12/10/2021
Alex Liang
Edgar Pineda
Xing Yang

COVID-19 Vaccination Analysis and Simple model prediction

Motivation

There are a few questions we want to ask our data:

1. How many vaccinations are offered daily by country?

2. What is the vaccine manufacturer breakdown by country?

3. How many people are vaccinated in specific countries?

4. What is the vaccine breakdown across the world?

5. What percent of each country is vaccinated?

6. What are the model that can use to predict when a country will be fully vaccinated?

In this project you will see the process of dealing with missing data and filling it with appropriate values.

You can also find exploratory data analysis about the questions along with visualization on Sunburst Charts, Choropleth Maps, Bar Charts, Pie Charts, and Treemap Charts in Python, the way we want to answer our questions would be through visualizations using our one combined data frame.

Moreover, at the end of the this you can find predictive model parameters choosing and making

Data

There are four pieces of data that we are using in the project.

Population_by_country_2020.csv: data related to the population of each country in year of 2020

https://www.kaggle.com/tanuprabhu/population-by-country-2020

Continents2.csv: data relate to the continents and sub-continent which is the region and sub-region of the country

https://www.kaggle.com/andradaolteanu/country-mapping-iso-continent-region

12/10/2021
Alex Liang
Edgar Pineda
Xing Yang

Country_vaccinations_by_manufacturer.csv: data related to the manufacturer of vaccine in worldwide

Country_vaccinations.csv: data related to the daily and total vaccination in each country

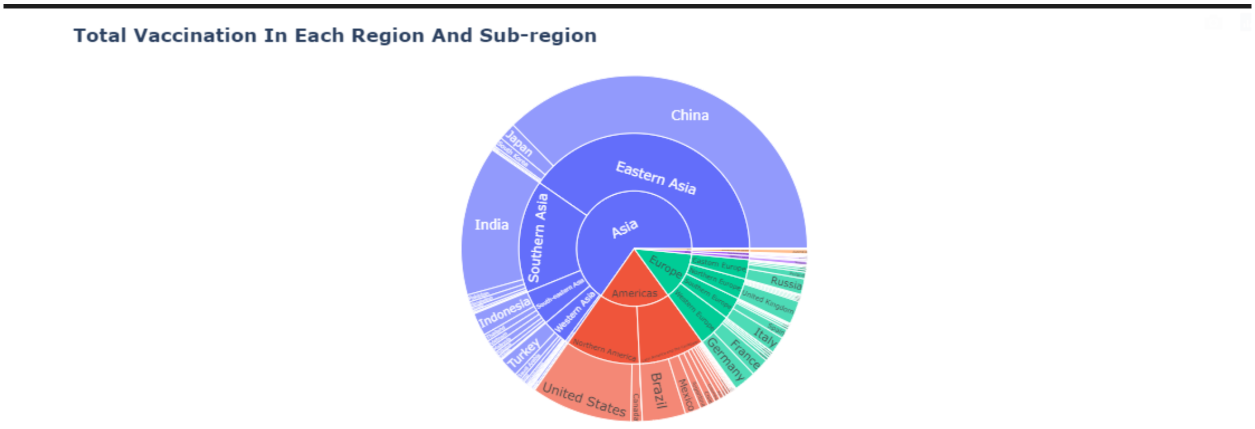
https://www.kaggle.com/gpreda/covid-world-vaccination-progress

The data will be transfer into Visualizations

Sunburst Charts in Python: https://plotly.com/python/sunburst-charts/

main.app.ipynb

This pie graph shows the total number of vaccinations in each region, sub-region and country. Each sub-region is grouped by the region and each country is grouped by the sub-region. In the inner layer, it is separated by different regions such as Asia, Americas and Europe etc.. In the middle layer, it is separated by different sub-region such as Northern America, Eastern Asia and Southern Asia etc.. And the outer layer is separated by each country. The area of each region, sub-region and country depend on the total number of vaccinations. For example, Southern Asia has less total vaccination than Eastern Asia.

