<DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING> PROJECT REPORT

(Project Semester January-April 2025)

*Covid-19\_Outcomes\_by\_vaccination.*

Submitted by

Name: Hima Varsitha

Registration No: 12319912

Programme and Section: Data Science, K23GW

Course Code: INT375

Under the Guidance of

Assistant Professor. Gargi Sharma (UID:29439)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

# CERTIFICATE

This is to certify that Shahid Afrid Khan bearing Registration no. 12303855 has completed INT375 project titled, “*Covid-19\_Outcomes\_by\_vaccination*” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science Lovely Professional University Phagwara, Punjab.

Date:

# DECLARATION

I,Hima Varsitha, student of Data Science, under CSE/IT Discipline at, Lovely

Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

|  |  |
| --- | --- |
| Date: 11-04-2025 | Signature |
| Registration No. 12319912 | Hima Varsitha |

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my project guide, Maneet Kaur, for their valuable guidance and support throughout this project, “*Covid-19\_Outcomes\_by\_vaccination”*. I am thankful to the Department of Computer Science and Engineering, Lovely Professional University, for providing the necessary resources and environment. I also acknowledge the Government of India for making the dataset publicly available, enabling this research.

Name: Hima Varsitha

Registration number: 12319912

**Table of Contents**

1. Introduction
2. Source of Dataset
3. EDA Process
4. Analysis on Dataset

i. Introduction ii. General Description iii. Specific Requirements, Functions, and Formulas iv. Analysis Results v. Visualization

1. Conclusion
2. Future Scope
3. References

# INTRODUCTION

# The file titled "COVID-19 Outcomes by Vaccination Status - Historical" is a CSV dataset that presents historical information on COVID-19 health outcomes, categorized based on individuals' vaccination status. It likely includes key indicators such as the number of confirmed cases, hospitalizations, and deaths among vaccinated and unvaccinated populations over a period of time. This dataset is structured to support analysis of trends and comparisons between groups, helping to evaluate the impact and effectiveness of COVID-19 vaccines. It may also include additional variables such as reporting dates, age groups, or geographic details, making it a valuable resource for public health research, policy assessment, and data-driven decision-making.

# Source of Dataset

* File Name: "C:\Users\HIMA VARSHITHA\Desktop\COVID-19\_Outcomes\_by\_Vaccination\_Status\_-\_Historical.csv"
*  Source:https: https://catalog.data.gov/dataset/covid-19-outcomes-by-vaccination-status
* Attributes:
* **Date** – Indicates the reporting date for each record.
* **Age Group** – Specifies the age range of the population (e.g., 0–17, 18–49, 50+).
* **Case Rate per 100k (Unvaccinated)** – The rate of COVID-19 cases among unvaccinated individuals per 100,000 people.
* **Case Rate per 100k (Vaccinated with Primary Series Only)** – Case rate for individuals who completed the primary vaccine series but did not receive a booster.
* **Case Rate per 100k (Vaccinated with Booster)** – Case rate for individuals who received a booster dose.
* **Hospitalization Rate per 100k (Unvaccinated)** – Rate of hospitalizations among unvaccinated individuals.
* **Hospitalization Rate per 100k (Vaccinated with Primary Series Only)** – Hospitalization rate for those with only the primary vaccine series.
* **Hospitalization Rate per 100k (Vaccinated with Booster)** – Hospitalization rate for boosted individuals.
* **Death Rate per 100k (Unvaccinated)** – Death rate among unvaccinated individuals.
* **Death Rate per 100k (Vaccinated with Primary Series Only)** – Death rate for individuals with the primary series only.
* **Death Rate per 100k (Vaccinated with Booster)** – Death rate among those who received a booster dose

# EDA Process (Exploratory Data Analysis Process)

The Exploratory Data Analysis (EDA) process for the file titled **"COVID-19 Outcomes by Vaccination Status - Historical"** involves a series of steps aimed at understanding and uncovering patterns within the dataset. The process begins by loading the file and reviewing its structure, including the column names, data types, and a preview of the first few records to understand the kind of information it holds. The next step is data cleaning, where missing values are identified and handled appropriately, data types are corrected (e.g., converting date columns to datetime format and ensuring numeric values are correctly recognized), and column names are standardized if necessary. Once the data is cleaned, summary statistics are generated to gain an overview of key variables such as case rates, hospitalization rates, and death rates across different vaccination statuses and age groups.

