ANAND INSTITUTE OF HIGHER TECHNOLOGY
OLD MAHABALIPURAM ROAD,
KALASALINGAM NAGAR,
KAZHIPATTUR- 603103

IBM - NAAN MUDHALVAN
DATA ANALYTICS WITH COGNOS
COVID VACCINES ANALYSIS

PHASE – 4

NAME : SUBATHRA E

REG No. : 310121104102

BRANCH : COMPUTER SCIENCE & ENGINEERING

YEAR/SEM : III / V

COVID VACCINES ANALYSIS

INTRODUCTION

- The pace of the COVID-19 Vaccine development process is unprecedented and is challenging the traditional paradigm of vaccinology science.
- The main pressure comes from the pandemic situation, but what makes it possible is a complex set of factors and innovative environments built along the times, which this manuscript aims to study.
- The world has witnessed an unprecedented series of events triggered by the pandemic of COVID-19(coronavirus disease),a disease caused by SARS-CoV-2, a new virus belonging to the Coronavideae family, of great impact on individual and collective health worldwide, and high impact implications for the global economy.
- On the other hand, it is possible to identify positive aspects in facing the pandemic, ranging from humanitarian solidarity aid actions to accelerating strategies for the development of vaccines, which assumes the position of main hope in solving this problem of global scope.
- The COVID-19 pandemic has spurred unprecedented efforts in vaccine development and distribution. As vaccines are administered to millions of people worldwide, it is crucial to monitor and optimize the distribution process while closely monitoring adverse effects. Advanced machine learning techniques can play a pivotal role in achieving these goals.

PHASE - 4 : { DEVELOPMENT PART - 2}

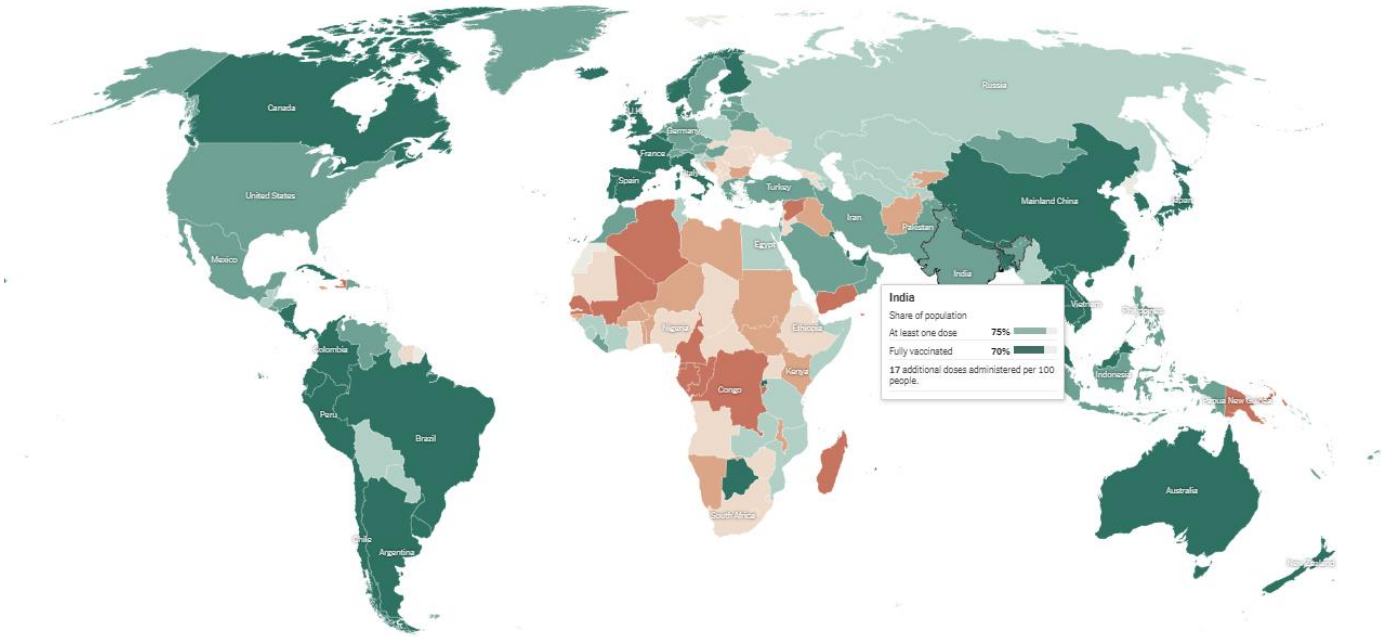
OBJECTIVES

- In this phase, we will continue building upon the foundation established in the earlier phases.
- The phase 4 will encompass two major components: statistical data analysis and data visualization.
- In this phase defines start to building the project by loading and preprocessing the dataset and perform different analysis and visualization using IBM Cognos.

PROBLEM STATEMENT

In this part we will continue to build our project by conducting the Covid-19 vaccines analysis by performing :

- Exploratory data analysis
- Statistical analysis
- Visualization



Data Collection :

COVID VACCINES ANALYSIS is done by using the Dataset of **"COVID-19 World Vaccination Progress"** provided by the dataset site www.Kaggle.com

DATASET:

https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress

Data Observation :

The dataset "COVID-19 World Vaccination Progress" on Kaggle is a collection of data related to the COVID-19 Vaccination efforts worldwide. It provides information about the progress of COVID-19 Vaccinations in various countries and regions. This dataset is designed to help researchers, data scientists, and analysts understand and analyze the progress of COVID-19 Vaccination campaigns across different countries. A second file, with manufacturers information is included.

Importance of Visualizing the Dataset :

The importance of visualizing datasets cannot be overstated in the realm of data analysis. Visualization serves as a powerful bridge between raw data and human comprehension. It enhances our ability to understand, interpret, and extract meaningful insights from complex datasets.

By transforming numbers and statistics into charts, graphs, and interactive displays, visualization offers several key advantages. Firstly, it enables us to detect patterns, trends, and outliers that might remain hidden in tabular data, facilitating more accurate and timely decision-making. Moreover, it supports data exploration by allowing users to interact with the data, making it easier to uncover specific details and refine analysis.

This feature is particularly valuable in the age of big data, where sifting through vast datasets can be a formidable challenge. It is a time-saving tool that provides a rapid overview of data, streamlining the analysis process.

