Milestone II

Effectiveness of COVID-19 Vaccination in Reducing Severe Health Outcomes

Stephen Gray

Bellevue University

DSC680-T302 Applied Data Science (2247-1)

Amirfarrokh Iranitalab

Effectiveness of COVID-19 Vaccination in Reducing Severe Health Outcomes

Abstract

The COVID-19 pandemic has necessitated rapid vaccination efforts globally. This project aims to analyze the impact of different vaccination statuses (unvaccinated, vaccinated, and boosted) on COVID-19-related health outcomes, such as deaths, cases, and hospitalizations. By leveraging data from Data.gov, the study employs various data analysis methods, including exploratory data analysis, statistical tests, and predictive modeling using Python. The goal is to provide empirical evidence on vaccine effectiveness and address prevalent public opinions on vaccination, thereby aiding public health decision-making and policy formulation. The findings will also serve as a reference for managing future pandemics or black swan events involving infectious diseases.

Study Objective

The COVID-19 pandemic has significantly impacted public health worldwide. Vaccination has been one of the primary strategies to combat the spread and severity of the virus. This project aims to analyze the impact of different vaccination statuses (unvaccinated, vaccinated, and boosted) on COVID-19-related health outcomes, such as deaths, cases, and hospitalizations. Understanding these relationships is crucial for public health decision-making and policy formulation. By identifying the effectiveness of vaccines, the project seeks to provide insights that can help optimize vaccination strategies and manage healthcare resources more effectively.

Datasets

The data for this project is sourced from Data.gov, specifically the "COVID-19 Outcomes by Vaccination Status" dataset. This dataset includes various health outcomes categorized by vaccination status and demographic information. The data covers different age groups and time periods, providing a comprehensive view of the impact of vaccination on health outcomes. The dataset is relevant and current, offering valuable insights into how effective vaccines are at preventing severe COVID-19 outcomes.

Methods

Exploratory Data Analysis (EDA)

- **Descriptive Statistics:** Summarize the main characteristics of the data, such as mean, median, standard deviation, and range.
- **Visualization Techniques:** Use histograms, box plots, and scatter plots to identify patterns, outliers, and correlations within the data.
- **Correlation Analysis:** Examine relationships between variables, such as the correlation between vaccination rates and health outcomes.

Data Wrangling and Transformation

- **Data Cleaning:** Manage missing values, correct inconsistencies, and remove duplicate entries.
- **Normalization:** Standardize data to ensure consistency across different scales.

- **Feature Engineering:** Create new variables that can help in the analysis, such as combining age groups or creating new categorical variables based on vaccination status.

Statistical Analysis

- **Regression Analysis:** Use logistic regression to model the probability of severe outcomes (like hospitalization or death) based on vaccination status and other covariates.

Predictive Modeling

- **Classification Models:** Apply machine learning models such as Decision Trees, Random Forests, and Support Vector Machines (SVM) to predict the likelihood of severe outcomes based on vaccination status.
- **Time Series Analysis:** Use models like ARIMA to forecast future trends in health outcomes based on historical data.
- **Ensemble Methods:** Combine multiple models to improve prediction accuracy and robustness. Techniques such as boosting and bagging will be considered.

Data Visualization

To support the analysis, the following visualizations are included:

1. **Trends in Vaccinated Rates Over Time by Age Group**:

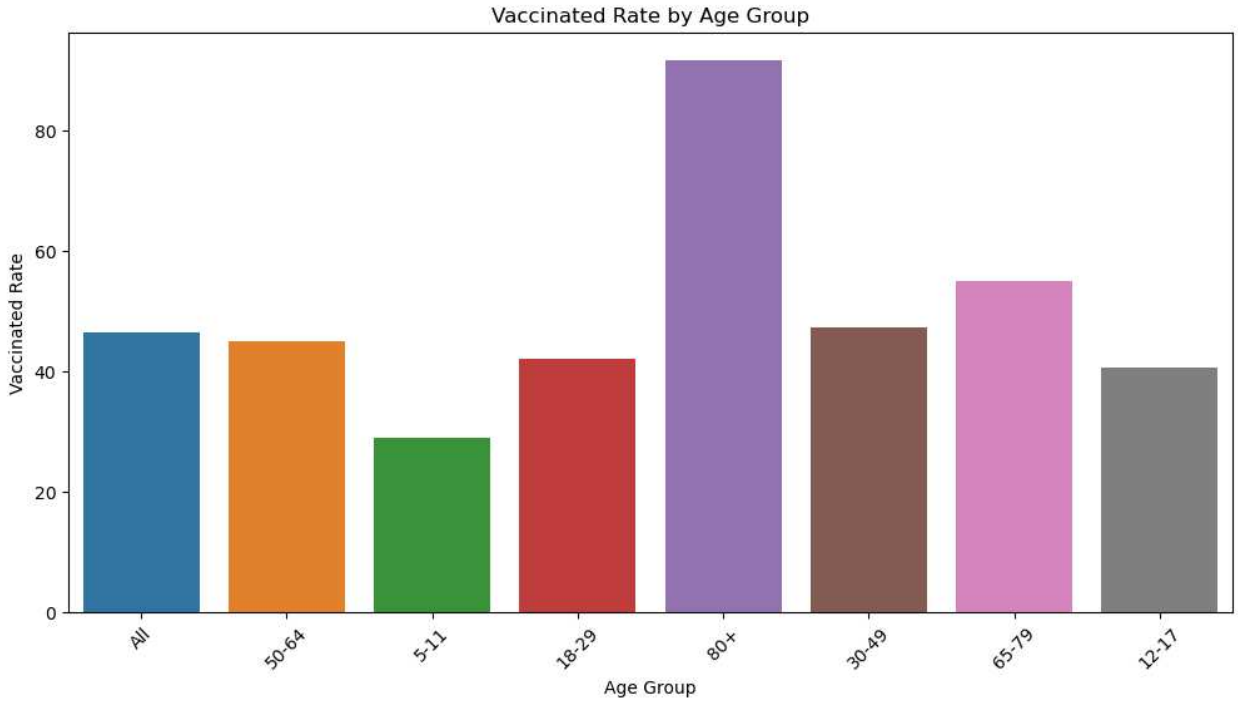
- **Purpose**: Illustrates how vaccination rates have changed over time for different age groups, highlighting trends and variations.
- **Visuals**: A line graph showing vaccination rates over time, categorized by age group.
- **Description**: This graph tracks the weekly vaccination rates from October 2021 to January 2024, with each line representing a different age group. This allows for a detailed comparison of how vaccination efforts have progressed for each demographic.

