Analyzing the Spread of Covid-19 and its Vaccinations

Jun Ahn

Summary of Research Questions

1. How is the world doing in terms of Covid cases?
    a. Covid-19 is still very active around the world. US is still leading in total covid cases
2. How is the world doing in terms of vaccination?
    a. North, South America, and Europe has the highest vaccination rate.
3. Is there any correlation between the spread of Covid-19 and vaccination status?
    a. Spread of Covid-19 is positively correlated with vaccination status
4. How effective is the vaccine? What is the mortality rate from Covid-19 among fully vaccinated individuals?
    a. Vaccine certainly reduces the risk of death from Covid-19

Motivation

It's been almost 2 years since the first Covid-19 case appeared in the US. Many of us are wondering, when will this pandemic coming to an end? When two pharmaceutical company: Pfizer and Moderna announced its release of Covid-19 vaccines, it seemed like the world has finally saw a light at the end of tunnel. Yet, a year later, we are still in the middle of one of the biggest pandemics in history.

As myself have not been keeping track with the Covid-19 and its vaccination statistics for a long time, I was curious to revisit those statistics as well as conduct analysis of the relationship between the vaccine and spread of Covid-19. By analyzing the relevant data visualization and its correlation, I hope this project provides insight on how we should move forward with the pandemic.

Dataset

https://github.com/owid/covid-19-data/tree/master/public/data

- Contains various metrics including Vaccinations, Tests & Positivity, Hospitalization, confirmed cases, and deaths among different countries
- How to download the dataset: Follow the link above, find README.md section (not the file) on the website and on the header, click "CSV" to download.

https://data.cdc.gov/Public-Health-Surveillance/Rates-of-COVID-19-Cases-or-Deaths-by-Age-Group-and/d6p8-wqjm

- Contains deaths and Age Group and Vaccination Status and Booster Dose
- How to download the dataset: find the "export" button near the top of the page, and export the file as csv

naturalearth_lowres dataset

- Geospatial dataset of the world and their respective GDP
- Downloaded on VSC via Geopandas library

Method

1. How is the world doing in terms of Covid cases?
    a. Import Panda, matplotlib and seaborn
    b. Using the panda library, filter out the relevant data: total covid , total population for from the owid dataset
    c. Use numpy library to clean up the null values (This will meet the "Messy Data" challenge goal)
    d. Visualize the statistics
    e. Create a bar chart with x-axis being each country and Y-axis being the percentage of population vaccinated (Y axis will need to be in percentage as all countries to ensure uniformity of a variable among countries with different population)
    f. Create a geospatial data to show the countries with leading covid cases

    Figuring out the current covid status of the world is crucial to conduct analysis in this project. Understanding the current situation of covid-19 cases around the world will give us basic knowledge on the up to date status of the pandemic.

2. How is the world doing in terms of vaccination?
    a. Import Panda, matplotlib and seaborn
    b. Using the panda library, filter out the relevant data: total vaccination, total vaccination per 100,000 from owid dataset
    c. Use numpy library to clean up the null values (This will meet the "Messy Data" challenge goal)
    d. Visualize the statistics
    e. Create a bar chart with x-axis being each country and Y-axis being the percentage of population vaccinated (Y axis will need to be in percentage as all countries to ensure uniformity of a variable among countries with different population)
    f. Create a geospatial data with vaccination status

    Figuring out the current vaccination status of the world is crucial to conduct analysis in this project. As vaccination status greatly differs among countries, we divide the countries into two groups: the wealthy and the poor. My logic is that wealth countries will have much higher vaccination status as they can afford to secure much more vaccines than the poor countries. By breaking up into specific groups, this will allow us to build more sophisticated analysis and ML model in the future question.