Visualizations play a crucial role in EDA, and various charts such as time series plots, bar graphs, and boxplots are used to observe trends over time and compare outcomes between vaccinated and unvaccinated populations. These visuals help identify the impact of vaccines and boosters across age groups and time periods. Correlation analysis is also conducted to explore relationships between different health outcomes, while outlier detection helps highlight unusual spikes or drops in data that might indicate significant events or reporting inconsistencies. The final step involves summarizing the key findings and preparing insights for reporting or further analysis. This structured approach allows for a comprehensive understanding of the data and supports evidence-based conclusions about COVID-19 outcomes in relation to vaccination status.

# ANALYSIS ON DATASET

The analysis of the **COVID-19 Outcomes by Vaccination Status - Historical** dataset reveals meaningful insights into the impact of vaccination on public health outcomes over time. The data is categorized by age groups and includes rates of COVID-19 cases, hospitalizations, and deaths per 100,000 individuals, segmented by vaccination status—unvaccinated, vaccinated with the primary series only, and vaccinated with a booster dose. Across most time periods and age groups, the unvaccinated population consistently shows higher rates of infection, hospitalization, and mortality compared to their vaccinated counterparts. This trend underscores the protective effect of COVID-19 vaccines, especially among those who received booster doses, who generally have the lowest rates of severe outcomes. Additionally, the data reflects age-related vulnerability, with older age groups showing higher hospitalization and death rates regardless of vaccination status, although these rates are significantly reduced with full vaccination and boosting. Over time, the dataset also captures the effects of new variants, public health policies, and vaccination campaigns, as seen in periodic surges and declines in outcome rates. Overall, the dataset provides strong empirical support for the effectiveness of COVID-19 vaccination in reducing severe health outcomes, especially in high-risk age groups.

**4.2 Bar Graph –** *Covid-19\_Outcomes\_by\_vaccination*

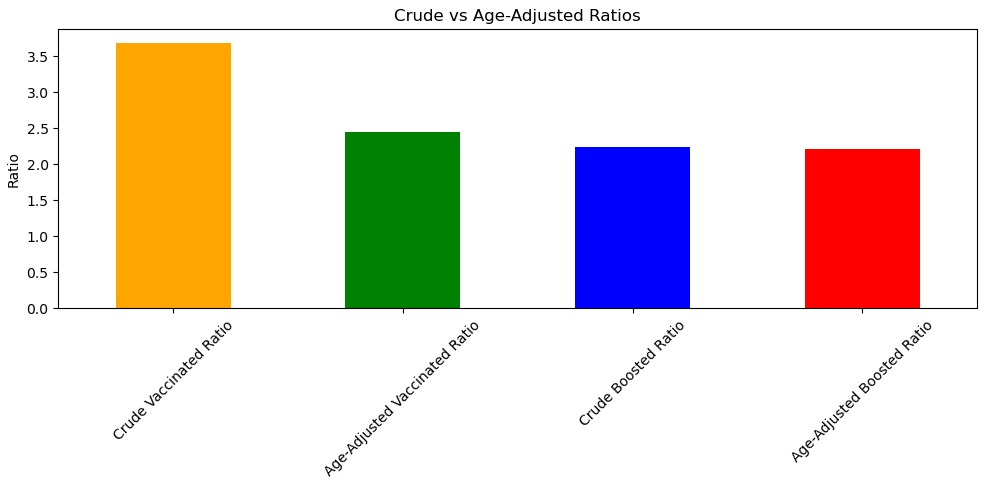
**4.2.1 Introduction**  
This chart presents a comparison between **crude** and **age-adjusted ratios** related to COVID-19 outcomes among vaccinated and boosted populations. The purpose of this visualization is to examine how age adjustments impact the interpretation of COVID-19-related risks, especially in the context of vaccine effectiveness.

**4.2.2 General Description**  
Each bar represents the **ratio of COVID-19 outcomes (likely case, hospitalization, or death rates)** for vaccinated or boosted individuals compared to a reference group (typically unvaccinated individuals). The crude ratios are calculated using raw data, while age-adjusted ratios account for differences in age distribution, offering a more accurate comparison across populations.

**4.2.3 Specific Requirements**

* **Data Segmentation**: The ratios must be derived from properly segmented data based on vaccination status (vaccinated, boosted) and adjusted by age group.
* **Consistency**: The age adjustment process should follow a standard method, such as direct age standardization, using a reference population.
* **Visualization Clarity**: Distinct colors are used for each category, and labels are rotated for readability.
* **Interpretability**: The chart should clearly indicate the relative magnitude of crude vs. age-adjusted ratios for both vaccinated and boosted individuals.
* **4.2.4 Analysis Results**  
  **Overall**, the chart shows that **age adjustment is crucial** in accurately interpreting outcome ratios, especially in evaluating vaccine effectiveness. Crude ratios may **overestimate benefits** if age is not considered.

