Advanced Analytics and Dashboard Design

6.1 Sourcing Open Data – Task 1

1. Data Source Summary: COVID-19 World Vaccination Progress

a) **Data Sourcing:** This is an external data source. The data is provided by Kaggle a subsidiary of Google, that provides access to a wide collection of data sets. This data can be considered as a trustworthy data source.

b) **Data Collection:** The data is collected from Our World in Data GitHub repository for Covid- 19, merged and uploaded. This data is having a source link from World Health Organization that means these data are pooled from numerous sources, including direct reports from Member States, WHO review of publicly available official data.

c) **Data contents:** This data contains the information on vaccinations doses for all countries and territories which include Country name, Total vaccinations, Date of vaccination, Daily vaccinations etc. the details on each column given below:

- **Country**- this is the country for which the vaccination information is provided.
- **Country ISO Code** - ISO code for the country.
- Date - date for the data entry; for some of the dates we have only the daily vaccinations, for others, only the (cumulative) total.
- **Total number of vaccinations** - this is the absolute number of total immunizations in the country.
- **Total number of people vaccinated** - a person, depending on the immunization scheme, will receive one or more (typically 2) vaccines; at a certain moment, the number of vaccinations might be larger than the number of people.
- **Total number of people fully vaccinated** - this is the number of people that received the entire set of immunization according to the immunization scheme (typically 2); at a certain moment in time, there might be a certain number of people that received one vaccine and another number (smaller) of people that received all vaccines in the scheme.
- **Daily vaccinations (raw)** - for a certain data entry, the number of vaccinations for that date/country.
- **Daily vaccinations** - for a certain data entry, the number of vaccinations for that date/country.
- **Total vaccinations per hundred** - ratio (in percent) between vaccination number and total population up to the date in the country.
- **Total number of people vaccinated per hundred** - ratio (in percent) between population immunized and total population up to the date in the country.
- **Total number of people fully vaccinated per hundred** - ratio (in percent) between population fully immunized and total population up to the date in the country.

- **Daily vaccinations per million** - ratio (in ppm) between vaccination number and total population for the current date in the country.
- **Vaccines used in the country** - total number of vaccines used in the country (up to date)
- **Source name** - source of the information (national authority, international organization, local organization etc.)
- **Source website** - website of the source of information.

An explanation for why I have chosen this data set:

I have chosen this data set because I found it interesting to work with. As, we all know how the world has suffered a lot from Covid-19 and how everything is coming back to normal with vaccinations going on in each country. This data has an information on vaccination and its worldwide progress.
Also, it meets the conditions for sourcing open data and its from trustworthy source. With the information available in data set, I believe that it will be useful to find out insights and trends from vaccination.

2. Data Profile:

2.1 Data cleaning and consistency checks

- **Dropping columns:** Original data set has **36063** rows and **15** columns. I observed the data dataset and found three columns which are not required for analysis. So, I dropped the columns: **'iso_code',' source_name',' source_website'**. So now dataset has **36063** rows and **12** columns.

- **Renaming the column names:** I have renamed some of the columns so that they can be easily understood after reading their names. Renamed columns have been mentioned in the Jupyter notebook.

- **Check for missing values in dataset:** Except for 'Date', 'Country' and 'Vaccines' columns, all other columns have missing or empty data. These columns represent the number of vaccinations or people who are vaccinated. Since these columns are cumulative, the null values from these columns won't affect our analysis.

- **Check for duplicate records:** No duplicates found in database.

- **Check for mixed-type data:** No mixed-type data found

2.2 Descriptive Analysis:

- Descriptive statistics like mean, min, max etc. has been calculated in Jupyter notebook.

2.3 Data Limitations and ethical considerations

- The data collected in different countries are from hospitals or medical institutions which can be automatic or human survey data. So, it may contain errors.

- The population estimates are used to calculate per-capital metrics are all based on the last revision of the United Nations World Population Prospects which is from year 2019.

- WHO has mentioned on their site that data published by third-party sites like Our World in Data have not been validated by WHO, and WHO cannot comment on accuracy or completeness.

3. Define questions to explore

With the data available for covid-19 vaccination progress, I would like to analyze it to get answers to below questions:

1. What are the top 10 countries in vaccination progress?(in term of total vaccinations)
2. Which 10 countries are lagging in vaccination? ?(in term of total vaccinations)
3. What are the top 20 countries that are having the highest and lowest of fully vaccinated people per population ?
4. What are the global average vaccinations by month?
5. What are the top 20 countries having highest count for daily vaccinations?