3. Is there any correlation between the spread of Covid-19 and vaccination Status?
    a. Using the data above, plot a scatter plot (exact plot function) for each country, with the x-axis being the covid cases and y-axis being number of vaccination

b. create best fit line of the scatter plot and try to look for any correlation using a 45 degree straightline (perfect fit) to quantify the correlation

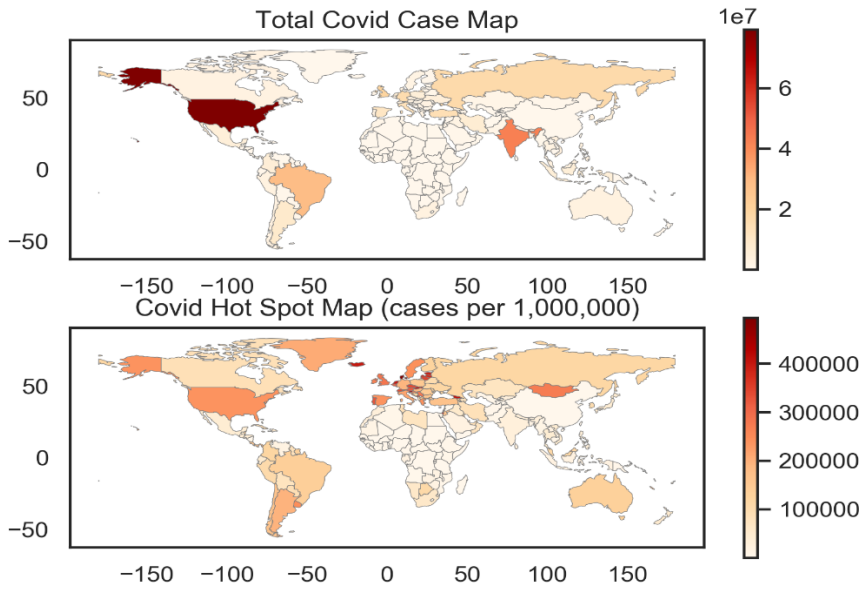
   By graphing the covid-19 cases and vaccination, we can directly measure the correlation between two variables. positive correlation will indicate that vaccination is not effective in spreading the virus whereas negative correlation indicates that vaccine is effective stopping the spread of covid.

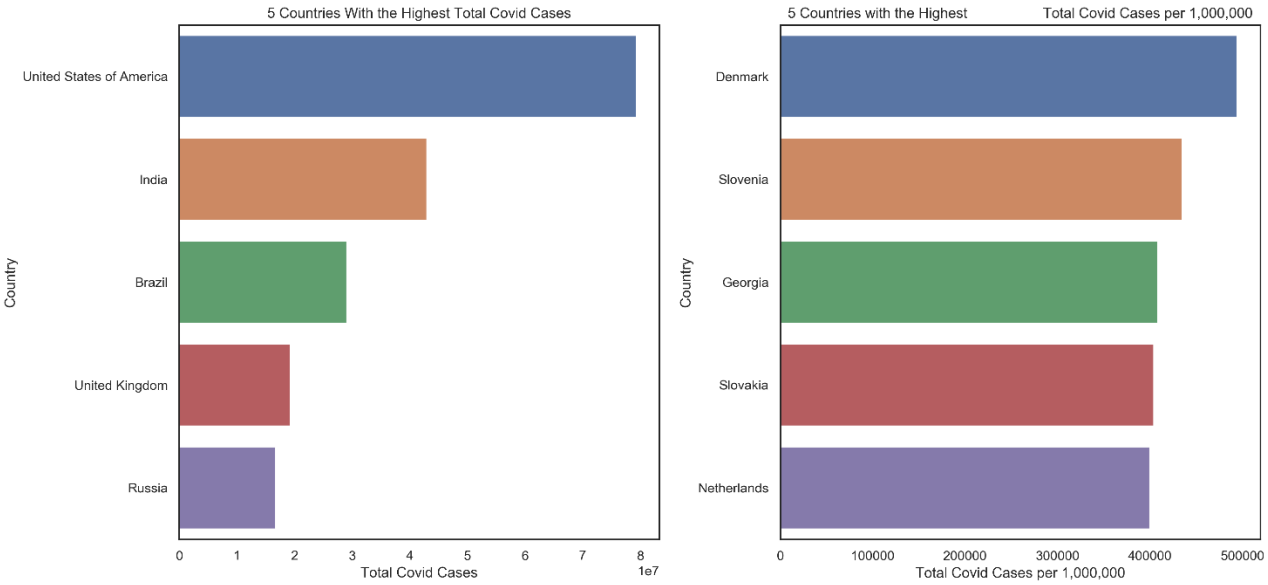
4. How effective is the vaccine? what is the mortality rate from Covid-19 among fully vaccinated individuals Vs. non-vaccinated individuals
   a. Clean the data
   b. Group by vaccine type and calculate mortality rate of vaccinated: crude_vax_ir /fully_vaccinated_population
   c. Create a barchart for mortality rate

