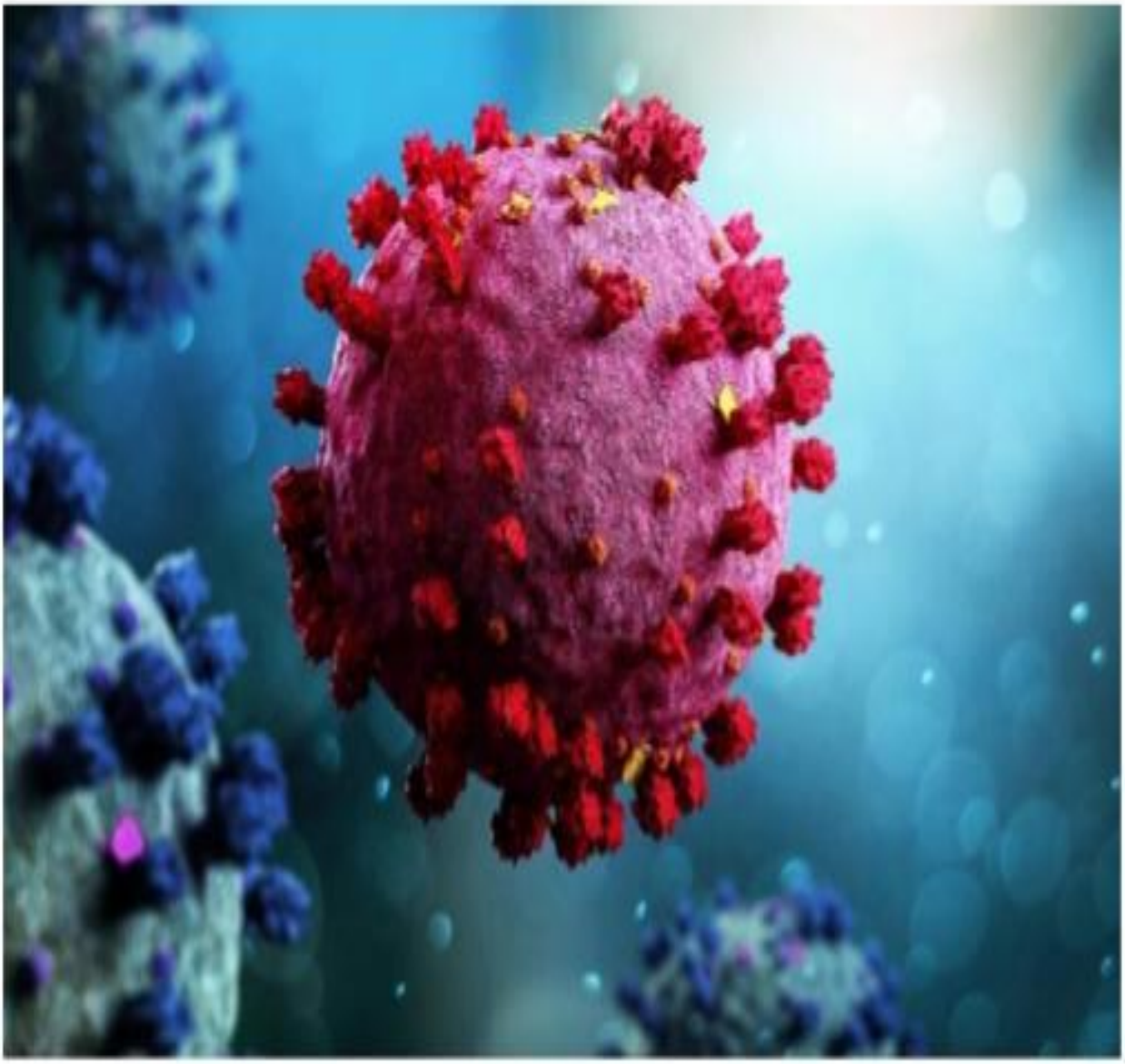
Covid-19 vaccine analysis

PHASE 5 SUBMISSION DOCUMENT

Name:Sruthikka.S

Reg No:712221104022

College:Park college of engineering and technology



PHASE 1

Problem Definition :

- The problem is to conduct an in-depth analysis of Covid-19 vaccine data, focusing on vaccine efficacy, distribution, and adverse effects. The goal is to provide insights that aid policymakers and health organizations in optimizing vaccine deployment strategies. This project involves data collection, data preprocessing, exploratory data analysis, statistical analysis, and visualization.