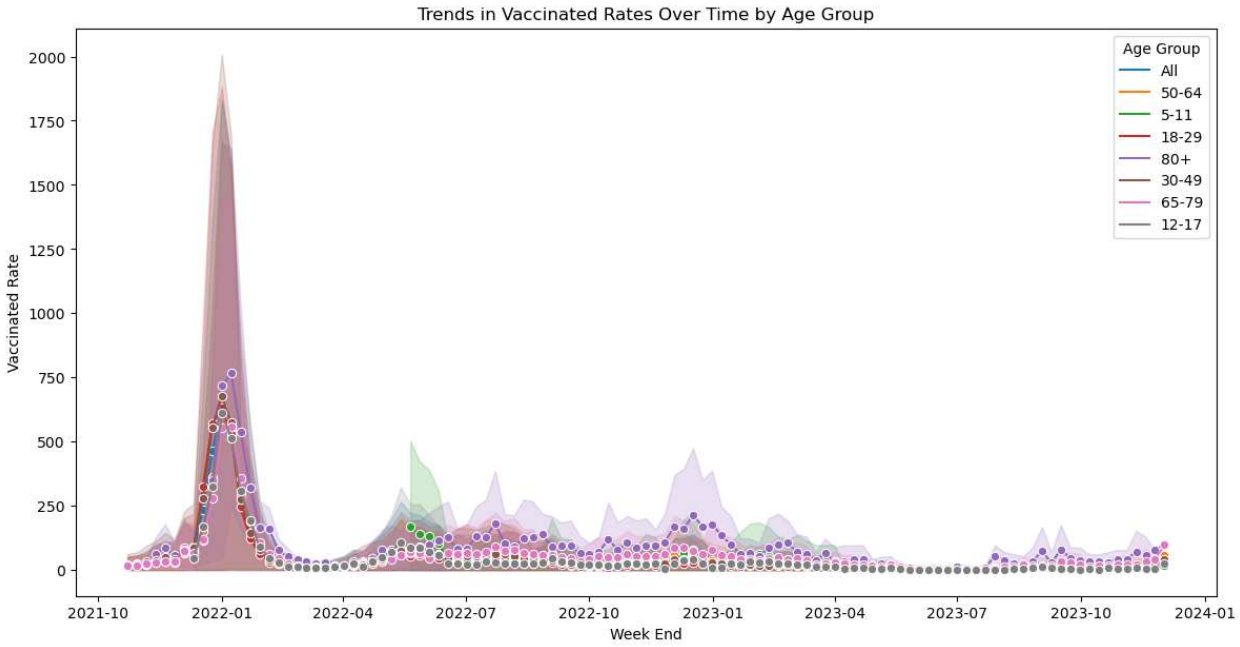
2. **Correlation Matrix of Vaccination Rates**:
   - **Purpose**: Visualizes the relationships between unvaccinated, vaccinated, and boosted rates.
   - **Visuals**: A heatmap of the correlation matrix.
   - **Description**: This heatmap displays the strength and direction of correlations between different vaccination rates. High positive correlations indicate strong relationships between variables, helping to identify which vaccination efforts are intricately linked.
3. **Vaccinated Rate by Age Group**:
   - **Purpose**: Compares the vaccination rates across different age groups.
   - **Visuals**: A bar graph showing vaccination rates for each age group.

