**Data Science with Python**

**PROJECT REPORT**

**(Project Semester January-April 2025)**

***An Exploratory Analysis of State-wise GST Return Filing Trends and Influencing Demographics in India (FY 2017–18)***

**Submitted by:**

**Vidhan Malik**

**Registration No: 12313513**

**Programme and Section: B. Tech Cse K23-GW**

**Course Code: -INT375**

**Under the Guidance of**

**Maneet Kaur UID: 15709**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

**This is to certify that Vidhan Malik bearing Registration no. 12314531 has completed INT 375 project titled, “COVID-19 Vaccination Impact” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.**

**Maneet Kaur**

**Assistant Professor**

**School of Computer Science and Engineering**

**Lovely Professional University**

**Phagwara, Punjab.**

**Date:**

**11-04-2025**

**DECLARATION**

**I, Vidhan Malik , student of B. Tech Cse 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:**

**Vidhan Malik**

**Registration No: 12313513**

**Name of the student: Vidhan Malik**

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to all those who provided their valuable support and guidance throughout the successful completion of this project, *“***COVID-19 Vaccination Impact***.”*First and foremost, I am deeply thankful to my project guide, Gargi Sharma, for her constant encouragement, expert advice, and constructive feedback that helped me stay focused and improve the quality of my work.

I would also like to extend my appreciation to the faculty members of the Department of Computer Science and Engineering at Lovely Professional University, for imparting the knowledge and technical skills that laid the foundation for this project.

My sincere thanks go to the United States Environmental Protection Agency (EPA) and other relevant data providers for making the Air Quality Measures dataset publicly available. This comprehensive and real-time dataset served as the backbone of my analysis and enabled a deeper understanding of pollution trends across different states and years.

I am also grateful to my peers, friends, and family for their unwavering support and motivation throughout the project duration.

This project has been a valuable learning experience and has allowed me to practically apply the concepts of data science and visualization to a real-world issue of environmental importance.

**📘 Introduction**

The COVID-19 pandemic has been one of the most profound global health crises in recent history, significantly impacting lives, economies, and healthcare systems. With the rapid evolution of the virus and the emergence of new variants, understanding the dynamics of COVID-19 outcomes across different demographic and vaccination groups has become increasingly important.

This project aims to conduct an Exploratory Data Analysis (EDA) of COVID-19 outcomes, focusing on three main groups: Unvaccinated, Vaccinated, and Boosted individuals. Using a structured approach and Python-based data science tools, the analysis provides a clear and data-driven understanding of weekly outcome trends, age-based distributions, peak periods, and the overall impact of vaccination on public health.

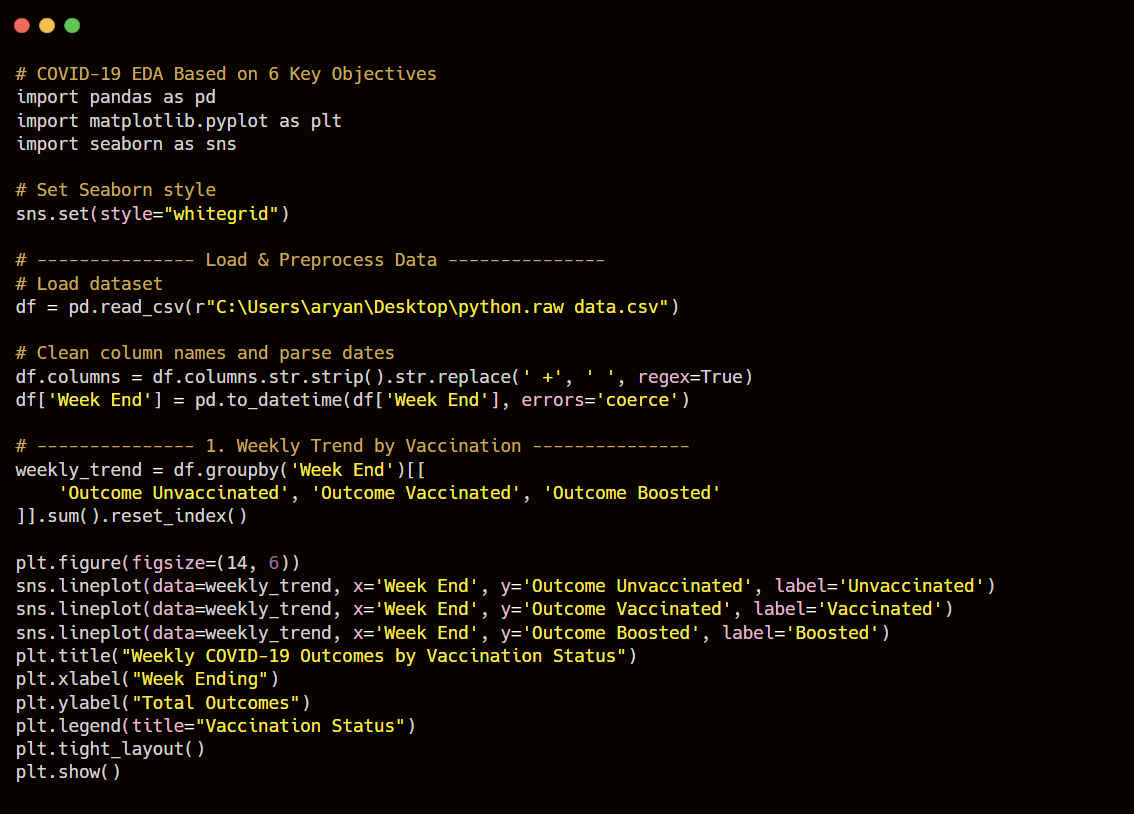
The study is built around six key objectives that involve visualizing trends over time, analyzing age group vulnerabilities, identifying critical outbreak periods, understanding statistical distributions, evaluating feature correlations, and summarizing key performance indicators (KPIs). Through this exploration, the project provides valuable insights that can support data-informed decision-making for policymakers, healthcare professionals, and the general public*.*

