Ningji Shen
DATA 512
November 10, 2021

A5: Extension Plan

1. **Problem Statement**

   Over the past two years, coronavirus disease 2019 reshaped our world. The datafication of the pandemic gives us the privilege to examine this pandemic, as well as many health related topics, from potentially many different perspectives, to understand how it has changed the society, and how we can control and prevent diseases in the future. By collecting, aggregating, and re-representing many aspects of the individual toll of the pandemic, organizations such as CDC are able to provide human-centered data to the public, and enables scientists and researchers to explore and gain insight from vast datasets. By conducting a human centered data science analysis of some available COVID-19 data, that is, putting the perspectives of the human — be it the end-user, the audience of the research, or the general public who benefits from this research, we intend to understand how various factors and public health policies contributes to the trends of COVID-19. For this project, I aim to discover the correlation between the vaccination rate and COVID-19 cases and deaths in Macomb County, Michigan.

   From the previous common analyzes, I learned the correlation between masking policies and the total number of confirmed COVID-19 cases in Macomb County, Michigan from January 1, 2020 to present. Obviously, besides the masking policies, there are many other factors contributing to the decrease of the number of confirmed cases, and vaccination would be one of the main contributors. Factors such as masking policies and social distance would only help with preventing some percentage of infection, but not preventing death once infected.   Vaccination, on the other hand, is supposed to prevent both infection and death to some degree. By analyzing the datasets related to the COVID-19 vaccination and the disease related data such as the community transmission historical changes and the death counts, I hope to learn if vaccination is a critical variable in the calculation to control the pandemic in Macomb County, Michigan.

2. **Research Question and Hypothesis**

Research question: How did vaccination change the progression of COVID-19 cases and deaths in Macomb County, Michigan in 2021?

Hypothesis: The high percentage of people who are fully vaccinated (have a second dose of a two-dose vaccine or one dose of a single-dose vaccine) leads to the decline in COVID-19 cases and deaths in Macomb County, Michigan.

3. **Data Used**

I will be using the following public datasets, and all of these abide by the ethical considerations because it does not involve any PII or PHI. By analyzing these datasets, I will be able to understand if there is a correlation between the vaccination rate and COVID-19 cases and deaths in Macomb County, Michigan, and how the time series changes.

- COVID-19 data from John Hopkins University from the Kaggle.
  - License: Attribution 4.0 International (CC BY 4.0)
  - This dataset allows me to access the number of confirmed cases and deaths by US county, from January 2020 to present, by US County by Day.
- COVID-19 Vaccinations in the United States,County from CDC.
  - License: Public Domain U.S. Government
  - This dataset allows me to access the total number and percentage of people who are fully vaccinated, and the percentage of county level fully vaccinated population from January 2021 to present, by US County by Day.
  - Note: No COVID-19 vaccines available to the public in 2020, so the assumption is the vaccination rate is zero for all counties in 2020.
- United States COVID-19 County Level of Community Transmission Historical Changes from CDC.
  - License: unknown
  - This dataset allows me to access the following weekly data points, from January 2020 to present, by US County:
    - Total number of new cases per 100,000 persons within the last 7 days
    - Percentage of positive diagnostic and screening nucleic acid amplification tests (NAAT) during the last 7 days
    - Community Transmission Level Indicator [low, moderate, substantial, high, blank]

4. **Unknowns and Dependencies**

There are many aspects of vaccination that are very hard to model. There is a delay between the time of injection and the effective date. Getting vaccinated may simply prevent some percentage of infection, or prevent death once infected. Different brands and types of vaccines have different effectiveness, and different populations within the same county may have different vaccination rates. Additionally, masking likely impacted the apparent effectiveness of vaccination. At this point, I will ignore all these aspects and potential issues of vaccination, and simply focus on the total number and percentage of people who are fully vaccinated.

Aside from the vaccination rate, there are many other potential factors that could be the causes of the decline in COVID-19 cases and deaths, such as the masking policies, the stay-at-home policies, and the different patterns of human interaction. There is also a great chance that although factors such as vaccination will help, vaccines alone will not be sufficient to end the pandemic. However, at this point, I do not have enough resources to research and address this issue in the time allotted.

5. **Methodology**

I will be gathering data for only Macomb County, Michigan from the above datasets. I will then clean and standardize the datasets, and discover if there are any potential issues with these datasets. My curated data will be stored into a single CSV, with all my data points (e.g. number of confirmed cases, number of deaths, percentage of fully vaccinated population, total number of fully vaccinated population, Community Transmission Level Indicator, etc.) consolidated into different columns in a single table, and the date being the unique identifier. During the data curation process, I will be asking myself a series of questions to discover useful information and inform conclusions.

To present my findings, I will create graphs that visualize how the course of the disease was changed by the increasing amount of vaccinations. During this process, I will model different aspects of this problem, and make assumptions if necessary. I will be using a time series showing the changes in the derivative

function of the rate of vaccination over time, and comparing it with the derivative function of the rate of death and infection. I will then analyze the sets of slopes in the time series, and test the difference in the derivative functions. In addition, my graph should also indicate days where the Community Transmission Level is low, moderate, substantial, or high, and whether the difference in the derivative functions was significant. In order to make my analysis reproducible and understandable, I will be documenting the whole process, along with all my code and data.

6. **Timeline to Completion**:

Data collection & curation	11/16/2021
Data analysis & visualization	11/30/2021
Presentation	12/7/2021
Final report	12/14/2021