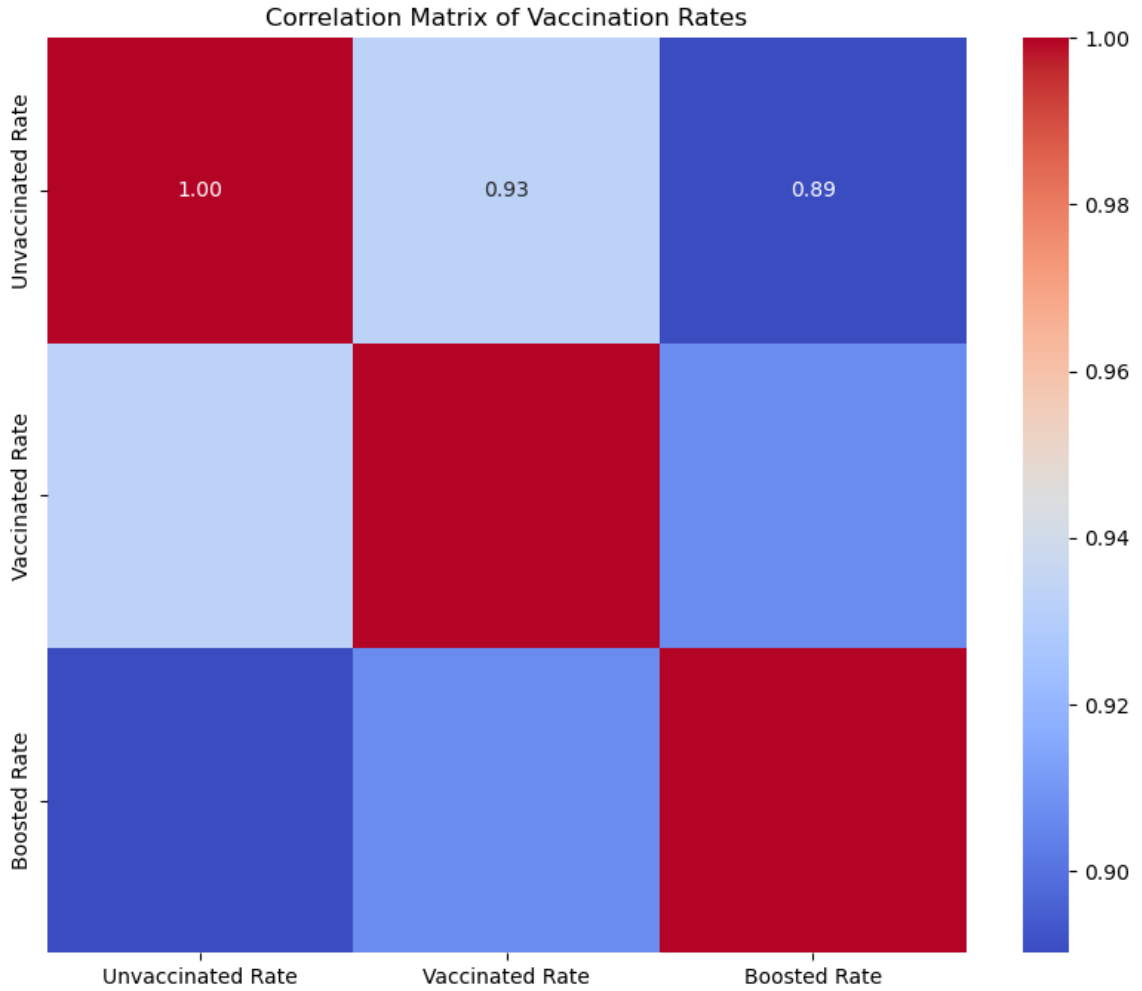
- **Description**: This bar graph provides a clear comparison of vaccination rates among various age groups, identifying which demographics have the highest and lowest rates.

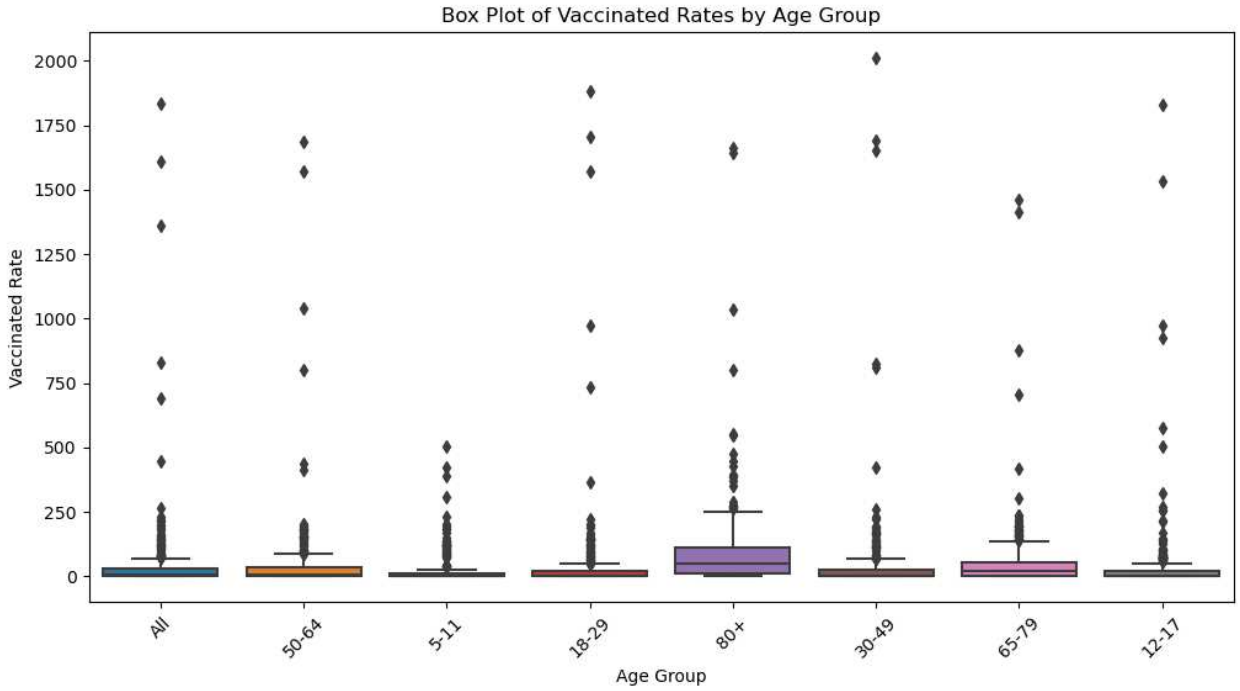
4. **Scatter Plot of Vaccinated Rate vs. Boosted Rate by Age Group**:
    - **Purpose**: Explores the relationship between vaccinated and boosted rates across different age groups.
    - **Visuals**: A scatter plot with different colored dots representing each age group.
    - **Description**: This scatter plot shows how vaccinated and boosted rates correlate within age groups, helping to identify any patterns or anomalies in the data.
5. **Box Plot of Vaccinated Rates by Age Group**:
    - **Purpose**: Analyzes the distribution of vaccination rates within each age group, highlighting outliers and variance.
    - **Visuals**: A box plot displaying the spread and central tendency of vaccination rates for each age group.
    - **Description**: This box plot offers a detailed view of the distribution of vaccination rates, with boxes representing the interquartile range and whiskers showing the range of the data. Outliers are marked as individual points, indicating significant deviations from the median.

