Effects of COVID-19 Vaccinations in the US - Project Proposal

The COVID-19 pandemic hit the US in March of 2020, disrupting the way of life as we know it. For months, scientists and pharmaceutical companies struggled to find preventative measures and cures for the virus as it affected millions of people around the country. Finally, late last year, a few vaccines - namely, Pfizer, Moderna, and Johnson & Johnson - were FDA approved for mass distribution with highly favorable efficacy. The country began the rollout of these vaccines, and after more than a year, the world is finally returning back to normal. But how much of this effect can be attributed to vaccinations alone? What is the correlation between the number of vaccine distributions and the number of new cases, and can the latter be predicted using the data?

I would like to forecast the number of new COVID-19 cases within 10% of the actual value, focusing on the number of vaccinations as input. This information would be extremely valuable to the pharmaceutical companies that developed the drugs, and policymakers / health officials who distribute doses and provide recommendations on a national level. The model, when carried out long term, will be able to predict how long it will take the US to return to “normal” life.

There are several limitations to consider in this problem. Firstly, the number of new cases is affected by a variety of different inputs, including lockdown policies, mask wearing, number of people who have already gotten the virus, etc. By looking at only the number of vaccinations, we may not be seeing the full picture. Also, the data is limited by how recent vaccination rollout began. We don’t have years of data to work with and we don’t know how complete the data is, so variability may be higher. It is also important to note that the effects of the vaccine are delayed by varying amounts of time depending on the type of vaccine received, and that we need to take into consideration the moderate efficacy of one dose of the two-dose vaccines. But nonetheless, if the model is able to predict new cases within a 10% margin of error, we can consider this model to be successful.

The datasets I will be using are from Kaggle - the first tracks the daily number of new cases by state in the US. [link here] The second dataset tracks the number of daily vaccinations in the US. [link here]

The way I would approach this problem would be to plot the daily number of new cases per state over the last 8 months and compare it to the daily number of fully vaccinated people per state. I would adjust for time delay by analyzing the correlation between the two plots at

various time delays and finding the most correlated. I would also compare the cumulative data for both cases and vaccination numbers to see if there are additional trends.