1. Project Background and Introduction

Coronavirus has changed our life in many ways. Since January 2020, millions of people have lost their beloved ones. But like darkness in the valley will always be shone upon, our hope for a normal life started increasing when the success in vaccine arrived.

Countries in the world have been working together to research various vaccines to fight against COVID-19. However, the rollout of the vaccines is highly unbalanced and there are still millions of people in the world facing the horror of losing their lives because of the coronavirus. Israel, one of the most-fully vaccinated countries in the world, starts showing deep plummet in daily new cases as almost half of its population have been vaccinated fully. On the other hand, under-developed countries in Africa or Latin America are desperate for vaccine distribution since almost all the vaccines were booked by countries in Europe and North America.

As a result, the motivation of this project is first to provide a tool for users to look up for specific country or specific date with its coronavirus cases, vaccination information and to prove that the distribution of vaccine across the globe is not even based on the economic situation.

2. Data source information

Three essential data sources: historical coronavirus cases information by country, daily vaccination update information by country and economy index data by country are used in this project.

Firstly, the historical coronavirus cases information by country could be retrieved via an API provided by volunteers. The URL of the API introduction is https://covid19api.com. Detailed documentation could be found on https://documenter.getpostman.com/view/10808728/SzS8rjbc. As shown in the webpage, this API is very powerful and could track both current update and historical data in different country for accumulated confirmed, recovered and death cases information. And it could also be filtered by date scope.

Secondly, by using selenium web scarping method, vaccination information worldwide could be acquired from the website: https://ourworldindata.org/covid-vaccinations On this website, there are several important tables that provide vaccination information including: *Share of people with at least one dose of COVID-19 vaccine, Cumulative COVID-19 vaccinations per 100 people and Daily COVID-19 vaccine doses administered per 100 people and Total number of people who have received at least one dose of the COVID-19 vaccine.*

Last but not least, the economy information could be retrieved via World Bank Database. World Bank provides API to connect with its database. The documentation for the API is on https://datahelpdesk.worldbank.org/knowledgebase/articles/898581-api-basic-call-structures To be more specific, by using indicator API query, it could easily acquire the economy index information you need for a specific country or all countries around the world. The documentation for indicator API query is on https://datahelpdesk.worldbank.org/knowledgebase/articles/898599-indicator-api-queries. The indicator that this project will be used is GDP per capita. In the API, it is represented by *NY.GDP.PCAP.CD*. World Bank explains that *" GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars."* As a result, it is more objective to reflect a country's economic development and living situation.

3. How does the program work?

The system works in a pretty straightforward way. It has one command line argument, '—static', which will skip the scraping part and open the database file directly and perform the analysis and visualization function. In general, however, it starts scraping data from three data sources mentioned above and then writes them into three csv files in order. This comes to the first challenge where second data source is a dynamic website but selenium package helped a lot.

Then the program will convert the csv files into a SQLite database file by using the pandas function. This will make data retrieving and manipulating more easily in SQLite. However, the second challenge popped up. Three data sources all have country name but they use different format. For example, COVID-19 API and World Bank refer America as United States of America but ourworldindata.org refers as United States. As a result, it could cause error when joining the tables because they do not have the same name. Luckily, 'pycountry' package in python solves this problem. It offers fuzzy search function so either USA or United States leads to US in ISO 2-digit country code. Therefore, it is possible to join the data from vaccine_info table and GDP_per_capita table to run regression analysis and draw scatter plot via ISO country code.

When running the program, the system will ask the user to input a country name and a date between 2021-01-01 and 2021-04-30 in the format of YYYY-MM-DD. If the user inputs a wrong country for example like "Disneyland" or "Wakanda" or the user inputs a wrong format date or date not in the range like '2020-02-02' or '20210303', it will keep asking until the user inputs a correct and existing country and date.

