Data Visualization

Lab 1 – Data relationship

1 Team information

No	Student ID	Student Name	Task	Completion (%)
1	19127242	Đỗ Vương Phúc	- Visualize crawled data to diagrams - Evaluate and give comments on data	100
2	19127142	Trần Thái Đức Hiếu	- Crawling data from source - Evaluate and give comments on data	100
3	19127360	Dương Thị Xuân Diệu	- Write report - Evaluate and give comments on data	100

2 Contents

3 Overview

3.1 About the problem

From about the end of 2019 and the beginning of 2020, a plague spread horribly around the world. Every day, thousands of people are infected and tens to hundreds of people die. Worldometer Organization (www.worldometers.info) has collected statistical data from many sources and many countries reporting daily to compile into a table.

3.2 Self-evaluation

In this lab, we have done all the given task including:

- Crawl data from website
- Statistics and visualization data
- Making unbiased comments at a continental level

Furthermore, we not only visualize and give comments on the result, but we also use linear regression model to extract more insights from the data.

3.3 Folder structure

From the submitted folder, there is a document file in root and folders:

- data: Contains the crawled data and cleaned data. The file whose suffix contains "-unclean" is the raw data, "-clean" is the pre-processed data.
- img: The images which generated from the visualization code.
- src: Including three notebooks:
    - "crawl.ipynb": The source code used to crawl data from the website
    - "preprocessing.ipynb": Used to clean and pre-process unclean data in "./data" folder
    - "visualize.ipynb": Used to visualize the chosen cleaned data

4 Data Collection & Preprocessing

4.1 Data Collection

Since the data is updated hourly, the collection in the **"now"** table on the website is likely to be missed due to underreporting by the country. So, we also collected in the table **"yesterday"** and **"2 days ago"** i.e. yesterday and the day before yesterday of the collecting time. We have collected from 2022/03/08 to 2022/03/20 to use in the following labs.

Due to the requirement of lecture, in this project, we used the data of **one day** (data/2022-03-08-clean.csv) to solve the problem.

4.2 Data Preprocessing

From the collected data, it is easy to see the type of data is mismatch as we expected.

```
RangeIndex: 218 entries, 0 to 217
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Country,Other     218 non-null    object
 1   TotalCases        218 non-null    object
 2   NewCases          167 non-null    object
 3   TotalDeaths       218 non-null    object
 4   NewDeaths         113 non-null    object
 5   TotalRecovered    208 non-null    object
 6   NewRecovered      146 non-null    object
 7   ActiveCases       208 non-null    object
 8   Serious,Critical  162 non-null    object
 9   Tot Cases/1M pop  218 non-null    object
 10  Deaths/1M pop     212 non-null    object
 11  TotalTests        209 non-null    object
 12  Tests/1M pop      209 non-null    object
 13  Population        218 non-null    object
 14  Continent         218 non-null    object
dtypes: object(15)
```

So, we listed down some head sample to see what is happening.

	Country,Other	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases
0	China	111,520	+325	4,636	NaN	102,832	+110	4,052
1	USA	81,024,903	+33,615	988,208	+1,299	55,221,462	+202,659	24,815,233
2	India	42,975,883	+4,575	515,386	+145	42,413,566	+7,416	46,931
3	Brazil	29,144,964	+75,495	652,936	+518	27,344,949	+165,757	1,147,079
4	France	23,164,872	+93,050	139,618	+167	21,836,839	+98,559	1,188,415

As can be seen, the **"TotalCases", "NewCases"** and other numeric columns contain plus sign and the comma. Furthermore, if we check the **"Country,Other"** column with the country provided by GeoPandas using FuzzyWuzzy, we can also spot out the leading and trailing space.

Because of that, we conduct the data processing in 2 problems including strip space and convert the columns to number.

4.2.1 Strip spaces

Since the data in two columns **"Country,Other"** and **"Continent"** has leading and trailing spaces, we used the strip space function to remove those spaces.

