

Report

Group:

CC - L21 - G3

Member:

510239417, ynie7952

520400070, scao2237

510248073, yzhe5295

520066234, zzha8989

What are the potential factors that affect the COVID-19 death rate?

Part A:

Overview

The aim of this project is to explore the potential factors that may affect COVID-19 death, and we put forward four factors that may have an impact on COVID-19 which is climate, vaccination, GDP and the population. The stakeholder of this project is the person who is interested in the factors that may affect COVID-19 death. To explore this topic, there are five datasets that we are focusing on that can make contributions to this question. The first dataset is the COVID-19 death population in a variety of counties, the second dataset is the latitude of different areas which is more related to the climate, the third one is the global vaccination that contains different countries' vaccination, the fourth one is the global economy which has the different countries GDP which is a significant factor that we are going to analyze and the last one contains different countries' population which is an important factor as well.

Metadata

The original data is from Kaggle.com. All of these datasets are authorized to be used. There are four separate datasets related to the following topic: global covid report, global latitude and longitude data, global vaccination, global economy and global population.

Global covid report:

<https://www.kaggle.com/datasets/imdevskp/corona-virus-report>

Latitude and Longitude: <https://www.kaggle.com/datasets/paultimothymooney/latitude-and-longitude-for-every-country-and-state>

Global Vaccination:

<https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>

Global economy:

<https://www.kaggle.com/datasets/shashwatwork/impact-of-covid19-pandemic-on-the-global-economy>

Global population:

<https://www.kaggle.com/datasets/tanuprabhu/population-by-country-2020/code>

The Relationship of GDP CAP and Covid-19 Death Cases

Dataset Description

This report aims to analyze the relationship between the GDP CAP and Covid-19 death cases and express a visualisation. The original data is downloaded from Kaggle.com. The datasets analyzed in this report are [Covid-19 Dataset, Impact of Covid-19 Pandemic on the Global Economy](#), and the integrated dataset of these two datasets from Stage 1.

There are 6 columns in the final integrated dataset:

- Country
- GDP CAP: *The GDP per capita for each country.*
- Deaths
- Deaths / 100 Cases
- Deaths / 100 Recovered
- WHO Region: *The World Health Organization divided the world into six WHO Regions for administration.*

Grouped-aggregate Summary

Import Pandas to proceed with the group aggregation for the dataset, using the below codes to aggregate the sum of death cases in different WHO Regions.

```
gb1 = data.groupby('WHO Region').Deaths.sum().to_frame(name = "Total Death  
Population").reset_index()
```

Aggregate the mean of GDP CAP in different WHO Regions using the following code, and I am able to produce a table shown below.

```
gb2 = data.groupby('WHO Region').GDPCAP.mean().to_frame(name = "Mean of  
GDPCAP").reset_index()
```

In Table1, it shows that the Americas is the region with the largest total death population, followed by Europe with the second largest death population. Western Pacific has the lowest death population.

Table 2 displays that Europe has the highest mean GDP CAP, while Africa has the lowest GDP CAP. Americas, Eastern Mediterranean, and the Western Pacific have similar mean GDP CAP.

WHO Region	Total Death Population
Africa	11904
Americas	194634
Eastern Mediterranean	38128
Europe	166294
South-East Asia	41343
Western Pacific	8242

[Table 1]

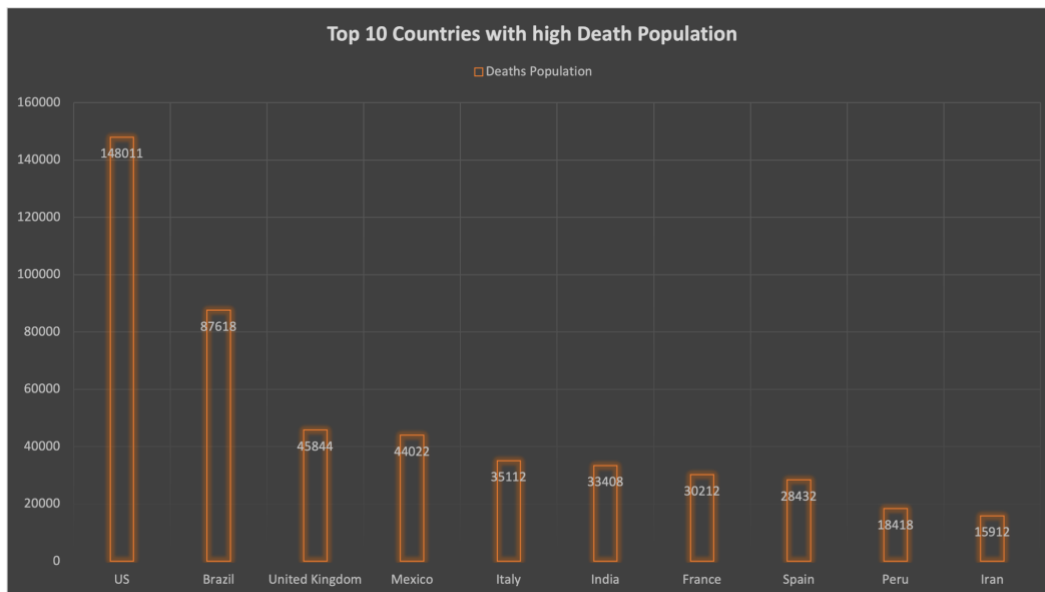
WHO Region	Mean of GDPCAP
Africa	7.960111
Americas	9.433212
Eastern Mediterranean	9.517796
Europe	10.053372
South-East Asia	8.977676
Western Pacific	9.730678

