Introduction

The COVID-19 outbreak has drastically transformed the way people go about their everyday lives. From lockdowns to social distancing regulations, the pandemic has shifted our perceptions regarding human interaction, work, meals, and sense of time. Recently, the introduction and distribution of vaccines has begun to give people a glimpse of normalcy. While the demand for vaccines is high and distribution continues to increase, the virus still remains a prevalent force that threatens the well-being of society. As a result, the ongoing pandemic has engendered our interest in the availability and accessibility of COVID-19 vaccines throughout the world. We explored a dataset containing COVID-19 data from across the world that was collected by an organization called Our World in Data. The dataset contains information collected daily such as the total number of COVID-19 tests, cases, deaths, and vaccinations available for 193 countries. Within the large dataset, we extracted 3 smaller files that would potentially allow us to explore vaccination progress around the world. In order to gain insight regarding COVID-19 vaccinations, we aimed to investigate three questions:

- What characteristics of a country (GDP, life expectancy, human development index, median age, etc.) best predict the vaccine distribution rate?
- How quickly are different vaccines (ex. Moderna, Pfizer, etc.) being used? Is there a relationship between the type of vaccine used in a country and the location/characteristics of the country?
- For which contact rates and recovery rates is vaccination the most effective or critical?

We first began to answer these questions through data visualizations we created. To continue our analysis of each question we used data analysis methods which included linear regression, logistic regression, and simulation.

Question 1:

What characteristics of a country (GDP, life expectancy, human development index, median age, etc.) best predict the vaccine distribution rate?

Visualizations:

To get a sense of which characteristics of a country may potentially affect vaccine distribution rates, we created multiple plots in Tableau that compare the vaccination rate of a country to a certain characteristic. Figure 1 shows vaccination rate being compared to median age for each country that had available data. Countries with larger dots have a higher vaccination rate compared to other countries, while the various shades of blue are indicative of a country's median age. Countries that are a lighter shade of blue have a smaller median age, while countries that are a darker blue have a larger median age. Using Figure 1 we are able to observe that South Korea's vaccination rate

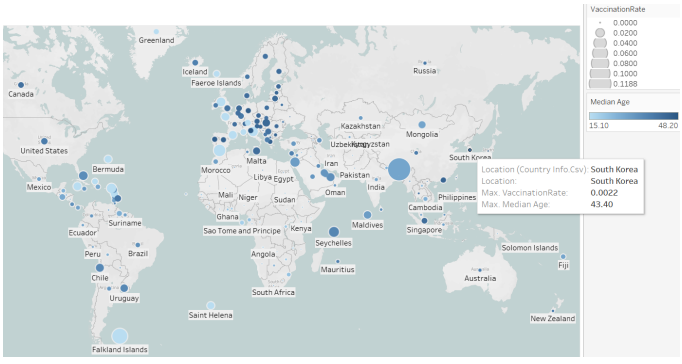

Figure 1: Tableau visualization of the world displaying a countries' median age and their vaccination rate

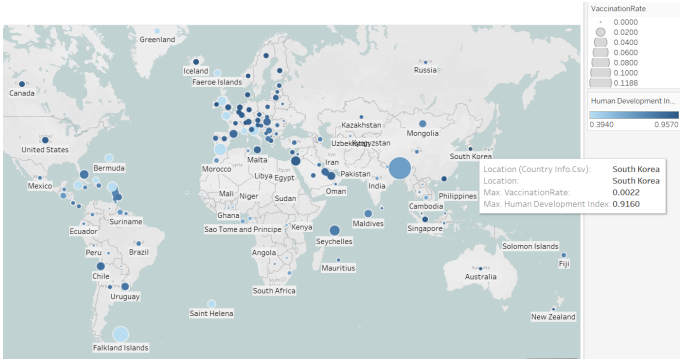

Figure 2: Tableau visualization of the world displaying a countries' HDI and their vaccination rate

.0022% is small, but the median age 43.40 leans toward the larger side. Looking at the size and color of Kenya's dot, Kenya's vaccination rate is also small, but the median age is on the smaller side. Therefore, it may be the case that median age is not a good predictor of vaccination rate. Figure 2 shows the vaccination rate being compared to HDI (Human Development Index). HDI is a measure of three fundamental components of a country's human development: life expectancy, education, and GNI (Gross National Income) per capita. This visualization is similar to Figure 1, with the only difference being that the shades of blue depict a country's HDI rather than median age. In

order to further delve into determining which characteristics may be a good indicator of vaccination rate, we decided to use linear regression.

