**VACCINATION SURVEY ANALYSIS**

1. AIM OF THE PROJECT:

This project is focused on analyzing immunization coverage data from the World Health Organization(WHO) and United Nations International Children's Emergency Fund(UNICEF). Immunization is one of the most effective public health interventions to prevent a range of infectious diseases, and improve global health outcomes. The primary aim is identifying trends, patterns through data analysis and visualization. The primary objectives of this project are to:

* Clean and preprocess the data to ensure accuracy and usability for analysis.
* Perform exploratory data analysis (EDA) to identify important trends and patterns that may influence global health outcomes.
* Visualize the data in accessible formats to help stakeholders better understand the key health indicators and their implications.
* Provide insights and recommendations based on the findings that can inform future policies and initiatives aimed at improving health conditions.

This data will contribute to a deeper understanding of immunization trends and to identify key factors affecting immunization coverage and provide access to life-saving vaccines.

2. BUSINESS PROBLEM OR PROBLEM STATEMENT :

The dataset consists of millions of rows (lakhs of data points) that are highly uncleaned and contain several issues such as missing values, inconsistent formats, duplicate records and irrelevant data. The goal of this analysis is to assess the effectiveness, accessibility. The variables such as vaccination rates, age groups, sex, and geographic areas, needs to be thoroughly analyzed to uncover patterns, trends, and potential gaps in vaccination coverage.

**Vaccination Coverage Analysis:**

* Assess the overall vaccination coverage by comparing the number of individuals vaccinated against the target population (denominator).
* Evaluate the effectiveness of vaccination campaigns across different regions and cohorts

**Demographic Breakdown:**

* Analyze vaccination rates by age group (ageVaccination, ageInterview, sex), and other demographic factors to uncover any disparities in access to vaccines.
* Determine if certain age groups or gender are more likely to receive vaccinations and investigate any gaps in vaccination coverage.

**Geographic and Regional Insights:**

* Assess the vaccination coverage in different geographical areas (coverage\_area) to identify regions with low vaccination uptake.
* Compare vaccination outcomes between countries or regions, identifying which areas require more resource allocation.

**Survey Data Validity:**

* Evaluate the quality and validity of the collected data by assessing variables such as "validity" and "cardsSeen," ensuring that survey results are accurate and trustworthy.
* Analyze how vaccination rates differ by cohortYear, helping to track improvements in vaccination coverage over time.

In addition, the project will evaluate the quality and validity of survey data to ensure that the results are reliable and actionable this analysis will not only contribute to improving future vaccination efforts but also help WHO, UNICEF, and other organizations optimize their strategies to ensure vaccines are delivered equitably and efficiently to all populations.

3. PROJECT DESCRIPTION:

The dataset includes key metrics related to vaccine coverage such as age, gender, geographic location, cohort year and the types of vaccines administered. By leveraging this data, the project seeks to derive insights into the effectiveness and reach of vaccination programs, identify gaps in coverage, and help improve future vaccine distribution strategies.

* Data Preprocessing and Cleaning
* Data Quality Assessment
* Vaccine coverage and effectiveness (EDA)
* Trend and Pattern Analysis
* Statistical Analysis

The main objective is to identify which demographic group especially by gender are receiving higher vaccination coverage based on evidence such as "cardsSeen" (i.e., the number of vaccination cards verified during the survey).

The **bar graph** for data completion will allow us to identify which organizations are consistently providing high-quality, complete data. Organizations with incomplete data may need targeted support or guidance to improve their survey processes and data collection methods.

To check whether the participants have provided the necessary information for the survey by visualizing the distribution of missing or incomplete data. Using a histogram plot (histplot), it will show the number of respondents who have incomplete responses across key fields like age, sex, vaccine type, and other essential survey data (e.g., "cardsSeen," "validity"). This will help to identify whether certain information is consistently missing, which could signal issues with survey quality or respondent compliance.

We will compare the data completeness across different organizations using a bar plot to identify which organization has the most complete and accurate data. By grouping the organizations (likely identified by the "surveyNameEnglish" or "ISO3" column), we will visualize how many records have complete data per organization, enabling us to assess the overall quality of survey data from different regions or organizations.

The **box plot** is used to identify outliers in vaccination data(e.g., "cohortYear").It helps us visually detect any extreme values such as unusually high or low vaccination rates in specific cohort groups. Outliers might indicate that some survey results are skewed, which could affect the reliability of the analysis. It highlight any unusual vaccination coverage patterns for specific cohort years such as cases where certain cohorts have exceptionally high or low vaccination rates. Investigating these outliers will help us determine whether they are caused by genuine factors (e.g., a successful vaccination campaign) or if they point to data quality issues (e.g., errors or misreporting).

Using a **pie chart**, we will analyze which vaccine is most commonly administered based on the vaccine column in the dataset. This will help us to understand vaccine distribution patterns, highlighting the most popular or widely distributed vaccines in different regions and cohorts. Understanding which vaccines are most prevalent can help WHO, UNICEF, and other stakeholders allocate resources and plan vaccination strategies more effectively.