4.2.2 Convert to number

For the remaining columns, we converted the **object type** to **float64 type** for easier comparison and visualization. The way to do it is as follows:

- Remove all plus signs "+" before the number
- Remove all comma ","
- Lastly, cast those columns as type float64

Finally, the processed data will look like this:

	Country,Other	TotalCases	NewCases	TotalDeaths	NewDeaths
0	China	111520.0	325.0	4636.0	NaN
1	USA	81024903.0	33615.0	988208.0	1299.0
2	India	42975883.0	4575.0	515386.0	145.0
3	Brazil	29144964.0	75495.0	652936.0	518.0
4	France	23164872.0	93050.0	139618.0	167.0

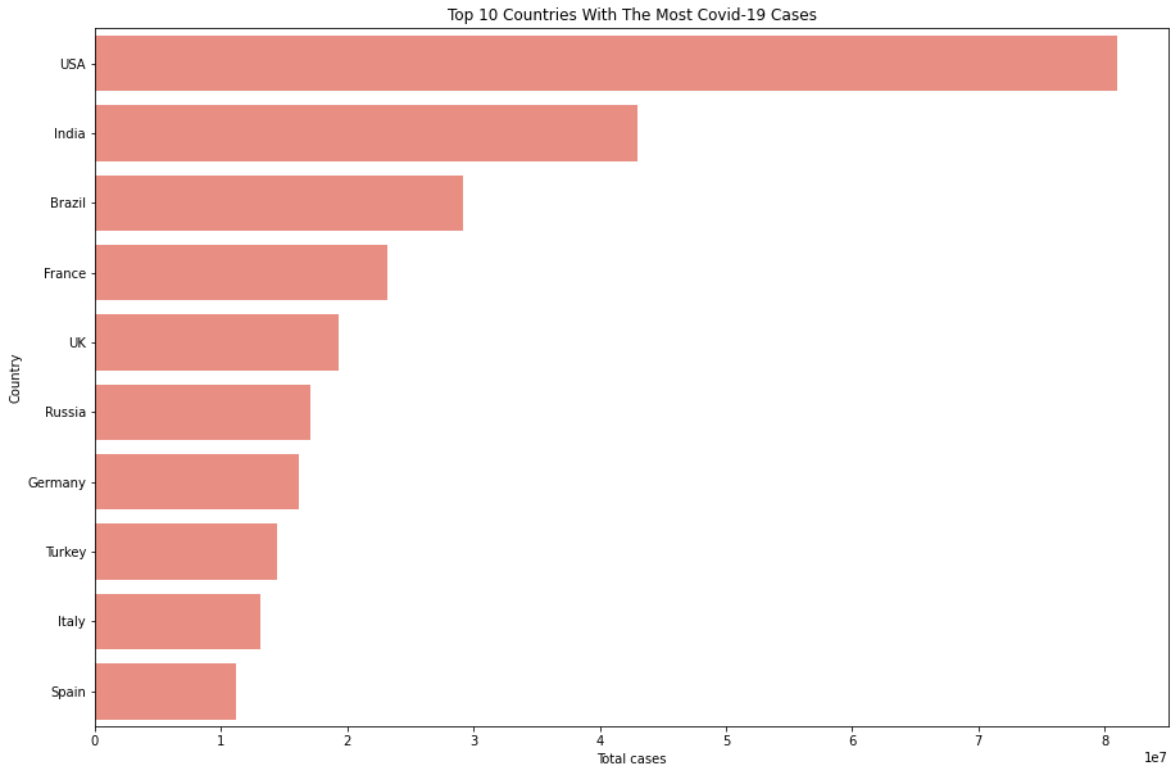
```
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Country,Other     218 non-null    object
 1   TotalCases        218 non-null    float64
 2   NewCases          167 non-null    float64
 3   TotalDeaths       212 non-null    float64
 4   NewDeaths         113 non-null    float64
 5   TotalRecovered    208 non-null    float64
 6   NewRecovered      146 non-null    float64
 7   ActiveCases       208 non-null    float64
 8   Serious,Critical  162 non-null    float64
 9   Tot Cases/1M pop  218 non-null    float64
 10  Deaths/1M pop     212 non-null    float64
 11  TotalTests        209 non-null    float64
 12  Tests/1M pop      209 non-null    float64
 13  Population        218 non-null    float64
 14  Continent         218 non-null    object
dtypes: float64(13), object(2)
```