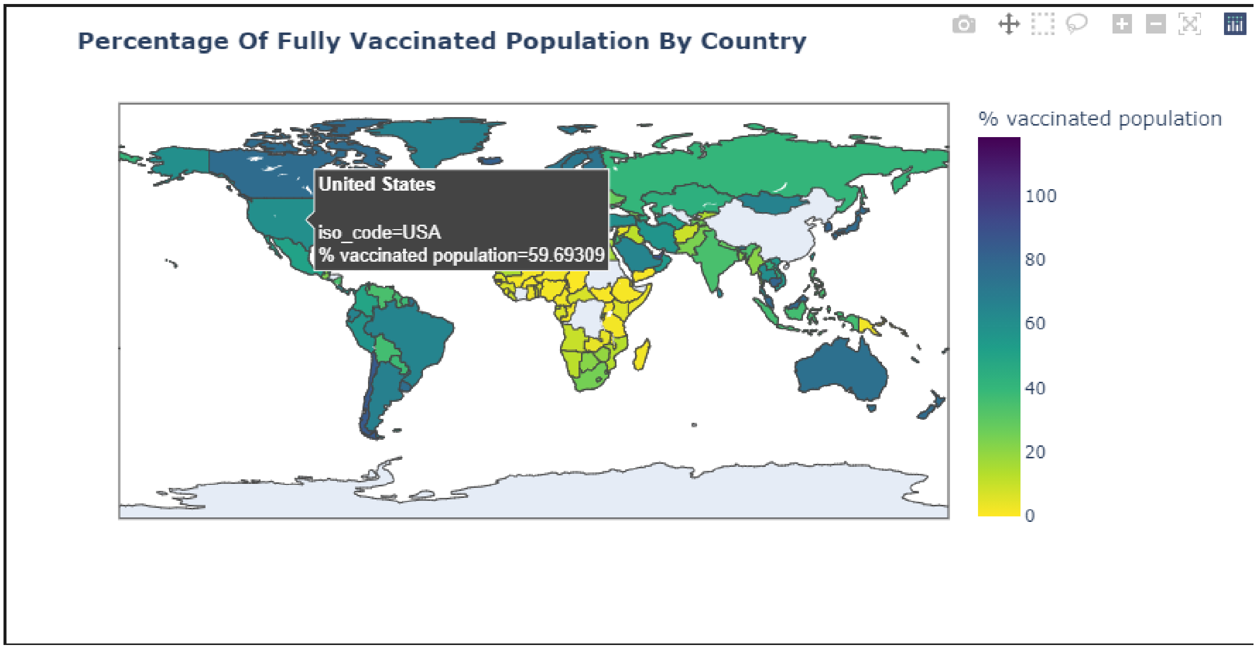


Choropleth Maps in Python: https://plotly.com/python/choropleth-maps/

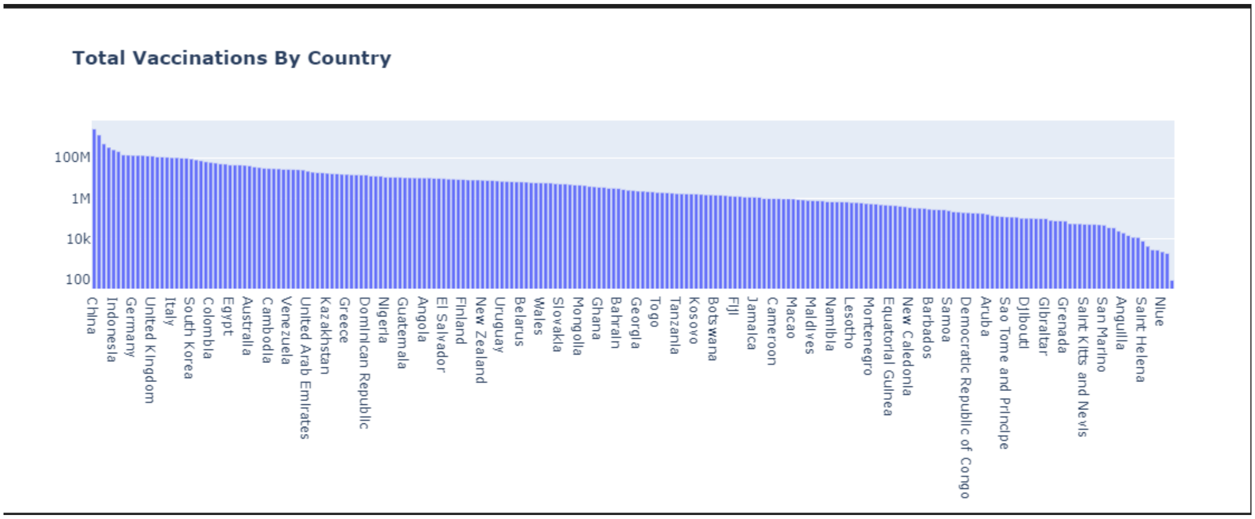
Vacci_percentage.ipynb

This map graph shows the percentage of people who are fully vaccinated in terms of the population of the country. The color is changed from light to dark, which is lower to higher percentage of people who are fully vaccinated. If the color is gray, it means