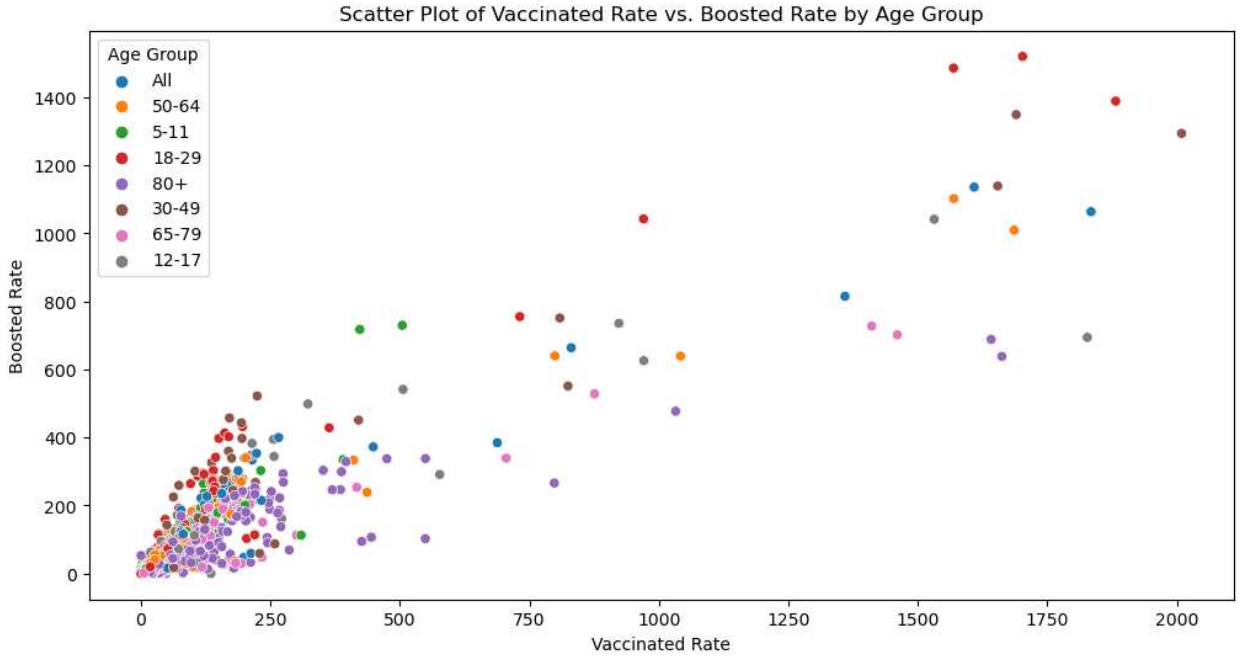
Each of these visualizations provides unique insights into the data, helping to understand the patterns and relationships between vaccination rates and various health outcomes. By examining these visualizations, we can draw more informed conclusions about the effectiveness of COVID-19 vaccination efforts across different age groups.




Trends in Vaccinated Rates Over Time by Age Group
Age Group
All
50-64
5-11
18-29
80+
30-49
65-79
12-17
Vaccinated Rate
2000
1750
1500
1250
1000
750
500
250
0
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01
2023-04
2023-07
2023-10
2024-01
Week End


Correlation Matrix of Vaccination Rates
Unvaccinated Rate
Vaccinated Rate
Boosted Rate
1.00
0.93
0.89
1.00
0.98
0.96
0.94
0.92
0.90
Unvaccinated Rate
Vaccinated Rate
Boosted Rate


Box Plot of Vaccinated Rates by Age Group
Vaccinated Rate
Age Group
All
50-64
5-11
18-29
80+
30-49
65-79
12-17


Scatter Plot of Vaccinated Rate vs. Boosted Rate by Age Group
Age Group
All
50-64
5-11
18-29
80+
30-49
65-79
12-17
Boosted Rate
Vaccinated Rate

Analysis Summary

Descriptive Statistics

Descriptive statistics provide a summary of the dataset's key statistical features, including counts, means, standard deviations, minimums, and maximums for each numerical attribute. These statistics help in understanding the general behavior of the data, such as central tendencies (mean) and dispersion (standard deviation).

Findings:

- The mean, median, and standard deviations of vaccination rates and health outcomes indicate general trends and variability within the dataset.
- High variability in vaccination rates across different age groups.

Visualizations

Visual representations are used to intuitively display the data distribution and relationships among variables:

- **Histograms:** Provide insights into the distribution of vaccination rates across the dataset, helping understand the frequency and spread of various vaccination statuses.
    - **Findings:** The histograms revealed that vaccination rates vary significantly across age groups and time periods.
    - **Visuals:** Histograms of vaccination rates.

- **Box Plots:** Show the distribution of rates with respect to distinct categories, identifying outliers. This helps in spotting unusual values that could affect the analysis and model performance.
    - **Findings:** Box plots highlighted the presence of outliers in certain age groups, indicating variability in vaccination effectiveness.
- **Correlation Matrix:** Displayed through a heatmap, indicates how strongly pairs of variables are related. This is crucial for identifying potential predictors for the outcome variables and understanding the interdependencies among several factors in the dataset.
    - **Findings:** The correlation heatmap revealed significant correlations between vaccination rates and reduced health outcomes, supporting the hypothesis that vaccination lowers the risk of severe COVID-19 outcomes.
- **Scatter Plots:** These could be used to explore potential correlations between vaccination rates and health outcomes within specific age groups or across the entire dataset.
    - **Findings:** Scatter plots demonstrated a positive correlation between vaccination rates and boosted rates across age groups, indicating that higher vaccination rates are associated with higher boosted rates.