And we have done this process for all the unclean data in "./data" folder

5 Data Visualization & Comment

In total, we have extracted 8 different insights from the data which listed as follows

5.1 Top 10 countries with the most covid-19 cases



5.1.1 Attribute

We choose 2 attributes: **“TotalCases"** and **“Country;Others"** to figure out the answer.

5.1.2 Why bar chart?

Bar chart will help organize and visualize data more clearly. In addition, the data has been sorted from the highest to the lowest, making it easy for viewers to compare the difference between countries.

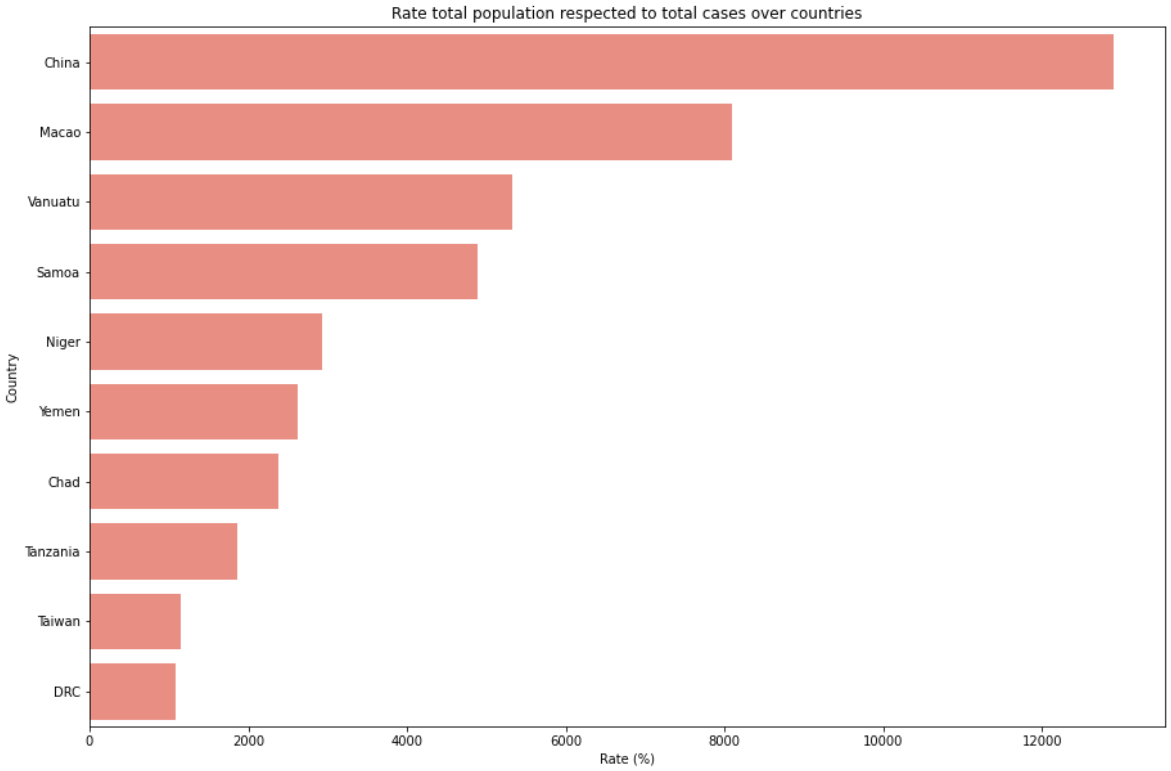
5.1.3 Comment and analysis

- Topping the list of total cases globally is the United States (USA) with nearly 80 million cases, accounting for a whopping 20 percent of all cases reported worldwide (twice as much as the second ranking country - India).
- 6/10 countries in the top 10 are in Europe (UK, Germany, France, Italy, Spain, Turkey). Possibly due to the new Omicron variant and the outbreak with the Delta variant in

European countries. So even though the population is not too high, the number of covid-19 cases in these countries is still among the top in the world.