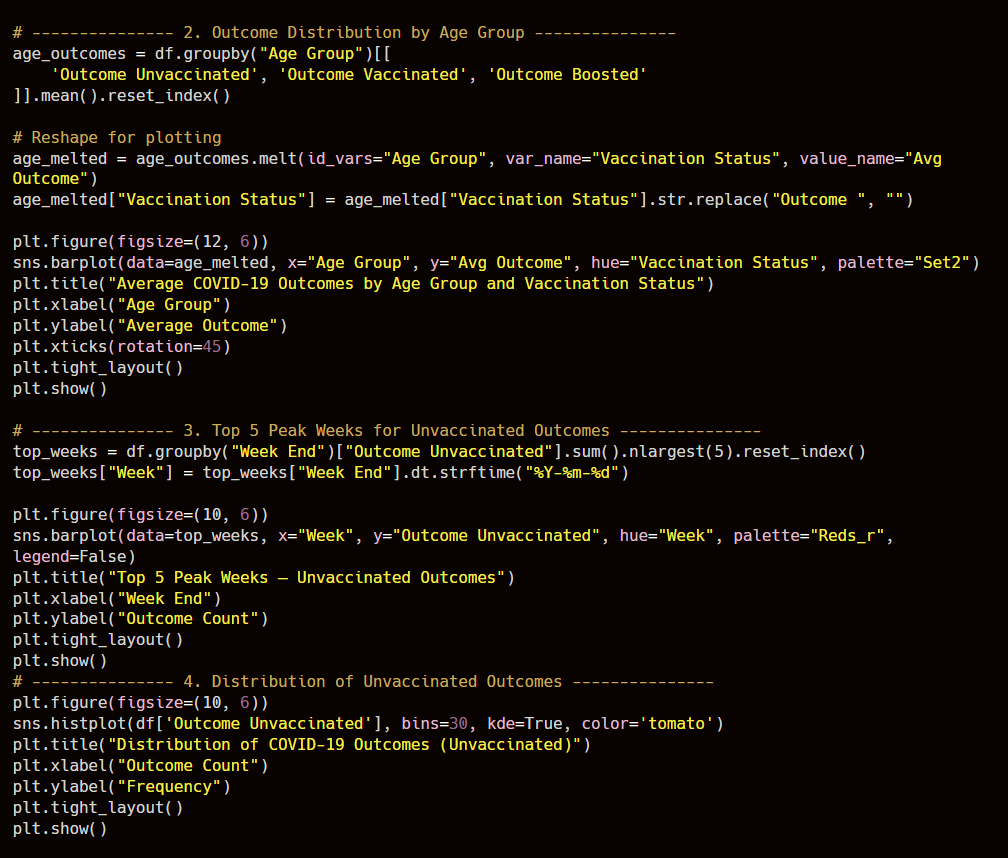
.

