Vaccine Equity Project

Leland Ling, Kristi Li, Brian Tao, Yu Yan

CS506 Data Science Tools and Applications

Professor Lance Galletti

December 5, 2022

I. Background & Motivation

It is no secret that the effects of SARS-COVID-19 have been devastating to communities around the world. 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. She had been working with a grassroots organization called La Colaborativa to better support the city of Chelsea. Together, they had done a lot serving the communities in Chelsea Massachusetts by doing outreach, specifically demystifying and providing access to vaccine rollout.

To explore the effects of La Colaborativa on Chelsea, we made our scope of our project to be Chelsea and compared it to five other cities of Massachusetts: Revere, Newton, Wellesley, Everett, and Springfield. Dr. Koehler categorized these cities into three groups: Revere, Everett and Springfield were classified as underserved communities, Chelsea as an underserved community with grassroots support, while Newton and Wellesley are classified as better-served communities.

The ultimate goal of this project was to be able to directly visualize the effects of grassroots organizations on vaccination roll out by discerning a qualitative difference between the communities that have grassroots support (Chelsea specifically in the scope of our project) and those that do not.

II. Previous Work

Since the emergence of SARS-COVID-19, there haven't been any specific attempts at quantifying this specific virus or its vaccine rollout response in the past. The only other similar outbreak was in 2003 with a similarly described SARS outbreak, in which development into a vaccine occurred yet never made it past the first trials (SARS and MERS)[1]. Interestingly enough, some academic literature around this time cautioned against the vaccine because of conflicting data on its effectiveness (Caution raised over SARS vaccine - PMC)[2]. 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.

III. Data Collection

Dr. Koehler had supplied us with two datasets, both collected by Massachusetts' Bureau of Infectious Disease and Laboratory Sciences. The first one was organized based on zip code while the other had been organized by county and city. Both of the datasets contain weekly logged vaccination statistics since the beginning of COVID. The datasets contained the number of individuals vaccinated with at least one dose of the vaccine, fully vaccinated, vaccinated with Booster one, and vaccinated with Booster two over different demographics, such as age, race, and gender identity. The data within was mostly complete, except for a couple of missing points to which we had cleaned. Both datasets, regardless of how they were organized, contained population data, number of individuals with at least one dose, fully vaccinated, boosted, and boosted twice.

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 did not 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 zeroes. 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.

IV. Data Visualization & Exploration

We quickly realized that analyzing every demographic contained within the datasets would be much too much for the timeline we were given, so we decreased the scope of our project to only race data. Furthermore, out of the races present in the datasets (AI/AN, Asian, Black, Hispanic, Multi, NH/PI, White, and Other/Unknown), we decided to only look at data for Asian, Black, Hispanic, and Caucasian, as the other race data was too small in scale – only having data for on average 150 data points. Asian data somewhat exhibits this issue as well, having a sample size of 810 individuals on average.

We normalized our data by dividing each metric (number of individuals with at least one dose, fully vaccinated and so forth) with population, getting the overall percentage. For example, the number of fully vaccinated Caucasian individuals of date 03/09/2021 living in the city of Chelsea was divided by the population of Caucasian living in Chelsea observed on the same date. This was done to be able to compare city data at the same scale, as the different cities had different populations for different demographics.

We began data visualization by plotting each city's data against time, separating the demographic data by race as shown below in Figure 1, a - d. We find that some of the city data have distributions similar to each other; Hispanic data shows that Chelsea matches Revere distribution during the initial stages of vaccination rollout, i.e. 03/2021 to 07/2021 and then surpasses the other distributions all together. Black data also seems to follow this trend as well. Chelseas' Caucasian data shows that the distribution has a much higher vaccination percentage on average than each of the other cities.

Asian population data in Chelsea had much higher values overall, even surpassing 100% fully vaccinated. While we initially thought that this was an issue on our part, the race data for smaller population races, such as the ones we deliberately decided to omit because of small sample sizes, sometimes would have populations less than the number of individuals vaccinated of that demographic. As mentioned above, all of the data of smaller sample sized populations would occasionally exhibit data points where the observed number of individuals would be more than the cumulative population of that demographic. This larger value would result in the above 100% data point. For these smaller population races, the difference between these values and the total population would be in the magnitude of 50-200 individuals. For Asian population data of Chelsea, with a total population of 810 for many of these points, the maximum difference was 200, and an extra 25% of the population.