Data Analysis: Linear Regression

Process: In order to determine which characteristics of a country best predict a country's total vaccination rate, we decided to use linear regression. We created a model by selecting eleven characteristics provided by our dataset and using them as predictors for 66 different countries. The predictors we included in our model are listed as follows: population, population density, median age, population 65 years old or greater, GDP (Gross Domestic Product) per capita, extreme poverty, cardiovascular death rate, diabetes prevalence, number of hospital beds per thousand, life expectancy, and HDI. We chose to include predictors such as GDP per capita and HDI which we think have potential to be an indicator of vaccination rate. However, we also chose to include the predictor diabetes prevalence which we think has no causal relationship to vaccination rate in order to later demonstrate what we can do with an insignificant predictor.

OLS Regression Results

Dep. Variable:	total_vaccinations_per_hundred	**R-squared:**	0.366
Model:	OLS	**Adj. R-squared:**	0.237
Method:	Least Squares	**F-statistic:**	2.832
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	0.00541
Time:	01:31:30	**Log-Likelihood:**	-291.73
No. Observations:	66	**AIC:**	607.5
Df Residuals:	54	**BIC:**	633.7
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9926	100.709	0.010	0.992	-200.918	202.903
population	-3.89e-10	2.94e-09	-0.132	0.895	-6.29e-09	5.51e-09
population_density	0.0267	0.012	2.141	0.037	0.002	0.052
median_age	-0.8407	1.378	-0.610	0.544	-3.604	1.922
aged_65_older	1.3547	1.711	0.792	0.432	-2.076	4.786
gdp_per_capita	2.292e-05	0.000	0.073	0.942	-0.001	0.001
extreme_poverty	0.0864	0.778	0.111	0.912	-1.474	1.646
cardiovasc_death_rate	-0.0111	0.038	-0.292	0.771	-0.088	0.065
diabetes_prevalence	0.7377	1.305	0.565	0.574	-1.879	3.355
hospital_beds_per_thousand	-2.4380	1.869	-1.304	0.198	-6.186	1.310
life_expectancy	-1.6616	1.445	-1.150	0.255	-4.559	1.235
human_development_index	201.8462	99.168	2.035	0.047	3.026	400.666

Omnibus:	61.933	**Durbin-Watson:**	1.914
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	342.233
Skew:	2.785	**Prob(JB):**	4.84e-75
Kurtosis:	12.665	**Cond. No.**	4.00e+10

Result: Observing the results of our fitted model summary, we focus on the column titled P > |t| which gives the p-value for each predictor. We see that population density and HDI are significant at the 95% level. Other predictors like GDP per capita and extreme poverty are likely less significant because their p-values are close to 1. This implies that a small change in population density or HDI presumably has an impact on the total vaccination rate of a country if the assumptions associated with linear regression hold.

Checking Assumptions of Linear Regression:
The results of our linear regression are only valid if the following four assumptions are satisfied:

1. Independent of the covariates: We check if the residuals are independent of the covariates by plotting the fitted values from our regression versus those for the actual vaccination rate for each of the predictors. Observing the plots for each covariate, the average value of the residuals in a narrow window over small ranges does not seem to display a nonlinear pattern, though some of the plots are a bit ambiguous. Six out of the eleven covariate plots are displayed in Figure 3. Overall, there appears to be no apparent relationship between the residuals and each predictor. Thus, all of the covariates seem to be independent of each other and our first assumption is satisfied.

