CSE163 PROJECT REPORT

Table of Contents

The Github repository link for our project

https://github.com/IrisDin/Covid_19_analysis

Title and author(s):

Title: Analysis about COVID-19 Vaccination, Cases, and Deaths Worldwide

Authors: Iris Ding & Tzuyang Lin

Summary of research questions and results:

1. What is the total vaccinations distribution worldwide?

   China and India are the top countries with total vaccinations. Indonesia, Brazil, and the United States are also countries of high vaccinations.

   And the countries with relative low vaccinations are Greeland, Democratic republic of congo, and Chad.

2. What is the distribution of total deaths cases distribution worldwide?

   United States has the highest number of deaths due to the COVID. Besides, Brazil, Russia, and India are all shown a relatively high number of death cases.

   And the country with relative low death cases are Greeland, Iceland, and China.

3. What is the relationship between vaccination and death cases?

   From the two scatter plots, low-p, and low-r squared number, it is unlikely that there is a linear relationship between vaccination and death cases.

4. What are the people fully vaccinated and people fully vaccinated situation for each country?

China, India, and United States are the top countries with their fully vaccinated population. However, Gibraltar, Pitcairn, and the United Arab Emirates are top countries with their full vaccination progress since these countries have high fully vaccinated per hundred ratios.

5. What is the usage of different types of vaccines distributed worldwide?

   Pfizer composed 66.2% of all the vaccinations brands are being used which show its dominant role. Moderna was the second-largest which was composed of 19.4%. Oxford/AstraZeneca, Sinovac, and Johnson & Johnson share roughly 15% of the world vaccine.

6. For this question, we mainly focus on the three variables, daily vaccinations, daily_new_cases, and dily_new deaths. To make a comparison, we first pick the period between October 1st, 2021 - January 1st, 2022. Since there are a lot of national holidays during that period like Halloween, Thanksgiving, Christmas, and New Year which might cause some interesting trends in the vaccination data, confirmed cases, and death cases. Here is a list of specific questions for this idea. To narrow down this question, we will only focus on the United State for this list of questions.
   - What's the daily vaccination trend from October 1st, 2021 to January 1st, 2022?
   - What are the newly increased cases from October 1st, 2021 to January 1st, 2022?
   - What's the daily increase in deaths cases from October 1st, 2021 to January 1st, 2022?
   - Making a comparison with the usual time (May 2021 to September 2021), what result or the relationship between the number of cases, both confirmed and death, and the number of people getting vaccinated we can find in that period? The

reason we include this question is to set a “contrast group” to compare this group with the holiday period.

With the comparison between the normal period and holiday period’s trends in new cases/deaths and vaccination, it shows an stable increasing trend of daily new cases during the holiday, a decrease in daily vaccinations, and an unstable change, frequently fluctuating in daily new cases.

Motivation:

The COVID-19 pandemic has led to the dramatic loss of human life and presents unprecedented challenges to not only public health but also individual health more importantly. Though people who had Covid recovered, the virus still has potential sequelae to the different organs. Thus, it is crucial and urgent to get vaccinated and make it universally accessible to ensure a safe condition for the general public. The pandemic is far from over, and vaccines are one of our best bets to help us stay safe. As more and more people get vaccinated, the community immunity would improve and further secure the invasion of the virus. We want to find a general patterns about the distribution of total vaccination, usage of different brands of vaccination, different periods of time’s vaccination pattern through our analysis. We also seek to disclose the serious situation of the global pandemic and find the correlation between covid-19 related factors(like deaths case and vaccination status).

Dataset:

The URL of our dataset:
https://www.kaggle.com/gpreda/covid-world-vaccination-progress

We will use the dataset from the Kaggle website and it was collected from the authoritative organization Our World in Data GitHub repository specifically for the covid-19, and it is continuing to update. There are two files of the dataset, one contains locations, also includes vaccination sources'

information. The second file is information about the manufacturers like Moderna and Pfizer etc. From the comprehensive vaccination information in different countries, people can see the total number of people who get vaccinated in their country which might let them feel safe. Also, from the vaccination information, people can visualize which country does not access adequate medical resources(vaccinations) that further helped them and facilitated the progress to ending the global pandemic.

Country_vaccinations.csv features:

- Country ISO Code
- country
- Date
- Total number of vaccinations
- Total number of people vaccinated
- Total number of people fully vaccinated
- Daily vaccinations (raw)
- Daily vaccinations
- Total vaccinations per hundred
- Total number of people vaccinated per hundred
- Total number of people fully vaccinated per hundred
- Daily vaccinations per million
- Vaccines used in the country
- Source name
- Source website

Country_vaccinations_by_manufacturer.csv features:

- Location
- Date
- Vaccine(vaccine type)
- Total number of vaccinations(total number of vaccinations / current time and vaccine type)

The URL of our dataset:

https://www.kaggle.com/josephassaker/covid19-global-dataset?select=worldometer_coronavirus_daily_data.csv

Another dataset we used mainly focused on the cases and deaths information in different countries. The data was shown on the Kaggle website and most of the data were collected from worldometers.info. Through the deaths and cases of the information provided by the dataset, we can compare the vaccination progress's relationship with the cases/deaths in different countries.

Worldometer_coronavirus_daily_data.csv features

- Date
- Country
- Cumulative_total_cases
- Daily_new_cases
- Active_cases
- Cumulative_total_deaths
- daily_new_deaths

Method:

Data Pre-processing:

At the very beginning, we will import the libraries such as Pandas, matplotlib, etc. Since we plan to construct different types of graphs and compare the specific data within the period. Also, we need some of the libraries to help us select the feature and clean the data. To clean our dataset, we also need to drop the NA data within the data set since there are so many null values. Also, we will filter out the redundant features, selecting and merging the features we want in the three different datasets. Then, we need to convert the date(string) into the Date time for processing.

The feature we do not need for our datasets:

Country_vaccinations.csv:

source, daily_vaccination_raw, daily_vaccination_per_million
Country_vaccinations_by_manufacturer.csv :/
Worldometer_coronavirus_daily_data.csv:
Cumulative_total_cases, Active_cases

Data Analysis / The method to solve our research question:

Question1(Generate a geo-map):

Dataset we will use:
Country_vaccinations.csv

We will group by country and then use the max() function to construct the dataframe with the max number of total vaccination in different regions since the total vaccination variable is cumulative. Then we will use a new library Plotly to construct an interactive map with legend to use color to differentiate the total vaccination distribution in various countries. The interactive map helps readers to view and examine the specific data in corresponding locations. The distinctive color helps readers to find the pattern and distribution of total vaccination vividly.