- Excluding European countries, the rest of the top 10 are all countries with the top population worldwide (USA, Brazil, India, Russia).

5.2 Top 10 countries with the highest Population-to-case ratio



5.2.1 Attributes

We choose 3 attributes: **“Population"**, **“Country;Others", “TotalCases"** to figure out the answer.

5.2.2 Why bar chart?

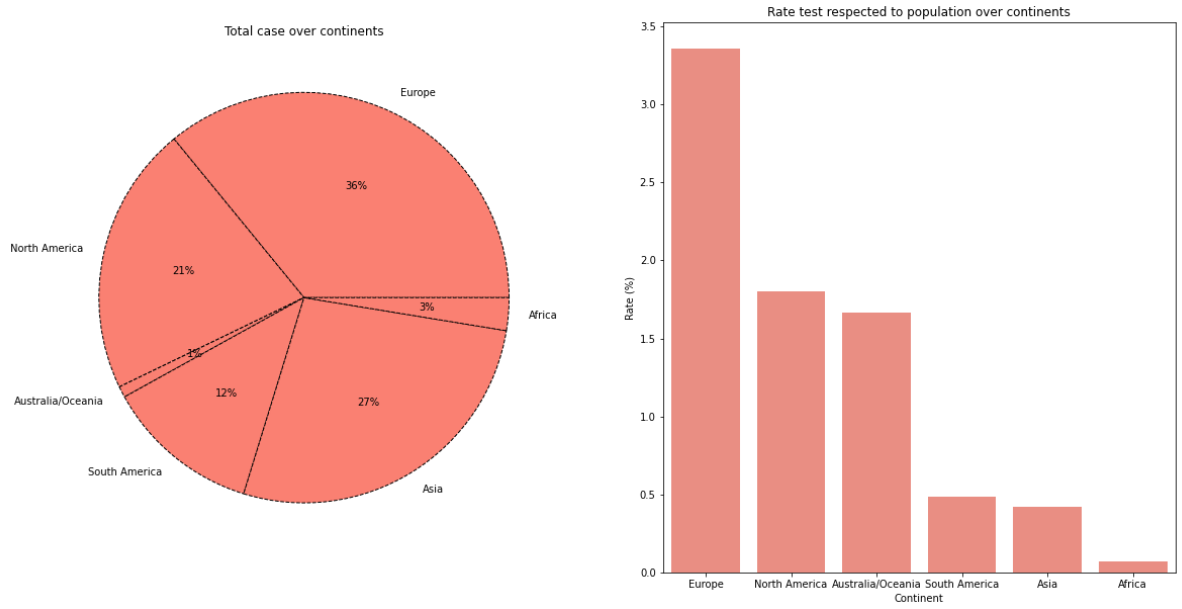
Bar chart will help organize and visualize data more clearly. In addition, the data has been sorted from the highest to the lowest, making it easy for viewers to compare the difference between countries.

5.2.3 Comment and analysis

- At first glance at the chart, it is noticeable that China has the highest population-to-case ratio. Besides, Macao and Taiwan - part of China are also in the top 10.
- It's quite strange, because China is the country with the highest population in the world, but the number of covid-19 cases is very low (ratio is 12,000 - over 12,000

people, only 1 person is positive for covid-19). Besides, covid-19 also came from China and the worldwide omicron & delta variation explosion. So, is the updated data of covid-19 cases in this country really true to reality?

5.3 Comparing the number of Covid-19 cases between continents



5.3.1 Attributes

We choose 5 attributes: **“TotalTests"**, **“Population"**, **“TotalCases"**, **“Test/Pop"**, **“Continents"** to figure out the answer.

