**Project Title**

**Reported Tuberculosis cases and the immunization coverage estimates by country in Europe Union/European Economic Area.**

**Abstract**

**This project involves a series of steps towards Extracting, Transforming and Loading (ETL) of data into a final database for the purpose of being used for future analysis.**

**Objective**

**The objective of this project is to source and combine data from multiple sources on the reported cases of tuberculosis and the coverage of Bacille Calmette-Guerin (BCG) immunization in Europe Union/ European Economic Area into a single, consistent data store and then load into a final database (data warehouse).**

**Introduction**

**According to the latest European Centre for Disease Prevention and Control (ECDC/WHO) report on diseases surveillance and monitoring, tuberculosis was considered as one of the deadliest infectious diseases worldwide, second only COVID-19. In 2020, tuberculosis was reported to have claimed an estimated 21, 000 lives, which equivalent to 2.3 deaths per 100, 000 people and with around 3, 800 (0.8 deaths per 100, 000 people) of these deaths was reported to be in the EU/EEA.**

**In order to contribute to the fight against tuberculosis, this group decided to embark on the process of data integration on reported cases of tuberculosis and immunization coverage in EU/EEA from 2008 to 2012 for the purpose of future analysis.**

**Findings**

**After reading and cleaning the BCG\_vaccine csv on pandas, from the Pandas data frame, it can be noticed that between 2008 and 2012 that the dataset covers, majority of the European countries has between 99% and 92% vaccine coverage. However, Sweden and Ireland have the lowest vaccination rate, 30 to 41% and 21% to 24% respectively.**

**From the second dataset (t*b\_rates.csv)*, Romania, United Kingdom, Poland, France, Germany has the highest number of cases respectively while, Iceland, Malta, Luxembourg, Cyprus and Slovakia have the lowest number of cases respectively.**

**Data Sources**

**The data we have chosen to use is on the reported TB cases in the EU/EEA, from 2008-2012 and the immunization coverage estimates by country in EU/EEA.**

**Our data sources were as follows:**

* **European Centre for Disease Prevention and Control (*https://www.ecdc.europa.eu/en/publications-data/reported-tb-cases-eueea-2008-2012*)**
* **The Official Site of the World Health Organisation** ([***http://immunizationdata.who.int/pages/coverage/bcg.html?CODE=eur***](http://immunizationdata.who.int/pages/coverage/bcg.html?CODE=eur)***)***

**Both datasets were in CSV format, and we have chosen to use Pandas to Transform the data and load it into PostgreSQL. We have chosen a relational database (POSTGRESQL) because both datasets are structured data. The data has been loaded into 2 tables within PostgreSQL.**

**This project involves all 3 parts of the ETL process: Extract, Transform & Load.**

**Data Extraction**

**We loaded the 2 CSV files into Pandas data frames.**

**Transform:**

* **Two comma-separated values files were fetched from the official website of the World Health Organisation and European Centre for Disease Prevention and Control and were named BCG\_vaccine and report\_cases.csv.**
* **These two csv files were then loaded into the Pandas DataFrame.**
* **Data contained in these csv files were displaced in Pandas DataFrames and which lead us into the data cleaning and transformation process.**

**Data Cleaning and Transformation**

**The extracted data were cleansed and transformed using Python Pandas. The following actions were executed in the process of transforming the data set.**

* **Dropping unwanted columns: we dropped some unwanted columns in order to be more intuitive & compatible when uploading them into PostgreSQL (i.e., ensure there are no spaces between column names).**
* **Renaming column headers: almost all column headers in the csv containing the reported tuberculosis cases were renamed in order to make it easier to be converted into our destination database.**
* **Creating a filtered DataFrame from selected columns: specific column headers were copied the data set containing the immunization coverage as not all the columns were needed.**
* **Dropping duplicates: duplicated values were also dropped in order to eradicate undue repetition of values.**
* **Dropping NaN: columns with missing values were also dropped.**

**Loading of Converted DataFrame into Database**

**Database: We decided to use the ProsgreSQL, (Relational Database) as our final database due to the relational model of our data.**

**Two tables were created in Prosgresql using progress query namely, “immunization and reported\_cases”, after which data were loaded into these tables from python Pandas using the sqlachemy engine (please see the ipynb file for details).**

**References**

* **World Health Organisation**

**(**[**https://www.who.int/**](https://www.who.int/)**)**

* European Centre for Disease Prevention and Control

**(**[**https://www.ecdc.europa.eu/en**](https://www.ecdc.europa.eu/en)**)**