# Global COVID19 and Worldwide Vaccination Progress Analysis

***CSE 163 Part 2 ~ Christopher Ku***

**Summary of Research Questions:**

1. Which countries has the highest rate of vaccinations in terms of its population?
   - Bar plot of developed countries with their rate of vaccinations

2. How does population density affect the growth rate of cases within different countries?
   - A regression plot of the average growth rate of COVID-19 in all countries against its population density.

3. How does the HDI of a country affect the mortality rate and the amount of testing within different countries?
   - A regression plot of death rate and testing rate in each country against its HDI.

**Motivation and background:**

The problem that I am currently investigating is highly important, because it is a global pandemic that has changed the lives of everyone drastically around the world. Especially for us as students we are forced into remote learning to combat the spread of COVID-19 on campus. This is a current and significant issue to us all, if we are able to analyze data that is made available to us we could possibly combat the virus with better policies, vaccine distribution methods, improved prevention measures etc.

**Dataset:**

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

**About the dataset:**

The dataset I would be using for my project would be a dataset that is compiled by Our World in Data. There are a total of 60 different columns within the dataset each containing different information about the current pandemic. Upon further inspection of the given CSV file I found that there are several missing data that needs to be cleaned out. The dataset is consistently updated and maintained, in which it contains data that dates as far back as the 23rd of January 2020.

**Methodology:**

Technology Stack: Plotly, Seaborn, Pandas, MatPlotLib

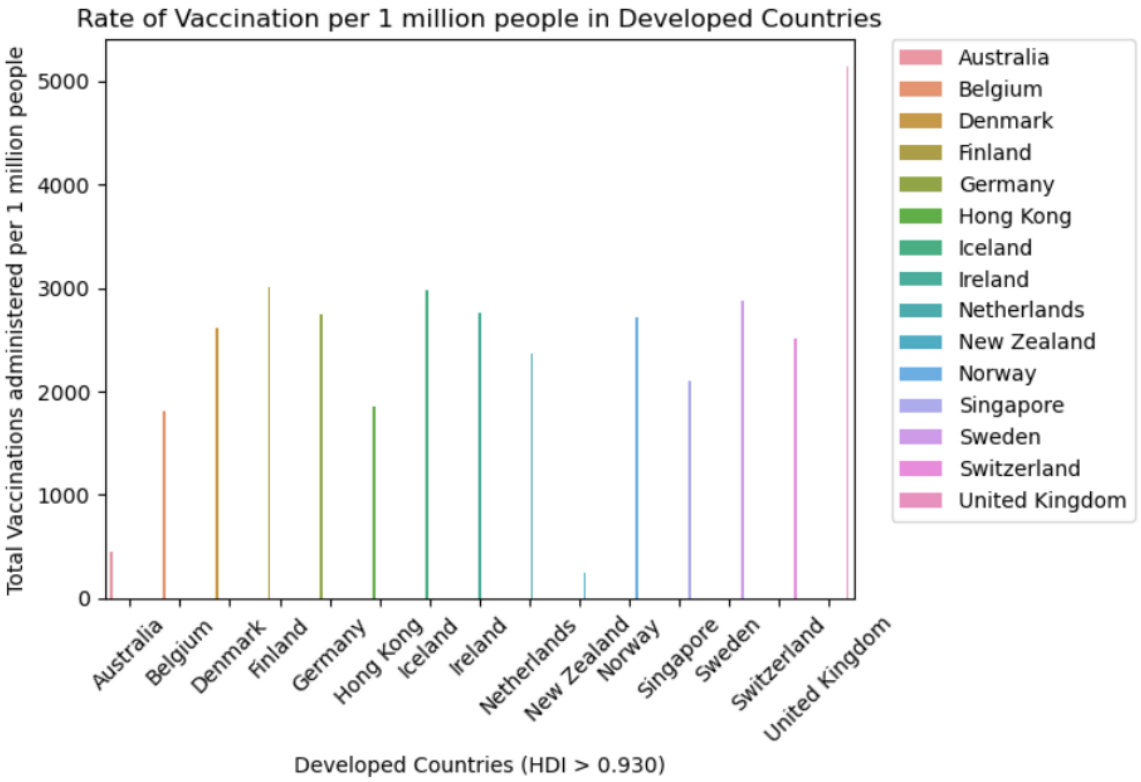
First and foremost, I would be performing data cleaning and this would begin by filtering out unneeded data such as rows with NA values, and also specific to my dataset since I am analyzing individual countries and their metrics I would need to also filter out confederations or groups of countries included within my dataset. So this first step relates to the challenge of messy data which is to understand the data and its impurities and filtering them out.

