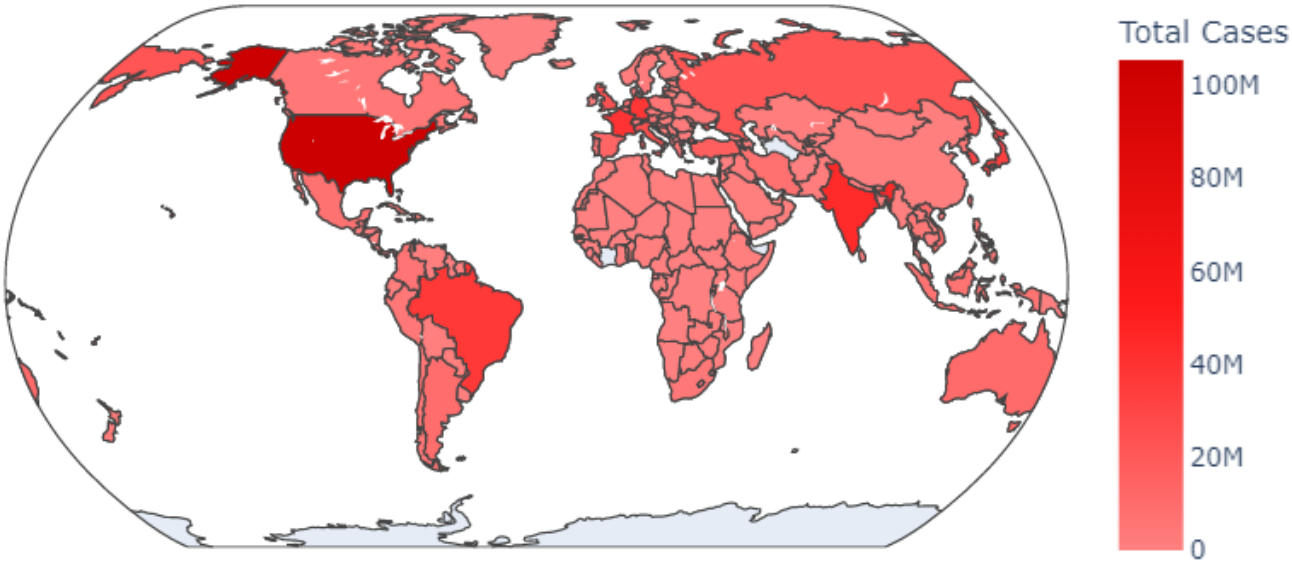
```python
df[df['Continent'].isna()]
country_to_continent = {
    "S Korea": "Asia",
    "UK": "Europe",
    "DPRK": "Asia",
    "UAE": "Asia",
    "Channel Islands": "Europe",
    "DRC": "Africa",
    "Faeroe Islands": "Europe",
    "Timor-Leste": "Asia",
    "CAR": "Africa",
    "Caribbean Netherlands": "North America",
    "Sint Maarten": "North America",
    "St Vincent Grenadines": "North America",
    "St Barth": "North America",
    "Saint Pierre Miquelon": "Europe",
    "Saint Helena": "Africa",
    "Diamond Princess": "Europe",
    "Vatican City": "Europe",
    "Western Sahara": "Africa",
    "MS Zaandam": "Europe"
}
```

Part 1

AN OVERVIEW:
THE WORLD'S CORONAVIRUS PANDEMIC

TAKE A LOOK AT WORLD'S TOTAL CASES, RECOVERED CASES, NEW CASES

WORLD HEATMAP ON TOTAL POSITIVE CASES UP TO 15/03/2023



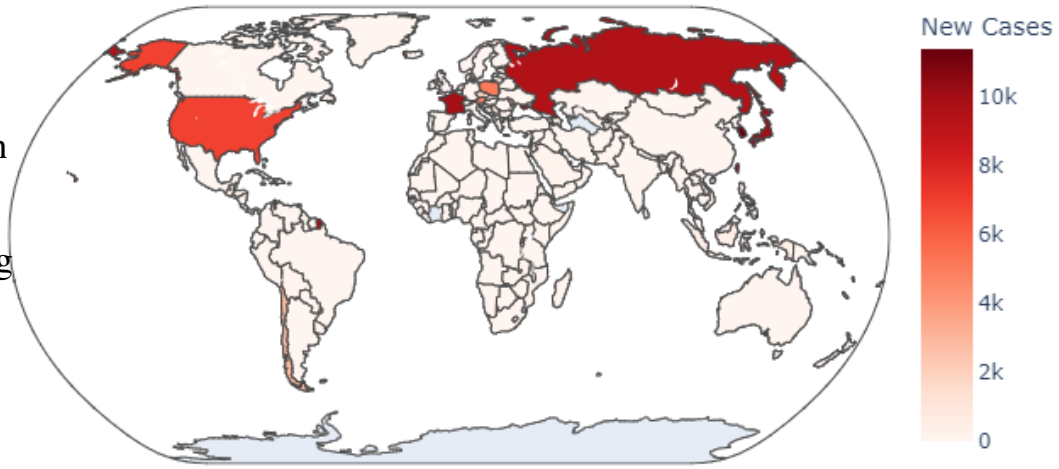
The World has been tremendously affected by this virus for the past four years, almost all countries were destroyed by this pandemic.

WORLD HEATMAP OF TOTAL RECOVERED CASES UP TO 15/03/2023

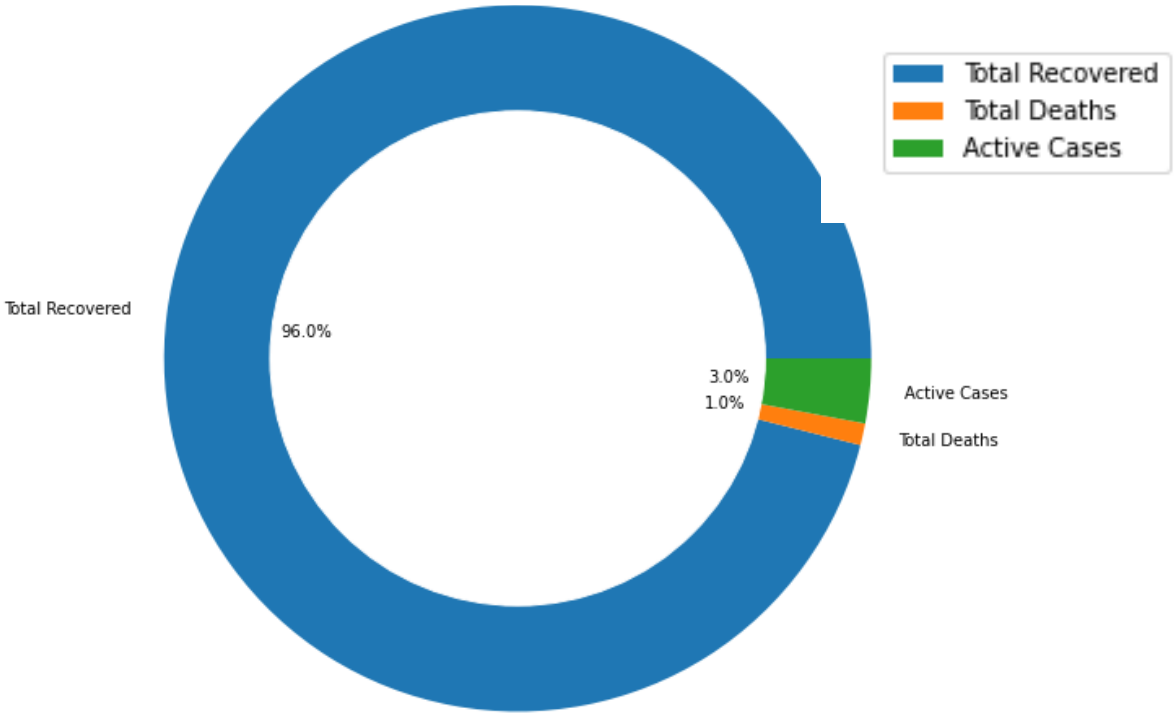
However, the number of worldwide recovered case posed positive signals with a huge number of people having overcome this disease. Hence, coronavirus may no longer be a big deal for any nations.