Then the system will search the database and draw three plots first showing the trend of Coronavirus confirmed, death and daily new cases number since 2021-01-01. After close the 3 figures, the system will run a regression analysis and draw a scatter plot. Then it will print out the regression statistical exam result including adjusted R^2, TSS, RSS, ESS.

The system will also print out a table that show the vaccine information for the country user inputs. At last, the system will generate and save three .svg world map files that shows the coronavirus information including total confirmed, deaths, and new cases number at the date user inputs. The svg file is interactive. User could open them in a browser and move the cursor into specific country location and it will show up the number of cases that country has.

4. Analysis and Conclusion

The analysis and output has two major part: statistical analysis and data visualization. For data visualization, there are several sample graphs below to show what information this project provides.

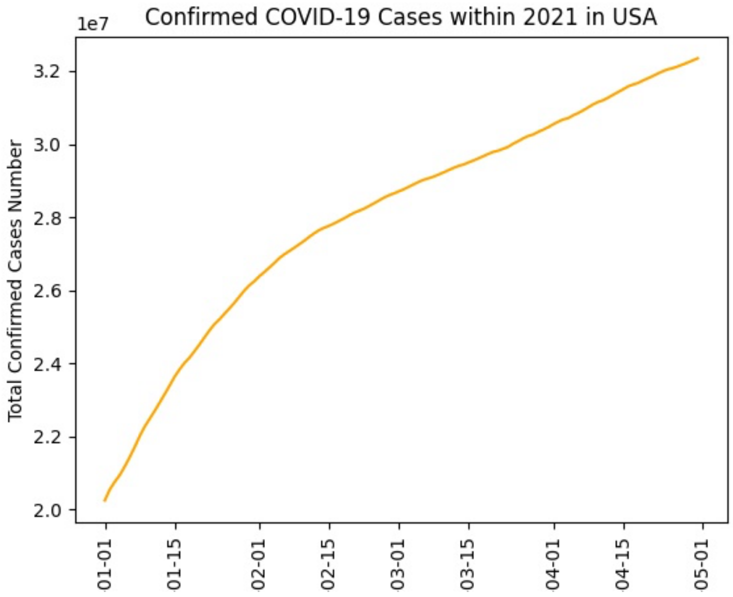


Figure 1

This figure 1 shows the trend of confirmed cases in USA since the first day of 2021. Since February, the speed of the increment of confirmed cases has been slowed down compared to January but on the other hand, the speed after February remain nearly steady and not slowing down dramatically in the United States of America while the vaccine keeps rolling out.

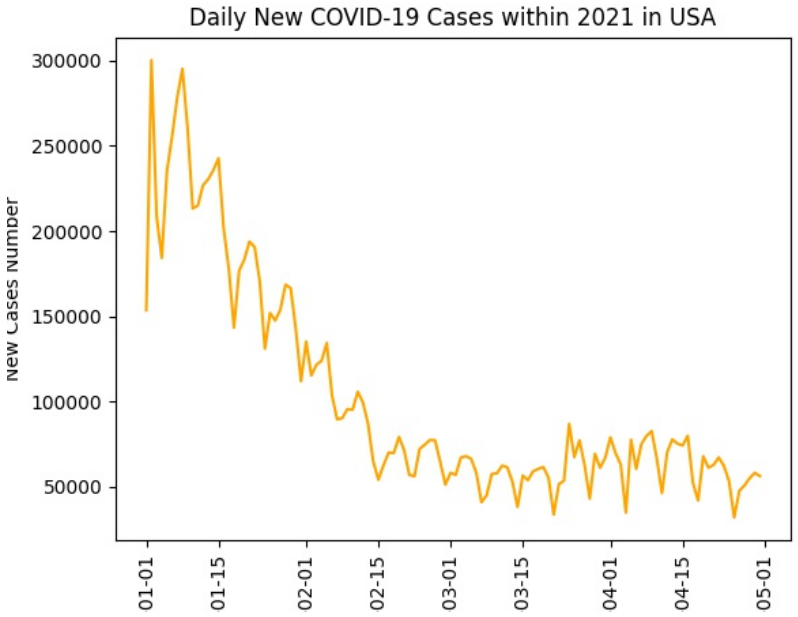


Figure 2

Figure 2 also proves the above analysis for United States of America but more straightforward and clearly. Starting February, the daily new cases dropped dramatically and remain oscillated in the range of 50000-100000 cases each day since then.