Question2(Generate a geo-map):

Dataset we will use and merge:
Country_vaccinations.csv
Worldometer_coronavirus_daily_data.csv

For this question, we need to merge two datasets country_vaccinations.csv(with the geographical feature) and Worldometer_coronavirus_daily_data(with cumulative deaths feature) and use these two datasets to create an interactive geo-map. To merge the dataset, we need to unify the format by changing the name of some countries. Then, we can merge two datasets with the common features, date, and country. The next step is to group

by the country and use the geo feature(iso-code) and cumulative death feature to create the interactive map by using the Plotly library.

Question 3(Generate scatter plots and test our hypothesis):

Dataset we will use and merge:
Country_vaccinations.csv
Worldometer_coronavirus_daily_data.csv

For this question, we will generate a scatter plot and find whether there is a linear relationship between immunization and death cases. We will use the cleaned and merged data frame as we used for question two. This dataset has both the vaccination features and deaths information. To generate the scatter, we will use the matplotlib library and we generate two plots. In the first graph, the x-axis would be the number of people fully vaccinated and the y-axis would be daily new deaths. In the second graph, the x-axis would be daily vaccinations and the y-axis would also be daily new deaths. Our null hypothesis is that there is a relationship between vaccination immunization and deaths. After generating plots to observe a pattern, then we calculate the p-value and r-squared to further test our hypothesis. The p-value can tell us whether we will reject the null hypothesis and the r-squared can tell us whether these variables have a linear relationship for this question.

Question 4(Generate bar charts):

Dataset we will use and merge:
Country_vaccinations.csv

For this question, we want to trace the fully vaccinated progress worldwide. To solve this problem, we would like to generate bar charts to find the top 10 countries with the fully vaccinated number of people and also the people fully vaccinated per hundred ratios (fully vaccinated people divide the country's total population). The bar charts can clearly show and rank different country's fully vaccinated progress. We need to group by our country and use the max function to

find each country's fully vaccinated population/ fully vaccinated per hundred ratios. After finding these numbers in each country, we will sort the values from descending order and use the matplotlib library to draw bar charts.

Question 5(Generate a pie-chart):

Dataset we will use and merge:
Country_vaccinations_by_manufacturer.csv

For this question, we will generate an interactive pie chart to show the popularity/percentage of the usage of different brands of vaccination. The pie chart is good at showing the proportion of multiple classes. To generate this graph, we first need to group by the different vaccination types and then use the sum function to find the total number of users worldwide. Through the processed dataframe, the plotly library can help us create a clear and organized pie chart.

Question 6 (Construct interactive line charts):

Dataset we will use and merge:
Country_vaccinations.csv &
Worldometer_coronavirus_daily_data.csv

To solve these problems, we construct six interactive line charts through the plotly library. We merge the Country_vaccinations.csv & Worldometer_coronavirus_daily_data.csv to get the vaccination, cases, and death cases information for analysis. The reason why we choose a line chart is that all the questions are seeking the vaccinations and cases' relationship in a period. Line charts are good choices in reflecting the trends within a time. With the interactive plot, the readers can know the number and trend more clearly by just looking at the hover information in the time they are interested. The main reason we create six individual graphs is that putting three different line charts, especially the line with lots of bumps, will be a mess to visualize.

- Q1 - Q3 are using the merged dataset (left-join) to construct the plots

Q1: Converting the date feature by using the DateTime in pandas, and using a mask to select the intended period of time and the target country(USA). Then, use the features daily_vacinations to plot the graph since total vaccination data is cumulative. Then we construct an interactive line chart by using the date and the daily deaths data. The x-axis is the date and the y-axis is the number of daily vaccinations.

Q2: Use the mask to select the intended period of time, then use the feature 'daily new cases' from the world meter dataset to find the daily new cases in that specific period of time. The next step is constructing an interactive line chart by using the above data. The x-axis is the date, the y-axis is the daily new cases.

Q3: Use a mask to select the intended period of time and target country, then use the feature 'daily new deaths from the world meter dataset to find the daily new cases in that specific period of time. We use the processed data above and construct an interactive line chart. The x-axis on the graph is a date, and the y-axis is the number of daily new deaths cases.

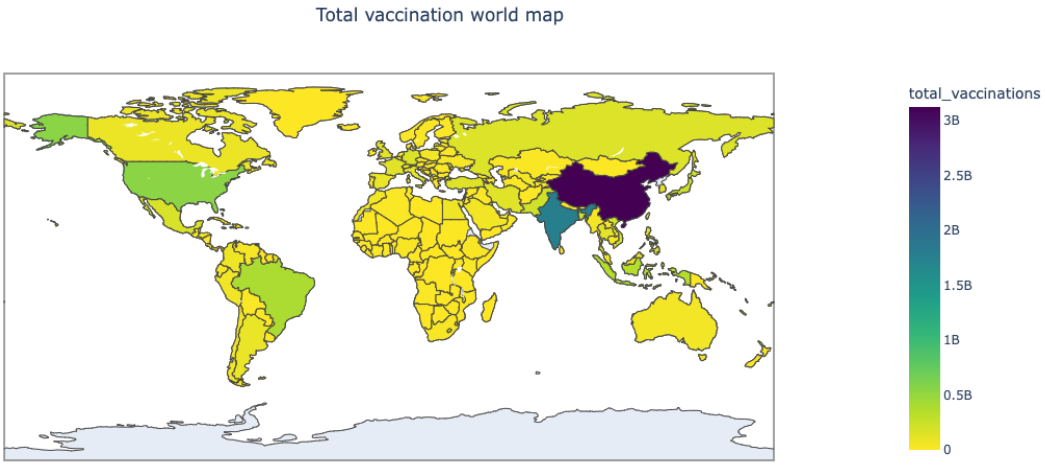
- Q4 is using a separate dataset each corresponding to the vaccination records data and COVID case data, since using in the left joined dataset, it didn't include the period from 05/01/2021 to 09/01/2021, rather trying to use the outer joined data will raise the iso_code error.

Q4:
Use the same method as above, the only difference is we use the mask to select the regular period of time. (May 2021 to September 2021). For this question, we need to compare the line chart generated in two different periods from the subquestion Q1-Q3. We can track whether there is an obvious change in the death cases / confirmed cases as people getting vaccinated increases from the graphs generated.