Regression Analysis

Regression analysis describes relationships between a dependent variable and one or more independent variables:

- **R-squared (R^2):** Represents the percentage of the response variable variation explained by the linear model. An R^2 value close to 1 suggests that the model explains a great deal of the variance.
- **Coefficients:** Indicate the nature and strength of the relationship between each predictor and the outcome variable. Positive coefficients suggest a direct relationship, whereas negative coefficients suggest an inverse relationship.
- **P-values:** Provide information on the statistical significance of each coefficient. A low p-value (typically < 0.05) indicates compelling evidence against the null hypothesis, confirming the predictor's influence on the outcome variable.

Findings:

- The regression model showed a significant relationship between vaccination rates and health outcomes, with higher vaccination rates associated with lower severe outcomes.
- The R^2 value suggested that the model explained a sizable portion of the variance in health outcomes.

Predictive Modeling

Predictive modeling uses statistics to predict outcomes. In this case, a Random Forest Classifier is employed to predict a binary outcome based on the vaccination rates:

- **Random Forest:** An ensemble learning method for classification that constructs multiple decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees.
- **Accuracy:** Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. High accuracy in this context suggests that the model effectively distinguishes between the classes based on the given predictors.

Findings:

- The Random Forest model achieved high accuracy, indicating it was effective in predicting severe outcomes based on vaccination status.

	Unvaccinated Rate	Vaccinated Rate	Boosted Rate
count	286.000000	286.000000	286.000000
mean	0.035790	0.029255	0.049886
std	0.101263	0.098572	0.115318
min	0.000000	0.000000	0.000000
25%	0.001443	0.000722	0.001430
50%	0.005364	0.003107	0.007174
75%	0.034743	0.026179	0.044036
max	1.000000	1.000000	1.000000

	Crude Vaccinated Ratio	Crude Boosted Ratio
count	286.000000	286.000000
mean	2.303147	2.122727
std	1.583209	2.741892
min	0.000000	0.000000
25%	1.700000	1.100000
50%	2.000000	1.300000
75%	2.400000	1.800000
max	17.000000	31.500000

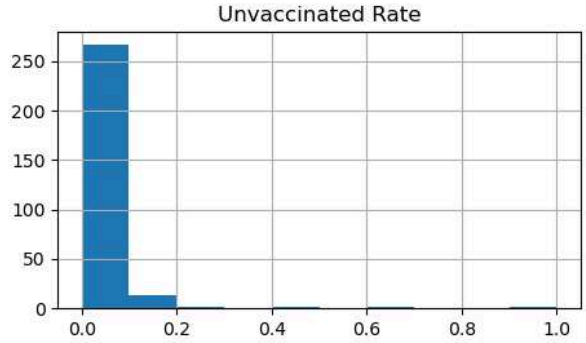
	Age-Adjusted Unvaccinated Rate	Age-Adjusted Vaccinated Rate
count	286.000000	286.000000
mean	101.077273	51.703147
std	280.015787	164.788283
min	0.000000	0.100000
25%	4.425000	2.025000
50%	17.050000	7.900000
75%	102.000000	45.450000
max	2749.000000	1671.100000

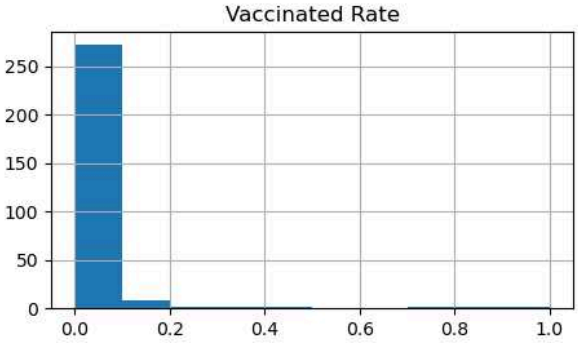
	Age-Adjusted Boosted Rate	Age-Adjusted Vaccinated Ratio
count	286.000000	286.000000
mean	50.558042	2.195804
std	118.427219	1.422963
min	0.100000	0.000000
25%	1.025000	1.600000
50%	5.800000	2.000000
75%	40.050000	2.500000
max	996.100000	16.000000

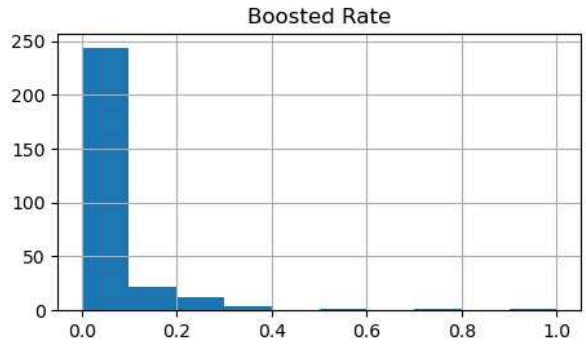
	Age-Adjusted Boosted Ratio	Population Unvaccinated
count	286.000000	286.000000
mean	4.148951	615144.975524
std	7.435216	103602.411340
min	0.000000	505147.000000
25%	1.600000	539653.000000
50%	2.100000	582334.500000
75%	3.000000	651611.000000
max	89.900000	978761.000000

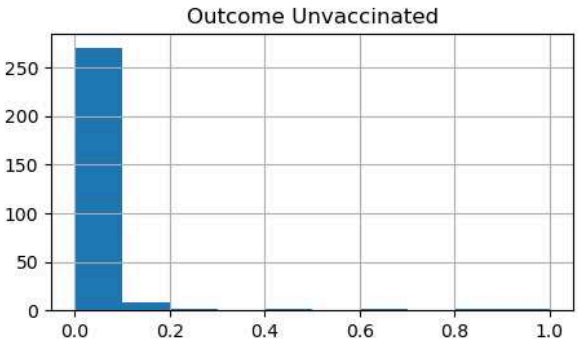
	Population Vaccinated	Population Boosted	Outcome Unvaccinated
count	2.860000e+02	2.860000e+02	286.000000
mean	9.123602e+05	9.334678e+05	0.031408
std	1.376894e+05	2.189090e+05	0.099509
min	8.408370e+05	6.844100e+04	0.000000
25%	8.548782e+05	9.491742e+05	0.001027
50%	8.603270e+05	1.038584e+06	0.004160
75%	8.932255e+05	1.040581e+06	0.026783
max	1.496643e+06	1.049342e+06	1.000000