5.3.2 Why pie chart and bar chart?

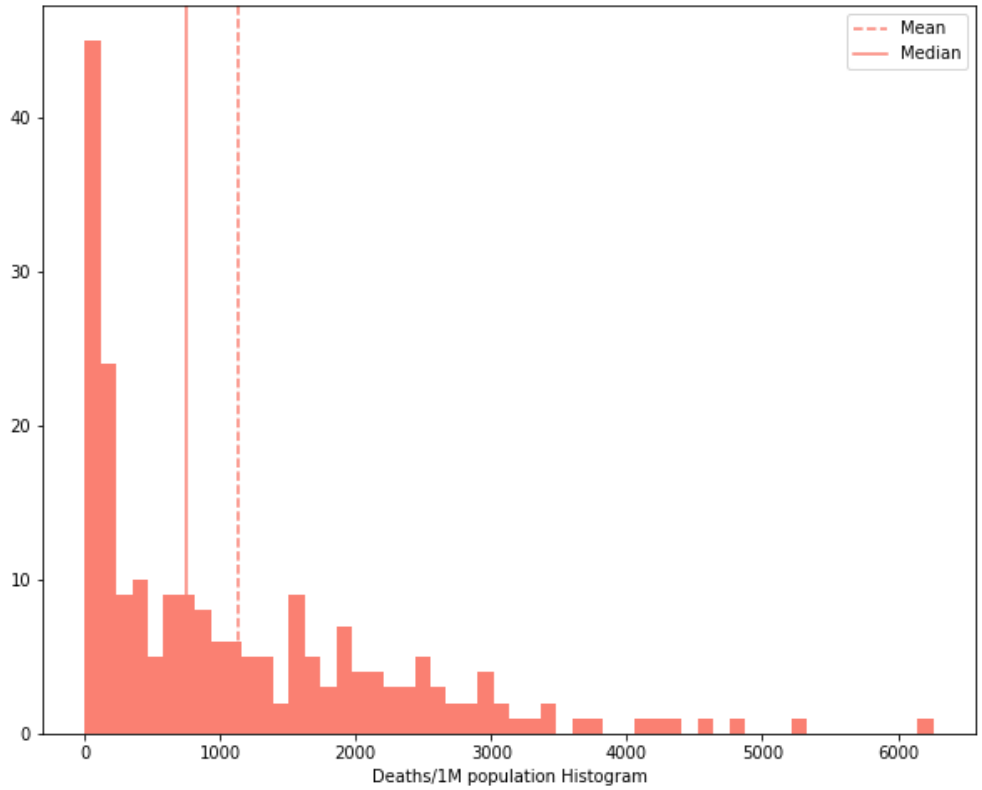
To solve this insight in the most clear and understandable way, we have used 2 types of charts - pie chart and bar chart:

- With the bar chart, we will visualize **the number of covid-19 cases** and with the pie chart we will visualize **the number of covid-19 tests** on these continents. Detailed reason, we will explain below.
- The purpose is to compare rates between continents, and no matter what the number it is. Pie charts make it easy to capture information about proportions between continents.
- Bar charts will help organize and visualize data more clearly. In addition, the data has been sorted from the highest to the lowest, making it easy to compare the difference between continents.

5.3.3 Comment and analysis

- At first glance at both charts, it is noticeable that the higher the number of tests, the higher the number of cases. In other words, the **more testing**, the **more cases appear**. This is true for all continents except Asia and Australia/Oceania.
- The total number of cases in Europe accounts for the majority with 36%, followed closely by Asia with 27%. The number of cases in Australia Oceania is insignificant, approximately 1%.
- The reason for the high number of cases in European countries is partly because the new variant Omicron spread and appeared first from this continent.
- It is quite surprising that Africa accounts for only a small part. With the African lifestyle (crowded concentration, poor medical conditions), the number of cases should have accounted for a very high percentage.

5.4 Death-to-case ratio between countries



5.4.1 Attributes

We choose 2 attributes: **“TotalDeaths"**, **“Deaths/1M pop"** to figure out the answer. The “bin" parameter is calculated by “len(df)/4”.