Secondly, I would be performing data filtering according to the three research questions I set myself for this project. Since all three of my research questions are country specific I would be making visualizations for each of the three research questions. After implementing I realized that there are exceptions that must be made, this is because the current data includes way too much countries in order for us find or identify any sort of discernable pattern from the plots that we make. Therefore, we would need to filter countries by Developed and Underdeveloped categories in order for better a better look at different trend lines. In this case we would be using > 0.930 as Developed Countries, and < 0.600 as Underdeveloped countries. Lastly, we would also need to remove any significant outliers for the purposes of regression plotting.

For the first question I will be making a plot of rates of vaccinations to population ratio against all of the countries in the dataset. I would need to perform extra calculations which would entail looking for a specific time period and dividing that time period by the difference in vaccinations between the time period for the rate of vaccinations create an appropriate visualization. For the second question, I would make a scatter plot of every country's population density against their specific growth rates of COVID cases. In this case growth rates would be calculated in terms of total difference in cases in a time period over a specific time period. Also, in order to identify which country each data point belongs to, we could use the Plotly library which would include interactive elements to my data visualization and satisfies my second challenge of in-cooperating and learning a new library. Lastly in order to answer my very last research question, I would need to calculate the mortality rates for each country individual based on the total deaths and total cases recorded within the country, then I would be able to make a scatterplot of mortality rates against HDI, for this I would also be using Plotly to show which country each of the data points belong to.

**Results:**

Which countries has the highest rate of vaccinations in terms of its population?