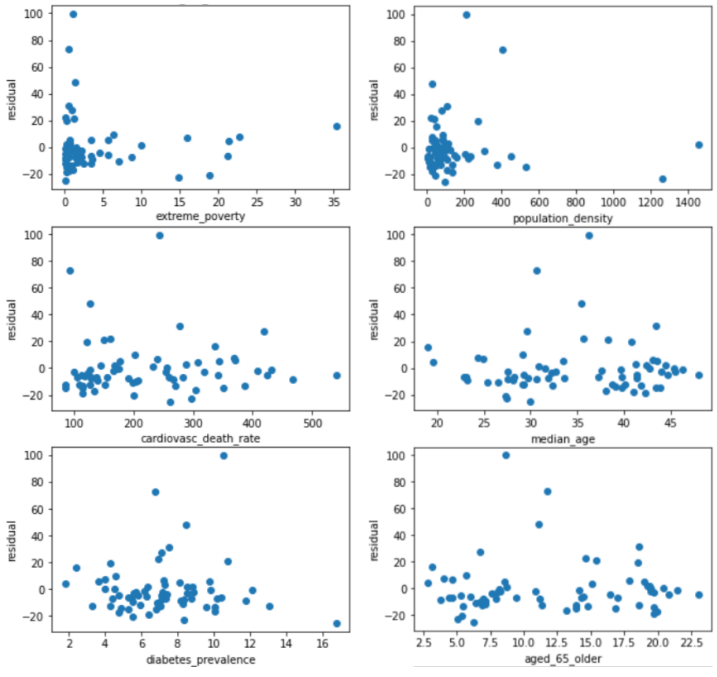

Figure 3: residual plots of 6 of the covariates

2. Mutually independent: We check whether the residuals are independent of each other by creating an autocorrelation plot. Observing Figure 4, we see the ticks at each lag are generally very close to 0 which indicates that the residuals are mutually independent. Thus, this assumption is not violated in our model.

3. Normally distributed: We use a qq-plot to check if the residuals are normally distributed for each covariate. Normally distributed residuals will form a line, where the slope of the line is dependent on the variance of the residual. In Figure 5 we observe that the qq-plot we generated seems to form a relatively straight line, indicating that the residuals are indeed normally distributed.

4. Constant variance: We plotted the absolute value of the residuals versus each covariate to check if there are any problems with non-constant variance. For each predictor, we observed whether or not the distribution was dependent on the predictor. Since it does not seem like the distribution depends on any of the covariates, the variance is constant. Figure 6 displays six out of the eleven plots we analyzed.

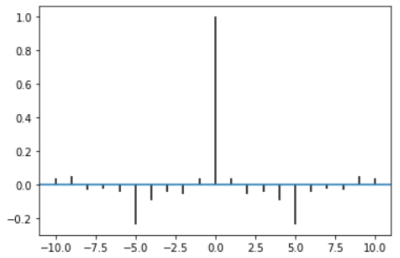

Figure 4: autocorrelation plot

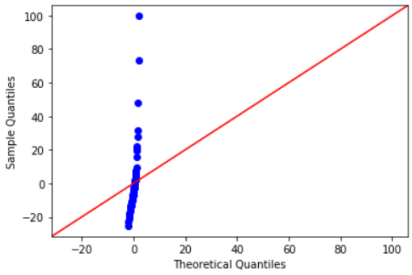

Figure 5: qq-plot

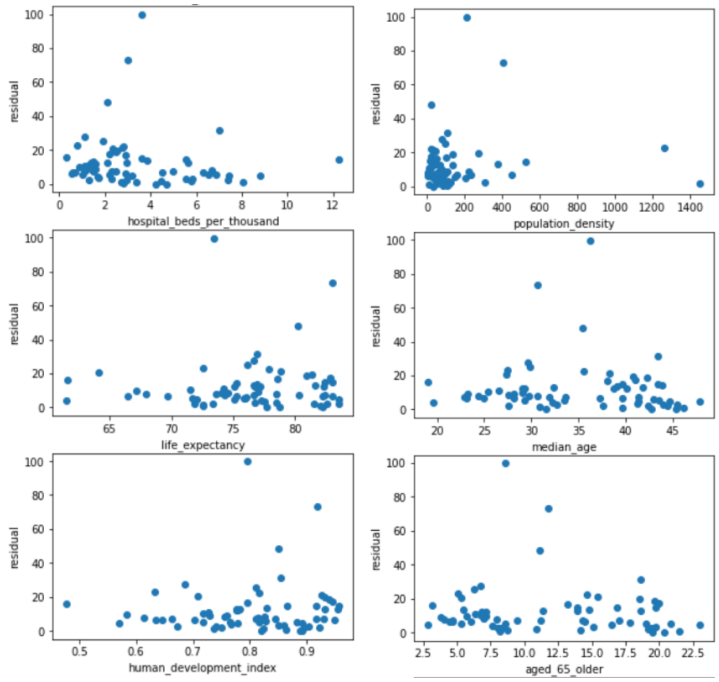

Figure 6: residuals versus covariates

Model Selection: We wanted to see if we could find a new model using fewer covariates because a simpler model with comparable predictive power is preferred. First, we split our data into training and testing sets so that we could test our models accurately. We used the AIC (Akaike Information Criterion) and p-values to help us select which covariates to include in our model. The AIC is a number computed from a model and a set of data that suggests how "good" the model is at explaining that data. The score is a combination of the sum of squared errors and the number of parameters in the model. A lower AIC indicates a better model. The p-value is a number that is used to test whether vaccination rate depends on a covariate. We iteratively removed covariates with big p-values because these are the covariates that have no influence on the model while checking the AIC of the new model.

