A5: Extension Plan

I. Problem Statement

Since the end of 2019, COVID-19 outbreak has lasted for almost two years, causing significant social and economic impacts. In addition to the direct deaths and spreading viruses, millions of people lost their jobs, students could not enjoy fully in-person education, and businesses are facing downtimes or even closure. With so many negative impacts, the only solution to solve this problem is to end the pandemic. One method largely supported by the government is via vaccination. Currently, there are multiple brands of vaccines approved by the Centers for Disease Control and Prevention (CDC). However, the vaccination rate has not increased much in recent months because some people do not accept these quickly developed vaccines and doubt its effect to prevent COVID-19 or any side effects it may cause. While the side effects may take longer terms, sometimes decades, to appear, we may examine these vaccines' effectiveness on disease prevention with our current data.

II. Research Question

In the common analysis, we perform analysis to see if wearing masks is effective in preventing the spread of coronavirus. In the extension, we perform analysis to see if getting vaccinated is effective in preventing the spread of coronavirus. Both analysis focus on the prevention aspect of COVID-19 to discover potential ways to stop the spread of disease.

III. Data

In addition to data used in common analysis, we also need CDC COVID-19 Vaccinations Data by County to answer our question. This data contains the percentage of population fully vaccinated vs. only one dose by age group and by county. This data will be very useful because it contains daily data, which is the same format as the confirmed cases data we used in our common analysis. We may use this information to visualize the progress of vaccination and compare it to the speed of COVID-19 spread. We may also possibly establish correlation between the two, since both are daily data so our data size is relatively large, even though the timeframe is limited in the two year pandemic.

This dataset has Public Domain U.S. Government license and is available for public use. The data does not contain any personal information for us to identify individuals, and due to the large population, the granularity of data is not too small for us to recover such information. Therefore, we do not anticipate any ethical issues related to the use of data.

IV. Unknowns and Dependencies

There are many factors that may affect the result of the analysis. For example, during the pandemic, there have been several variants being more fatal and spreading faster. It has been assumed that vaccines are less powerful against these variants, as the vaccines were developed and made by researching the original COVID-19 virus. These variants have also caused the

confirmed case number to increase at some points of the time series. However, we cannot control this or include such an effect into our analysis because we practically cannot quantitatively determine the exact number of COVID variant prevention by vaccines.

Another factor is whether the vaccines delay or prevent infections. While we may surely make comparisons in the long term to hopefully rule out the possibility of vaccines delaying infections, we can never be sure of this not happening. There have also been news (might be right or wrong) saying that vaccines are not as effective as time goes on. We have no control over this and without enough evidence we are not able to account for this information into our analysis.

## V. Methodology

The analysis consists mainly of two parts.

The first part is using Fulton County (the assigned county) data and plot the time series of daily confirmed cases and accumulative vaccinations. The goal here is to hopefully establish the negative correlation between them from the visualization. Of course, visualization itself may not be sufficient to support a correlation, so a regression with the two factors might be necessary. If we take this approach to do a regression analysis, then we may also need to perform data manipulations to reduce the dependencies among different data points.

The second part is to use data from another county to compare it with Fulton County. Depending on the vaccination status of the entire Fulton County, we may choose another county with significantly less or more vaccination rate, and plot the confirmed cases. While not as statistical as the first part, doing so does help us remove lots of other variables, such as time sensitive events, from our analysis.

This section is temporary and may change in future depending on the actual data.

## VI. Timeline

Rough timeline:

- 11/16: Complete data collection and cleaning.
- 11/23: Complete first part of analysis.
- 11/30: Complete second part of analysis.
- 12/7: Finish preparing for presentation.

## VII. References


- https://www.researchgate.net/publication/244588315_The_Correlation_Coefficient_An_Overview
- https://par.nsf.gov/servlets/purl/10062095