KINGSTON
ENGINEERING COLLEGE

COLLEGE CODE: 5113

APPLIED DATA SCIENCE

Project No.5- COVID -19 VACCINE ANALYSIS

BATCH MEMBERS:

1. K.HARISH-au511321104028-
   harishkumar251603@gmail.com

2.ASVINDHAN-au511321104008-
   asvindhanelangovan@gmail.com

3.HEMNATH AJAY-au511321104030-
   hemnathajay51@gmail.com

INTRODUCTION:

Analyzing COVID-19 vaccines is a critical component of the global response to the ongoing pandemic caused by the novel coronavirus, SARS-CoV-2. These vaccines have been developed at an unprecedented pace and are essential tools in controlling the spread of the virus and reducing the severity of the disease. Vaccine analysis involves a multifaceted approach that encompasses various aspects, including efficacy, safety, distribution, public acceptance, and their impact on the pandemic One of the primary aspects of analyzing COVID-19 vaccines is assessing their efficacy and effectiveness.

ABOUT THE DATA:

Where did we get the dataset?

Kaggle:

The dataset provided on Kaggle,

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

offers a valuable resource for our project aimed at forecasting covid-19 vaccine analysis.

Dataset Details:

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;

PERFORMING DIFFERENT ANALYSIS:

Performing different types of analysis on a dataset depends on the goals of your analysis and the nature of the data. Here are some common types of analysis that you might perform on a dataset:

Descriptive Analysis:

Summarize and describe the main characteristics of the dataset, including measures of central tendency, dispersion, and visualizations such as histograms, box plots, and bar charts.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) of COVID-19 vaccine data involves a systematic examination of various facets of vaccine distribution and efficacy. Researchers begin by collecting and cleansing data from various sources, such as government reports, clinical trials, and global vaccination databases. Key EDA tasks include summarizing demographic information of

vaccine recipients, assessing vaccine coverage across different regions, and tracking vaccination timelines.

Researchers also scrutinize adverse event reports to identify potential safety concerns and investigate disparities in vaccine distribution. Visualization tools like bar charts, heatmaps, and time series plots aid in spotting trends and patterns. EDA of COVID-19 vaccine data plays a crucial role in providing insights for public health decision-makers, enabling them to optimize vaccination campaigns, address equity issues, and continuously monitor vaccine performance and safety.

Statistical Analysis:

Statistical data analysis in COVID-19 vaccine research involves the application of advanced quantitative techniques to derive meaningful insights from vaccine-related data. Researchers employ statistical methods to assess vaccine efficacy through clinical trial results, calculating efficacy rates and confidence intervals. They also analyze large-scale vaccination datasets to understand factors influencing vaccine coverage and effectiveness, utilizing regression analyses, hypothesis testing, and survival analysis to identify significant associations. Additionally, statistical techniques are vital in evaluating vaccine safety by identifying adverse event signals, conducting risk-benefit assessments, and monitoring rare side effects. These

analyses are fundamental for evidence-based decision-making in vaccine distribution, regulation, and public health policy.

CODE:

Analyzing a COVID-19 dataset involves various tasks such as loading data, cleaning it, visualizing trends, performing exploratory data analysis and performing statistical analysis. Here's a Python code example that demonstrates how to perform basic COVID-19 data analysis using a sample dataset as provide for us.

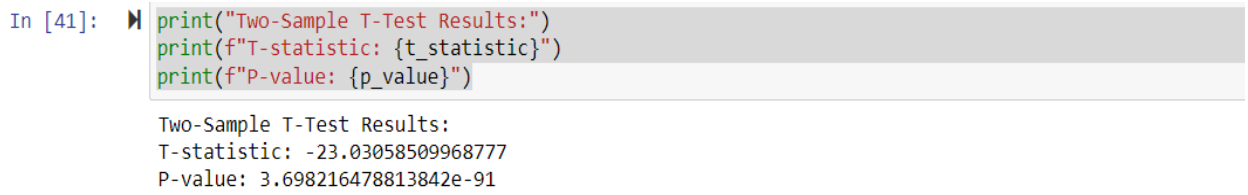
INITIAL LOADING AND CLEANING PROCESS ARE BEEN PERFORMED IN PHASE 3 ,LET'S PERFORM EXPLORATORY DATA ANALYSIS , STATISTICAL ANALYSIS AND DATA VISULATION:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats

# Load your dataset
data = pd.read_csv("C:/Users/haris/OneDrive/Desktop/country_vaccinations.csv
    (2)/country_vaccinations.csv")
```

CODE SECTION OF EXPLORATORY DATA ANALYSIS AND STATISTICAL ANALYSIS:

```
# Data types and missing values
data_info = data.info()
data.fillna(0, inplace=True)

# Example 1: Two-sample t-test between two groups (e.g., two countries)
group1 = data[data['country'] == 'Afghanistan']['total_vaccinations']
group2 = data[data['country'] == 'India']['total_vaccinations']
t_statistic, p_value = stats.ttest_ind(group1, group2)
print("Two-Sample T-Test Results:")
print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")
```

OUTPUT:




```
if p_value < 0.05:              # You can choose a significance level (e.g., 0.05)
    print("There is a significant difference between the two groups.")
else:
    print("There is no significant difference between the two groups.")
```

 # Example 2: One-way ANOVA to test differences among multiple groups (e.g., multiple countries)

```
groups = [data[data['country'] == 'Afghanistan']['daily_vaccinations'],
        data[data['country'] == 'Albania']['daily_vaccinations'],
        data[data['country'] == 'India']['daily_vaccinations']]
f_statistic, p_value = stats.f_oneway(*groups)
print("\nOne-Way ANOVA Results:")
print(f"F-statistic: {f_statistic}")
print(f"P-value: {p_value}")
if p_value < 0.05:  # You can choose a significance level (e.g., 0.05)
   print("There is a significant difference among the groups.")
else:
   print("There is no significant difference among the groups.")
```

OUTPUT:

In [42]:

```
if p_value < 0.05:  # You can choose a significance level (e.g., 0.05)
    print("There is a significant difference between the two groups.")
else:
    print("There is no significant difference between the two groups.")

# Example 2: One-way ANOVA to test differences among multiple groups (e.g., multiple countries)
groups = [data[data['country'] == 'Afghanistan']['daily_vaccinations'],
          data[data['country'] == 'Albania']['daily_vaccinations'],
          data[data['country'] == 'India']['daily_vaccinations']]

f_statistic, p_value = stats.f_oneway(*groups)

print("\nOne-Way ANOVA Results:")
print(f"F-statistic: {f_statistic}")
print(f"P-value: {p_value}")

if p_value < 0.05:  # You can choose a significance level (e.g., 0.05)
    print("There is a significant difference among the groups.")
else:
    print("There is no significant difference among the groups.")
```

```
There is a significant difference between the two groups.

One-Way ANOVA Results:
F-statistic: 1158.0734307553596
P-value: 6.482446870836429e-287
There is a significant difference among the groups.
```

 # Summary statistics