WORLD HEATMAP OF NEW CASES ON 15/03/2023

Though, the pandemic today has overcome the most serious period, some of big nations such as Russia, USA, Japan, South Korea, Taiwan are still be influenced.
Also, some nations sharing borders or close to one another have high indicates of New Cases and Active Cases like Russia, Japan, South Korea, Taiwan, etc. as well as Poland is so closed to Russia.

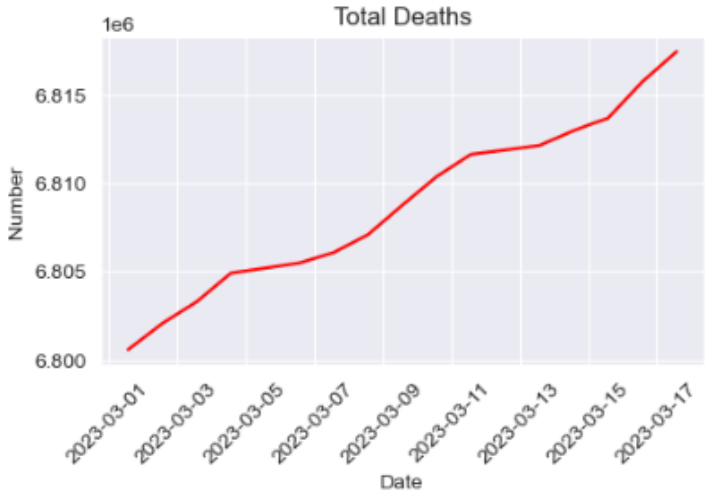


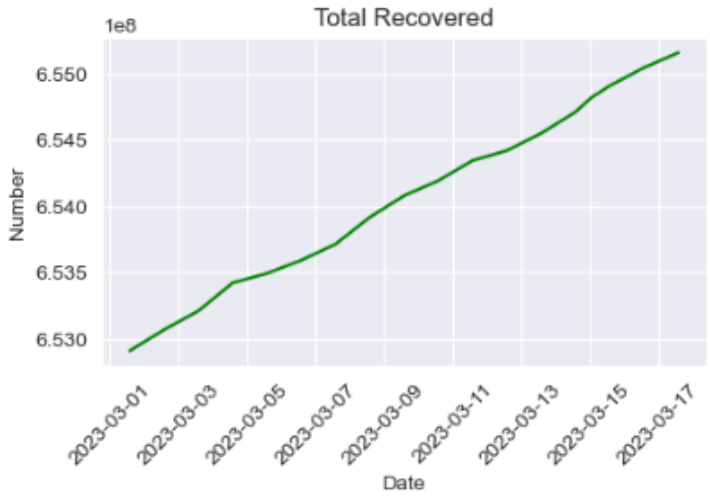
DONUT CHART OF RATIO OF ACTIVE, DEATH AND RECOVERED ON TOTAL CASES AROUND THE WORLD

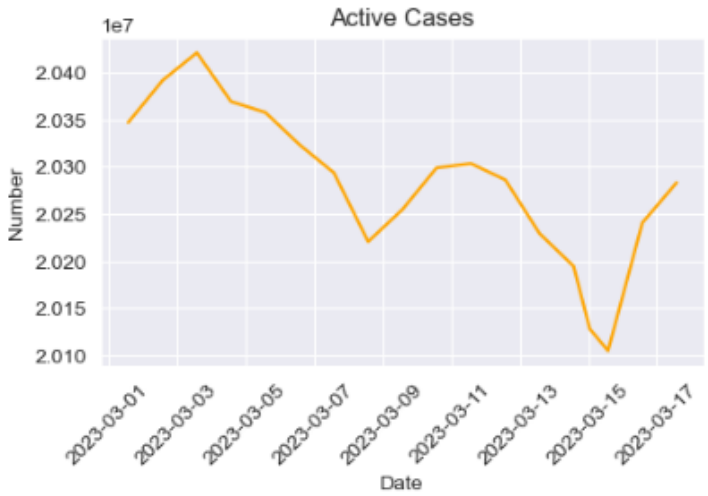


By observations, though spreading rapidly, almost all patients have been recovered from Covid-19, representing 96% of confirmed case, Meanwhile, total death made up only 1% of cumulative cases, indicating a low mortality risk. The current active cases only comprised 3% of the world total cases, equivalent to more than 20 millions people worldwide still infected by Coronavirus but not yet recovered.
The following analysis would dig deep into these figures.

LINE CHART OF THE TOTAL CASES, TOTAL DEATHS, TOTAL RECOVERED, ACTIVE CASES IN THE WORLD DURING THE RESEARCHING TIME







The reason why line chart is chosen for visualization in this section:

- The type of the dataset (Total Cases, Total Deaths, Total Recovered, Active Cases) is numeric.
- The dataset is constantly changing over time (data of 14 latest days is collected).
- We want to display the trend of covid status.

The time plots for the cumulative cases, deaths, recovered respectively shows a steady rise over 17 observation days. This trend matches up our expectation since the variables are the cumulative cases, which have consistently increased as new confirmed, fatal, and recovered cases have been continuously reported. Specifically,