Visualizing the dataset :

- Visualizing a dataset in Python is a fundamental aspect of data analysis and interpretation. It involves creating graphical representations of data to uncover patterns, relationships, and insights, making it an essential tool in data-driven decision-making and storytelling.

- Visualizing a dataset in Python is the process of using data visualization libraries like Matplotlib, Seaborn, or Plotly to create graphical representations of data. The goal is to gain insights, identify patterns, and present data in a visual format that is easy to understand.

Identify The Dataset:

The first step is to identify the dataset that you want to load. This dataset may be stored in a local file, in a database, or in a cloud storage service.

Load The Dataset:

Once you have identified the dataset, you need to load it into the machine learning environment. This may involve using a built-in function in the machine learning library, or it may involve writing our own code.

Preprocess The Dataset:

Once the dataset is loaded into the machine learning environment, you may need to preprocess it before you can start visualizing the dataset. Because, the raw data may contain numerous Null values and anomaly values. So it could be preprocessd also for training and evaluating the model. This may involve cleaning the data, transforming the data into a suitable format, handling the missing values.

Let's see, How the covid vaccines dataset is loaded and visualized using Python Jupyter Notebook and IBM cognos.

EXPLORATORY DATA ANALYSIS :

BASIC INFO ABOUT DATASET :

```
print('Data point starts from:',df.date.min(),'\n')
print('Data point ends at:',df.date.max(),'\n')
print('Total no of Countries in the data set:',len(df.country.unique()),'\n')
print('Total no of unique Vaccine Schemes in the data set:',len(df.vaccines.unique()),'\n')
```

```
Data point starts from: 2020-12-02

Data point ends at: 2022-03-29

Total no of Countries in the data set: 219

Total no of unique Vaccine Schemes in the data set: 84
```

DATAFRAME DESCRIBE :

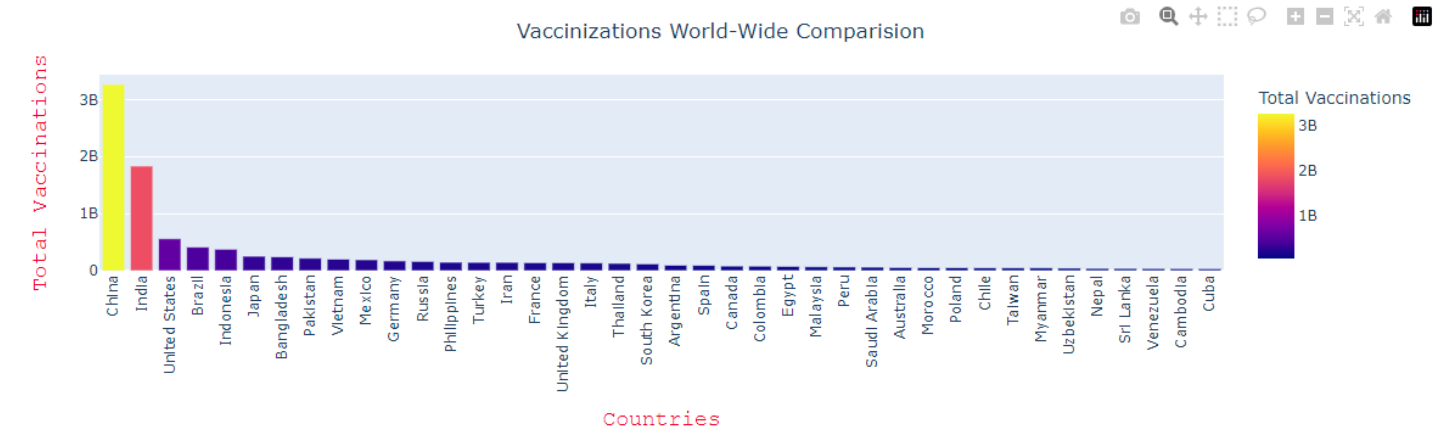
```
In [7]: df.describe()
```

Out[7]:

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred	people_vaccina
count	8.651200e+04	8.651200e+04	8.651200e+04	8.651200e+04	8.651200e+04	86512.000000	
mean	2.315117e+07	8.451007e+06	6.341251e+06	1.106083e+05	1.308517e+05	40.419616	
std	1.611037e+08	4.969867e+07	3.890729e+07	7.864756e+05	7.669487e+05	62.707869	
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000	
25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	8.770000e+02	0.000000	
50%	1.008000e+03	0.000000e+00	0.000000e+00	0.000000e+00	7.245000e+03	0.010000	
75%	3.697554e+06	1.843103e+06	1.137869e+06	1.280625e+04	4.370450e+04	68.750000	
max	3.263129e+09	1.275541e+09	1.240777e+09	2.474100e+07	2.242429e+07	345.370000	

VACCINIZATIONS WORLD-WIDE COMPARISION :

```
fig = px.bar(country_data[:40], x = 'Country', y = 'Total Vaccinations', color = 'Total Vaccinations')

fig.update_layout(title = dict(text = 'Vaccinizations World-Wide Comparision', x=0.5, y=0.95))
fig.update_xaxes(title = 'Countries', title_font = dict(size=18, family='Courier', color='crimson'), tickangle=-90)
fig.update_yaxes(title = 'Total Vaccinations', title_font = dict(size=18, family='Courier', color='crimson'))

fig.show()
```

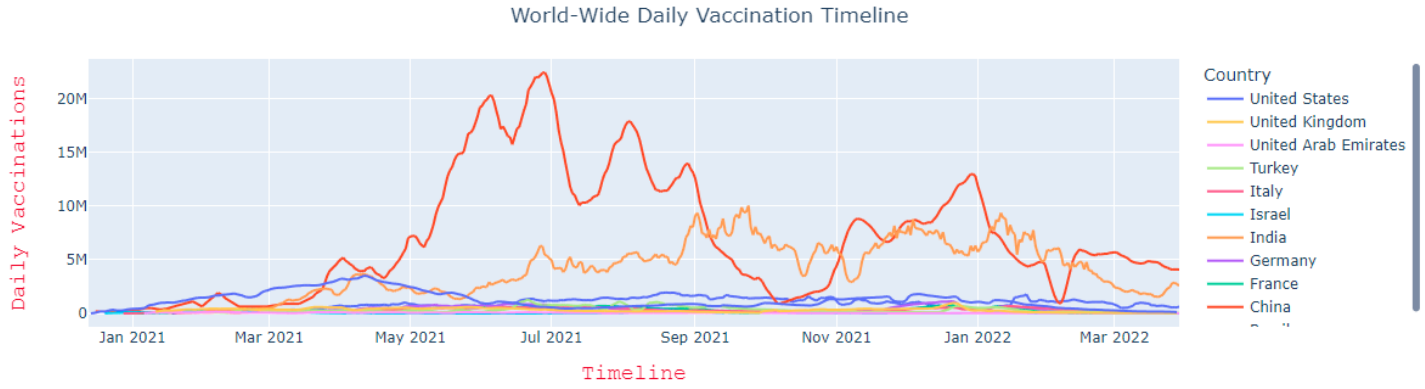


From the plot, some interesting facts stand out:

- The **United States**, despite having the highest number of people affected by Covid-19, has the highest number of vaccinated people.
- **China**, from where the virus started spreading, is at second.
- **India**, who has been supplying vaccines to the world is at 3th position.
- **UK**, where we have found a new variant strain of the virus, is right next.
- Following that, we have **Israel**, **UAE**, **Brazil**, **Germany** and others

COUNTRY WISE DAILY VACCINATION :

```
top_countries = ['USA','CHN','GBR','IND','ISR','ARE','BRA','DEU','TUR','ITA','FRA']
fig = px.line(df[df.iso_code.isin(top_countries)], x='date', y='daily_vaccinations', color='country')

fig.update_layout(title = dict(text = 'World-Wide Daily Vaccination Timeline', x=0.5, y=0.95),
                  legend = dict(title = 'Country', traceorder = 'reversed'))
fig.update_xaxes(title = 'Timeline', title_font = dict(size=18, family='Courier', color='crimson'))
fig.update_yaxes(title = 'Daily Vaccinations', title_font = dict(size=18, family='Courier', color='crimson'))

fig.show()
#Country wise daily vaccination
```

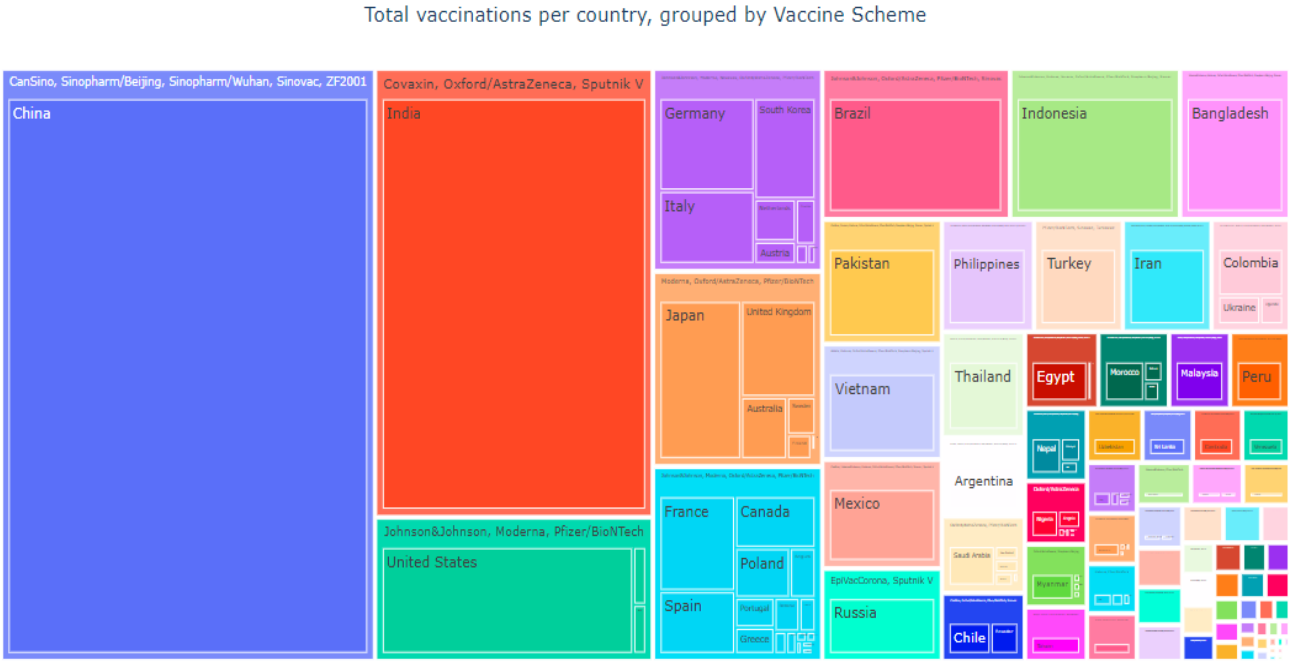


From the plot, we can deduce:

- The Line plot for China is composed entirely of straight lines. This can be attributed to the CCP which tries to restrict flow of information in and out of China. Thus, information from China usually comes in intervals and can be taken with a grain of salt.
- Comparatively, the plot of vaccinations in the USA is better plotted. We can also see that while the USA was heavily affected by the virus, its vaccination drive is highly effective.
- Others like the UK have a steady increase in Daily Vaccinations and India, while supplying to many countries, maintains a respectable 3th position.

TREEMAP OF TOTAL VACCINATIONS PER COUNTRY, GROUPED BY VACCINE SCHEME :