[Table 2]

Visualization

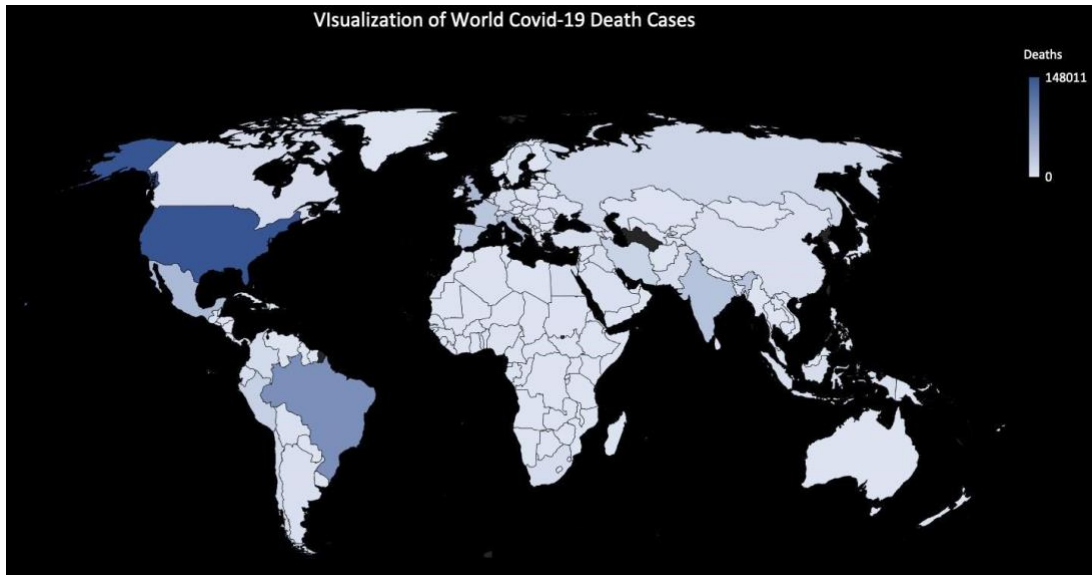
Death Case

To visualize the death case (numerical variable) and country (categorical variable), a barplot from Excel is used. I sort the data by death population in descending order. Thus, I am able to see the countries with the highest death population and apply them to the bar plot. (Figure 1)



[Figure 1]

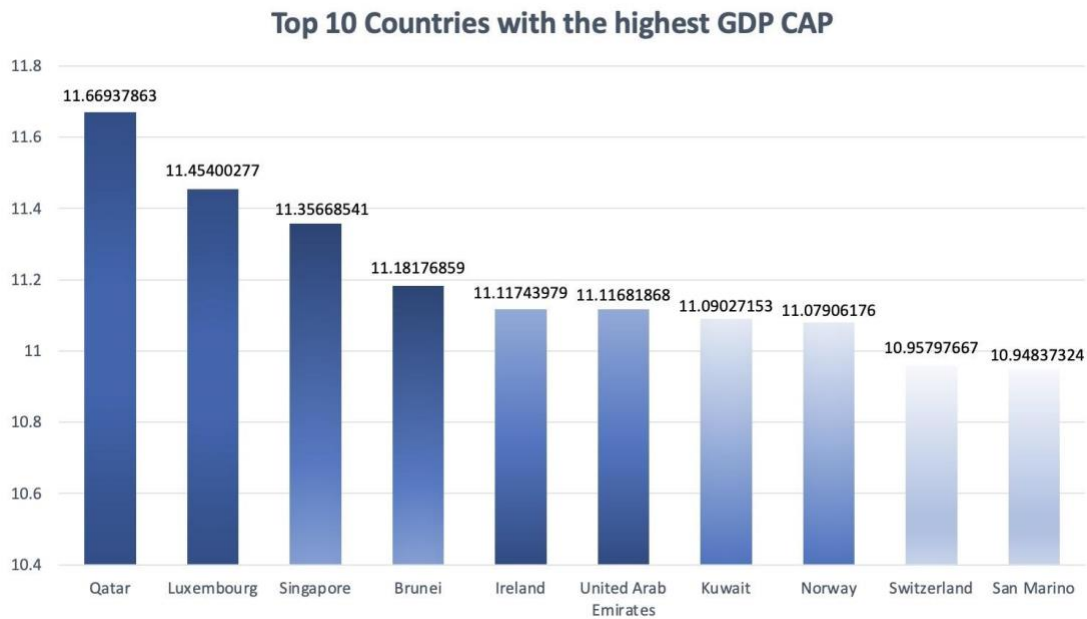
Besides that, I also use the map template in Excel to visualize the death cases on the world map(*Figure 2*). Even though this is not a chart to estimate the relationship among data, it provides a direct visual expression. This diagram shows that the region with the highest number of death cases concentrates in the United States, and Brazil.



