COVID-19 Vaccine Disparity Amongst Age Groups in Massachusetts

Jingyi Li, Kush Vekeria, Liyu Qu, Zhou Ye
Data Science Tools and Applications
Professor Lance Galletti
16th of December, 2022

I. Background and Objectives

The COVID-19 virus is an acute respiratory virus which has contributed to an increase in mortality rate across the world and has brought to light several disparities among various groups in their access to proper healthcare and reputable information about the virus and its vaccination. Specifically, disparities predominantly exist between individuals of certain age groups and races, with Black/Latino groups and adolescents reporting some of the highest incidence rates in the State of Massachusetts. As Covid-19 rates nationwide continue to increase, it is important to acknowledge the gravity of these factors and steps we can take to mitigate such disparities.

This project seeks to further understand the important role the age plays in vaccination trends in six Massachusetts cities: Chelsea, Revere, Springfield, Everett, Newton and Wellesley. Specifically understanding the effects of the virus on the vaccination rate of adolescent populations. Moreover, this project aims to understand the role of grassroots organizations, such as La Collaborativa, in increasing vaccination rates in underserved communities, such as Chelsea, Revere and Springfield.

II. Data Preprocessing and Challenges

Our team analyzed two datasets provided by Dr. Julia Koehler and her colleagues at Boston Children's Hospital. One of these datasets, labeled 'Grouped', grouped individuals by age group, city and vaccination status (boosted, fully-vaccinated, etc.). The other dataset, labeled 'Zipcode', grouped individuals by zip code, age group and vaccination status.

To process the 'Grouped' dataset, we split the dataset into several smaller datasets (ages 0-19, 20-29, 30-49, 50-64, 65-75 and 75+) for each of the six cities. To process the 'Zipcode' dataset, we split the dataset into smaller datasets grouped by age group for each city. One of the initial challenges we faced when processing the dataset was the lack of data. Specifically, in the 'Zipcode' dataset several columns contained little to no data. In an attempt to remedy this problem, our group employed the technique of interpolation. Using interpolation we were able to preserve the linear trends that the data displayed through predicting values for the empty spots.

Lastly, to get a better understanding of general trends in the data and remove unfairness, we chose to normalize the data. For the 'Grouped' dataset, we additionally divided the number of people in each group for each city by the total population of that age group and subtracted this rate by the previous rate in order to ensure we obtain the cumulative increase. For the 'Zipcode' dataset, we additionally took the cumulative sum for each age group to capture the overall increase in vaccination rates.

III. Methods

To extract value from our datasets, we chose to make use of three main methods. In order to select features to set as the object of our focus, we made use of random forest regression.

Through this method, we were able to map the importance of each vaccination status' contribution to overall vaccination rates.

Using the 'Grouped' dataset, we visualized the data into several graphs: One Dose (of a two-dose regimen vaccine) vs. Date and Full Vaccination vs. Date. Using the 'Zipcode' dataset, we visualized the data into several graphs for each of the underserved communities of Chelsea, Revere and Springfield. For each of these three cities, we had two graphs: One Dose (of a two-dose regimen vaccine) vs. Date and Full Vaccination vs. Date.

Lastly, to understand the contribution of grassroots organizations, such as La Colaborativa, we made use of hypothesis T-tests. We ran a hypothesis t test on the key times that La Colaborativa started their adult vaccination, pediatric vaccination, and adult booster in Chelsea. To make sure the data we use in the test is credible, we choose to use the increased rate instead of the increased numbers, so we can prevent interference from the population. Moreover, we need to make sure the data is from the same time range so that there are no problems about unfairness (if the data comes from different time ranges, the increase rate of the earlier one may be much more than the others since earlier the time, more the vaccination will be taken. The hypothesis t test will also run on every age range since we do not want to make all age groups add together and omit some possible success.

Our team first chooses to analyze the effect of La Colaborativa on pediatric vaccination. We separate the data into different age groups as 6-11 years old and 12-15 years old to better figure out whether the vaccination site has an effect or not. Since La Colaborativa starts pediatric vaccination in Jan 2022, we choose the time range from Jan 2022 to June 2022 to maximize the effect of this event. After running t-test with Chelsea and other five cities (Revere, Springfield, Everett, Newton, Wellesley)

IV. Data Visualization and Analysis

1. Feature importance

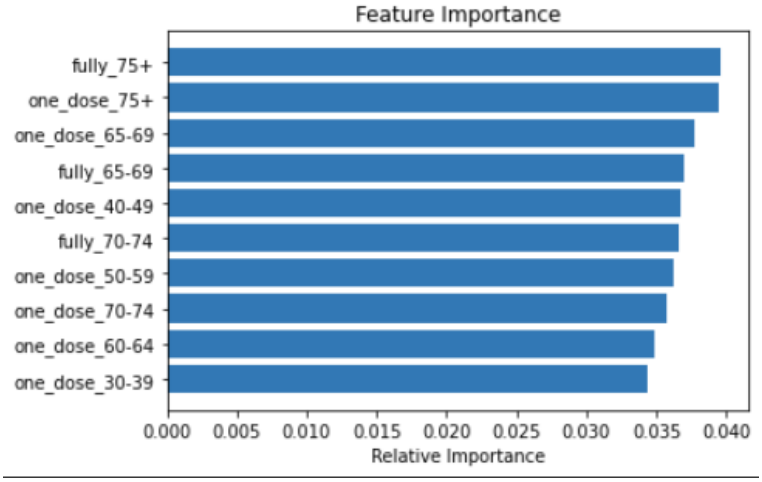

Figure 1

The above graph (Figure 1) displays the relative importance of vaccination status (one-dose of a two-dose regimen, fully vaccinated, etc.) on overall vaccination trends for each of the six cities. The graph simplifies these two the top 10 highest contributing factors. Based on these graphs, we can see that the majority of individuals who reported being vaccinated across all age groups tended to be individuals who only received one dose of a two-dose regimen. This result is of particular interest due to the fact the majority of our conclusions will be based on individuals only having one dose alone. We can also see that the majority of older age groups, 65+ years of age, tended to have their full vaccination and thus our conclusions are based on individuals with a complete set of vaccinations for this age group.

The lack of individuals in young age groups (0-19 years of age) and mid-age individuals (30-49 years of age) receiving their full set of vaccinations could be contributed to a lack of incentivization to be vaccinated in these six cities. The majority of older age groups having received their full set of vaccinations aligns with our initial predictions being that this is the age group that is most vulnerable to infection and a weaker immune system/higher mortality rate may have provided sufficient incentivisation to ensure this age group is fully vaccinated. Ultimately, these results highlight the gravity of this ongoing situation and the lack of sufficient incentivisation for younger age groups, particularly adolescents, to receive a complete set of vaccinations.