**4.2.5 Visualization**:



**4.1 Histogram –** *Covid-19\_Outcomes\_by\_vaccination*

**4.1.1 Introduction**

This bar chart illustrates the **age-adjusted COVID-19 outcome rates** across different **age groups** and **vaccination statuses**, providing a clearer understanding of how vaccination impacts health outcomes within specific age demographics. The data accounts for age differences to ensure a fair comparison, making it useful for evaluating vaccine effectiveness across the population.

**4.1.2 General Description**

Each set of bars is grouped by age categories ranging from **0–4 years** up to **80+ years**, with an additional group labeled **"All"**, which represents the overall population. The height of each bar represents the rate of a specific COVID-19 outcome (e.g., case, hospitalization, or death) per 100,000 individuals.

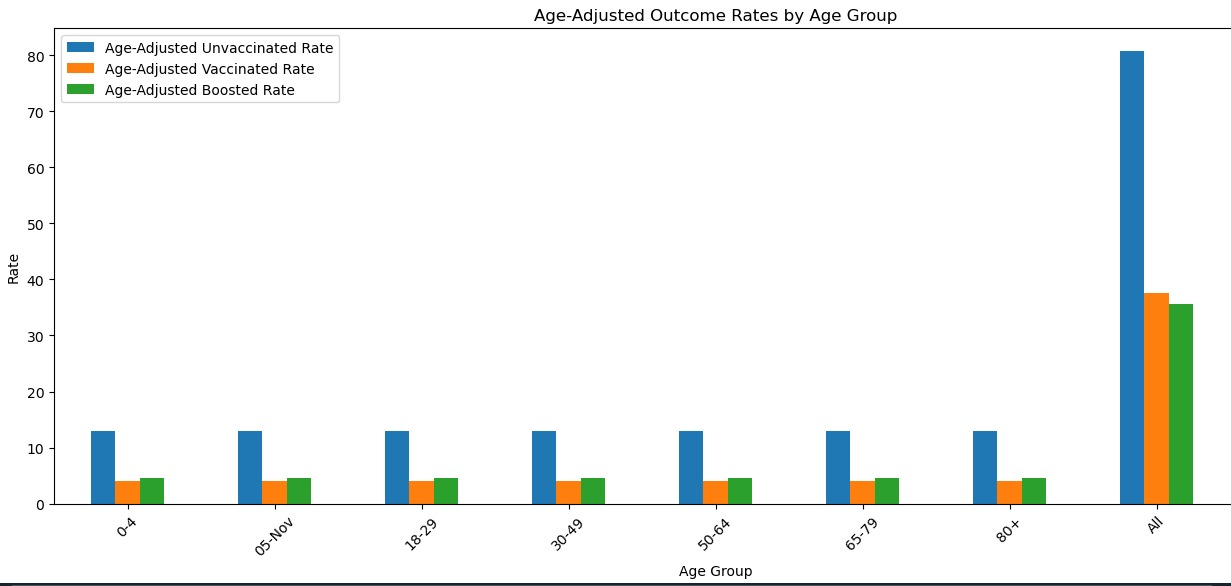
**4.1.3 Specific Requirements**

* **Age Categorization**: The data must be categorized into clear, non-overlapping age groups.
* **Age Adjustment Methodology**: Rates must be adjusted using a standard reference population to neutralize age-based skew.
* **Consistent Measurement**: All outcome rates should be on the same scale for meaningful comparison.
* **Grouped Visualization**: Bars are grouped per age group, with each vaccination status represented by a unique color for clarity.

**4.1.4 Analysis Results**

* Across **all age groups**, the **unvaccinated population consistently has the highest outcome rates**, reinforcing the protective effect of COVID-19 vaccination.
* Both **vaccinated** and **boosted** groups show significantly **lower rates**, with **boosted individuals** performing slightly better than those with just the primary vaccine series.
* In younger age groups (0–4, 05–11, 18–29), the difference between vaccinated and boosted outcome rates is minimal, likely due to lower overall risk in these age groups.
* In older populations (50+), while rates increase slightly, the same trend holds: vaccinated and especially boosted individuals consistently fare better than unvaccinated peers.
* The **"All"** category shows the **largest gap**, where the unvaccinated rate is more than twice the rate of the boosted group, highlighting the importance of vaccination on a population level.