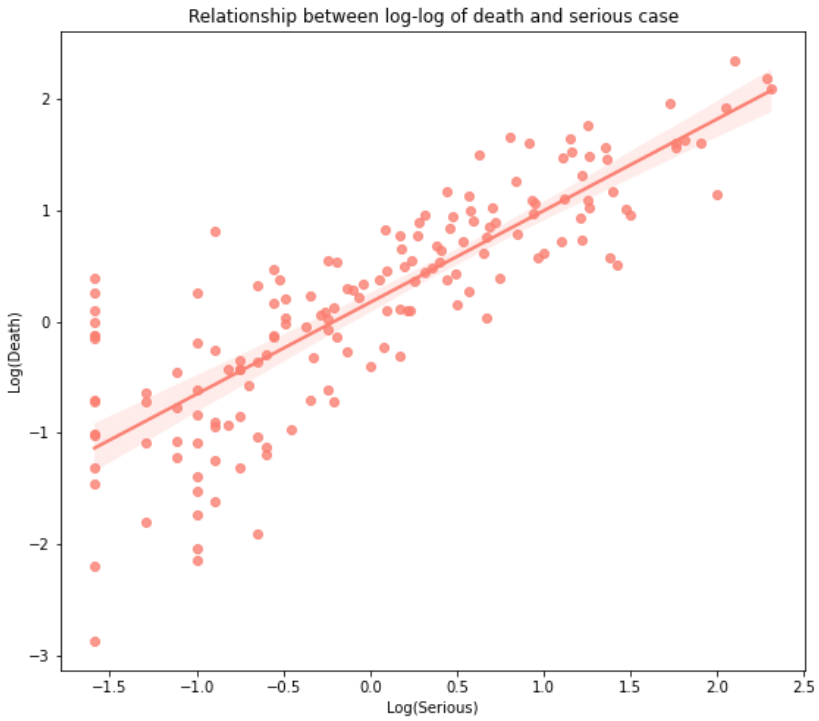
5.4.2 Why histogram chart?

Choosing the histogram because we wanted to see how this data morphed into distribution. Is it skewed or not?

5.4.3 Comment and analysis

- The data shows that death-to-case ratio is mostly at 0.015. And the data is positively skewed.
- Due to some rather large data (an outlier), the data is skewed positively.
- We also know that only 50% of countries have a death-to-cases ratio < 1.5%, which means, over every 200 people positive for covid, there are 3 deaths.
- Data of the countries located to the right of the median line need to review the epidemic prevention and medical conditions.

5.5 Relationship between death and serious case



5.5.1 Attributes

We choose 2 attributes: **“Serious,Critical"**, **“TotalDeaths"** to figure out the answer. In addition, we use **z-score standardization** to process data for the regression plot.

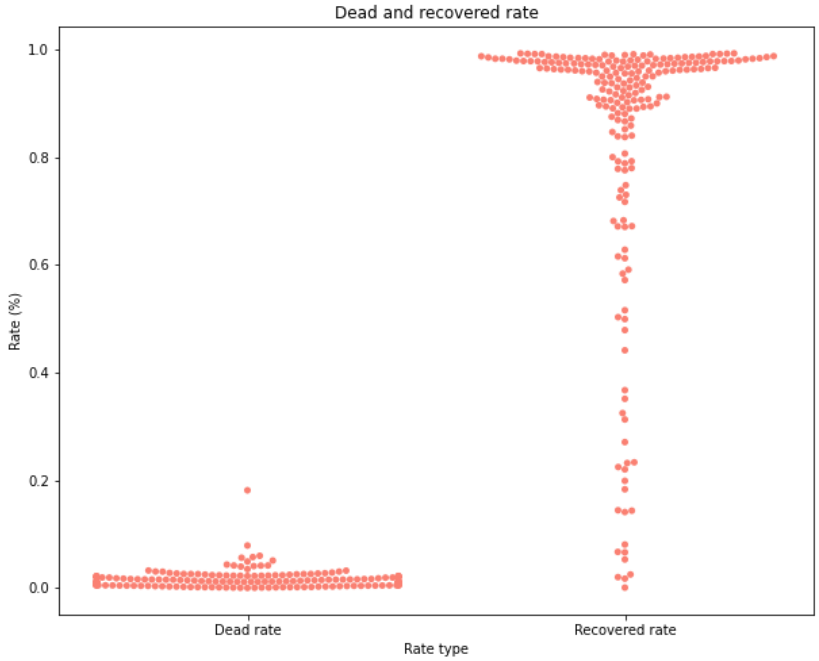
5.5.2 Why regression plot?