2. Fully Vaccination in Six Cities with Different Age Ranges

We graphed the following fully vaccinated rate graphs for six age ranges (age 0-19, age 20-29, age 30-49, age 50-65, age 60-75, age 75+) and each line in the graphs corresponds to a city. The first graph includes the vaccination rate in all age ranges. The purpose is to figure out the characteristics and discrepancies of vaccination rates between different cities in each age group.

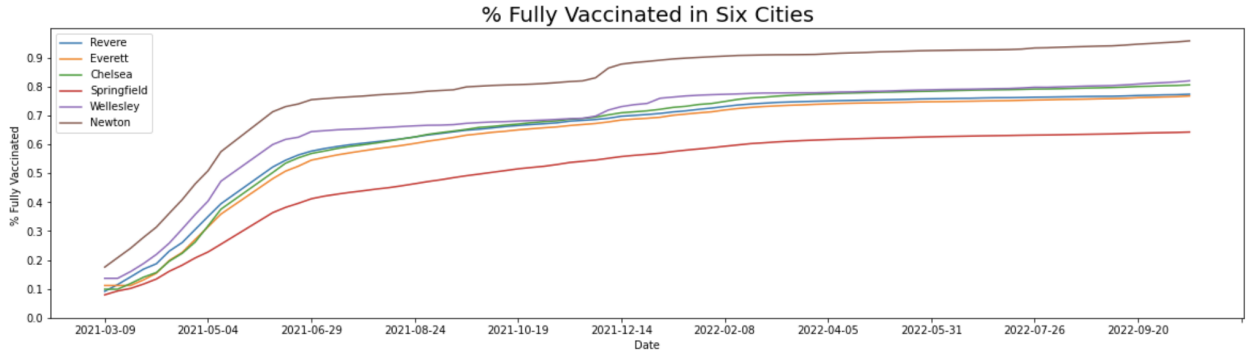

Figure 2-a

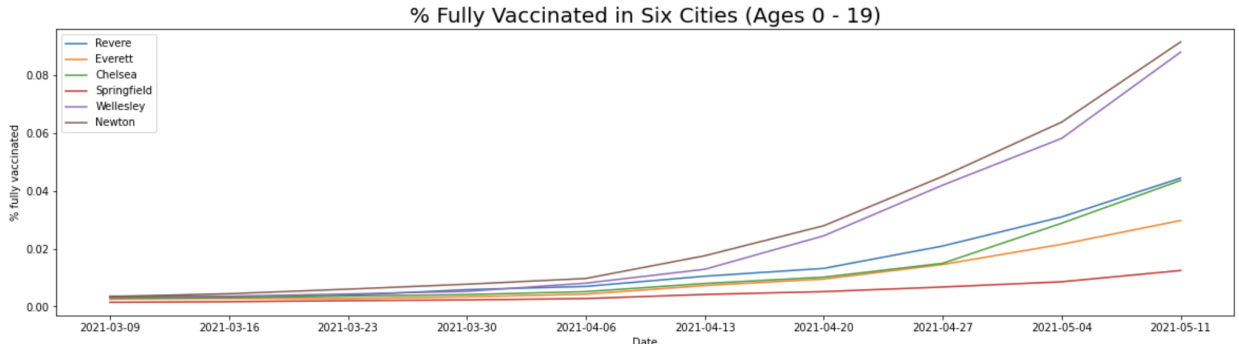

Figure 2-b

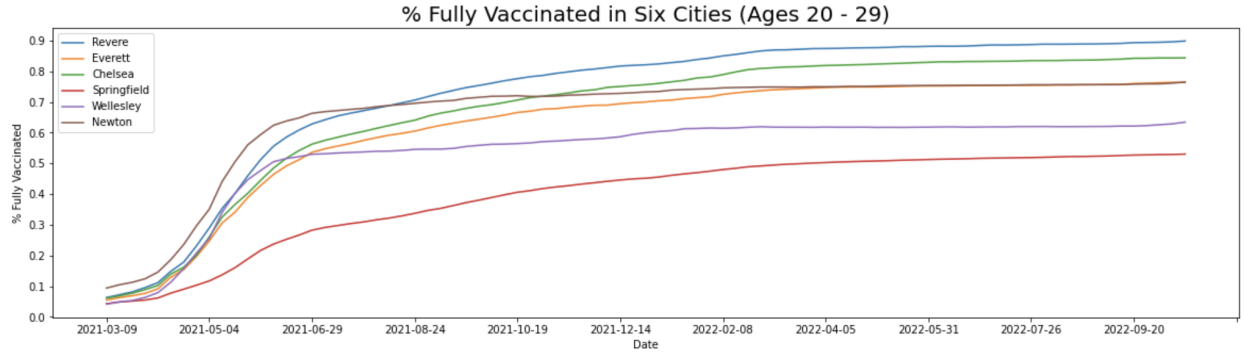

Figure 2-c

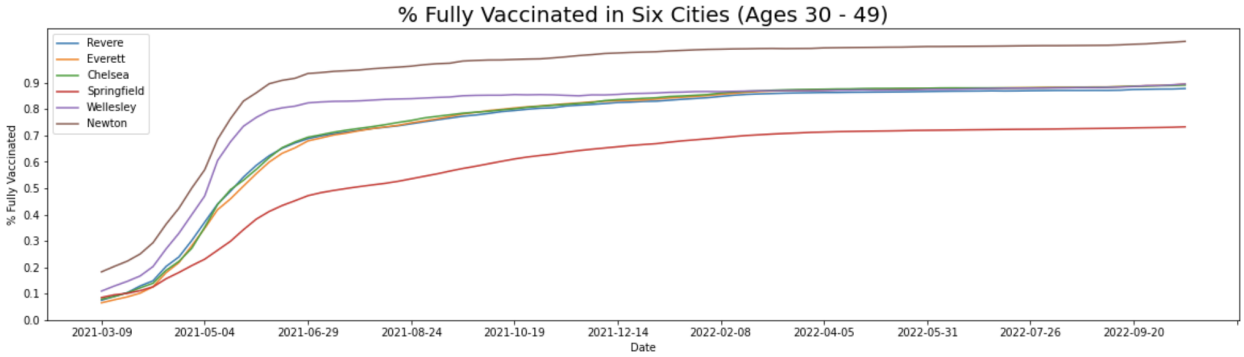

Figure 2-d

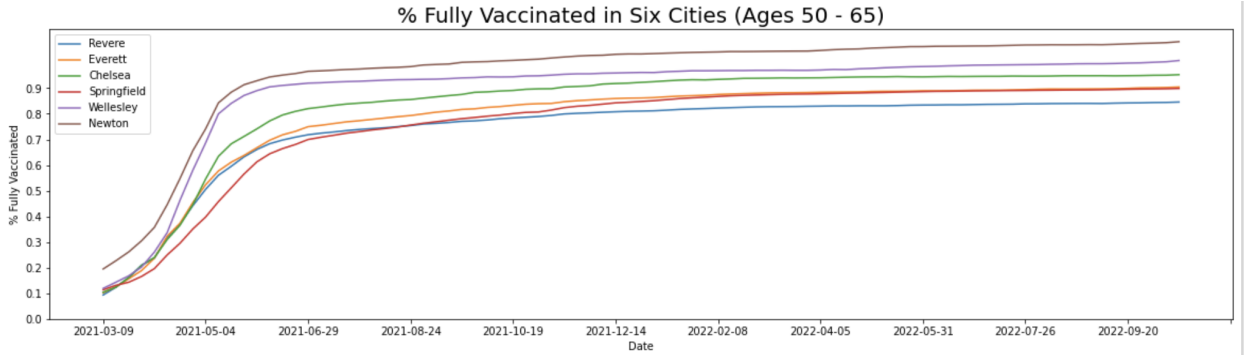

Figure 2-e

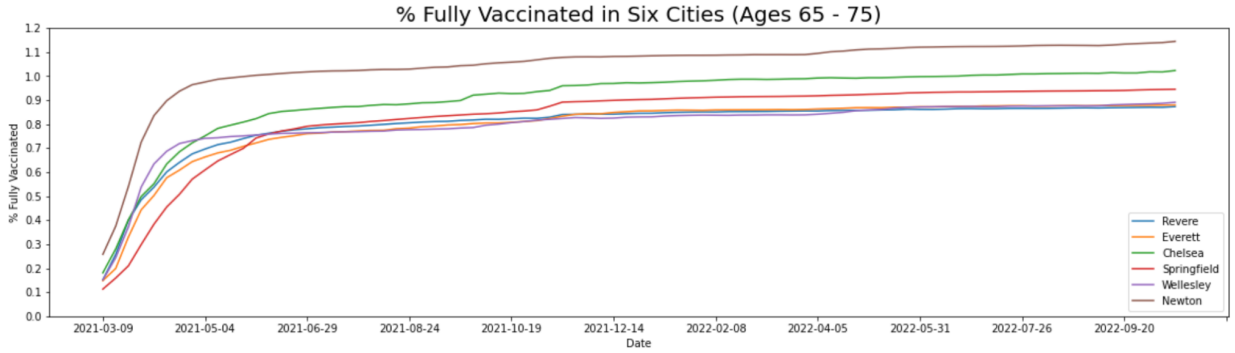

Figure 2-f

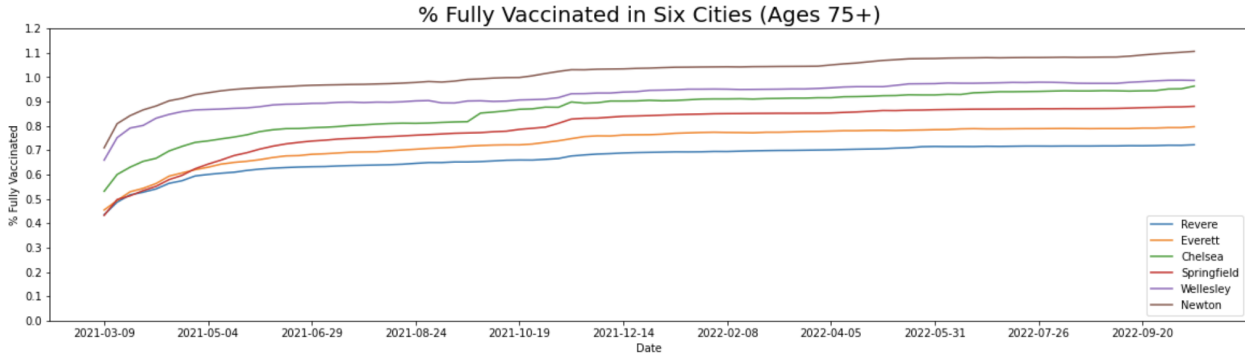

Figure 2-g

Figure 2-a includes people of all ages and it presents the fully vaccinated proportion in six cities (Revere, Chelsea, Springfield, Everett, Newton, and Wellesley). June 2021 is a turning point, the date from which the growth rate of fully vaccinated starts to slow down, especially for Revere, Everett, Chelsea, Wellesley, and Newton. Overall, Newton had the highest fully vaccinated rate and Springfield had the lowest fully vaccinated rate. Chelsea's fully vaccinated growth rate was slightly higher than Everett, Wellesley, and Revere's.

Figure 2-b shows the change in the proportion of fully vaccinated children and adolescents (aged 0-19) in six cities. Due to the lack of numerous data, the interpolation method cannot fix too large missing values. So, this plot is only from March 19, 2021 to May 11, 2021. Around April 25, 2021, Revere, Chelsea and Everett had a significant increase in the proportion of fully vaccinated people aged 0-19.

In the age group 20-29, except for Springfield, the fully vaccinated proportion in the other five cities has increased significantly from April 2021 to June 2021. After June 2021, the proportion of fully vaccinated in Chelsea, and Revere had a significant increase. They became cities with the two highest fully vaccinated rates.

In the age group 30-49, the fully vaccinated proportion has increased significantly from April 2021 to June 2021. After that, the increase rate in all cities did not have a significant

increase. Chelsea, Everett, Wellesley and Revere had the same fully vaccinated rate which was 80%.

In the age group 65-75 and 75+, the proportion of fully vaccinated in all cities has increased significantly from March 2021 to May 2021. Compared with other cities, Chelsea had a higher growth rate and a relatively high vaccination rate after April 2021.
Across the age groups, Newton had the greatest percentage of complete vaccinations among the six cities, and Springfield had the lowest percentage of complete vaccinations. However, the vaccination rate of 50-75+ older people in Springfield is the highest among the six cities. Chelsea and Revere have similar fully vaccinated rates and growth rates, but in the 65-75+ age group, Chelsea had higher rates of complete vaccination and growth rates than Revere. Based on the average vaccination rate of each age group, Chelsea had a relatively high vaccination rate in the 20-29, 65-75 and 75+ age groups.

