Client Based Projects: All data should have been collected. All project questions should have been reviewed, answered, and submitted in a written document outlining findings. You will also be asked to submit the associated data and a README explaining what each label/feature in your dataset represents. Your team should meet with the client before this deliverable.

Checklist

- ☐ All data is collected
- ☐ Refine the preliminary analysis of the data performed in PD1&2
- ☐ Answer another key question
- ☐ Attempt to answer overarching project question
- ☐ Create a draft of your final report
- ☐ Refine project scope and list of limitations with data and potential risks of achieving project goal
- ☐ Submit to Gradescope with the above report and modifications to original proposal

It is no secret that the effects of COVID-19 have been devastating. However, it isn't as well known how disproportionate the effect across different communities and demographics has been. Dr. Julia Koehler approached us to explore these discrepancies, especially in the context of vaccinations.

To do so, we made the scope of our project to be the cities of Revere, Chelsea, Newton, Wellesley, Everett, and Springfield. Revere, Chelsea, Everett and Springfield were classified as underserved communities, while Newton and Wellesley are classified as better-served communities. Dr. Koehler elaborated that there had been a grassroots organization called La Colaborativa serving the communities in Revere, dispelling rumors about vaccines and providing access to these vaccines.

The ultimate goal of this project was to be able to directly visualize the effects of La Colaborativa and discern a qualitative difference between the communities that have grassroots support (or just Revere as we don't have the knowledge of whether the other communities are supported by grassroots organizations) and those that do not.

Since this is the first emergence of SARS-COVID-19, there haven't been any specific attempts at quantifying it or its vaccine response in the past. The only other time that a similar outbreak occurred was in 2003 with a similarly described SARS outbreak, in which development into a vaccine occurred yet never made it past the first trials(https://www.bcm.edu/departments/molecular-virology-and-microbiology/emerging-infections-and-biodefense/specific-agents/sars-mers). Yet, academic literature cautioned against the vaccine (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7094954/). Due to these past actions, the emergence of COVID-19 and its vaccine response has been extremely novel, and as such its vaccine rollout can't be compared to other events.

Dr. Koehler had supplied us with two datasets. The first one was organized based on zip code while the other had been organized by county and city. In both of them, the Mass government

had been logging vaccination numbers since the beginning of COVID. The data was mostly complete, except for a couple of missing points to which we had cleaned.

During the earlier times, booster two vaccination data is missing. This makes sense as the booster was not out yet during that time period. We didn't make any changes to them, and we considered these time periods completely separate when plotting them. During the later times, we figured that some of the partially vaccinated data was missing. We cleaned that part of the data by subtracting the fully vaccinated data from at least one dose data. We also realized that for most times, the infant data is missing. We simply filled them out with 0. Lastly, the partially vaccinated data of Wellesley city from line 2396 to 2403 seems to be shifted up by one row, so we shifted them down by one row.

To ensure that the different cities had a statistical difference, we used the T test, finding that the data between cities and races is significantly different as p value is lower than 0.05 when comparing them to each other.

We decided to make line graphs, because they can compare changes over the same time period for more than one group. By splitting up the timeline into 5 parts — early, mid, late vaccine rollout, booster one rollout, booster two rollout — we made the data easier to view and work with. Early mid and late vaccine rollout time intervals were calculated by splitting the timeline from the beginning of the data to booster one rollout into thirds. Booster one rollout ends when booster two rollout begins.

[MISSING THE OTHER GRAPHS THAT DR KOEHLER WANTS US TO DO (CUMULATIVE, AND KRUSKAL)]

We graphed the following four graphs for all 5 time periods. In the first two graphs (Data *vs.* Rate of At Least 1 Dose by Region, Data *vs.* Rate of Fully Vaccinated by Region), each line corresponds to a region and each graph corresponds to a race. The purpose is to figure out the discrepancies of vaccination rate between the regions that we are interested in. In the last two graphs (Data *vs.* Rate of At Least 1 Dose by Race, Data *vs.* Rate of % Fully Vaccinated by Race), each line corresponds to a race and each graph corresponds to a region. They helped us visualize the discrepancies of vaccination rate among the different ethnicities.

Based on the data visualizations we made, we were able to answer the following questions outlined in the project document:

1. **Do grassroots organizations have a statistically significant effect on vaccination rates in hispanic populations?**
    - I think we might have to run the T test on individual points.
2. **What are the factors that affect vaccination rates in the data that we are seeing**
    - Race seems to be a factor that affects vaccination rates, and the effect varies from city to city.

3. **Does the rate of vaccination rate change as la colaborativa was able to do more work in the community?**
    - Cannot tell because of no timeline
4. **Are there changes in the vaccination rate?**
    - Yes. They will be explained in the next paragraph.

After constructing all the graphs for vaccination rates of “at least one dose” and “fully vaccinated”, we are able to compare the vaccination rates across cities and races. Based on the plots of early vaccine rollout corresponding to regions, we can see that Chelsea has the highest vaccination rate of “at least on dose” for each race, but the trend does not decrease or increase over time. For the middle vaccine rollout, the graph begins to fluctuate. For hispanic data, Revere and Chelsea have generally the highest increasing vaccination rates. Similarly, for African American data, Revere and Chelsea have the highest rate of covid vaccinations. For Causian and Asian data, the vaccination rate of Chelsea is much higher than that of Revere, Wellesley and Newton. By observing the plots of late vaccine rollout, we can see that there is a significant spike appearing between 11/01/2021 and 12/01/2021 for each race, which could be contributed by flu season. Also, we found out that the rates of death tend to be increasing in this time period, which could be the reason that people attempt to get vaccinated. For booster 1 rollout, we discovered that the trend of vaccination rate are similar across cities, but there is significant drop for Causian data at the time interval 01/2022 to 02/2022. We need to do further investigation on this. For booster 2 rollout, Chelsea city has the lowest vaccination rate at the middle of July in 2022 except for the African American population. Then, the vaccination rates of “at least one dose” become the same for all cities after the latter half of July. (9:53)

While we can discern that there is a qualitative difference between Revere and the other graphs, we quickly realized that without a timeline to La Colaborativa’s efforts, we cannot pin these differences to specific events in the timeline we observed.

Some suggestions for the future of the project is to look at other aspects of neighborhoods in which residents are in, such as analyzing the effect of public transportation and economic resources. This could be possible if provided with a more detailed dataset (such as neighborhood ridership, resident’s annual income, etc.)