To quantify how similar each city's vaccination data is to one another, we used Hypothesis Testing to compare corresponding race data. Our hypothesis was that the two cities' corresponding race data are similar - using the p-value generated from t-test as our metric. Our hypothesis was the data being compared are similar distributions, while the null hypothesis was that the distributions were not. The code to this can be found in the github repository's hypothesis testing file[3].

With the T test, if the resulting p value is above 0.05, we fail to reject the null hypothesis and conclude that the data distributions being compared are statistically similar. Below we show significant observations, or T test comparisons that result in p values above 0.05, and omit the ones that result in a p value under 0.05.

When comparing Chelsea and Newton, we found that the Hispanic demographic data of both cities had a p value of 0.911, meaning that they statistically are probably similar distributions. White demographic data when compared had a p value of 0.111, meaning that they probably are somewhat similar as well. Black and Hispanic data distributions of Chelsea and Revere result in p values of 0.470 and 0.535 respectively when performing the T test. Chelsea and Springfield's Black data reveal a p value of 0.463 when performing the T test on the two distributions.

We did compare the rest of the cities with each other using the T test, and those results can be found within the same github link above.

To further understand the similarities and differences between Chelsea and the other city's racial data, we explored the rates of change in percent of fully vaccinated data of each race. We calculated the rate of change of each demographic data by applying the python package Pandas' dataframe differential function to the data points. This function essentially calculates the difference between a data point and its previous neighboring data point, as we are only analyzing the rate of change over two points (Figure 2, a - d).

Plotting the data on the timeline from 03/2021 to 11/2022 made it difficult to observe these minor increases and decreases in rate of change. Therefore we split up the timeline into 5 parts — early, middle, late vaccine rollout, booster one rollout, booster two rollout — making the data easier to view and analyze. Early, middle, 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.

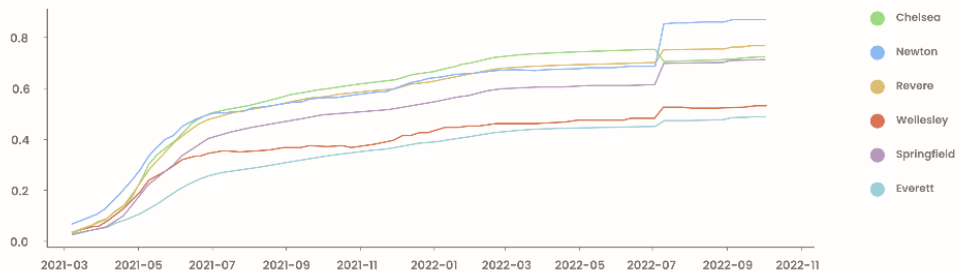
When we applied the same hypothesis testing that we did above to the rate of change, we found no significant findings. Each application of the T test as a comparison results in a p value of above 0.05, ranging from 0.21 to 0.98, meaning that all compared distributions were statistically similar. These findings can also be found in the same github link with the other hypothesis testing.

CITY DATA *VS*
% OF INDIVIDUALS FULLY VACCINATED

In the following graphs, each line corresponds to a region,
and each graph corresponds to a race.