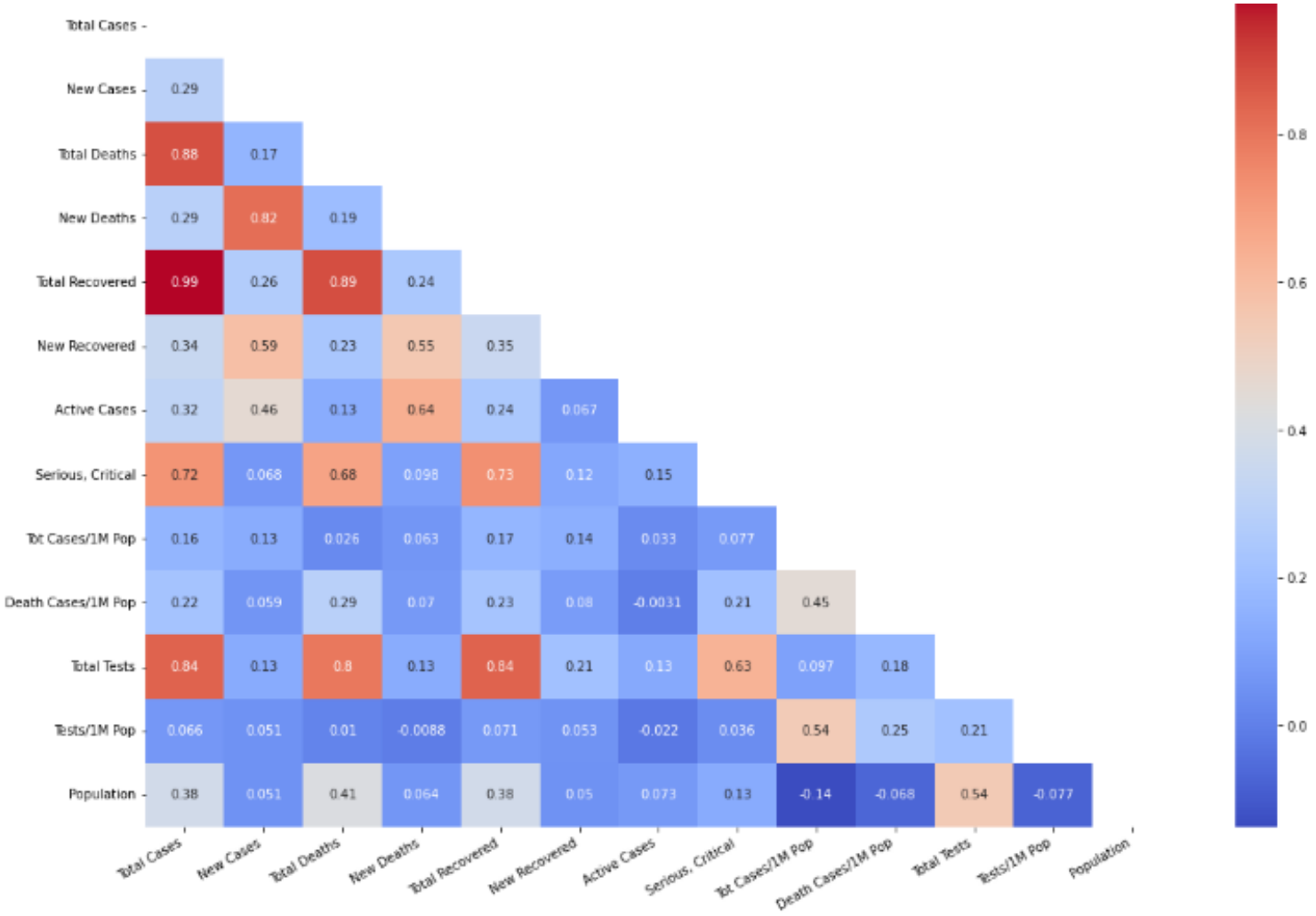
- The first line chart about *Total Cases* in the world exposed an increasing trend over 17 days which means we still can't stop the disease, people still can caught Covid-19 up till now. The disease is still strong enough to make people sick with Covid-19 despite efforts to control its spread.
- The upward trends in the second line chart about *Total Deaths* in the world revealed the disease is still dangerous enough to even kill people. So we can't be subjective and we need to protect ourselves carefully, maintain healthy life by wearing mask.
- The growth in *Total Recovered* (third line) in the world indicated human's capability to control the covid epidemic with better resistance (through vaccines), more thorough knowledge and practices to treat this disease.

However, the above three line chart failed to clearly disclose latest trends. As indicated in the final line chart about *Active Cases,* the trend is not followed any shape over 17 days which means, it is continuously increasing and decreasing which means the epidemic situation is quite complicated over the world. The following line charts would discuss this in-depth

LINE CHART ABOUT THE WORLD'S NEW CASES, NEW DEATHS, NEW RECOVERED IN LATEST 17 DAYS (17 MAR 2023)

The additional (new) confirmed, deaths, and recovered cases' time plots demonstrated fluctuations, suggesting a sporadic rise and fall in case volume per day. Three line chart shared the same nearly the same pattern during 14 days, notably, all three variables peaking on Mar 15 2023, followed by a significant drop in the next two days indicating a positive relationship among those three figures, However, to better detect the seasonal patterns to produce meaningful forecast, a longer time series of data is required.

