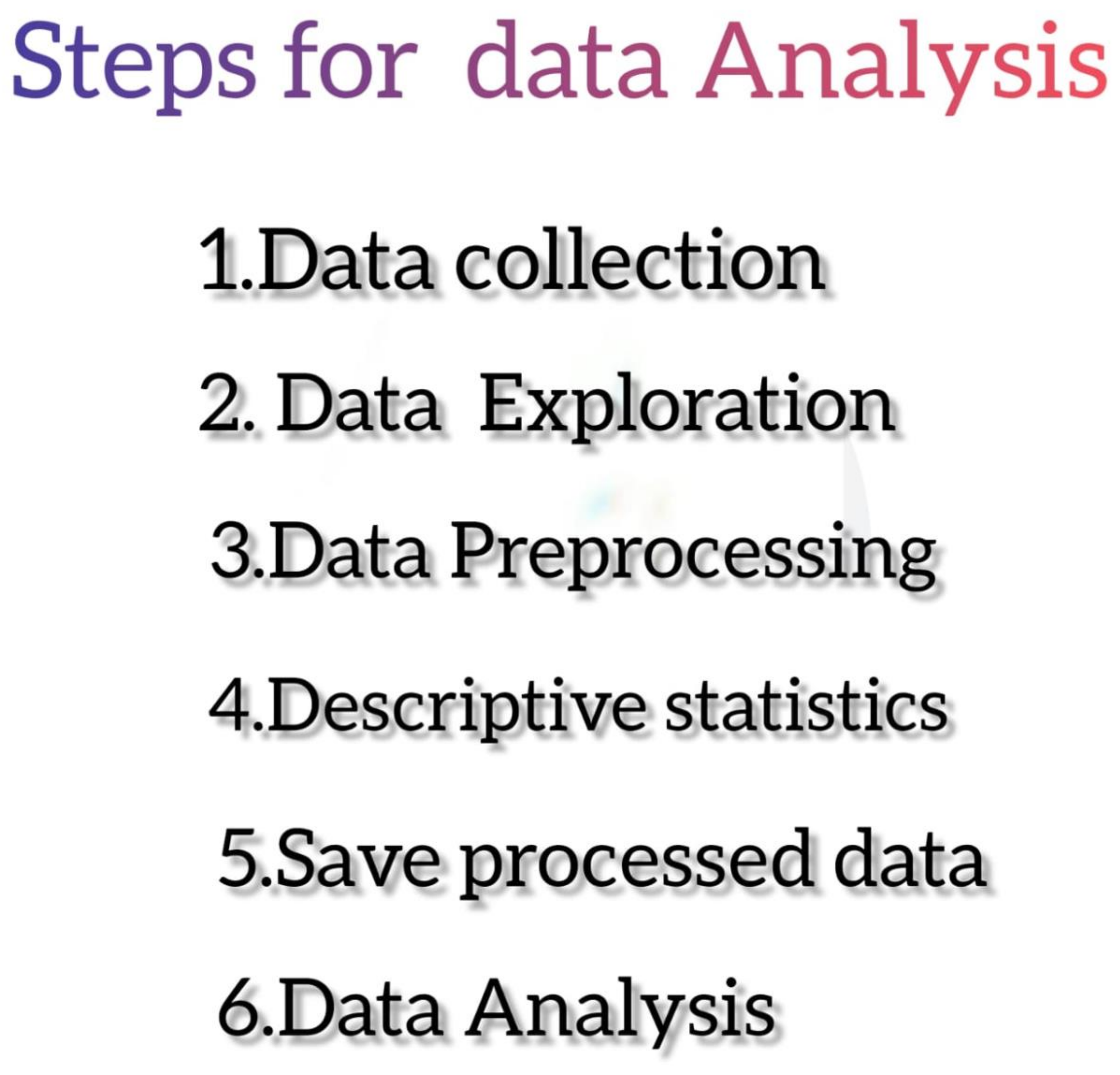
Collect And Preprocess The COVID Vaccine Analysis