Results:

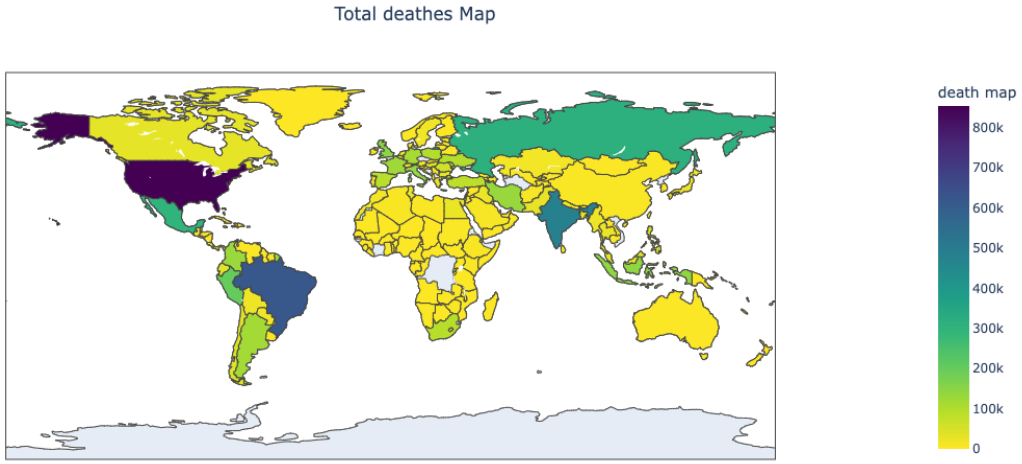
- **Question 1:**What is the total vaccinations distribution worldwide?

From the Choropleth graph being generated, based on the color differentiation on the legend, we can tell that China has the most number of total vaccinations, around 3.124118 billion, and India is the country with the second most number of total vaccinations, around 1.776674 billion. In addition to that, the countries with the relatively high number of total vaccinations are Indonesia, Brazil, and the United States, all within the level of 300 to 500 million. As for the rest of the countries, most of them are within 100 million level, especially for the developed countries like Russia, Canada, and the United Kingdom, the accessibility of the vaccinations didn't meet the demand of the general public. The potential reason why China, India, and the United States have such a large number of total vaccinations is that they share a large portion of the production of vaccines, and possess the relatively mature in making the vaccines. Thus, provides a comparative advantage in providing opportunities to the general public to get vaccinated.



- **Question 2:** What is the distribution of total deaths cases distribution worldwide?

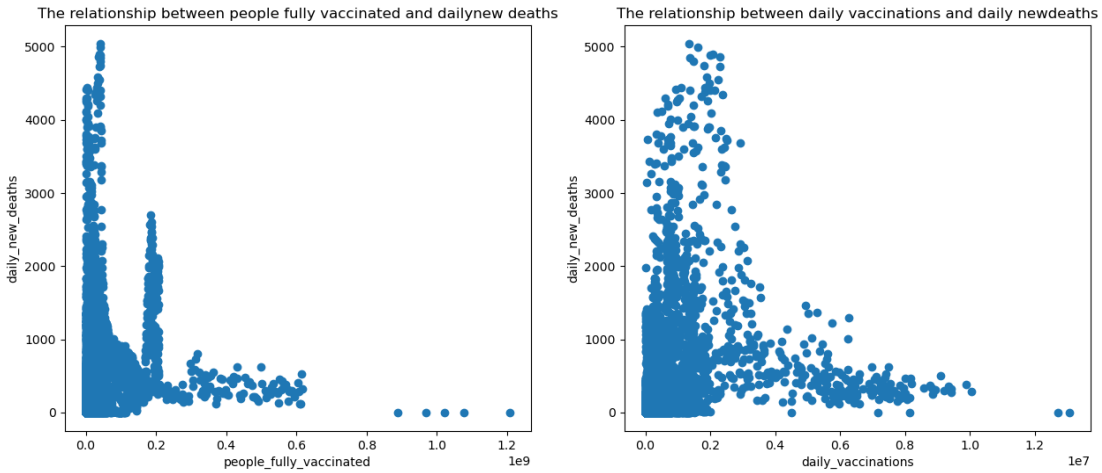
  Looking through the Choropleth we made, it is remarkably shown that the United States has the highest number of deaths due to the COVID, roughly 850k. Also, Brazil, Russia, and India are all shown a relatively high number of death cases in the world, at a level between 300k to 600k. From the overall layout of the death cases, we can identify the general condition of the COVID control is severe in regions with darker color contrast to the other regions of the world. One major factor that results in such an enormous death number in the country with the high number of total vaccinations is due to the local loosen mask policy and regional regulation in different countries. In some regions, people didn't realize the seriousness of the COVID and didn't wear masks in the public areas, and even reluctant to get treatment when they got infected, thus, leads to increasing deaths.



- **Question 3:** What is the relationship between immunization and death cases?

  We generate scatter plots to find the pattern between immunization and the relationship between immunization and death cases.

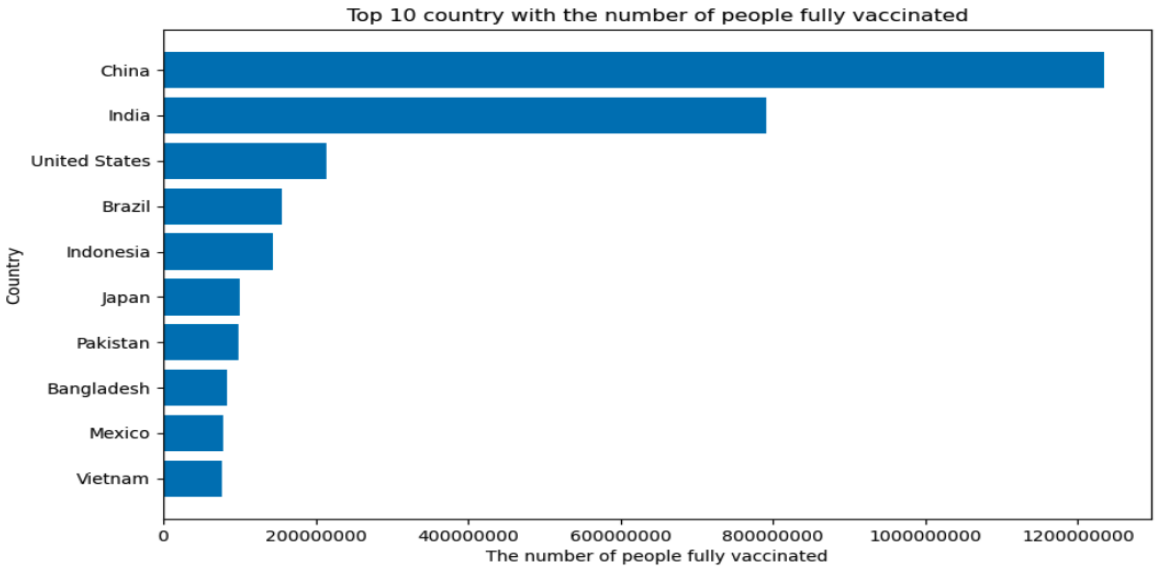
Our null hypothesis is that there is a relationship between these two variables: vaccination(people fully vaccinated, daily vaccinations) and death. I generate two scatter plots and these two graphs show no clear correlation. And the next step I did was calculating the r squared and p-value for these two linear regression models. The two p-values I calculated are 1.387260761795476e-256 and 0.0. The two r-squared values I calculated are 0.05180982981503136 and 0.1076206127701334. The two p-values are low and the two r-squared values are also low. The low p-values indicate we reject our null hypothesis that there is some relationship between people fully vaccinated/daily vaccinations and daily new death. And the two low r-squared numbers indicate our variables don't follow a linear relationship. To sum up, there is no strong linear correlation between people fully vaccinated/daily vaccinations and death.