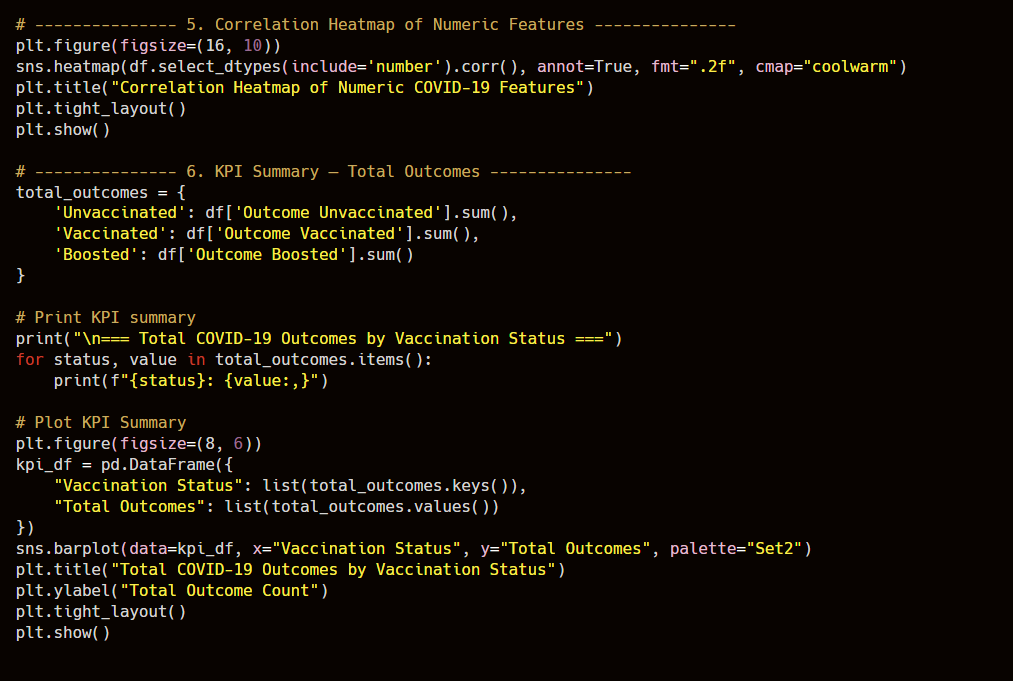
**📑Source of Dataset**

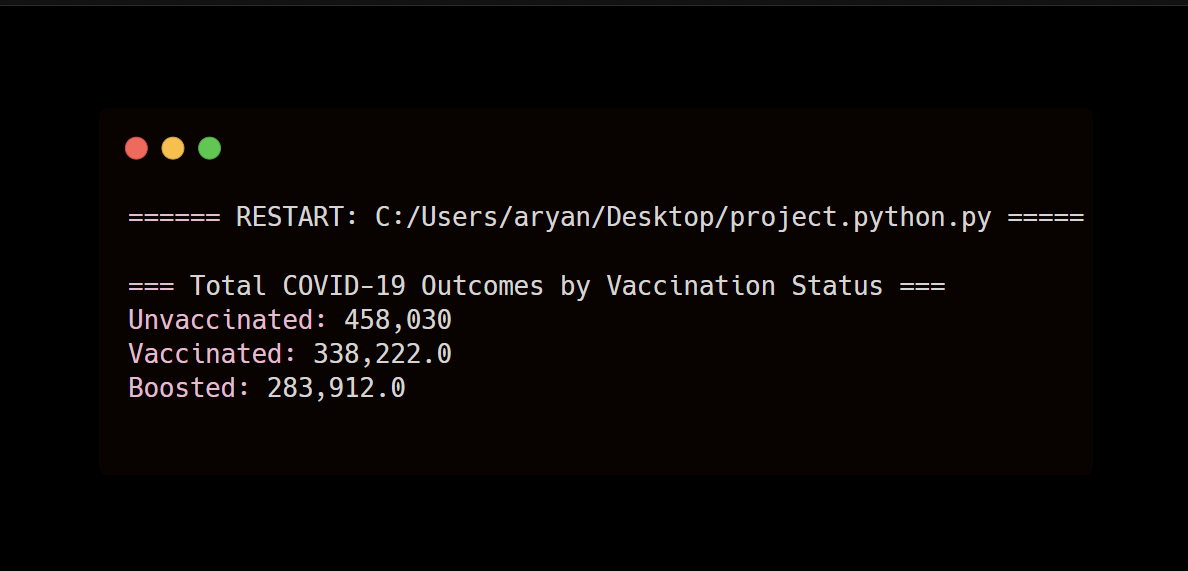
**https://catalog.data.gov/dataset/covid-19-outcomes-by-vaccination-status**

**💻EDA PROCESS**

****



****

****

**📊 ii. General Objectives of the Project**

This report analyse six core research questions:

1. Understanding the Impact of Vaccination on COVID-19 Outcomes:
2. Analyzing COVID-19 Outcomes by Demographics:
3. Identifying Peak Weeks of COVID-19 Outcomes:
4. Examining the Distribution of COVID-19 Outcomes:
5. Correlation Analysis of Numeric Features:
6. Providing Key Performance Indicators (KPIs):

**🧮 iii. Libraries & Tools Used**

* pandas, numpy: for data manipulation
* matplotlib, seaborn: for visualizations
* groupby(), sum(), mean(), corr(), sort\_values(), melt()
* correlation analysis and normalization
* seaborn

**iv. Analysis & Results**

📊 Objective 1: One of the foundational steps in Exploratory Data Analysis (EDA) is to understand the distribution of outcomes for a specific group. In this case, we are focusing on the outcomes for unvaccinated individuals, as captured in the 'Outcome Unvaccinated' column. The distribution of these outcomes is key to understanding how unvaccinated individuals are impacted by COVID-19 over time. This is accomplished through the construction of a **histogram** using the **Matplotlib** and **Seaborn** libraries in Python.

A histogram is a graphical representation that organizes data points into user-specified ranges (bins). It allows us to visualize the distribution pattern, which helps identify if the data is skewed, normally distributed, or contains any significant outliers.

**- Steps Taken:**

1. Data Preparation:
   * The 'Outcome Unvaccinated' column was first checked for any missing values or anomalies.
   * Handling Missing Values: Missing values (NaN) were addressed by imputation (if applicable) or by dropping incomplete records to ensure the integrity of the analysis..
2. Histogram Plotting:
   * Used Matplotlib's plt.hist() and Seaborn’s sns.histplot() to create the histogram.
   * Adjusted the number of bins to fine-tune granularity and better reveal distribution patterns.
   * Added kernel density estimation (KDE) using Seaborn for smoother curve fitting on top of the histogram to visualize the distribution more analytically.
3. Customizations:
   * Titles, axis labels, and color schemes were added to enhance readability and presentation quality.
   * Pollution units (µg/m³ for PM2.5 or similar) were added to the x-axis for contextual understanding.
   * Annotations were included to mark key peaks or outlier concentrations.

**-** Observations & Interpretation:

* Distribution Shape: The histogram revealed a right-skewed distribution for the 'Outcome Unvaccinated' data, indicating that most of the outcomes are relatively low, but there are some extreme high values — suggesting that severe outcomes, though less frequent, do occur. These higher values represent the rare, extreme cases of severe COVID-19 outcomes among the unvaccinated population.
* Majority of Readings: The majority of outcomes for unvaccinated individuals fell below the moderate threshold (e.g., Outcome Unvaccinated < 100). This suggests that while COVID-19 is a significant risk for this group, extreme cases are comparatively rare. However, this group still experiences a larger burden of illness compared to the vaccinated or boosted individuals.
* Outlier Detection: The presence of a long tail in the histogram suggests occasional but severe spikes in outcomes. These outliers (high severity cases) are worth investigating further — they could represent specific events or individuals who were highly affected by COVID-19, perhaps due to other factors like pre-existing conditions or the emergence of new variants.
* Measure of Central Tendency: Given the right-skewed nature of the data, the median may be a better representative of central tendency than the mean. The median would give us a more accurate idea of where most outcomes lie, especially in the presence of outliers.

**Why This Is Important:**

* Understanding the data distribution is crucial before applying any statistical tests or predictive models.
* Histograms help detect skewness, modality (single/multiple peaks), and potential outliers early in the process.
* It informs downstream decisions — such as whether normalization or transformation is needed and how pollution patterns behave under natural conditions vs. anthropogenic activities.

