

Python Project Report
Project 9: COVID-19 Analyser

COVID-19 Analyser

Team ID	63
Cassandra Fu Si Yun	2201540
Davin Lim Yi Han	2201898
Leo En Qi Valerie	2202795
Shu HaiJie	2202050
Nicholas Ng Jing Ngee	2201727

Project 9: COVID-19 Analyser

Team ID:63 Python Project Report

Abstract- This report is on the ongoing COVID-19 global pandemic. COVID-19 is a deadly viral pandemic that causes severe implications on the lives of numerous people globally. Governments worldwide have initiated measures and protocols to help reduce the spread of the coronavirus, such as imposing lockdowns, mandatory mask enforcement in public, and encouraging people to take the COVID-19 vaccine. The purposes are to gather, compile and analyse the data obtained from the datasets and present it for the governments to determine if the current countermeasures are effective in mitigating COVID-19. The research scope focuses on the COVID-19 vaccination of the general people as well as to determine whether vaccines can reduce the mortality rate due to COVID-19. After obtaining and sifting through available open-source datasets, the data extracted were cleaned through methods such as the handling of null columns, and also, pre-processed. The data were then plotted into graphs and charts in the Graphical User Interface (GUI). The GUI enables users to understand the data in a more simple and concise manner. Key analyses identified in this study involve the correlation between vaccination and the COVID-19 mortality rate. Therefore, this study can better aid the governments in advocating the usage of the COVID-19 vaccines for the general public as vaccination can minimise and reduce the number of deaths due to COVID-19.

I. INTRODUCTION

A. Background of COVID-19

COVID-19 is a deadly viral disease identified in Wuhan, China, that has caused numerous deaths globally. The World Health Organisation (WHO) declared it a global pandemic on 30 January 2020 [1]. Beta-Coronaviruses (Beta-CoV) have the highest mortality rate and can cause the most severe illness among the seven types of pathogenic coronavirus. COVID-19 is caused by the SARS-CoV-2 virus, from the Beta-CoV variant. It is similar to the SARS outbreak caused by SARS-CoV [1], [2], [3]. SARS-CoV-2 is suspected to originate from bats and went through multiple mutations as it spreads through other species of mammals [4].

B. Transmission of COVID-19

COVID-19 is transmitted via aerosols frequently caused by sneezing and dry coughing of the host. Various reports have shown that the SARS-CoV-2 virus can remain in the air for at least 3 hr. Hence, the virus is very contagious and can infect new hosts [5], [6]. Infected people might unknowingly release the airborne virus to infect others nearby [5], [7], [8].

C. Symptoms

The majority of symptoms can range from mild to extremely severe and even fatal after an incubation period of

about 4 to 14 days [6], [7]. Loss of smell and loss of taste are also commonly observed in people infected with COVID-19 [8]. Other symptoms include sore throat, headaches, and runny nose [5]. An estimation of 55% of infected people has mild to severe illness. Patients with severe symptoms frequently deteriorate during the second week of their sickness. These patients often require hospitalisation 7 or 8 days after contracting SARS-CoV-2 [9], [10]. 20% of hospitalised patients with COVID-19 might have rapidly worsened symptoms following the beginning of dyspnoea and develop respiratory failure [5], [11]. 30% of individuals that contracted COVID-19 do not exhibit any symptoms [12].

D. Why is COVID-19 a serious issue?

Mao et al. [13] reported that neurological symptoms were present in the individuals with COVID-19 evaluated, with 46% displaying significant neurological impairments. Several other studies have also reported the prevalence of neurological symptoms in COVID-19 patients [14]. The coronavirus has been discovered to have the ability to enter the brain via the blood-brain barrier and might even cause irreversible brain damage [15].

E. Problem Statement

By March 2020, there was a record of 52 million websites spreading false information and conspiracies regarding COVID-19, which caused mistrust among the general public and the medical profession [16]. According to surveys, only 38% of Russians and 29% of Americans trust information about the epidemic from the official media. Despite the overwhelming amount of news and reports showing the devastating effect of COVID-19 on the population around the globe, many people are still uncertain whether COVID-19 actually exists or are against vaccination. This might be due to multiple rumours and conspiracy theories about the COVID-19 vaccines [16], [17].

F. Importance of COVID-19 vaccinations and Immunity

Immunity against COVID-19 can be enhanced through vaccinations. Vaccinations can minimise the rate of viral transmissions and reduce the amount of COVID-19 related deaths [18]. Some vaccines require multiple doses, given weeks or months apart, to enable the immune system to produce more memory cells and antibodies to combat future similar pathogenic infections. Thus, the body will be able to combat the disease more efficiently.

G. Report Objectives

The research scopes of this report are to determine whether vaccination status and other factors affect the COVID-19 mortality rate in different countries. The objectives of this report are to a) develop an interactive and engaging program using Python and providing analysis to further aid the Government and Healthcare sectors of the respective countries

for more efficient handling and management of their citizens, and b) for any interested users to have a clearer insight into the recent global COVID-19 pandemic.

The hypotheses are:

Null Hypothesis (H_0): Countries with higher vaccination will not have a lower COVID-19 mortality rate

Alternative Hypothesis (H_a): Countries with higher vaccination will have a lower COVID-19 mortality rate

The results collected can be used to help improve awareness of the general public, and to encourage the Governments to push out more initiatives to promote vaccination amongst the general public and to manage the current COVID-19 pandemic and other future outbreaks.

II. RELATED WORKS

Background research was conducted on similar studies as this study's goal is to determine if vaccination of the population will reduce the mortality rates due to COVID-19.