[Figure2]

GDP CAP

Using the same method as visualising the death cases, I also use the barplot in Excel. I sort the GDP in descending order to find the top ten countries with the highest GDP values.



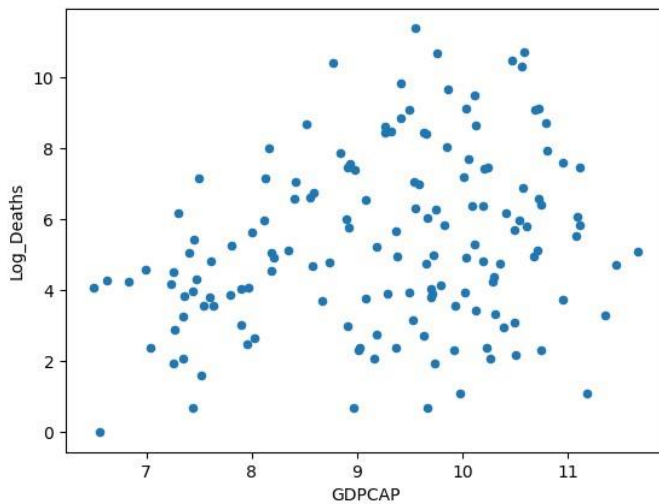
[Figure 3]

The relationship between Death cases and GDP CAP

Death cases and GDP CAP are both numerical variables. Thus, I use the scatter plot to express the relationship between these two variables visually.

However, the number of death cases is much larger than the GDP CAP and it caused the problem shown visually in the graph. The large values skew the graph. To solve this problem, I use a logarithmic scale for death case data. The code and graph shown as below:

```
data['Log_Deaths'] = np.log(data['Deaths'])
data.plot(kind = 'scatter', x = 'GDPCAP', y = 'Log_Deaths')
plt.show()
```



It seems that there is a positive trend between GDP CAP values and death cases through this scatter plot. To have a further analysis of the correlation between these two variables, I use the code `data.corr()`

[Figure 4]

The output is shown below:

	GDPCAP	Deaths	Deaths / 100 Cases	Deaths / 100 Recovered	Log_Deaths
GDPCAP	1.000000	0.143782	0.068667	0.135235	0.268158

The correlation between the GDPCAP and the Log_Deaths is 0.268 and it is considered a weak correlation. In addition, the correlation between GDPCAP and other death-related data is not strong as well.

In conclusion, the GDP CAP and the Covid-19 death cases have a weak correlation, and there is no obvious direct relationship between these two variables.

Unikey: zzha8989

Dataset Description

There are four columns in the final integrated dataset:

- Country
- Deaths
- People_vaccinated ● Population2021

Summary

```

merge1.csv
country,Confirmed,Deaths,Recovered,total_vaccinations,CODE,GDPcap,latitude,people_vaccinated
Afghanistan,36263,1269,25198,5751015,0,AFG,7.497754494,33.93911,5082824.0
Albania,4880,144,2745,2754244,0,ALB,9.376145531,41.153332,1278902.0
Algeria,27973,1163,18837,13704895,0,DZA,9.540639325,28.033886,7461932.0
Angola,950,41,242,17535411,0,AGO,8.668968767,-11.202692,11235059.0
Antigua and Barbuda,96,2,55,125206,0,ATG,9.97538687,17.060816,63836.0
Argentina,167416,3,Col 4: Recovered,0,ARG,9.848709615,-38.416097,49007186.0
Armenia,37390,711,26665,2088962,0,ARM,9.08109464,40.069999,1113472.0
Australia,15303,167,9311,56242913,0,AUS,10.70658069,-25.274398,22202366.0
Austria,20558,713,18246,18131115,0,AUT,10.72407512,47.516231,6812569.0
Bahrain,30446,423,23242,13425932,0,AZE,9.67061927,40.143105,5322427.0
Bahamas,382,11,91,334155,0,BHS,10.22983178,25.03428,164961.0
Bahrain,39482,141,36110,3421273,0,BHR,10.67569323,25.930414,1232740.0
Bangladesh,226225,2965,125683,243642749,0,BGD,8.167347447,23.684994,127544055.0

```

From the original data file. The data set has a summary of vaccinations and deaths. However, the vaccination rate cannot be calculated. So I found another new dataset containing the 2021 world population. And merge the two datasets. This dataset includes world population records from 1960 to 2021. The file name is

'World-population-by-countries-dataset.csv.'

```
df2 = df2.rename(columns = {'Country Name' : 'country'})
result = pd.merge(df1, df3, how = "inner")
merge = result.to_csv("merge1.csv", index = False)
result2 = pd.merge(result, df2, how = "inner")
result2 = result2.rename(columns = {'2021' : 'population2021'})
result2.to_csv("merge2.csv", index = False)
```

This code uses pandas' merge function. The first is to change the title's name to match the name of the country. For example, the file name is merge2.csv.