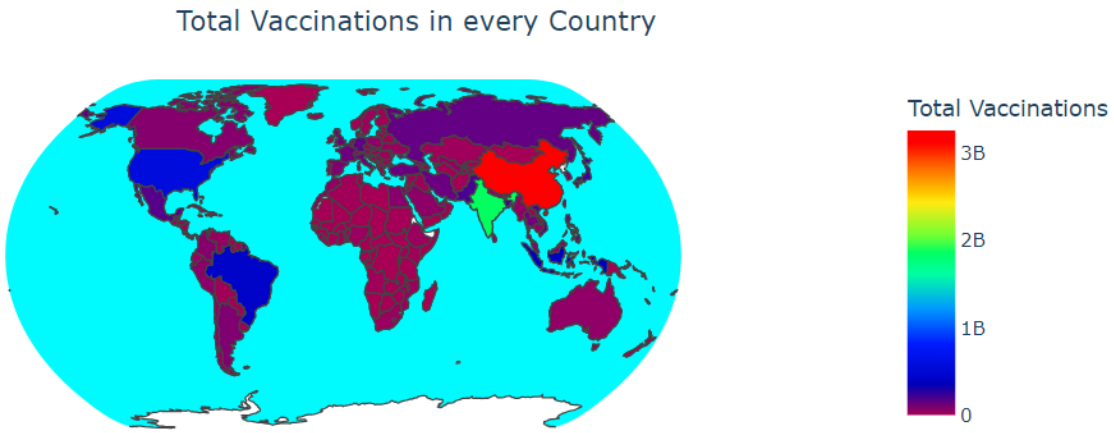
```
fig = px.treemap(country_data, path = ['Vaccines', 'Country'], values = 'Total Vaccinations', height = 650,
                 custom_data = ['Country', 'Vaccines', 'Total Vaccinations'])

fig.update_layout(title = dict(text = 'Total vaccinations per country, grouped by Vaccine Scheme', x=0.5, y=0.95))
fig.update_traces(hovertemplate = 'Country: %{customdata[0]}<br>Vaccine: %{customdata[1]}<br>Total Vaccinations: %{customdata[2]}')
fig.show()
```



- From the above Treemap we can realise that a Bar and Pie Plot may often only show a part of the information that can be observed, whereas a Treemap can accurately show the share of a particular vaccine world-wide, the countries that are using the said vaccine and can even show comparisons between all the countries.
- As the Treemap shows so much information at a time, it can help one understand the data much more accurately.

TOTAL VACCINATIONS IN EVERY COUNTRY :

```
fig = px.choropleth(country_data, locations = 'Country', color = 'Total Vaccinations',
                    locationmode = 'country names', color_continuous_scale = 'rainbow',
                    hover_name = 'Country', projection = 'natural earth')

fig.update_layout(title = dict(text = 'Total Vaccinations in every Country', x=0.5, y=0.95),
                  geo = dict(showocean = True, oceancolor = "#7af8ff", showland = True,
                             landcolor = "white",showlakes = False, showframe = False))

fig.show()
```



- In the above visualisation, we can see the countries and the total vaccinations they have completed.

Statistical Analysis :

Which countries started vaccinations first?

In [5]:

```
# Find out which countries started vaccinations earliest
vacc['date'] = pd.to_datetime(vacc['date'], utc=True)
vacc_start = vacc.loc[vacc[vacc.total_vaccinations > 0].groupby('country')['date'].idxmin()].sort_va
lues('date')
vacc_start.head(5)
```

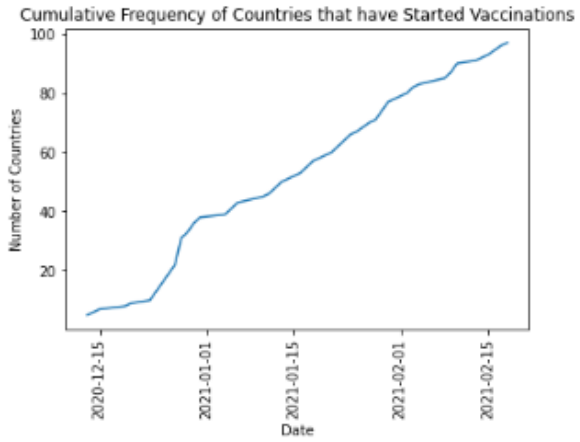
Out[5]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw
3487	Wales	NaN	2020-12-13 00:00:00+00:00	8212.0	8212.0	NaN	NaN
3357	United Kingdom	GBR	2020-12-13 00:00:00+00:00	86265.0	86265.0	NaN	NaN
999	England	NaN	2020-12-13 00:00:00+00:00	55437.0	55437.0	NaN	NaN
2807	Scotland	NaN	2020-12-13 00:00:00+00:00	18993.0	18993.0	NaN	NaN
2271	Northern Ireland	NaN	2020-12-13 00:00:00+00:00	3623.0	3623.0	NaN	NaN

How have the cumulative number of countries adopting covid-19 vaccinations evolved over time?

In [7]:

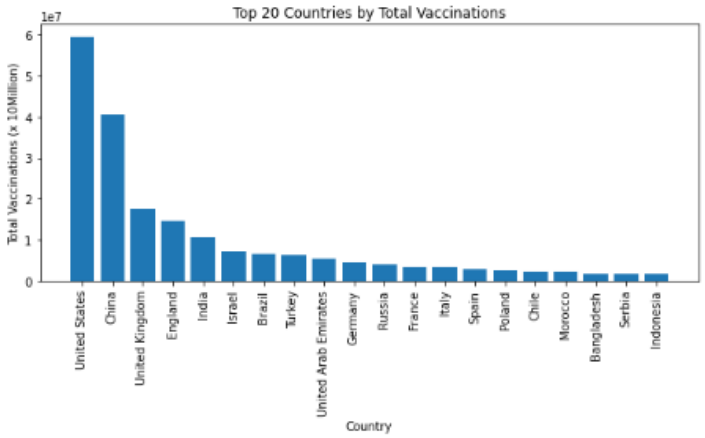
```
# Cumulative distribution of vaccination start dates
events = pd.Series(vacc_start.date.value_counts())
events.index = pd.to_datetime(events.index)
events.sort_index(inplace=True)

plt.plot(events.cumsum())
plt.xticks(rotation=90)
plt.title('Cumulative Frequency of Countries that have Started Vaccinations')
plt.xlabel('Date')
plt.ylabel('Number of Countries')
plt.show()
```



What are the top 20 countries in terms of total number of vaccines administered?