3. At Least One Dose in Six Cities with Different Age Ranges

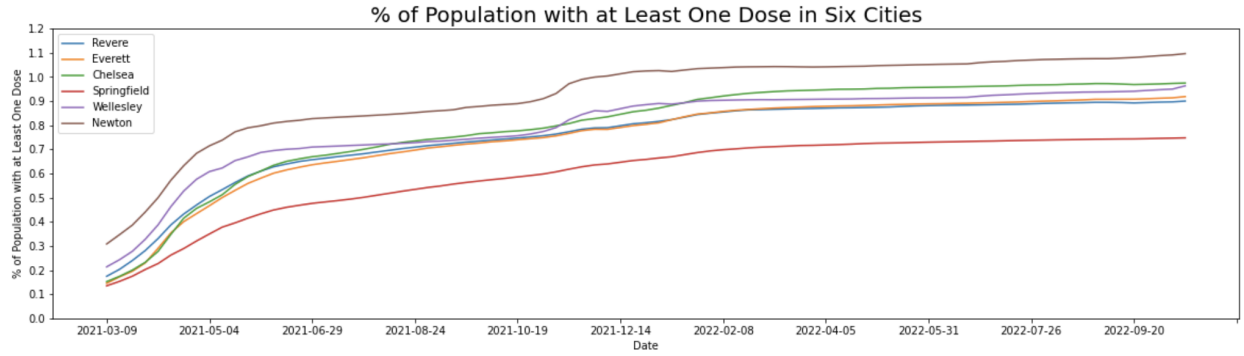

Figure 3-a

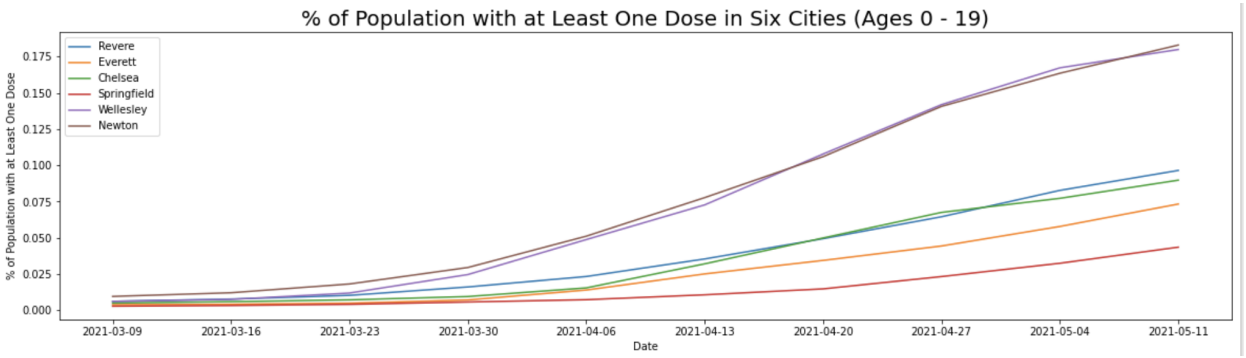

Figure 3-b

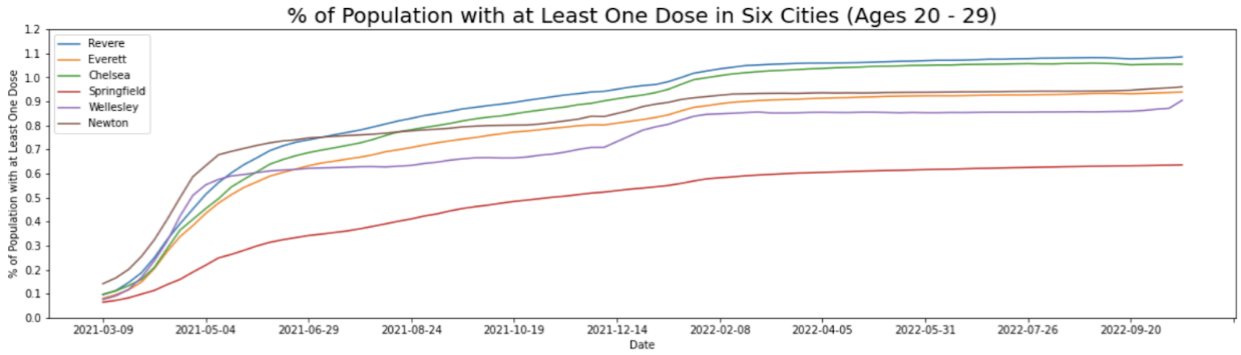

Figure 3-c

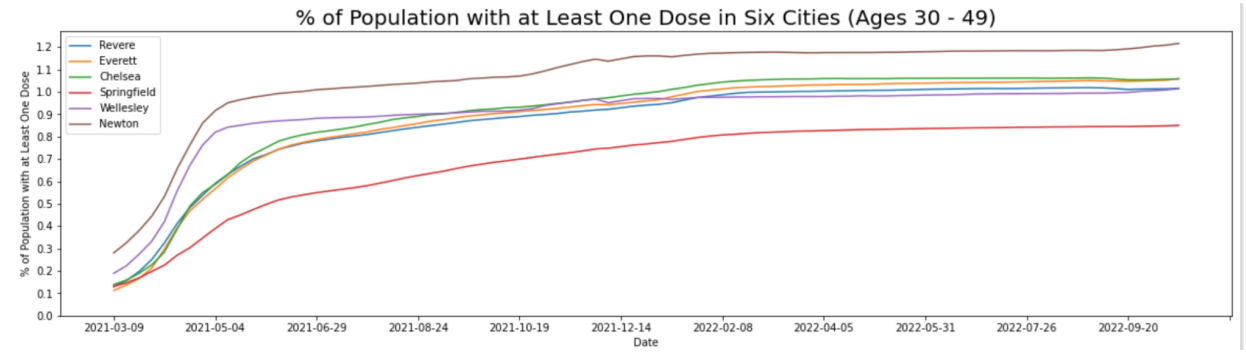

Figure 3-d

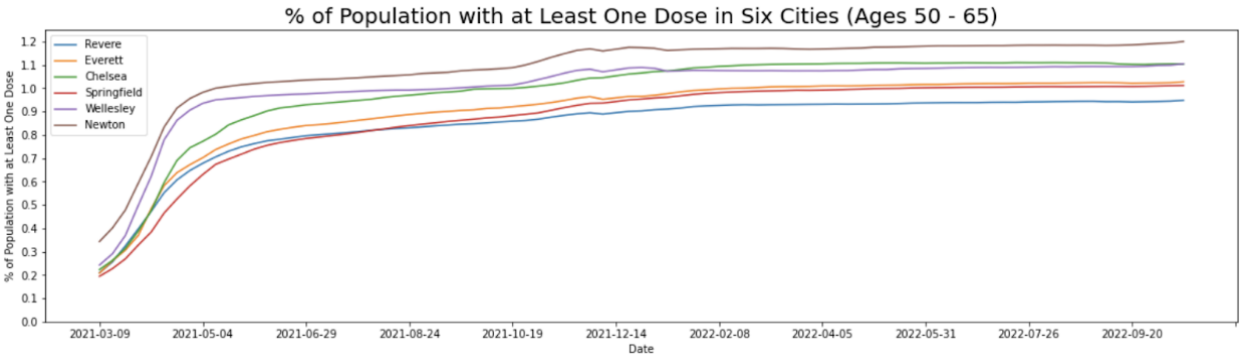

Figure 3-e

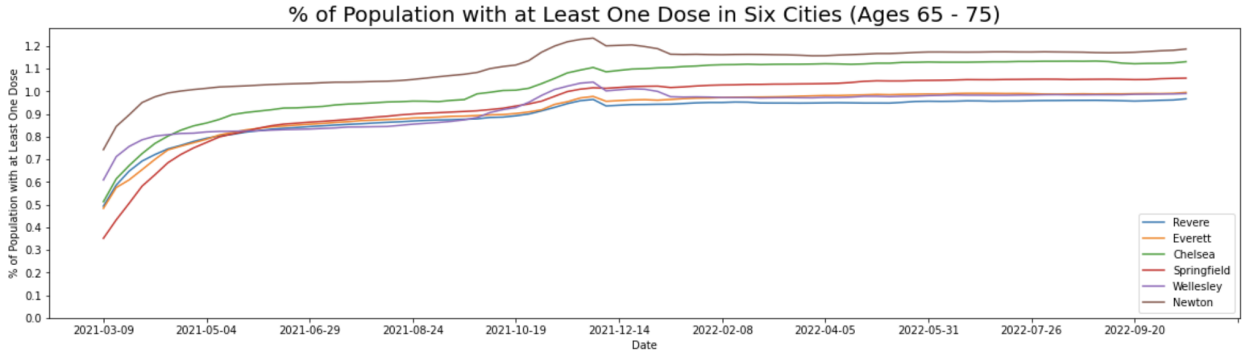

Figure 3-f

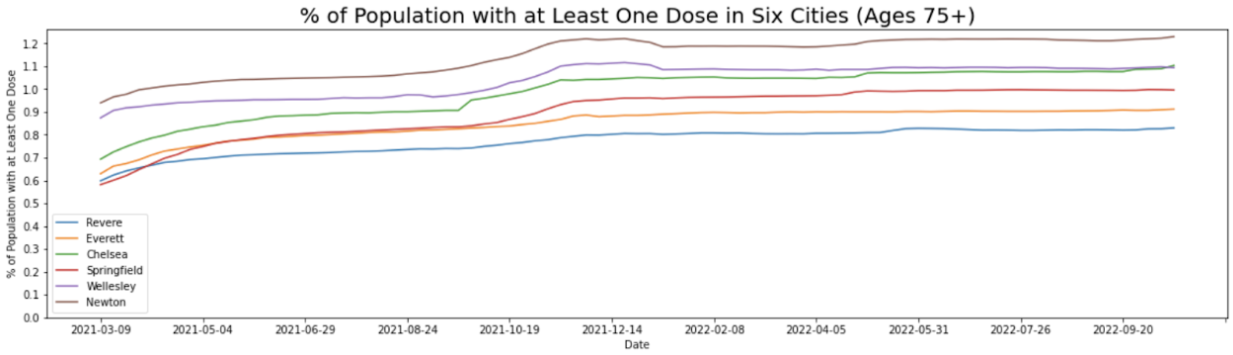

Figure 3-g

In all the graphs of percentage of population with at least one dose, the proportion of people of all age groups who got at least one dose at the end of September 2022 is higher than 100% in Newton and Chelsea, which means that the number of people with at least one dose was more than the total population of the city. The reason might be that Newton and Chelsea had good publicity for vaccination, so many people from other cities would go to these two cities to get vaccinated.

In the 65-75 and 75+ age groups, there was a strange rise and fall in the vaccination rate around December 2021, which was due to the sudden increase and decrease in the number of people with at least one dose during that time. The reason might be that the data was modified after the registration error or there were special events during this period and caused this phenomenon.

In Figure 3-a, which includes all age groups, we can find that around September 2021, Chelsea's vaccination rate was similar to that of other cities, but after that, Chelsea's vaccine growth rate increased a lot, and finally it was the city with the second highest vaccine rate among the six cities.

In the age group 20-29, Revere and Chelsea had high vaccination rates. Chelsea had higher vaccination rates than Revere in all age groups except the 20-29 age group. Overall, Chelsea's vaccination rate was the highest among the six cities, second only to Newton.