   Visualization will showcase the mortality rate among the ones who are vaccinated which can be used to measure the effectiveness of vaccine in terms of the sickness

Results

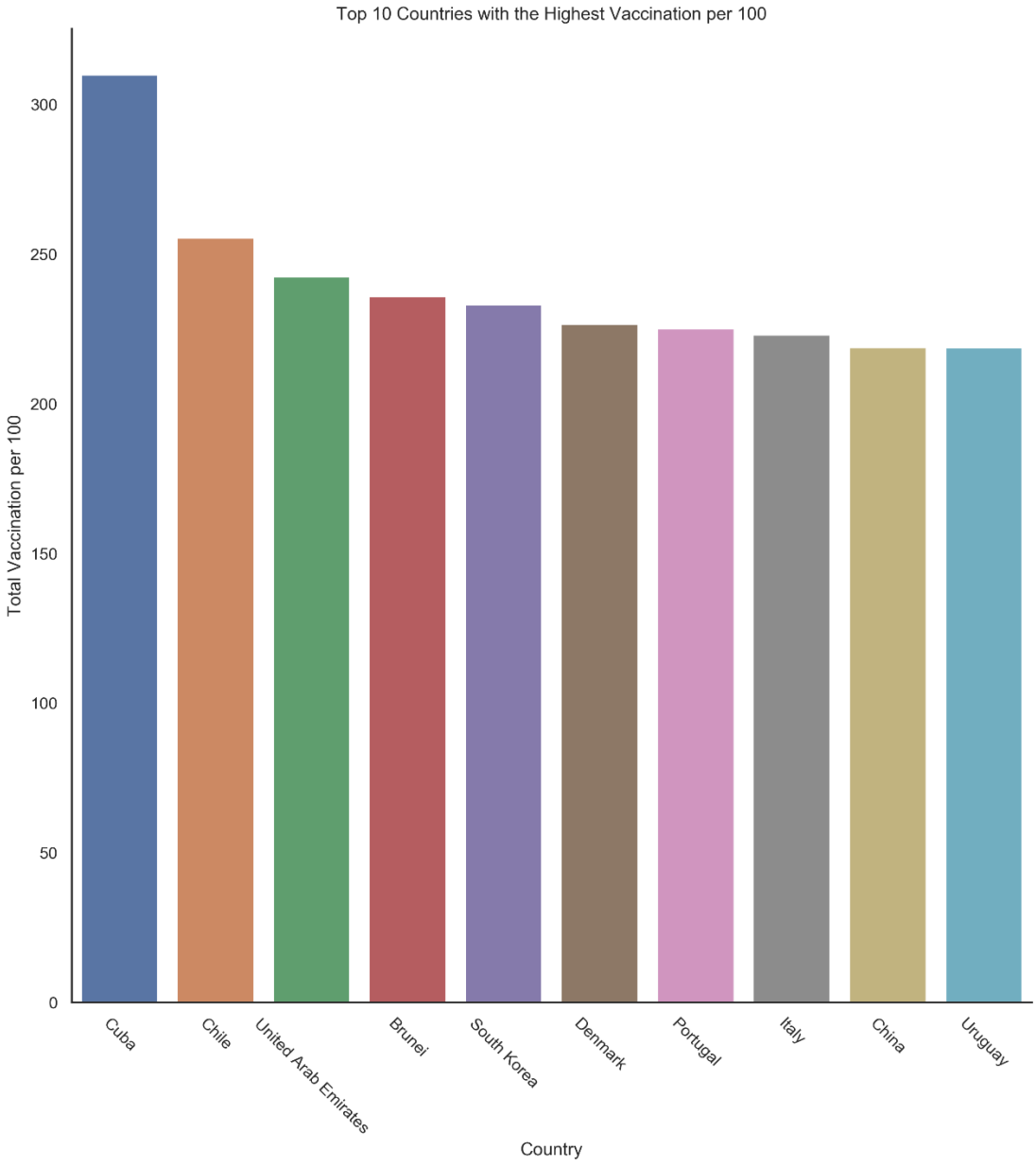
1. How is the world doing in terms of Covid cases?