Steps for data Analysis

1.Data collection

2. Data Exploration

3.Data Preprocessing

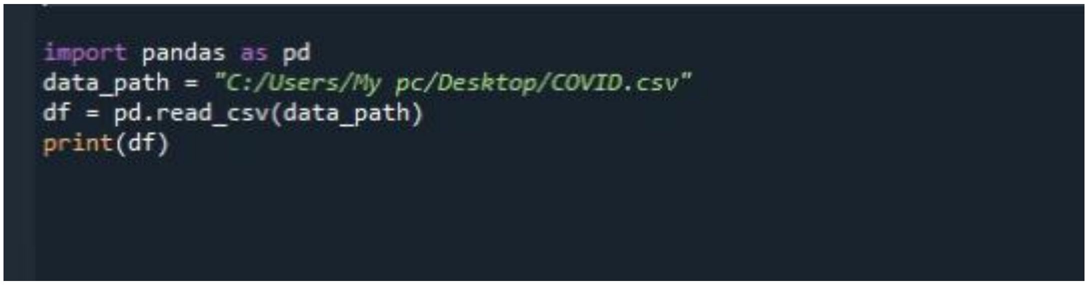
4.Descriptive statistics

5.Save processed data

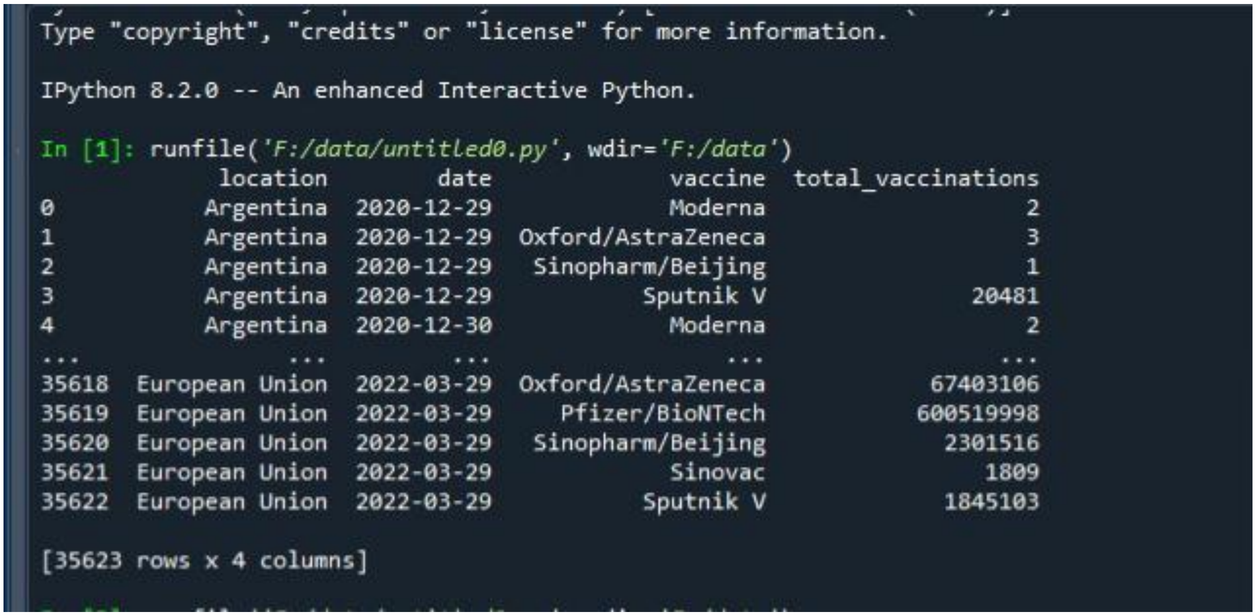
6.Data Analysis

Data Collection:

- Find a reliable source for COVID-19 vaccine data.Common sources include government health agencies, reputable research institutions, or datasets on platforms like Kaggle.
- Download or access the dataset in a format that's compatible with your analysis tools (e.g., CSV, Excel, JSON).



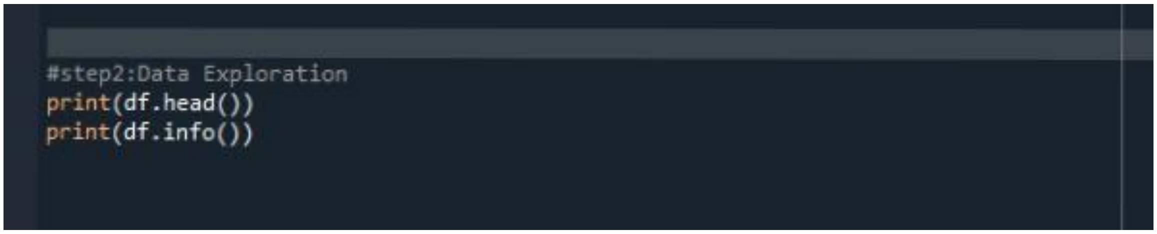
```
import pandas as pd
data_path = "C:/Users/My pc/Desktop/COVID.csv"
df = pd.read_csv(data_path)
print(df)
```