```
summary = data.describe()
print(summary)
```

```
In [9]: # Summary statistics
        summary = data.describe()
        print(summary)
```

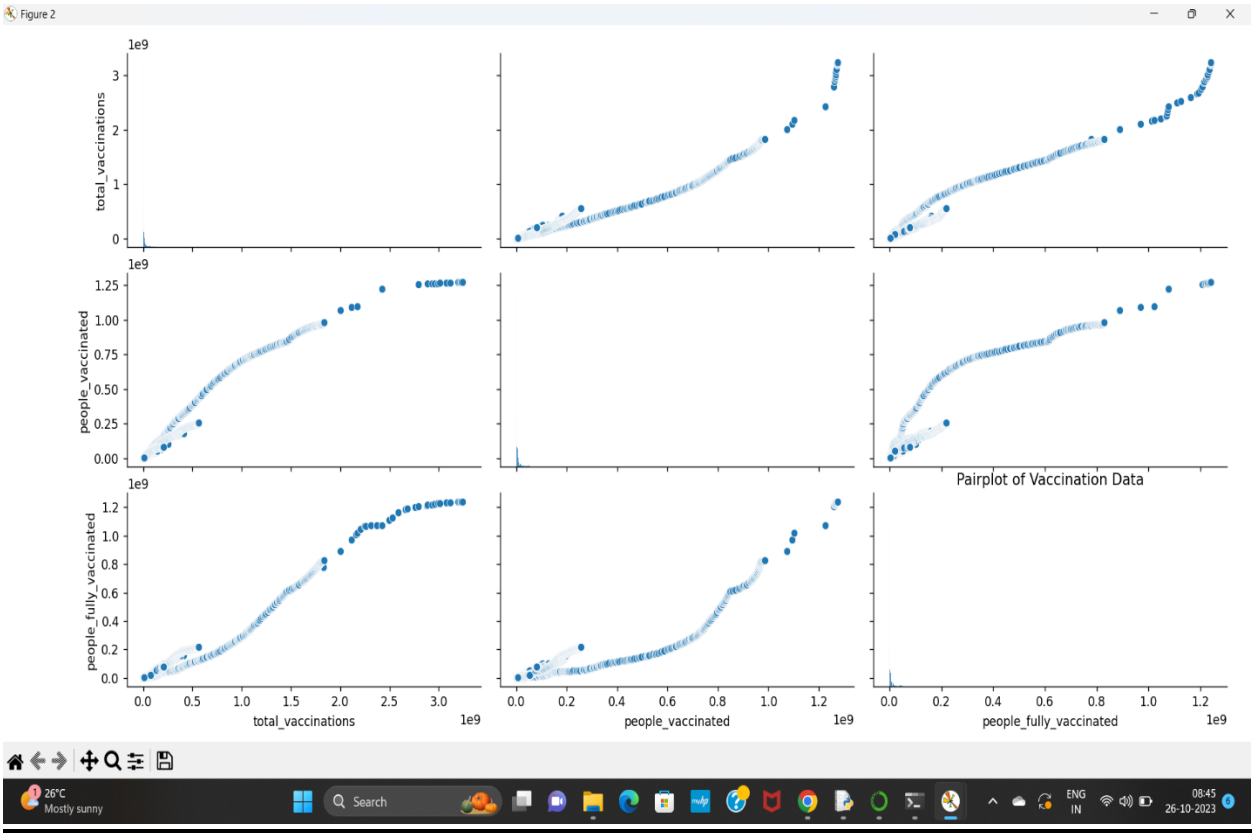
```
       total_vaccinations  people_vaccinated  people_fully_vaccinated  \
count        4.360700e+04       4.129400e+04              3.880200e+04
mean         4.592964e+07       1.770508e+07              1.413830e+07
std          2.246004e+08       7.078731e+07              5.713920e+07
min          0.000000e+00       0.000000e+00              1.000000e+00
25%          5.264100e+05       3.494642e+05              2.439622e+05
50%          3.590096e+06       2.187310e+06              1.722140e+06
75%          1.701230e+07       9.152520e+06              7.559870e+06
max          3.263129e+09       1.275541e+09              1.240777e+09

       daily_vaccinations_raw  daily_vaccinations  \
count            3.536200e+04        8.621300e+04
mean             2.705996e+05        1.313055e+05
std              1.212427e+06        7.682388e+05
min              0.000000e+00        0.000000e+00
25%              4.668000e+03        9.000000e+02
50%              2.530900e+04        7.343000e+03
75%              1.234925e+05        4.409800e+04
max              2.474100e+07        2.242429e+07

       total_vaccinations_per_hundred  people_vaccinated_per_hundred  \
count                    43607.000000                   41294.000000
mean                        80.188543                      40.927317
std                         67.913577                      29.290759
min                          0.000000                       0.000000
25%                         16.050000                      11.370000
50%                         67.520000                      41.435000
75%                        132.735000                      67.910000
max                        345.370000                     124.760000
```

CODE SECTION OF VISUALIZATION**:**

 # Data distribution and visualization

