A5: Extension Plan

Problem statement:
In this extension plan, I want to see the trends in Covid Cases with Vaccination i.e., how the number of cases in the county Illinois changes based on vaccine status. I feel this is interesting and also important to know the number of cases increases as people take vaccines, to see the effect of vaccine and how far it is protecting people. This problem infers to the human-centered data science as it tells us how important the vaccines are from seeing the trends, Similar to how mask mandate reduces the effect of virus spread, we can see how vaccines are important.
To manage the spread and control the pandemic, vaccination is important, knowing its effect in the number of cases help us to take necessary actions against the spread of Covid.

Research questions and/or hypotheses:
The questions I am trying to answer from this Extension Plan is:

- To check if there is any effect of vaccination on Covid (By Hypothesis Testing).
- Also, by using Statistical Methods like Linear regression I want to check how significant is the Vaccination data of different age groups in the county (As mentioned in the dataset).
- The basic question I want to research is to check trends in covid based on vaccination in the county using visualization. (Whether it increases or decrease with time)
- I want to see the visualization of the county vaccination based on different age groups, which may support the regression analysis.
- Check how different areas in the county react to the vaccination.

Data Used:
The Data used in this scenario is taken from the CDC "Link". The data set is large and has 32 columns and 1.09M rows.
The columns in the data set are Date, FIPS, Country of Residence, State of Recipient, etc. the detailed list of 32 columns is found in the dataset link. Below is the table of the columns which I want to use in this scenario.

Column Names
Date
FIPS
Series_Complete_pop_Pct(totoal population percentage)
Recip_County(CountyName)
Recip_State(state Name)
Series_Complete_yes(Total population got vaccinated)

Series_Complete_12Plus (population above !2+plus who got vaccinated)
Series_Complete_12Pluspop_Pct (percentage of 12+ population who got vaccinated)
Series_Complete_18Plus (18+ population who got vaccinated)
Series_Complete_18PlusPop_pct (Percentage of 18+ population)
Series_Complete_65Plus (65+ population who got vaccinated)
Series_Complete_65PlusPop_Pct (percentage of 65+ population who got vaccinated)
Metro_status (If it is a metro or non-metro)

The dataset is publicly available and can be downloaded. The data consists of different age groups' vaccination count and their percentage of vaccinations for a particular county in a given time. The Data is available from 12/13/2020 to 11/09/2021. We can easily filter the data based on the county and combine it with the previous data set based on date and county.
As the contains columns like the total number of people above 12+ vaccinated and total people above 18+ vaccinated it would be helpful to see the regression statistics. Also, based on the county and date we can see the different visualizations trends of the vaccine in different areas in a particular county. It is a huge data set and also, we have the total number of people vaccinated in the county at a particular time with which we can also have hypothesis tests to answer our questions.

Unknowns and dependencies:
The current data set is from 12/13/2020 to 11/09/2021, our A4 common analysis is from February 1, 2020, through October 15, 2021, so there is a gap in the period, I couldn't find a dataset exactly matching this period. So, that is the unknown part of the data.

Methodology:

Step1:
In this step, I will combine the two datasets i..e, the one in A4_common_analysis and the Vaccine_by_county dataset based on the date and county name columns.

Once the data is merged, I would clean the data for any dummy values. Also, will rename the columns to make it easy understanding the requirement.

Step 2:
To check if the Vaccines have any effect on the Covid, I would conduct a null hypothesis saying that there is no effect of the vaccination on the Covid and would conduct a Z-test as we have large data and two components to compare Z test would be the most logical and suitable one. Then I would check the P-value (<0.05) to see if the null hypothesis is rejected or not. (My assumption is it should be rejected)

Step 3:
Using Linear regression, I will check the significance of the vaccine data by taking different age group vaccine count as different factors. i..e considering the columns Series_Complete_12Plus, Series_Complete_65Plus, Series_Complete_18Plus as different factors for a number of daily confirmed cases. Based on the p values and coefficient values we get from regression we can find the most significant factor among the three age groups.

Step 4:
By plotting a visualization graph between the percentage of people vaccinated in a county vs the total number of cases in a county, we can see the trends in covid like whether vaccine reduces the cases or not for the given period. Also, we can plot different graphs for different factors like regions vs total vaccine count in those regions, different age groups taking vaccine vs the total number of cases. By seeing such trends, I wanted to check if these analyses support the output of our statistical methods(Hypothesis and Regression).

Timeline to completion:

The time lines for task submission is as follows:
Step1: Clean and merge the data (11/18/2021)
Step2: Null hypothesis (11/22/2021)
Step3: Linear Regression (12/26/2021)
Step4: Visualizations (11/3/2021)
Step5: Documentation (12/5/2021)