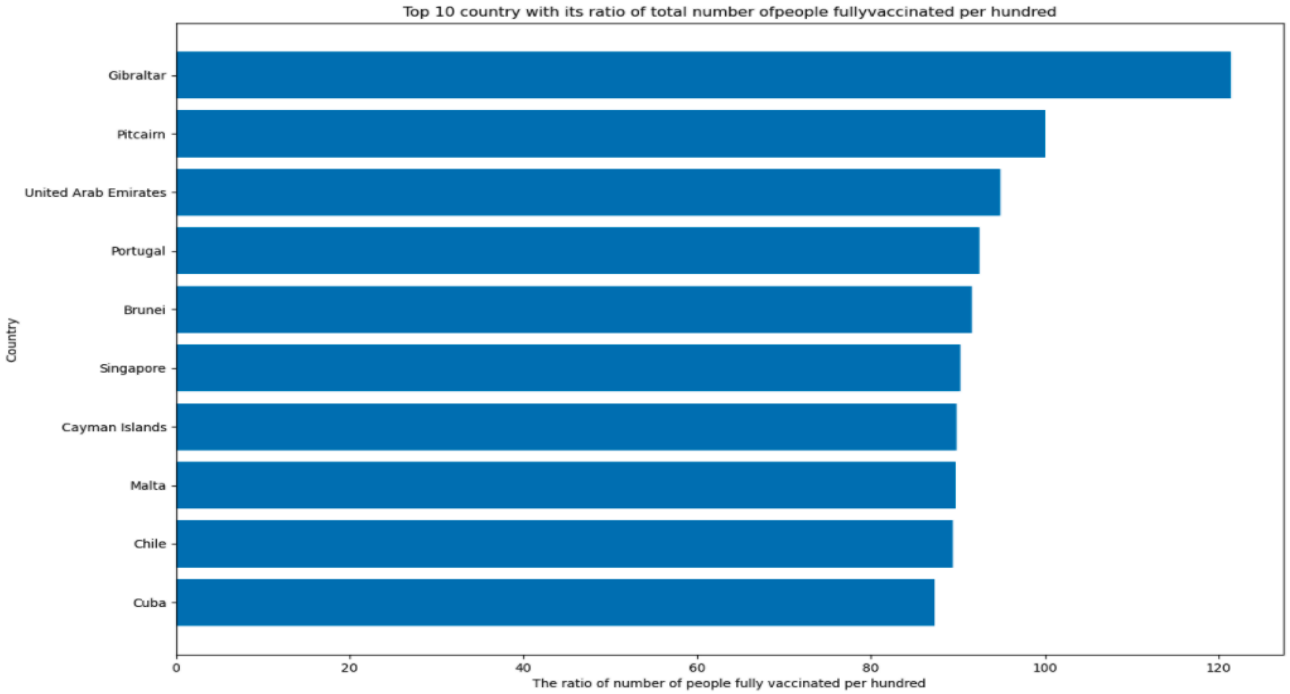


- **Question 4:** What are the people fully vaccinated and people fully vaccinated for each country?

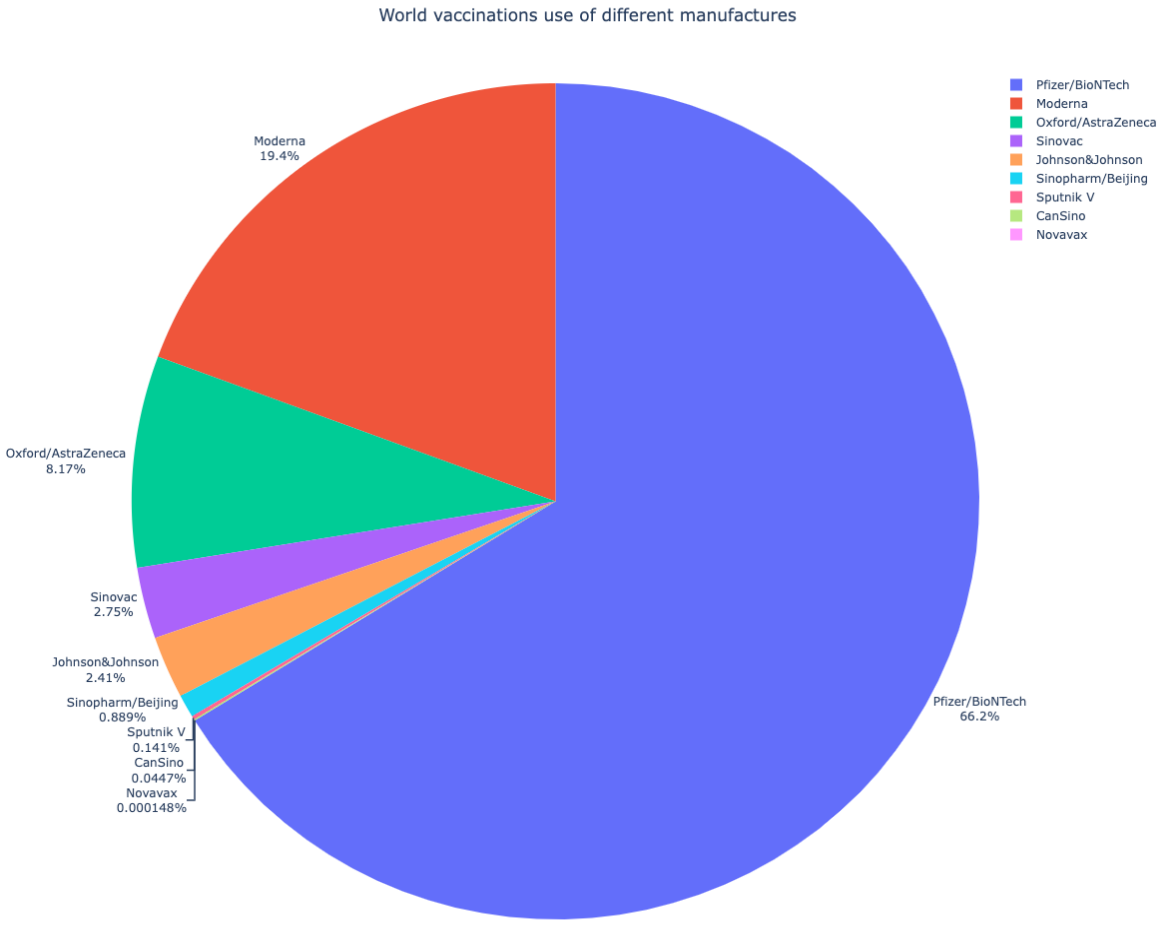
  To begin with, we generated a bar chart that shows the top country with people fully vaccinated. We could see that China, India, and the United States have more total people that are fully vaccinated. However, we found the total fully vaccinated population is hugely related to the country population since these countries I mentioned above are countries with a huge population. The total number of people fully vaccinated can not be a good indicator to compare different countries' fully vaccination progress. Thus, we continue to explore different countries' full vaccination progress by using the feature “people_fully_vaccinted_per_hundred”. That feature indicates the ratio between populations fully immunized and total population up to the date. After constructing the bar charts to find the top country with “people fully vaccinated per hundred”, it is surprising that the country shown on the bar charts dramatically changes. We can see Gibraltar, Pitcairn, and the United Arab Emirates are countries with the top fully vaccination per hundred and also the top country with fully vaccination progress.

