Analyzing the Correlation between COVID-19 Vaccination Rate and Mortality

Final Project Report

Chris Wang, Claire Tsui, Xiaopeng Wang

- A statement of the problem investigating

The COVID-19 pandemic has posed a significant health crisis and severe economic and technological disruption around the globe, leading to high mortality and economic recession. As one of the most effective ways to stop the spreading of the virus, vaccines greatly mitigate the health issues brought by the virus and protect the elders and children from deadly effects. This project aims to investigate the correlation between COVID-19 vaccination rates and mortality rates. In particular, we seek to determine whether higher vaccination rates are associated with lower mortality rates due to COVID-19.

- Why is this problem relevant?

This problem is relevant due to the crucial role of vaccinations in preventing diseases and lowering death rates. Understanding the correlation between vaccination rates and death rates can provide valuable insights into the effectiveness of public health interventions. The COVID-19 pandemic has highlighted the role of vaccines in saving lives, making this analysis timely and significant. With statistical analysis, this project aims to provide empirical evidence supporting the life-saving benefits of vaccines.

- Where did you get the data from?

The primary dataset for this project is the "covid19postvaxstatewidestats.csv" file, which contains comprehensive statistics on COVID-19 vaccination rates and death rates across California. This dataset includes post-vaccination statewide data, capturing key insights such as the number of vaccinated individuals, COVID-19 death rates, and other relevant health indicators. The data is sourced from a reliable and authoritative database, data.gov, ensuring its accuracy and relevance. The dataset will be downloaded directly and cleaned to handle any missing or inconsistent values, preparing it for detailed statistical analysis.

- Description of the data

This dataset provides a basis for analyzing the correlation between vaccination rates and death rates due to COVID-19 in California. The data includes daily records of COVID-19 cases, hospitalizations, and deaths among both unvaccinated and vaccinated populations.

Key Variables for Analysis

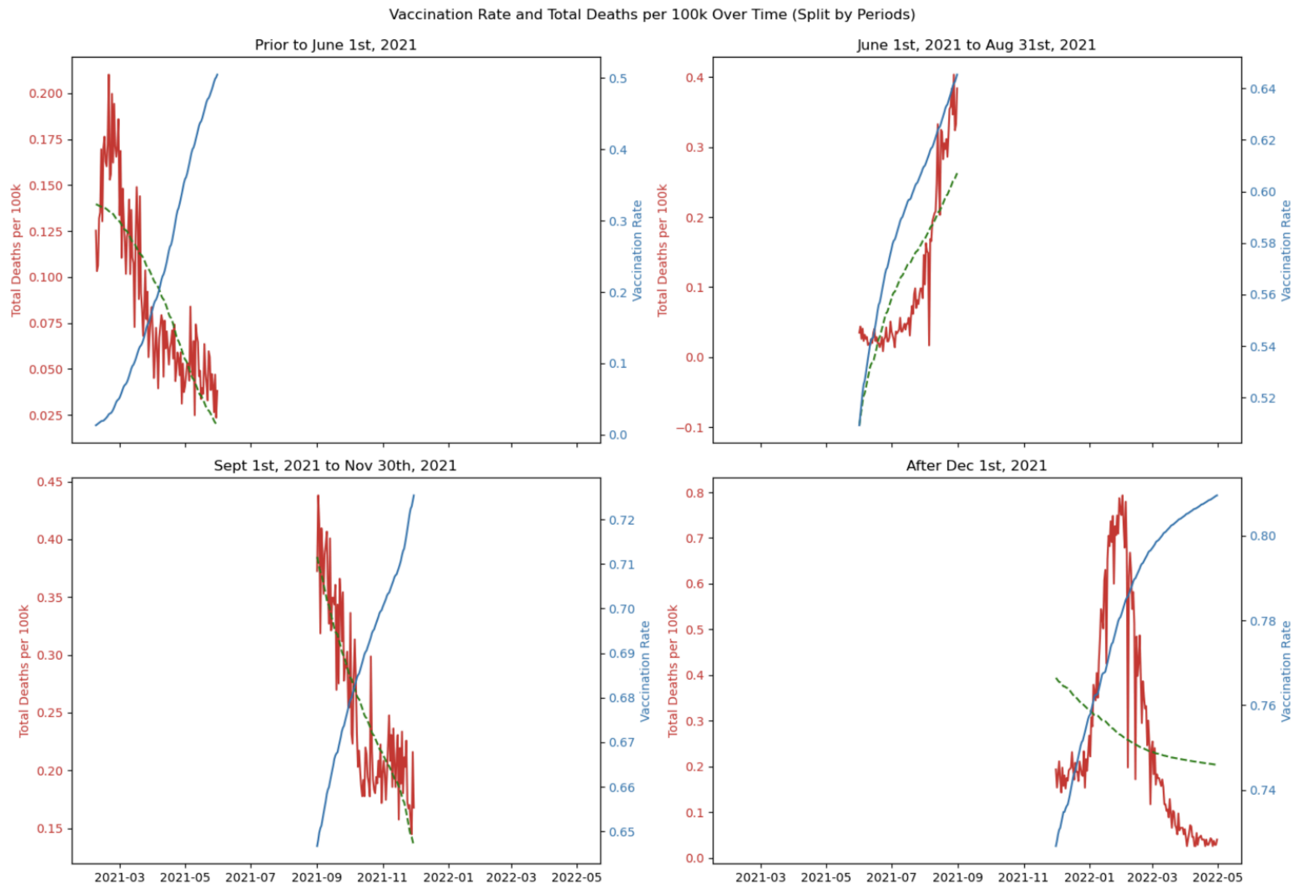
1. **Date**: The specific day for each record, which will help track changes over time.
2. **Population Sizes**:
   - **population_unvaccinated**: The total number of unvaccinated individuals.
   - **population_vaccinated**: The total number of vaccinated individuals.
3. **Health Metrics**:
   - **unvaccinated_deaths**: Number of deaths among the unvaccinated population.
   - **vaccinated_deaths**: Number of deaths among the vaccinated population.
4. **Derived Metrics**:
   - **unvaccinated_deaths_per_100k**: Deaths per 100,000 unvaccinated individuals.
   - **vaccinated_deaths_per_100k**: Deaths per 100,000 vaccinated individuals.