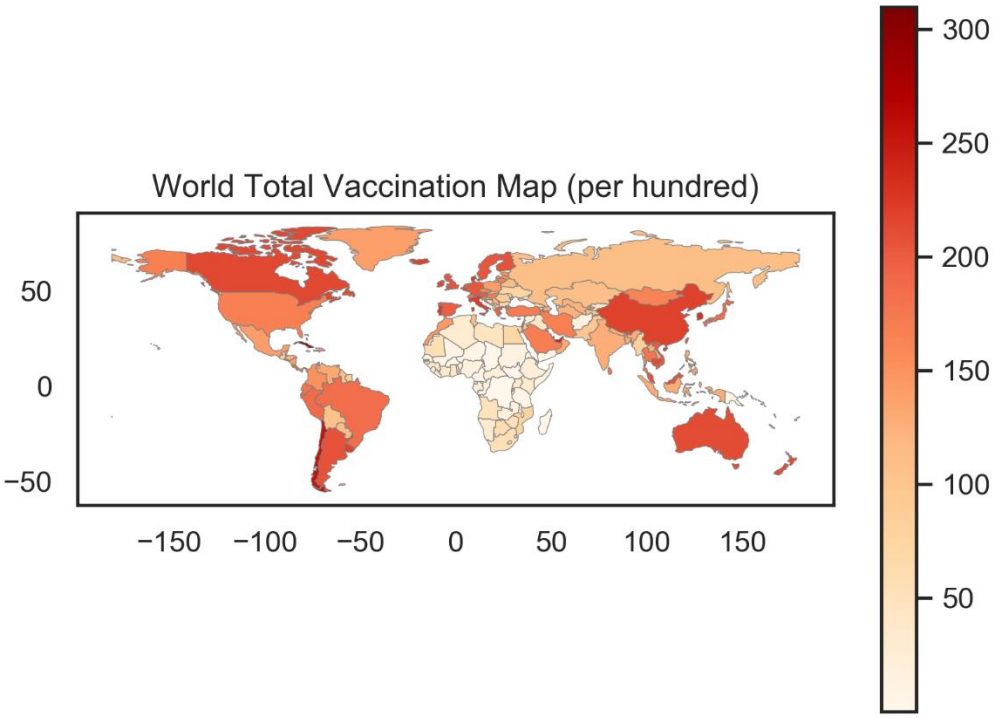




As we can see in the geospatial map and the bar chart, US is leading in terms of the total Covid cases, almost doubling India. However, Denmark has the highest total covid cases when measuring cases per 1,000,000. What this essentially means is that Denmark has the highest percentage of population infected with Covid, thus one will have higher chance of being infected with Covid-19 if they were in Denmark than the U.S.