In [11]:

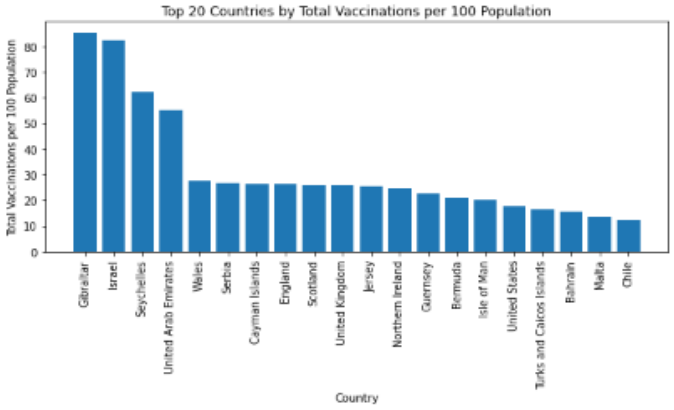
```
# Plot out which countries have performed most vaccinations in descending order
plt.figure(figsize=(10, 4))
plt.bar(vacc_total.country[0:20], vacc_total.total_vaccinations[0:20])
plt.xticks(rotation=90)
plt.title('Top 20 Countries by Total Vaccinations')
plt.xlabel('Country')
plt.ylabel('Total Vaccinations (x 10Million)')
plt.show()
```



What are the top 20 countries in terms of number of vaccines administered per 100 population?

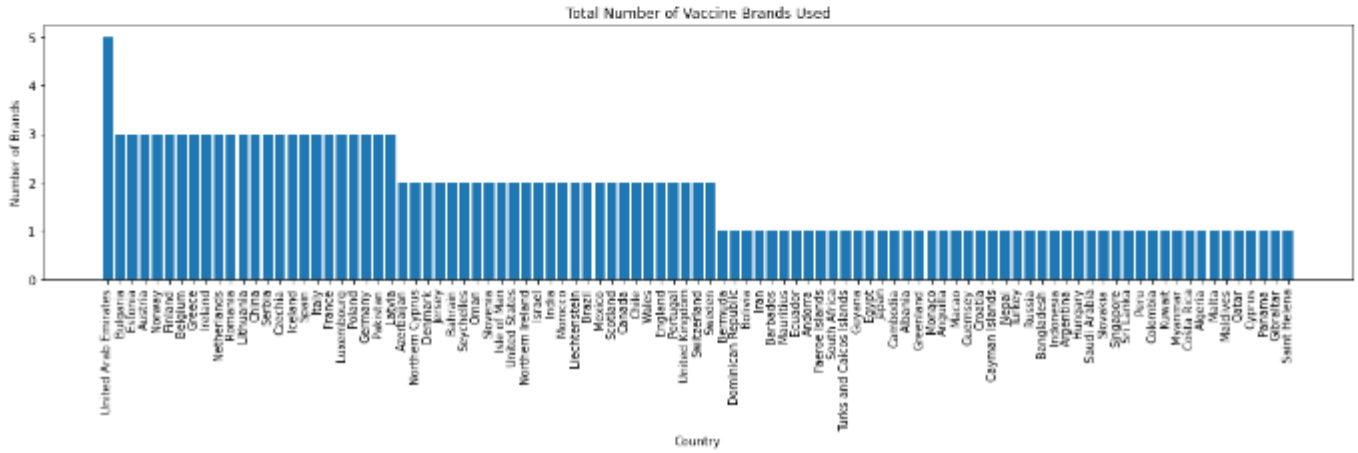
In [14]:

```
# Find out which countries have performed most vaccinations with respect to their populations

plt.figure(figsize=(10, 4))
plt.bar(vacc_total.country[0:20], vacc_total.total_vaccinations_per_hundred[0:20])
plt.xticks(rotation=90)
plt.title('Top 20 Countries by Total Vaccinations per 100 Population')
plt.xlabel('Country')
plt.ylabel('Total Vaccinations per 100 Population')
plt.show()
```



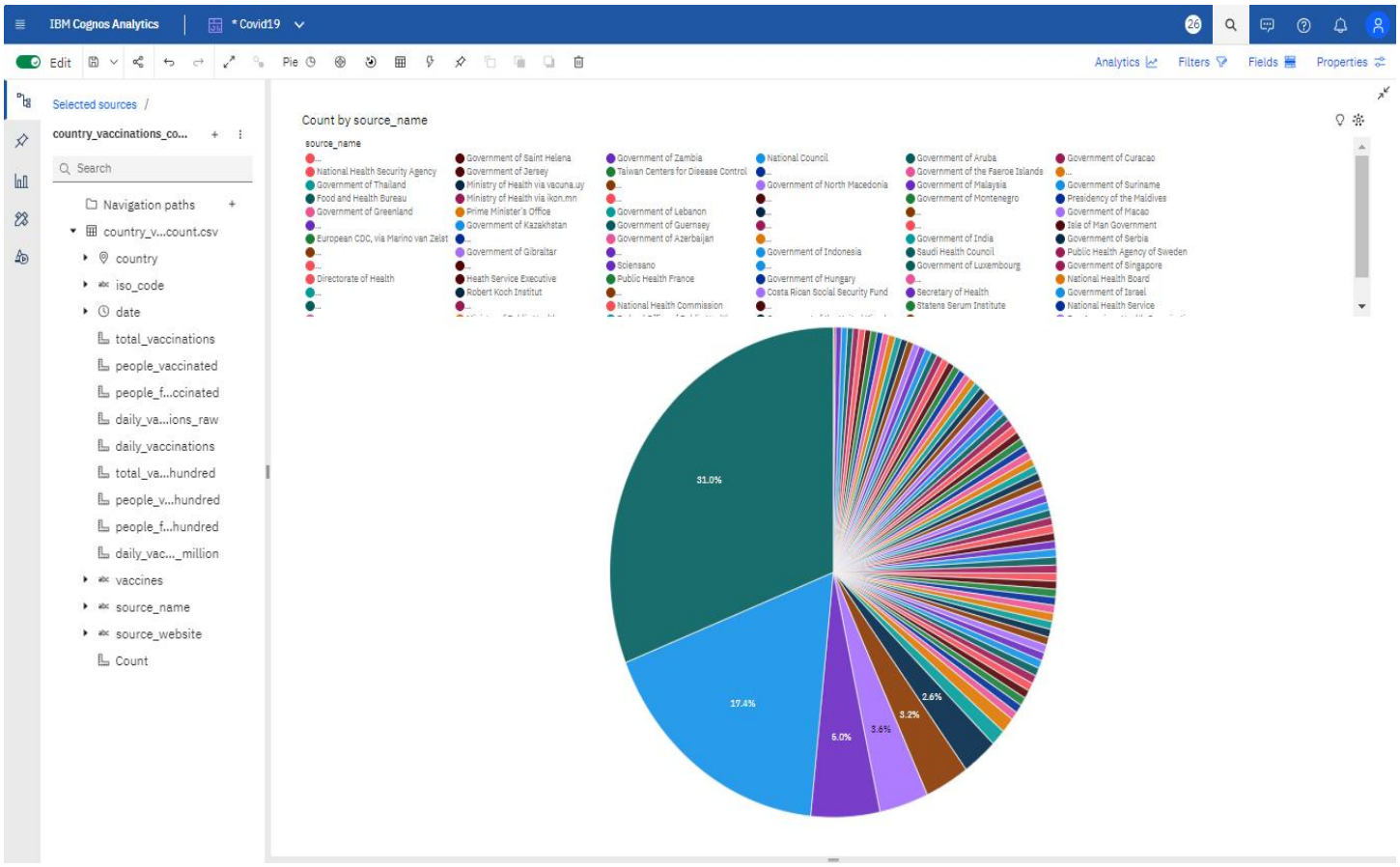
Which countries are using more than 1 type of vaccine, and how many are they using?

```
# How many vaccines is each country using

vacc_total['vacc_brands'] = vacc_total.iloc[:,-5:].apply(lambda x: (5 - x.isnull().sum()), axis='col
umns')
vacc_total = vacc_total.sort_values('vacc_brands',ascending=False)
plt.figure(figsize=(20, 4))
plt.bar(vacc_total.country, vacc_total.vacc_brands)
plt.xticks(rotation=90)
plt.title('Total Number of Vaccine Brands Used')
plt.xlabel('Country')
plt.ylabel('Number of Brands')
plt.show()
```



