Reflection Statement

Collaboration in Data Science projects is usually helpful, especially for solving complex, fuzzy problems. For this assignment, it was very important to get and give input for people blocked at any point in the analysis. Different perspectives for approaching the problem, coming up with valid assumptions, design considerations for the solution, and adopting various tools/techniques were really helpful in coming up with the final visualization.

Starting with data-related concerns, I filtered the data for this visualization for Maricopa County in Arizona. One of the primary concerns was there were no masking mandates issued by the State during the given time period. Hence, none of the counties in the state have that data. My first thought was to try extrapolating the data from the “Voluntary masking” data which is given for a 2-week period. Now, this data cannot be representative of the entire time period. Additionally, it is an estimate from a small sample size. One finding was, people “always” wearing masks comprise more than 73.4% and the ones wearing masks more often than not comprise 95% of the population. This masking survey data collected has a problem. The estimation is taken on the basis of the 200 nearest responses. I doubt the generalization of masking on the basis of a few data points. Also, they weigh the nearest and farthest points, where the weights are not mentioned, if it’s an arbitrary way of doing it, all the approximation and estimation could lead to incorrect calculations and misinterpretation. Some of my classmates faced the same issue, hence we discussed the advantages and disadvantages of having such a dataset. Hence, modeling the voluntary masking data was the primary concern. I then decided to take the national policy route which was validated by my classmates. Therefore, collaboration helped. To adopt that, there were certain assumptions made on that with data collected between July 2 to July 14, 2022, where the CDC had strong recommendations for policies for face coverings. Essentially, meaning if 95% of people were wearing masks in Maricopa more than “Sometimes” during the 2-week they inherently followed the CDC’s national guidelines in absence of State mandated masks. With this simplified assumption backed up by data, I was able to model voluntary masking compared to state mandates.

As another question speaks about modeling for population compliance using voluntary masking data. One way of modeling based on probabilities could be where we take the weighted sum of the entire population with respect to mask compliance. For example, Never can be assigned weight 0, Rarely-0.25, Sometimes - 0.5 until Always - 1.0. This weighted sum would help us model population compliance. The underlying assumption is that the data is representative of the entire population and can be extrapolated for more than 2 weeks as if it was collected without any additional changes. This was just a discussion and not included as a part of the current analysis.

One important thing to note here, is we do not consider hospitalizations, herd immunity, recovery, vaccinations, or other implicit factors in our modeling process. For example, one interesting finding is that after vaccines (Late December 2020) were available, the CDC changed to a closed-space masking policy but still saw an increase in cases reason being people traveling across states during Christmas and New Year, which is another factor that was left out and could have biased our analysis in some way. In our case, we handle the open-air and closed-air with CDC guidelines dates. We basically disintegrate the problem and see the impact of different masking policies on the cases. The first month of the pandemic is a cold start as the cases and infection rates could take time, to get the medical equipment in place. As the research question speaks about the impact starting February 1, 2020, we automatically leave the time series for the first 15 days. We see one month gap between masking and the derivative function of infection rates near first 4 change points. For modeling vaccinations, one way could be to take into account susceptible population and their percentage of vaccination and non-susceptible ones, then take a sum of them. The resulting value can be subtracted from the denominator of the percentage of the population infected to understand the derivative function of change infection rates. A similar idea was suggested by Tharun Kumar Reddy Karasani.

On the other hand, my lack of experience with time series tools helped me understand a lot of change point detection options. We used the Pelt Search method and Prophet tool suggestions by Eli Copron and Charles Reinertson respectively. I adapted some of the business logic shared by Eli for plotting the graphs. Discussing some of the terminologies-related time series was also helpful in analyzing the data to an extent. One part where this collaboration hindered my thinking was there could be easier ways to understand the changes in time series without going to complex change-point detection methods. When it comes to the explainability of your data science solutions, it is always better to have simpler methods.

In summary, collaboration and brainstorming with people helped me quite a bit to come up with new ideas, validate my assumptions and get help with tools and techniques as it helped me progress on the roadblocks in the assignment. Some places where it hindered progress, were when many ideas were on the table and it was tough to weigh the pros and cons of each to come up with the best deterministic solution.