Design Thinking

- Data Collection: Collect customer data, including attributes like purchase history, demographic information, and interaction behavior.
- Data Preprocessing: Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.
- DesignClustering Algorithms: Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

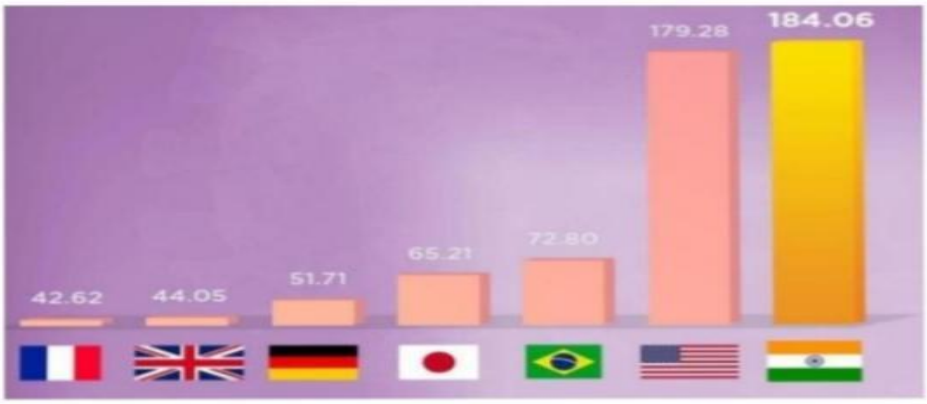
Data set:

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

Phase 2

Innovation:

DATA SET:



India Leading the world covid vaccination

More than 184 million individuals fully vaccinated so far

Types of covid -19 vaccine india

Vaccination coverage in states

Percentage of eligible population receing two doses

DATA SET

https://in.docworkspace.com/d/sIKSisfjMAY6OmakG

The following link contains the dataset of country vaccinations

Phase 3

Covid-19 vaccine Analysis

Introduction:

- The analysis of COVID-19 vaccines is a critical area of study, especially in the context of the global response to the pandemic. This analysis involves assessing various aspects of COVID-19 vaccines, including their development, efficacy, safety, distribution, and impact on public health.

- Researchers and experts have been evaluating data from clinical trials, real-world studies, and post-market surveillance to understand how these vaccines perform in different populations.

- The goal of such analysis is to provide evidence-based information to guide vaccination strategies, monitor vaccine safety, and contribute to the ongoing effort to control the spread of the virus. In this analysis, factors such as vaccine technology, effectiveness against different variants, vaccine hesitancy, and equity in distribution have also been subjects of significant interest and investigation.

Given data set:

	location	date	vaccine	total_vaccinations
0	Argentina	2020-12-29	Moderna	2
1	Argentina	2020-12-29	Oxford/AstraZeneca	3
2	Argentina	2020-12-29	Sinopharm/Beijing	1
3	Argentina	2020-12-29	Sputnik V	20481
4	Argentina	2020-12-30	Moderna	2
5	Argentina	2020-12-30	Oxford/AstraZeneca	3
6	Argentina	2020-12-30	Sinopharm/Beijing	1
7	Argentina	2020-12-30	Sputnik V	40583
8	Argentina	2020-12-31	Moderna	2
9	Argentina	2020-12-31	Oxford/AstraZeneca	3

Necessary step to follow:

1. Import Libraries:

Start by importing the necessary libraries:

Program:

```
import pandas as pd
import plotly.express as px
import plotly.graph_objects as go
from folium.features import Choropleth
import folium
from folium.features import Tooltip
import seaborn as sns
```

2.Load the Dataset:

Load your dataset into a pandas Dataframe.you can typically find covid-19 vaccine analysis dataset in CSV format ,but you can Adapt code to other format as needed.