The predictors that the AIC selected were population density, population 65 years old or greater, number of hospital beds per thousand, life expectancy, and HDI. The mean squared error for this model was 1167.9801. When we trained a model on the training set that used all of the predictors, the mean squared error was 1173.1325. We can conclude that we prefer the simpler model because it has fewer predictors and a lower mean squared error.

Conclusion: Using the new AIC model we created, we rechecked all the linear regression assumptions we had performed on our original model in a similar manner. We found that the residuals were independent of the covariates, the residuals were mutually independent and normally distributed, and there were no problems with non-constant variance. Therefore, we identified that the 5 predictors in the AIC model are the characteristics that best indicate a country's COVID-19 vaccination rate. Furthermore, the p-value of HDI in our new model is 0.003 which is significant at the 95% level. The relationship between HDI and vaccination rate could stem from the fact that countries with a higher HDI may have more available access and resources to the production and distribution of the COVID-19 vaccines. Our finding that HDI is significant is interesting because life expectancy and GDP (which is related to GNI) were individual predictors we had included in our original model, yet their p-values were not found to be significant. Thus, it could be the case that there is an underlying factor such as the education component within HDI or the way that HDI is calculated that is causing the relationship between HDI and vaccination rates to exist.

Question 2:

How quickly are different vaccines (ex. Moderna, Pfizer, etc.) being used? Is there a relationship between the type of vaccine used in a country and the location/characteristics of the country?

Visualization 1: To get an initial idea of what vaccines are being used and how popular they are, we plotted the number of countries that use each vaccine. From this plot, we saw that the two most popular vaccines are Oxford/AstraZeneca and Pfizer/BioNTech. This was able to tell us that these two vaccines were being distributed to the most countries, but it didn't tell us how the vaccines were being used in these countries.

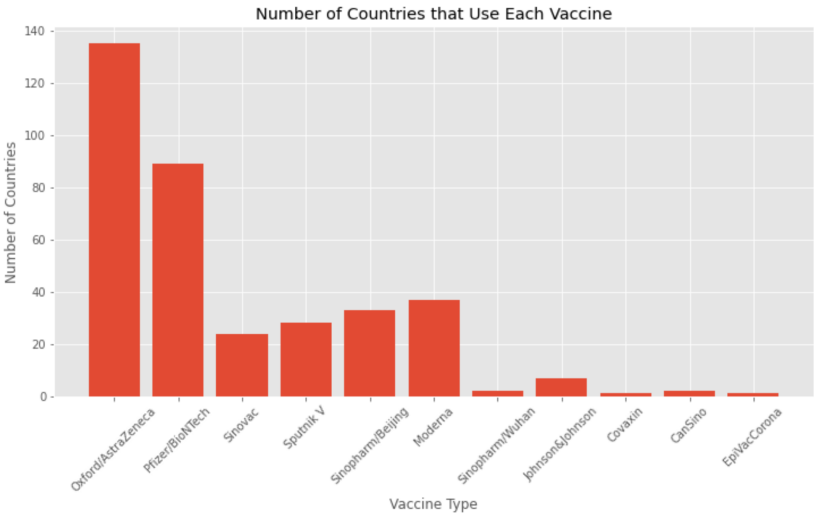

Figure 7: Number of countries that distribute the 11 types of COVID-19 vaccines

Visualization 2: We combined the table about which countries use which vaccines with the table about the vaccination progress by country. Then, for each date, we found the average of the number of people vaccinated per hundred in each country across all of the countries that use each type of vaccine. Since we didn't have information about the proportion of vaccine use for countries that use more than one type of vaccine, we made an assumption that these countries use each of their corresponding vaccines in an equal amount, and divided the number of people vaccinated per hundred in those countries to evenly contribute to each of the vaccines that each of those countries uses.