Reference [19] explored the impact of vaccination on COVID-19 mortality, similar to the objectives of this report. The research was conducted through the collection of electronic medical records (EMRs) of adults, aged 18 and above, with COVID-19 diagnosis. Comparisons between mortality rate of vaccinated patients and unvaccinated patients were made to identify if there was a correlation between the COVID-19 vaccine and the mortality rate. Data were collected over a period of time, specifically from the 5th of March, 2020 to the 5th of February, 2022. The results determined that out of 10,000 patients, those vaccinated patients had a higher rate of survival than those unvaccinated patients [19].

Reference [19] and this report contains some key differences in the methodology used to determine the conclusion. This study used Python to create visualisations for the analysis of correlation between vaccinations on COVID-19 mortality rate cases. Reference [19] utilised a different approach by conducting a statistical analysis to determine the impact of vaccinations on the mortality rate of COVID-19 cases. The fundamental difference lies in the type of analyses and their respective assumptions. This study aims to identify if there exists any correlation between the two factors. However, [19] already assumed the presence of a correlation between the two factors.

III. THE PROPOSED APPROACH

A. System Overview

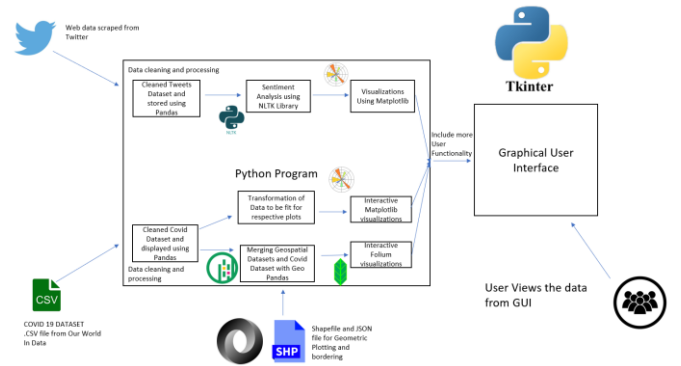


Fig. 1. System Overview

Based on Fig. 1, the system overview was outlined to provide a simple program breakdown. The tweets dataset was scraped from the popular social media application, Twitter, using the Snsrape module. The data were cleaned and pre-processed before being used for analysis. These analyses were fitted into visualisations and outputted into a graphical user interface for easier viewing by the users.

Data pre-processing and data cleansing were performed by handling null columns. Where applicable, the null columns were assigned 0. Visualisation methods, like charts and graphs, showed the relationship and correlation between selected data variables from the dataset. These included the rate of fully vaccinated individuals and the mortality rate. A Graphical User Interface (GUI) using Python was also implemented to enhance the user experience when reading or searching for the data. The sections below describe the methods employed to clean the data.

B. Twitter Data

1) Data cleaning

Data cleaning is performed on the datasets to ensure data quality and increase the accuracy of the analyses. As the tweets dataset was scraped from Twitter, the tweets were extremely messy, and the data were of low value. In order to increase the quality of the data, data cleaning has to be conducted. The data were replaced or removed where necessary, the processes taken will be further described below.

```

0 All entertainment venues such as pubs, cinemas,
1 Inside Australia's hidden conspiracy networks |
2 It's astounding to me, how hard governments an...
3 To all egovernments &amp; corporations attempt...
4 @gillon_sing @Charanator @ANINewsIP @ANI_Ba...
5 All the people who supported this inefficient...
6 A medical institution in the #Chika Prefecture...
7 Shaoguan, in Guangdong province, reported Chin...
8 No, we deserve safe air, mandated masks, VPH...
9 Don't miss our upcoming releases over the next...
10 "Infectious Diseases: Are We Ready? An Edmonto...
11 Going to RPA hospital for annual metanoma chec...
12 Having ended #COVID19 by not reporting it, Gov...
13 Government injects funding boost for cutting-e...
14 "The booster offers 'some' protection against...
15 @KarenRo08012406 @David_Coland None of us cou...

```

Fig. 2. Original Twitter Dataset

Based on Fig. 2, the original dataset was not cleaned. Firstly, the tweets columns were put through the .lower() function to convert the strings to lowercase and then the drop_duplicates function is used to drop all duplicate tweets so that only unique tweets should be considered to allow the NLTK machine learning algorithm to process the polarity score of each tweet. Finally, all empty rows were removed, as null tweets hold no value to the analysis.

```
#Data cleaning
#change all letters to lowercase
tweets_df['Tweet'] = tweets_df['Tweet'].str.lower()
#Remove all duplicate tweets
tweets_df = tweets_df.drop_duplicates(subset=["Tweet"], keep=False)
#Remove rows with NULL
tweets_df = tweets_df.dropna()
Fig. 2.1. Data cleaning
```

2) Data Pre-Processing

Afterwards, the tweets were put through the NLTK regular expression tokenizer to remove all special characters and the sentences were split into individual words instead. This allowed the algorithm to consider the polarity score of each word.

The list was then filtered to only contain words which appear more than 3 times, in order to choose words with a significant sample size and not one-off words, where it could be a name or an insignificant word.

Finally, the word lemmatizer from NLTK was used to obtain the root words of each of the words, this decreased the noise within the algorithm and allowed for multiple words to be mapped to have the same meaning for consideration of the algorithm.

C. COVID-19 Dataset

1) Data Exploration

The dataset was explored to identify which fields will be used in the analysis. The fields chosen were shown in Section 3) Data collection/ Introduction of the detail of the data used. There are 215172 rows and 67 columns in the dataset.

The fields used were further described in Table I.