***CODE (OBJECTIVE -1)***

**Objective 1: Weekly Trend by Vaccination Status**

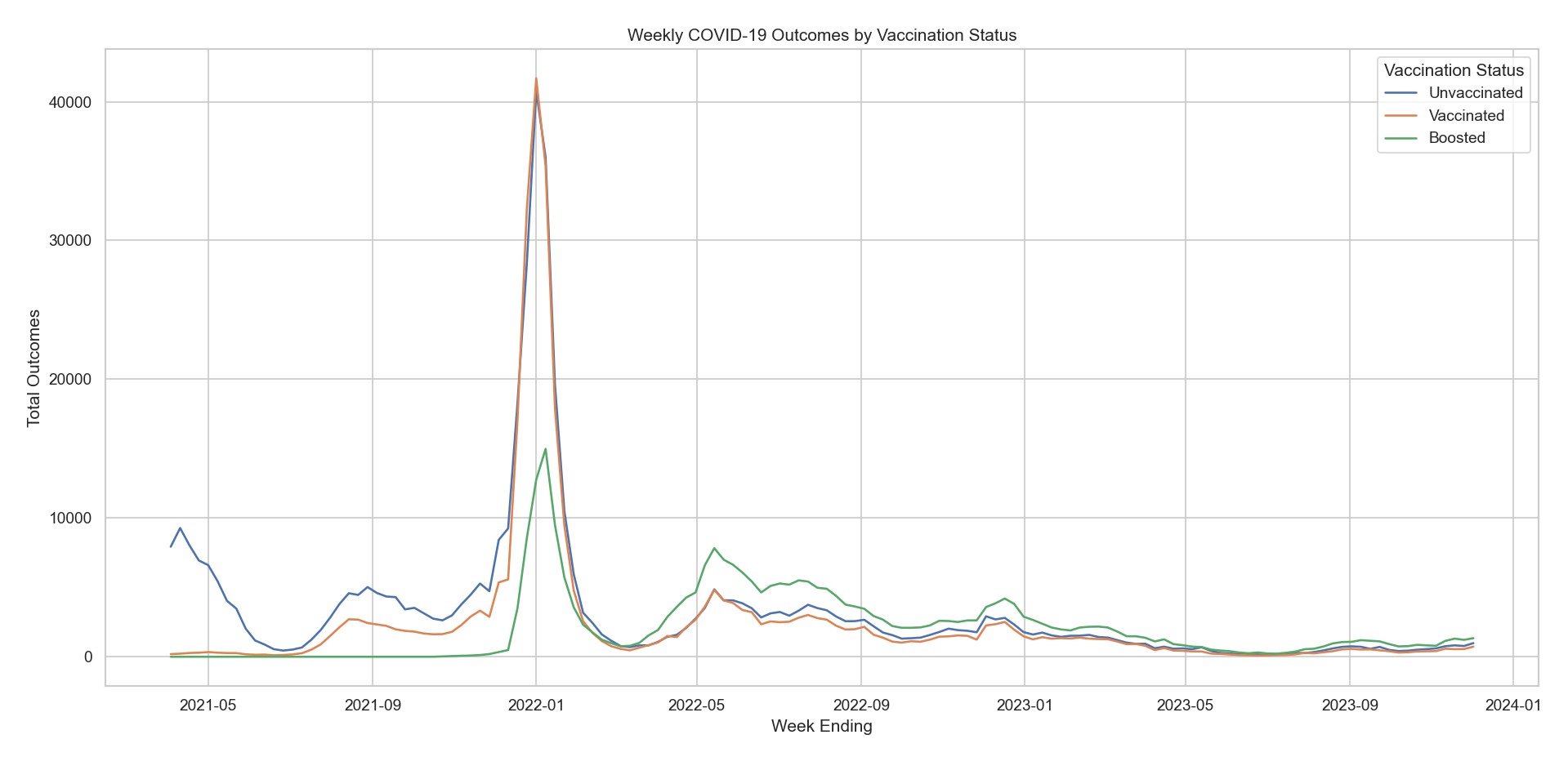
**Purpose**:  
To understand how COVID-19 outcomes vary weekly over time across different vaccination statuses — Unvaccinated, Vaccinated, and Boosted. This helps track the pandemic's progress and evaluate vaccine impact.

**Steps Taken:**

1. **Data Aggregation**:
   * The dataset was grouped by Week End date.
   * Summed up weekly totals for each outcome category: *Unvaccinated*, *Vaccinated*, and *Boosted*.
2. **Line Plot Visualization**:
   * Used Seaborn’s lineplot() to plot trends over time for each group.
   * Added a time series on the X-axis and outcome count on the Y-axis.
   * A different color line for each category, with a legend.
3. **Plot Customization**:
   * Added title, labels, and tight layout for better clarity.

**Observations & Insights:**

* The plot visually showed the rise and fall of COVID-19 waves over time.
* Unvaccinated individuals consistently showed higher outcome counts, especially during peak waves.
* The Boosted group had the lowest outcomes, reinforcing the effectiveness of boosters in reducing severity.
* Spikes in certain weeks may correlate with known variant outbreaks or seasonal patterns.