that region doesn’t have the percentage of the fully vaccinated（missed data). The United States and several countries from southern America have the higher percentage of people who are fully vaccinated.

12/10/2021
Alex Liang
Edgar Pineda
Xing Yang

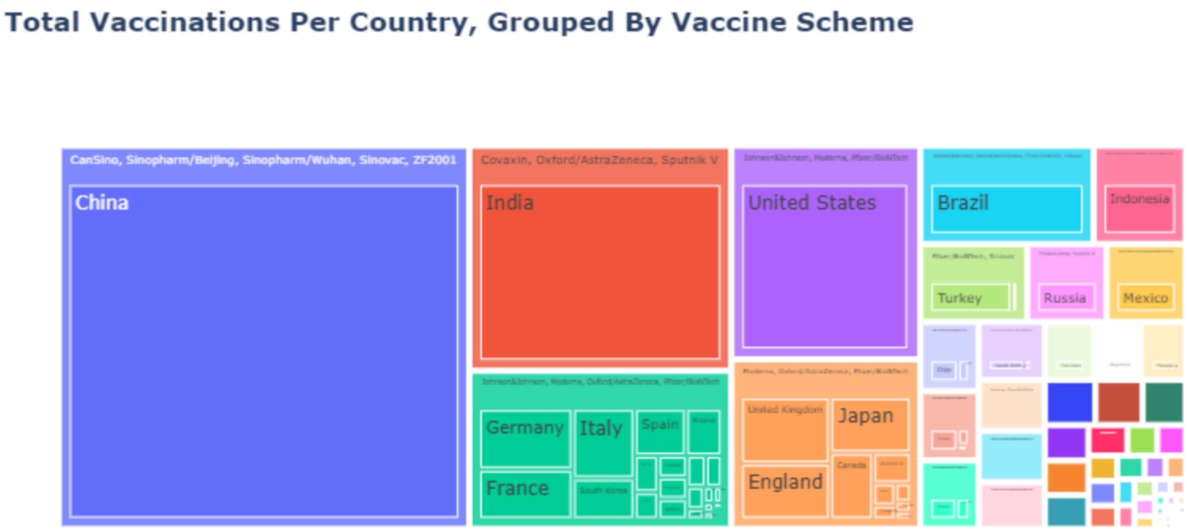

Percentage Of Fully Vaccinated Population By Country

Bar Charts in Python: https://plotly.com/python/bar-charts/
Covid_vax_vis_bar_graph.ipynb