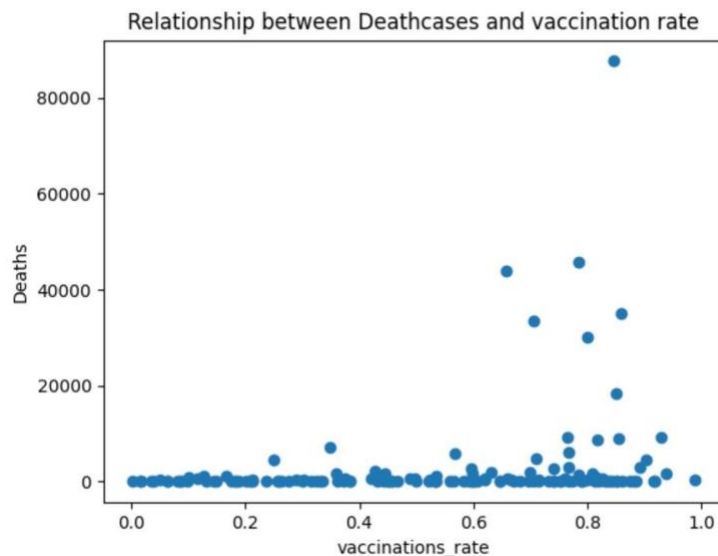
```
df = pd.read_csv ("merge2.csv")
vaccination_rate = df['people_vaccinated']/df['population2021']
deaths = df['Deaths']
```

The vaccination rate is calculated using the number of vaccinations per country divided by the total population of each country.

Visualization

```
plt.scatter(vaccination_rate, deaths)
plt.title("Relationship between Deathcases and vaccination rate")
plt.xlabel('vaccinations_rate')
plt.ylabel('Deaths')
plt.savefig("leo.png")
```

I used matplotlib's pyplot to draw my scatterplot. The ordinate is "deaths," and the abscissa is "vaccination_rate", The output photo is as shown.



Are vaccinations directly linked to COVID-19 deaths? My first hypothesis is that the higher the vaccination rate, the lower the mortality rate. To test this conjecture, I make the following analysis. According to the chart, vaccination rates did not directly affect COVID-19 mortality. The more profound implication is that high vaccination rates do not reduce mortality. But the higher the vaccination rate, the higher the mortality rate. There are no direct factors to test this

view. The data were not comprehensive enough to prove that the deaths were caused by vaccination. But increased vaccination rates will not directly affect COVID-19 mortality.

country	Deaths	people_vaccinated	population2021
Afghanistan	1269	5082824	39835428
Albania	144	1278902	2811666
Algeria	1163	7461932	44616626
Angola	41	11235059	33933611
Antigua and Barbuda	3	63836	98728
Argentina	3059	40907186	45808747
Armenia	711	1113472	2968128
Australia	167	22202366	25739256
Austria	713	6812569	8956279
Azerbaijan	423	5322427	10145212
Bahrain	141	1232740	1748295
Bangladesh	2965	127544055	166303494
Barbados	7	161076	287708
Belarus	538	5801653	9340314
Belize	2	236015	404915
Benin	35	2955274	12451031
Bhutan	0	687549	779900
Bolivia	2647	7043449	11832936
Bosnia and Herzegovina	294	943394	3263459
Botswana	2	1454775	2397240
Brazil	87618	181078067	213993441
Bulgaria	347	2081276	6899125

The table shows that COVID-19 vaccination does not directly affect mortality.

China	4656	1275541000	1412360000
-------	------	------------	------------

Especially in the data from this, China has many vaccinations, but the number of deaths is only 4,656. However, this data is not comprehensive because each country has different vaccinations. It will affect effectiveness. The second is the existence of variants of COVID-19. This resulted in varying lethality rates. Too many variable factors lead to this data, which cannot be regarded as very reliable data. The following two figures illustrate this phenomenon.

country	Deaths	country	people_vaccinated
Brazil	87618	Burundi	10372
United Kingdom	45844	San Marino	26331
Mexico	44022	Dominica	32403
Italy	35112	Grenada	43292
India	33408	Antigua and Barbuda	63836
France	30212	Seychelles	84327
Peru	18418	Djibouti	145050
Chile	9187	Barbados	161076
Germany	9125	Haiti	163710

The table above shows the left is the top 10 countries with the new crown mortality rate. The one on the proper ranks the countries with the least vaccination coverage. is not directly related to mortality.