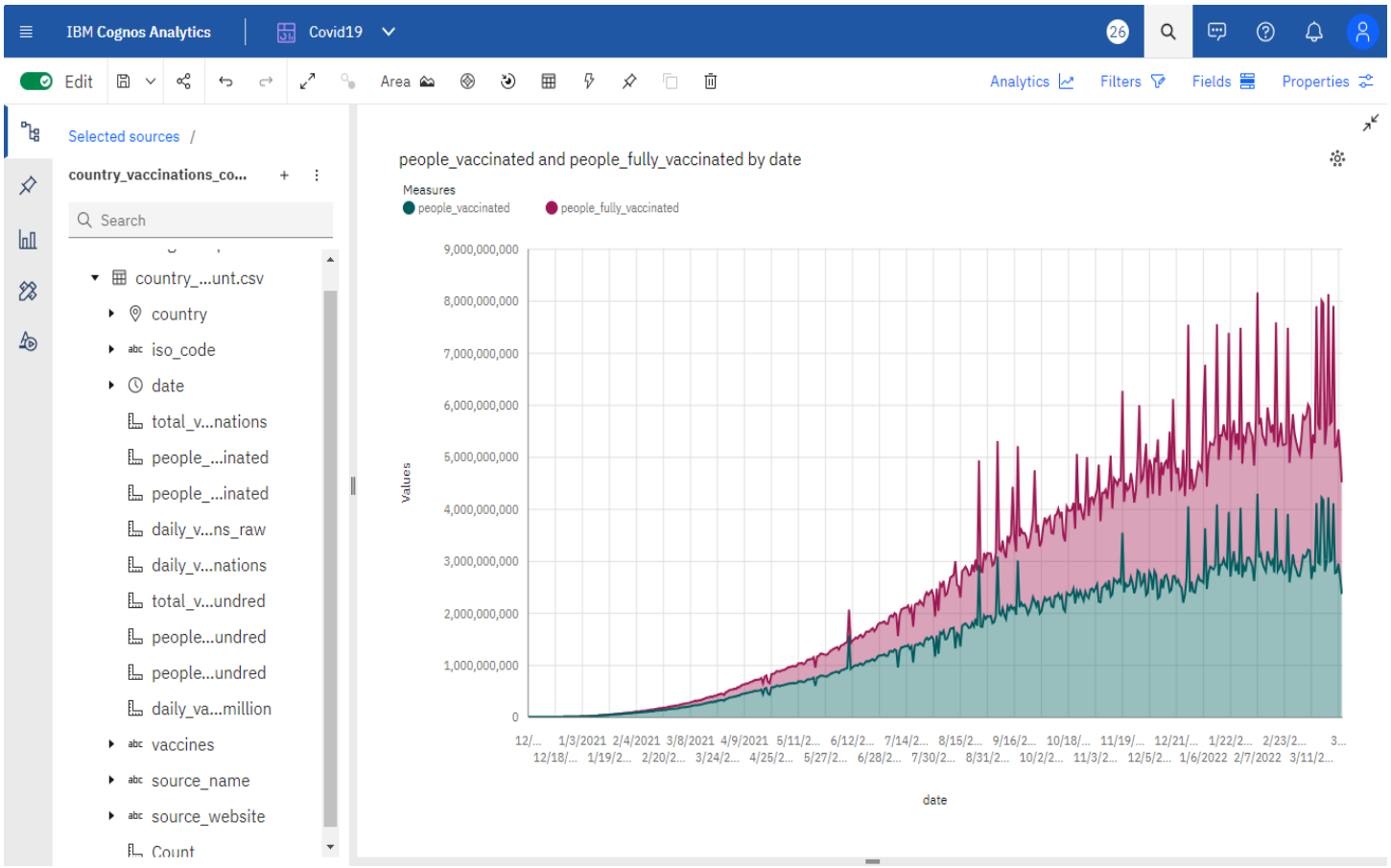
Visualization of the dataset Using Cognos

Visualization using pie plot :