Program:

```
df = pd.read_csv(covid-world-vaccination-progress/country_vaccinations_by_manufacturer.csv")
```

```
df['date'] = pd.to_datetime(df['date'])
```

In our dataset, the Total Vaccinations represent the cumulative sum of vaccinations up to that date. To express the usage of different vaccines by countries, we need to clean the dataset and transform it.

```
Data=pd.DataFrame(columns=['Country', 'Vaccine', 'Total_vaccine'])
For country in df["location"].unique():
    For vaccine in df["vaccine"].unique():
        Filtered_data = df[(df['location'] == country) & (df['vaccine'] == vaccine)]
        Total_count = filtered_data['total_vaccinations'].max()
        Data = pd.concat([data, pd.DataFrame({'Country': [country], 'Vaccine': [vaccine], 'Total_vaccine': [total_count]})], ignore_index=True)
```

Challenge involved in loading and preprocessing a covid-19 vaccine analysis dataset:

- **Data Source and Format**: The dataset may be scattered across multiple sources, in various formats, such as CSV, JSON, or APIs. Gathering and harmonizing these sources can be complex.

- **Data Transformation**: Preprocessing might involve scaling, normalization, or encoding categorical variables to make the data suitable for machine learning algorithms.

- **Data Integration**: Combining COVID-19 vaccine data with other relevant datasets, such as demographic or geographic information, can be complex but can provide valuable insights.

How to overcome the challenge involved in loading and preprocessing a covid-19 vaccine analysis data set:

- **Data Collection and Integration**: Collect data from reliable sources and ensure it's well-documented. Combine multiple data sources into a single, cohesive dataset.

➢ **Data Cleaning**:

- Handle missing data by imputation or removal, depending on the extent of missing values.
- Detect and address errors or inconsistencies.
- Handle duplicates and outliers appropriately.

Loading the dataset:

❖ **Data Source**: Identify the source from which you want to obtain the dataset. Common sources include government health agencies, research institutions, or publicly available datasets on platforms like Kaggle or GitHub.

❖ **Load the Dataset**: Use code to load the dataset into your chosen programming environment. For example, in Python with Pandas, you can use `pd.read_csv()` to load a CSV file.

Program:

```
data_2=pd.DataFrame(columns=['Country', 'Vaccin
e'])
data["Total_vaccine"] = pd.to_numeric(data["Tot
al_vaccine"], errors="coerce")
for country in data["Country"].unique():
    new_data = data[data["Country"] == country]
    max_vaccine = new_data.loc[new_data["Total_
vaccine"].idxmax(), "Vaccine"]
    data_2 = pd.concat([data_2, pd.DataFrame({'
Country': [country], 'Vaccine': [max_vaccine]})
], ignore_index=True)
```

```
data_2.head()
```

```
data.dropna(axis=0,inplace=True)
```

```
data.head(20)
```

Loading dataset:

Output:

	Country	Vaccine	Total_vaccine
0	Argentina	Moderna	6507561
1	Argentina	Oxford/AstraZeneca	25977231
2	Argentina	Sinopharm/Beijing	28322602
3	Argentina	Sputnik V	20405678
4	Argentina	CanSino	610540
5	Argentina	Pfizer/BioNTech	14681054
10	Austria	Moderna	1585063
11	Austria	Oxford/AstraZeneca	1588222
15	Austria	Pfizer/BioNTech	14584985
16	Austria	Johnson&Johnson	363548
17	Austria	Novavax	3682
20	Belgium	Moderna	4267394
21	Belgium	Oxford/AstraZeneca	2846716
25	Belgium	Pfizer/BioNTech	17451842
26	Belgium	Johnson&Johnson	425639
27	Belgium	Novavax	36
30	Bulgaria	Moderna	491663
31	Bulgaria	Oxford/AstraZeneca	478541
35	Bulgaria	Pfizer/BioNTech	2852218
36	Bulgaria	Johnson&Johnson	511702

Preprocessing the Dataset:

- **Data Cleaning**: Remove or handle missing values, duplicates, and outliers. Ensure data consistency and accuracy.
- **Data Transformation**: Perform necessary transformations, like normalization or scaling, to ensure all data is on the same scale.