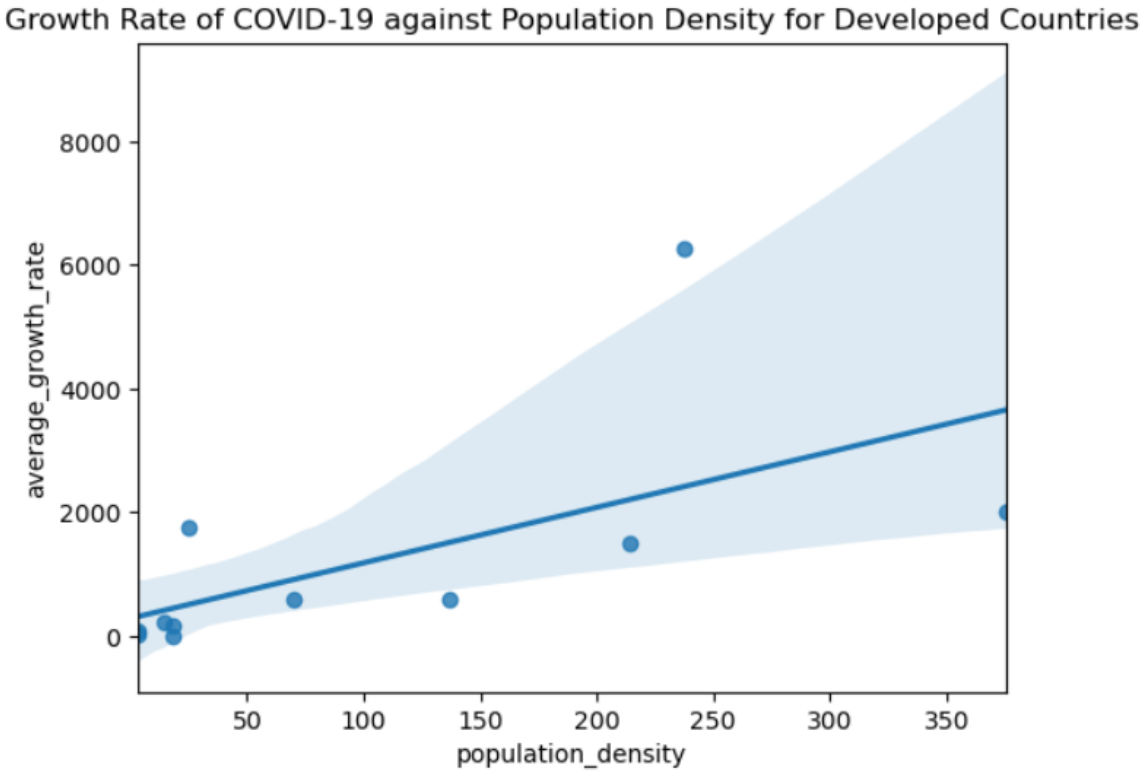


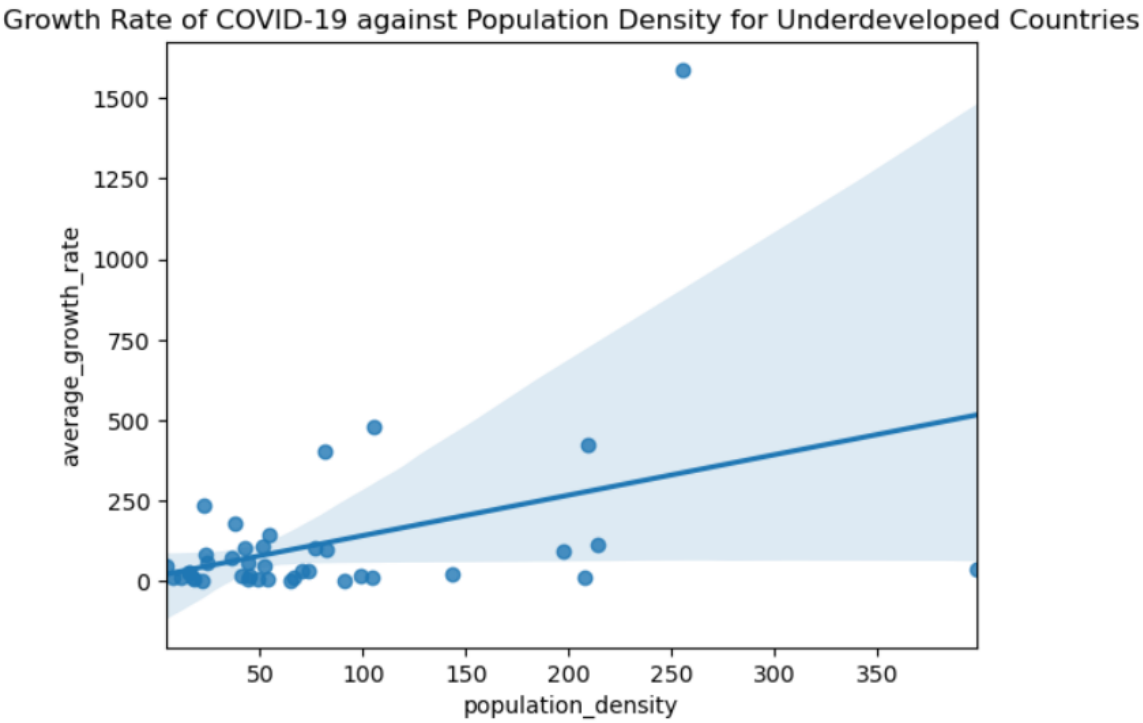

For the first research question, my code yielded the 2 graphs as shown on the previous page. We can see that within developed countries the United Kingdom is the possesses the highest rate of vaccination.