2) Data Cleaning

```
# takes the given dataframe and removes all rows where location is in the rowlist i.e. list of countries to remove
def remove_rows(dataset):
    c_to_remove = 'World,North Korea,North America,South America,Oceania,Africa,Asia,Europe,European Union,' \
    'High Income,International,Low Income,Lower middle income,Upper middle income'
    rowlist=c_to_remove.split(',')
    cases_df = dataset.loc[dataset["location"].isin(rowlist) == False]
    return cases_df
```

Fig. 2.2. Data cleaning code sample

The Data Processor Module was created to clean and process the data to ensure the quality of the dataset. Some of the functions created were:

1. `replace_null_values(dataset,column_name)`: `replace_null_values()` method takes in a dataframe and replaces all null values in one specified column (`column_name`) in the dataframe with 0.
2. `drop_columns(dataset,col_l)`: `drop_columns()` method takes in a dataframe (`dataset`) and drops all columns in the list `col_l` from dataset
3. `remove_rows(dataset)`: `remove_rows()` method takes in a dataframe (`dataset`), removes all rows where the location is

in a list of unnecessary countries to (`c_to_remove`) from dataset, and returns the remaining rows as another dataframe (`cases_df`)

The countries that were removed were those where there was either no reported data, or were continents or categories wrongly classified as countries.

3) Data Collection / Introduction to the data used

The data used was obtained from the 'COVID-19 Data Repository' by Our World in Data which collates data from reputable sources such as Johns Hopkins University Centre for Systems Science and Engineering (JHU CSSE) [20], [21]. The dataset is updated daily, and the data metrics used in this report were as listed:

1. Vaccination status
2. Confirmed Deaths
3. Mortality rate
4. Gross domestic product (GDP)
5. New deaths per million
6. Population density
7. Country stringency index

TABLE I. DATA METRICS AND DEFINITIONS

Data Variables	Definitions
Country stringency index (stringency_index)	Composite measure based on the country's response indicators. ^d
Gross Domestic Product (GDP) (gdp_per_capita)	Gross domestic product at purchasing power parity
Total COVID-19 cases (total_cases)	Total number of confirmed COVID-19 cases. Probable cases are also included when reported. ^b
Total Deaths (total_deaths)	Total deaths attributed to COVID-19. Probable deaths are also included when reported. ^a
Mortality rate of COVID cases	$Mortality\ rate\ (\%) = \frac{Total\ COVID-19\ deaths}{Total\ COVID-19\ cases} \times 100\%$
New death per million (new_deaths_per_million)	New deaths attributed to COVID-19 per 1 million people. Probable deaths are also included when reported. ^a
Population (population)	Latest available population values. ^c
Population density (population_density)	$Population\ density\ (km^2) = \frac{Number\ of\ people}{Land\ area}$
Vaccination status (total_vaccinations)	Total number of COVID-19 vaccination doses administered in the country. ^a

^a. Our World in Data Team's official data information listed in the dataset

^b. John Hopkins University CSSE's official data information listed in the dataset

^c. United Nation's official data information listed in the dataset

^d. University of Oxford's COVID-19 Government Response Tracker official data information listed in the dataset

4) Data Analysis and algorithm design

Data Analysis is the process of cleaning all the data, transforming it into understandable form, and then modelling data to extract some useful information for various use cases. Several Python libraries were imported/used in modelling the data into useful visualisations:

TABLE II. PYTHON LIBRARIES

Python Libraries	Purposes
Folium	Folium makes it easy to visualise on an interactive map. It enables both the binding of data to a map for choropleth visualisations as well as passing rich vector/HTML visualisations as markers on the map
Pandas	Pandas is mainly used for converting data into tabular form; this is to make the data more structured and easier to read.
Matplotlib	Matplotlib is a data visualisation and graphical plotting package for Python and it is used to plot the graph in this report.
Mplcursors	Mplcursors provides interactive data selection cursors for Matplotlib.
NLTK	NLTK is a library that is built for Natural Language Processing (NLP). It contains many resources that allow for one to manipulate, process and analyse strings.
Seaborn	Seaborn is a Python data visualisation package based on Matplotlib that is tightly connected with Pandas data structures. It is used to plot more complex graphs.
Tkinter	Tkinter is a library that provides a simple Graphical User Interface (GUI) using Python. It has many sub-libraries such as ttkwidgets, which allows for better customisation.

Based on the analysis of the data, the following outputs were generated and shown below.

5) Graphical User Interface (GUI)

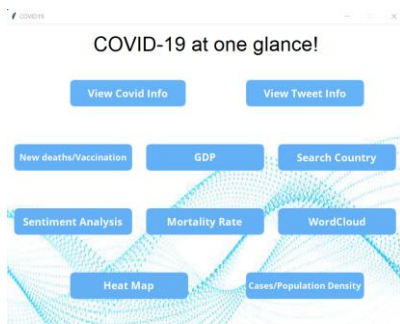


Fig. 3. GUI

A Tkinter GUI was created to facilitate the ease of viewing the data for users. This GUI contains interactive buttons for users to click which will open up a new page that contains the information stated on the button, the information shown will be in the form of graphs, a heatmap, and data sheets. Users can easily view the data/visualisations of the data analysis in the report. In addition, the GUI is a much neater way of displaying the data for users with no prior background knowledge of Python to view the output, and this allows for a more interactive experience for users to access the graphs and understand the data in the report.

6) COVID-19 Dataset

Fig. 4. COVID-19 Dataset

A Pandas table was created for the user to view the dataset, as shown in Fig. 4. In the Python program, the datasets were parsed through the data processor module, which displays the cleaned data for users. As such, it allowed the user to view and manipulate the dataset to their interests. Within this table, users can choose the columns to filter with a simple right-click to view the data they desire. Additionally, it also serves as a form of fact-checking to prove that the report visualisations align with the data shown.

Through displaying the data in an interactive format, data analyses were made easier as users have the ability to explore the data and identify areas of interest, such as the new cases of a certain date or the total number of deaths in a country.