A **line chart** is used to track the **data completion pattern** across time or other demographic variables, such as cohort year or age group. By visualizing the percentage of completed survey records over time or by group, we can understand whether certain periods, regions or demographic groups have more complete or reliable data. This will helps to identify trends in how people engage with the survey and highlight any areas where data completion is suboptimal.

The **scatterplot** is used to examine the **density of vaccine coverage across different geographical areas.** This visualization explore the relationship between vaccination coverage and geographic factors such as region or coverage area. It analyze whether some areas show a higher concentration of vaccinated individuals, potentially indicating a higher density of healthcare services or more successful vaccination campaigns. The scatterplot will highlight areas with low vaccine coverage, which could inform future intervention strategies.

Based on the visualization and statistical findings, this project offer recommendations on improving gender equity in vaccination programs. For example, if females are significantly less vaccinated than males, specific outreach efforts, such as awareness campaigns or providing female-focused healthcare services could be proposed. It should includes the findings from the bar plot, additional visualizations, statistical tests, and a detailed breakdown of vaccination coverage by gender, age, region, and other demographic factors.

4. FUNCTIONALITIES:

The key functionalities of this vaccination survey data focus on cleaning, exploring, and deriving insights to assess vaccine coverage and effectiveness. Through various techniques like data preprocessing, exploratory analysis, cohort tracking, and statistical testing, the project aims to uncover patterns and inform targeted public health interventions.

### Data Preprocessing and Cleaning****:****

The Data Preprocessing and Cleaning functionality ensures that raw data is cleaned, standardized and made ready for analysis. It involves addressing common issues like missing data, duplicates, inconsistent data formats and outliers. This is the foundational step to ensure the data's integrity and reliability before any deeper exploration or statistical analysis is performed. The preprocessing includes handling standardizing formats for categorical and date variables and dealing with outliers that could distort analysis.

* Standardization of data formats for categorical columns like vaccine, age, and sex.
* Transformation of data types (e.g., converting dates into a consistent format).
* Filling Missing Values in Object-Type Columns with Mode and replacing Values Based on Case-Insensitive Matching
* Removing Non-Breaking Spaces from the 'ageVaccination' Column
* Reducing text redundancy and ensures consistency in the dataset.
* Data Quality Assessment

It evaluates the integrity of the dataset by examining the completeness, accuracy, and consistency of the data. This ensures that the analysis is based on valid, reliable information. Data quality checks assess the frequency of missing values, accuracy of categorical data and consistency of numerical columns. It helps to detect potential errors, such as out-of-range values, invalid categories or illogical relationships which are crucial to address before moving forward with any further analysis.

* Checking for incomplete data in key variables like ageVaccination, cardsSeen, and coverage\_area.
* Analyzing the distribution of missing data to identify patterns.
* Validating the consistency of values (e.g., ensuring **sex** is coded as "Male" and "Female" and not misspelled).
* Identification and treatment of outliers (e.g., extremely high or low values in cardsSeen or ageVaccination).
* Vaccine Coverage and Effectiveness:

It uses Exploratory Data Analysis (EDA) techniques to understand vaccine coverage and effectiveness across different populations and regions. By visualizing key variables, such as the number of cardsSeen (representing the number of people vaccinated) and denominator (the total population eligible for vaccination), we can calculate and compare vaccination rates. EDA also uncovers trends and patterns, if certain demographics are underserved. It helps to identify gaps in vaccine distribution and highlights areas where coverage needs to be improved.

Purpose of the Visualization:

* Bar plot visualizes which organization has the highest completion of data by comparing the number of 'cardsSeen' for each 'ISO3' organization code. The visualization allows for a quick and clear understanding of how data has been processed across different organizations.
* Pie chart visualizes the proportion of different vaccine categories in the dataset. Each slice represents a vaccine type, and the percentage labels on each slice show the proportion of that category relative to the whole. The colors and labels make it easy to compare the contributions of each category to the total dataset. This visualization helps to quickly grasp the distribution of vaccines and see which category has the largest or smallest share.
* Line plot provides a visual representation of the cardsSeen column, sampled to show how data completion progresses. By displaying the sampled values as a connected line with markers, the plot makes it easy to identify trends or patterns in the data over time (or index order). It highlights how the accomplishment of data varies and helps to visualize any noticeable peaks, dips, or steady trends in the completion process.
* Boxplot compares the distribution of the validity variable across different categories defined by CohortYear. Outliers, which are represented as individual points outside the whiskers, indicating extreme values in the dataset. By looking at the boxplot, you can quickly identify how the validity distribution differs across various CohortYear categories, and spot any outliers or unusual values. This visualization is particularly useful for understanding variations within each cohort and detecting anomalies in the data.
* Trend and Pattern Analysis:

This analysis helps to understand temporal changes in vaccination coverage and effectiveness. By comparing vaccination rates for different cohortYears, we can detect patterns such as whether younger or older populations are more likely to receive vaccinations. Trend analysis can be used to forecast future vaccination trends.