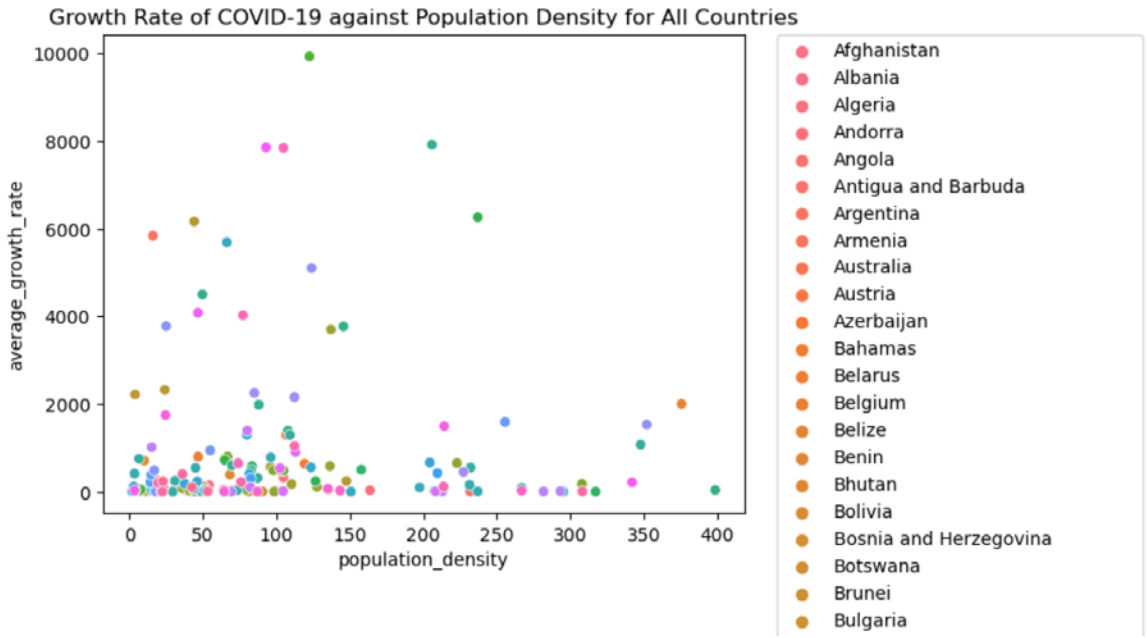
One reason for this to be the case is because of their involvement in manufacturing one of the world's major vaccines. As an effect domestic production may had made me for accessible for the population in the United Kingdom. The reason why we made the assumption that the highest rate of vaccination would exist in a developed is because they are the only countries in the world that are either distributors or buyers of the vaccine. Therefore, we could only assume that the country with the highest rate of vaccinations would be amongst them. So instead of looking over all of the countries we could just look at a given range of countries that we know where the answer lies.

The nation with the highest vaccination rate could act as an example for the rest of the world to follow. In terms of their policies, they could greatly guide other nations in distributing and administering vaccine to their local populations.

How does population density affect the growth rate of cases within different countries?






For the second research question, the graph above and the 2 graphs before are what my code yielded. For this section we would be splitting our analysis in for each country in terms of developed and underdeveloped counties.