- What types of analyses did you perform?

The analysis of the correlation between COVID-19 vaccination rate and mortality involves few key analyses. The dataset is segmented into four distinct time periods to observe environmental and viral changes over time. Plotting out the line graphs allows us to visualize the total deaths per 100k individuals and vaccination rates over distinct time periods. Statistical analyses, including linear regression and correlation analysis, are conducted to determine the relationship between vaccination rates and death rates. The regression analysis provides a p-value indicating the statistical significance of this topic, while the correlation analysis quantifies its strength and direction. Overall, these analyses suggest that higher vaccination rates are not necessarily associated with higher COVID-19 mortality due to various factors, with more detailed interpretations below.

- How do you interpret the results from these analyses?

Model Fit and Summary Statistics



Based on four line graphs we generated through our codes, we can interpret the results in four periods:

1. **Pre-Delta Period** (Prior to June 1st, 2021):
    a. **Coefficient for variable "vaccination_rate"** = -0.2447
    b. **P-value** = 5.13e-30

    The above results represent that, on average, one unit increase in vaccination rate before June 1st, 2021 leads to a 0.2447 decrease in total deaths per 100k. With an extremely small p-value, the correlation is proved to be statistically significant. This phenomenon is most likely because that before the appearance of delta mutant, the newly invented vaccine, which was first available at Dec, 2020, was proved to have significant efficacy against covid-19 virus, which represented by the significant decline in deaths as the vaccine was popularized within California.

```
Regression Summary for Prior to June 1st, 2021:
                            OLS Regression Results
==============================================================================
Dep. Variable:     total_deaths_per_100k   R-squared:                       0.687
Model:                               OLS   Adj. R-squared:                  0.684
Method:                    Least Squares   F-statistic:                     245.8
Date:                   Sun, 09 Jun 2024   Prob (F-statistic):           5.13e-30
Time:                           16:02:35   Log-Likelihood:                 252.89
No. Observations:                    114   AIC:                            -501.8
Df Residuals:                        112   BIC:                            -496.3
Df Model:                              1
Covariance Type:               nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const                0.1428      0.004     33.574      0.000       0.134       0.151
vaccination_rate    -0.2447      0.016    -15.677      0.000      -0.276      -0.214
==============================================================================
Omnibus:                        2.223   Durbin-Watson:                   0.827
Prob(Omnibus):                  0.329   Jarque-Bera (JB):                2.261
Skew:                           0.311   Prob(JB):                        0.323
Kurtosis:                       2.700   Cond. No.                         6.59
==============================================================================
```

