A5: Extension Plan

By Venkata Sai Muktevi

1. Motivation

The original analysis question was to show the significance of introducing mask mandates during peak COVID times to curb the spread of the virus in Harris County, Texas. This was where being restricted to only the 3 datasets made it difficult to find any significant relation between mask mandates and infection rates. The question is interesting in itself but we may need to dissect the particulars of what exactly the mask mandates had an influence on by utilizing other sources of data and statistical analysis. Many reasons could have contributed to the rise and fall of COVID infection rates so it would be more helpful to measure each potentially contributing factor against the other to actually find out how much mask mandates helped on relative grounds. As an extension we may look into how other policies introduced during the COVID pandemic affected aspects of our lives or if these policies actually helped in reducing the rates of infection observed through some correlation in the data.

Upon researching various aspects and policy changes during the pandemic I've decided on the following approach to delve into understanding certain factors that affected COVID infection rates or were influenced by COVID themselves in certain ways. I will be looking into more data on infection rates because initially I had trouble coming up with a good definition of "at risk" population as well as in defining infection rate. This will support my approach to take a second deep dive into mask mandate data but with a more human centered outlook and a more broadened perspective. Instead of looking at the mask mandate period I think it's important to gauge compliance intent of COVID protocols in general based on survey data of people in and around Harris County. This would give us a much more in-depth understanding of people's decisions in actually following protocols to also see to what degree they were followed. The research scope is now more human centred with the data points representing various aspects of people's intent from various demographic backgrounds. For example, front-line workers may yield different data compared to IT professionals.

The next aspect I'd like to consider is regarding mobility data provided by Google and Apple. I believe the lockdowns and restricted travel played a big part in reducing transmissions. I would like to explore mobility in Harris County alongside COVID infection rates to look for any relationships between them.

2. Research Questions

There are multiple research questions I would like to explore in this study that I will briefly discuss here which are mainly two-pronged.

I. **Survey Data -** *Is there more to the mask mandate than its policy execution period? How many seem like they will comply more with the mask mandate, can we put this side-by-side with the infection rates time series to see any correlation?* Data will consist of COVID-19-related public opinion, demographic, and symptom prevalence. Upon observing the results of A4, a deep dive into survey data on mask mandate would support the human centered aspect of understanding how people responded to these types of policy changes in and around Harris County.

II. **Mobility Data** - We tried to see initially if there was a direct relation between mask mandates and infection rates but are there other factors here that can be looked into. I chose to look into mobility which is interesting because it gives us an idea of travel and movement to and from the county. This could link to transmission coming from outside the county or vice versa. The available Google and Apple mobility data is quite comprehensive. *Is there a correlation between covid cases over time and mobility? Can we see spikes in COVID cases when visualized alongside large mobility changes? Did travel restrictions from city to city help reduce covid alongside observing mask mandates?*

3. Data Used

I plan to mostly use the C3.ai COVID-19 API source for analysis of their Survey data and Mobility data. This dataset is open source and integrates data from 40 different disparate sources, then models and presents those data in a unified, cohesive structure to be accessible by API. The C3 AI COVID-19 Data Lake is available at no cost to the global research community. The data is accessible via any utility that supports a RESTful interface including commonly used tools such as Python, R, and Microsoft Power BI. They mention in their terms of use about giving them credit in case of any publications or research results that are derived in full or in part from the C3.ai COVID-19 Data Lake. We have to credit the C3.ai COVID-19 Data Lake by referencing the case study at https://c3.ai/customers/covid-19-data-lake/.

I chose this data source because it's very well maintained and has exactly what I needed on COVID surveys and mobility. It's also easy to access through their python API using notebooks. The two datasets I will be utilizing are the Survey data and the Mobility data. The survey data stores COVID-19-related public opinion, demographic, and symptom prevalence data collected from COVID-19 survey responses. There are more than 1000 records for Harris County and I plan to use more from neighboring counties to gauge the perspectives of the overall population in the region. The mobility data from Apple and Google provide a view of the impact of COVID-19 and social distancing on mobility trends.