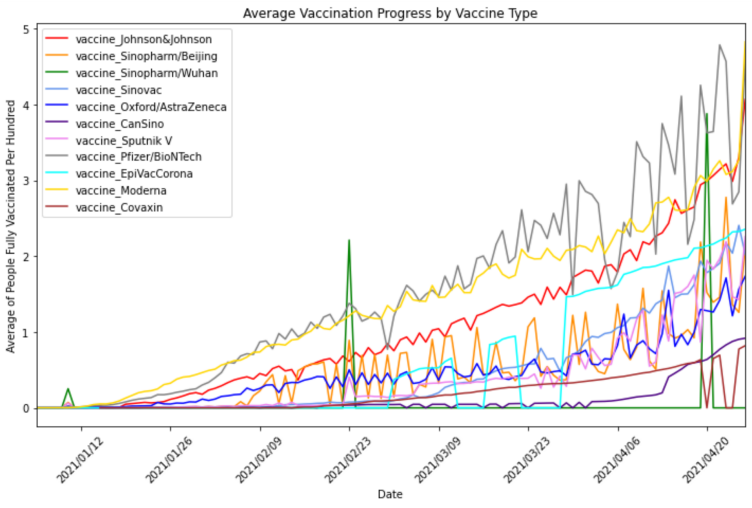

Figure 8: Average number of people receiving each different vaccine over time

This visualization tells us that despite the Oxford/AstraZeneca vaccine (dark blue line) being used by the most countries, not as many people have been becoming fully vaccinated with this vaccine. Meanwhile, the Pfizer, Moderna, and Johnson & Johnson vaccines have consistently been the top 3 vaccines in terms of getting high percentages of people in their respective countries fully vaccinated.

Visualization 3:

To gain a better understanding of the relationships between using each type of vaccine type, we observed the correlation between demographic and socioeconomic factors of the countries included in our data, and the types of vaccines that are used in those countries. We did this visually by creating 11 indicator columns for each possible vaccine type in "country_info.csv", where a 1 means that the country uses a vaccine and a 0 means it does not. We also used a Python library called geopy to add two new columns corresponding to the latitude and longitude of each country. We then created a correlation matrix between all of the columns, and plotted a heatmap that shows the correlation between the columns that were not indicators and the columns that are.

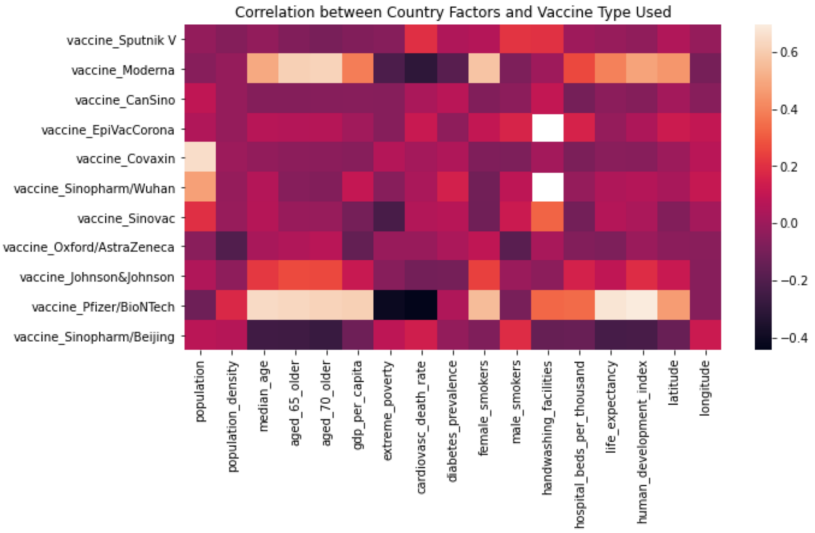

Figure 9: Heatmap that demonstrates the correlation between several chosen factors for each different vaccines

From this heatmap, we can visually see that there is a high positive correlation (at least 0.6) between a country using the Pfizer vaccine and factors like the median age, number of people aged 65 and older and 70 and older, life expectancy, and human development index. These characteristics also have a relatively high positive correlation with a country's use of the Moderna vaccine.

Data Analysis: Logistic Regression

Process: We wanted to see if any of the factors that describe a country have a statistically significant relationship with whether a country uses a certain vaccine. We did this by using Logistic Regression, creating eleven different models for each of the vaccines. We used stepwise variable selection to find models that minimize the AIC.

