**Motivation:**

The start of 2020 was the beginning of a global pandemic. Scientists and doctors have been developing cures and medicine that may slow down or even stop COVID-19 from spreading. Though medicine has not been created, vaccines have been made by many countries to increase immunization towards the virus. Even though some people were vaccinated right away, many are skeptical towards the vaccine.

Various facilities have been collecting data on vaccines. For example, the Pennsylvania Department of Health has recorded data since the start of the pandemic. However, the data sets are scattered through the websites, which makes it difficult for the public to make sense of it.

The motivation of this project is to provide the public with a view of the effectiveness of the vaccine and whether getting vaccinated is a solution to the pandemic.

**Specific Objectives:**

The objective of this project is to organize the massive amount of data aggregated since the pandemic first started. Create an entity-relationship diagram and perform SQL queries to navigate through the data. We will perform analysis to answer the following questions:

a. Is there a correlation between vaccination rate and the political party a county support?

b. Which counties have better recovery rate?

c. Are there any correlation between vaccination rate and new cases, hospitalize and death rates?

**Data Sources:**

The data we will use is provided by the Pennsylvania Department of Health.

1. "COVID-19 Vaccinations by Day by County of Residence Current Health": This dataset contains the record of the date and number of citizens that received the first dose and second dose of covid vaccines. immunocompromised.[1]
2. “COVID-19 Aggregate Death Data Current Daily County Health”: This table records the total death count and additional deaths every day by county.[2]
3. “COVID-19 Aggregate Hospitalizations Current Daily by County Health”: This table record the hospitalized patients because of COVID-19. In our project, we will be using the COVID-19 ICU record to perform our analysis.[3]
4. “COVID-19 Aggregate Cases Current Daily County Health”: This table records the new additional cases every day by county.[4]
5. “Pennsylvania Current Vote Stats”: This tables records the voting results in Pennsylvania by county.[5]

**Task followed:**

1. Set questions we want to answer.

2. Browse through dataset provided Pennsylvania Department of Health and identified datasets that can help answer questions proposed.

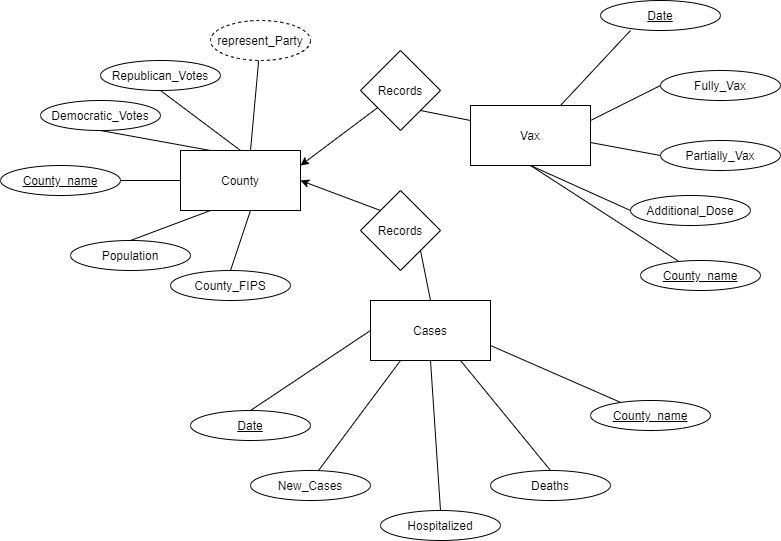
3. Create database and ER-diagram to store and organize useful data.

4. Perform two queries in MYSQL to acquire data for exploratory analysis.

5. Utilize My SQL to query data from the database and perform time series analysis.

6. Create visualization to present our findings and draw conclusion to answer question proposed

**ER-diagram and Summary of Database:**



Graph 1. Project ER-Diagram

Our database is designed to have three tables recording data on County (County), Cases records (Cases), and Vaccination records (Vax). The county table has a primary key county\_Name to uniquely identify each row. In the table, we keep information on the votes for different parties, populations, and FIPS of each county.

In the Cases table, we set county\_Name and Date as our composite primary key. In this table, we kept data on new cases, cases that are hospitalized, and deaths. The records from all three attributes start and end on different dates. To preprocess the data, we decided to set the same range on the dates (from 2020 April 1 to 2021 November 15) of all three attributes in Excel before importing the data into MySQL.