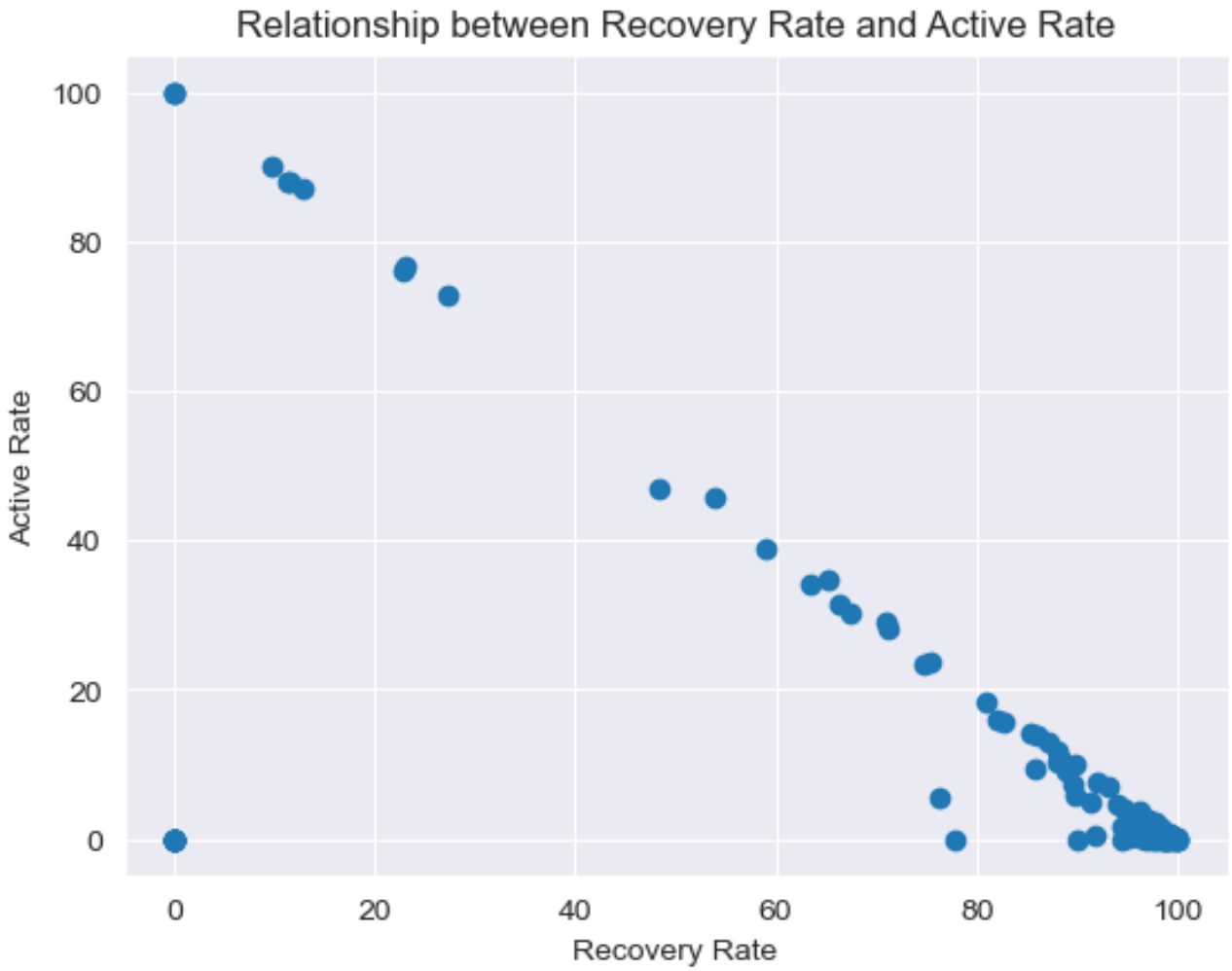
CORRELATION HEATMAP



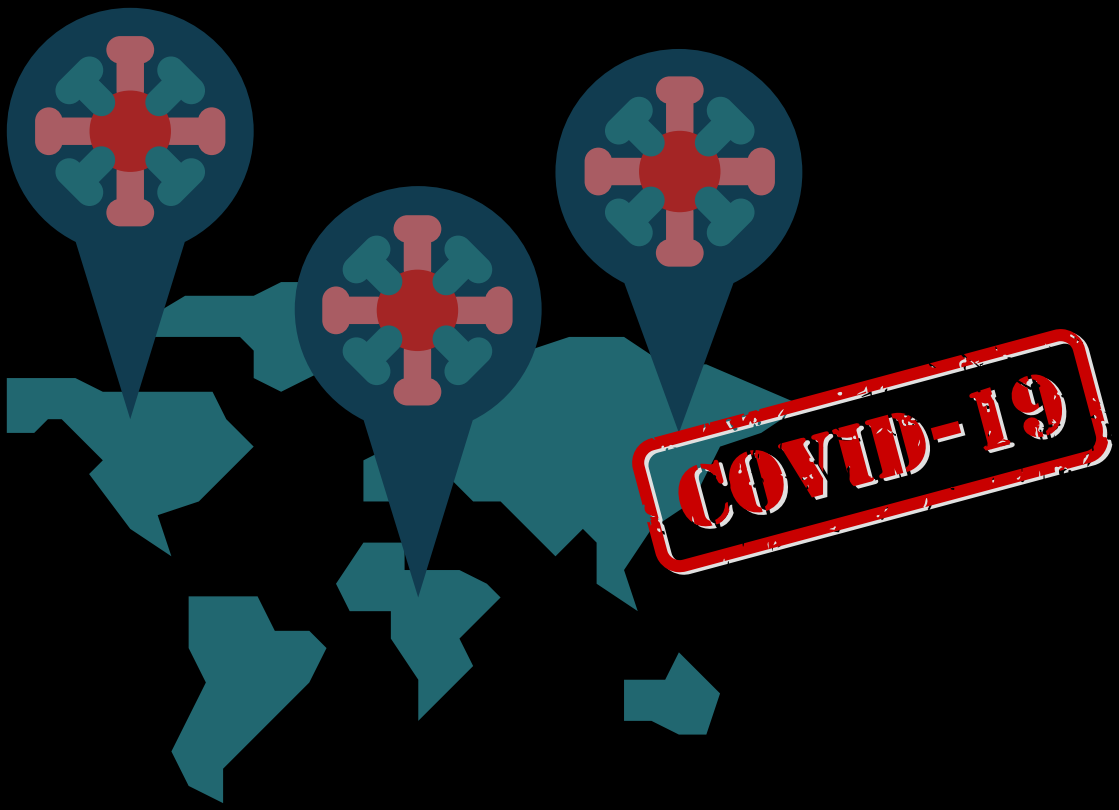
By using Correlation Heatmap, there are some pairs of attributes to really positively correlated to each other. That means these pairs of attributes show strongly about Cause-Effect context.

+ For Example: Total Recovered - Total Cases (0.99), Total Death - Total Cases (0.88),...

Through this correlation matrix, interesting insights among those pairs will be derived in the following parts.

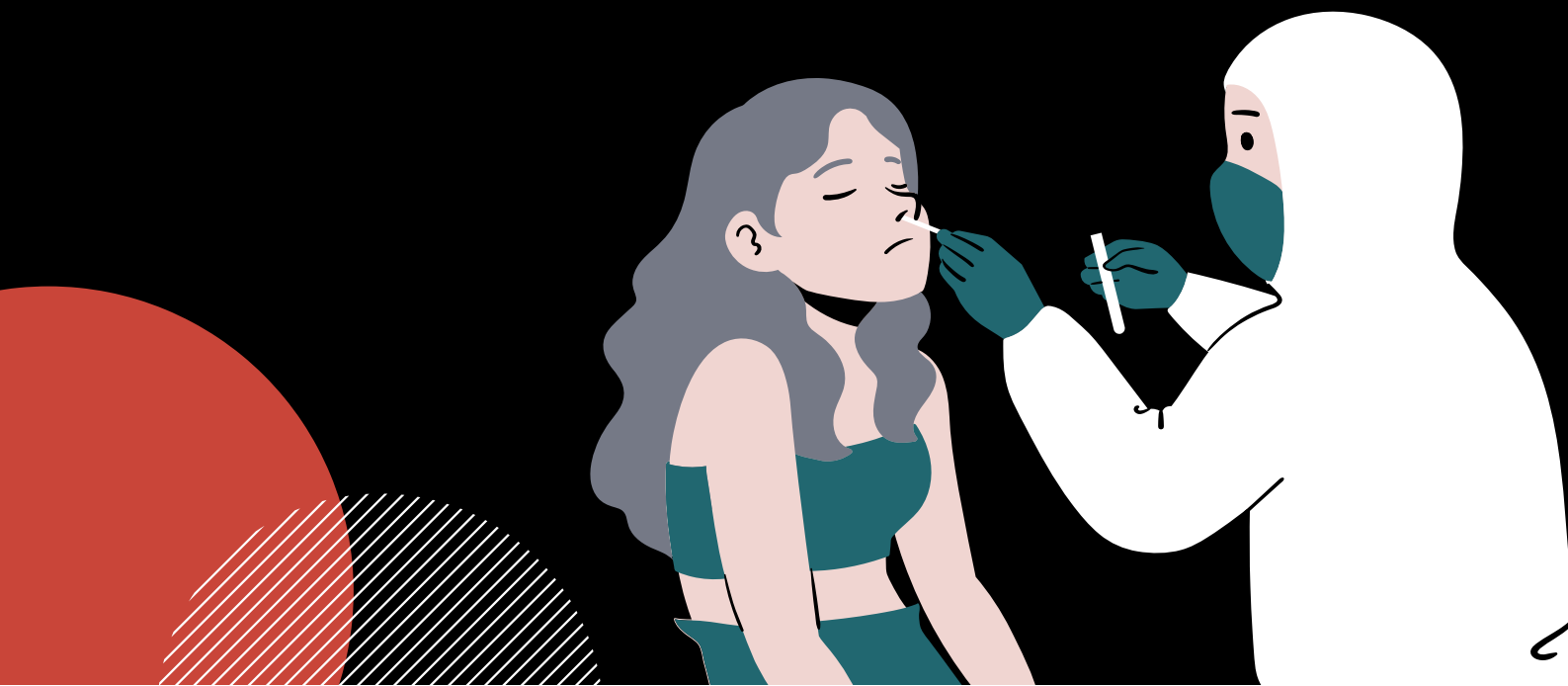


This scatter plot has emphasized on the cause-effect relationship between two variables: Active rate and Recovery rate.
Country with high recovery rate should have employed appropriate policies in combating Covid-19 pandemic with either high healthcare quality or high vaccine coverage rate and so on. Hence, they could utilize these practices to maintain a low active rate.
Adversely, countries with current high active rate meaning that they have not figured out the way to control the pandemic which lack capabilities to treat Covid-19 disease or only recently impacted by Coronavirus. Without a proper healthcare level and experience to control Covid, the recovery rate are accordingly low.
Hence, the cause-effect relation is clear in Active rate-Recovery rate pair.

