World Based COVID-19 Vaccination Tracker Preprocess and Analysis

ABHISHEK PUPPALA
UID : 11596928
PROJECT - 3
INFO 5709 DATA VISUALIZATION AND COMMUNICATION

World Based COVID-19 Vaccination Tracker Preprocess and Analysis

Introduction:

One of the most devastating infectious disease crises ever recorded is a public health disaster like COVID-19. Over millions of people have died and suffered as a result of this worldwide. The best chance or likelihood of stopping the pandemic is improving the immunity of every person by providing the vaccines to each person. Tracking and analyzing the development of vaccination around the world is the most important asset for evaluating the effectiveness or improvement of immunity for everyone. One of the easiest and best ways to understand or display data is through visualizations. Researchers can detect or identify the developments within vaccine data much more quickly and highlight the areas of interest for more analysis and research by visualizing it. The data gathered can be utilized to create health regulations to prevent future pandemic outbreaks, once the patterns have been analyzed. The COVID-19 vaccination tracker data helps to identify the number of vaccines administered, and the number of vaccine percentages that are required for the world population, number of people vaccinated with single dose and fully vaccinated. At what rate, vaccines are administered in each country all over the world. The primary objective of this project is to consistently track and evaluate worldwide updates on vaccinations that are being administered, which helps to improve immunity and prevent covid.

KeyWords: Covid-19, Vaccination, Visualization, Analysis, Tracker's data.

Related Work:

For project-3, I have used the dataset, which is COVID-19 vaccination tracker data, that I found on Kaggle, which contains a huge amount of data that helps us to generate effective visualizations. Also I found some of the paperwork and articles related to my covid-19 data.

According to Mathieu et al. (2021), the collection of data includes daily immunization rates, and breakdown of the total number of COVID-19 vaccinations given in each country according to number of doses. Combining these variables can assist users in determining the scope and rate of vaccines administered related to population. Comprehending variations in administered vaccine rates across countries and analyzing the importance of countries with 3 dose schedules against 2 dose schedules. The above dataset is used by journalists, researchers and the public. Therefore, this dataset is used by the World Health Organization (WHO) to develop the COVID-19 dashboard (https://covid19.who.int/info). Moreover policymakers also use this dataset to compare international country vaccination strategies. The WHO has also been able to quantify the discrepancies in access to vaccines around the world using the dataset and

they are using these results to support for more funding for COVAX, the WHO sponsored worldwide program to increase availability to covid-19 vaccines.

According to Li et al. (2021), everyday recently recorded COVID-19 cases and fatalities in every country worldwide were compared according to vaccination rates. Everyday, the data on recently vaccinated, completely vaccinated, new cases, deaths and recovered cases for COVID-19 have been obtained over 187 countries. Using a additive generalized model (GAM), the study analyzed the connection between daily vaccination recipients and COVID-19 death rates and new cases. Moreover, the worldwide pooled results were also calculated using a random effective meta analysis.

According to Alkan et al. (2021), while performing the biometric analysis , a VOS viewer visualization tool was utilized to highlight how articles on COVID-19 vaccines are rapidly shifting at a time, when the most effective vaccination has not been discovered. These publications were written in a total of 9 different languages, in which, most of the part was written in English(97%). On an average, a total of 7.73 citations were found. At first, the identification of covid-19 was found in a very complex way. As the pandemic was rapidly increasing, it led to providing vaccinations with single and full doses. After the immediate breakout of covid-19, as there were no vaccines, they started with the evaluation of developments over vaccines.

Methods:

Exploratory data analysis:

The EDA term refers to exploratory data analysis, which is used for data cleaning and removing null values etc. The strategy of this is to analyze the data before creating the visualizations. Examples such as finding null values, duplicated data and outliers are key goals of this data analysis. Before making any type of decision, data scientists or analysts can find the patterns in data using this kind of analysis. Thanks to EDA, through which, businesses can learn more about customers and expand their business by taking well taken decisions.