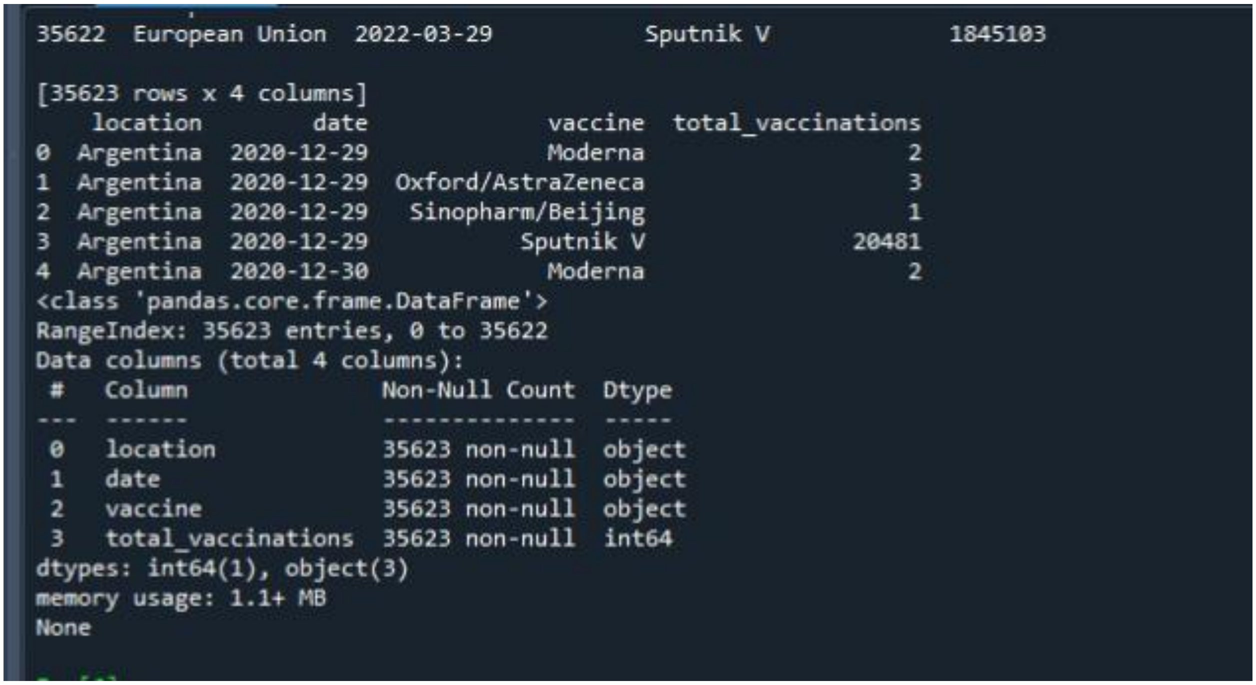
```
Type "copyright", "credits" or "license" for more information.

IPython 8.2.0 -- An enhanced Interactive Python.

In [1]: runfile('F:/data/untitled0.py', wdir='F:/data')
             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
...               ...         ...                 ...                 ...
35618  European Union  2022-03-29  Oxford/AstraZeneca            67403106
35619  European Union  2022-03-29     Pfizer/BioNTech           600519998
35620  European Union  2022-03-29   Sinopharm/Beijing             2301516
35621  European Union  2022-03-29             Sinovac                1809
35622  European Union  2022-03-29           Sputnik V             1845103

[35623 rows x 4 columns]
```

➢ Data Exploration:

❖ Load the dataset using a data manipulation library such as Pandas for Python or a tool that fits your preference.Examine the dataset’s structure, column names, and the type of information it contains.



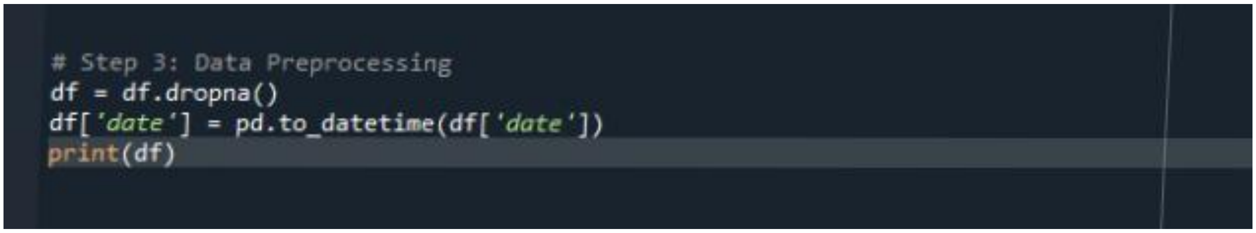
```
#step2:Data Exploration
print(df.head())
print(df.info())
```



