Part 2 - An Extension Plan

1. Motivation/problem statement:

Covid-19 has in fact impacted many lives over the past 2 years. These years brought in a state of panic and uncertainty! With the goal of finding answers to questions, backed with the available technology, humongous data of all sorts was collected. We were able to gather information like - how many cases are being reported every day, how many people are recovering from covid, what is the number of deaths every day, what is the availability of hospitals and medicines and so on.

All this data in its raw stage is just numbers, which can often be misleading. A number of visualizations were shared across social media displaying the increasing number of cases, irrespective of the measures taken by them. Looking at these charts, people believed that covid is inevitable and also that taking vaccines or wearing masks will not make a difference. I believe that simply presenting raw numbers, cumulative values or increasing time series is unjustified as that provides incomplete information to people.

Hence, with this project, I want to collect and combine multiple datasets, and perform data analysis to answer commonly asked questions and bring to light the actual information. I hope to show that covid 19 did impact our lives, and the actions like use of masks, or getting vaccinated did make a difference. The objective of this project will be to take the data generated by people, and answer questions that people have about covid 19 and its impact.

2. Research questions and/or hypotheses:

While performing the common analysis I used two datasets - the data about the number of cases in Allegheny County and the masking policies within the county during that time. Based on the limited data I was able to infer that the number of cases had two significant spikes, even when the masking policies were in place. Aside from the masking policies, a number of other factors were in play to define the spread of infection such as the masking policy, or availability of hospitals and medicines etc.

Research Question: How do the actions, such as vaccination, hospitalization and masking policy, impact the number of cases and deaths during covid.

3. **Data to be used**:

While performing the common analysis, I did feel that the data is not sufficient to understand how the masking policies had an impact on the number of cases. There was no visible trend or insight from the plot that could confirm or deny the fact that there was a change in the rate of infection. Hence adding more data for this analysis seems like a good decision.

Selecting the right data source is a crucial step for testing a hypothesis or answering a research question. The data needs to come from a trusted and reliable source. Based on the county I was assigned (Allegheny County), I focussed on looking through official government websites to find any relevant data source that would accompany my research question.

The dataset that I will be using for my research is taken from the website - OpenData Commonwealth of Pennsylvania. There are a number of datasets available on this site and I will be referring to the following datasets in particular - COVID-19 Vaccinations by Day by County of Residence Current Health. The dataset consists of the following variables -

Date	Date the vaccine was given.
County Name	County name
Partially Vaccinated	Quantity of vaccine doses administered that provide partial coverage against COVID disease
Fully Vaccinated	Quantity of vaccine doses administered that provide full coverage against COVID disease
First Booster Dose	Quantity of individuals who have received an additional vaccine dose #1 against COVID disease since August 13, 2021.
Second Booster Dose	Quantity of individuals who have received an additional vaccine dose #2 against COVID disease since March 29, 2022
Bivalent Booster 1	Bivalent booster doses include vaccinations using the bivalent formulation of Moderna and Pfizer vaccines.

This data would help me understand how the vaccination policy changes over time and if that had any impact on the spread of infection. The website does provide a Public Domain U.S. Government license for data usage.

4. **Unknowns and dependencies**:

Apart from how covid had an impact on our health and lives, it also had a major impact on the world economy and the environment. For instance, the unemployment rate increased, new jobs in the market reduced while the pollution supposedly decreased. These are not related to my original research question and hence I considered these as supplementary research questions.

Since they are not directly related to the original question, within the limited time I will only focus on these if time permits. I did find some data around unemployment in the allegheny county from this website - Fred Economics

5. **Methodology**:

I was able to access the dataset through the OpenData Commonwealth of Pennsylvania and downloaded a csv file. Based on my current understanding, here is the plan for going forward with this project-

1. Data Cleaning and Preprocessing

   At the first glance I see a lot of zeros in the data. Moreover the data is for the Pennsylvania state as opposed to just Allegheny County. Hence there would definitely be some data cleaning and data preprocessing required before I can use the data.

2. Data Analysis

   Having this additional data, I would be able to perform a more detailed analysis. To investigate the listed research question, I plan on creating time series plots and look for trends and patterns in both the datasets.
   I will also attempt to train a forecasting model (prophet or ARIMA) and use it to forecast what the trend could have been without masking policy or vaccination, and then compare it with actual values to see how these made an impact on the spread of infection.

3. Presentation and Data Visualization

   All the analysis will only be useful if I can communicate this idea to people. I plan on adding comparison visualizations. More specifically, I will be building

2 time series within a single plot for comparison and use hue for categorization further. The focus will be on displaying results from the analysis and model in the form of comparison plots so as to better understand the impact.

6. **Timeline to completion**:

Week 7 - November 10th

- Work on Hypothesis Generation
- Collect Data Sources
- Create a methodology plan
- Submit Part 2 -extension analysis report

Week 8 - November 17th

- Clean and Preprocess the data
- Preliminary Data Analysis
- Build simple visualizations

Week 9 - November 24th

- Perform extensive data analysis
- Build Models to predict and compare
- Begin working on presentation

Week 10 - December 1st

- Analyze model results
- Create Data Visualization for comparison
- Complete presentation

Week 11 - December 8th

- Final in class presentation
- Start writing the report

Week 12 - December 15th

- Complete and submit the report