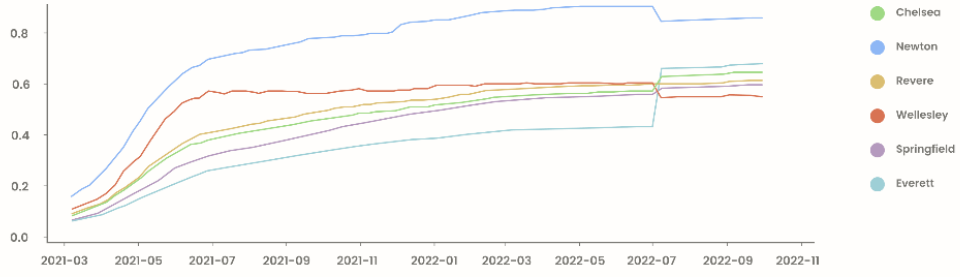
HISPANIC

Figure 1a



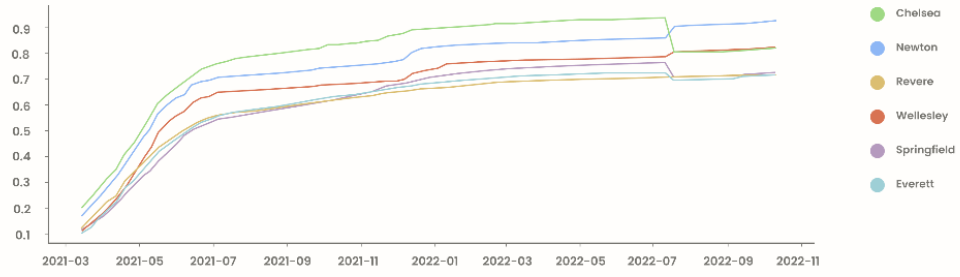
BLACK

Figure 1b



WHITE

Figure 1c



ASIAN

Figure 1d

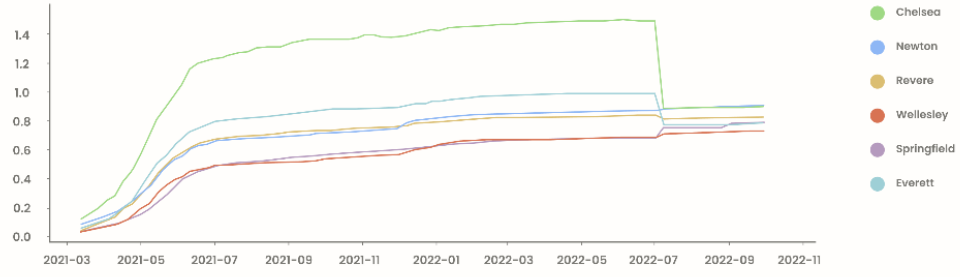


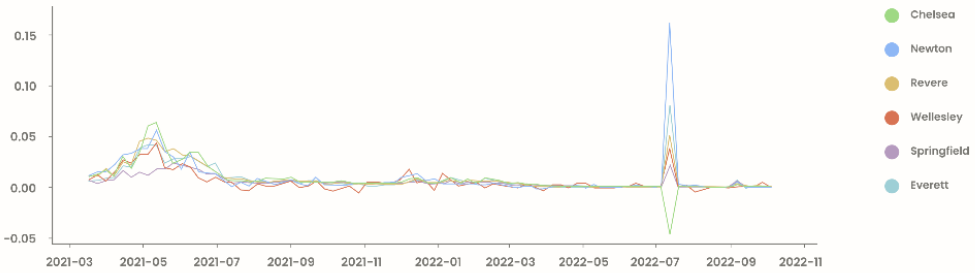
FIGURE 1

CITY DATA *VS*
RATE OF CHANGE IN PERCENTAGE OF INDIVIDUALS FULLY VACCINATED

In the following graphs, each line corresponds to a region,
and each graph corresponds to a race.