2. **Delta Period** (June 1st, 2021 to Aug 31st, 2021):
    a. **Coefficient for variable "vaccination_rate"** = 2.6459
    b. **P-value** = 2.61e-21

The above result demonstrates that, on average, one unit increase in vaccination rate in the Delta period leads to an increase in total deaths per 100k with a very low p-value, which is statistically significant. To better understand this counterintuitive result, one must combine the outbreak of delta variant to precisely interpret the coefficient. Increasing death cases might be due to the high transmissibility and severity of the Delta variant while the current vaccine had no or minor efficacy on delta variant, so the spike in death cases in this period was potentially due to lag in vaccine coverage or effectiveness against Delta. Thus, based on the correlation in this period, we fail to conclude that the vaccine was ineffective since there was no targeting vaccine at that time.

```
Regression Summary for June 1st, 2021 to Aug 31st, 2021:
                            OLS Regression Results
==============================================================================
Dep. Variable:     total_deaths_per_100k   R-squared:                       0.633
Model:                               OLS   Adj. R-squared:                  0.629
Method:                    Least Squares   F-statistic:                     155.4
Date:                   Sun, 09 Jun 2024   Prob (F-statistic):           2.61e-21
Time:                           16:02:35   Log-Likelihood:                 112.25
No. Observations:                     92   AIC:                            -220.5
Df Residuals:                         90   BIC:                            -215.5
Df Model:                              1
Covariance Type:               nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              -1.4447      0.125    -11.514      0.000      -1.694      -1.195
vaccination_rate    2.6459      0.212     12.466      0.000       2.224       3.068
==============================================================================
Omnibus:                       11.528   Durbin-Watson:                   0.237
Prob(Omnibus):                  0.003   Jarque-Bera (JB):                5.047
Skew:                           0.328   Prob(JB):                       0.0802
Kurtosis:                       2.059   Cond. No.                         38.0
==============================================================================
```

3. **Post - Delta Period** (Sept 1st, 2021 to Nov 31st, 2021):
    a. **Coefficient for variable "vaccination_rate"** = -3.1789
    b. **P-value** = 1.85e-28

The above results showcase a strong correlation between the vaccination rate and death cases per 100k, in particular, on average, one unit increase in vaccination rate leads to a decrease of around 3 death cases, which has a very low p-value, so the correlation in this period is statistically significant. To better understand this phenomenon, combining the release date of the first booster vaccine – Dec 2021 – we can conclude that the effect of booster shot on delta variant is conspicuous that as the vaccination rate increased, the total death rate decreased significantly. Thus, in this period, we can conclude that the booster shot is effective on delta covid virus which leads to a huge decline in deaths.

```
Regression Summary for Sept 1st, 2021 to Nov 30th, 2021:
                            OLS Regression Results
==============================================================================
Dep. Variable:     total_deaths_per_100k   R-squared:                       0.749
Model:                               OLS   Adj. R-squared:                  0.746
Method:                    Least Squares   F-statistic:                     265.7
Date:                   Sun, 09 Jun 2024   Prob (F-statistic):           1.85e-28
Time:                           16:02:35   Log-Likelihood:                 168.00
No. Observations:                     91   AIC:                            -332.0
Df Residuals:                         89   BIC:                            -327.0
Df Model:                              1
Covariance Type:               nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               2.4405      0.134     18.182      0.000       2.174       2.707
vaccination_rate   -3.1789      0.195    -16.300      0.000      -3.566      -2.791
==============================================================================
Omnibus:                       13.344   Durbin-Watson:                   1.319
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                3.993
Skew:                          -0.064   Prob(JB):                        0.136
Kurtosis:                       1.982   Cond. No.                         71.0
==============================================================================
```

4. **Post - Omicron Period** (After Dec 1st, 2021):
    a. **Coefficient for variable “vaccination_rate”** = -2.2999
    b. **P-value** = 2.042e-3