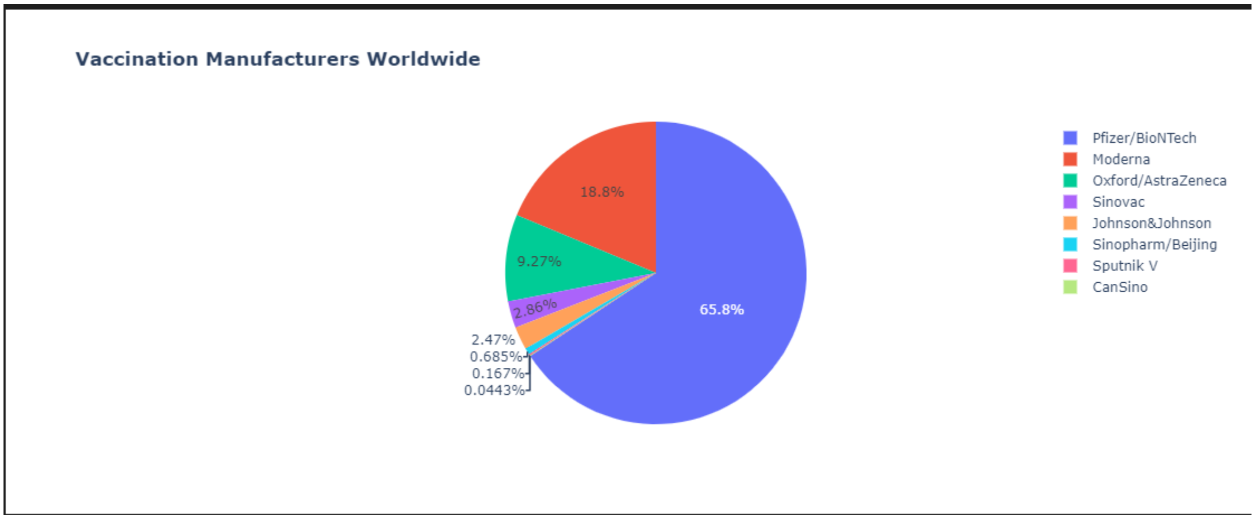

Total Vaccinations By Country

The pie chart shows the ratio between people who are vaccinated once and the people who are fully vaccinated in the top 10 countries. The blue part is showing how many percent of people are fully vaccinated. The gray part is showing how many people were vaccinated once. In the best case, the people who received fully vaccinated has more percentage of the people who received vaccinated once since it indicates many people are fully vaccinated. From the above graph, even though the United states has the highest number of people who are vaccinated, Israel has the best percentage of people who take the fully vaccinated.

Treemap Charts in Python: https://plotly.com/python/treemaps/
main.app.ipynb
The tree map shows the vaccine schema that is used in each country. For example, only the United States used three different kinds of vaccines: Johnson & Johnson, Moderna and Pfizer/BioNTech. The size of the boxes depends on the total number of vaccinations in that country. And The United States has the most vaccination in terms of the rest of the world. From the graph, we could see what the schema that each country used and which countries are using the same vaccines schema.



Pie Charts in Python: https://plotly.com/python/pie-charts/
Pie_chart_for_manufacturer.ipynb
The pie chart shows the ratio between people who are vaccinated once and the people who are fully vaccinated in the top 10 countries. The blue part is showing how many percent of people are fully vaccinated. The gray part is showing how many people were vaccinated once. In the best case, the people who received fully vaccinated has more percentage of the people who received vaccinated once since it indicates many people are fully vaccinated. From the above graph, even though United states has the highest number of people who are vaccinated, Israel has the best percentage of people who take the fully vaccinated.



Building the Model - Model.ipynb

First we need to choose a proper model for the question we are currently trying to answer which is “How many people will be vaccinated after 53 days?”. Since this question is asking to predict a number from a range it is best if we use a regression model. Before picking a regression model we have to set up the data to be numerical numbers. The two regression models we choose for this question are linear regression and decision tree regressor.