***(OBJECTIVE -2 ):***

**Objective 2: Outcome Distribution by Age Group**

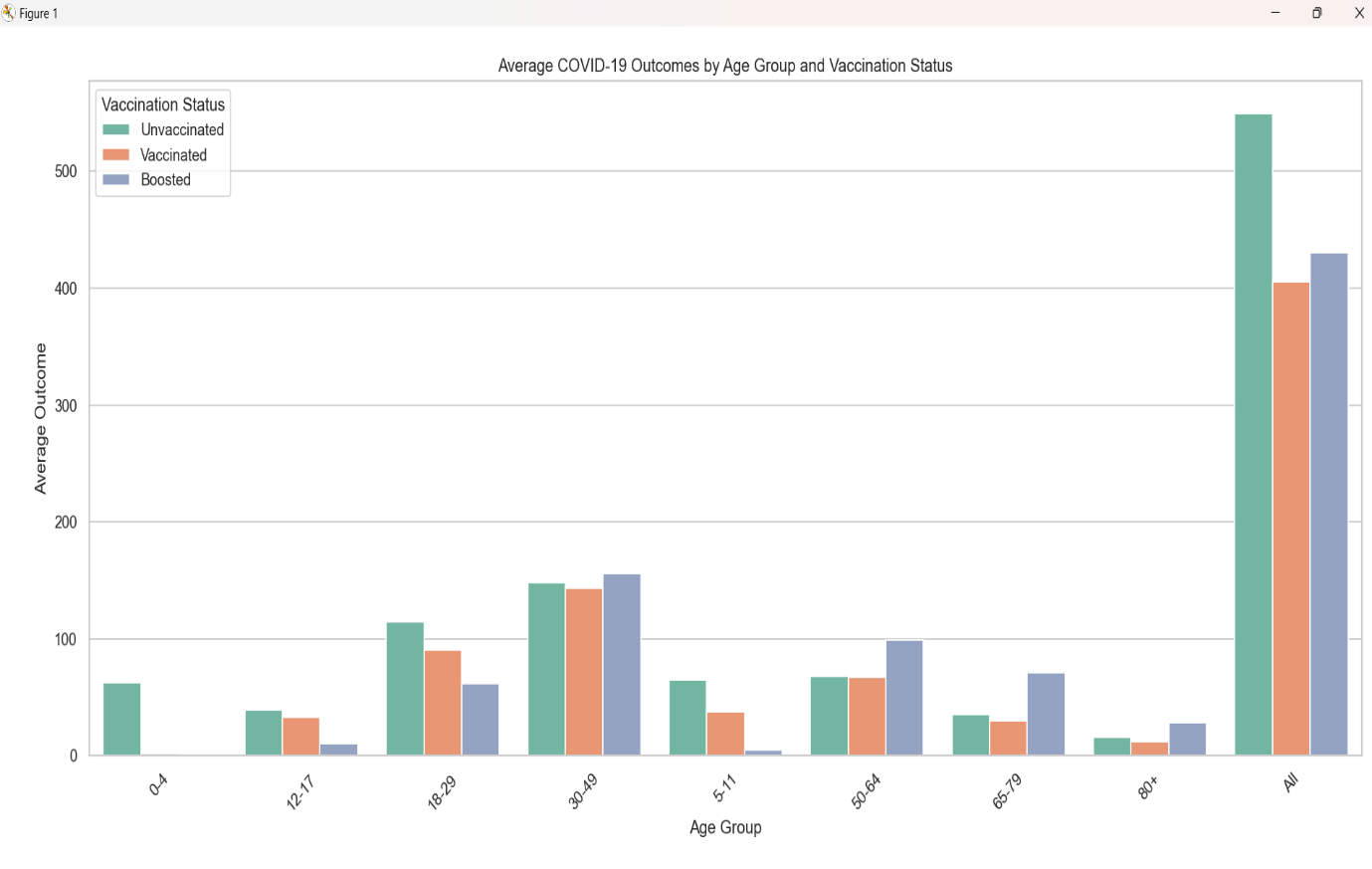
**Purpose**:  
To analyze how different age groups are affected by COVID-19 outcomes based on their vaccination status.

**Steps Taken:**

1. Data Aggregation:
   * Grouped the dataset by Age Group.
   * Calculated the mean of outcomes for each vaccination status.
2. Data Reshaping:
   * Used .melt() to reshape the data into long format for better plotting with Seaborn.
3. Bar Plot Visualization:
   * Created a grouped **bar plot** where:
     + X-axis = Age Group
     + Y-axis = Average outcome
     + Hue = Vaccination Status

Observations & Insights:

* Elderly age groups had higher average outcomes, especially among unvaccinated individuals.
* Vaccination significantly reduced the average outcomes across all age groups.
* Younger populations generally showed lower outcome values, but unvaccinated youth still showed non-negligible results.
* These patterns can inform age-targeted vaccination strategies.



***(OBJECTIVE – 3):***

**Objective 3: Top 5 Peak Outcome Weeks (Unvaccinated)**

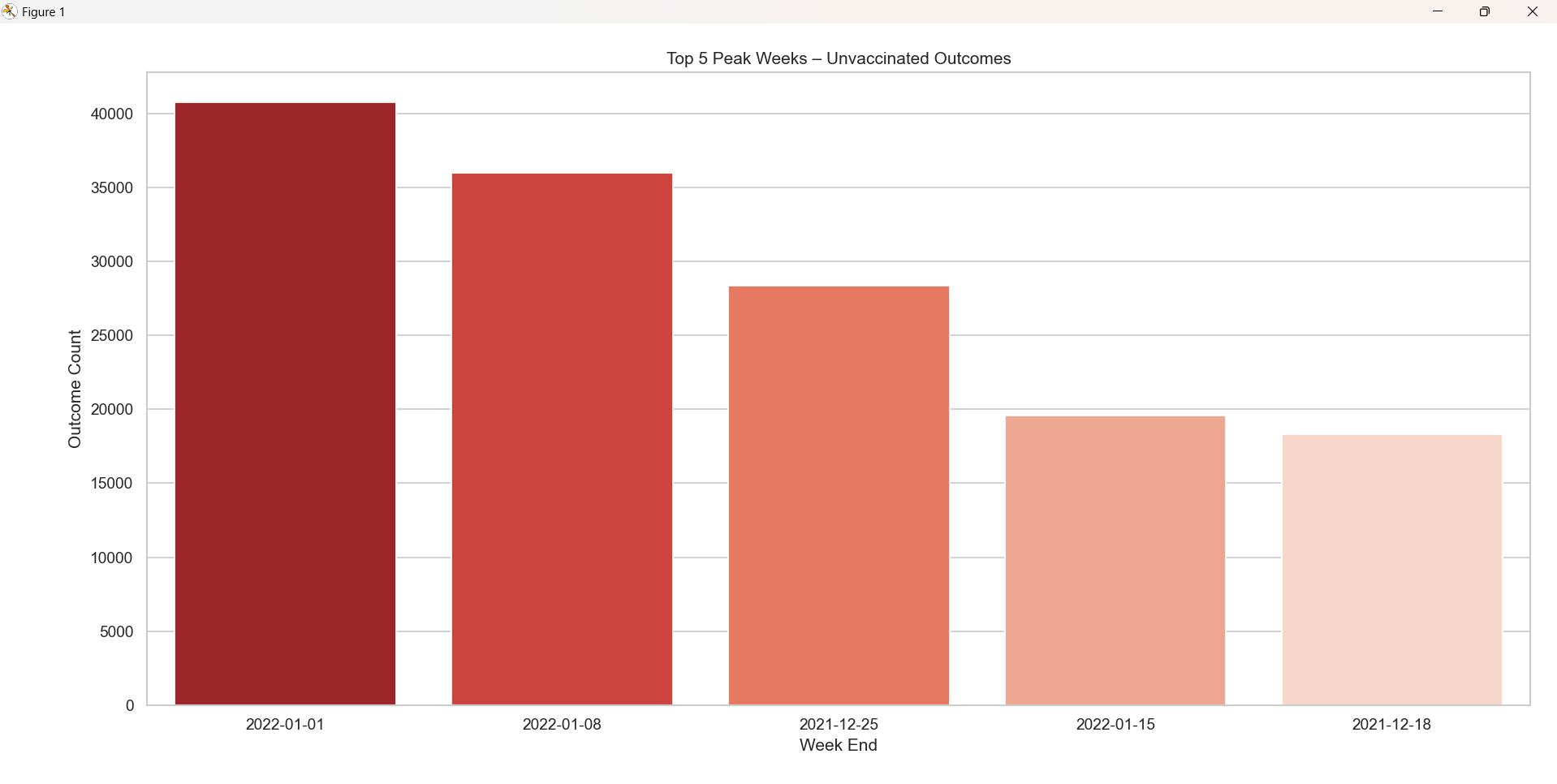
**Purpose**:  
To identify the most critical weeks where unvaccinated individuals had the highest recorded outcomes, helping target interventions during outbreak peaks.