Lastly, the Vax table has a primary key of county\_Name and Date. This table is separated from the “Cases” table because vaccination records are recorded after the invention of the vaccine. Therefore, if we put them in the same table, we will create a database with most data being NULL data points.

The relation of the tables is shown in the ER diagram. Each County records many cases and vaccination records, and every case and vaccination record is recorded by one County.

**Queries, Time Series Analysis, and Insights:**

First query: Our first query is to calculate the fully vaccination rate of individual counties and show the relationship between vaccination rate and the county’s political representative party.

一張含有 文字 的圖片

自動產生的描述

|  |  |  |  |
| --- | --- | --- | --- |
| High Vax\_rate | | Low Vax\_rate | |
| County Name | Represented Party | County Name | Represented Party |
| Montour | rep | Franklin | rep |
| Forest | rep | Tioga | rep |
| Lehigh | demo | Snyder | rep |
| Lackawanna | demo | Susquehanna | rep |
| Allegheny | demo | Juniata | rep |
| Chester | demo | Bradford | rep |
| Northampton | demo | Bedford | rep |
| Montgomery | demo | Potter | rep |
| Butler | rep | Fulton | rep |
| Luzerne | demo | Philadelphia | demo |

Table 1. County Represented Party Order by Vaccination Rate

Insight: From the first query, we can see that the vaccination rate of any county does not exceed 70%. Furthermore, out of the 10 counties that have the highest vaccination rates, 7 of the counties are represented by Democrats. In addition, of the 10 counties with the lowest vaccination rates, only one is represented by Democrats.

Second query: Our second query will focus on listing out the 10 counties that have high recovery rate (we define recovery rate as 1-total Deaths/total Hospitalized). This allows a preliminary insight on how well these counties take care of their patients, and which county’s treatment is more effective to lower death rates.

一張含有 文字 的圖片

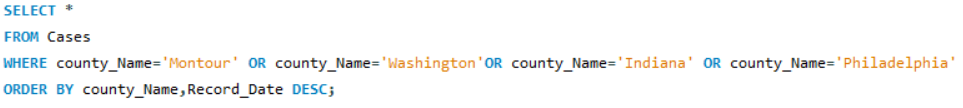
自動產生的描述

Insight: From the second query, we found the list of counties that has the top ten highest recovery rate.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| County | Montour | Braford | Erie | Union | Allegheny | Lehigh | Dauphin | Crawford | Washington | Clearfield |
| Recovery rate | 0.9883 | 0.9577 | 0.9357 | 0.9319 | 0.9214 | 0.9115 | 0.9099 | 0.9002 | 0.8932 | 0.8882 |

Table 2. Top Ten County order by Recovery Rate

**Time Series Analysis:**

In our time series analysis, we will be performing data visualization to show the correlation between vaccination rate and new case rate, hospitalization rate, and death rate individually. We will be choosing four counties with different vaccination rates, with Montour (0.6865) and Philadelphia (0.1254) having the highest and lowest rates and Washington (0.5546) and Indiana (0.4009) having close to average rates among counties. After querying the data needed with the query below, we used Excel to plot our visualization.