7) Interactive Heat Map



Fig. 5. Interactive Heat Map

Fig. 5 was created using the folium library. In this heat map, the countries which have the highest amount of total cases are highlighted in red, while countries that have null values for total cases are highlighted in grey, such as Antarctica. These countries either have limited amounts of data or have no globally available data published.

The heat map was generated using the shapefile, which is in a geospatial vector data format. The shapefile from Natural Earth Data was used in order to determine the geometries for each of the countries, which allows for the borders of each country to be shown in Fig. 5 [25]. A temporary dataset using values from the COVID-19 dataset was used in this report and then merged with the shapefile using the ISO-CODES. Both of the data's ISO-CODES were in ISO3 format, which meant that there was no need for any conversion of ISO-CODES. Using the ISO3 codes as a Primary key, the datasets were combined using a left outer join where the geo-spatial data frame was on the left. After combining the datasets, the next step was to plot the merged dataset. After using the Folium's *Choropleth* and *Geo_Json* functions, the map was created and interactive features were added, such as showing the values of each country on hover. This heat map was created as a way to

allow for users to have more interactivity of identifying how many cases each country has, using colours to allow for the users to easily pinpoint which countries have the highest total number of COVID-19 cases. Additionally, they can then hover over said country to see if the country has taken stringent measures or protocols to minimise the spread of COVID-19. For example, in Fig. 5, United States of America (USA) was highlighted, and it shows that the USA has one of the highest number of cases amongst all countries and that they have not taken strict protocols against COVID-19.

8) Sentimental Analysis towards the Governments

Sentiments towards Government for covid 19 as of 10/10/2022

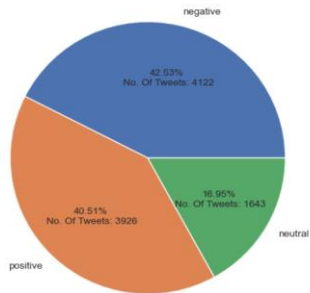


Fig. 6. Sentimental Analysis

Fig. 6. shows the percentage of sentimental tweets by users on Twitter that were directed toward the Government, regarding COVID-19 regulations. The analysis was done using the NLTK sentiment analysis algorithm.

To complete the process, the data were firstly scraped using the SNScrape library. Tweepy was considered as an alternative to SNScrape. However, as Tweepy was limited to 100 tweets, it did not fit the criterion of 10000 tweets, and thus, Snscape was chosen instead. The data were first cleaned by processing all letters in the tweets to lowercase, duplicates were removed to ensure that each tweet is only used once in the algorithm. The tweets were dropped if any of the rows include null values as null tweets are not useful. The tweets were then put through a regular expression tokenizer to take each word of the tweet to be analysed. Stopwords were then removed from the tweets as stopwords provide little to no unique information that can be used for the classification of tweets. Next was identifying the frequency of the words which appeared more than 3 times to allow for words with significant sample sizes to be analysed. Finally, the words were lemmatized over stemmed as lemmatization is widely regarded to be more useful than stemming and the graph was plotted using Matplotlib.

9) Mortality rate of COVID-19

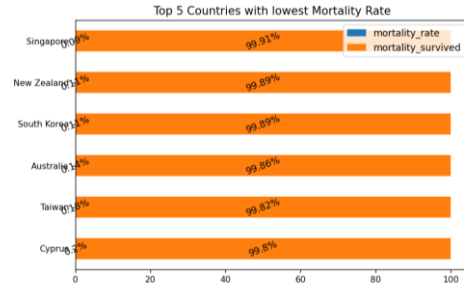


Fig. 7. Top 5 Countries with the lowest COVID-19 Mortality Rate + Singapore

	location	gdp_per_capita
3	Singapore	85535.383
22	Australia	44648.710
33	New Zealand	36085.843
35	South Korea	35938.374
41	Cyprus	32415.132
223	Taiwan	NaN

Fig. 7.1. Countries with the lowest mortality rate + Singapore

The stacked bar plot identifies the top 5 countries of mortality rate. The formula used was:

$$\text{Mortality rate (\%)} = \frac{\text{Total COVID-19 deaths}}{\text{Total COVID-19 cases}} \times 100\% \quad (1)$$

These countries are the countries that were proficient in handling COVID-19 cases and were successful in lowering the rate of death. The data used in this bar plot was grouped using location as the Key and using the total deaths of each country to plot the percentages.

Data pre-processing and filtering were performed on the data to allow for more significant analyses. A temporary dataset was first created using only columns required to lower the processing time of the datasets for more optimised code. It was then cleaned through the removal of any null values. The dataset was then filtered to only countries which had more than 1000 deaths to prioritise countries with a significant sample size of deaths. Finally, the percentages of survival and deaths were calculated and plotted above.

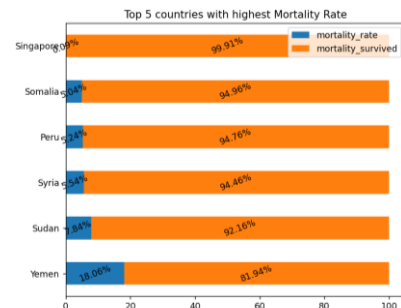


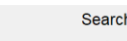
Fig. 8. Top 5 Countries with the highest COVID-19 Mortality Rate + Singapore

Fig. 8.1. Countries with the highest COVID-19 Mortality Rate + Singapore

The data used in Fig. 8. and Fig. 8.1. are the same as the ones used in Fig. 7. and Fig. 7.1. However, the top 5 countries with the lowest Mortality rate were identified and used as the variable for the y-axis of the graph instead.