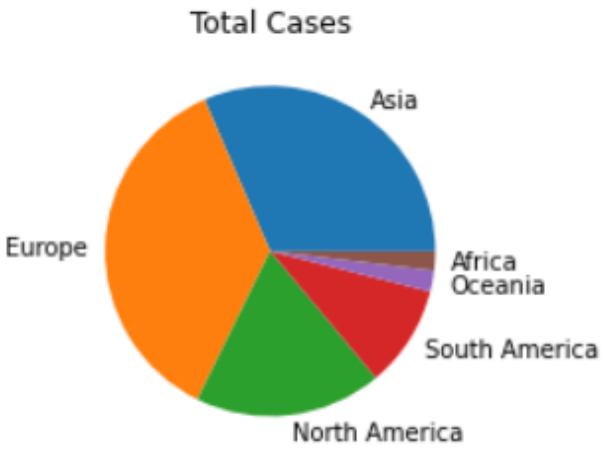


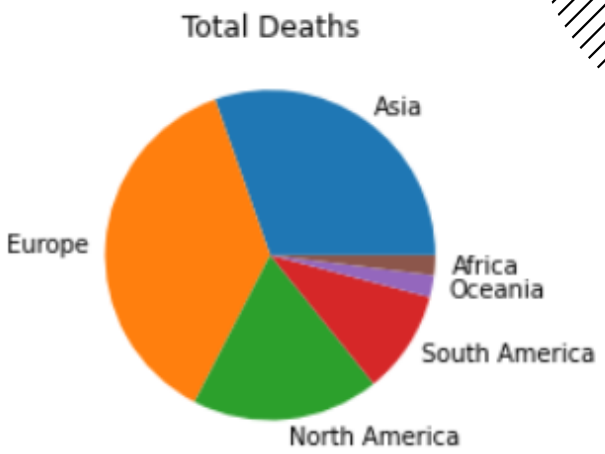
Part 2

COMPARISONS AMONG CONTINENTS



TOTAL CASES, TOTAL DEATHS, TOTAL RECOVERED AND ACTIVE CASES COMPARISON BY CONTINENTS



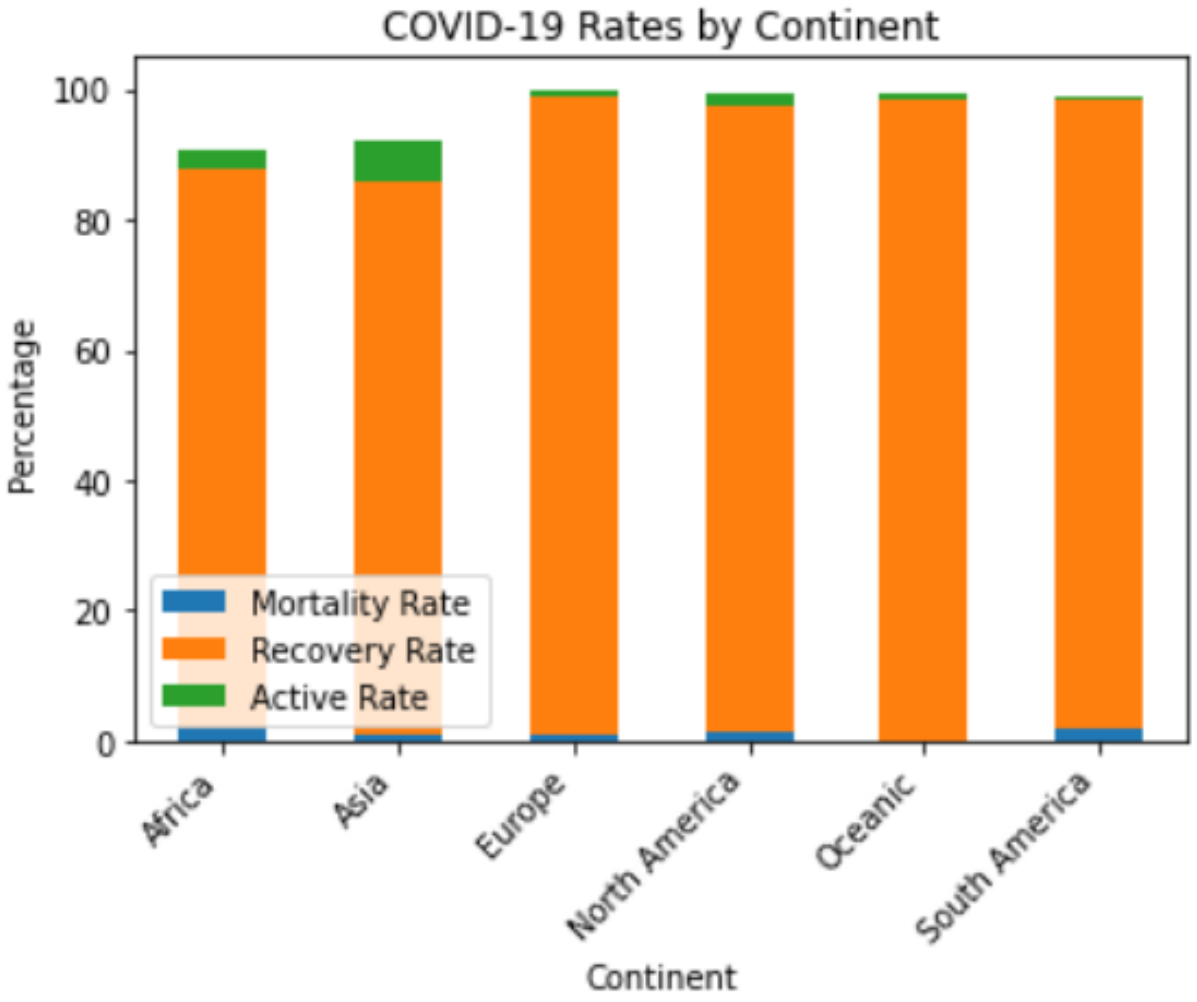


The reason why pie chart is chosen for visualization in this sections:

- The type of the dataset (Total Cases, Total Deaths, Total Recovered, Active Cases) is numeric.
- We want to compare the effect of one factor on different categories.
- The maximum categories in this question is 6 which are 6 continents.