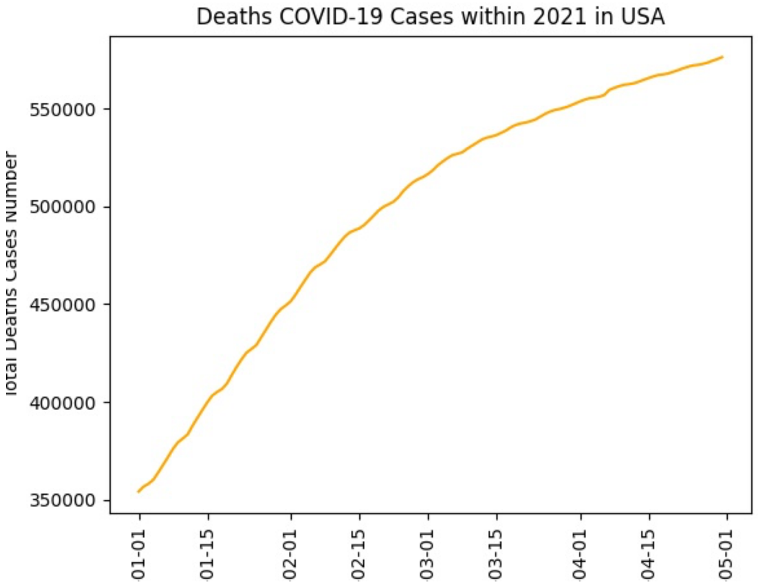
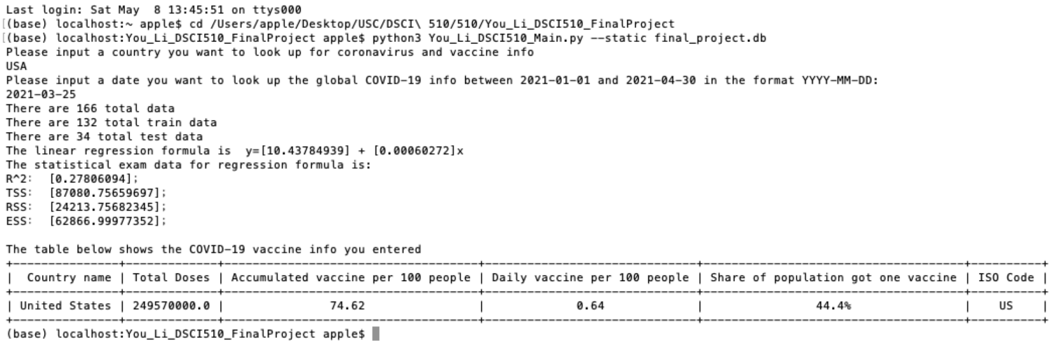