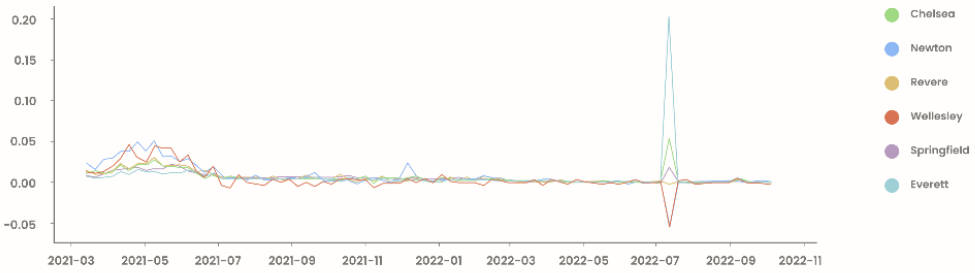
HISPANIC

Figure 2a



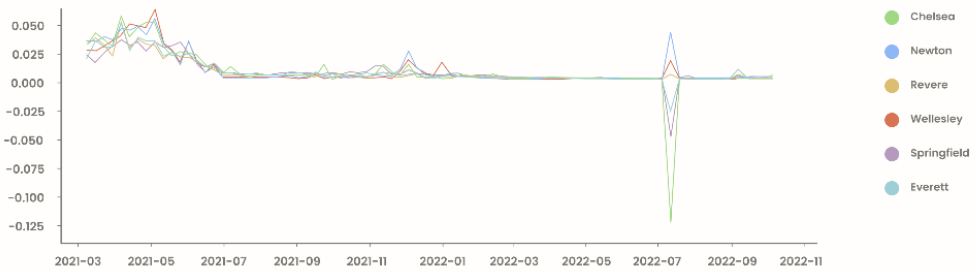
BLACK

Figure 2b



WHITE

Figure 2c



ASIAN

Figure 2d

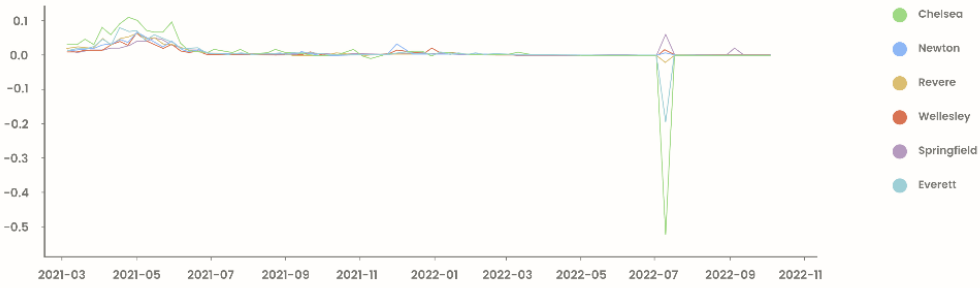


FIGURE 2

V. Results, Interpretations & Limitations

Using the results from the hypothesis testing to examine the similarity between Chelsea and its influence from La Colaborativa and the other cities of Massachusetts, we found many correlations.

The cities that we chose to analyze fall into three categories - underserved, underserved with grassroots efforts, and better served. While there could be another category of Better Served with Grassroots efforts, to our knowledge we do not know of any of these non-governmental organizations (NGO) that support better served cities that we are analyzing in terms of vaccination rollout. Regardless, an NGO serving a better served city would be included in the definition of better served. Newton and Wellesley were selected as examples of better served communities. Revere, Everett and Springfield were selected as underserved communities. Chelsea has La Colaborativa as an NGO and therefore would fall into the category of underserved with grassroots efforts.

When considering the results of the hypothesis testing of the distributions of percentage of individuals fully vaccinated, we find that only a select few of the p values are above 0.05. Therefore, we decided to use these p values as a metric to similarity. As mentioned above, the fact that Chelsea's Hispanic data when compared to Newton's results in a p value of 0.911, suggests that Chelsea's Hispanic vaccination data follows similar trends as Newton's. This correlation possibly could be the result of La Colaborativa's influence, as their work was done mostly in Spanish.

Chelsea and Revere's Hispanic and Black populations also have correlation, as indicated above. Chelsea and Revere are both underserved communities, and as such it is no surprise that they have such a statistical similarity – having p values of 0.470 and 0.535. Springfield also falls in this category and its Black population exhibits similarity with Chelsea as well, with a p value of 0.463.

While Chelsea's hispanic population data has similarities with both Newton and Revere, the p value resulting from comparing the Chelsea and Newton hispanic populations with the T test versus that of comparing Chelsea and Revere suggests that the distributions of Chelsea and Newton are more similar than the Chelsea and Revere distributions.

It is important to note that Chelsea does not exhibit similarity with Everett, another city that supposedly falls into the same category. This lack of similarity could possibly be because Chelsea's majority consisted of Hispanic identifying individuals while Everett is mostly of non Hispanic identifying individuals.

Qualitatively, there are a number of points within the rate of change graphs that could be correlated to some real world event. The graphs appear to show a gradual spike in the rate of change of percentage of individuals that are fully vaccinated across all races and cities. This trend seems to begin around early April 2021, peaking at mid May 2021, and gradually decreasing to zero after, until July 2021. This gradual spike could be explained by the gradual increase in availability of the vaccine and then a large number of individuals getting the vaccines.

However, comparing Chelsea's hispanic rate of change data to Newton, we find that the rate of change values of Chelsea are below Newton's until late April. Chelsea's values surpass Newton's after late April until mid May, where the two distributions seem to become similar. This trend is shown below in Figure 3. Revere also follows this trend to a lesser extent, but does not peak as high as Chelsea.