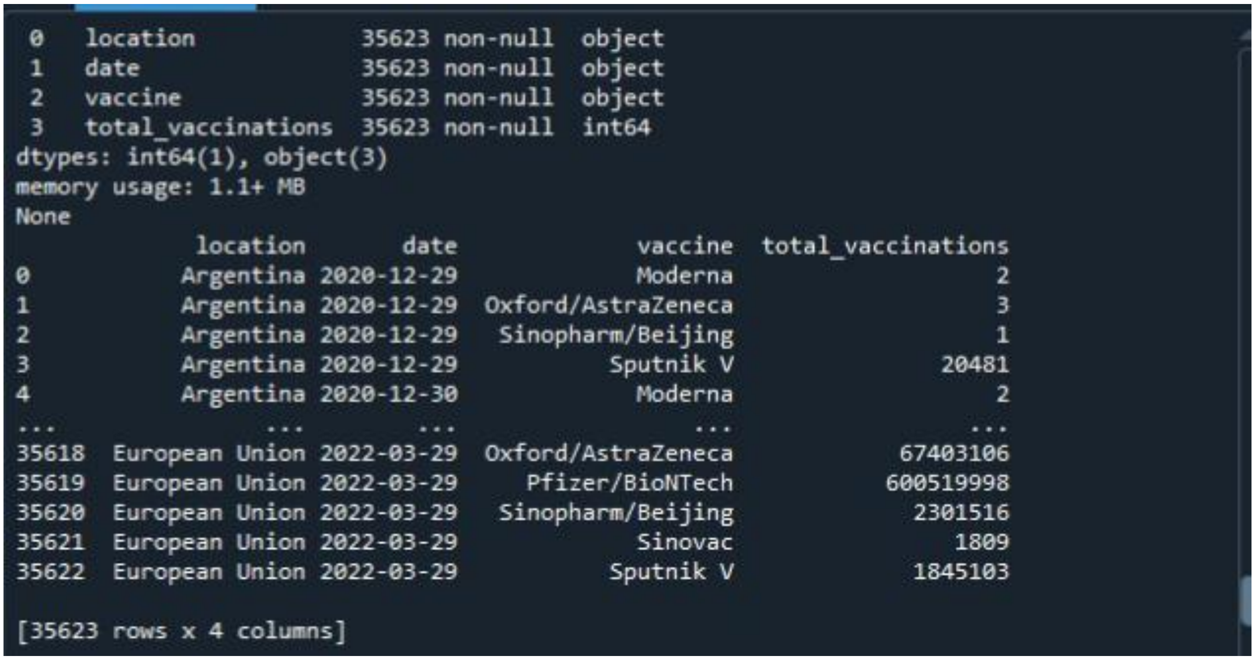
```
35622  European Union  2022-03-29          Sputnik V          1845103

[35623 rows x 4 columns]
    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
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35623 entries, 0 to 35622
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   location            35623 non-null  object
 1   date                35623 non-null  object
 2   vaccine             35623 non-null  object
 3   total_vaccinations  35623 non-null  int64
dtypes: int64(1), object(3)
memory usage: 1.1+ MB
None
```

➢ Data Preprocessing:

- ❖ Handle missing data: Check for missing values and decide on an appropriate strategy, like imputation or removal of incomplete rows.