In analyzing the data, the very first step of the process is exploratory data analysis. Error detection, testing for assumptions, choosing the appropriate models, figuring out how different explanatory factors relate to each other and measuring the link between them are important aspects of EDA. It also helps in data set cleaning as well as understanding its structure.

It consists of the following elements:

- Interpret the variables or data

- Cleaning the data
- Analyzing the relationships between each other features

a) **Interpret the data**: From the Kaggle site, I have taken the dataset called "Global_Covid-19_Vaccination_Tracker". It contains the features such as countries and regions, Doses administered, Doses enough for % of people, % of population with at least 1 dose, % of population fully vaccinated, daily rate of doses administered.

b) **Cleaning the data**: I have cleaned the data by first analyzing the data with finding the missing values for each feature and then filling the missing values with appropriate operation according to the data distribution. Then validated, if any duplicate instances exist.

c) Analyzing the relationships between each feature: I have analyzed the data for the whole dataset. And found the features which shows the better visual representation and gives us the insight of the data for our research questions.

Code:

PORJECT-3 : INFO_5709_COVID_VACCINE_TRACKER_VISUALIZATION

In [1]:
```
# importing required libraries for plotting the data visuals and data processing
# importing pandas for python data analysis operations
import pandas as pd
# importing numpy for numerical python calculations
import numpy as np
# importing matplotlib library for plotting the visuals
import matplotlib.pyplot as plt
# importing seaborn for plotting the visuals
import seaborn as sns
```

In [2]:
```
import codecs # importing codecs to open the files with encoding utf-8
# filename: path of the file, we are opening
file_name = "Global_COVID_Vaccination_Tracker.csv"
# opening the file with help of with statement and utf-8 as encoding and ignoring other errors
with codecs.open(file_name, 'r', encoding='utf-8',
                 errors='ignore') as fdata:
    # reading the csv data file.
    df = pd.read_csv(fdata)
# displaying the first five rows of the data.
df.head()
```

Out[2]:

	Countries and regions	Doses administered	Enough for % of people	Percentage of population with 1+ dose	Percentage of population fully vaccinated	Daily rate of doses administered
0	Global Total	5.663213e+09	NaN	NaN	NaN	33380378.0
1	Mainland China	2.129833e+09	76.1	NaN	69.3	6454714.0