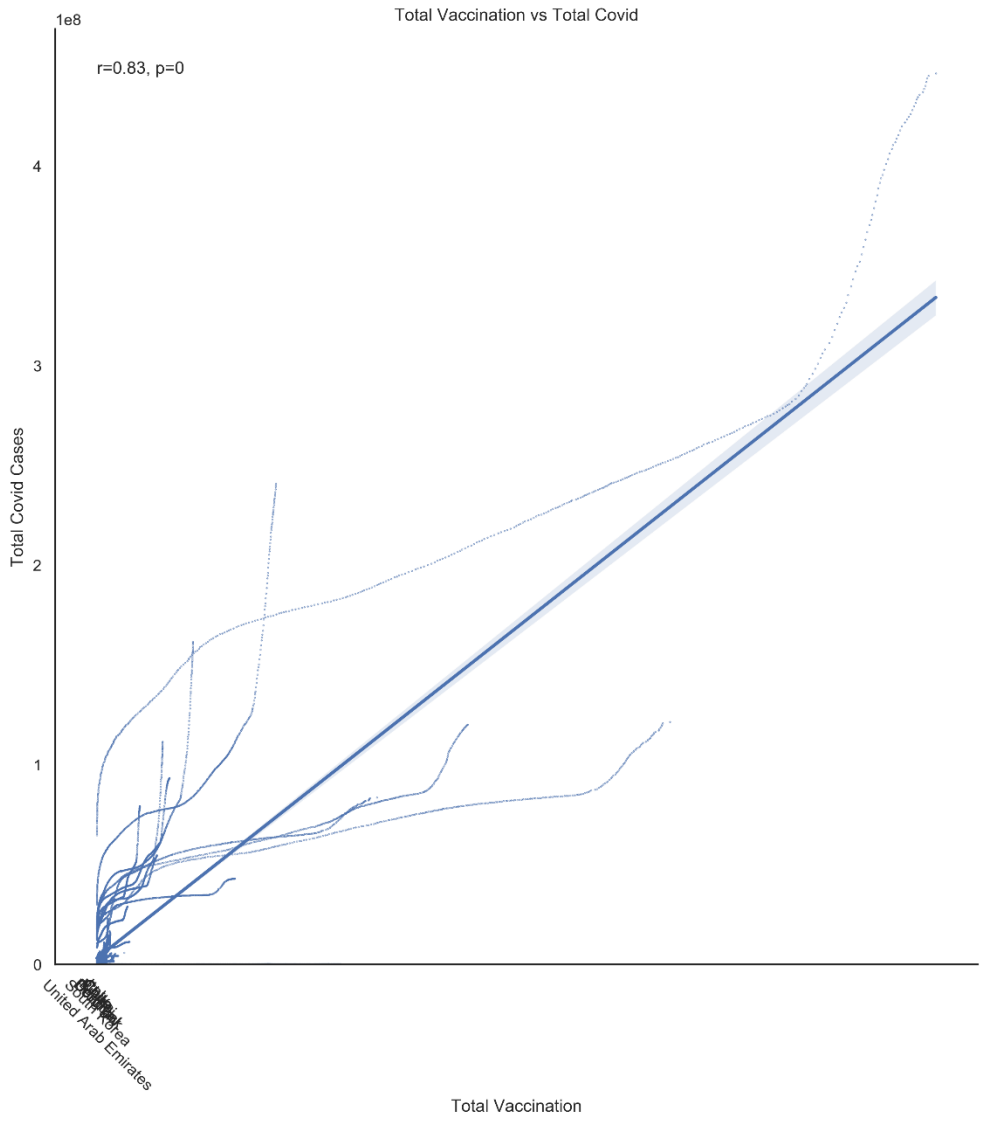
2. How is the world doing in terms of vaccination?