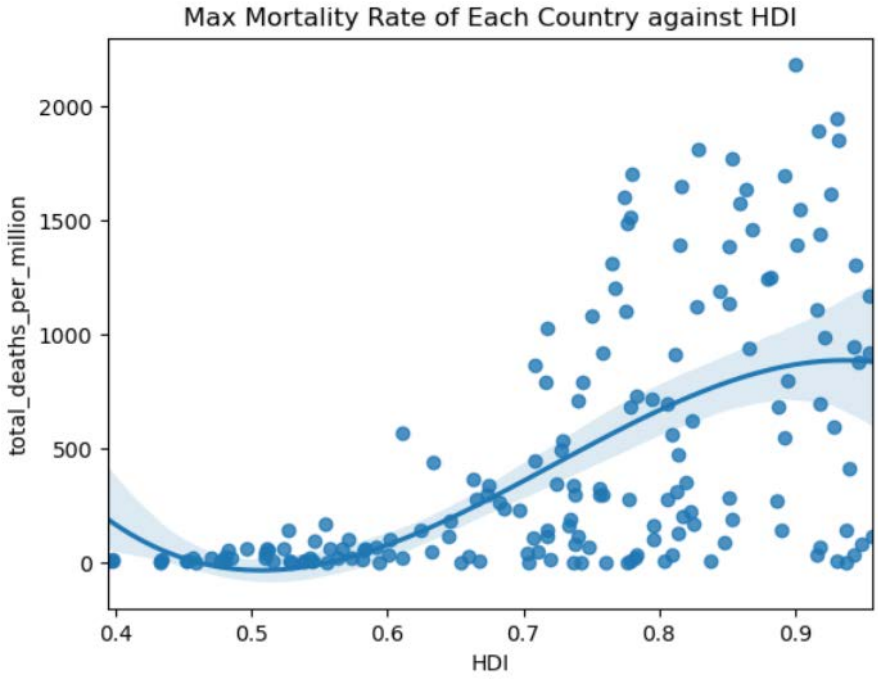
Since the average growth rate of COVID is not given to us for each country we would need to calculate that ourselves. To calculate that we just simply take the most recent number of cases recorded subtracted by the earliest number of cases recorded divided by the time period between the 2 for each country.

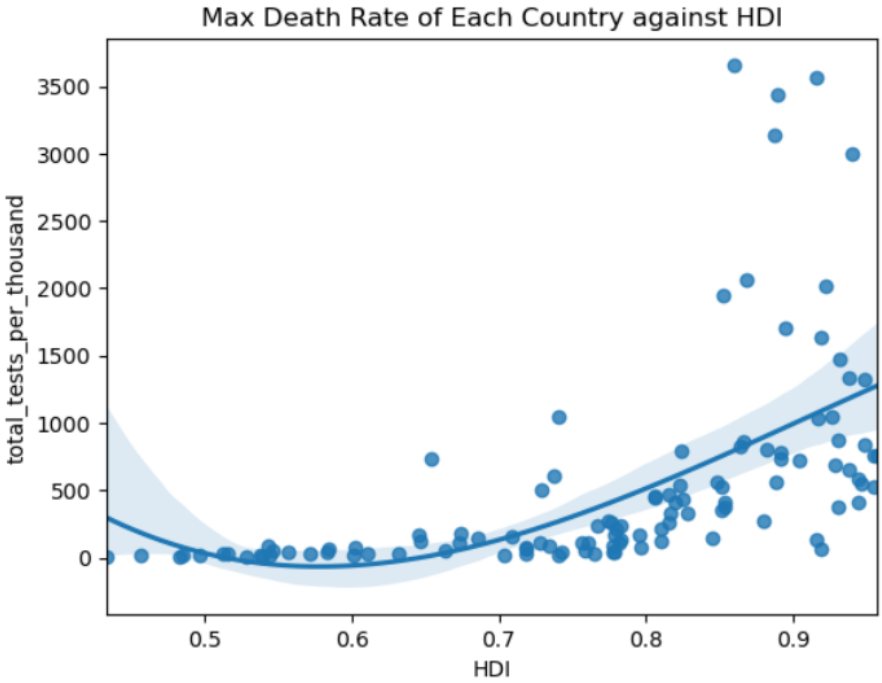
The results we got for both Developed and Underdeveloped countries are just what I expected, and this proves the assumption that as population density increases the average growth rate of the virus as increases.