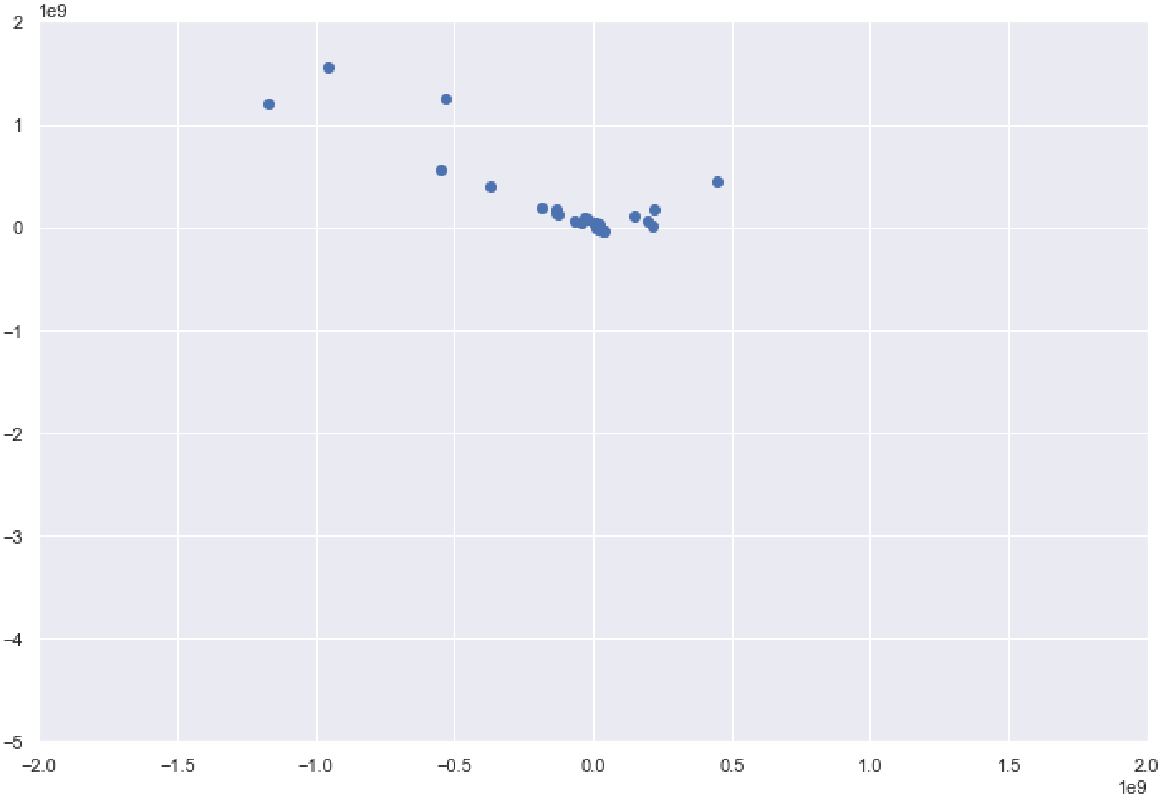
The first step is to set up the data converting three columns ‘Yearly Change’, ‘Urban Pop %’ and ‘World Share’. These columns were strings that will be converted into floats that will properly represent the data. I just used the apply function to the columns passing in a function which handled some cases for these columns. The next step is to handle the columns ‘country’, ‘region’ and ‘sub-region’. To handle these columns we used a label encoder to transform those strings into integer value data. The reason we want to do the first two steps is for we can have the ability to use the columns if we choose to. The final step before splitting the data into train and test. I would like to normalize the data except for the column that we are predicting which would be ‘total_vaccinations_x’.