  What makes the graph different is closely related to the population base of the country. That is to say, even though the country has lots of people getting vaccinated, as the total number of people divided the areas of geographic location, the people's vaccinations per unit will vary notably.


Top 10 country with the number of people fully vaccinated
China
India
United States
Brazil
Indonesia
Japan
Pakistan
Bangladesh
Mexico
Vietnam
Country
0
200000000
400000000
600000000
800000000
1000000000
1200000000
The number of people fully vaccinated


Top 10 country with its ratio of total number ofpeople fullyvaccinated per hundred
Gibraltar
Pitcairn
United Arab Emirates
Portugal
Brunei
Singapore
Cayman Islands
Malta
Chile
Cuba
Country
0
20
40
60
80
100
120
The ratio of number of people fully vaccinated per hundred

Question 5: What is the vaccination brand usage distribution worldwide?



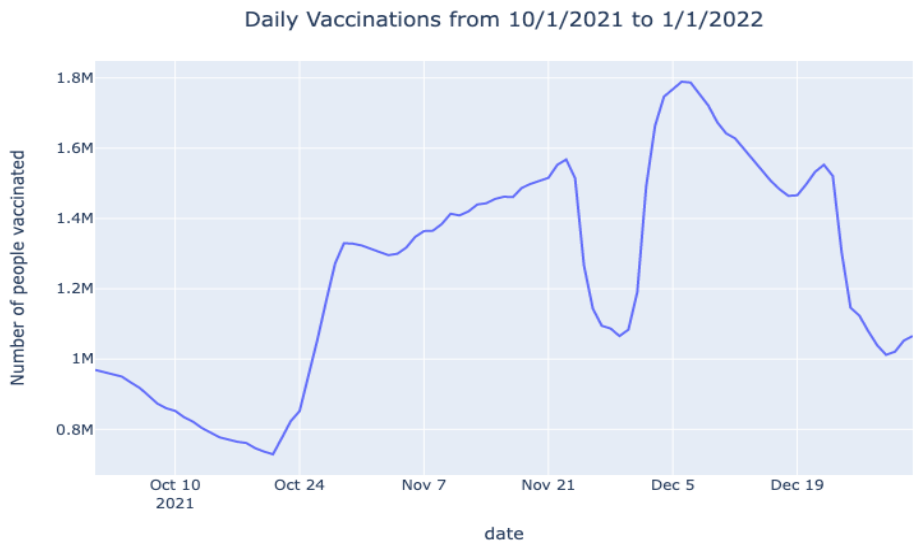
From the pie chart being generated, it is very clear that the vaccine takes up a large portion of the graph, also the one used the most worldwide is the Pfizer/BioNTech. It takes around 66.2% among the total usage of different brands of vaccines. Moreover, the second most used vaccine is the Moderna, which takes around 19.4 percent of all vaccines. We can also tell that Pfizer/BioNTech and Moderna are not only the most prevalent vaccines in the world but also are the vaccines that are trusted by the general public. Other than that Oxford/AstraZeneca, Sinovac, and Johnson & Johnson share roughly 15% of the world vaccine market. Pfizer and Moderna have huge market share primarily due to their advanced mRNA technology, and they are the very first granted vaccine

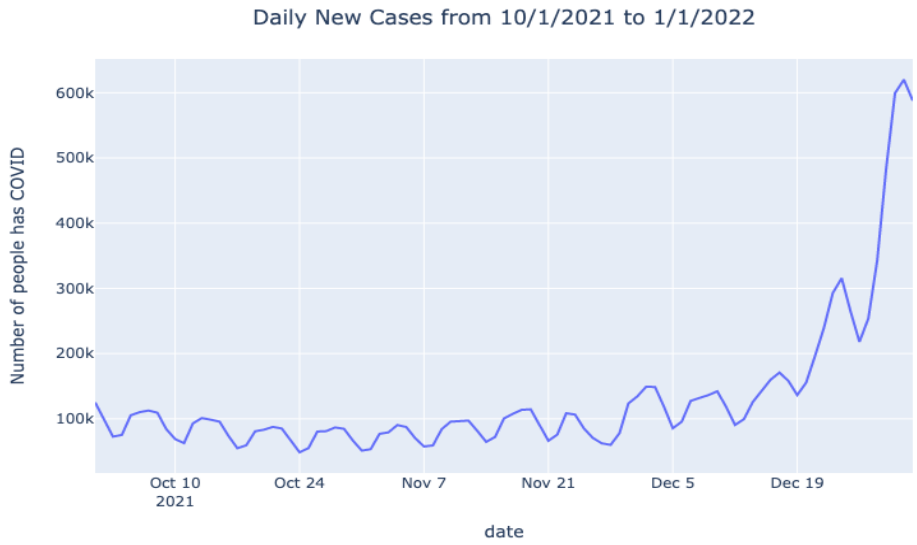
being used around the world possessed with fairly good quality, so to lay the consumer foundation since then.