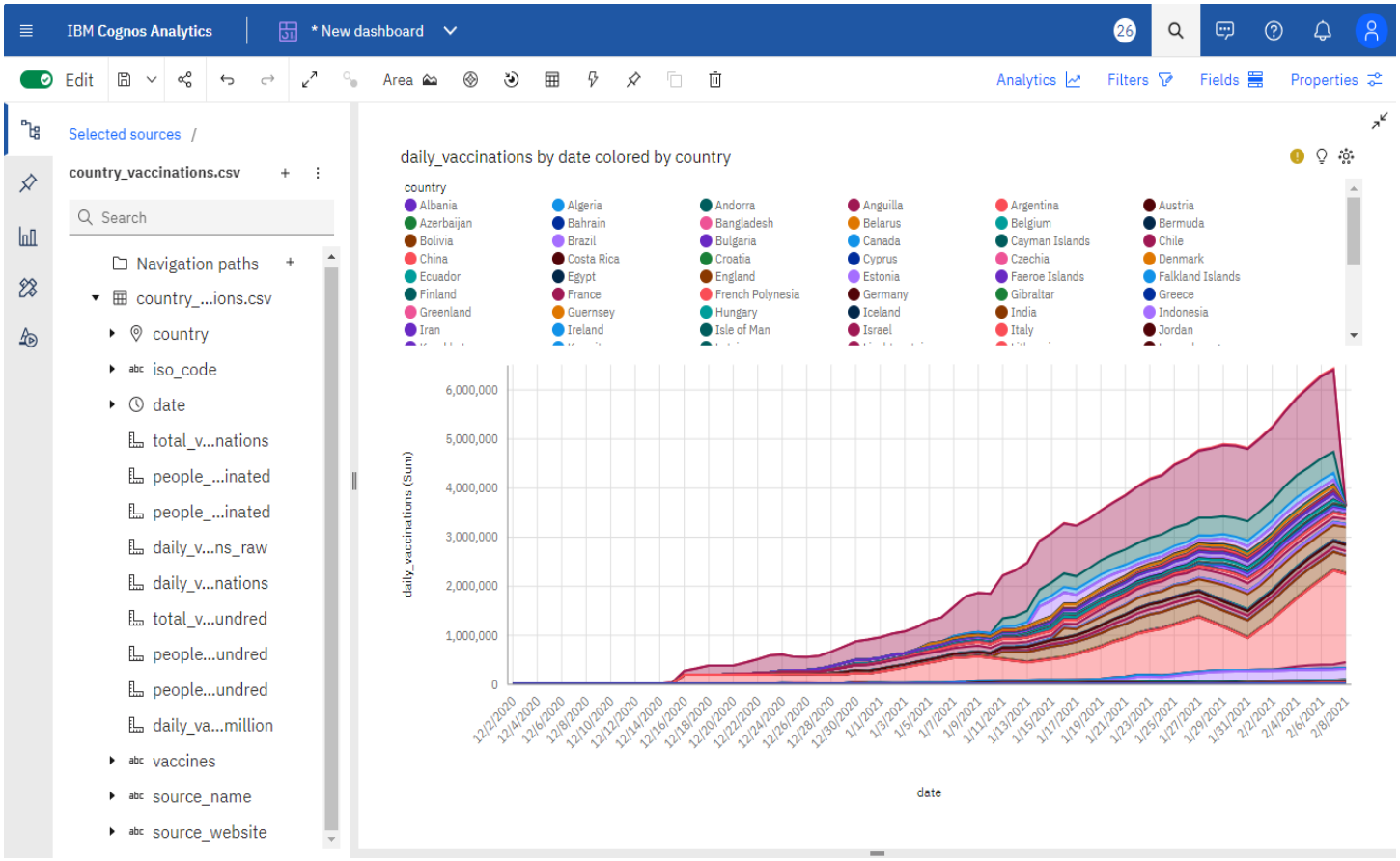
Visualizing the percentage of Source Names for Vaccines using the Pie Chart provided in the Cognos.

Visualizing using Line Chart :



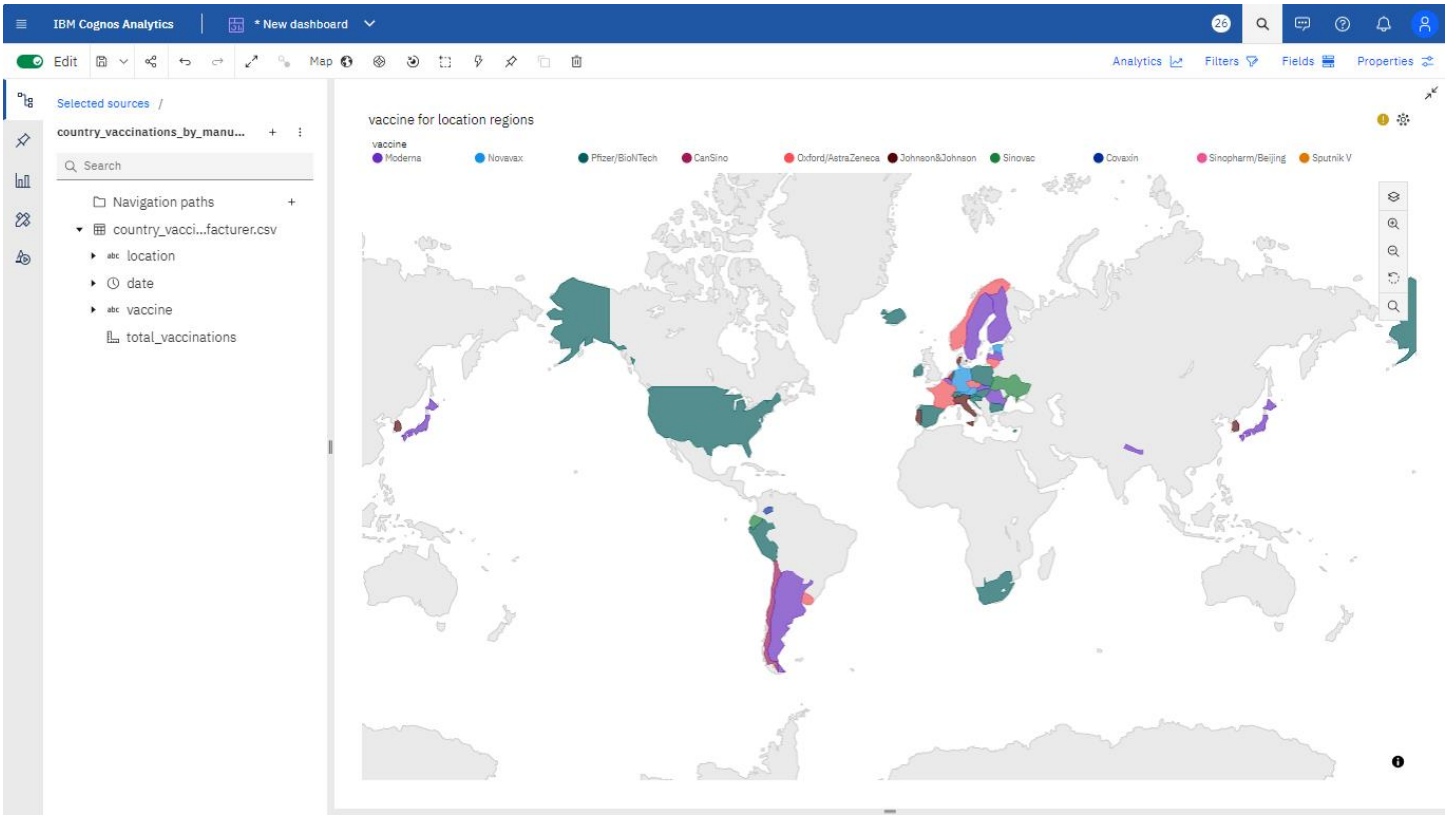
Visualizing the people_vaccinated and people_fully_vaccinated by the date using the Line Chart provided in the Cognos.