Figure 3

On the other hand, Figure 3 shows the total death cases of COVID-19 in Untied States of America in 2021. The curve became more and more horizontal since March, showing that new deaths cases in US is reducing since then.



```
Last login: Sat May  8 13:45:51 on ttys000
[(base) localhost:~ apple$ cd /Users/apple/Desktop/USC/DSCI\ 510/510/You_Li_DSCI510_FinalProject
[(base) localhost:You_Li_DSCI510_FinalProject apple$ python3 You_Li_DSCI510_Main.py --static final_project.db
Please input a country you want to look up for coronavirus and vaccine info
USA
Please input a date you want to look up the global COVID-19 info between 2021-01-01 and 2021-04-30 in the format YYYY-MM-DD:
2021-03-25
There are 166 total data
There are 132 total train data
There are 34 total test data
The linear regression formula is  y=[10.43784939] + [0.00060272]x
The statistical exam data for regression formula is:
R^2:  [0.27806094];
TSS:  [87080.75659697];
RSS:  [24213.75682345];
ESS:  [62866.99977352];

The table below shows the COVID-19 vaccine info you entered
+---------------+-------------+------------------------------------+------------------------------+-------------------------------------+----------+
|  Country name | Total Doses | Accumulated vaccine per 100 people | Daily vaccine per 100 people | Share of population got one vaccine | ISO Code |
+---------------+-------------+------------------------------------+------------------------------+-------------------------------------+----------+
| United States | 249570000.0 |               74.62                |             0.64             |                44.4%                |    US    |
+---------------+-------------+------------------------------------+------------------------------+-------------------------------------+----------+
(base) localhost:You_Li_DSCI510_FinalProject apple$ 
```

Figure 4

For vaccine information in the United States of America, we could find out that nearly half of the population have been administrated one shot vaccine and 74.62 shots have been administrated per 100 people so it is relatively fast considering the large amount of American population.

Besides basic plot figure, there are also three world map SVG files showing the current COVID-19 cases statistics information at the date user inputs. It is mostly a descriptive and informative visualization tool and could provide little analysis so they will not be shown in this report but could be accessed in the ZIP file via sample_output folder.

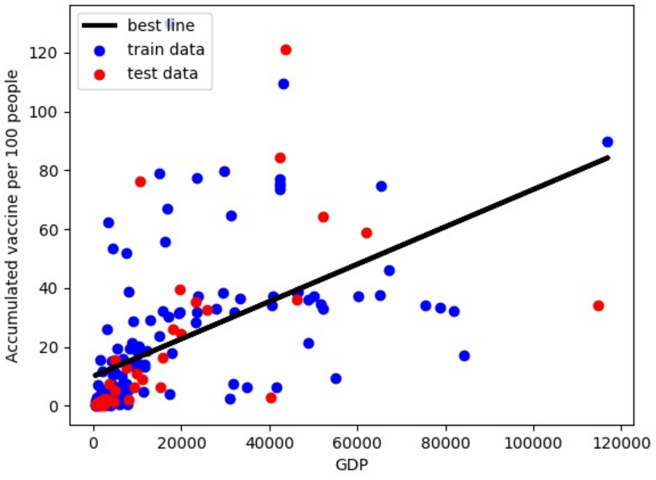


Figure 5

More importantly, this project has run a single variable regression analysis between

GDP per Capita and Accumulated Vaccine per 100 hundred people to examine if there is relationship between a country's economy development and vaccine distribution speed. The conclusion is that there is not a strong relationship between two variables from Figure 4, especially the adjusted R^2 is below 0.3 so the regression formula could not well-explained two variables linear relationship. There are multiple reasons to explain this phenomena but most essential one is that often small population country is much easier to make all people get shot than large country. As a result, country like Israel, U.A.E. and Seychelles is faster than country like Britain or even United States. Other reasons might include that vaccine distribution varies among countries. Like Japan or Taiwan is developed but they have not enough vaccine to administrate their people. Besides, like China, the virus has been well-controlled and people do not rush to get shot because there is simply no need to protect themselves since there is no severe virus threat within China. Therefore, to better explain the distribution speed of vaccine, other important variables must be considered and the model needs to be more complex and well-designed.

However, from Figure 5, we could easily find out in the scatter plot that there is a cluster near the original point in the coordinate. It shows that a lot of poor country with low GDP per Capita now is very slow in the distribution of COVID-19 vaccine in their country. Only small portion of people have been administrated in poor countries. As a result, the unbalance between rich country and poor country in vaccine distribution still exists and must be solved by immediate measures.