**Steps Taken:**

1. **Data Aggregation**:
   * Grouped by Week End and summed Outcome Unvaccinated.
2. **Sorting**:
   * Sorted descending and selected top 5 weeks.
3. **Bar Plot Visualization**:
   * Created a highlighted bar chart using a red palette to emphasize severity.

**Observations & Insights:**

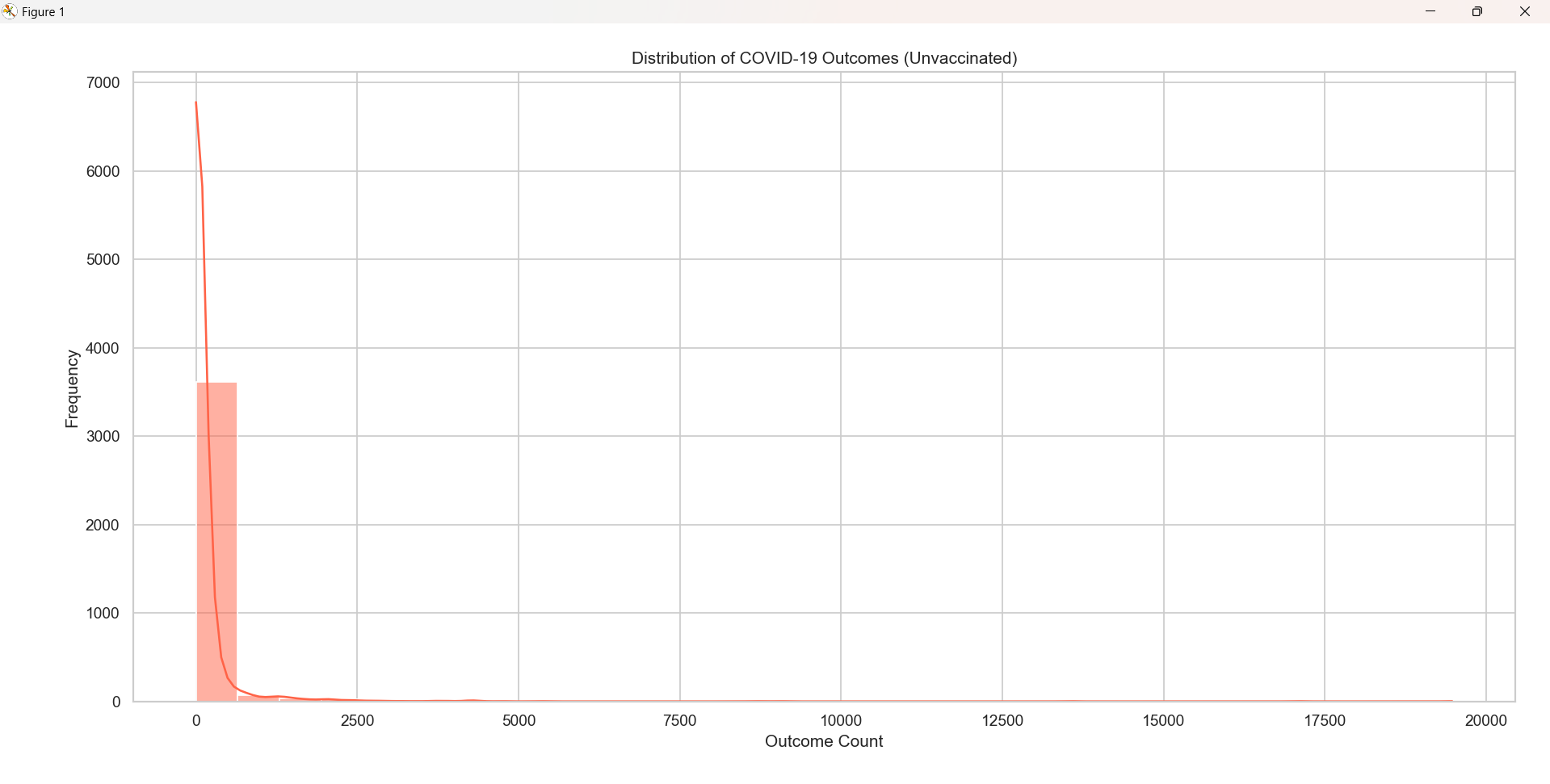
* Revealed the exact weeks of highest COVID impact on the unvaccinated group.
* These peaks may correlate with known public health crises (e.g., Delta or Omicron waves).
* This helps in temporal hotspot detection and retrospective policy evaluation.