Visualization using Area :



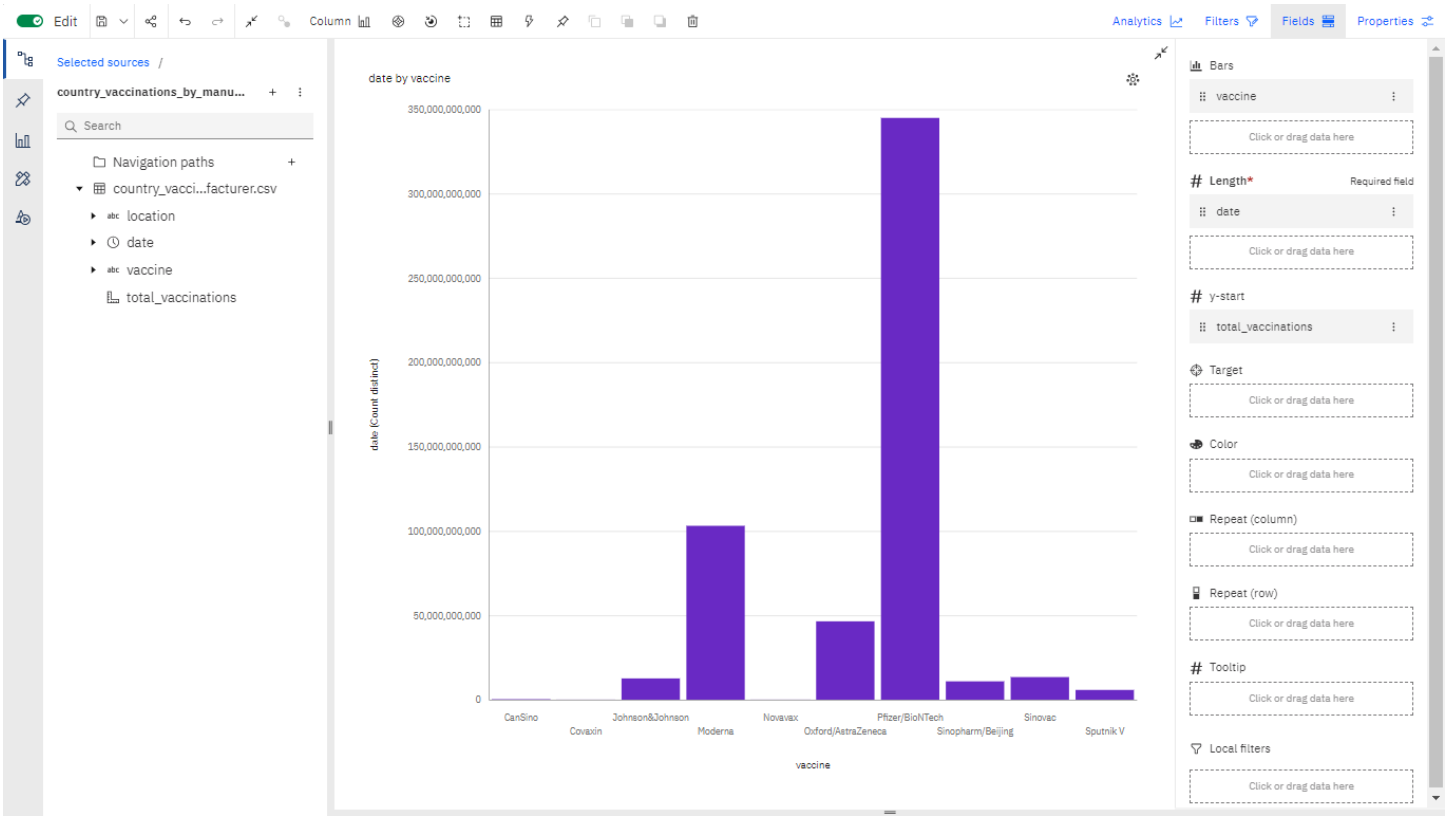
Visualizing the daily_vaccinations by the date in all the country using the Area Chart provided in the Cognos.

Visualization using Map :



Visualizing the most vaccines used in the locations using the Map Chart provided in the Cognos.

Visualization using Bar Chart :



Visualizing the amount of vaccines used in the date using the Bar Chart provided in the Cognos.

CONCLUSION

- In this phase, we have built upon the foundation established in the earlier phases.

- The phase 4 encompassed two major components:
  statistical data analysis and data visualization.

- This defines project by loading and preprocessing the dataset and performed different analysis and visualization using IBM Cognos.

- The need to rapidly develop a vaccine against COVID-19 occurs at a time of great excitement in basic scientific understanding, as well as strategies learned in the past by industry and optimization of regulatory pathways. It is expected that these factors, arising from the global emergency, may redirect the R&D processes for new drugs, especially in times of pandemic.

- In this part we have built our project by conducting the Covid-19 vaccines analysis by performing :
  - Exploratory data analysis
  - Statistical analysis
  - Visualization

IBM DATA ANALYTICS WITH COGNOS

TEAM NAME : Proj_229800_Team_1

PROJECT : 3101-COVID Vaccines Analysis

TEAM MEMBERS :

- PAVITHRA V
- SHARU DHARSHINI S
- SUBATHRA E
- SHALINI S

TEAM LEADER : PAVITHRA V