**ETL PROJECT**



**Project title : Data Analysis for COVID-19 and Economic**

**Statistics for Countries**

**Team Members:**

**Aroosa Sawati**

**Presit Kaur**

**Nooshin kooshyar**

**Github Link:**

**About the Project**

The outbreak of pandemic COVID-19 across the world has completely disrupted the political, social, economic, religious, and financial structures of the world. The aim of this project is to present how well COVID-19 correlated with economic growth through gross domestic products (GDP). The statistical analysis for this project uses publicly available data from different sources. This project is used for tracking the impact of COVID 19 on economic variation and to see how well and how far in advance the prediction holds true, if at all. This ETL project can support business strategy makers, and create awareness among general public to understand the situation.

Data is critical to support countries in managing the global coronavirus (COVID-19) pandemic. This ETL project provides an array of real-time data, statistical indicators, and other types of data that are relevant to the coronavirus pandemic. Data is drawn from different authoritative sources which is then cleaned and updated.

# ETL Process

# **ETL** is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system (Postgres , MongoDB)

# Extract

Information for this report was extracted from a variety of reliable online sources on March 11th 2021. For the economic data provided, it was sourced from ourworldindata.org. The data was freely available in CSV format and was subsequently downloaded for transformation and analysis purposes. This file contained information relating to the economic statuses of each country as of the current date as well as some information regarding COVID statistics of each country listed.

Information regarding COVID specifically was taken from the World Health Organisation’s specific site dedicated to coverage of the virus. The CSV file extracted not only displayed the cumulative cases and deaths so far but also displays the most prevalent transmission classification for each country, giving some reasoning to the results displayed in the table. The site is regularly updated to ensure that correct information is available for display.

Population data was extracted to provide reasoning for trends listed in the data above. It would help answer questions such as why the data displayed for the country had a transmission classification of “sporadic” or why some countries had higher cases than others. Kaggle user Tanu N Prabhu provided a public CSV file that displayed the population and density statistics for several countries. This file was extracted for analytical purposes.

# Transform

Once the files were downloaded from their respective sources, they were imported into Jupyter Notebooks for transformation purposes. Cleaning of a majority of the data included deleting unnecessary columns and renaming others to enhance readability. As well as that, all of the tables extracted were checked for obsolete values within the rows. If any were found (such as global statistics) they were removed as the analysis is to focus on specific countries rather than a cluster of countries.

# Load

# Loading data into the target data warehouse database is the last step of the ETL process. Connecting to PostgreSQL data looks just like connecting to any relational data source. Create a connection string using the required connection properties. For this project , a connection string is passed as a parameter to the “create engine” function. To connect to PostgreSQL, set the Server, Port (the default port is 5432), and Database connection properties and set the User and Password you wish to use to authenticate to the server. If the Database property is not specified, the data provider connects to the user's default database.

### **Create a SQL Statement to Query PostgreSQL**--- Use SQL to create a statement for querying PostgreSQL.

### **Extract, Transform, and Load the PostgreSQL Data**--- With the query results stored in a Data Frame, we can use extract, transform, and load the PostgreSQL data. In this project, we extract PostgreSQL data, sort the data and load the data into a CSV file.

# Analysis

# Population Data

# 

## **COVID-19 Data**

## This table shows deaths per million from COVID-19 and the transmission classification.

## **Economic Data**

## 

This table shows the effect on GDP per capita against the population for all of the countries.

# Technical Issues