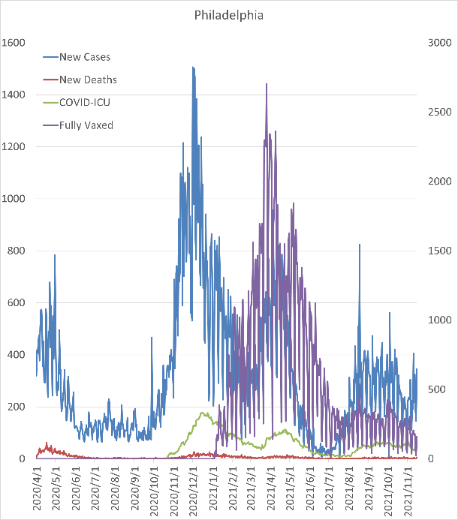
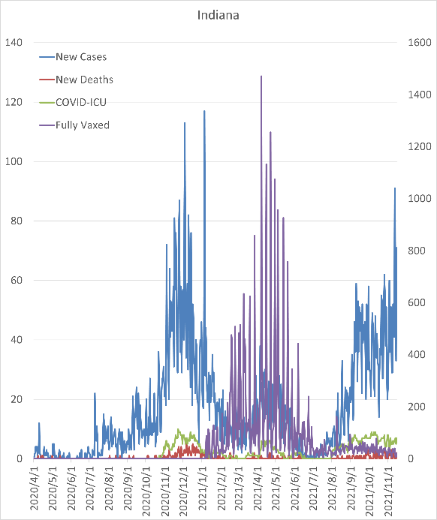
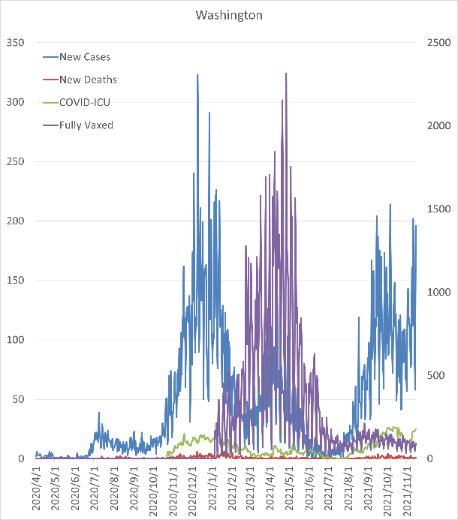
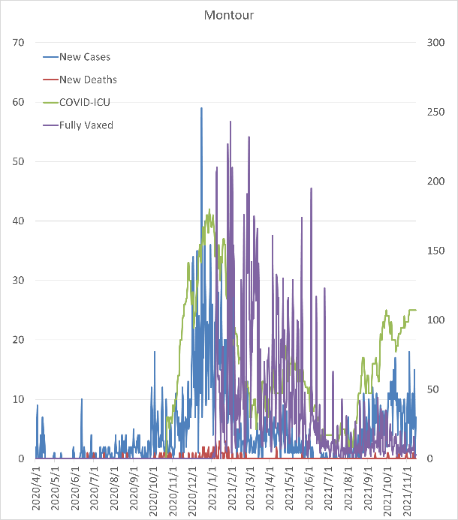


Figure 1. Time Series Visualization of Four Counties with Different Vaccination Rates

In the graphs above, we can see that for all four counties, daily new cases rise from the start of April 2020 and reaches a peak in January 2021. In addition, daily new cases drop significantly after fully vaccination rates increases. However, new cases rates start to rise again in August 2021, which supports the claim that the effectiveness of vaccines lowers after six months of vaccination and booster shots are needed.

While we cannot make a definite conclusion that vaccination rate is the only factor that reduces daily new cases, we can still draw information from the data that vaccination rate has a high chance of lowering daily new cases. Further research will have to be conducted to draw a more accurate conclusion.

**Conclusion:**

We set out to answer three questions for this project. a. Yes, counties that are represented by democrats tend to have a higher vaccination rate. b. The three counties that have the best recovery rate are Montour, Bradford, and Erie. c. Though definite causation can’t be concluded, the vaccination rate does have a high chance of reducing daily new cases according to the data collected, and this causes hospitalization and death cases to drop.

Source:

[1] "COVID-19 Vaccinations by Day by County of Residence Current Health"-

<https://data.pa.gov/Covid-19/COVID-19-Vaccinations-by-Day-by-County-of-Residenc/bicw-3gwi>

[2] “COVID-19 Aggregate Death Data Current Daily County Health”-

<https://data.pa.gov/Covid-19/COVID-19-Aggregate-Death-Data-Current-Daily-County/fbgu-sqgp>

[3] “COVID-19 Aggregate Hospitalizations Current Daily by County Health”-

<https://data.pa.gov/Covid-19/COVID-19-Aggregate-Hospitalizations-Current-Daily-/kayn-sjhx>

[4] “COVID-19 Aggregate Cases Current Daily County Health”-

<https://data.pa.gov/Covid-19/COVID-19-Aggregate-Cases-Current-Daily-County-Heal/j72v-r42c>

[5] “Pennsylvania Current Vote Stats”-

<https://www.dos.pa.gov/VotingElections/OtherServicesEvents/VotingElectionStatistics/Pages/VotingElectionStatistics.aspx>