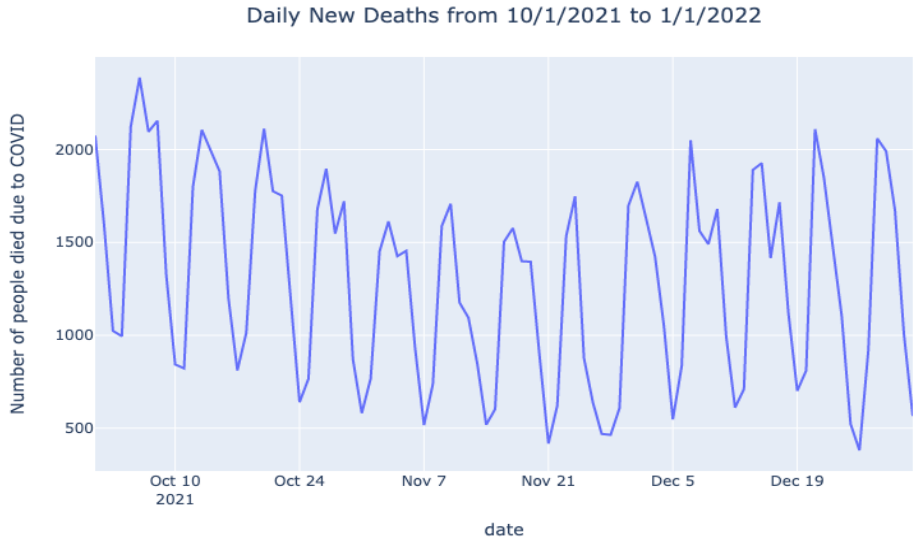
- Question 6:
    - What's the daily vaccination trend from October 1st, 2021 to January 1st, 2022?
    - What are the newly increased cases from October 1st, 2021 to January 1st, 2022?
    - What's the daily increase in deaths cases from October 1st, 2021 to January 1st, 2022?
    - Making a comparison with the usual time period (May 2021 to September 2021), what interesting result or the relationship between the number of cases, both confirmed and death, and the number of people getting vaccinated, can we find in that period of time?

      From the line chart generated during the period 10/1/2021 to 1/1/2021, it is the period with lots of holidays like Halloween, Thanksgiving, and Christmas. The general trends of the daily vaccination from October to December have three turning points occurring at 10/20/2021(decreasing till the date), 11/29/2021(increasing till the date), and 12/6/2021(decreasing), which shows an unstable state. The peak occurs on 12/6/2021 with 178687 people vaccinated during the day. As for the daily_new_cases, it, even more, fluctuated with multiple sharp bumps in that period, thus, we couldn't get any useful information from the graph.

      For the daily_new_deaths, it shows a stable stage at first till Christmas Day, it proliferates almost 400k within 4 days and reaches the peak at 12/31/2021. Overall, there is a remarkable increase of daily new cases in the holiday period of time, a decrease in daily vaccinations, and an unstable change in daily new deaths.


Daily Vaccinations from 10/1/2021 to 1/1/2022
Number of people vaccinated
1.8M
1.6M
1.4M
1.2M
1M
0.8M
Oct 10
2021
Oct 24
Nov 7
Nov 21
Dec 5
Dec 19
date

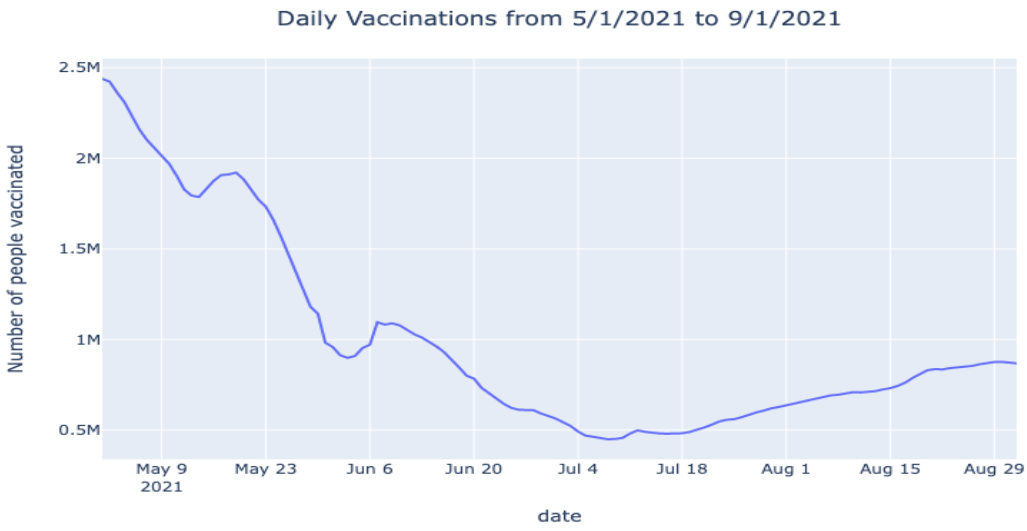

Daily New Cases from 10/1/2021 to 1/1/2022
Number of people has COVID
600k
500k
400k
300k
200k
100k
Oct 10
2021
Oct 24
Nov 7
Nov 21
Dec 5
Dec 19
date

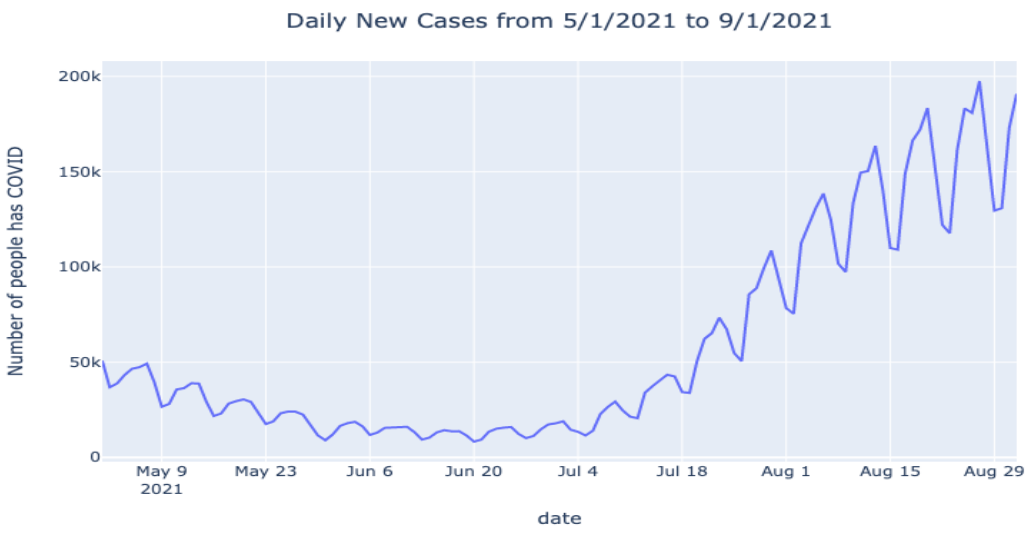

Daily New Deaths from 10/1/2021 to 1/1/2022
Number of people died due to COVID
2000
1500
1000
500
Oct 10
2021
Oct 24
Nov 7
Nov 21
Dec 5
Dec 19
date