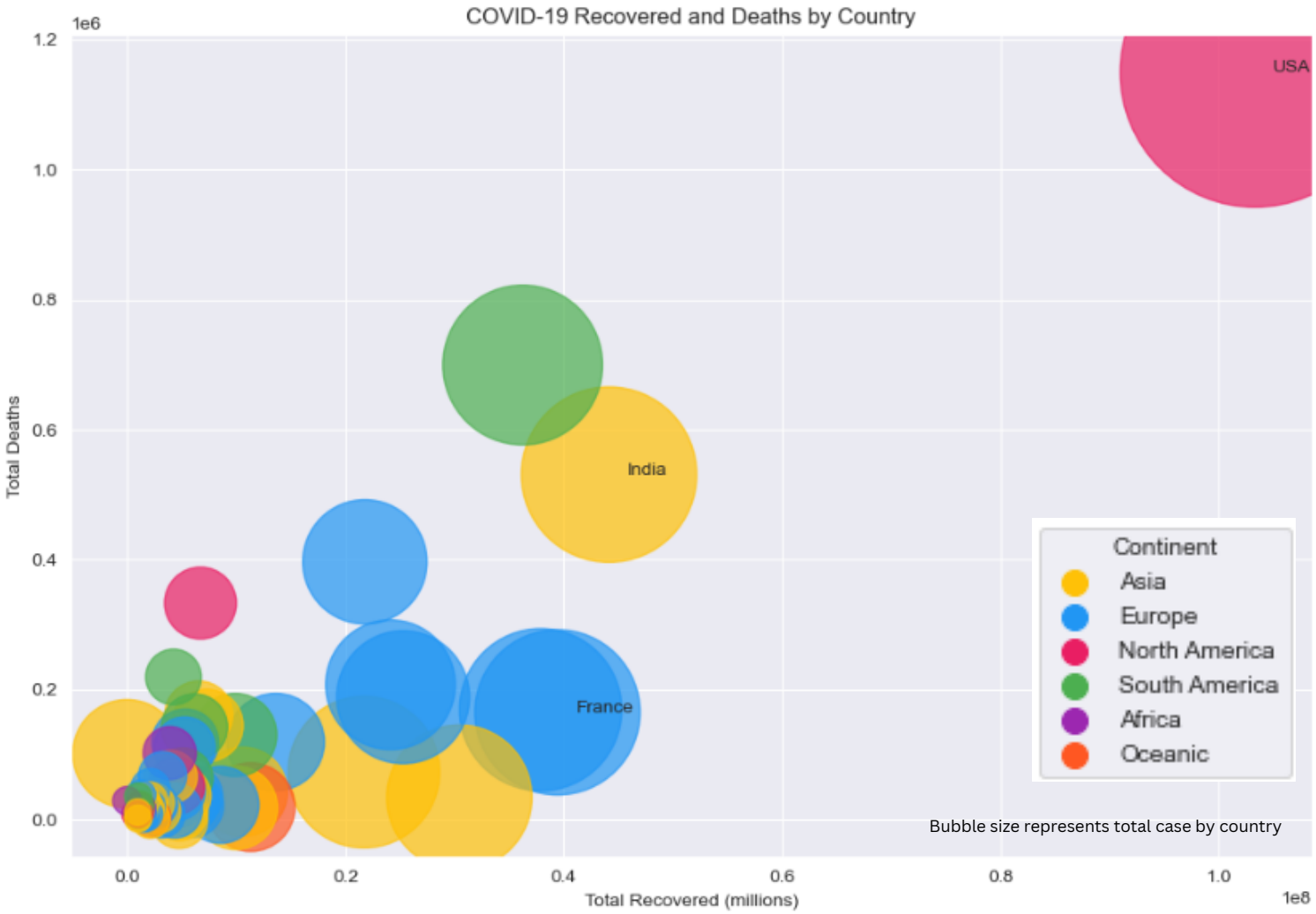
Through these 4 pie charts, we can see that the impact of covid disease on Asia and Europe is quite serious. These 2 continents account 2 large parts in the chart. The Europe has not only the most Total Cases in the world but also the most Total Deaths and the most Total Recovered.

- Europe, NA, SA, Asia are 4 continents that account 4 large parts about Total Recovered maybe because of having a good medical background.
- Asia has the largest part in the Active Cases because of Japan and South Korea are still having tens of thousands of people are infected every day.
- Africa and Oceania are 2 continents that are least been affected by the Covid-19.



From the chart above,

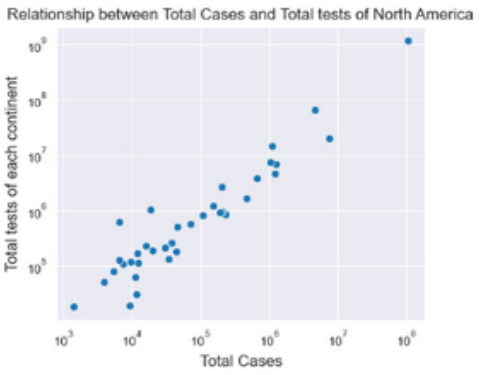
- Asia has the highest Active Rate right now, this mean Covid-19 in other continents except Asia is under control.
- Highest Mortality Rate are Africa and South America probably due to the fact that most people of these continents live in poverty and do not have access to health care.
- Europe has largest recovery rate as they are mostly in high-income status with better medical facilities and early access to Covid-19 vaccine, hence, they could guarantee both low mortality and active rate till now.

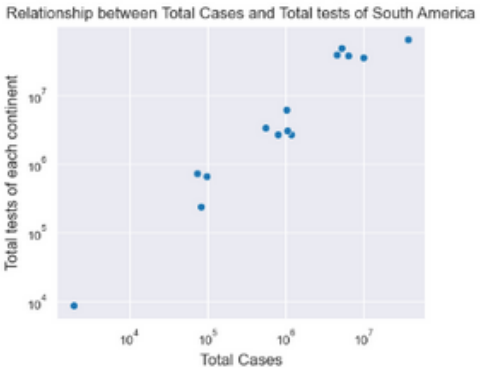


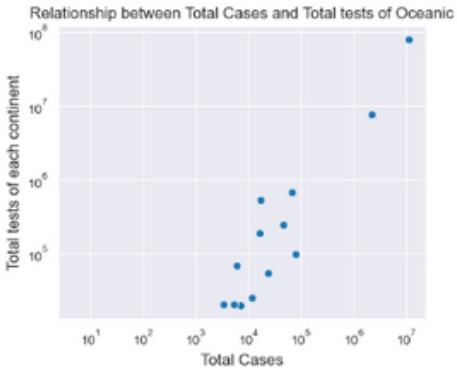
This bubble plot shows the relationship between the total number of COVID-19 recoverd and deaths for countries with at least 1 million cases. The size of the bubbles corresponds to the total number of cases, while the color of the bubbles corresponds to the continent that each country belongs to.

As we can see from this bubble plot, USA (North America), India (Asia) and France (Europe) have the biggest bubble size that means Covid spreads all over the world.
Furthermore, we can only detect few purple and orange bubbles, also, the size of it is quite small which prove that total case of Covid-19 in Africa and Oceanic continent is low (compared to other continents)
We can also explore this bubble plot that countries with high total cases also have the high total recover, besides that, total deaths increase lightly, too.
Therefore, we can find out that Asia, Europe and America need to have more preparation in health care.