Purpose:

* Violin plot provides a comprehensive view of the distribution of the CardsSeen variable across the sex categories. It consists of shape that shows the density of the data at different values. Wider sections of the violin represent more frequent values, while narrower sections indicate less frequent values.The thick black bar in the middle of the violin represents the interquartile range, which contains the middle 50% of the data.The violin plot not only reveals the central tendency and spread of the data but also provides insights into the distribution's shape, which is particularly useful for comparing distributions between categories.
* A stacked bar plot is ideal for visualizing how different genders contribute to the overall count of valid and invalid data. The stacked bars represent the total count of valid data, but they are divided into different segments to show the gender distribution. The colored segments within each bar represent the distribution of different genders (e.g., male, female, both) within each validity category. By comparing the height of the bars for different validity statuses, we can see how the total counts of valid and invalid data differ across genders. The relative size of the colored segments allows us to identify which gender has a larger or smaller proportion of valid data within each validity category.
* The line plot provides a clear visualization of how the vaccination coverage differs across age groups and genders and allowing for easy comparison of vaccination coverage by gender across age group. This plot is particularly useful for understanding trends in vaccination coverage and identifying differences between genders in different age groups. It helps to visualize which gender and age group combinations are receiving more or less coverage, and can be used to spot any disparities or trends in the vaccination process.
* The heatmap serves as an efficient way to visualize the relationships between multiple numerical variables. Variables that increase or decrease together (values closer to +1).Variables that do not show a clear relationship (values closer to 0).Variables that move in opposite directions (values closer to -1), like the negative correlation between CardsSeen and denominator. It have a negative correlation between CardsSeen and denominator, this could imply that higher values of CardsSeen are associated with lower values of denominator. Depending on the context of your data, this relationship might need further exploration to understand its cause or significance.
* The bar plot helps visualize the average vaccination coverage area for each organization, based on the ISO3 codes. This plot is particularly useful for comparing the coverage area across different organizations and quickly identifying which organization covers the largest or smallest area.The length of the bars allows us to compare the coverage area of different organizations.The ISO3 codes on the y-axis make it clear which organization corresponds to each bar.The plot is styled with gridlines for easier reading, and labels are added with appropriate font sizes for clarity.
* Statistical Analysis

This functionalityadds rigor to the insights derived from the data. It involves applying various statistical techniques to test hypotheses, explore relationships between different variables and derive actionable insights. Common tasks include running descriptive statistics (e.g., mean, median, standard deviation), performing hypothesis testing (e.g., t-tests, chi-square tests, anova), and building regression models to understand how different factors (e.g., age, sex, vaccine type, coverage\_area) affect vaccination coverage. Statistical analysis will validate the patterns and trends observed in the previous functionalities and provide evidence for decision-making.

* Calculating descriptive statistics for key variables like ageVaccination, cardsSeen, and denominator.
* Running hypothesis tests (e.g., comparing vaccination coverage between different genders or regions using t-tests or chi-square tests, anova).
* Performing regression analysis to understand the impact of demographic and geographic variables on vaccine coverage.
* Deriving confidence intervals, p-values and z-scores to assess the significance of observed patterns in the data.

5. RESULTS AND OUTCOMES:

The vaccination survey analysis project has yielded significant insights and outcomes, highlighting critical trends, gaps, and opportunities for improving vaccine coverage and public health strategies. Through comprehensive data preprocessing and cleaning, the project ensured that the dataset was free from inconsistencies and errors, setting a solid foundation for further analysis. The gathered insights are,

* Identifies which age groups and genders have higher or lower coverage areas.
* Organizations which have higher data completion rates and improvement. It reveals any gender-based disparities or patterns in data completion.
* It helps identify patterns or anomalies in how individual data is completed over time.

Moreover, a closer look at demographic variables, such as gender and age, uncovered disparities in vaccination coverage. For example, while male participants had slightly higher vaccination rates than females in some regions. Gender-specific outreach strategies were also recommended to address the observed disparities in vaccine uptake. Based on the analysis, it is recommended that vaccination efforts focus on improving access in rural or underserved regions, addressing gender disparities, and enhancing data collection mechanisms to ensure more accurate assessments in future surveys.

Overall, the project has provided actionable insights that can inform more effective vaccination strategies, optimize resource allocation, and support equitable vaccine distribution globally. These results are essential for guiding future public health initiatives and ensuring that vaccination efforts reach all populations, regardless of geographic location, gender, or age.

6. CONCLUSION:

This analysis emphasized the importance of understanding the **temporal trends** in vaccine uptake, showing that vaccination coverage can fluctuate based on factors such as seasonal public health campaigns, vaccine availability, and awareness programs. It reinforces the idea that improving healthcare infrastructure and increasing access to vaccines, particularly in remote or resource-poor areas, is essential for achieving universal vaccine coverage.

The vaccination data analysis serves as a valuable tool for guiding public health interventions, identifying gaps in vaccine distribution, and informing future strategies to enhance immunization coverage. By leveraging these insights, stakeholders can better allocate resources, optimize vaccination efforts, and ultimately improve public health outcomes. The project not only contributes to a deeper understanding of vaccine coverage in the region but also provides a framework for improving future surveys and enhancing the overall effectiveness of vaccination programs.