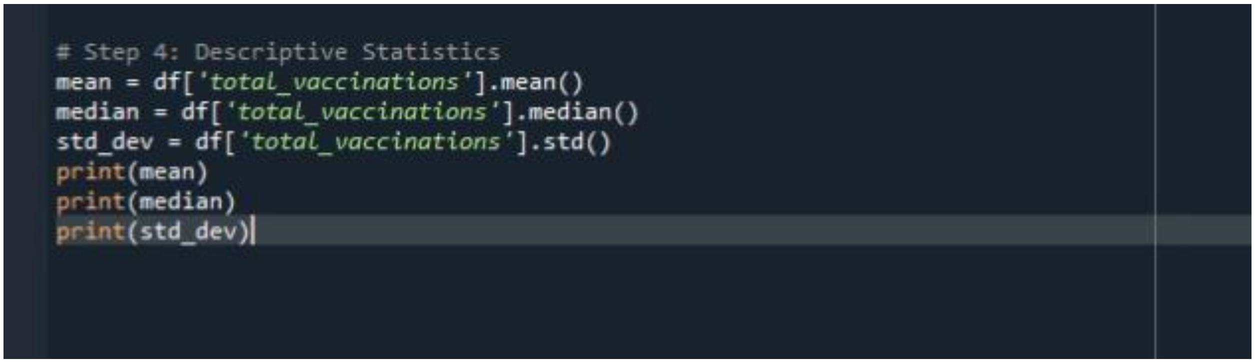
```python
# Step 3: Data Preprocessing
df = df.dropna()
df['date'] = pd.to_datetime(df['date'])
print(df)
```



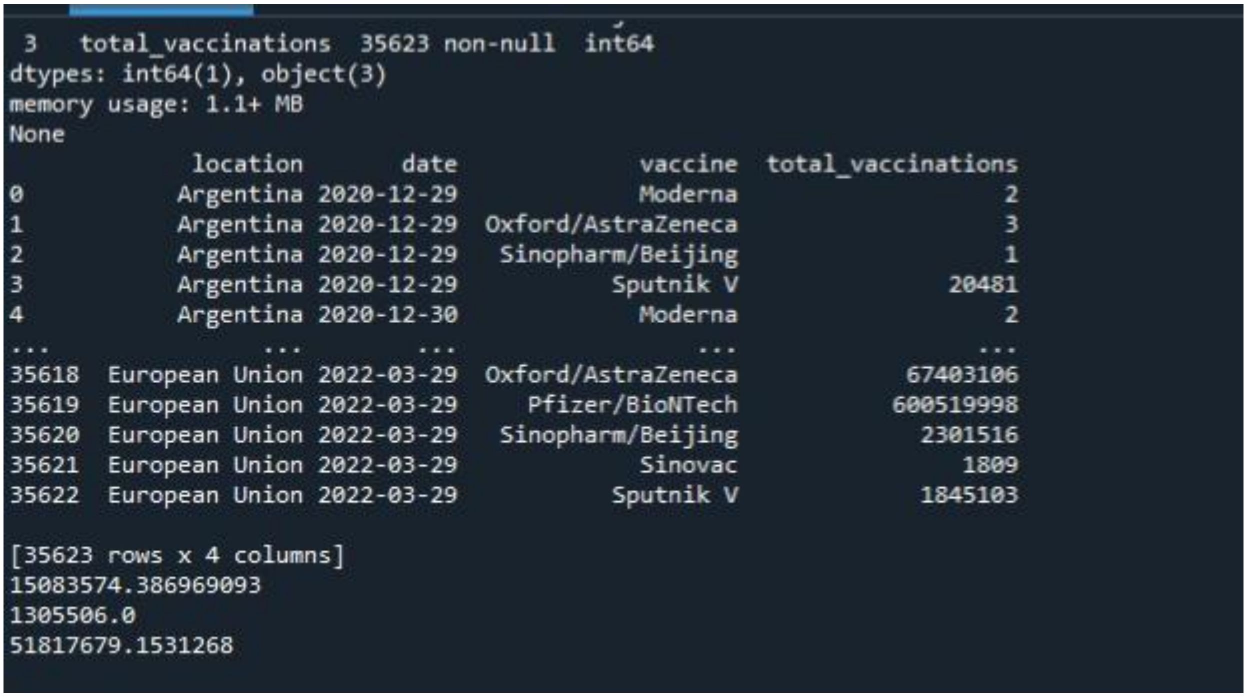
```
 0   location            35623 non-null  object
 1   date                35623 non-null  object
 2   vaccine             35623 non-null  object
 3   total_vaccinations  35623 non-null  int64
dtypes: int64(1), object(3)
memory usage: 1.1+ MB
None
             location        date             vaccine  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
...               ...         ...                 ...                 ...
35618  European Union  2022-03-29  Oxford/AstraZeneca            67403106
35619  European Union  2022-03-29     Pfizer/BioNTech           600519998
35620  European Union  2022-03-29   Sinopharm/Beijing             2301516
35621  European Union  2022-03-29             Sinovac                1809
35622  European Union  2022-03-29           Sputnik V             1845103

[35623 rows x 4 columns]
```

➢ Descriptive Statistics:

- ❖ Calculate basic statistics like mean, median, and standard deviation to understand the central tendencies and variability of the data.



