# DSC-540: Term Project – Milestone 1

## Introduction:

As part of Data Preparation Term project, I have selected COVID-19 as topic of my choice.

COVID-19 has impacted the world very badly and has been a topic of discussion and analysis for over two years.

COVID-19 cases and deaths show how the countries are impacted due to COVID-19, COVID-19 testing and vaccinations are very important and gives a picture about the countries efforts to get back to normalcy.

## Data Sources:

**CSV File:** CSV file consists of data related to the global vaccinations administered so far. This data is per day and needs aggregation to get total vaccinations per country.

Link: https://github.com/somas1986/Data_Preparation/blob/main/vaccinations.csv

**Website:** COVID testing is equally important as vaccinations. Wikipedia has good amount of data related to testing statistics under the section "Testing statistics per Country"

Link: https://en.wikipedia.org/wiki/COVID-19_testing

**API:** COVID 19 API gives details related to overall deaths and confirmed cases to guage the pandemic situation world over.

Link: https://api.covid19api.com/summary

## Relationship between Data Sources:

CSV File: Contains global vaccination details by location, country iso code, date, total vaccinations, people vaccinated, people fully vaccinated, total boosters, daily vaccinations raw, daily vaccinations, total vaccinations per hundred, people vaccinated per hundred, people fully vaccinated per hundred, total boosters per hundred, daily vaccinations per million, daily people vaccinated, daily people vaccinated per hundred.

Website: Contains COVID 19 testing details Country or region, Date, Tested, Units, Confirmed (cases), Confirmed/tested %, Tested/population %, Confirmed/population %, Ref.

API: Contains the COVID cases per country - contains ID, Country, CountryCode, Slug, NewConfirmed, TotalConfirmed, NewDeaths, TotalDeaths, NewRecovered, TotalRecovered, Date and Premium.

All these data sources are related to COVID-19 pandemic and all these datasets are connected by country name/ location.

The CSV file has a 1 to many relationships with the Website by location and has a one to one relationship with the API data by Country as well.

**Project Description:**

In the 5 milestones of this Term project, I plan to perform the below tasks but not limited to:

1. Data Cleansing
2. Identify Outliers and bad data
3. Aggregations
4. Data Merging
5. Data Visualizations using different graphs.

All the data being used is related to COVID-19, its impact and steps being taken to overcome this pandemic.

As we are sourcing data from different sources and in different formats, we might face issues to merge the data into one final format. As I am sourcing data from very liable resources, not sure if I will have much scope for data cleansing but I need to see to implement different functions to accomplish the task.

**Conclusion:**

As part of Milestone 1, I have gathered data from different sources and in different formats. I will use this data in the upcoming milestones and will try to accomplish the goal of data cleansing and data visualization and will store this data in a DB.