Based on reality, the more serious cases will lead to more deaths. So, we will use data to check if the above statement is correct or not? Is the number of deaths proportional to the number of serious cases? And the **regression plot** is a pretty good choice to find out this relationship.

5.5.3 Comment and analysis

Based on the regression plot, it is easy to see that the more serious cases, the higher the number of deaths. In other words, the number of deaths is proportional to the number of serious cases.
With this result, in order to reduce the number of deaths, it is imperative that we reduce the number of serious cases. This is also the reason why countries rush to produce and distribute vaccines around the world.

5.6 Death rate and recovered rate



5.6.1 Attributes

We choose 2 attributes: **"TotalDeath", "TotalRecovered"**, to figure out the answer.
Calculate the ratio by:

- Death rate = total death / total case
- Recovered rate = total recovered / total case

5.6.2 Why swarm plot

A swarm plot is another way of plotting the distribution of an attribute or the joint distribution of a couple of attributes. And here we have 2 related attributes, so it makes sense to use swarm plot

5.6.3 Comment

At first glance, the recovery rate accounts for about 80% and the death rate accounts for about 20%. If we keep the death rate as low as possible, it can be considered that we have controlled this virus.

So, we can also predict that "if a person is infected by COVID, he/she will have 80% to survive".

Furthermore, to estimate the accuracy of the prediction, we use a linear regression test:

OLS Regression Results

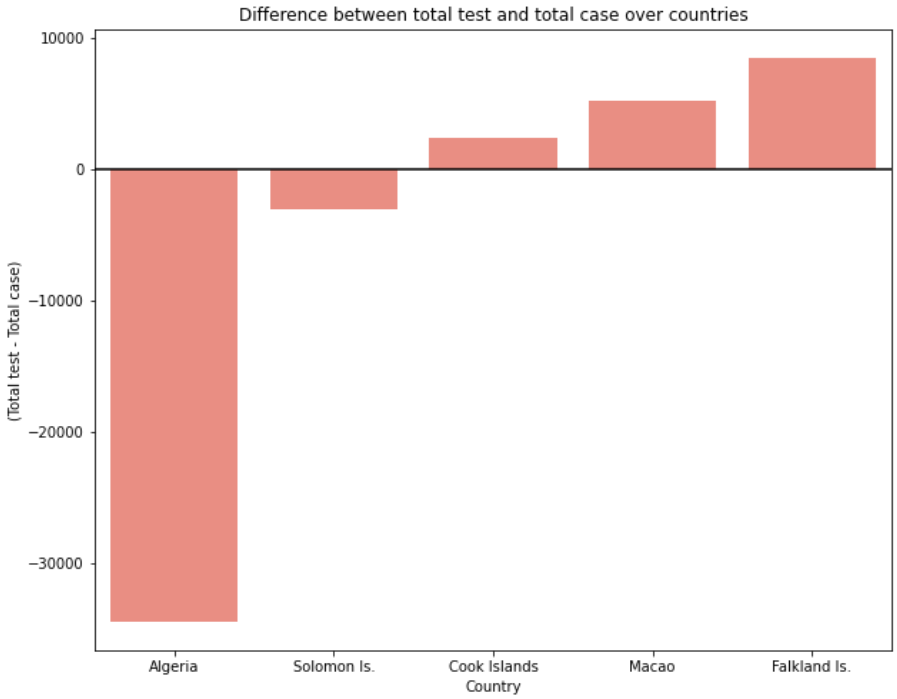
Dep. Variable:	TotalCases_Std	**R-squared:**	0.972
Model:	OLS	**Adj. R-squared:**	0.972
Method:	Least Squares	**F-statistic:**	3492.
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.96e-156
Time:	02:16:28	**Log-Likelihood:**	69.273
No. Observations:	203	**AIC:**	-132.5
Df Residuals:	200	**BIC:**	-122.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.0091	0.012	0.751	0.453	-0.015	0.033
TotalRecovered_Std	0.8582	0.032	26.437	0.000	0.794	0.922
TotalDeaths_Std	0.1593	0.032	4.906	0.000	0.095	0.223