Visualization and preprocessing of data:

In[1]:

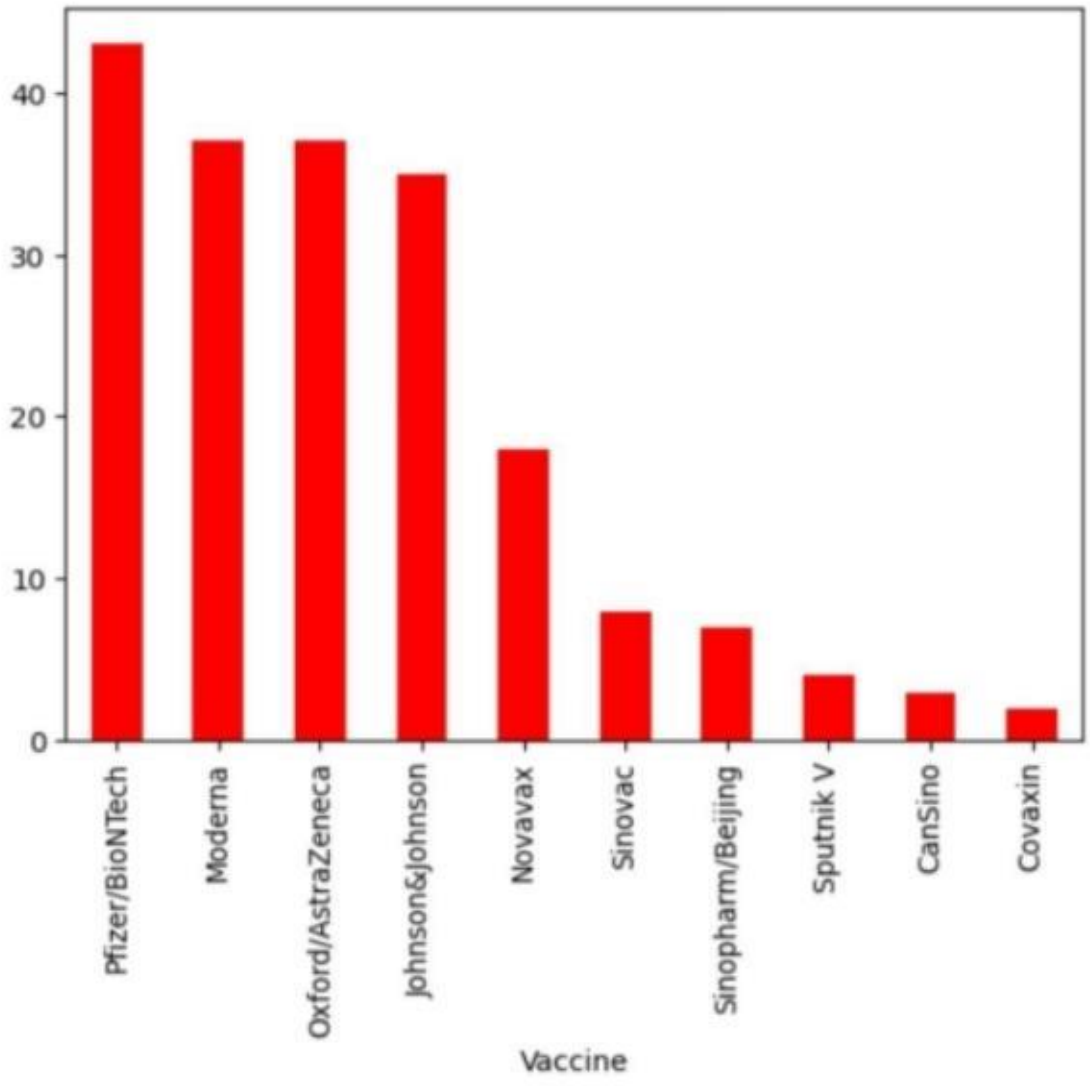
```
    Number _of_vaccines = data.groupby('Vacci
ne')['Country'].nunique()
```

In[1]:

```
  number_of_vaccines.sort_values(ascending=Fal
se).plot(kind="bar",color="r")
```

Out[1]:

```
<Axes: xlabel='Vaccine'>
```



In[2]:

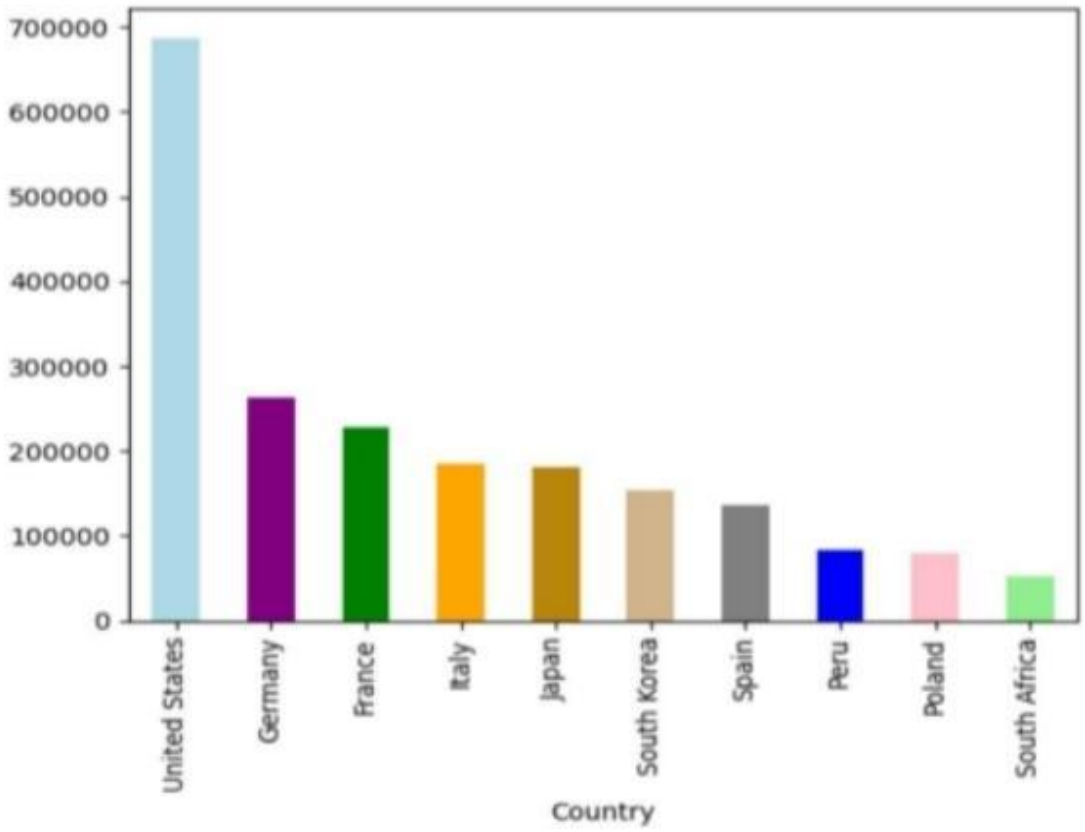
```
Dtfrm.set_index("Country",inplace=True)
```

In[2]:

```
   color=["Lightblue","Purple","Green","Orange"
,"darkgoldenrod","tan","Gray","Blue","Pink","Li
ghtgreen"]
dtfrm["average_vaccination_count"].sort_values(
ascending=False).head(10).plot(kind="bar",color
=color)
```

Out[2]:

```
<Axes: xlabel='Country'>
```



Phase 4

COVID-19 VACCINE ANALYSIS

Introduction:

In this part you will continue Building the project

It continues conducting the covid-19 vaccine analysis by :

- Performing exploratory data analysis
- Statistical analysis
- Visualization

PERFORMING EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis (EDA) for COVID-19 vaccines involves analyzing and visualizing data to gain insights and understand the distribution, patterns, and trends. Here are some steps you can follow:

1. Data Collection: Gather data related to COVID-19 vaccines. This could include vaccine distribution, administration, adverse events, and effectiveness data.

2. Data Cleaning: Check for missing values, duplicate records, and inconsistencies in the data. Clean the data to ensure it's ready for analysis.