From the results in the previous page we could see that regardless of an Underdeveloped or Developed nation they both have an upwards trending correlation for the 2 variables. However, what surprised me was how the trend in Underdeveloped nations are isn't as pronounced as the ones in Developed nations.

This may suggest the fact that population density within Developed Nations are way higher than the ones in Underdeveloped nations, which shows why the trend is more pronounced in developed nations. This shows how urban and a country's inherent characteristics could also play major role in determining the state of the pandemic in the country

How does the HDI of a country affect the mortality rate and the amount of testing within different countries?



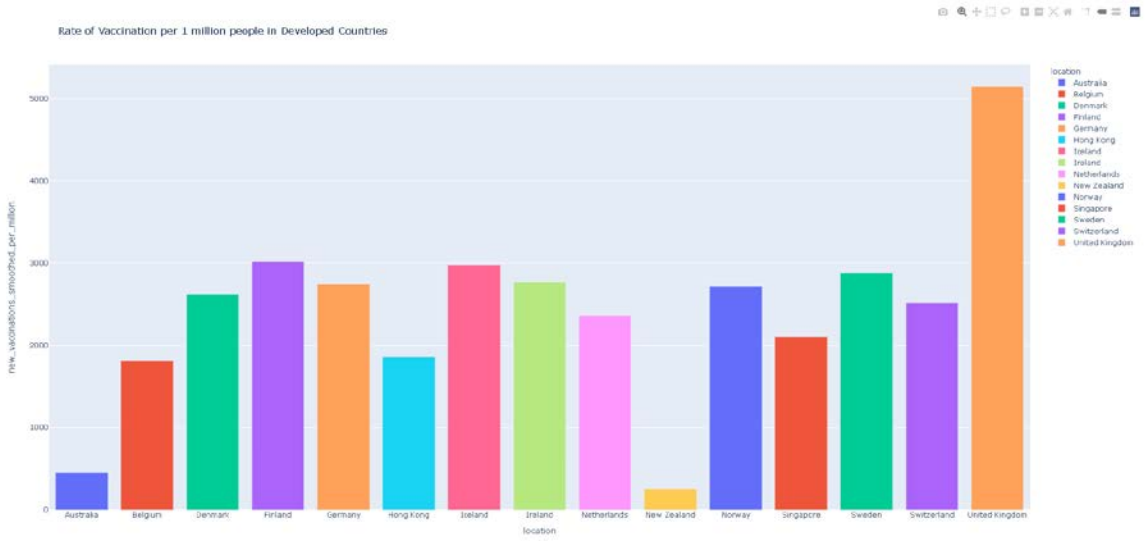
For the third research question, the 2 graphs in the previous page are what my code produced. The reason I decided to use an order 3 regression is because given the shape of the data points I believe it would be more appropriate, and it indeed yielded a better approximation.