Result: For each vaccine type, the statistically significant factors and their effects on the use of that vaccine were:

- Sputnik V: life expectancy (negative)
- Sinopharm/Beijing: male smokers (positive)
- Sinopharm/Wuhan: no results; only used in China and UAE
- Sinovac: no statistically significant factors after model selection, the closest is extreme_poverty (negative)
- Oxford/AstraZeneca: aged_70_older (positive), gdp_per_capita (negative), diabetes prevalence (positive)
- CanSino: no results; only used in Mexico and Pakistan
- Pfizer: aged_65_older (positive), life expectancy (positive), human development index (positive), latitude (positive)
- EpiVacCorona: no results; only used in Russia
- Moderna: life expectancy (positive), latitude (positive)
- Covaxin: no results; only used in India
- Johnson & Johnson: no results; only used in seven countries

For some of the vaccines, we were unable to fit a logistic regression model because these vaccines are used by a very small number of countries, which made it easier for some combination of predictors to perfectly separate the dependent variable corresponding to each of these vaccines.

Conclusion: From our logistic regression analysis, we see that life expectancy and country latitude are both statistically significant factors that have a positive effect on whether the Pfizer vaccine or the Moderna vaccine is being used in a country. Other statistically significant factors with a positive effect on whether a country uses the Pfizer vaccine are the human development index and the percent of the population that is 65 or older. Our observations here support our observations from our correlation heatmap, because these factors also have a relatively high positive correlation with the Pfizer and Moderna vaccines. The factor of latitude having a positive effect on both of these vaccines suggests that geographically, countries in the northern hemisphere are more likely to be using the Pfizer and Moderna vaccines. When the other factors are high, this tends to indicate that the country is more wealthy and has a higher standard of living.

From our plot displaying the average proportion of people getting vaccinated by country for each of the vaccines, we can see that Pfizer and Moderna consistently have the highest average vaccination rates. This suggests that countries that are wealthier, more advanced in technology and infrastructure, and have a higher standard of living overall are more likely to get more Pfizer and Moderna shots in peoples' arms. These findings also connect to our previous question, because we saw that human development index has a statistically significant effect on vaccination rate, which also corresponds to a higher standard of living.

The vaccines with much lower average country vaccination rates also seem to have statistically significant factors relating to a poorer standard of living. For example, life expectancy has a negative effect on predicting if a country uses the Sputnik V vaccine, and GDP per capita has a negative effect on predicting if a country uses the AstraZeneca vaccine. This is also evident in the countries that we weren't able to find significant factors for, just by looking at the vaccination progress of vaccines unique to a 1 or 2 countries. For example, India, a country that has been in the news for having increasingly worse conditions, only uses Covaxin, so the line corresponding to Covaxin in the line graph directly corresponds to India's vaccination rate over time, and it is very low, with the average vaccination rate not even rising above 1 in every 100 individuals. The clear divide between vaccination rates and vaccine types for wealthy, more developed countries and poorer, less-developed countries raises concerns that we are seeing today that COVID-19 may be transforming into a "third-world problem" as developed countries escape the pandemic.

Question 3:

For which contact rates and recovery rates is vaccination the most effective or critical?

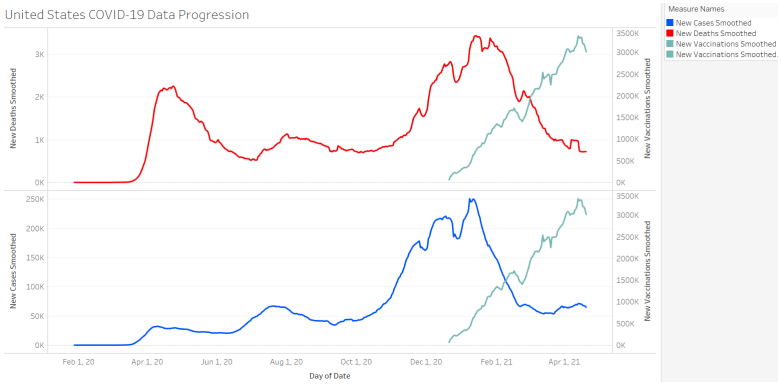

Figure 10: Two graphs showing the plotting of new vaccinations in the United States versus new deaths and new cases

Visualization: In order to gain a proper understanding of the trend between the number of new COVID-19 cases, new deaths, and new vaccinations during the pandemic, we created two line graphs depicting these exact trends in the United States. Looking at Figure 10, we see that once vaccinations started to occur in the United States, the number of cases and deaths subsequently decreased. We also observe that the vaccination rate appears to be increasing linearly. Note that the scaling of the left and right axis for both line graphs

are not the same because the number of new vaccinations is substantially larger than the number of new cases and new deaths. If we were to make the y-axis ranges equal, new deaths and new cases would be a horizontal straight line. However, the difference in scaling does not affect the quality of the visualization, as we are mainly observing the general trend of the data in the United States so that we are able to create an accurate simulation.