This parallel possibly is important to note as Revere and Chelsea fall in the same category. While this trend could be explained by vaccine availability becoming more widespread during the month of April, the slightly higher magnitude of Chelsea's hispanic data distribution could be the result of La Colaborativa.

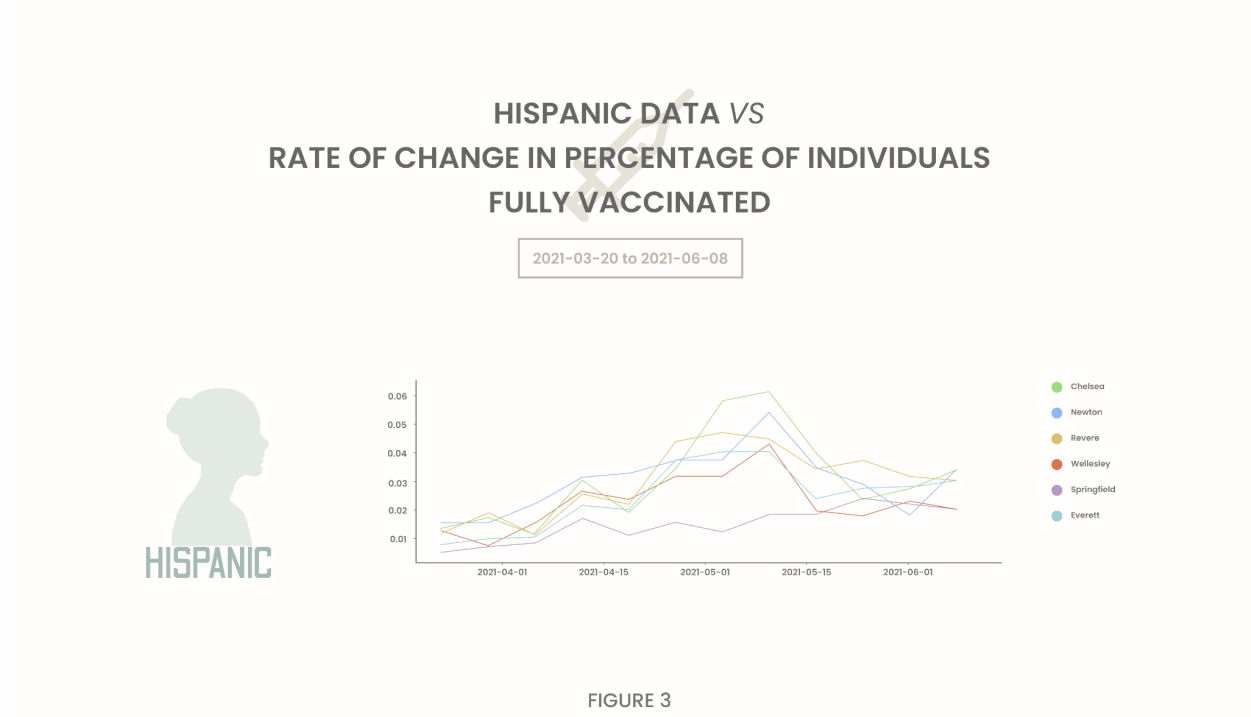


FIGURE 3

It seems before the holiday season in mid late November 2021 and after in January 2022 until gradually tapering off until mid February 2022, there also appears to be a gradual increase in the rate of change for percentage of individuals that are fully vaccinated for all demographics (shown in Figure 4, a - d). This increase could be explained by the flu season and the increase in rate of death tend to be increasing in this time period, which could be the reason that people attempt to get vaccinated. This trend is pretty consistent over all racial data.

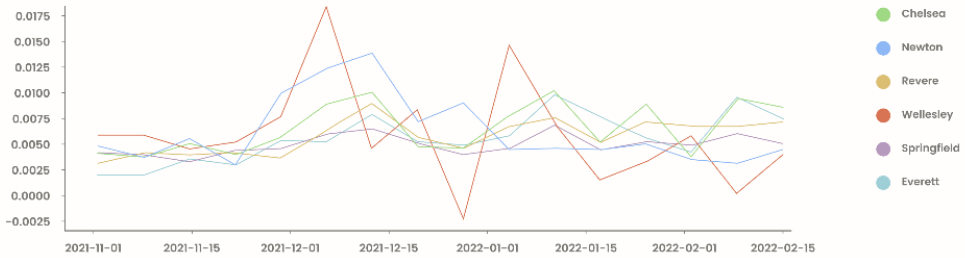
There possibly could be more dates of interest within these time plots. However, while we do know that generally La Colaborativa began their efforts around March 2021, it is difficult to extrapolate the actual effects of this NGOs work on the city as we are missing a full timeline to La Colaborativa's work. On their official website, there are snippets that try to quantify their effort, such as in their October 2021 Blog[4] post that discusses -- how through their effort, Chelsea is rank #1 out of 20 cities targeted by the Vaccine Equity Initiative. However, there is no set start date besides announcing 2-4-21 as the start of their vaccination campaign[5].