4. Detailed Analysis in Revere, Chelsea and Springfield

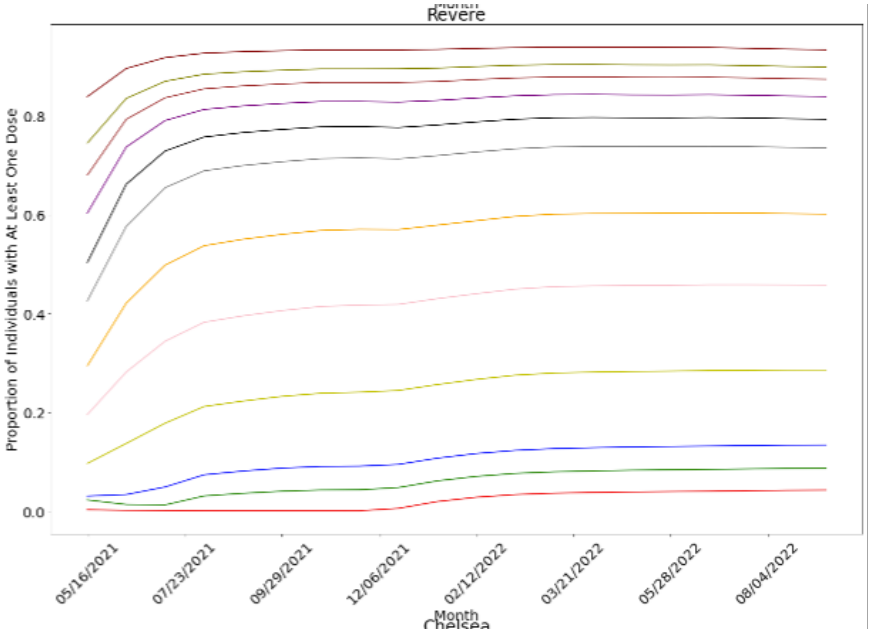

Figure 4-a

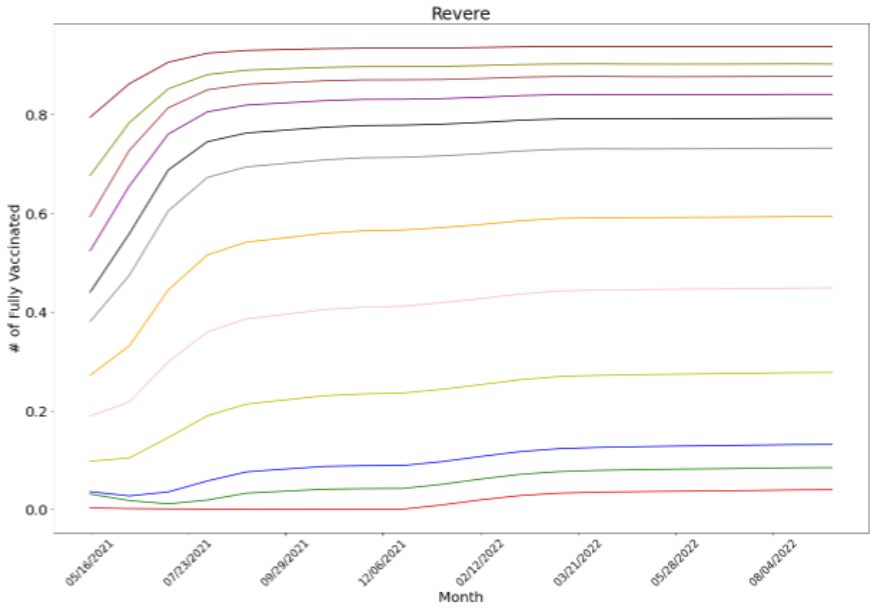

Figure 4-b

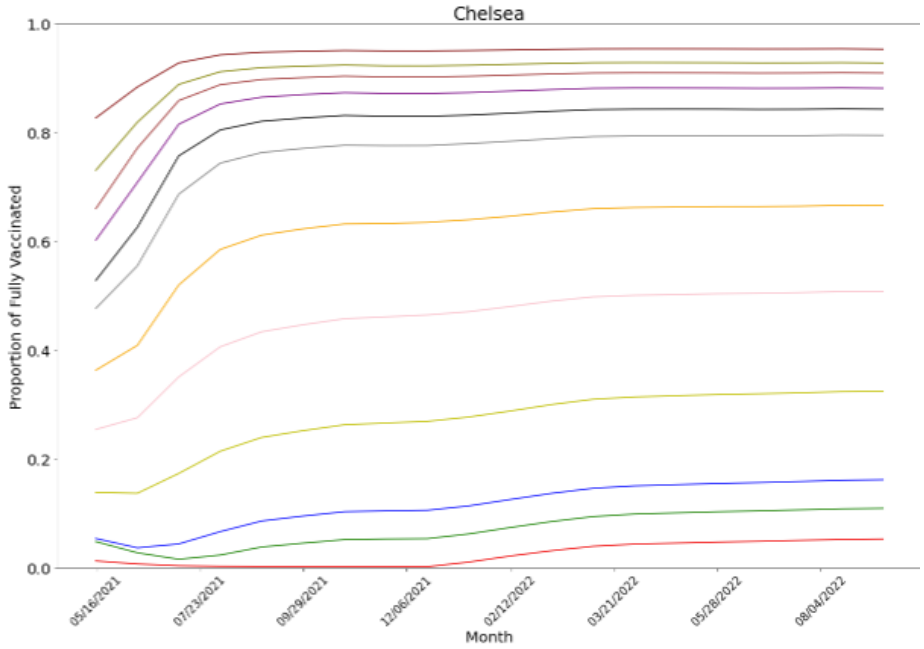

Figure 5-a

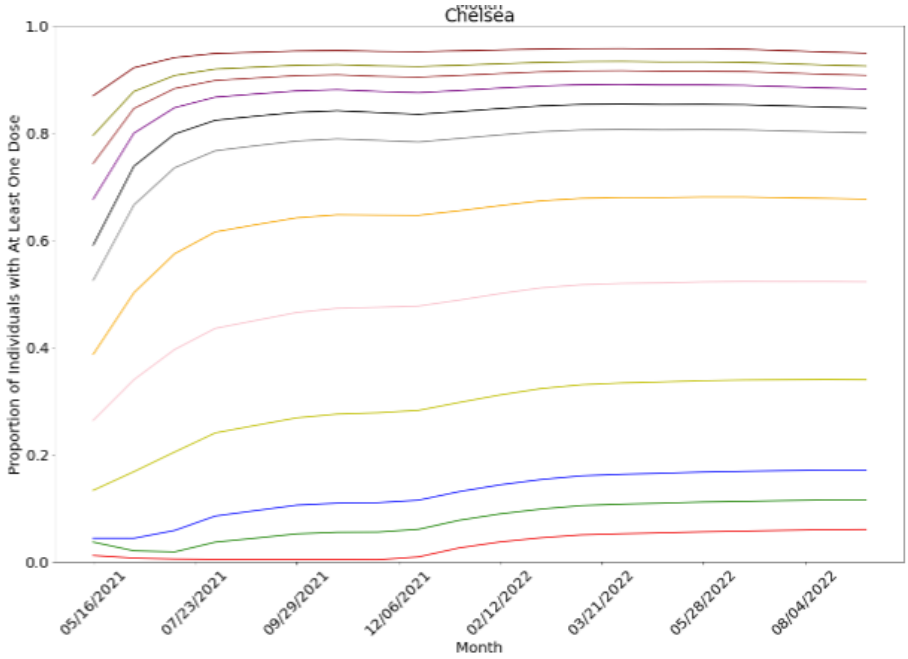

Figure 5-b

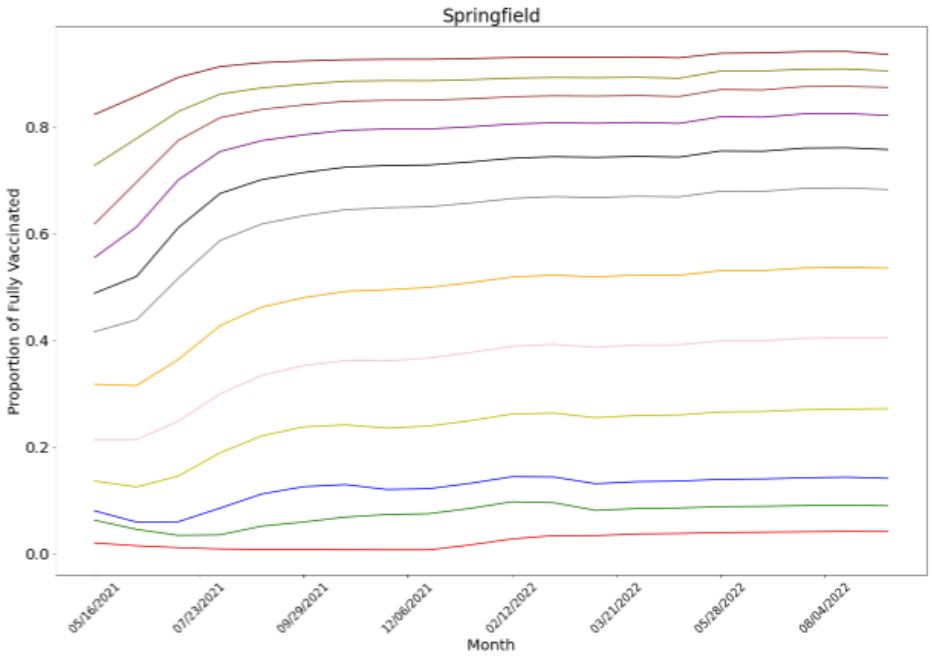

Figure 6-a

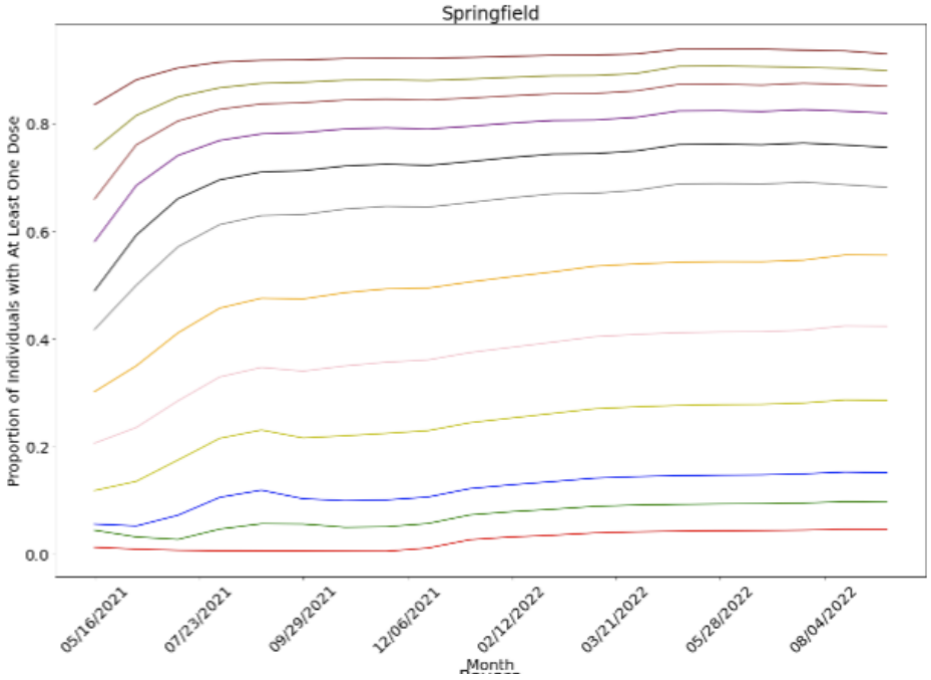

Figure 6-b

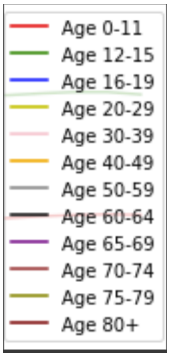

Figure 7

The above figures (4-a, 4-b, 5-a, 5-b, 6-a, 6-c), with figure 7 as their legend, each compare the proportion of fully vaccinated individuals to the proportion of individuals who have received only one vaccination for all age groups for each of the three cities of Revere, Chelsea and Springfield respectively. The graphs display generally similar results with more variation among age groups in the one-dose graph as compared to the fully-vaccinated graph. It is important to note that in general, there exists two groups that are at extreme ends of the spectrum. The proportion of adolescents that have been vaccinated is extremely low. Conversely, the proportion of individuals in older age groups (65+ years of age) that are vaccinated is extremely high. This trend can be seen in both groups of individuals who have only received one dose of a two-dose vaccination and individuals who are fully vaccinated. In general, the graphs show that initially the majority of older individuals had received their vaccination (both full vaccination and one dose) and thus for these populations, their vaccination trend started off extremely high and then leveled off very quickly due to the quick rate at which the elderly had received their vaccinations. In contrast, the proportion of adolescents vaccinated started off low and did not show any notable increase until far into the pandemic. In general, Chelsea has the highest proportion of individuals vaccinated over the course of the pandemic as compared to the other two cities. However, Chelsea also shows the most variability between age groups as compared to the other two cities. As a result, Chelsea has the lowest proportion of adolescents vaccinated out of the three cities, while Springfield has the highest.