The relationship between latitude and covid-19 Death Cases

Introduction

As for our group topic, which factors affect the number of deaths in the COVID-19 pandemic, I would like to talk about this topic from the perspective of the geographical location-latitude of the countries of the world to analyse the relationship between latitude and death cases. The original data is downloaded from Kaggle.com. The datasets analysed in this report are the world country latitude Dataset, the COVID-19 dataset, and the integrated dataset of these two datasets from Stage 1.

Data cleaning

Initially, I extracted the death count and latitude columns from the overall database which merged data from our stage 1 and tabulated them to explore the connection between the two.

Data cleaning

Initially, I extracted the death count and latitude columns from the overall database which merged data from our stage 1 and tabulated them to explore the connection between the two.

```
import pandas as pd
a = pd.read_csv(r"C:\Users\HP\OneDrive\桌面\DATA1002\scao2237_stage2\raw data.csv")
choose_data = a[['country', 'Deaths', 'latitude']]
print(choose_data)
choose_data.to_csv("Final.csv", index=False)
```

Armenia	711	40.069099
---------	-----	-----------

```
import pandas as pd
df=pd.read_csv(r"C:\Users\HP\OneDrive\桌面\DATA1002\scao2237_stage2\final.csv")
df.to_html("Table_stage.htm")
html_file = df.to_html()
```

Belarus	538	53.709807
---------	-----	-----------

As for the large size of the table, I only showed part of the table

data. From this data table, I can preliminarily judge that most of the countries with collected data are located between the middle and low latitudes of the earth.

Visualization

Next, I take the merged dataset from our group project stage 1 to plot the relationship between the number of deaths and the latitude of the countries of the world. As the data points are mostly concentrated below 20000, I change the range of the y-axis to make the charts clear.

