**INTRODUCTION**

I came across the data set online and was really fascinated by the richness of the data and I wanted to practice my skills of data cleaning, analysis and visualization on this rich dataset.

Power BI concepts applied:

* Data cleaning: column cleaning and merging
* Data Modelling: Snowflake model
* DAX concept: Calculated column, calculated measures, custom column, Year(),Table creation

**PROBLEM STATEMENT**

* which countries had the highest number of people vaccinated?
* Which country had the highest number of people fully vaccinated?
* Which countries received the highest vaccination supplies?
* What is the vaccination - supply rate per country?
* Which institution supplied the most vaccines?

**DATA SOURCE**

I downloaded this CSV file from kaggle.com and extracted it into power BI for cleaning, analysis and visualization. It contains two sheet/tables:

1. Country\_Vacination:16 columns 86513 rows
2. Country\_Vacination\_by\_manufacturer: 4 columns 35624

**DATA TRANSFORMATION/CLEANING**

Data was efficiently cleaned and transformed with the power query editor of Power BI. (a screenshot of the applied steps) some of the applied steps included:

* Remove columns that were not relevant to the study.
* Reordered columns
* Removed duplicates
* Some of the Data types were changed from text to whole number.

**DATA MODELLING**

I created four Dimensional tables ‘DimCountry’, ‘DimDate’, ‘DimVaccine’ and ‘DimVaccine2’ from the Fact table ‘Covid Country Vaccination’ thereby allowing Power BI to automatically connect the related tables resulting in a snowflake model.

* Created new column in the” DimVaccine” dimensional table to include various vaccine names.
* Created a separate table “DimVaccine2” to create a snowflake model relationship with the “DimVaccine” table.
* Used DAX concept to create calculated columns and measures.

**DATA ANALYSIS & VISUALIZATION**

**CONCLUSION & RECOMMENDATION**

The data (country vaccinations) contains the following information:

* **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 vaccination 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 vaccination for that date/country;
* **Daily vaccinations** - for a certain data entry, the number of vaccination 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;
* **Number of vaccinations per day** - number of daily vaccination for that day and 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;

There is a second file added recently (country vaccinations by manufacturer), with the following columns:

* **Location** - country;
* **Date** - date;
* **Vaccine** - vaccine type;
* **Total number of vaccinations** - total number of vaccinations / current time and vaccine type.