Fig. 8.1. Countries with the highest COVID-19 Mortality Rate + Singapore

The data used in Fig. 8. and Fig. 8.1. are the same as the ones used in Fig. 7. and Fig. 7.1. However, the top 5 countries with the lowest Mortality rate were identified and used as the variable for the y-axis of the graph instead.



Search Country Data

Search by country

enter country name:

Singapore

Confirm

Export to CSV

Location	Continent	Date	Population	New_cases
----------	-----------	------	------------	-----------

Fig. 9. Search Country Data

Fig. 9. shows the Tkinter window for one to input the desired Country and export the results of all the countries to a .csv format. In order to facilitate user experience, some features were added for ease of usage, such as auto completion of input, allowing users to have a deeper look into the dataset if the users want to analyse the data in Microsoft Excel.

By implementing user controls within the program, it allows for users to search their desired country and provides a quick way for users to filter their data to a specific country. For instance, if governments are trying to identify their mortality rates, they can easily do so by using this feature. A treeview is constructed for the preview of the data within a Tkinter window. A filter function was developed to allow for only the selected countries' data to appear within the treeview. Users are able to choose the file path if they want to export the data as csv.

Error handling was also implemented in the program. If the users decide to not export and click the cancel button while on the export page, there will be an error message showing that the export was unsuccessful and return the users to the search page.

[illegible]

Fig. 10. Word Cloud

As shown in Fig. 10, the top words mentioned inside tweets were displayed. The word ‘government’ is quite big in this word cloud which could indicate that many people have thoughts on what the governments are doing in handling COVID-19.

Daily new deaths per million vs Cumulative Covid-19 vaccinations

The graph displays the relationship between cumulative COVID-19 vaccinations (x-axis) and daily new deaths per million people (y-axis) for ten countries. The x-axis ranges from 0 to 30,000,000, and the y-axis ranges from 0 to 5. Each country is represented by a colored line with a shaded confidence interval. The lines generally show a downward trend, indicating that as cumulative vaccinations increase, the daily new deaths per million people tend to decrease. The Netherlands shows the steepest decline, starting at approximately 1.5 deaths per million at 0 vaccinations and dropping to about 0.5 deaths per million at 20 million vaccinations. India starts at the lowest point (around 0.5 deaths per million) but shows a slight upward trend as vaccinations increase. The United States and Brazil start at higher points (around 3.5 and 4.5 deaths per million respectively) and show a moderate decline. The United Kingdom and Germany show a more gradual decline. The confidence intervals are wider for countries with fewer data points or higher variability.

Country	Approx. Start (0 Vaccinations)	Approx. End (20M+ Vaccinations)
United States	3.5	2.8
India	0.5	0.2
France	4.5	1.8
Brazil	3.0	1.5
Germany	2.8	1.8
United Kingdom	1.8	1.6
Italy	4.0	1.8
Argentina	3.0	1.0
Netherlands	1.5	0.5

Fig. 11. Daily new deaths vs total vaccination graph

After reading the data into a pandas dataframe, the data were cleaned by removing unnecessary rows, and null values were replaced with 0. Next, a new column (Total vaccinations administered per million) within a country is calculated using the formula below:

$$\text{Total vaccinations per million} = 10^6 \times \left(\frac{\text{Total vaccinations}}{\text{Total country population}} \right) \quad (2)$$

As shown in Fig. 11, the total vaccinations per million (i.e. cumulative vaccinations) was plotted against the daily new deaths per million people, in a linear regression graph, for the top 20 countries with the most COVID-19 cases as of 10 September 2022.

A bar chart titled "Mortality Rate (%) vs Top 5 GDP Per Capita" showing the mortality rate percentage for five countries. The y-axis is labeled "Mortality Rate Percentage (%)" and ranges from 0.0 to 4.0 in increments of 0.5. The x-axis is labeled "Countries" and lists Singapore, Brunei, Qatar, Luxembourg, and Macao. The bars are blue, and the exact percentage value is displayed above each bar.

Countries	Mortality Rate Percentage (%)
Singapore	0.09%
Brunei	0.10%
Qatar	0.16%
Luxembourg	0.39%
Macao	0.76%

Fig. 12. Mortality rate of the top 5 countries with the most GDP

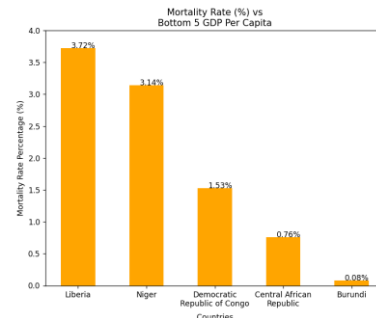


Fig. 13. Mortality rate of the bottom 5 countries with the least GDP per capita comparison bar graph

Based on Fig. 12 and 13, the Mortality rate for the Top 5 countries with the most GDP and the Bottom 5 countries with the lowest GDP are shown respectively. For data cleaning,

unnecessary rows are removed and null values in the columns Total Cases and Total Deaths are replaced with zero.

The GDP per capita, total deaths, and total cases of all countries were taken from the dataset and used to calculate the mortality rate using (1). Mortality rates of the top 5 countries with the highest and lowest GDP were filtered from data rows and shown in their respective bar charts.

