**EDA on Covid-19 World Vaccination Progress**

Dataset that was selected for Exploratory Data Analysis is Covid-19 World Vaccination. CSV file that was used for analysis is country\_vaccinations.csv

**Outcome of EDA**

Based on initial analysis that is during data cleaning, it was found that there was a duplicate data for United Kingdom i.e., one for United Kingdom and separately for Scotland, England, Wales and Northern Island. If we add the data of Scotland, England, Wales and Northern Island we are getting the data of United Kingdom. Therefore, I have deleted the data related to the Scotland, England, Wales and Northern Island.

From the initial analysis after cleaning the data, when I tried to plot the histogram in accordance with the total vaccinations, observed that China is leading but when histogram was plotted against total number of people vaccinated, I do not see China in the histogram. When verified the same, I am seeing that there is no data available for total number of people vaccinated in China. From all the histograms that were plotted United states, India, United Kingdom, Brazil and Germany are the top 5 countries leading in vaccinating their population.

Pfizer is widely used in the United States, France and Germany. AstraZeneca is widely used in India and used less in France and Germany.

**What do you feel was missed during analysis?**

Due to miss in analysis of the data during the data cleaning. Some of the plots like scatter plots were not represented as expected.

**Were there any variables you felt could have helped in the analysis?**

I do not see any other extra variables that could have helped in the analysis. Data set has all the columns that were required in the analysis.

**Were there any assumptions made you felt were incorrect?**

Assumed that the data set does not have any outliers when histograms were plotted against total number of people vaccinated and some other variables. But when scatter plots were represented seeing some outliers.

**What challenges did you face, what did you not fully understand?**

I faced issues during scatter plots, I was not able to scale the X axis and Y axis with the proper units which might have helped in correct and best representation of scatter plots. I did not understand how scatter plots can be represented effectively by removing the jittered noise that was caused due to the mismatch in the units of the variables.