Relationship between Total cases and total tests by continents





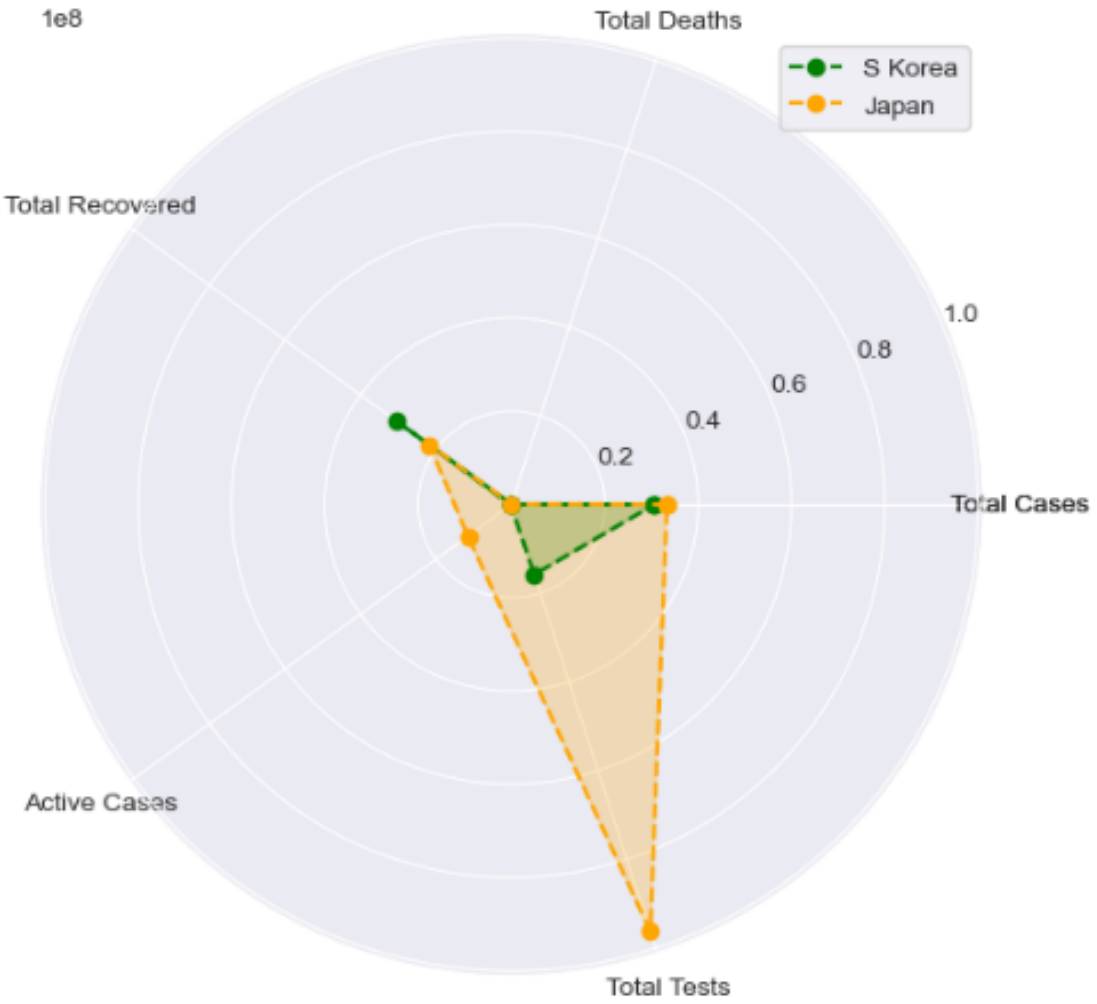


Surprisingly, all 6 continents have the linear relationship between Total Cases and Total Test. It may be a cause - effect relationship, because when the total cases increases which mean this case was confirmed by test, so the it also makes the number of tests rise.

Part 3

Insights into:
COUNTRY CHARACTERISTICS ASSOCIATED WITH COVID-19 STATISTICS

Radar Chart for South Korea and Japan Coronavirus Figures



The reason why we chose radar chart for visualization in this question:

- The type of the dataset (Total Cases, Total Deaths, Total Recovered, Active Cases, Total Tests) is numeric.
- We want to compare the statistics between 2 countries which still having thousands, tens of thousands of people are infected every day.

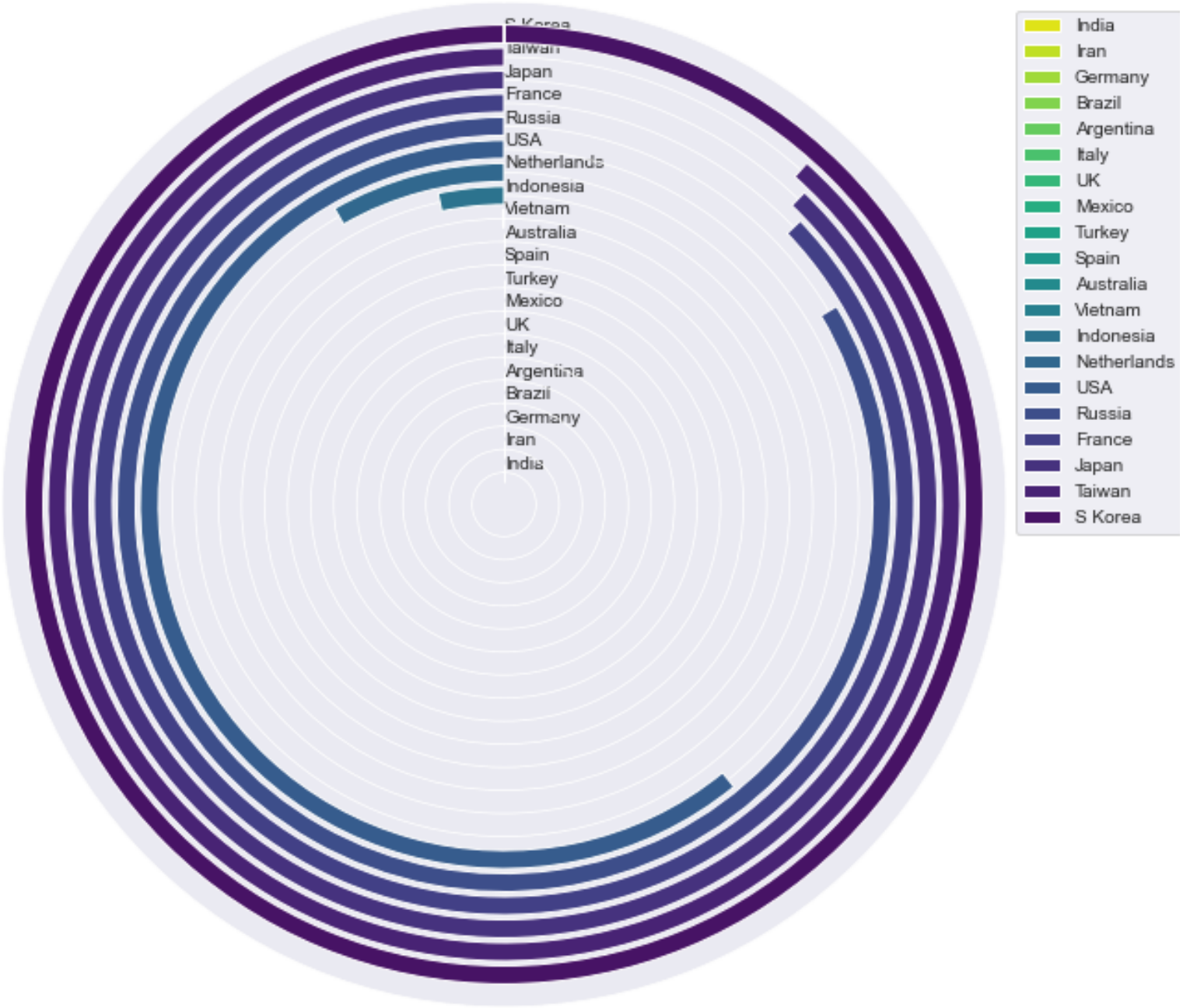
Derived from the chart above, despite having an approximately equal number of total cases and recording tens thousands of new confirmed daily case, they did not share the same patterns in total test, total recovered and active cases.

- Despite total test of Japan being five times more than that of South Korea, the active cases in South Korea are superior Japanese's one. Hence, test more could not guarantee the efficiency of combating this pandemic.
- With a significantly huge gap in active cases, and total recovered given the nearly the same total case and new daily case as mentioned, it is concluded that South Korea has done better than Japan in treating and controlling Covid-19 disease.