```
plt.figure(figsize=(12, 8))
sns.pairplot(data[['total_vaccinations', 'people_vaccinated',
    'people_fully_vaccinated']])
plt.title('Pairplot of Vaccination Data')
plt.show()
```

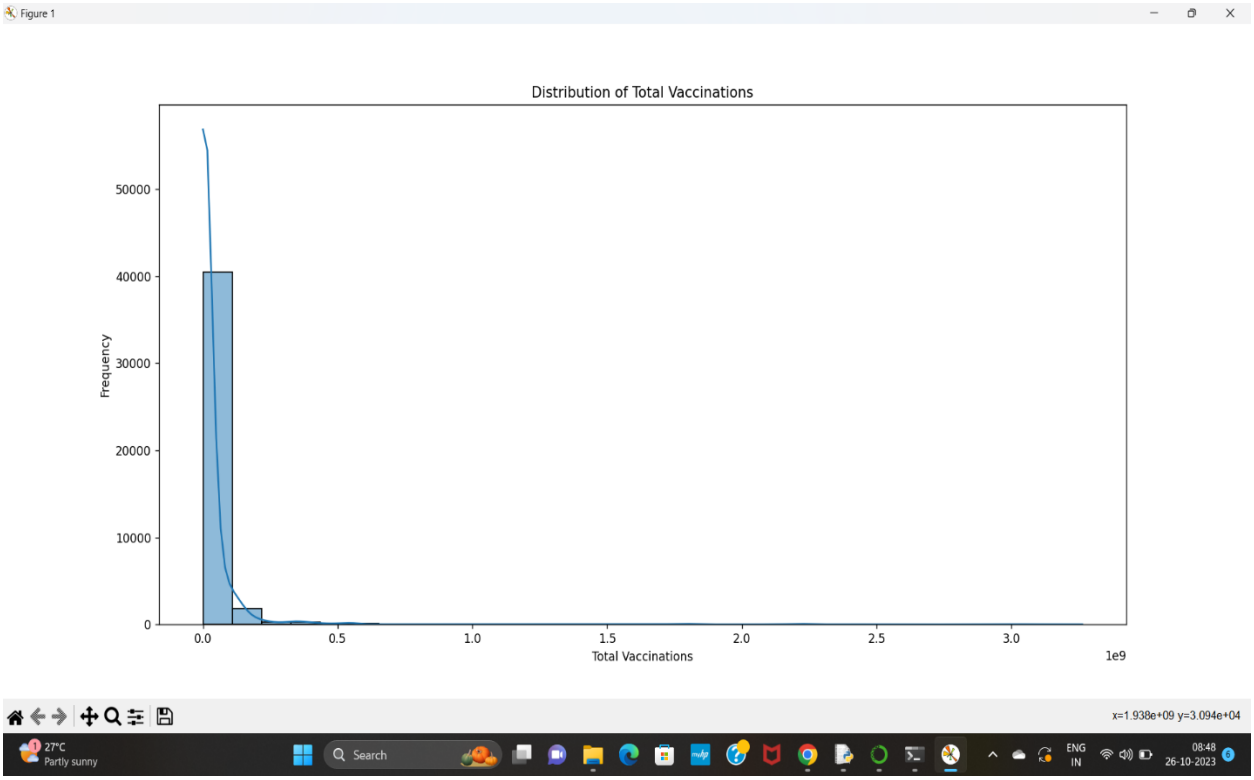
OUTPUT:



 # Histograms for selected columns

```
plt.figure(figsize=(12, 8))
sns.histplot(data['total_vaccinations'], kde=True, bins=30)
plt.title('Distribution of Total Vaccinations')
plt.xlabel('Total Vaccinations')
plt.ylabel('Frequency')
plt.show()
```

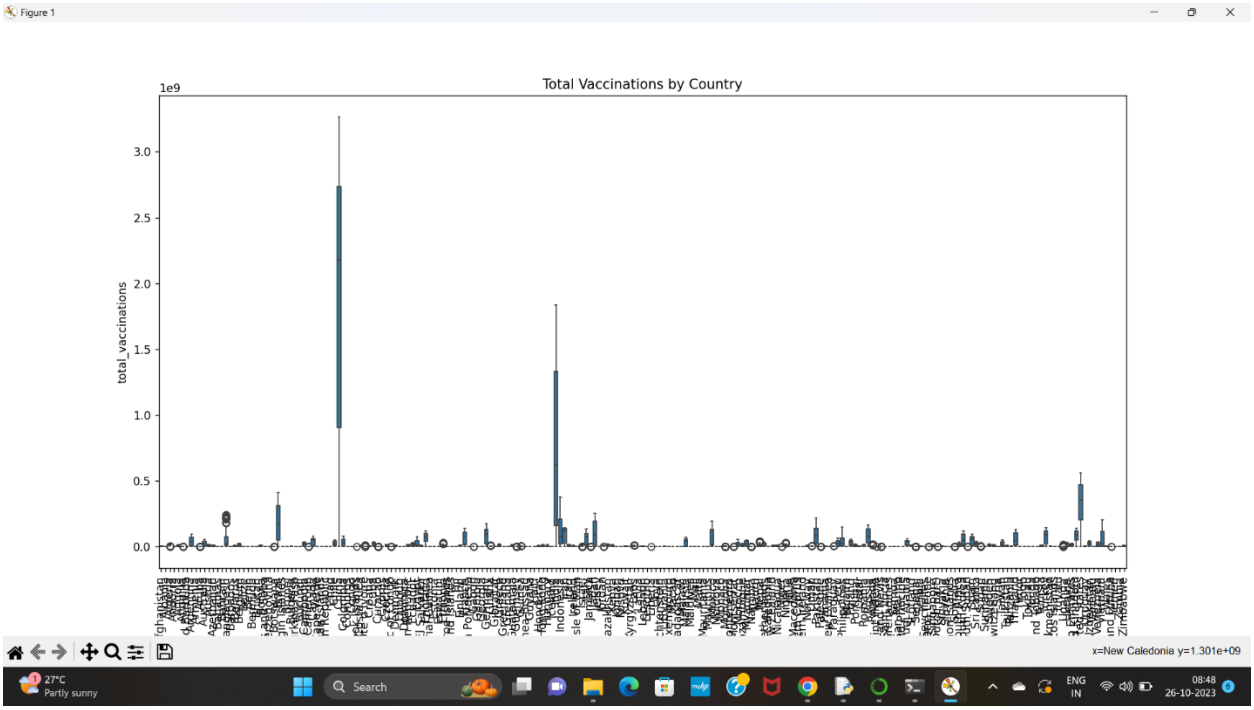
OUTPUT:



 # Boxplot for total vaccinations by country:

```
plt.figure(figsize=(12, 8))
sns.boxplot(x='country', y='total_vaccinations', data=data)
plt.title('Total Vaccinations by Country')
plt.xticks(rotation=90)
plt.show()
```

OUTPUT:



FULL CODE:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats

# Load your dataset
data = pd.read_csv("C:/Users/haris/OneDrive/Desktop/country_vaccinations.csv
    (2)/country_vaccinations.csv")
# Data types and missing values
data_info = data.info()
data.fillna(0, inplace=True)
```

 # Example 1: Two-sample t-test between two groups (e.g., two countries)

```
group1 = data[data['country'] == 'Afghanistan']['total_vaccinations']
group2 = data[data['country'] == 'India']['total_vaccinations']
t_statistic, p_value = stats.ttest_ind(group1, group2)
print("Two-Sample T-Test Results:")
print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")
if p_value < 0.05:              # You can choose a significance level (e.g., 0.05)
   print("There is a significant difference between the two groups.")
else:
   print("There is no significant difference between the two groups.")
```

 # Example 2: One-way ANOVA to test differences among multiple groups (e.g., multiple countries)

```
groups = [data[data['country'] == 'Afghanistan']['daily_vaccinations'],
        data[data['country'] == 'Albania']['daily_vaccinations'],
        data[data['country'] == 'India']['daily_vaccinations']]
f_statistic, p_value = stats.f_oneway(*groups)
print("\nOne-Way ANOVA Results:")
print(f"F-statistic: {f_statistic}")
print(f"P-value: {p_value}")
if p_value < 0.05:  # You can choose a significance level (e.g., 0.05)
   print("There is a significant difference among the groups.")
else:
   print("There is no significant difference among the groups.")
```

 # Summary statistics

```
summary = data.describe()
print(summary)
```

 # Data distribution and visualization

```
plt.figure(figsize=(12, 8))
sns.pairplot(data[['total_vaccinations', 'people_vaccinated',
    'people_fully_vaccinated']])
plt.title('Pairplot of Vaccination Data')
plt.show()
```

 # Histograms for selected columns

```
plt.figure(figsize=(12, 8))
sns.histplot(data['total_vaccinations'], kde=True, bins=30)
plt.title('Distribution of Total Vaccinations')
plt.xlabel('Total Vaccinations')
plt.ylabel('Frequency')
plt.show()
```

 # Boxplot for total vaccinations by country

```
plt.figure(figsize=(12, 8))
sns.boxplot(x='country', y='total_vaccinations', data=data)
plt.title('Total Vaccinations by Country')
plt.xticks(rotation=90)
plt.show()
```

CONCLUSION:

In summary, COVID-19 vaccine analysis, whether through exploratory data analysis (EDA) or statistical data analysis, is pivotal in our fight against the pandemic. EDA helps us uncover critical patterns in vaccine distribution, efficacy, and safety, highlighting the need for equitable vaccine access and assisting in the identification of adverse events. It is the lens through

which we gain a comprehensive understanding of the evolving vaccination landscape.

Statistical data analysis, in turn, provides the scientific rigor required to assess vaccine efficacy and safety with precision. It allows us to quantify the real-world impact of vaccines, set benchmarks, and identify influential factors. The combination of EDA and statistical analysis equips healthcare professionals and policymakers with evidence-based insights to make informed decisions, optimize vaccine deployment, and safeguard public health.

In the midst of a global health crisis, these analytical approaches are our compass, guiding us towards a safer and more secure future. They play an instrumental role in our collective efforts to combat COVID-19, ultimately contributing to the well-being of individuals and communities worldwide.