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv(r"C:\Users\HP\OneDrive\桌面\scao2237_stage2\final.csv")
data.plot(kind='scatter', x='latitude', y='Deaths')
plt.show()
```

From the three scatter plots, the positive

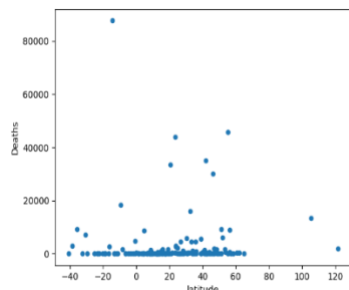


Figure 1

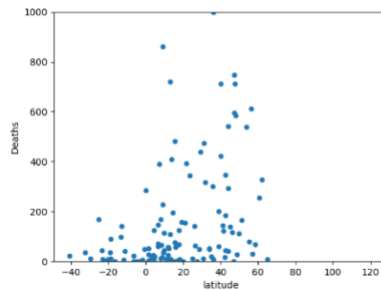


Figure 2

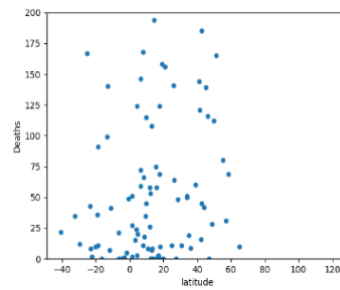


Figure 3

value of

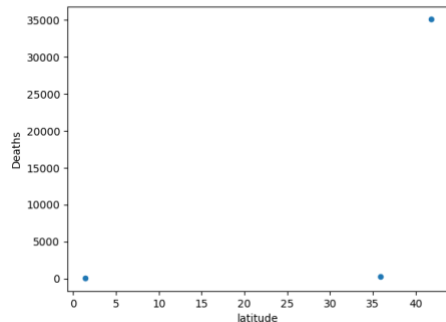
the latitudes means the country is in the north of the equator and the negative latitudes mean the country is in the south of the equator. And I can see that the points are mostly located between the middle and low latitudes of the Earth and the number of deaths is concentrated below 200. But specifically, by looking at the location of the data points in the graph, I find that the number of deaths in the countries near the equator, where the x-axis is zero, is generally lower than the number of deaths in the countries far from the equator. This also suggests that the number of deaths may be related to regional climatic conditions, as countries are in different latitudes and geographically.

Data analysis

However, to analyze the relationship between the latitudes and death cases from a more specific perspective, I extracted three countries from the database, which are Singapore, Italy, and South Korea. Singapore is in the low-altitude region, and Italy and South Korea are in the middle-latitude range. The following is the table of the corresponding data of the three countries I extracted from the dataset.

	country	Confirmed	Deaths	latitude
0	Italy	246286	35112	41.871940
1	Singapore	50838	27	1.352083
2	South Korea	14203	300	35.907757

In this analysis of the data, I ignored the effects of population density and other factors which are difficult to eliminate, but I still only looked at the effects of latitude on the number of deaths.



I use the scatter plot to show the relationship between the latitude and deaths of the three selected countries. I can see from this scatter plot that it's more severe in the middle latitude countries as the death cases are much higher compared to the low latitudes. Not only that, if I look at the number of cases, the number of deaths in Singapore is 0.05 percent of the total confirmed cases, but the death rate in Italy is approximate 14 percent of those diagnosed and in South Korea, the death rate is