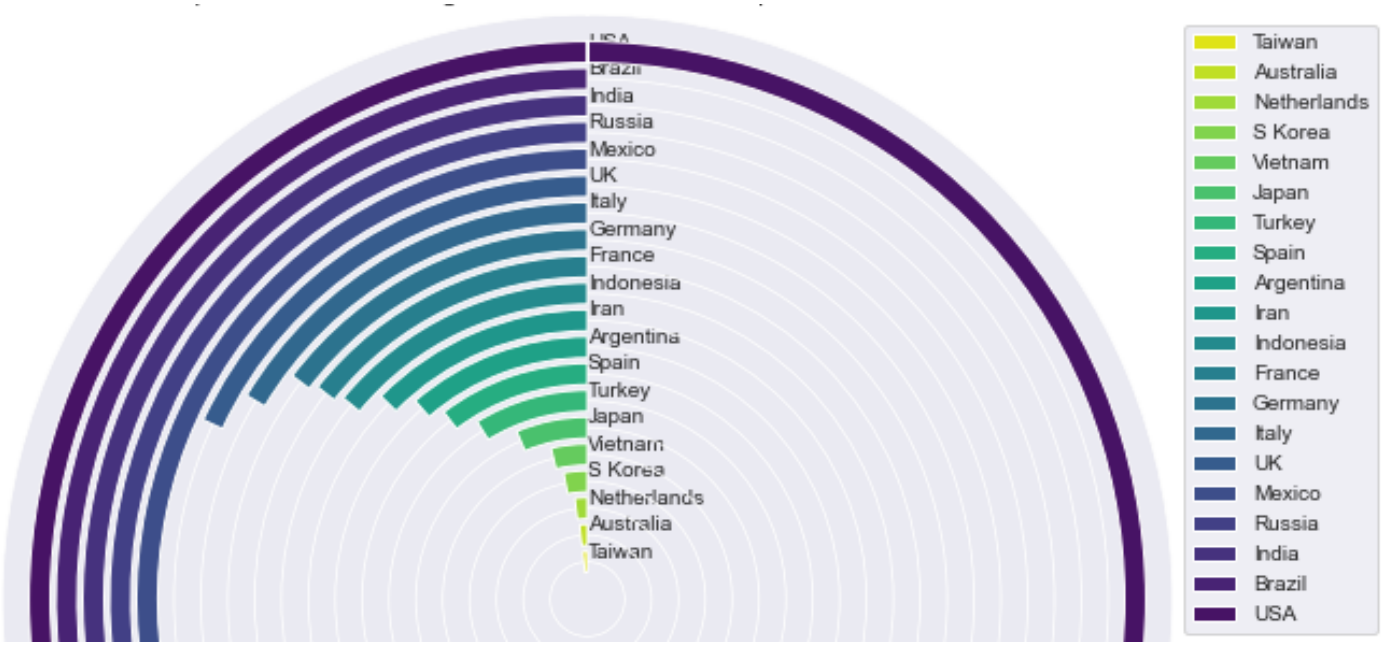
New Cases And New Recovered By Top 4 Highest New Cases Countries

- These are 4 top countries with highest indicate of New Cases.
- The Recovered Cases are really much more than New Cases that makes these countries facing a big trouble of how to deal with increasing number of patients every day.
- Especially Japan, New Recovered Cases are dominant to New Cases, Japanese Government should have some solutions to scale down the number of New Cases.

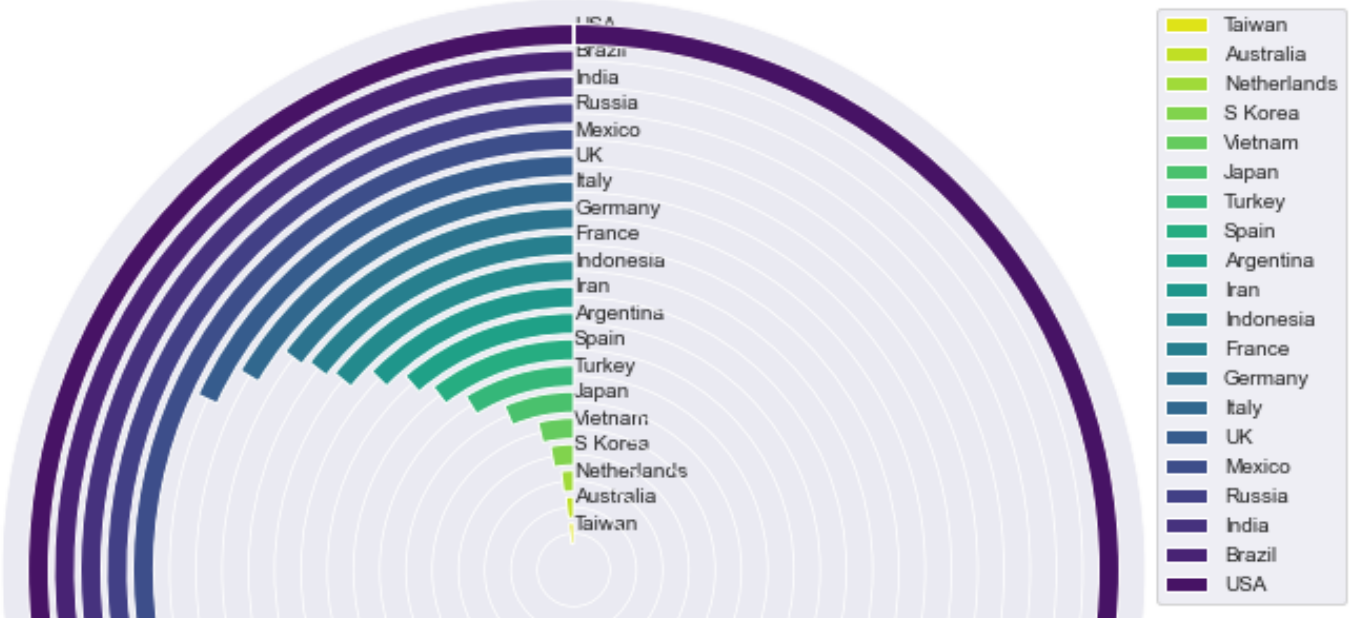
Top 20 Countries with highest New Cases updated on 15/03/2023



Top 20 Countries with highest Total Deaths Cases updated on 06/03/2023



Top 20 Countries with highest Total Deaths Cases updated on 06/03/2023

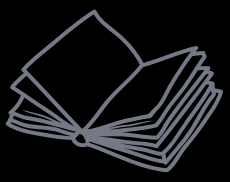


- Following 3 Circular bar charts: USA shows they are destructively impacted by Corona Virus as we can see they always lead in Total Cases and Death Cases as outlier to other nations
- In 15/03/2023, there are still some countries have really high indicate in New Cases like South Korea, Japan, Russia, USA, France which are some of top Power Countries

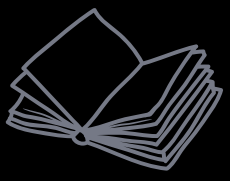
REFERENCES

Lecture slide by Dr. Bui Tien Len

Theory lesson every Friday of Dr. Bui Tien Len



Visualization Analysis & Design - Tamara Munzner (2014)



The Big Book of DashBoards

Coronavirus Statistics Worldometers (2023)

IT'S THE END

DATA RELATIONSHIP