When comparing the proportion of fully vaccinated individuals as compared to individuals who had only received one dose of a two-dose vaccination, there is not any significant difference in the graphs, however it is important to note that these exists a slightly higher variability for individuals that have only received one dose as compared to two, such as shown for Revere. This may be attributed to initial hesitancy by individuals in these communities to get vaccinated. It is also important to note that vaccinations only spiked in the initial months of the pandemic and remained relatively stable throughout the rest of the pandemic, suggesting the lack of efforts later on in the pandemic to get more individuals of these communities vaccinated.

In general, these results emphasize the importance/necessity to get younger populations vaccinated as this age group has some of the lowest vaccination rates as well as a general effort to ensure that communities receive reputable information so that more age groups are likely to start their initial vaccinations and/or get fully vaccinated.

V. Hypothesis Test

1. Object and Idea

To verify whether the grassroots organizations like La Colaborativa have significant influence on vaccination, our team chooses to run the hypothesis t test on the key times that La Colaborativa started their adult vaccination, pediatric vaccination, and adult booster in Chelsea. To make sure the data we use in the test is credible, we choose to use the increase rate instead of increase numbers, so we can prevent interference from population. Moreover, we need to make sure the data is from the same time range so that there are no problems about unfairness (if the data comes from different time ranges, the increase rate of the earlier one may be much more than the others since earlier the time, more the vaccination will be taken. The hypothesis t test will also run on every age range since we do not want to make all age groups adding together and omit some possible successful cases.

2. Data modification

Due to the modification to our data, we choose to divide the number of people by the population of that age range. To get the increase rate (%), our team also needs to subtract the data by the previous rate. After getting the increase rate for every age group for all time, we will also choose a reasonable time range for different circumstances.