Data Analysis: Simulation with SIR Modeling

Process: For our SIR model, we created a step function where the number of new infections and new recoveries comes from a binomial distribution. Our binomial distribution was binom(n = S(t), p = beta * i(t)) for new infections and binom(n = I(t), p = gamma) for new recoveries, where S(t) is the number of susceptible people, I(t) is the number of infectious people, and i(t) is the proportion of infectious people. We selected a subset of beta values [0.1, 0.15, 0.2, 0.25], and a set of gamma values [0.04, 0.06, 0.08, 0.1], and simulated the pandemic progress across every combination of values for 300 days, starting with 0.01% of the population being infectious. This allowed us to visualize a baseline of what a pandemic would look like with our simplified model if there were no vaccinations in those 300 days.

Then, in order to incorporate vaccination into our SIR model, we fit a quadratic model to estimate the number of people fully vaccinated in the United States as a function of the date (right). We decided to use the United States to model a general vaccination rate because we have had a solid vaccination program. We then added the number of new vaccinations per day into the "recovered" portion of the SIR state, to represent that being fully vaccinated takes someone out of the susceptible pool. Then, for each pair of beta and gamma, we tried to find the latest possible day (since our start date, not necessarily since the beginning of the pandemic) such that if a vaccination program started that day and followed the model that we fitted, the peak of the proportion of infectious people would be 80% of the peak if there were no vaccinations.

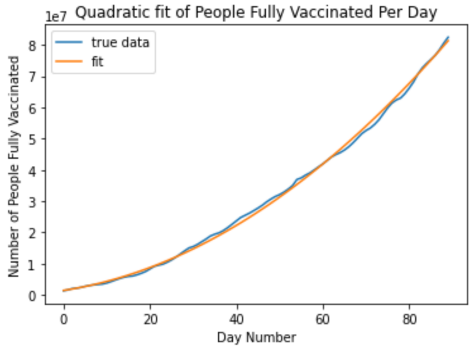

Figure 11: our quadratic fit of people vaccinated compared to the true data

Result: Of the 16 combinations of beta and gamma values that we tested, below are the ten pairs where vaccination would need to start as soon as possible in order to reduce the peak of infections by 20% (we filtered out the pairs where even starting vaccinations at the first state would not reduce the peak of infections by 20%).

```
Latest Starting Days to Reduce Infection Peak by 20%
beta = 0.25, gamma = 0.08: original peak = 0.3257265289020483, latest starting day: 0
beta = 0.2, gamma = 0.06: original peak = 0.3475996802669074, latest starting day: 9
beta = 0.25, gamma = 0.1: original peak = 0.24095074763767677, latest starting day: 9
beta = 0.2, gamma = 0.08: original peak = 0.23959630529691806, latest starting day: 22
beta = 0.15, gamma = 0.04: original peak = 0.38811926257707524, latest starting day: 24
beta = 0.2, gamma = 0.1: original peak = 0.15719007608026772, latest starting day: 37
beta = 0.15, gamma = 0.06: original peak = 0.2378986904739164, latest starting day: 44
beta = 0.15, gamma = 0.08: original peak = 0.13376071359544148, latest starting day: 69
beta = 0.1, gamma = 0.04: original peak = 0.23646303125998866, latest starting day: 87
beta = 0.15, gamma = 0.1: original peak = 0.06397745594070732, latest starting day: 105
```