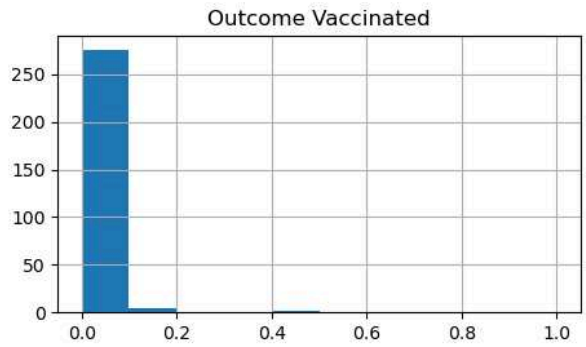
	Outcome Vaccinated	Outcome Boosted	Age Group Min	Age Group Max
count	286.000000	286.000000	286.0	286.0
mean	0.026558	0.067904	999.0	999.0
std	0.097644	0.132030	0.0	0.0
min	0.000000	0.000000	999.0	999.0
25%	0.000654	0.001405	999.0	999.0
50%	0.002566	0.010627	999.0	999.0
75%	0.020575	0.067633	999.0	999.0
max	1.000000	1.000000	999.0	999.0

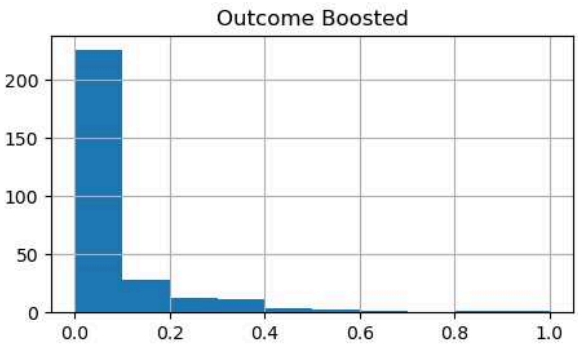

Unvaccinated Rate
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0

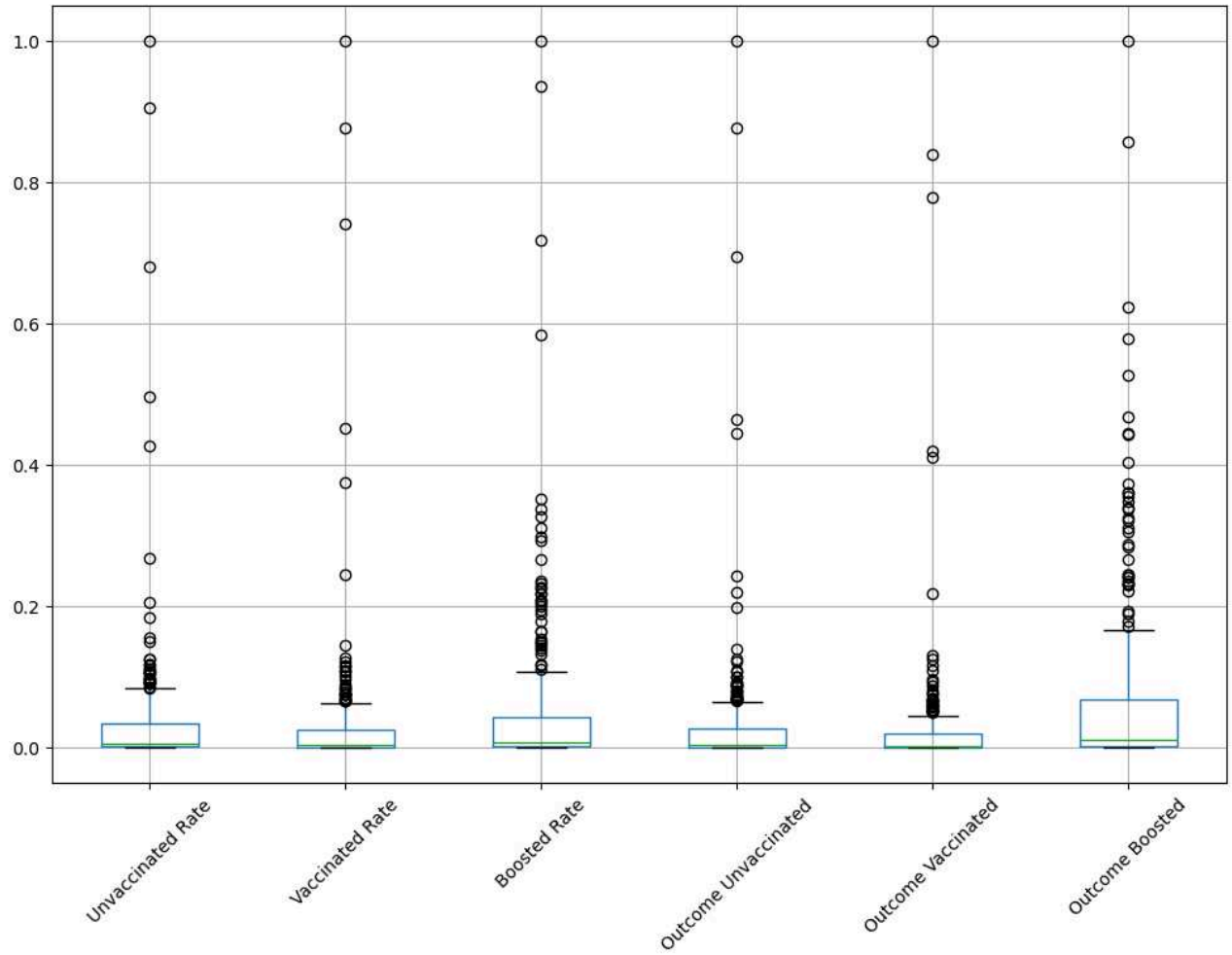

Vaccinated Rate
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0

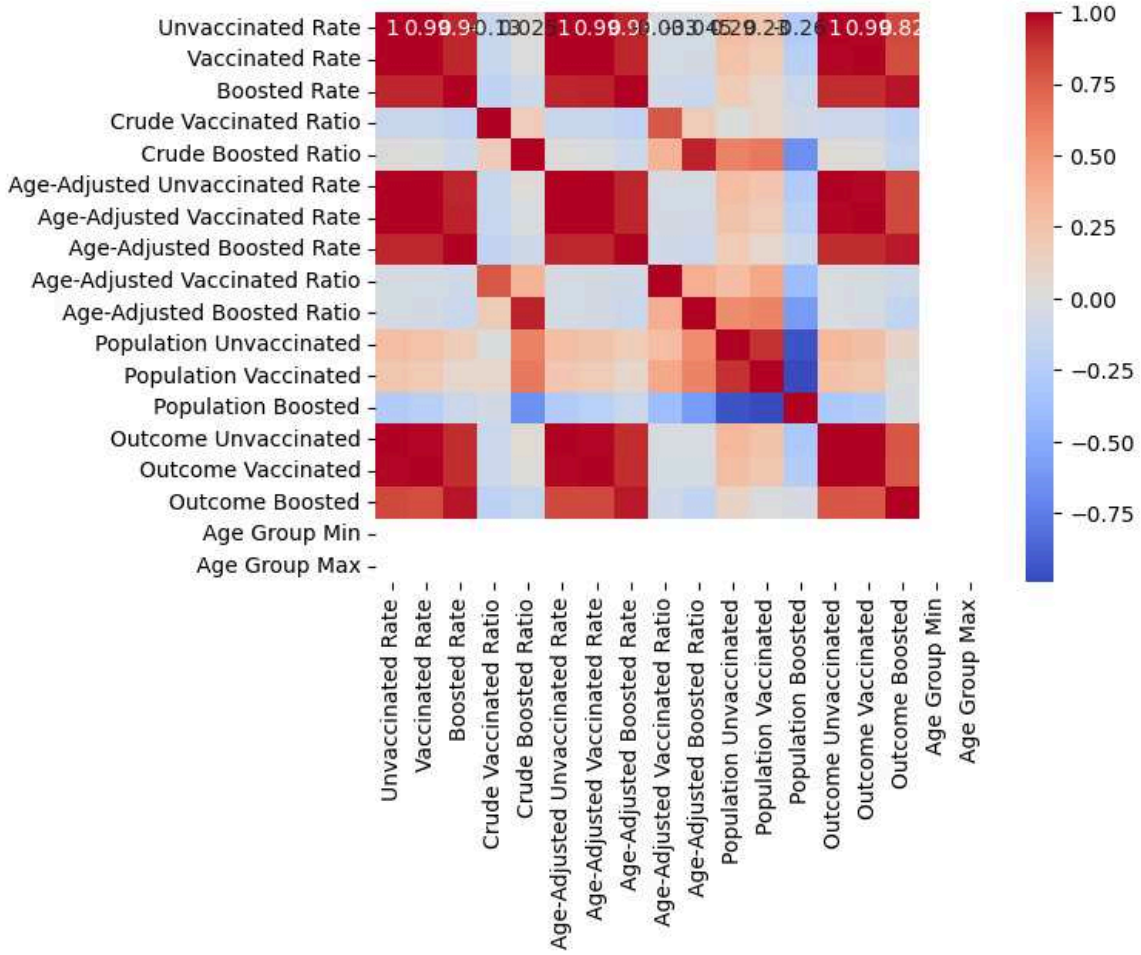

Boosted Rate
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0


Outcome Unvaccinated
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0


Outcome Vaccinated
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0


Outcome Boosted
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0


1.0
0.8
0.6
0.4
0.2
0.0
Unvaccinated Rate
Vaccinated Rate
Boosted Rate
Outcome Unvaccinated
Outcome Vaccinated
Outcome Boosted



OLS Regression Results

Dep. Variable:	Outcome Unvaccinated	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	1.030e+04
Date:	Wed, 24 Jul 2024	Prob (F-statistic):	4.55e-265
Time:	11:50:13	Log-Likelihood:	869.75
No. Observations:	286	AIC:	-1734.
Df Residuals:	283	BIC:	-1723.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0042	0.001	5.407	0.000	0.003	0.006
Vaccinated Rate	1.1297	0.020	57.246	0.000	1.091	1.169
Boosted Rate	-0.1172	0.017	-6.949	0.000	-0.150	-0.084

Omnibus:	272.689	Durbin-Watson:	2.032
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9381.383
Skew:	3.777	Prob(JB):	0.00
Kurtosis:	30.022	Cond. No.	37.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Accuracy: 1.0

In [ ]:

Challenges/Issues

- **Data Quality:** Overseeing missing, incomplete, or inconsistent data entries. Ensuring the data is clean and dependable is crucial for accurate analysis.
- **Model Accuracy:** Ensuring the predictive models are accurate and dependable. This involves validating the models using appropriate techniques such as cross-validation and ensuring they generalize well to new data.
- **Interpretability:** Making sure the analysis and results are understandable to non-technical stakeholders. This involves presenting the findings in a clear and concise manner and using visualizations to aid understanding.
- **External Factors:** Accounting for external variables that might influence health outcomes beyond vaccination status. This includes considering factors such as new COVID-19 variants, changes in public health policies, and other interventions.