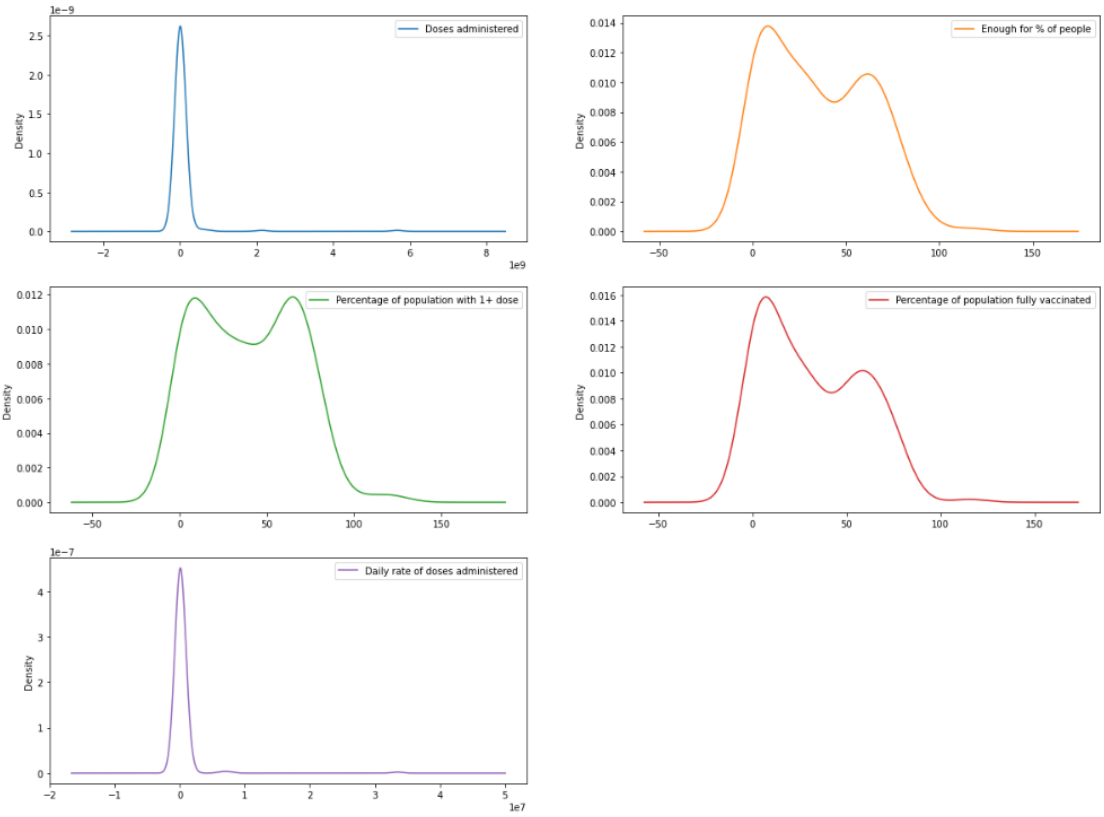
EDA : Exploratory Data Analysis

In [3]:
```
# validating all the columns having null values
df.isna().sum()
```

Out[3]:
```
Countries and regions                          0
Doses administered                             1
Enough for % of people                         5
Percentage of population with 1+ dose          7
Percentage of population fully vaccinated     10
Daily rate of doses administered               4
dtype: int64
```

In [4]:
```
# displaying the kernel density distribution for all numerical columns
df.plot(kind='kde', subplots=True, layout=(3,2), sharex=False, sharey=False, figsize=(20,15))
```

Out[4]:
```
array([[<AxesSubplot:ylabel='Density'>, <AxesSubplot:ylabel='Density'>],
       [<AxesSubplot:ylabel='Density'>, <AxesSubplot:ylabel='Density'>],
       [<AxesSubplot:ylabel='Density'>, <AxesSubplot:ylabel='Density'>]],
      dtype=object)
```



As we can see,

- **Doses administered** has normal distribution, so replacing null values with mean.
- **Enough for % of people** has skewed distribution, so replacing null values with median.
- **Percentage of population with 1+ dose** has skewed distribution, so replacing null values with median.
- **Percentage of population fully vaccinated** has skewed distribution, so replacing null values with median.
- **Daily rate of doses administered** has normal distribution, so replacing null values with mean.

In [5]:
```
# filling the missing values according to the distribution of the data.
df['Doses administered'].replace(np.NaN, df['Doses administered'].mean(), inplace=True)
df['Enough for % of people'].replace(np.NaN, df['Enough for % of people'].mean(), inplace=True)
df['Percentage of population with 1+ dose'].replace(np.NaN, df['Percentage of population with 1+ dose'].mean(), inplace=True)
df['Percentage of population fully vaccinated'].replace(np.NaN, df['Percentage of population fully vaccinated'].mean(), inplace=T
df['Daily rate of doses administered'].replace(np.NaN, df['Daily rate of doses administered'].mean(), inplace=True)
```

In [6]:
```
# re-validating if all the columns null values are filled.
df.isna().sum()
```

Out[6]:
```
Countries and regions                       0
Doses administered                          0
Enough for % of people                      0
Percentage of population with 1+ dose       0
Percentage of population fully vaccinated   0
Daily rate of doses administered            0
dtype: int64
```

HYPOTHESES:

Hypothesis-1: What is the range of difference between % of population fully vaccinated v/s % of population partially vaccinated per each country?

The best strategy for halting the transmission of covid-19 is vaccination. In this first statement of hypothesis, first I have stated a range of differences between % of population fully vaccinated v/s % of population vaccination per each country.