about 2 percent. This is a straightforward illustration of latitude as a factor determining the large differences between countries in the number of death cases during the COVID-19 pandemic.

Conclusion

In general, the number of deaths in the countries near the equator is generally lower than the number of deaths in the countries far from the equator, which proves that latitude is an important factor in the number of death cases. However, although latitude is directly related to the number of death cases, this conclusion is not absolute. We need to combine the country's economic conditions, population density, vaccination probability, and other factors to make a comprehensive comparison. A single comparison based on latitude is not persuasive.

The relationship between the countries' population and death cases

Dataset Introduction

This report is aimed to explore the relationship between populations and death cases in different countries. There are two datasets that we will use. The first one is the dataset related to the death which is used at Stage 1. The other is the population dataset which is the main factor that we are going to analyse. I merge these two datasets to one new dataset.

The dataset is from Kaggle. The dataset is newly added into this report which has 11 columns and 235 rows. There are variables that I mainly use as following:

Country: This column contains different countries' name

Population: This column contains the population of different countries

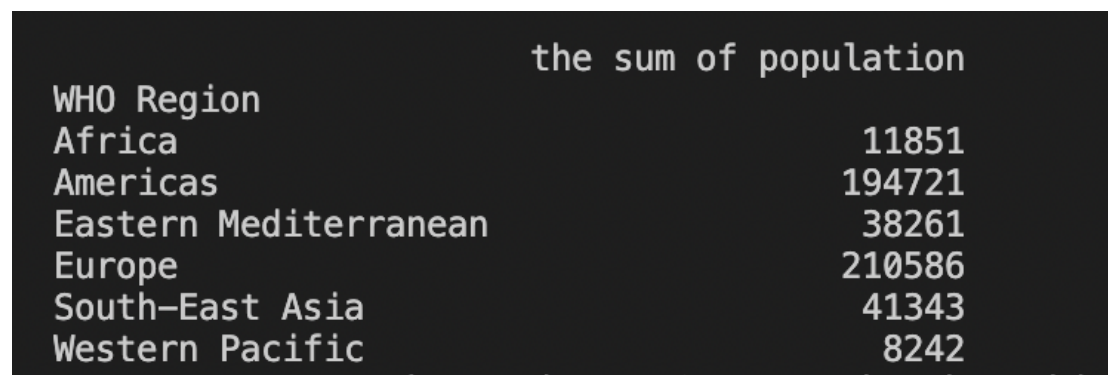
Deaths: This column contains the death of different countries

Confirmed: This column contains the confirmed of different countries

Recovered: This column contains the recovered of different countries

Grouped-aggregation summary

For the grouped aggregation, I choose to check the population in different WHO regions. We get the sum of the population by the WHO region. The table is as follows.



the sum of population	
WHO Region	
Africa	11851
Americas	194721
Eastern Mediterranean	38261
Europe	210586
South-East Asia	41343
Western Pacific	8242