CITY DATA *VS*
RATE OF CHANGE IN PERCENTAGE OF INDIVIDUALS FULLY VACCINATED

2021-11-01 to 2022-02-15

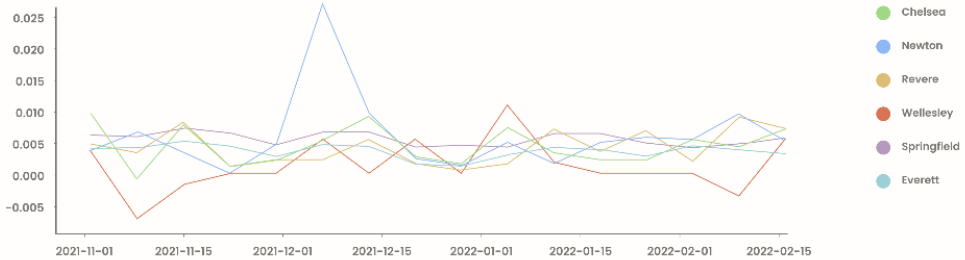
HISPANIC

Figure 4a



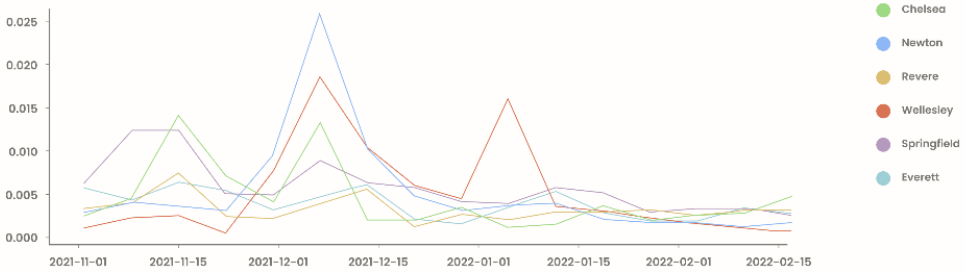
BLACK

Figure 4b



WHITE

Figure 4c



ASIAN

Figure 4d

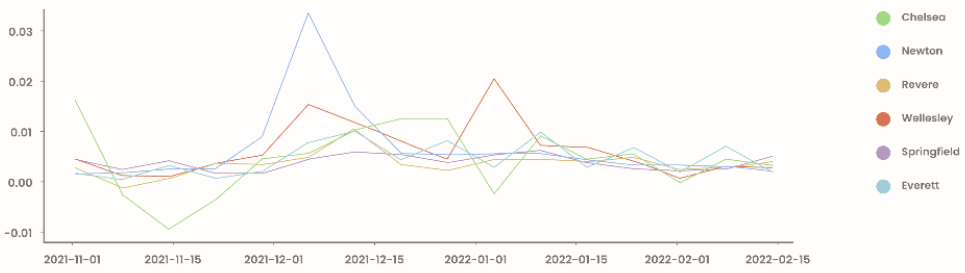


FIGURE 4

VI. Questions Answered

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 rollout?
2. What are the factors that affect vaccination rates in the data that we are seeing
3. Does the rate of vaccination change as La Colaborativa was able to do more work in the community?
4. Are there changes in the vaccination rate resulting from La Colaborativa's efforts?

As shown above in section four, Chelsea's distribution of data matches Newton's more than the other cities, possibly because of La Colaborativa's effects. Therefore it is possible that NGOs have a statistically significant effect on vaccination rollout. However, correlation is not causation and we must not err to assume that La Colaborativa is the reason that this similarity exists.