Code:

Showing Visualization for percentage of population with atleast 1 dose(partially) and percentage of population with fully vaccinated per each country and region.

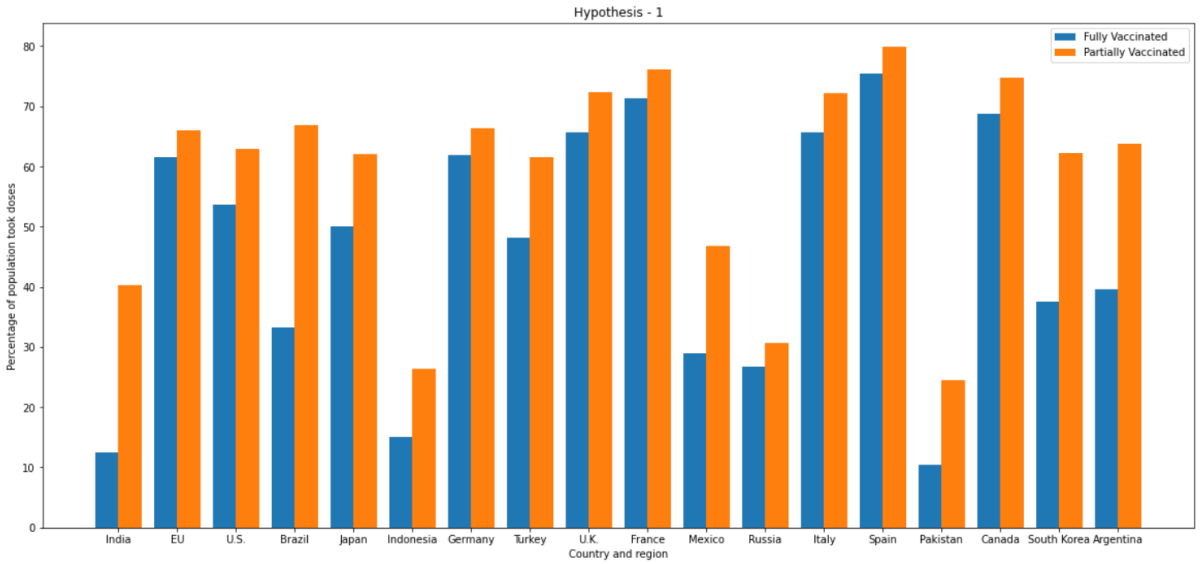
In [9]:
```
# plotting the vertical bar graph
# taking top 20 countries and regions of data
x_axis = df['Countries and regions'].head(20)
# taking the percentage population accordingly with country data
y_fully_axis = df['Percentage of population fully vaccinated'].head(20)
y_partial_axis = df['Percentage of population with 1+ dose'].head(20)

# length of num of x labels(countries)
X = np.arange(len(x_axis[2:]))
plt.figure(figsize=(20,9))
# plotting bar graph with fully vaccinated as well as partially vaccinated
plt.bar(X-0.2, y_fully_axis[2:], 0.4, label='Fully Vaccinated')
plt.bar(X+0.2, y_partial_axis[2:], 0.4, label='Partially Vaccinated')

# labeling x axis values with labels of countries
plt.xticks(X, x_axis[2:])
# xlabel and ylabel and title of bar graph
plt.xlabel("Country and region")
plt.ylabel("Percentage of population took doses")
plt.title("Hypothesis - 1")
plt.legend()
plt.show()
```

Choice of visualization:

Type**:** barplot.
Tool**:** Jupyter Notebook using Python
Colors**:** Blue, Orange
Scale: On x-axis, countries and regions
On the y-axis, % of population with 1+ dose, % of population fully vaccinated.

