**ETL PROJECT**



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

**Statistics for Countries**

**Team Members:**

**Aroosa Sawati**

**Presit Kaur**

**Nooshin kooshyar**

**Github Link:** https://github.com/presitkaur/ETL-COVID-vs-Economic-Stats

**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

# The following tables were created and uploaded to SQL.

# Population Data

# COVID-19 Data

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

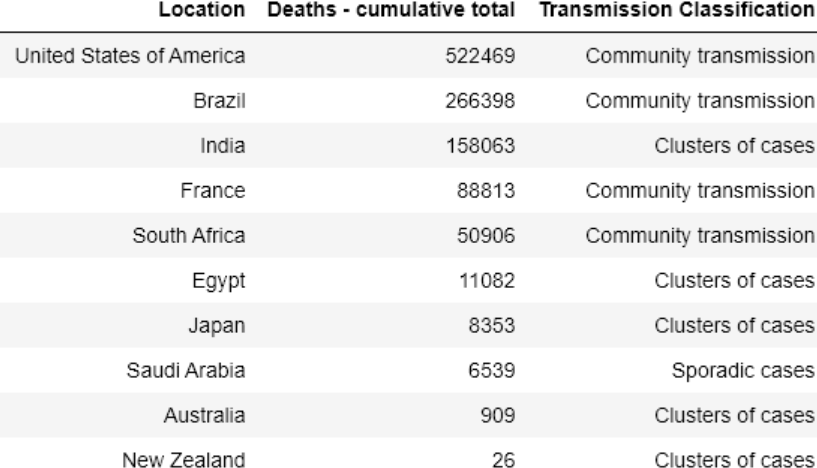
## **Economic Data**

## 

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

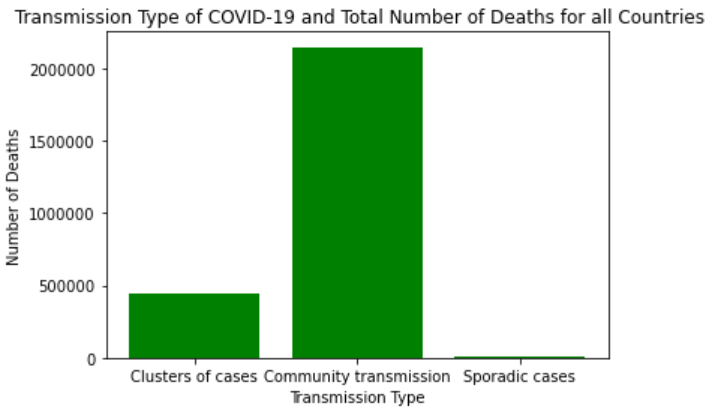
**Analysis**

The data provided can be used to make multiple conclusions about COVID-19 and economic statistics for various countries. Information about population, economic and case/death rate can not only look at the trends seen but also explain why they may occur. For example take *Figure 1:*



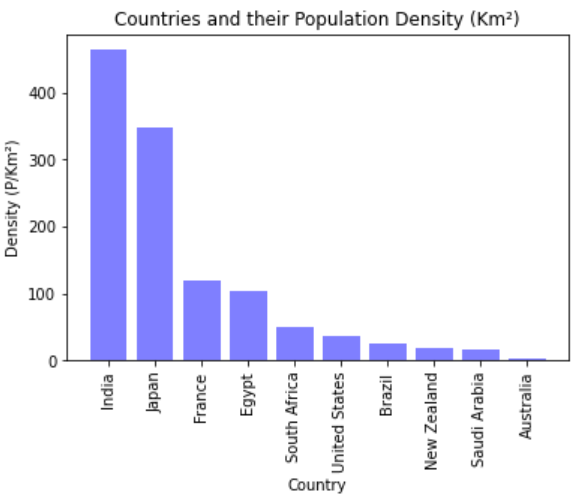
*Figure 1: A table displaying the total deaths from COVID-19 in selected countries and transmission classification*

The information extracted was from two different tables and displays the total deaths from COVID-19 and the mode of transmission of the virus. This data alone cannot provide a definitive conclusion about the relationship between the total deaths and the transmission type of the virus. Therefore the following table would be looked at to see if there is even a relationship to test.

*Figure 2: The total number of COVID-19 deaths for 236 countries based on their transmission classification of the virus*

*Figure 2* strongly suggests that there is a relationship between the transmission type and the total number of deaths from COVID-19. It displays that community transmission causes the most amount of deaths and sporadic cases cause the least. This information is also reflected with the ten sample countries listed in *Figure 1* as a similar trend can also be observed when performing calculations with the information. Therefore it would be reasonable to conclude that there is a distinct relationship between the mode of transmission of the virus in a country and the total number of deaths caused by it.

The conclusion reached above is satisfactory and does provide valid information, however further analysis can be conducted. This means exploring any lurking variables that may exist within the findings. By looking for such pieces of information, the research is able to find a factor that can be manipulated to decrease the deaths or indirectly manipulate the  dependent variable in their inquiry. Examples for this particular analysis include looking at the median age, stringency index, or the Human Development Index for the selected countries. Below, a bar-chart displaying the population density of the selected countries can be seen. It’s goal is to possibly help explain if the population density causes the transmission type of the virus.

*Figure 3: The population density (people per kilometre squared) of the ten selected countries*

The chart displays that India and Japan have significantly higher population densities than other countries. Australia has the lowest population density. India and Japan are both listed to have “clusters of cases” listed as their transmission type. Egypt, which is ranked fourth in population density in the bar chart is also listed to have “clusters of cases” in *Figure 1*.  Australia and New Zealand, however, are on the lower end of the bar chart with significantly lesser population densities. Therefore it could be concluded that whilst population density does have somewhat of an effect on the transmission type, another factor should be explored to reach a stronger conclusion for lurking variables.

**Technical Issues**

As with all analytical projects, this ETL project was not without some initial complications. The first of which was the timeframe provided. Given more time, a more in depth analysis could be achieved as well as a more comprehensive end database. Communication also became a significant issue as the small amount of time provided caused little time to discuss specific tasks of the project to be undertaken by team members.