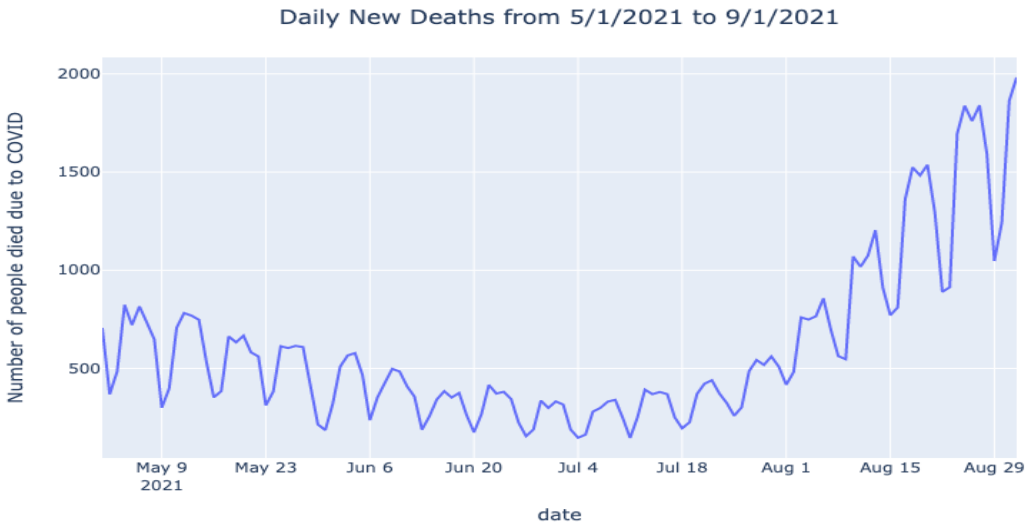
From the line chart generated during the period 5/1/2021 to 9/1/2021, the general trend of the daily vaccination from May to September has shown steady decrease overtime, till 7/17/2021(increase a bit since then). Based on the charts, we can tell that the daily vaccination reaches the lowest point at 7/8/2021, daily new cases have the peak at 8/27/2021, and daily new deaths meet its vertex at 9/1/2021. Referring to the chart of the daily_new_cases, it illustrates a notable increase so as the line chart of daily_new_deaths exhibits.

In contrast to the period with many holidays, mainly the date around Oct 31st, 2021, Nov 24th, Dec 25th, and the Jan 1st, along with the data presented by the line charts, unfortunately, we are not able to explore a clear pattern for which increase in the daily vaccination may lead to decrease to either death cases or covid cases. Mainly because, the trend is shown on the graph is different at the holiday time, and most of the time, it fluctuated multiple times with extremely high variability, therefore, it shows no clear correlation between the daily vaccinations and daily new deaths/cases.Despite the fact that the daily new cases during the two periods all showed an upward trend, still, the difference between the daily new cases was enormous. To be more specific, during the holiday session, the number of daily new cases raises from 100k to 600k, in contrast to the normal session, it only uplifts from 50k to 200k. In this case, it is possible to infer that especially during the period with many holidays, the daily new cases are very likely to proliferate.

As we spot on the graph, the abrupt increase in the daily new cases mainly connects to the population movement in the holidays. People move around from place to place, which provides a “good” environment and intensely enhances the risks of being infected.


Daily Vaccinations from 5/1/2021 to 9/1/2021
Number of people vaccinated
2.5M
2M
1.5M
1M
0.5M
May 9
2021
May 23
Jun 6
Jun 20
Jul 4
Jul 18
Aug 1
Aug 15
Aug 29
date


Daily New Cases from 5/1/2021 to 9/1/2021
Number of people has COVID
200k
150k
100k
50k
0
May 9
2021
May 23
Jun 6
Jun 20
Jul 4
Jul 18
Aug 1
Aug 15
Aug 29
date


Daily New Deaths from 5/1/2021 to 9/1/2021
Number of people died due to COVID
2000
1500
1000
500
May 9
2021
May 23
Jun 6
Jun 20
Jul 4
Jul 18
Aug 1
Aug 15
Aug 29
date

Impact and Limitations:

Through the data visualization we made from three datasets, we can draw insights into which types of vaccine are widely used the most, the percentage of each vaccine among them all, also what's the status-quo of the cumulative total deaths due to the COVID globally. What's more, from the bar chart and line chart constructed, the audience can access the information about which country has the most number of people_fully vaccinated, and what the daily_vaccination, daily_new_deaths, and daily_new cases change over time during two different periods in the US. The intended audience of the analysis is focused on the general public because COVID information is closely related to their daily life within the community.

One limitation we found is the naming convention used for the country is different, also, due to the identity of the data(global data), it is essentially too large to handle, especially with lots of NA data(country_vaccination.csv and country_vaccinations_by_manufactures.csv). As a result, it makes it difficult for us to make a comprehensive analysis for the vaccination track and pattern change between the daily_new_deaths, daily_new_cases with daily_vaccinations.

To address our project idea more broadly and to get more insights on Covid-19, it would be better if we can further obtain and analyze the spread rate of the pandemic after people are vaccinated. After then, we will know which vaccine is the most efficient one compared with others so that certain kinds of vaccines could be highly recommended to the public. In addition, since the progress of dealing with the pandemic is closely related to each country's policy and culture, it is not an easy thing to ensure each country's cooperation and to get authentic data. Besides, it would be a massive project and would cost a lot of time if we'd like to address our idea more comprehensively. All this analysis and collection of data should provide people with an overview of the epidemic situation, to help the public to know the popularity of each vaccination.

Challenges Goals:

Multiple Dataset:

For this project, we plan to use three datasets. The first one is country_vaccinations_by_manufacturer which contains the information of the daily usage of the vaccinations of different brands throughout various regions. And the second dataset is country_vaccinations which keeps records of the daily vaccinated data and tracks down the percentage of people who received the vaccination out of the whole population. The third dataset we use is worldometer_coronavirus_daily_data, which mainly keeps the records of the daily new increase cases and daily new increase deaths over the period of time globally. We will use a joint method to merge the dataset so that we can analyze the relationship between different brands' vaccination, overall vaccination information, and deaths/confirmed cases.

Messy Data:

After viewing the dataset, we found there is some messy part of the dataset Country_vaccinations.csv. To begin with, there are a lot of missing values in the front part of this dataset. However, when we scroll down the dataset, there are only a small amount of missing values remaining. We think it is necessary to do pre-processing of the dataset to filter the missing value. And we still need to pay attention to the influence of the missing values. For example, if the country has too many missing values of their vaccination data and we choose to keep these data, the result generated would be biased. If we just end up removing an entire country from the dataset because they don't have vaccination data, that might tell us there might not have many total vaccinations or only a few people get vaccinated. Also, another potential problem with the dataset is that there are over 70000 rows of this dataset which contain all different countries' vaccination information. This is a large dataset that might be hard to process and filter the core information and pattern we want. In addition, the content of data also makes the visualization hard to interpret since the figure will be extremely large. The

potential solution we came up with is that we need to drop the feature or data we do not need. We need to clean and condense our dataset further to get better visualization. Besides, setting scales of the data we want to analyze. For example, we can make emphasis on the vaccination data for the United States in 2021 only.