For visualizing this data, I used a bar plot with 0.4 width, which is a suitable way to display the comparison between fully and partially vaccinated in the top 20 countries. The bar plots are used to effectively express information because they provide the data as rectangular bars with heights that match to the highest values. Also, it makes it easier to comprehend the summary of the data presented. As the colors represent the % of the population of fully vaccinated blue and blue for partially vaccinated. The following below image displays the comparison between fully and partially vaccinated per each country.

Result:



Analysis of Hypothesis - 1:

From the above visualization, we observe that, % of population with partially vaccinated dominates over fully vaccinated in all top 20 countries. Also, Spain has more % of population vaccinated partially and fully compared to others, and Pakistan country has the least rate of % of population that are vaccinated. Both India and Indonesia have a smaller proportion of population

who have received all recommended vaccinations than Europe, Germany and France. All of the nations are having the highest % of people with 1+ dose, with the exception of indonesia. Additionally, fewer individuals in the top 20 countries have received 1+ doses of vaccination than those who have received all of the suggested doses.

Conclusion: In almost all of the top 20 countries, my prediction that the proportion of persons with 1+ dosage will be larger than the proportion of fully immunized people came true.

Hypothesis - 2: Number of doses administered for top 15 countries.

Even those who have already had COVID-19 are protected by the COVID-19 vaccine, even from being hospitalized for a subsequent infection. Getting vaccinated against COVID-19 is a safer and more dependable way to increase immunity than becoming ill with COVID-19. The COVID-19 vaccine serves to protect you by eliciting an immune response without the need for you to contract a disease, possibly even a serious one. I therefore came up with this idea since I was intellectually inquisitive about the top 15 nations that were in the running to secure the most doses administered.

Code:

Showing the visualization for top 10 countries that have been doses administered.

In [10]:
```
# plotting the line graph
# taking top 15 countries and regions of data
x_axis = df['Countries and regions'].head(15)
# taking Num of doses administered per each country accordingly
y_axis = df['Doses administered'].head(15)

# adjusting figsize and keeping xlabel, ylabel and title
plt.figure(figsize=(12,5))
plt.plot(x_axis[2:], y_axis[2:])
plt.xlabel("Country and region")
plt.ylabel("Doses administered")
plt.title("Hypothesis - 2")
plt.show()
```

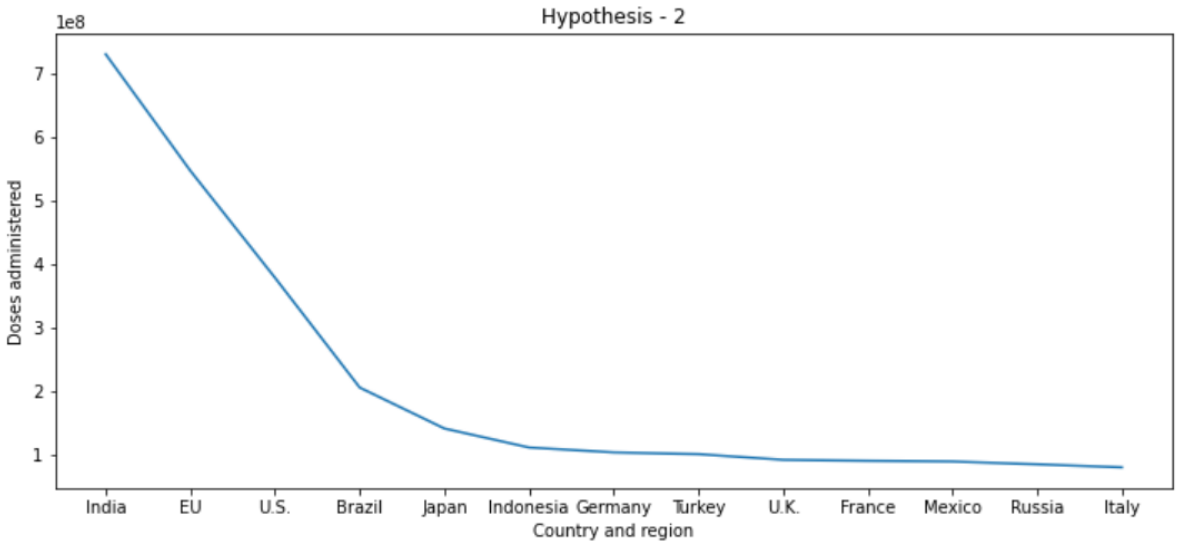
Choice of visualization:

Type: Line Plot
Tool**:** Jupyter Notebook using Python
Colors**:** Blue
Scale: On x-axis, countries and regions
On the y-axis, doses administered.