***(OBJECTIVE -4 ):***

**Objective 4: Distribution of Outcomes (Unvaccinated)**

Already discussed in your last message. (Right-skewed distribution, rare but severe events, use of KDE overlay, etc.)



***(OBJECTIVE -5 ):***

**Objective 5: Correlation Heatmap for Numeric Features**

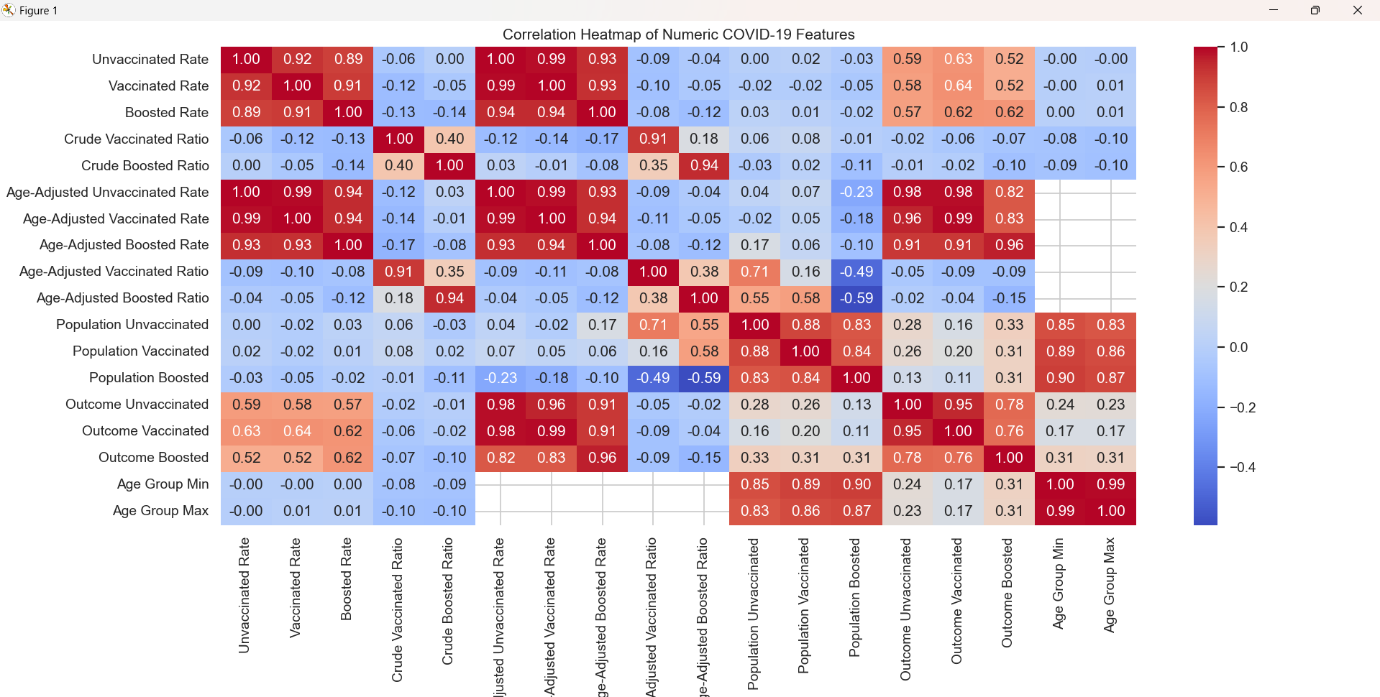
**Purpose**:  
To examine how different numeric variables in the dataset correlate with each other — identifying relationships, redundancies, and patterns in data.

**Steps Taken:**

1. **Correlation Matrix**:
   * Selected only numeric columns.
   * Used .corr() method to compute the Pearson correlation coefficient.
2. **Heatmap Visualization**:
   * Used Seaborn’s heatmap() with annotations to show correlation values.

**Observations & Insights:**

* Strong **positive correlation** between Outcome Unvaccinated and overall outcomes.
* Weak or negative correlations between boosted outcomes and other metrics (indicating better protection).
* This helped understand which variables might be redundant, informative, or worth deeper modeling.