The above results showcase a strong correlation between the vaccination rate and death cases per 100k, in particular, on average, one unit increase in vaccination rate leads to a decrease of around 2.3 death cases, which has a very low p-value, so the correlation in this period is statistically significant. To better understand this phenomenon, it is crucial to consider the impact of the Omicron variant. The Omicron variant, which emerged around November 2021, became the dominant strain by December 2021. Despite its higher transmissibility, the Omicron variant is generally associated with less severe illness compared to previous variants such as Delta. However, the initial surge in cases led to a temporary spike in death rates, particularly among the unvaccinated or those without booster doses. During this period, the vaccination rate continued to rise steadily, which likely contributed to reducing the severity of COVID-19 infections and the death rate. Vaccines, especially with the administration of booster doses, have proven effective in mitigating severe outcomes even with the Omicron variant. Therefore, the increase in vaccination coverage correlates strongly with the observed decline in death rates. In conclusion, the data for this period underscores the effectiveness of the vaccination campaign in reducing COVID-19 mortality rates during the spread of the Omicron variant. The negative coefficient highlights the protective effect of higher vaccination rates against severe outcomes, reaffirming the importance of widespread vaccination and booster doses in controlling the pandemic's impact.

```
Regression Summary for After Dec 1st, 2021:
                            OLS Regression Results
==============================================================================
Dep. Variable:     total_deaths_per_100k   R-squared:                       0.062
Model:                               OLS   Adj. R-squared:                  0.056
Method:                    Least Squares   F-statistic:                     9.855
Date:                   Sun, 09 Jun 2024   Prob (F-statistic):            0.00204
Time:                           16:02:35   Log-Likelihood:                 13.053
No. Observations:                    151   AIC:                            -22.11
Df Residuals:                        149   BIC:                            -16.07
Df Model:                              1
Covariance Type:               nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               2.0653      0.573      3.604      0.000       0.933       3.198
vaccination_rate   -2.2999      0.733     -3.139      0.002      -3.748      -0.852
==============================================================================
Omnibus:                       22.422   Durbin-Watson:                   0.137
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.823
Skew:                           1.060   Prob(JB):                     5.51e-07
Kurtosis:                       2.710   Cond. No.                         64.9
==============================================================================
```

- What are some potential limitations and shortcoming of your analyses?

While our analysis of the correlation between COVID-19 vaccination rates and death cases provides valuable insights, there exists some important limitations and shortcomings in our dataset and the analytical approach. The dataset may suffer from underreporting or inconsistent reporting across different regions and times, leading to inaccuracies in recorded cases, hospitalizations, and deaths. Additionally, the completeness and accuracy of vaccination data are crucial; discrepancies or underreporting can skew the results. The analysis demonstrates a significant correlation but does not establish causation, as other factors like changes in public health policies, healthcare capacity, and behavioral changes are not taken account for. The categorization of variants (Delta and Omicron) overlooks the different characteristics each variant had such as their fatality rate and main symptoms. Furthermore, variations in effectiveness of different booster brands is not considered, and temporal dynamics such as lag effects and seasonal variations are not captured, which can only be assumed from the graphs. These limitations highlight the need for more granular data and comprehensive analysis to better understand the impacts of vaccination and other mitigation measures on COVID-19 mortality.

- Notes on relation to project proposal

After thorough consideration and discussion, we've decided to alter the direction of our initial proposal, which was 'Analyzing the impact of social media usage on mental health'. The initial proposal presented challenges in terms of quantification and sourcing concrete data sets, making it difficult to produce and visualize a solid result for our research. By pivoting to a new direction, the objectives can be more measurable and attainable, ensuring the project's quality and reliability. The new approach allows us to gather and analyze the data in more accurate and actionable insights.

- Previous analysis on the same or similar dataset

There has been research done on the similar datasets and it derived similar results that the vaccine had a significant effect on the mortality rate of COVID-19. However, rather than perform data analysis on the general population like we did, other researchers performed analysis on a dataset with limited population but more specific in terms of the objects, ex: race, sex, age group, etc. The article stated that: "The covariate-adjusted mortality rates were 5.1% and 8.3% for vaccinated and unvaccinated patients hospitalized with COVID-19, respectively, in the whole analysis sample", which is similar to our results.

Reference: https://link.springer.com/article/10.1007/s11606-022-08007-0