3. Algorithm and formula

Indeed, for those three circumstances, since we need to verify whether La Colaborativa has significant influence on vaccination, the null hypothesis and alternative hypothesis will be the same:

H_0: La Colaborativa doesn't have a significant influence on (adult/pediatric/booster) vaccination in Chelsea. (Which means $\mu_C = \mu_{othercity}$)

H_a: La Colaborativa has a significant influence on (adult/pediatric/booster) vaccination in Chelsea. (Which means $\mu_C > \mu_{othercity}$)

So here we need to run a one-tailed hypothesis t-test, and the output, which is p-value, once is less than 0.05, then we can believe that it is credible to reject the null hypothesis and accept our alternative hypothesis–the grassroot organizations have significant influence on vaccination. The reason that p-value should be less than 0.05 for us to reject the null hypothesis is that the exceed in mean may be caused by accident or coincidence, and our test works to prevent our result from accidents. If p-value is greater than 0.05, it means that the possibility that our conclusion is caused by accidents is greater than 5%, which violates the rigorous rule of statistics to think the evidence is credible enough to reject the null hypothesis.

To run the hypothesis t-test, we need to compute the mean, standard deviation of each group of data, and also a t table to get p-value. The formation we use is below.

$$t = \frac{\overline{x_1} + \overline{x_2}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where s means standard deviation, x-bar means mean and n means population.

By the t value we compute from the formula, we can use a t table to compute the p-value with 95% confidence interval. There is also an easier way to compute the p-value–by applying the stats package in python, we can use a function to compute p-value with two cities' data.

4. Real Case

First, our team chooses to analyze the effect of La Colaborativa on pediatric vaccination. We separate the data into different age groups as 6-11 years old and 12-15 years old to better figure out whether the vaccination site has an effect or not. Since La Colaborativa starts pediatric vaccination in Jan 2022, we choose the time range from Jan 2022 to June 2022 to maximize the effect of this event. After running t-test with Chelsea and other five cities (Revere, Springfield, Everett, Newton, Wellesley), we only find four tests' output p-value is less than 0.05. The output p-value is $1.94 * 10^{-5}$ and $1.01 * 10^{-6}$for Newton and Wellesley at an age range from 5 years old to 11 years old. Similarly, at the age range of 12 to 15 years old, Newton and Wellesley's p-value is 0.001 and 0.0067.Such a result means only these two cities' result shows that the vaccination site helps Chelsea about the increase of pediatric vaccination. The output of the test shows that La Colaborativa has some influence on pediatric vaccination, but in many cities the hypothesis tests do not pass, so La Colaborativa has, but not significant influence on pediatric vaccination in Chelsea.

The next step for our team is to analyze the influence of La Colaborativa on adult vaccination. To figure out the effect on adult vaccination, we focus on the data of one-dose for adults between Mar 2021 to Aug 2021. After applying t test, a surprising result comes out–unlike our assumption that La Colaborativa has great influence on Chelsea's vaccination, only Newton and Wellesley at 75+ years old have outputs of p-values less than 0.05, with a division of age groups as 16-19, 20-29, 30-39, 40-49, 50-59, 60-64, 65-69, 70-74, and 75+. This result shows that only a tiny part of our data shows the vaccination cites help Chelsea with adult vaccination. Although by comparing these groups mean we can always find that the mean increase rate of Chelsea is more than that of Revere or other cities, we cannot say that this is a credible evidence that the organization helps adult vaccination in Chelsea since the difference of mean increase rate may be caused by accident, and passing hypothesis test can eliminate this possibility. Thus, by the result of the hypothesis test, La Colaborativa does not have a significant influence on adult vaccination.

The final step of the hypothesis test is to run the t test on booster data since the vaccination site of the organization starts adult booster from summer 2022. Our team selects the data from Jun 2022 to Oct 2022 since this time range can best represent the effect of La Colaborativa in Chelsea. By running the hypothesis test, we can find a lot of results' p-values less than 0.05–for all age ranges from 16 to 69, Chelsea's booster vaccination performs better than that of Newton and Wellesley. Moreover, in the age ranges of 16-19, 20-29, 60-64, and 60-69, Chelsea also performs better than Revere for booster vaccination. Such a result implies that the organization does have a significant effect on booster vaccination in Chelsea from our data, and this is credible evidence that the increase rate of boosters in Chelsea is greater than most of other cities in our dataset.

Thus, by verifying with the hypothesis test, La Colaborativa has some influence on pediatric vaccination, almost no influence on adult vaccination for the first dose, and a great effect on adult booster vaccination.

VI. Discussion and Limitations

1. Discussion

According to the graphs and analysis we made, 4 questions listed in the document of project description can be answered.

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 for the first question, Chelsea has higher increasing rates than the average rates in 6 six cities according to the graphs listed above. Also, the total vaccination rate of Chelsea is at the forefront among six cities, sometimes only falling behind Newton. But the gap between Chelsea and other 5 cities is not so enormous. Therefore we may be able to say that grassroots organizations had a statistically effect on vaccination rates, but not so significant. Also, the timeline of the operations of La Colaborativa is unknown, so we can not totally match the increase with the grassroots organization.

For the second question, it seems that age is also another factor that affects vaccination rates. Picture 2-c, 2-f, 3-c, 3-f have better figures than the others, which means that people around 20-29 and 65-75 have higher vaccination rates.

For the third and fourth question, we can not make a precise conclusion that La Colaborativa's efforts directly affect the increase of vaccination rates. Without the timeline of the actions of La Colaborativa, we are not able to attribute the turning point in the graph to the grassroots organization.

2. Limitation

The work we've done also has some limitations. The first is that there are a lot of missing values in the original data, which are centered around ages under 19. Therefore, the way we dealt with the data would more or less affect the result of the graph. This may be one of the reasons why curves in Chelsea don't have a distinction from other cities in the graphs above.

Moreover, without the help of the timeline of La Colaborativa, we can not interpret the increment in the graph easily. For instance, we don't know whether at the specific time , La Colaborativa vaccinated people in the same age period. Therefore, we can not provide a clear understanding of the efforts of La Colaborativa.

VII. Future Work

The next step for our team is to analyze the influence of La Colaborativa on adult vaccination. Also, we hope that we can collect the complete data around ages under 19 so that we can analyze the vaccination situation in all age groups better.

VIII. Sources

1. Assoumou, Sabrina A, et al. "Addressing Inequities in SARS-CoV-2 Vaccine Uptake: The Boston Medical Center Health System Experience." Annals of Internal Medicine, vol. 175, no. 6, 2022, pp. 879–884.
2. Hunter, David J, et al. "Addressing Vaccine Inequity — Covid-19 Vaccines as a Global Public Good." The New England Journal of Medicine, vol. 386, no. 12, 2022, pp. 1176–1179.
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
    b. COVID-data_Massachusetts-vaccines_grouped, https://docs.google.com/spreadsheets/d/e/2PACX-1vSeubClCewwoXfr3kRwLMH1E_2Ue4NLJPM2EBKKxGfMOZY599En3sa0Sga6Q6HaHBcvKo43mxStrSdi/pubhtml