Vaccination Data Analysis and Visualization

Project Overview

This project aims to analyze global vaccination data to uncover trends in vaccination coverage, disease incidence, and vaccination effectiveness. The project uses a robust data cleaning pipeline, stores clean data in a structured SQL database, and showcases insights via Power BI interactive dashboards. The analysis supports public health strategies, disease prevention initiatives, resource allocation, and global health policy formulation.

Skills Acquired

- Python scripting for data cleaning and transformation
- Exploratory Data Analysis (EDA) to identify key patterns
- Data modeling and normalization in SQL
- Building interactive dashboards in Power BI
- Domain knowledge in Public Health and Epidemiology

Problem Statement

Global vaccination efforts require detailed quantitative analysis to evaluate their success and identify gaps in coverage that impact disease control. By analyzing a combination of vaccination coverage, disease incidence, and vaccine introduction data, the project provides actionable insights for stakeholders to improve vaccination programs, strategize resource deployment, and influence policy-making.

Business Use Cases

Use Case Category	Description
Public Health Strategy	Assess program effectiveness, pinpoint low vaccination areas for intervention.
Disease Prevention	Detect diseases with high incidence despite vaccination, guide new vaccine policies.
Resource Allocation	Identify regions needing resource prioritization, forecast vaccine demand for supply management.
Global Health Policy	Offer evidence-based recommendations for vaccine introduction and boosters, aid international health efforts.

Data Sources and Variables

Table	Key Variables
Coverage Data	Country Code, Year, Vaccine Code, Coverage %, Target Number, Doses Administered, Coverage Category
Incidence Rate	Country Code, Year, Disease Code, Incidence Rate, Denominator
Reported Cases	Country Code, Year, Disease Code, Cases
Vaccine Introduction	Country Code, Year, Vaccine Description, Introduced (Boolean)
Vaccine Schedule	Country Code, Year, Vaccine Code, Schedule Round, Target Population, Age Administered

---

Data Cleaning Approach

- Use Python scripts to handle missing values by imputing or removing incomplete records.
- Normalize column names, ensure consistent data types, and handle year formatting uniformly.

- Fill missing numerical data with 0 where appropriate (e.g., doses, cases).
- Final cleaned datasets saved as CSV files and prepared for database ingestion.
- Load cleaned data into a normalized SQL database with referential integrity and optimized schema.

SQL Database Design

- Separate lookup tables for Countries, Vaccines, and Diseases to avoid redundancy.
- Fact tables for CoverageData, DiseaseIncidence, and ReportedCases capturing annual, country-level granular data.
- Additional tables for VaccineIntroduction and VaccineSchedule to incorporate vaccine rollout timing and scheduling details.
- Use of foreign keys and primary keys to maintain data integrity and enable complex queries.

Analytical Methods and Exploratory Data Analysis (EDA)

- Statistical summaries for vaccination coverage and disease incidence by region and year.
- Correlation analysis between vaccination rates and disease incidence reduction.
- Trend analysis pre- and post-vaccine introduction.
- Disparity analysis by demographics (urban/rural, gender, education) and geographic regions.
- Visualization of seasonal uptake patterns and booster dose trends.

Data Visualization with Power BI

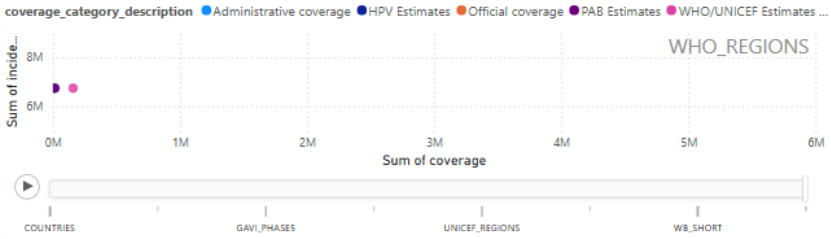
- Geographical heatmaps showcasing vaccination coverage and disease incidence.
- Trend lines and bar charts to display changes in vaccination rates and disease outbreaks over time.
- Scatter plots correlating vaccination coverage and disease incidence by country.
- KPI indicators to track progress toward health targets like 95% coverage for measles.
- Interactive dashboards with filters and slicers to allow deep dive exploration by region, disease, or vaccine.

Results and Insights

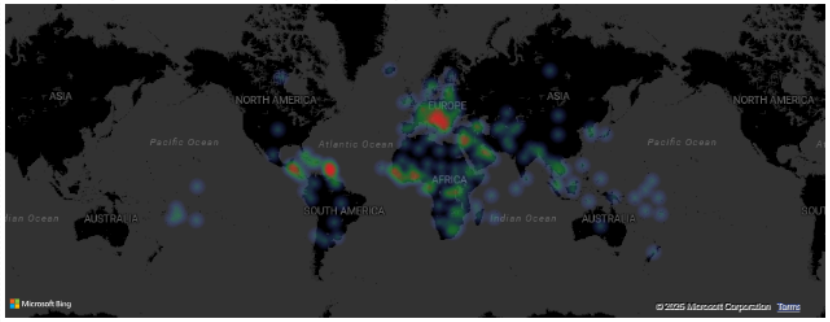
- Dataset cleaning and SQL database design ensure high data quality, ready for advanced analysis.
- Power BI reports visualize complex multidimensional data interactively, enabling better decision-making.
- Identified regions with low vaccination coverage and high disease incidence for urgent intervention.
- Observed positive correlation between vaccine introduction and reduction in disease cases for many antigens.
- Highlighted gaps in booster dose uptake and seasonal variations to inform campaign timing.
- Found disparities in vaccine coverage linked to socioeconomic and demographic factors.
- Supported strategic resource planning and policy recommendations for vaccine rollout enhancements.

Key Visualizations

Sum of coverage and Sum of incidence_rate by coverage_category, coverage_category_description and group

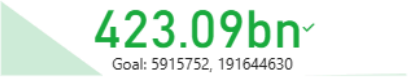


Sum of coverage and Sum of incidence_rate by countryname



Sum of year and Sum of year by disease

Sum of target_number, Sum of coverage and Sum of year by coverage_category_description



Sum of incidence_rate, Sum of year and Sum of target_number by disease_description

Sum of coverage and Sum of incidence_rate by year