We did find that race heavily influences vaccination rates as shown in the graphs above in Figures 1 - 4. As the scope of our project was refined to be more to be comparing Chelsea with the other cities, we did not explore trends between races of other cities as deeply. We did find that Hispanic demographics between Chelsea and Newton have similar distributions when comparing the rate of change in percentage of population that is fully vaccinated.

Questions three and four could not be answered because of the same issue with question one. Without a clear timeline of La Colaborativa's efforts, we cannot directly analyze the effects of work done by this NGO. Therefore it is inconclusive as to whether La Colaborativa directly influenced Chelsea's hispanic population.

VII. Challenges Faced

Overall, we majorly had two issues. The issue with data collection mentioned above in section four with Chelsea's Asian data stands to be a minor issue with data visualization and the lack of a timeline to La Colaborativa's efforts. It does appear that the later points in Chelsea's Asian data have a higher population value, but without knowing exactly how it changed, or why it changed it is not reasonable to select the larger population to overwrite previous population values.

We do recognize that this data collection error could be an issue for all demographics – however, there is no easy way to fix this for all data points as we do not know what to use as ground truth for the population. If we were to use the observed difference in values and population as an overall collection error margin, this problem is not significant in the larger populations as this error margin would be between 2.5% to 0.67% of the population observed and therefore be negligible to visualizing the data trends of those populations. As such, we decided to continue with the data we had and visualize as is.

While we can discern that there is a qualitative and quantitative difference between Chelsea 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.

VIII. Next Steps

There are a multitude of ways this project can be continued in the future. Here we will only mention a couple. Our hypothesis testing was not performed on time intervals, and rather was done on the entire dataset. Doing the same testing on smaller intervals makes more sense as we can more meaningfully find similarity on smaller scale time periods and perhaps could result in more findings.

This project could be also further expanded upon by doing a direct analysis of La Colaborativa's timeline and its effects upon Chelsea's populations. Doing so could further elucidate possible effects of NGOs on underserved populations. Performing the same analysis we did on the zip code dataset that we also have could also reveal more about the discrepancies between differently served neighborhoods of these cities. Data organized by zip code could have more meaningful takeaways, as zip code might be a better way to group demographic data as neighborhoods within a city could have discrepancies as well.

Works Cited:

1. "SARS and MERS", *Baylor College of Medicine: Department of Molecular Virology and Microbiology*,https://www.bcm.edu/departments/molecular-virology-and-microbiology/emerging-infections-and-biodefense/specific-agents/sars-mers
2. Health, Department of Public, and Executive Office of Health and Human Services. "Covid-19 Response Reporting." *Mass.gov*, https://www.mass.gov/info-details/covid-19-response-reporting.
    a. COVID-data_Massachusetts-vaccines, https://docs.google.com/spreadsheets/d/1FaXay9eNR48sZ2rdeivH8vldMBZsuWagdbF3ub9y-tU/edit#gid=1069676523
3. Health, Department of Public. "Massachusetts Covid-19 Vaccination Data and Updates." *Mass.gov*, https://www.mass.gov/info-details/massachusetts-covid-19-vaccination-data-and-updates.
    a. COVID-data_Massachusetts-vaccines_zipcodes, https://docs.google.com/spreadsheets/d/1MbJO8Va82DHuzvxtNSwEWju8ZUhiZJgzZblrSVrD4Yk/edit#gid=1096893525
4. Pearson, Helen. "Caution raised over SARS vaccine." *Nature*, 10 Jan. 2005, doi:10.1038/news050110-3
5. BU-Spark. "DS-Vaccine-Equity/hypothesistesting_consolidated.Ipynb at Dev · Bu-Spark/DS-Vaccine-Equity." *GitHub*, https://github.com/BU-Spark/ds-vaccine-equity/blob/dev/fall22-team-1/Deliverable%204/hypothesisTesting_Consolidated.ipynb.
6. "From COVID 'Hotspot' to #1 in Vaccinations" *La Colaborativa* 7 Oct. 2021, https://la-colaborativa.org/2021/10/07/from-covid-hotspot-to-1-in-vaccinations/
7. "Opening Our Doors: Covid Vaccines Now Available at La Colaborativa." *La Colaborativa*, 4 Feb. 2021, https://la-colaborativa.org/2021/02/04/opening-our-doors-covid-vaccines-now-available-at-la-colaborativa/.