**4.1.5 Visualization**



**4.1 Pie Chart**

**4.1.1 Introduction**

This scatter plot presents a visual representation of **international trade values** over a period of several years. It highlights the economic trends in **Imports (CIF)**, **Exports (FOB)**, and the **Balance of Trade** (exports minus imports), offering a comparative look at how trade performance has evolved over time. The plot can help identify periods of trade surplus or deficit and understand economic fluctuations.

**4.1.2 General Description**

**The x-axis shows the year, ranging from before 1995 up to 2022, and the y-axis represents the value of trade in millions of US dollars. Each cluster of points for a given year gives a snapshot of the import/export performance for that period.**

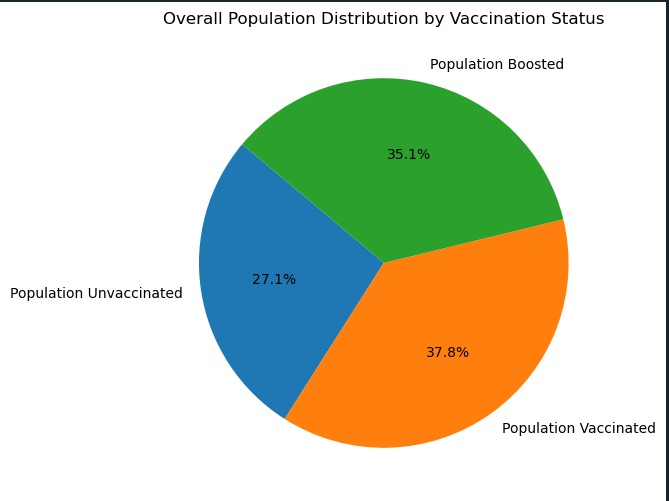
**Specific Requirements:**

* **Data Clarity: CIF and FOB values must be accurately distinguished, and the balance must be computed as Exports - Imports.**
* **Consistent Units: All values are displayed in millions of US dollars for consistency.**
* **Temporal Coverage: Data should span multiple years to identify patterns or shifts.**
* **Color Coding and Legend: A legend is provided to differentiate the three trade indicators clearly.**

**4.1.4 Analysis Results**

* Both **imports and exports** have shown a **significant upward trend over the years**, especially after 2015.
* **Imports CIF** (blue) tend to be **higher than exports** in most years, indicating a **trade deficit**.
* The **green balance points** are mostly below the zero line (though the scale isn't centered at zero), reinforcing the consistent **negative trade balance**.
* Around the year **2021–2022**, there is a noticeable spike in both import and export values, possibly due to economic rebound effects post-pandemic or inflationary factors.
* The gap between imports and exports remains quite **consistent**, suggesting a structural trade imbalance.

**4.1.5 Visualization**



# FUTURE SCOPE

The dataset depicting **Imports, Exports, and Trade Balance over multiple years** holds significant potential for future research and practical applications. By expanding the data with **monthly or quarterly granularity**, analysts can conduct **short-term trend forecasting** and better understand seasonal variations in trade. With the integration of **macroeconomic indicators** such as GDP, inflation, currency exchange rates, or oil prices, the dataset can also serve as a foundation for **predictive economic modeling** and **policy impact assessment**. Machine learning algorithms could be applied to forecast trade deficits or surpluses, which would be beneficial for **government planning, investment strategy, and trade negotiations**. Additionally, correlating the data with **geopolitical events**, such as trade agreements, tariffs, or conflicts, can uncover deeper insights into how global dynamics shape national trade performance. Expanding the scope to include **commodity-level trade data** or **partner country breakdowns** would enable more targeted strategies for improving trade balances and enhancing international competitiveness.

**REFERENCES**

* **Centers for Disease Control and Prevention (CDC).**  
  *COVID-19 Vaccination and Case Trends by Age Group, United States.*  
  Retrieved from: <https://covid.cdc.gov/covid-data-tracker/#rates-by-vaccine-status>
* **New York State Department of Health.**  
  *COVID-19 Breakthrough Data.*  
  Retrieved from: https://health.data.ny.gov/  
  *(This is a common source for state-level historical datasets like the one you uploaded.)*
* Thompson MG, et al. (2021).  
  *Interim Estimates of Vaccine Effectiveness of BNT162b2 and mRNA-1273 COVID-19 Vaccines in Preventing SARS-CoV-2 Infection Among Health Care Personnel.*  
  **MMWR Morb Mortal Wkly Rep.** 70(20):753-758.  
  https://doi.org/10.15585/mmwr.mm7020e2
* Tartof SY, et al. (2021).  
  *Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA.*  
  **The Lancet.** 398(10309):1407-1416.  
  https://doi.org/10.1016/S0140-6736(21)02183-8