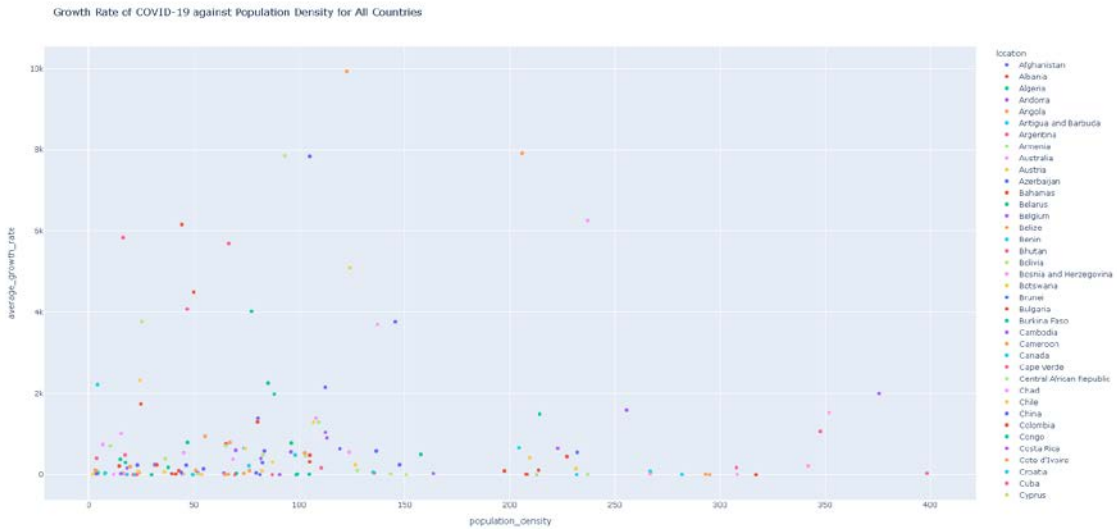
So for both graphs we can see that there is a clear upwards trend in the fact that as HDI increases both testing rate and mortality rate increases. I was surprised by how mortality rate also increases as HDI increases. This is because an increasing HDI means a more developed nation with more resources to handle to deal with the pandemic, but according to the data we have that's not the case.

The higher the country's HDI implies that governments within those countries are better and has more resources to administer tests, which makes sense why as HDI increases the testing rate within a country also increases.

A possible reason as to why mortality rate also increases as HDI increases, is because high HDI countries tend to have a higher population density meaning that the chance of infection is higher which implies a higher mortality rate.

*Interactive Visualizations*






**Challenge Goals:**

- Messy Data

As mentioned before the dataset I would be using for this project contains quite a lot of missing data and one major issue is that there are countries included within the country column of the data that aren't individual countries but rather groups or confederation of countries, so for having an in depth and correct analysis on the dataset I would be using for this project it would be crucial for me to clean the data and get rid of existing messy data. I believe my project completes this checkpoint you would be able to see in each method I would need to filter and clean data specifically tailored to what is being answered.

- New Library

In class we have only explored the seaborn data visualization library, which mostly produces static visualizations and plots. However, there are also interactive data visualization libraries that allows you to alter the data from a user level. For this project I want to explore more about these new interactive visualization libraries like Plotly and create data visualizations that are tangible. I believe there are numerous data features contained within the dataset that could be explored and used in conjunction with these interactive data visualizations libraries, which can allow users to explore the data with a more hands on experience. For my project I decided to explore the Plotly library which is a tangible way to represent my data visualizations, and I have successfully done so for 2 of my plots.

**Work Plan Evaluation:**

In terms of my work plan I will divide it up in terms of my three research questions

- Data Cleaning
  **Approximately 30 minutes**

- Which countries has the highest rate of vaccinations in terms of its population?
  **Approximately 2 hours**

- How does population density affect the growth rate of cases within different countries? – For this I would be making more subplots and additional visualizations
  **Approximately 3 hours**

- How does the HDI of a country affect the mortality rate and the amount of testing within different countries? – Takes time to learn a new library
  **Approximately 3 hours**

I believe my estimates were pretty spot on a good in the sense that it gave me an appropriate amount of time for each section to fully address the research questions to the best of my ability.

**Testing:**

I used the code that was given to us in all of our assessments for writing asserts. I managed to come up with tests for the last 2 methods in my code. However, I wasn't able to come up with tests for the first method using assert because of an out of place bug. For some reason my code runs fine in main but whenever I run the first method in my testing file it always throws a ValueError. Therefor for the first method I just use print statements to make sure that each dataframe processing at each step is correct and matches the outcome I would expect. For the other 2 methods I wrote my own CSV file and calculated the values on my own to test it with what my 2 methods returns in an assert statement. So I believe my code is pretty reliable and accurate as my tests all passes.

**Collaboration:**

- For this project I worked on my own, and whenever I am stuck I always used Stack Overflow and all of the relevant course materials for the section that I am stuck on.