14) Population Density vs Total cases

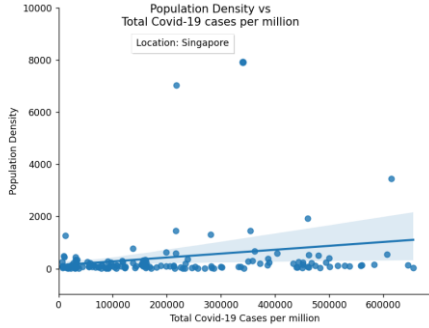


Fig. 14. Population density scatterplot

Fig. 14 above shows the population density and total COVID-19 cases per million. The population density and total cases per million was calculated using (3) and (4), then were extracted from the datasets with 1 data row per country.

$$\text{Population density (km}^2\text{)} = \frac{\text{Number of people}}{\text{Land area}} \quad (3)$$

$$\text{Total cases per million} = \frac{\text{Total COVID-19 cases}}{1,000,000 \text{ people in population}} \quad (4)$$

The countries with population density greater than or equal to 10 are filtered from the data rows to ensure reliability. The top 150 countries with the highest population density and total number of COVID-19 cases per million people were filtered from the data rows and shown on an interactive graph, as shown in Fig. 14. The interactive graph showed which country each data point represents by hovering the cursor over the data point. The graph would then show how close the data for each country was to the trendline, thereby indicating how well the country was handling the number of COVID-19 cases.

IV. RESULT ANALYSIS

The expected outcome is that our alternative hypothesis is true, whereby fully vaccinated people will have a lower COVID-19 mortality rate. Graphs and charts were used to visualise the comparison of factors as stated in the hypotheses.

A. Sentimental Analysis

Based on Fig. 6, there were slightly more negative sentiments than positive sentiments, which supports the problem statement that the general public perception of the way the governments worldwide handle COVID-19 is less than favourable. This could imply that the general public is unsatisfied with the current governmental protocols implemented.

B. Mortality Rate of COVID-19

Based on Fig. 7 and Fig. 7.1, the policies of these countries regarding the management of COVID-19 cases can be used as a reference for other countries to follow as these are extremely impressive numbers showing that the management of COVID-19 cases is important in preventing the deaths of the population. Further analysis can also be done by using these graphs along with any graphs that utilise other factors, allowing for users to possibly conclude that GDP per capita could play a factor in the COVID-19 Mortality rate as these countries are all in the top 50 GDP per capita.

Based on Fig. 8 and Fig 8.1, we observe that Yemen has an extremely high mortality rate. Even within the bottom 5 mortality rates of covid cases, Yemen has an 11% worse mortality rate. This means that they have not been prudent in handling COVID-19 cases. If we check their GDP, it shows that Yemen has the 11th lowest GDP. While they may have low GDP, they have not been shrewd in controlling the number of deaths within their COVID-19 cases and this could possibly be a correlation to a lower GDP and worse handling of COVID-19 Mortality rates. However, we need to perform further analyses to conclude any results.

Governments can change or enhance their existing policies by investing more of their annual national financial budget to increase the budgets of the Healthcare sector and the Research & Development (R&D) sector. Thus, increasing the quality of national healthcare.

C. Trendline of daily new deaths vs total vaccinations

Based on Fig. 11, the general trendlines for each of the countries show that as more vaccination doses are administered within the country, the number of new deaths from COVID-19 per day will decrease. This indicates that the vaccines are effective in reducing the number of people dying from COVID-19 infections. The gradient of the trendline for Germany is the least steep but still in a negative direction. This means that the vaccine is effective but at a slower rate.

Compared to Brazil, the United Kingdom, and India, the gradient of the trendlines for these countries is steeper, which means that the number of daily new deaths per million is decreasing at a faster rate than for these countries.

D. Mortality rate (%) vs GDP per capita

Based on Fig. 12, the top 5 countries with the highest GDP per capita have a lower mortality rate, except for Macao with a 0.37% difference from the next lowest mortality rate. This might indicate that Macao's policies to regulate vaccinations may not be effective to reduce the mortality rate. Singapore has the lowest mortality rate at 0.09%. This might indicate that Singapore's policies might be effective in reducing COVID-19 mortality rate.

Based on Fig. 13, the countries with a lower GDP per capita generally have a higher mortality rate, as compared to the countries with a higher GDP per capita. Burundi has the lowest mortality rate at 0.08%. This may be because measures against COVID-19 in Burundi are effective to curb the mortality rate of COVID-19 cases. Therefore, other countries with similar GDP can learn from Burundi's governmental policies. With

reference to the results displayed in Figure 12 and Figure 13, it can be inferred that the countries that have the highest GDP per capita have a significantly lower mortality rate than those of lower GDP per capita. Higher real GDP leads to an increase in purchasing power, and often more money for the national budget. With an increase in the national healthcare financial budget, more money can be allocated to improve the quality of healthcare provided within the countries. Additionally, richer countries can afford to buy more vaccinations for their citizens, as compared to their poorer counterparts. Buying more COVID-19 vaccinations implies a higher vaccination coverage, with an increased number of citizens gaining access to more COVID-19 vaccines.

E. Population Density vs Total cases

Based on Fig. 14, despite Singapore's high population density of around 8000 people/km², the total number of COVID-19 cases per million is not as high as in other countries with more than 400,000 COVID-19 cases per million population. This could indicate that Singapore is doing relatively well in handling the number of COVID-19 cases within the country. As viral transmission occurs most effectively in close contacts, a higher population density value was predicted to have an increasing correlation trendline to the total number of cases.

V. ACKNOWLEDGEMENTS

C.F.S.W., D.L.Y.H., L.E.Q.V., S.H.J., and N.N.J.N. would like to express sincere gratitude towards the Singapore Institute of Technology's ICT Cluster faculty, especially Prof. Daniel Wang Zhengkui, who has provided them with advice, professional guidance, and supervision throughout the development of this project. They are grateful for this opportunity to test their Python skills and attempt to create a fun yet meaningful project.

Conflict of Interest: None declared.

