**Project: Vaccination Data Analysis and Visualization**

|  |  |
| --- | --- |
| **Project Title** | **Vaccination Data Analysis and Visualization** |
| **Skills** | Python script, Data Cleaning , EDA, MySQL, Power BI |
| **Submitted by** | Vengamma V, DS-WD-E-B23 |

# Project Statement:

Analyze global vaccination data to understand trends in vaccination coverage, disease incidence, and effectiveness. Data will be cleaned, and stored in a SQL database. Power BI will be used to connect to the SQL database and create interactive dashboards that provide insights on vaccination strategies and their impact on disease control.

# Approach:

1. Data was given in 5 different Excel Sheets as Coverage, Incidence, Reported, Vaccine Intro and Vaccine Scheduled.
2. To go through all the excel file - data and reference sheets.
3. Write a Python Script to import the data from each table and analyze the data.
4. After data analysis using Python Pandas – Data frame, the data was exported to MySQL Database WHO\_Vaccination.
5. 5 table’s coverage, incidence, reported, vacintro and vacsch were created in the database.
6. Data Normalization was done on all tables.
7. Power BI was used to create visual analysis using the MySQL database – WHO\_Vaccination.

# Process:

Python:

Python Script was used to write the program to import the excel sheets into DataFrame, analyze the data, connect to MySQL, export the data using DataFrames to the Database – 5 different tables, and finally Normalize the data.

Data Analysis:

* There were no duplicate values across all tables
* Although null values were present, the null values were not removed as there was large amount of data associated with the null values.
* In coverage table the column COVERAGE\_CATEGORY\_DESCRIPTION was dropped as the COVERAGE\_CATEGORY column was self-descriptive.
* The columns Code, group and Name have repeated in tables in coverage, incidence and reported.
* The columns Disease and Disease Description have repeated in incidence and reported.
* The columns ISO\_3\_CODE, COUNTRYNAME, WHO\_REGION repeated in vaccine intro and vaccine schedule

MySQL Database:

Normalization – create new Tables:

* Data was exported to MySQL Database – 5 different tables.
* 3 new tables were created for the common data found across all the tables.
* Distinct Values of Code, group and Name from coverage table was inserted into the new table regioncode(Code Primary Key) .
* Distinct Values from incidence table of Disease and Disease Description was created in a new table – Diseases(Primary Key - Disease).
* Distinct Values ISO\_3\_CODE, COUNTRYNAME, WHO\_REGION from Vaccine Intro was used to create a new table who\_region (Primary Key - ISO\_3\_CODE)

Normalization – create Foreign Keys:

* Foreign Key with column ‘Code’ was created from coverage, incidence and reported to regioncode table.
* Foreign Key with column ‘Disease’ was created from incidence and reported to diseases table.
* Foreign Key with column ‘ISO\_3\_CODE’ was created from Vacintro and Vacsch to who\_region table.

Normalization – Drop Columns:

* Group and Name columns dropped from coverage, incidence and reported tables.
* Disease description column dropped from incidence and reported tables.
* COUNTRYNAME, WHO\_REGION columns dropped from Vacintro and Vacsch table.

Power BI:

* MySQL Database was connected to Power BI.
* All tables were present in the Power BI.
* The Visuals were created based on the questionnaire given in the project requirement and self-analysis.

# Challenges and Solutions:

1. The Normalization of the database:

* During the normalization process, I was unable to create a Foreign Key with the column “ISO\_3\_CODE” from Vaccine Schedule table to the Who\_region (Primary Key - ISO\_3\_CODE)table.
* This was because the table Vaccine Schedule had more distinct values in the column “ISO\_3\_CODE”.
* The remaining distinct values were Inserted into the table Who\_region using the SQL query: “INSERT INTO who\_region(ISO\_3\_CODE,COUNTRYNAME,WHO\_REGION) SELECT DISTINCT ISO\_3\_CODE,COUNTRYNAME,WHO\_REGION from who\_vaccination.vacsch where ISO\_3\_CODE NOT IN (SELECT ISO\_3\_CODE FROM who\_vaccination.who\_region)”

1. Power BI:

* Initially I didn’t notice the “Filters” Pane in the Visual and hence I had written DAX Queries for Male,Female,First Vaccine Dose, Subsequent Vaccine.
* However, I found Filter Pane very useful through-out the Power BI Visual creation.
* I was unable to use columns from 2 tables even though they were related.

To solve the issue, I then used the Modeling – Manage Relationship – New Relationship option.

Thank you