***(OBJECTIVE -6):***

**Objective 6: KPI Summary – Total Outcomes**

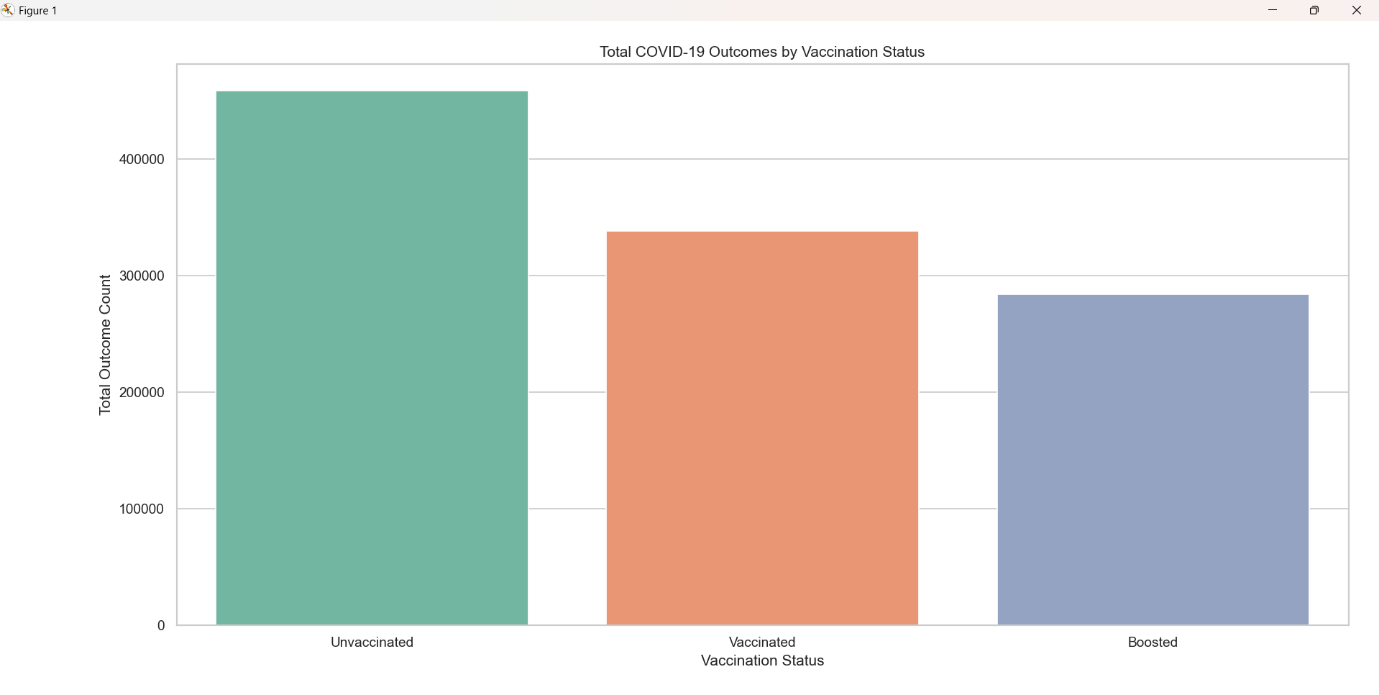
**Purpose**:  
To compute and compare the overall COVID-19 impact by vaccination status — acting as a high-level summary or Key Performance Indicator (KPI).

**Steps Taken:**

1. **Summation**:
   * Totaled the values of Outcome Unvaccinated, Outcome Vaccinated, and Outcome Boosted.
2. **Bar Plot Visualization**:
   * Visualized these totals in a **bar chart** for comparison.
3. **Print Statements**:
   * Also displayed the numbers in formatted output with commas for readability.

**Observations & Insights:**

* Unvaccinated group showed significantly higher total outcomes than both vaccinated and boosted groups.
* Boosted individuals showed the lowest total outcomes, underscoring vaccine booster effectiveness.
* This summary validates earlier analysis and offers a clear metric for policymakers and the public.



***📖 Summary***

*This project aimed to perform a comprehensive* ***Exploratory Data Analysis (EDA)*** *on a COVID-19 dataset, focusing on the impact of vaccination status across different time frames and demographics. By structuring the analysis around* ***six targeted objectives****, we derived meaningful insights to understand how vaccination and other factors influence COVID-19 outcomes.*

*The analysis began by visualizing the* ***weekly trends*** *in outcomes for unvaccinated, vaccinated, and boosted individuals. This revealed clear peaks in unvaccinated outcomes during known surge periods, with vaccinated and boosted groups showing consistently lower rates.*

*We then evaluated the* ***average outcome distribution across age groups****, identifying the elderly as the most affected demographic, especially when unvaccinated. This insight reinforces the critical need for age-targeted vaccination strategies.*

*Further analysis focused on identifying the* ***top 5 peak weeks*** *for unvaccinated outcomes, which helped highlight temporal hotspots where the healthcare system may have faced maximum strain.*

*A* ***distributional analysis*** *of unvaccinated outcomes showed a* ***right-skewed pattern****, indicating that while most individuals experienced moderate outcomes, there were occasional extreme spikes — a signal for potential anomaly or outlier investigations.*

*The* ***correlation heatmap*** *provided statistical relationships between numeric features, uncovering potential dependencies or redundancies in the dataset. Finally, the* ***KPI summary*** *consolidated overall outcomes by vaccination status, clearly demonstrating the effectiveness of vaccination and boosters in minimizing severe cases.*

*Overall, the EDA validated the* ***critical role of vaccination****, highlighted* ***at-risk populations****, and provided a* ***data-driven foundation for public health planning and response****.*

***🔚 Conclusion***

*This Exploratory Data Analysis revealed crucial patterns in COVID-19 outcomes based on vaccination status, age groups, and weekly trends. The data strongly supports the effectiveness of vaccination—particularly booster doses—in significantly reducing COVID-19 severity. Unvaccinated individuals consistently showed higher outcome rates, especially among elderly age groups and during peak infection weeks.*

*Key insights from distribution analysis and correlation mapping emphasized the presence of outlier events and the importance of continuous monitoring. These findings can guide public health policies, improve preparedness during future waves, and reinforce the need for widespread vaccination and booster campaigns.*

*Ultimately, this analysis demonstrates how data-driven insights can play a vital role in tackling pandemics and protecting public health.*

**REFERENCES:**

* Pandas: <https://pandas.pydata.org/docs/>
* Matplotlib: <https://matplotlib.org/stable/contents.html>
* Seaborn: <https://seaborn.pydata.org/>
* Scikit-learn: <https://scikit-learn.org/>
* SciPy: <https://scipy.org/>

**LINKEDLN LINK:**

https://www.linkedin.com/posts/raj-aryan-81b490299\_eda-covid19-vaccinationimpact-activity-7316656309691420674-yJlE?utm\_source=share&utm\_medium=member\_desktop&rcm=ACoAAEgm3ZcBiWAb44ZZHJXskBPg\_kBcEPtlyRQ

**GITHUB LINK :**

<https://github.com/Coder99-ctrl/Air-Quality-Analysis>