VI. CONCLUSION

This report aims to determine if vaccination status and other factors affect the COVID-19 mortality rate in different countries. Based on the findings, we can conclude that the alternative hypothesis (H_a) is true whereby fully vaccinated people will have a lower COVID-19 mortality rate.

There were multiple limitations and challenges faced when conducting this study, an example being that the Python library used to create the Graphical User Interface (GUI), Tkinter is old and outdated, it is considered a flexible library, but it was relatively lacking in the design elements of the GUI. It was also quite troublesome to debug and contained little to no advanced widgets or designer tools. The other limitation faced throughout the project was the lack of certain elements and values in the dataset used in this project such as the average age and number of people with illnesses before contracting COVID-19, etc. Such information would have been proved useful for this study as there could have been a correlation between these factors and the COVID-19 pandemic. In addition, there are multiple features of the data shown in the GUI, such as the population density against the total number of COVID-19 cases, GDP of countries against mortality rate, as

well as an interactive heat map to allow users to easily access and visualise the findings made.

Through the identification of these correlations, our reports provide further value to the governments in identifying the key areas they can focus on in their management of COVID-19.

Therefore, various Governments can utilise the information gathered to form more effective regulations and countermeasures in the event of future possible pathogenic outbreaks. This can be achieved through implementing more Governmental initiatives to continue pushing for the vaccination of their citizens against current or future virulent pathogen outbreaks.

This could be an area of interest for future studies and the focus will be on obtaining these data which will provide further value to the analysis, therefore this would allow the study to investigate new possible factors that could affect mortality rates such as the average age of the individuals in a country and individuals with current lung illnesses and studies can be conducted to identify the side effects of the COVID-19 vaccines.

In the event of future works, an interactive website which utilises real-time reporting of live data could be developed. Users would be able to compare the data based on their preferred criteria of factors which could be more important to their research.

This project was achieved with the collective help of all the group members of Team 63. The team was able to cooperate and work together in the development of this program. The group had worked together in the cleaning and pre-processing of the data. The tasks allocated to its members are listed as follows. Hai Jie was entrusted with the development of the GUI and Filter function of data. Nicholas was assigned to develop and plot the bar charts for the comparisons of COVID-19 Mortality Rate % vs top and bottom 5 GDP per capita. Valerie was tasked with the development and plotting of stacked bar charts for the top and bottom 5 countries in handling of COVID-19 Mortality rate%. Cassandra was entrusted to develop the visualisations of the trendline of daily new deaths per million vs Cumulative COVID-19 vaccinations per million and the scatter plot of population density vs total COVID-19 cases per million with cursor. Davin was assigned to the development of the web scraper, sentiment analysis of Twitter data, as well as the interactive heat map.

With the conclusion of the project, the group has learnt a valuable lesson in teamwork and discovered the importance of collaboration. As a team, they were able to understand how to collaborate and utilise every individual's strength to complete the task.

Throughout the project, the team was motivated and focused on the goal and development of the program. Even when there were disagreements and concerns with the direction of the project, they were able to compromise and agree on the direction of the project. As such, the group was able to complete this project with valuable lessons learnt and were satisfied with the product.