The easiest way to illustrate the comparison of the number of doses administered (in billions) in the top 15 countries, in my opinion, is using a line plot. When displaying continuous data that has changed over time, line plots are very much useful. For displaying a relationship between one or more variables, line graphs work perfectly. This kind of plot does a good job of depicting trends and change. Additionally, it is the best way to illustrate trends and changes over time.

The segments are further categorized and the presentation is made more appealing because the colors show the amount of doses administered. The images show a comparison of the number of doses administered (in billions) in the top 15 countries worldwide.

Result:



Analysis of hypothesis - 2:

The top 15 countries with the greatest number of doses, according to the visualization above, are India, Europe, the United States, Brazil, Japan, Indonesia, Germany, and Turkey. India is the nation with the most doses administered among the top 15, while Turkey is the nation with the fewest doses administered.

Conclusion:

The goal of this hypothesis is to identify the top 15 nations by the total number of doses administered. The top 15 nations with the most dosages administered include India, Europe, the USA, Brazil, Japan, Indonesia, Germany, and Turkey.

Hypothesis - 3: In industrialized nations against those in underdeveloped ones, the daily rate of doses provided will be higher.

On average, developed nations invest more than developing nations. As a result, wealthier nations spend more money manufacturing and distributing vaccines to their own population. Therefore, I assume that developed countries administer the most vaccine doses daily compared to developing countries.

Code:

Showing visualizations for daily rate of doses administered per each country or region

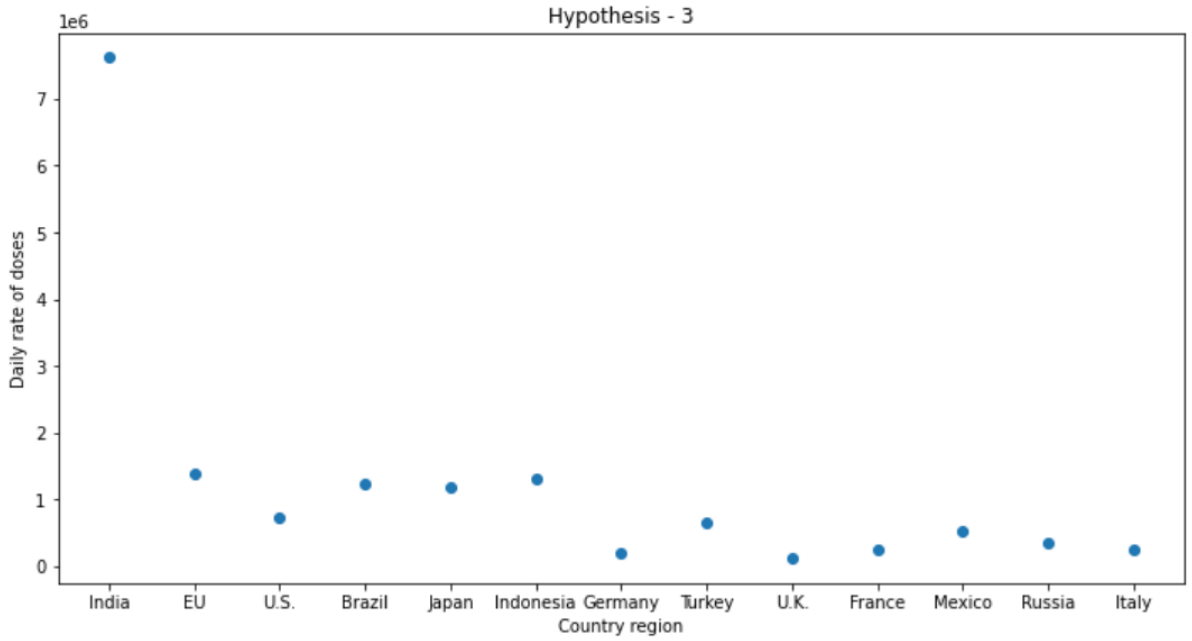
11]:
```
# plotting the scatter plot
# taking top 15 countries and regions of data
x_axis = df['Countries and regions'].head(15)
# taking daily rate of doses administered per each country
y_axis = df['Daily rate of doses administered'].head(15)

# adjusting figsize and keeping xlabel, ylabel and title
plt.figure(figsize=(12,6))
plt.scatter(x_axis[2:], y_axis[2:])
plt.xlabel("Country region")
plt.ylabel("Daily rate of doses")
plt.title("Hypothesis - 3")
plt.show()
```