Visualisation

There are many countries in this dataset, we firstly look at the countries that have more populations and the recovered cases, confirmed cases and death cases. In the first figure, I choose to draw a bar plot by using pandas and matplotlib which are helpful for our visualisation. The dataset is sorted by the population value. We firstly create a new data frame according to the population by using pandas. And then we get a bar plot using matplotlib. There are two kinds of

data: the first is qualitative data and the other one is quantitative. That is why we use bar plots here. According to the figure, we find that some highest population countries do not have many death cases, even almost close to zero death cases, not only the death cases, but also, they have low confirmed cases if we observe this figure. However, that is only ten countries, and they all have lots of population. Therefore, that may not be representative for the analysis. We cannot simply conclude that there is no relationship between populations and death cases.

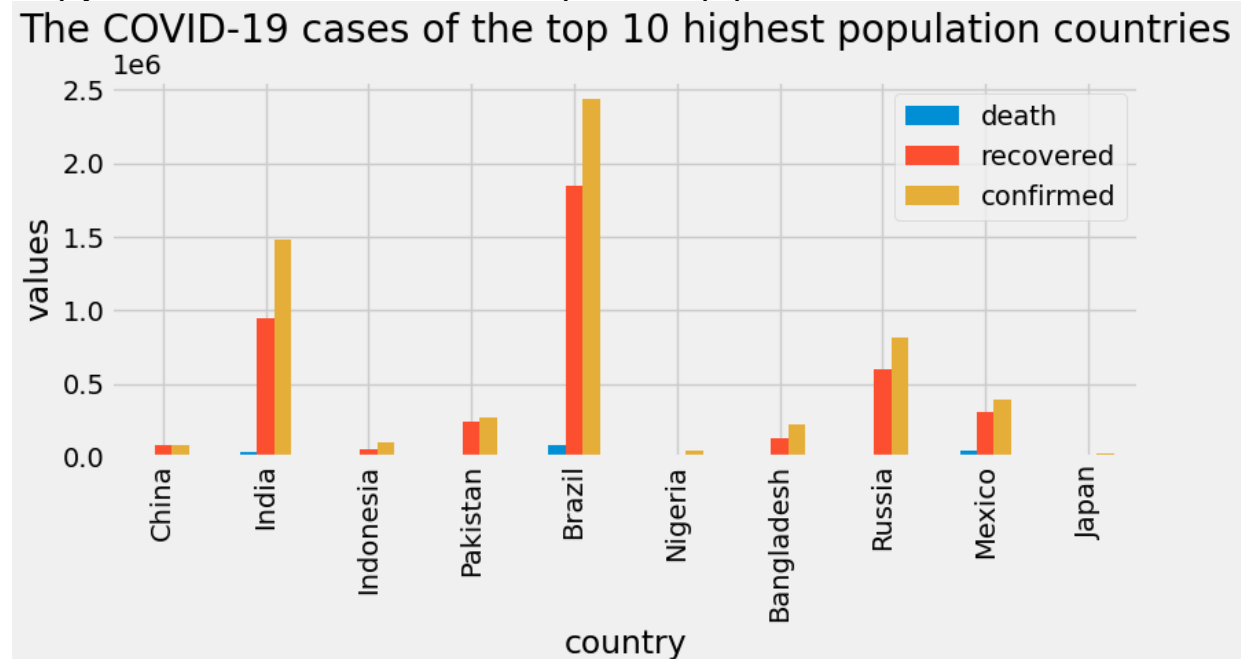


Figure 1

Due to the limitation of the above analysis, I decided to draw another figure that shows the top ten lowest population countries. The figure is shown as follows. According to this figure, it seems that they all have relatively low death cases and low confirmed cases except for San Marino and Iceland. But they all have low death cases. However, we still cannot find an obvious relationship between populations and death cases. Therefore, we got pretty much the same thing in these two figures.

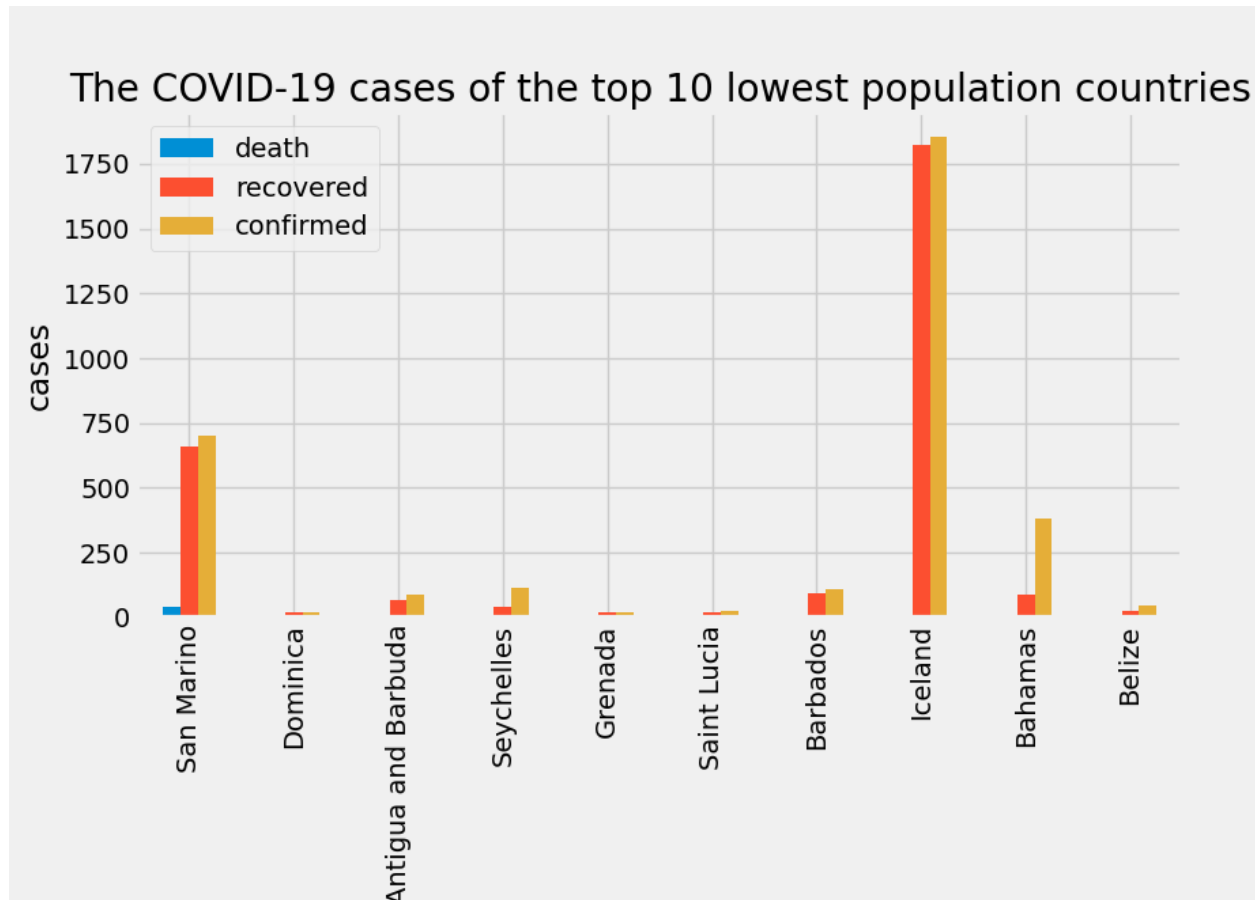


Figure 2

Finally, let us check the scatter plot using the whole countries that we have. The reason why I use scatter plots is that the populations and death cases are both quantitative data. By using scatter plot, we can directly check the relationship between these two factors.

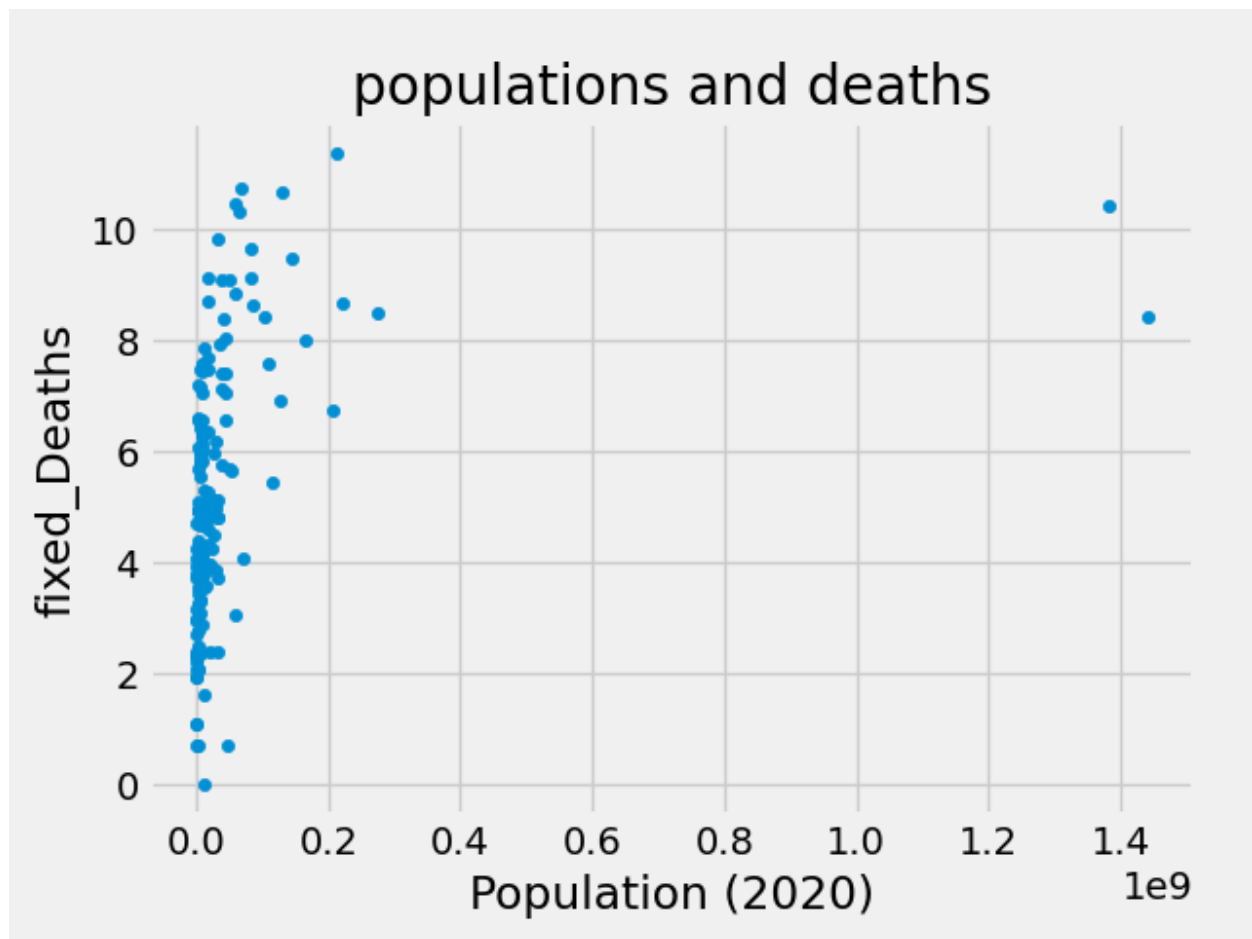


Figure 3

Conclusion

According to this scatter plot, we can conclude that there is no clear trend between the populations and death cases. Therefore, there is no relationship between the populations and death cases in different countries according to the figure3.