VII. REFERENCES

- [1] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Tropical Medicine & International Health*, vol. 25, no. 3, pp. 278–280, Feb. 2020.
- [2] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, vol. 22, no. 13, Mar. 2017.
- [3] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao, and Z.-L. Shi, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, Feb. 2020.
- [4] G. H. Thomas, "Microbial musings – January 2020," *Microbiology*, vol. 166, no. 1, pp. 1–3, Jan. 2020.
- [5] V. S. Salian, J. A. Wright, P. T. Vedell, S. Nair, C. Li, M. Kandimalla, X. Tang, E. M. Carmona Porquera, K. R. Kalari, and K. Kandimalla, "Covid-19 transmission, current treatment, and future therapeutic strategies," *Molecular Pharmaceutics*, vol. 18, no. 3, pp. 754–771, Jan. 2021.
- [6] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, and Z. Feng, "Early Transmission Dynamics in Wuhan, China, of novel coronavirus–infected pneumonia," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1199–1207, Mar. 2020.
- [7] W.-jie Guan, Z.-yi Ni, Y. Hu, W.-hua Liang, C.-quan Ou, J.-xing He, L. Liu, H. Shan, C.-liang Lei, D. S. C. Hui, B. Du, L.-juan Li, G. Zeng, K.-Y. Yuen, R.-chong Chen, C.-li Tang, T. Wang, P.-yan Chen, J. Xiang, S.-yue Li, J.-lin Wang, Z.-jing Liang, Y.-xiang Peng, L. Wei, Y. Liu, Y.-hua Hu, P. Peng, J.-ming Wang, J.-yang Liu, Z. Chen, G. Li, Z.-jian Zheng, S.-qin Qiu, J. Luo, C.-jiang Ye, S.-yong Zhu, and N.-shan Zhong, "Clinical characteristics of Coronavirus Disease 2019 in China," *New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, Apr. 2020.
- [8] A. Giacomelli, L. Pezzati, F. Conti, D. Bernacchia, M. Siano, L. Oreni, S. Rusconi, C. Gervasoni, A. L. Ridolfo, G. Rizzardini, S. Antinori, and M. Galli, "Self-reported olfactory and taste disorders in patients with severe acute respiratory coronavirus 2 infection: A cross-sectional study," *Clinical Infectious Diseases*, vol. 71, no. 15, pp. 889–890, Jul. 2020.
- [9] J. F.-W. Chan, S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. Poon, H.-W. Tsoi, S. K.-F. Lo, K.-H. Chan, V. K.-M. Poon, W.-M. Chan, J. D. Ip, J.-P. Cai, V. C.-C. Cheng, H. Chen, C. K.-M. Hui, and K.-Y. Yuen, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster," *The Lancet*, vol. 395, no. 10223, pp. 514–523, Feb. 2020.
- [10] Z. Wu and J. M. McGoogan, "Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China," *JAMA*, vol. 323, no. 13, pp. 1239–1242, Feb. 2020.
- [11] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, and B. Cao, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497–506, Feb. 2020.
- [12] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, "Radiological findings from 81 patients with covid-19 pneumonia in Wuhan, China: A descriptive study," *The Lancet Infectious Diseases*, vol. 20, no. 4, pp. 425–434, Feb. 2020.
- [13] L. Mao, H. Jin, M. Wang, Y. Hu, S. Chen, Q. He, J. Chang, C. Hong, Y. Zhou, D. Wang, X. Miao, Y. Li, and B. Hu, "Neurologic manifestations of hospitalised patients with coronavirus disease 2019 in Wuhan, China," *JAMA Neurology*, vol. 77, no. 6, pp. 683–690, Apr. 2020.
- [14] G. Conde Cardona, L. D. Quintana Pájaro, I. D. Quintero Marzola, Y. Ramos Villegas, and L. R. Moscote Salazar, "Neurotropism of SARS-COV 2: Mechanisms and manifestations," *Journal of the Neurological Sciences*, vol. 412, p. 116824, May 2020.
- [15] A. M. Baig, A. Khaleeq, U. Ali, and H. Syeda, "Evidence of the COVID-19 Virus Targeting the CNS: Tissue Distribution, Host–Virus Interaction, and Proposed Neurotropic Mechanisms," *ACS Chemical Neuroscience*, vol. 11, no. 7, pp. 995–998, Mar. 2020.
- [16] A. Mian and S. Khan, "Coronavirus: The spread of misinformation," *BMC Medicine*, vol. 18, no. 1, Mar. 2020.
- [17] Mylan, Sophie, and Charlotte Hardman, "Covid-19, Cults, and the Anti-Vax Movement," *The Lancet*, vol. 397, no. 10280, 2021, p. 1181., [https://doi.org/10.1016/s0140-6736\(21\)00443-8](https://doi.org/10.1016/s0140-6736(21)00443-8).
- [18] Ministry of Health, Singapore, "COVID-19 VACCINATION," *Ministry of Health*, 13-Oct-2022. [Online]. Available: <https://www.moh.gov.sg/covid-19/vaccination>. [Accessed: 20-Oct-2022].
- [19] M. Stepanova, B. Lam, E. Younossi, S. Felix, M. Ziayee, J. Price, H. Pham, L. de Avila, K. Terra, P. Austin, T. Jeffers, C. Escheik, P. Golabi, R. Cable, M. Srishord, C. Venkatesan, L. Henry, L. Gerber, and Z. M. Younossi, "The impact of variants and vaccination on the mortality and resource utilisation of hospitalized patients with COVID-19," *BMC Infectious Diseases*, vol. 22, no. 1, Aug. 2022.
- [20] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rod s-Guirao, "A global database of COVID-19 vaccinations," *Nature Human Behaviour*, vol. 5, no. 7, pp. 947–953, 2021.
- [21] J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie, "A cross-country database of COVID-19 testing," *Scientific Data*, vol. 7, no. 1, Oct. 2020.
- [22] F. Fernandes. (2016). Python Data, Leaflet.js Maps (Version 0.13.0) [Source code]. <https://github.com/python-visualization/folium/blob/main/examples/data/world-countries.json>
- [23] J. Kirenz, "Text mining and sentiment analysis with NLTK and pandas in Python," *Jan Kirenz*, 19-May-2022. [Online]. Available: <https://www.kirenz.com/post/2021-12-11-text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/text-mining-and-sentiment-analysis-with-nltk-and-pandas-in-python/>. [Accessed: 20-Oct-2022].
- [24] A. Lee, "Extracting data and labels from a DataFrame," *Extracting data and labels from a DataFrame - mplcursors 0.5.2 documentation..* [Online]. Available: <https://mplcursors.readthedocs.io/en/stable/examples/dataframe.html>. [Accessed: 20-Oct-2022].
- [25] Natural Earth, "Admin 0 – Countries," *Natural Earth*, 25-Sep-2009. [Online]. Available: <https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/>. [Accessed: 20-Oct-2022].
- [26] D. Rustagi, "Stacked percentage bar plot in Matplotlib," *GeeksforGeeks*, 01-Oct-2020. [Online]. Available: <https://www.geeksforgeeks.org/stacked-percentage-bar-plot-in-matplotlib/>. [Accessed: 20-Oct-2022].
- [27] "Folium map issue in PyCharm," *Stack Overflow*, 08-Apr-2020. [Online]. Available: <https://stackoverflow.com/questions/61099551/folium-map-issue-in-pycharm>. [Accessed: 20-Oct-2022].
- [28] "How do I plot percentage labels for a horizontal bar graph in python?," *Stack Overflow*, 17-Aug-2020. [Online]. Available: <https://stackoverflow.com/questions/63444546/how-do-i-plot-percentage-labels-for-a-horizontal-bar-graph-in-python>. [Accessed: 20-Oct-2022].
- [29] "How to plot trendlines on multiple line plot?," *Stack Overflow*, 24-Oct-2020. [Online]. Available: <https://stackoverflow.com/questions/64514297/how-to-plot-trendlines-on-multiple-line-plot>. [Accessed: 20-Oct-2022].
- [30] "How to plot each year as a line with months on the x-axis," *Stack Overflow*, 18-Nov-2021. [Online]. Available: <https://stackoverflow.com/questions/70016547/how-to-plot-each-year-as-a-line-with-months-on-the-x-axis>. [Accessed: 20-Oct-2022].