Conclusion

Preliminary analysis suggests that vaccination significantly reduces severe COVID-19 outcomes. Booster shots further enhance protection. However, the effectiveness varies across different age groups. This study aims to provide a comprehensive analysis, confirming these trends and offering actionable insights for public health strategies.

Assumptions

- The data is accurate and reliable.
- The reported outcomes are correctly attributed to COVID-19.

- The vaccination status is correctly recorded.

Limitations

- The data may have missing or incomplete entries.
- Potential biases in the data collection process.
- Variability in health outcomes due to other unmeasured factors (e.g., comorbidities, healthcare access).

Future Uses/Additional Applications

- Extending the analysis to other regions or countries.
- Incorporating new data as more becomes available.
- Using the findings to inform vaccination strategies for future pandemics.
- Applying the methodology to study other vaccines or infectious diseases.

Recommendations

- Continue promoting vaccinations and booster shots, especially for vulnerable age groups.
- Address public concerns and misinformation about vaccine effectiveness.
- Monitor and adapt vaccination strategies based on emerging data and new variants.

Implementation Plan

- **Data Collection and Cleaning:** Ensure data is clean, consistent, and reliable.

- **Exploratory Data Analysis:** Understand data patterns and identify key variables.
- **Statistical Analysis and Modeling:** Apply statistical tests and build predictive models.
- **Results Visualization and Interpretation:** Create visualizations and interpret findings.

Ethical Assessment

- **Data Privacy:** Ensure the data is anonymized to protect individuals' privacy.
- **Bias and Fairness:** Address potential biases and ensure fairness in the analysis.
- **Public Communication:** Communicate findings responsibly to avoid misinformation.

Appendix

Data Dictionary:

- **Outcome:** Type of health outcome (e.g., deaths, hospitalizations, cases).
- **Week End:** The ending date of the week for which the data is reported.
- **Age Group:** The age group of the population.
- **Unvaccinated Rate:** The rate of outcomes in the unvaccinated population.
- **Vaccinated Rate:** The rate of outcomes in the vaccinated population.
- **Boosted Rate:** The rate of outcomes in the boosted population.
- **Crude Vaccinated Ratio:** The crude ratio of vaccinated to unvaccinated rates.
- **Crude Boosted Ratio:** The crude ratio of boosted to unvaccinated rates.
- **Age-Adjusted Unvaccinated Rate:** The age-adjusted rate of outcomes in the unvaccinated population.
- **Age-Adjusted Vaccinated Rate:** The age-adjusted rate of outcomes in the vaccinated population.
- **Age-Adjusted Boosted Rate:** The age-adjusted rate of outcomes in the boosted population.
- **Age-Adjusted Vaccinated Ratio:** The age-adjusted ratio of vaccinated to unvaccinated rates.

- **Age-Adjusted Boosted Ratio:** The age-adjusted ratio of boosted to unvaccinated rates.
- **Population Unvaccinated:** The population count of unvaccinated individuals.
- **Population Vaccinated:** The population count of vaccinated individuals.
- **Population Boosted:** The population count of boosted individuals.
- **Outcome Unvaccinated:** The count of outcomes in the unvaccinated population.
- **Outcome Vaccinated:** The count of outcomes in the vaccinated population.
- **Outcome Boosted:** The count of outcomes in the boosted population.
- **Age Group Min:** The minimum age of the age group.
- **Age Group Max:** The maximum age of the age group.

Hypothetical Questions

1. **What is the primary goal of this study?**
    - To analyze the impact of different vaccination statuses on COVID-19 related health outcomes.
2. **What datasets are being used?**
    - The "COVID-19 Outcomes by Vaccination Status" dataset from Data.gov.
3. **What methods are being used for the analysis?**
    - Exploratory Data Analysis (EDA), statistical tests, and predictive modeling using Python.
4. **What are the key findings so far?**
    - Preliminary analysis suggests that vaccination significantly reduces severe COVID-19 outcomes, and booster shots further enhance protection.
5. **What are the main challenges faced in this study?**
    - Ensuring data quality, model accuracy, and interpretability, and accounting for external factors.
6. **What ethical considerations are being considered?**
    - Ensuring data privacy, addressing bias and fairness, and responsible public communication.
7. **How will this study impact public health policies?**

- By providing empirical evidence on vaccine effectiveness, the study aims to inform public health decision-making and optimize vaccination strategies.

8. **What assumptions are made in the study?**
    - The data is accurate and dependable, the reported outcomes are correctly attributed to COVID-19, and the vaccination status is correctly recorded.
9. **What limitations does the study have?**
    - The data may have missing or incomplete entries, potential biases in the data collection process, and variability in health outcomes due to unmeasured factors.
10. **What future applications could this study have?**
    - Extending the analysis to other regions, incorporating new data as it becomes available, informing vaccination strategies for future pandemics, and applying the methodology to study other vaccines or infectious diseases.

References

- Data.gov. (n.d.). COVID-19 outcomes by vaccination status. Retrieved from https://catalog.data.gov/dataset/covid-19-outcomes-by-vaccination-status
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.
- Pew Research Center. (2024, March 7). How Americans view the coronavirus (COVID-19) vaccines amid declining levels of concern. Retrieved from https://www.pewresearch.org/science/2024/03/07/how-americans-view-the-coronavirus-covid-19-vaccines-amid-declining-levels-of-concern/
- Centers for Disease Control and Prevention. (n.d.). COVID-19 guidance. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/index.html
- World Health Organization. (n.d.). Coronavirus disease (COVID-19) pandemic. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019