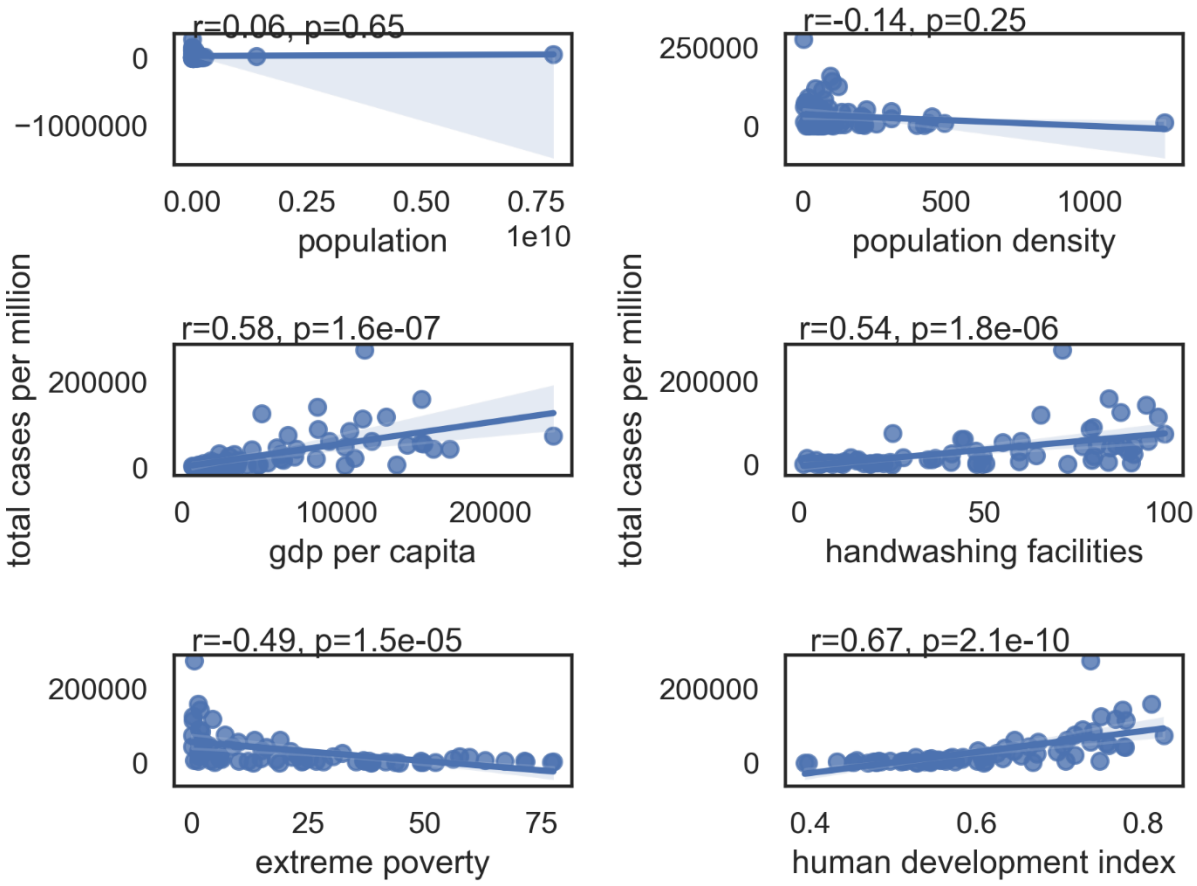


Based on the geospatial map of the vaccination status, North, South America and Europe have the highest vaccination status (as indicated by dark red color) and countries in Africa have the lowest vaccination status. What was surprising is that when measuring the vaccination status per 100 people, US is not even in the top 10 despite the fact of being the first distributor of the vaccine. Cuba is in the lead with 300 vaccinations per 100 (this means that most of people in Cuba has received a booster shot).