12/10/2021
Alex Liang
Edgar Pineda
Xing Yang

Evaluation- Model.ipynb

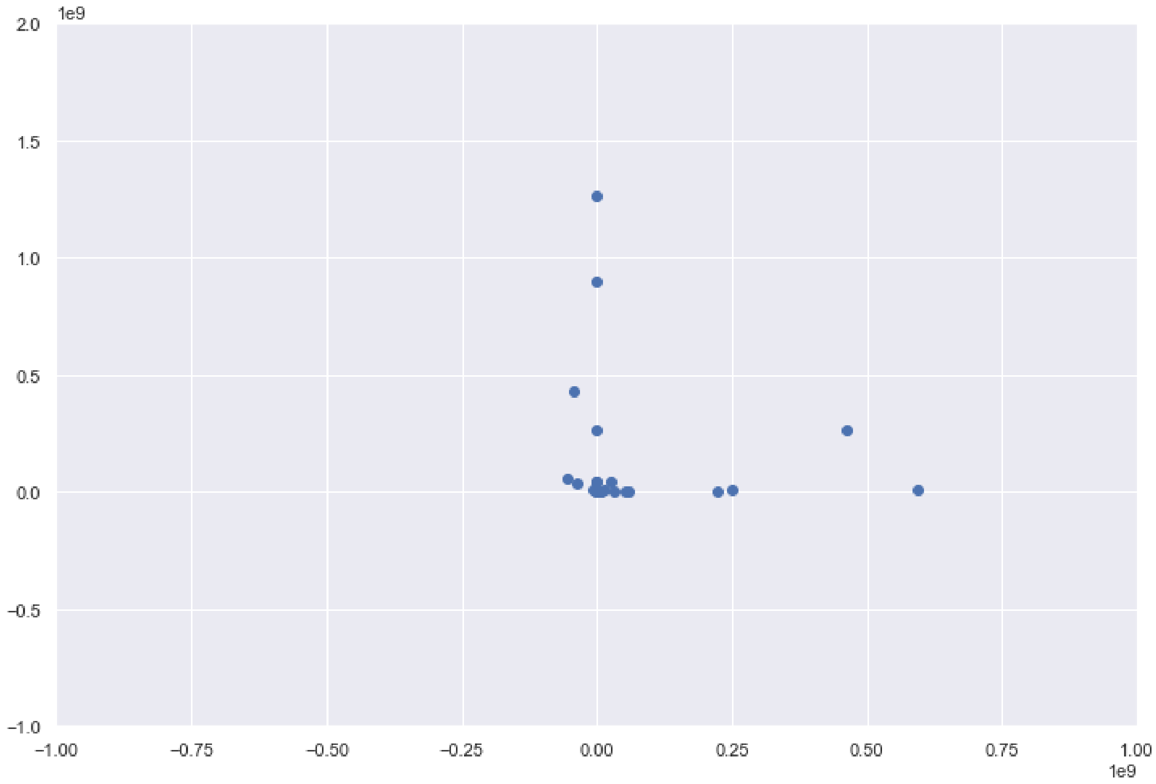
For both models we are using the following metrics to measure the performance of the model: residual plot, R2 score, and mean squared error. The residual plot shows us how the model fits the data. R2 score measures of how close the data are to the fitted regression line. Mean squared error averages the squared difference between the predicted and actual value.

Residual Plot Linear Regression



As seen above the residual plot for the linear regression model seems to have a random pattern which indicates that the model provides a decent fit to the data. The r2 and mean squared error for the linear regression are 0.6199456194864913 and 3.819808283106056e+18. The R2 score on the testing set can have a better score if we tune the hyperparameters of the model.

Residual Plot Decision Tree Regressor



As seen above in the graph, the decision tree regressor also seems to have a random pattern in the residual plot. The r2 and mean squared error for the decision tree regressor are 0.9386528681214432 and 6.165809276501198e+17.

Results for both r2 and mean squared error are better on the decision tree model regressor. The reason we are getting huge numbers for mean squared error is because there are thousands of people vaccinated and the prediction will not always be near that number. One way to reduce the results from mean squared error would be if we change the prediction from the number of vaccinated people in a country after 53 days to percent vaccinated of that country after 53 days.

12/10/2021
Alex Liang
Edgar Pineda
Xing Yang

Future Work

Currently, our data only contains information vaccination details between a 53 day range. So the next steps would be to gather more information on people that have vaccination between different day ranges. The reason we would want more data with more day ranges is to predict the amount of vaccinated people after a certain amount of days. Another would be instead of predicting the number of vaccinated people of a country after 53 days. We would like to change it to predict the percentage population of that country vaccinated after 53 days.If we have more time, we will also like to get data from the number of cases and number of death by each country and try to compare and predict when vaccinations will be fully vax. With these data, we can also analyze how many percentage of people will be vaccinated, which will lead to a significant reduction in the number of new cases.