Choice of visualization:

Type: Scatter Plot
Tool: Jupyter NotBook
Color: Blue
Scale: On x-axis, Countries and regions
On y-axis, Daily rate of doses administered

Line plots are the most effective technique, in my opinion, to depict the precise location of the daily rate of doses administered in the top 15 nations when visualizing this data. Scatter plots are frequently used to observe and graphically illustrate the relationship between two variables. The values of the variables are represented by dots. The value of the data point is determined by the placement of the dots on the vertical and horizontal axes.

Result:



Analysis of hypothesis - 3:

From the image above, it can be seen that wealthier nations like the USA provide less doses per day than developing nations like India and Brazil. India is the country with the greatest daily dose administration rate among the top 15; Europe is second with the highest daily dose administration rate, and Germany is last with the lowest daily dose administration rate.

Conclusion:

My assumption that affluent countries administered more vaccination doses daily than underdeveloped countries was incorrect. Compared to developing nations like India and Brazil, wealthy nations like the United States administer fewer doses everyday.

Discussion:

This visualization's fundamental purpose is to do exploratory data analysis and display the data visualization that offers tracking and analysis of updates on vaccines across the globe. It provides details on the number of vaccination doses given, the number of vaccine doses required to vaccinate the entire population, the proportion of the population who have received at least one dose, the proportion of the population who have received all recommended doses, and the daily dose rate in each nation on the globe.

According to the first of the three hypothesis 16 listed above, more persons in the top 15 nations have received one or more doses of the vaccine than have received all the recommended doses. The top ten nations with the most doses administered will be shown in the second hypothesis. The third hypothesis showed a negative result for our assumption that rich countries administer the most vaccination doses daily compared to underdeveloped countries. Compared to developed countries, poor countries administer more doses everyday.

Future work:

I was successful in conducting my analysis using the available dataset. However, if the dataset contained information on the proportion of people who had not received the vaccine, that could aid in identifying which nation or region is most at risk of contracting COVID-19. If this dataset were to accrue any additional data in the future, it would be beneficial to carry on with the study.

References:

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*, *5*(7), 947–953. https://doi.org/10.1038/s41562-021-01122-8

Li, Z., Liu, X., Liu, M., Wu, Z., Liu, Y., Li, W., Liu, M., Wang, X., Gao, B., Luo, Y., Li, X., Tao, L., Wang, W., & Guo, X. (2021). The effect of the COVID-19 vaccine on daily cases and deaths based on global vaccine data. *Vaccines*, *9*(11), 1328. https://doi.org/10.3390/vaccines9111328

Alkan-Ceviker, S., Onturk, H., Alıravcı, I. D., & Sıddıkoğlu, D. (2021). Trends of covid 19 vaccines: International Collaboration and visualized analysis. *Infectious Diseases and Clinical Microbiology*, *3*(3), 129–136. https://doi.org/10.36519/idcm.2021.70