Mobility data is a time series of various forms such as Walking, Driving, Parks, Residential, Transit, etc. The measure is with respect to deviation from a Baseline metric.

Survey data can often be biased depending on collection methods and there could be an imbalance in the demographics of people surveyed. I'm not sure how reliable the information could be in terms of accurately capturing the larger population of Houston's views on COVID. The survey data also holds political views of people which can be harmful if misrepresented. The mobility data respects privacy and anonymity but it is still an ethical dilemma on Big Tech conglomerates collecting private information like location.

More details on the data are available here in the API Documentation.

4. Unknowns and Dependencies

I plan to prioritize my approach to work with the Survey data then Mobility data. I planned for multiple research questions just in case some do not provide satisfactory results or if the data just turns out to be "bad" such as having less variance to work with or to find any meaningful trends. Some factors that may be influencing the data in certain ways may be out of my hands and my initial hypotheses or research questions might not be fruitful approaches from the very beginning. The mobility data may be influenced by only those who use Google and Apple products and therefore have a specific bias associated in this manner. The survey data results are at risk of being misinterpreted due to the politically sensitive features and must be dealt with carefully.

I plan on an iterative approach to course correct along the way in case I have an issue with the data. Since I will be focussing on one county there often seems to be less data for such a specific region or within a longer time frame. This will affect outcomes as well but I will try my best to perform a significant analysis with available data.

5. Methodology

Data will be gathered using the API available and processed accordingly. Survey data will have a lot of categorical data that will need to be factorized by labels and processed. I will also need to filter out the time frame under consideration for the Mobility time series data. The API is well documented and allows for filtering based on requirements.

Survey Data Analysis - The survey data will be gathered using the aforementioned API by filtering for Harris County and Bexar county (San Antonio) for comparison. The survey data consists of various categorical and ordinal information on what each person would deem their level of intent to follow certain COVID protocols like social distancing, masks, washing hands. They also record their detailed demographic information on various aspects like religion, political inclination, etc. I plan to use this information and compare these fields with the Participant's response to the question: On a scale from 0 to 10, "how concerned

are you about the coronavirus?" (0: Not at All, 5: Somewhat, 10: Extremely Concerned). I want to see if the proportions of responses to each of these fields is related with the COVID Concern filed using statistical analysis. The Chi-square test is a good way to measure these proportions to see if there is any relationship between various fields and COVID Concern. This could help confirm the hypothesis that more concerned people would generally follow all advised protocols. This could also help uncover any latent biases to observing protocols if there are any relations regarding demographics such as certain political inclinations. The end goal is to develop a list of statistically significant parameters that may help determine what demographic characteristics often dictate COVID concern levels from direct survey data.

Mobility Data Analysis - This data will also be gathered using the API mentioned. We have data on Walking, Driving, Parks, Residential, Transit, etc. mobility provided by Google and Apple. Using this data we can plot alongside it our time series on COVID cases from A4. I will aim to find any relationships between this mobility data and COVID cases to see if people moving in/out of the county are affecting the infection rates. It will be interesting to observe the correlation using the ARIMA model because these data are time series.

I will be presenting my findings in the following ways. A table visualizing the important features related to COVID Concern based on demographics. A time series visualization comparing different types of mobility and infection rates. A presentation compiling salient findings.

6. Timeline

The timeline to completion spans roughly ***3.5 weeks***. Within this timeframe I will split the work in the following manner each week:

1. Survey Data Analysis - Process survey data and perform Chi-square tests, look into any other relevant statistical analysis, document findings and approaches.
2. Mobility Data Analysis - Process the data, perform ARIMA model and find relations between mobility data and infection rates, explore other models given time, document findings and approaches.
3. Consolidation of FIndings - Document findings into a final report and presentation, compile visualizations to tell the story of the analysis, list failures to comply with this extension plan and reasoning, taking feedback and reiterating.
4. Finishing touches on the final report.

Thank You