3. Is there any correlation between the spread of Covid-19 and vaccination status?

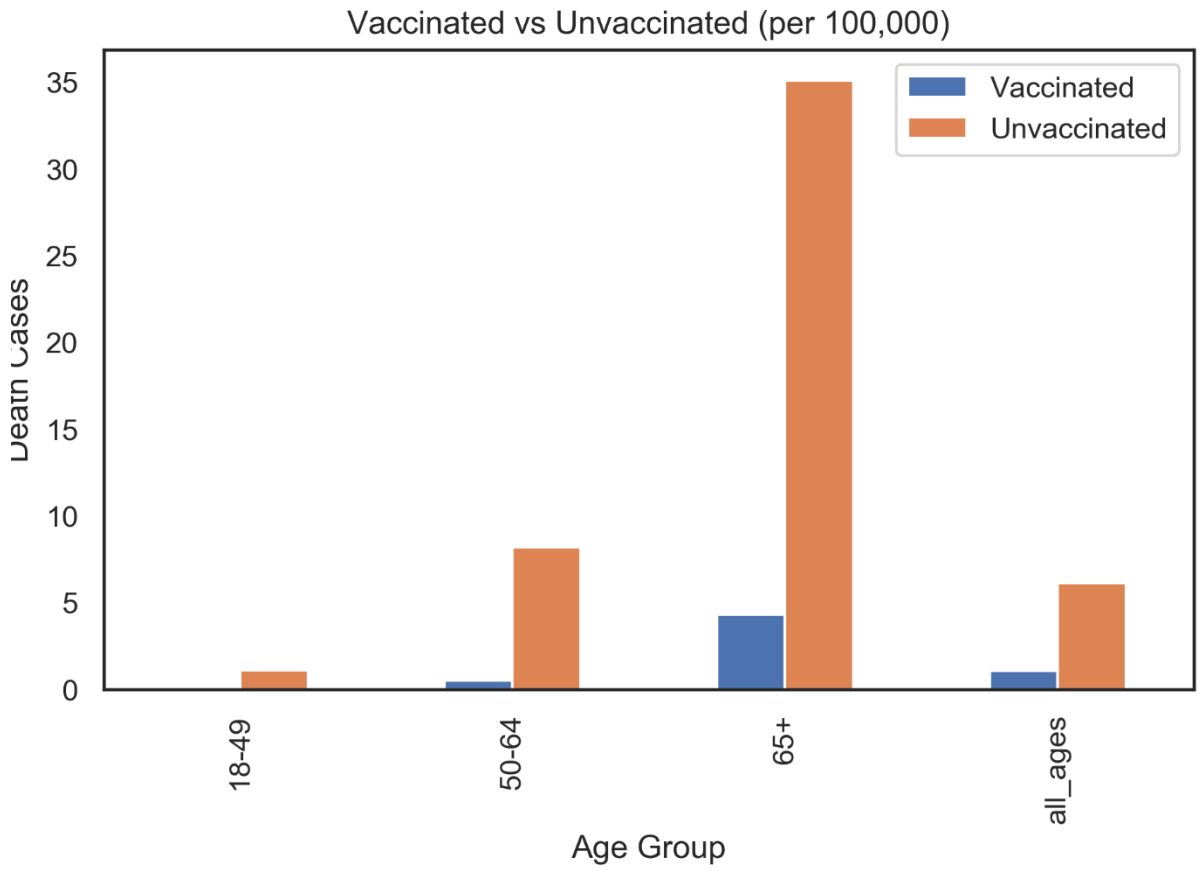


Yes. There is certainly a correlation between two variables. As seen in the graph below, when the number of Total Covid Cases and the total Vaccination of 138 countries are graphed against each other, we can observe that the general trend of the graph is upward. In fact, linear regression line is positive, which is an indication that two variables are indeed positively correlated. Lastly, the correlation coefficient r is 0.83, which is another strong indicator that two variables are positively related to each other. As the number of vaccinations increases, the Total Covid case also increases, which means is that vaccination is not effective when it comes to preventing the spread of Covid-19.



In addition, I've also plotted different factors that may effect the spread of Covid-19. One Interesting outcome of this model is that the more developed a country is, the more Covid Cases it tends to have. This can be found in the GDP per Capita and Human Development Index Model as these two metrics are positively correlated with Covid cases per million.

4. **How effective is the vaccine? What is the mortality rate from Covid-19 among fully vaccinated individuals?**



From the previous analysis, we concluded that Covid Vaccination is not effective in terms of preventing the spread of disease. However, the vaccination is certainly effective when it comes to preventing death from the virus. As we can see in the chart above, there is a large difference between the death cases of someone who is vaccinated and unvaccinated. The difference gets even bigger for the older age group who is far more vulnerable to the virus.

Impact and Limitations

As everyone's lives have been impacted by Covid-19, I believe whoever reading this paper can benefit from this project. It never hurts to know the most up to date statistics about the pandemic, and the vaccination status. However, readers should be warned that correlation does not mean causation. For example, because the number of Total Vaccination is positively correlated with the number of Total Covid cases, this does not mean vaccination causes Covid-19 to spread rapidly and vice versa. Thus, any correlation analysis is not a definitive measure to define the cause of covid-19 spread.

Challenge Goals

1. **Messy Data**: the dataset I picked is very messy and will need a lot of clean up. As there are over 50 columns 160,000+ rows, there are many rows with missing data which surely will have to be cleaned up.
2. **Machine Learning**: I am planning on using machine learning to hopefully predict the covid trend in the future. It will be a difficult task to build an accurate model due to the unstable nature of the cases, but I hope to try my best.

I had to drop Machine Learning challenge Goal as It was too difficult to for my model to fit into. However, I ended up expanding the Multiple Datasets challenge Goal as I used 3 different datasets for this project. In addition, I also used a New library scipy for regression and correlation analysis which also meets the Result Validity and New Library challenge goal.

Work Plan Evaluation

Pick relevant details needed for the analysis (ex. labels and features for machine learning model)

- Predicted: 2~3 hours
- Actual: 2 hours

Set the structure of the code (importing libraries, reading in the csv file, defining relevant functons with method headers)

- Predicted: 2~3 hours
- Actual: 4 hours

Clean up the data and filter relevant columns

- Predicted: 3~4 hours
- Actual: 2 hours

Implement each function

- Predicted: 10 hours
- Actual: 15 hours

When developing the code, I will try to be efficient as possible by not creating unnecessary loops, using functions to avoid redundancy. Testing is on the tricky side as I do not know the discrete answer to the question (for example, graph is hard to test as I do not know what the graph is supposed to look like). However, I can certainly run test on filtering and data clean up to check for any null values and correct columns have been filtered. I am working on this project by myself, but I will reach out to the TAs and stack overflow will come in handy when I run into problems.

Collaboration

I did not collaborate with anyone.