Program:

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt

# Load COVID-19 data
data = pd.read_csv("covid19_data.csv")

# Basic EDA
print(data.head())  # Display the first few rows of the dataset
print(data.info())  # Display information about the dataset
```

STATISTICAL ANALYSIS:

Statistical analysis for COVID-19 vaccines involves using statistical methods to draw conclusions from data related to vaccine effectiveness, safety, and distribution. Here are some key statistical analyses commonly performed in this context:

1. Vaccine Efficacy Analysis:

- Calculate vaccine efficacy based on clinical trial data, comparing infection rates in the vaccinated and placebo groups.

- Use relative risk reduction, absolute risk reduction, and number needed to treat (NNT) metrics.

- Perform hypothesis tests to determine if observed differences are statistically significant.

2. Vaccine Safety Analysis:

- Analyze adverse event data to assess the safety of vaccines.

- Calculate the incidence of adverse events and their confidence intervals.

- Use statistical tests to identify any significant associations between adverse events and vaccination.

3. Distribution Analysis:

- Analyze vaccine distribution data to assess whether vaccines are being distributed equitably.

- Use descriptive statistics to summarize vaccine allocation and coverage across regions or demographic groups.

- Employ statistical tests to detect significant disparities in distribution.

VISUALIZATION:

Visualizations play a crucial role in conveying information about COVID-19 vaccines effectively. Here are some types of visualizations commonly used in the context of COVID-19 vaccines:

1. Vaccine Distribution Map: Create a map showing how vaccines are distributed across regions or countries. Use different colors or shades to represent the number of doses administered in each area.

2. Vaccination Progress Line Chart: Plot the daily or weekly progress of vaccination, showing the cumulative number of doses administered over time. This helps track the speed of the time.

3. Vaccine Type Pie Chart: Use a pie chart to illustrate the distribution of different vaccine types or manufacturers in a particular region or globally.

4. Bar Charts:

- Compare vaccination rates between different regions, age groups, or demographic categories using horizontal or vertical bar charts.

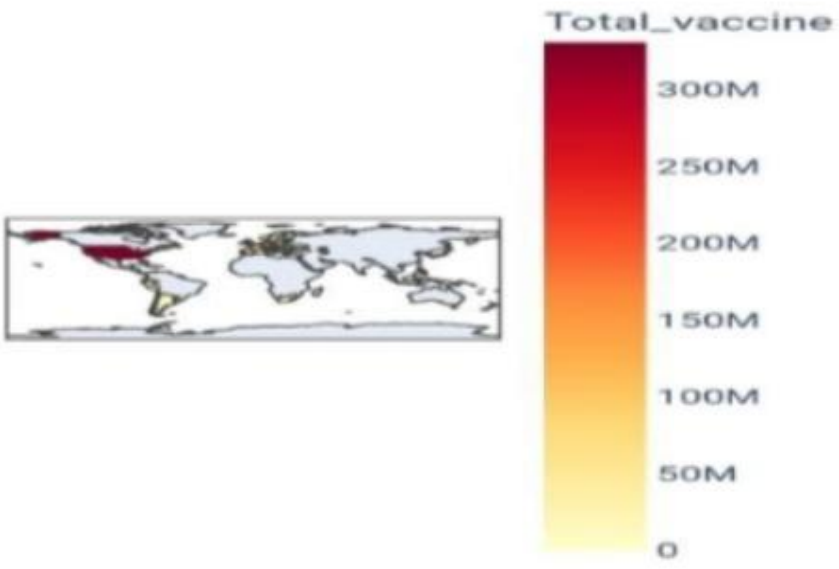
- Show the percentage of the population that has received one or both doses of the vaccine.

Program:

```
fig = px.choropleth(data_frame=dtfrm,
                    locations=dtfrm.i
ndex,
                    locationmode='cou
ntry names',
                    color='Total_vacc
ine',
                    color_continuous_
scale='YlOrRd',
                    title='Ülkelerde
Yapılan Biontech Aşıları')

fig.update_layout(title_x=0.5)
```

Output:



CONCLUSION

The above project contains the all four phases of Covid-19 vaccine analysis.

It contains,

- Data set
- Bar graph
- Problem definition
- Innovation

Here by the covid -19 vaccine analysis is concluded.

Thank you!