New library:

Plotly:

Plotly is one powerful library that allows us to make interactive visualization through python. We can drill down more details, patterns and organize clearer labels by making interactive plots.

Scipy:

Scipy is a library that contains a lot of mathematical and statistical operations. For our project, we need to generate some values with statistical significance(p-value, r-squared) to verify and test whether there is a correlation between the variables we want to test. The scipy library has a lot of built-in functions which make us calculate these values in a few lines of code.

Work Plan & Evaluation:

We insert a chart below that displays our overall work plan for the final project and plan to work together based on the schedule and meet weekly depending on the working load. On top of that, we separating the work equally and review of each other's work/code to make sure we are working in the same place. We used google co-lab and vscode to collaborate together and review each other's python code/jupyter notebook. Then we upload all of our code, paper, testing files to Github. For our proposed work plan, it was as accurate as we estimated. The only difference is the testing and reviewing part takes a longer time than we expected. We think the main reason is that different people have different

coding styles and it takes some time to understand each other's code. Moreover, it is also challenging to make test cases for our dataset.

Working part	People on duty	Schedule /estimate time
Processing / Filtering the Data	Frank/Iris	We plan to meet twice a week on zoom/in person, this part takes 3-5 hours
Applying the libraries to construct the data visualization	Frank/Iris	We plan to meet twice a week on zoom/in person, this part takes 5 hours
Testing/Debugging code, revising code	Frank/Iris	We plan to meet twice a week on zoom/in person, this part takes 2 hours
Constructing the analysis for the question listed	Frank/Iris	We plan to meet twice a week on zoom/in person, this part takes 3-5 hours
Submitted the work on the grade scope(final review)	Frank/Iris	Depending on the requirement

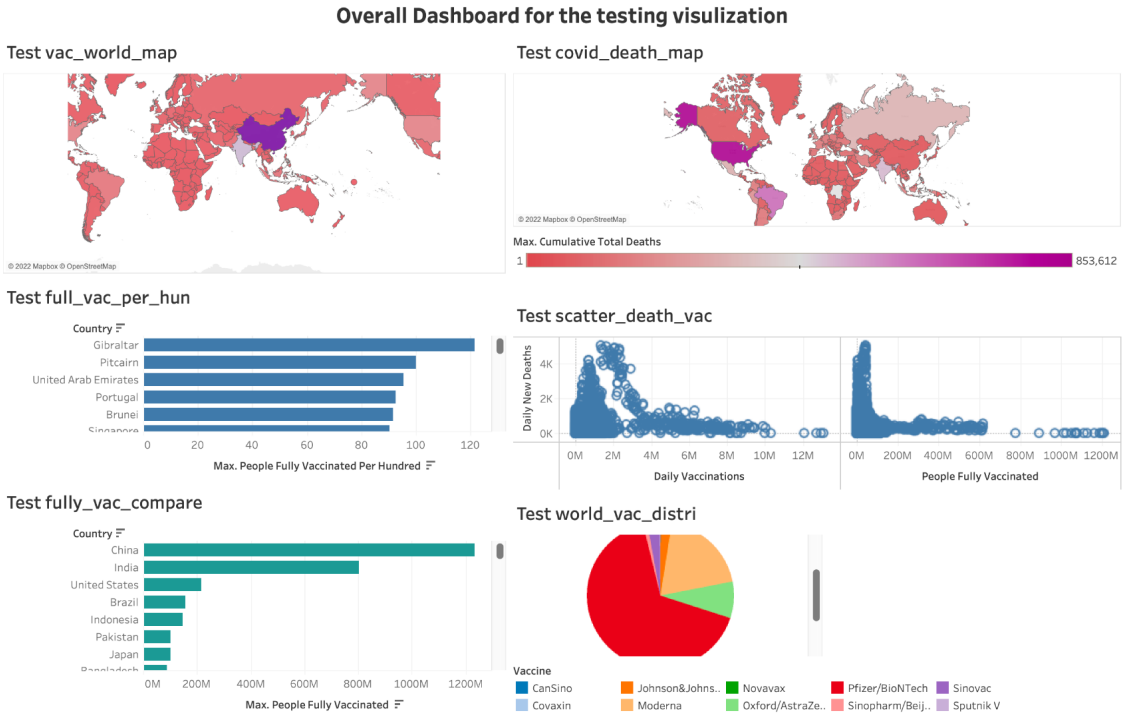
Testing:

To test the accuracy of the number/data being shown on the plots, we chose the five countries, China, India, Greenland, Congo, and United States, and two dates(12/31/2021 and 05/01/2021) to double-check the result presented on the graph matched the actual data in the files.

During the process, we mainly use the assert function to examine our output. Since the graphs we created are interactive, especially on the two total_vaccinations map and daily_new_cases map, the number is stored in the scientific notation, which means we are only to test the number with visible digits. Thus, we use the round function and math.isclose as well to ensure the correctness of the result. Besides, since we narrow down the country(only focus on the U.S.) and time period to only two days, it is relatively obvious to identify the difference between the actual data and data on the chart.

In the method.py we used the merged dataset to create the plot. To better check the accuracy, rather than using the merged data, we extract the specific data value from the corresponding dataset to make a comparison with the data in the merged dataset, which is the number being shown on the graph. It turns out, the data of daily_vaccinations, daily_new_cases, and daily_new_deaths matched perfectly on the assigned date. Moreover, question 3 was tested by another software.

Besides the assert statements, we also use another software called Tableau to test our data visualization. Tableau is a powerful visualization tool that can create beautiful interactive visualizations by simply dragging the features to the corresponding fields. We used Tableau to test the visualization for questions 1-5. Question 6 is tested by the assert function. Since the dataset we choose to continue to update, the visualization and the number on the visualization might show a slight difference depending on the input dataset. Overall, all the visualizations from questions 1-5 look the same in both Tableau and python. Here is the overall dashboard for the testing visualizations we generated. We also attached our Tableau workbook which can be downloaded and opened in Tableau Desktop to view all the interactive charts.



Collaboration:

This project is done by Yunyi(Iris)Ding and Tzuyang(Frank) Lin. We would like to thank for the help, support, and advice provided by our project mentor Suh Young Choi and the course instructor Hunter Schafer. This insightful advice provides new perspectives and also helps us to critically reflect on some potential issues related to our projects.

Aside from the course staff and team members, we search google to look for ways and new libraries to make interactive bar charts, change features of the visualization to make the chart clear, calculate statistical values(p-value, r-squared) through python.