From the list of values above, we can see that while the gamma values, which represent recovery rates, are more varied, the beta values, which represent infection rates, are consistently high. This becomes more clear when we look at the ten pairs of values that have the earliest days needed to reduce the peak of infections by 10%. Below, all but 2 pairs have a beta value of 0.2 or above, while the gamma values still vary.

```
Latest Starting Days to Reduce Infection Peak by 10%
beta = 0.25, gamma = 0.04: original peak = 0.5616628626679835, latest starting day: 4
beta = 0.25, gamma = 0.06: original peak = 0.43066239970189724, latest starting day: 12
beta = 0.2, gamma = 0.04: original peak = 0.48938720500534244, latest starting day: 19
beta = 0.25, gamma = 0.08: original peak = 0.3256988705129599, latest starting day: 19
beta = 0.25, gamma = 0.1: original peak = 0.24111429315760122, latest starting day: 27
beta = 0.2, gamma = 0.06: original peak = 0.34750115235392665, latest starting day: 29
beta = 0.2, gamma = 0.08: original peak = 0.23961831720548146, latest starting day: 40
beta = 0.15, gamma = 0.04: original peak = 0.38819183299371024, latest starting day: 45
beta = 0.2, gamma = 0.1: original peak = 0.15712808866057204, latest starting day: 54
beta = 0.15, gamma = 0.06: original peak = 0.2379631431306953, latest starting day: 62
```

Figure 12 shows a comparison between pandemic progress when there are no vaccinations and when there are vaccinations starting at the 9th state. We see that even though infections hit a peak at around the same time, the proportion of people who are infectious at this peak is less as vaccinations begin earlier.

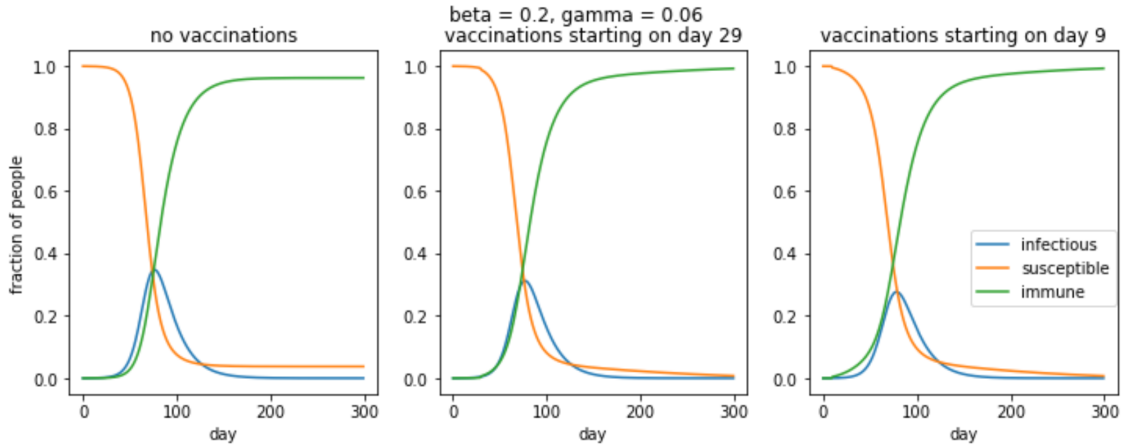

Figure 12: three different models showing the COVID-19 progression with different vaccination starting dates

Conclusion:

We observed above that higher beta values lead to vaccinations needing to start earlier in order to lower the peak of infections. This makes sense when we observed the effects that both beta and gamma have on infection progress (see our Appendix), because while the beta value contributed to the number of infectious people, the gamma value contributed to the number of recovered people; changing beta means that the model needs more "help," in the form of incorporating vaccinations, to lower the peak of infections. Since beta represents the average number of contacts a person has in some period of time, we conclude that the more contacts there are between individuals, the more critical vaccines are to keeping infections low.

Final Reflection

We gathered data and were able to see how different vaccination information could be used to answer some interesting questions. We used various forms of data analysis to provide these insights. We had a large number of data points and some information was missing due to certain countries' lack of resources but we were still able to gather conclusions. Some of the conclusions we were able to draw are that countries with more wealth, a higher standard of living, and more resources are more likely to have high vaccination rates, and also tend to use the same vaccines. We were also able to take a mathematical approach to simulate different scenarios of diseases to see how vaccination could play a role in these scenarios. It has been more than a year since COVID-19 started and we are beginning to hopefully see an end to the pandemic with the increased number of vaccine distributions.

Bibliography

Dataset:

Our World In Data

- https://github.com/owid/covid-19-data/tree/master/public/data
- https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/vaccinations.csv
- https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/locations.csv

Additional Resources:

- United Nations Development Programme Human Development Reports
  - http://hdr.undp.org/en/content/human-development-index-hdi
- Geopy library
  - https://pypi.org/project/geopy/