Omnibus:	181.214	**Durbin-Watson:**	2.117
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23870.117
Skew:	2.545	**Prob(JB):**	0.00
Kurtosis:	55.879	**Cond. No.**	5.19

From the result, we can accept this prediction with a 97%.

5.7 The difference between total test and total case over countries



5.7.1 Attribute

We want to compare the difference between the **"Total tests"** and the **"Total cases".**
Calculate the difference:
Diff = **Total tests** – **Total cases**

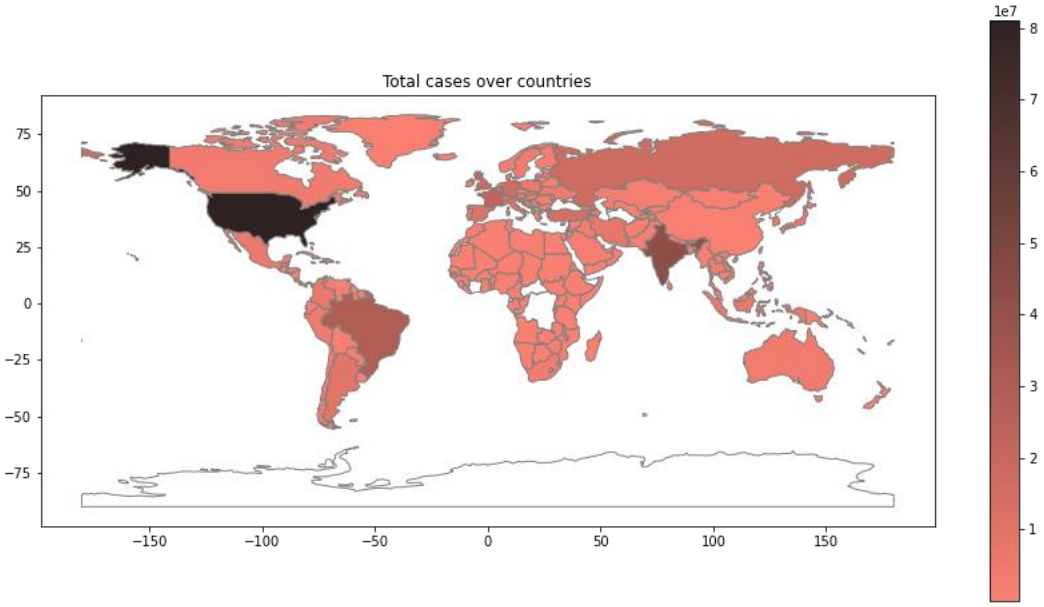
5.7.2 Why bar chart?

A bar chart can show the column of each country clearly. In additional, we want to compare the differences between countries, which bar chart will do effectively.

5.7.3 Comment and analysis

First, we sort the "Difference" of each country that we have calculated in the ascending order. Then we plot the 5 smallest "Difference" countries.
As we can see in the bar chart, only 2 countries **Algeria and Solomon** have more "Total cases" than "Total tests".

5.8 The distribution of cases over continents



5.8.1 Attribute

We choose 2 attributes: **"TotalCases", "Country,Other"**, to see how distributed of cases.

5.8.2 Why maps

Using map can help us analyze the information on the geographical view. From this view, we can see whether a country may affect another one.

5.8.3 Comment and analysis

As can be seen the USA and Alaska have a very high number of cases. However, the nearby countries have a quite low number. So, we can conclude that the nearby countries may have some policy to prevent the diseases such as lockdown.