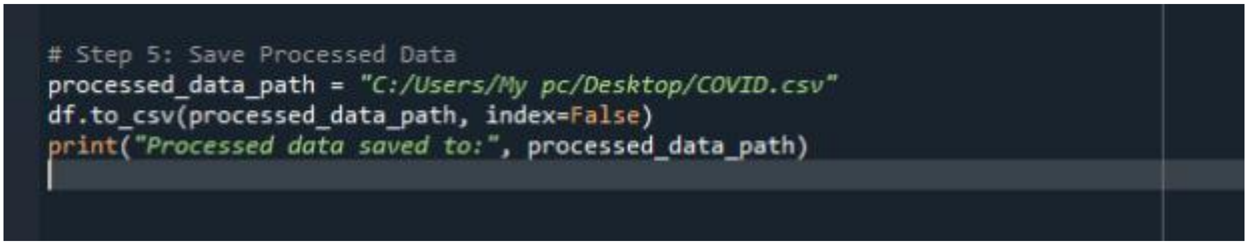
```
# Step 4: Descriptive Statistics
mean = df['total_vaccinations'].mean()
median = df['total_vaccinations'].median()
std_dev = df['total_vaccinations'].std()
print(mean)
print(median)
print(std_dev)
```



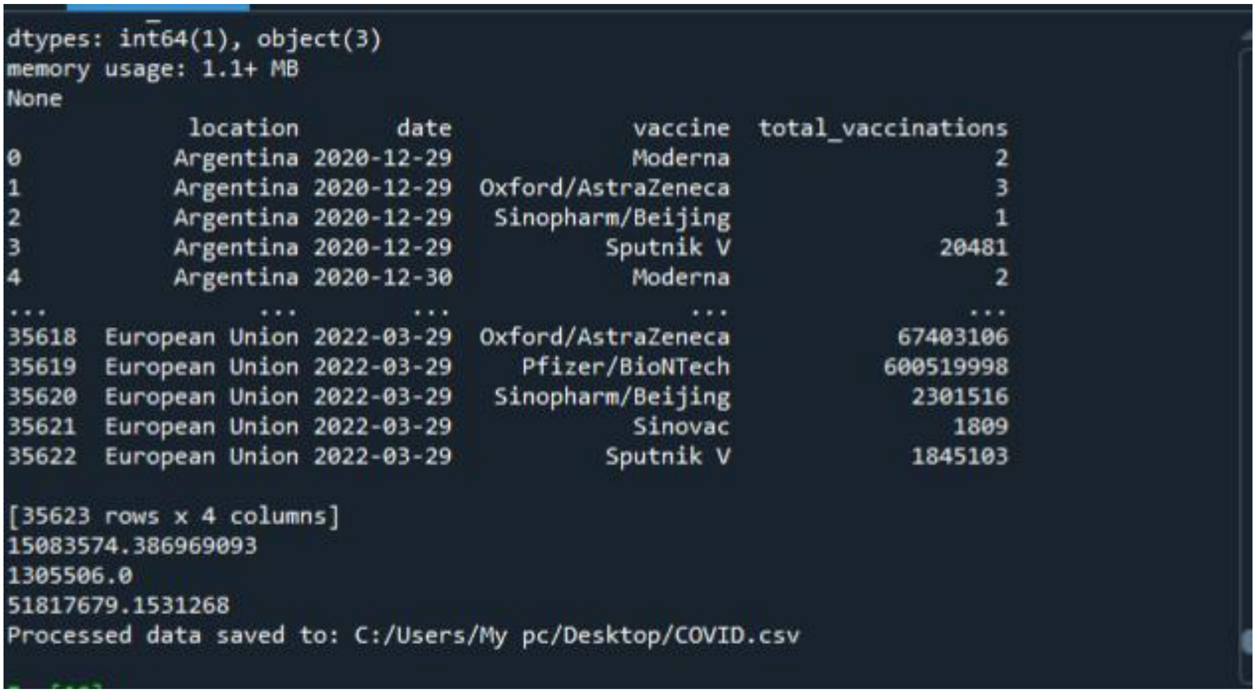
```
 3   total_vaccinations  35623 non-null  int64
dtypes: int64(1), object(3)
memory usage: 1.1+ MB
None
             location        date             vaccine  total_vaccinations
0           Argentina  2020-12-29             Moderna                   2
1           Argentina  2020-12-29  Oxford/AstraZeneca                   3
2           Argentina  2020-12-29   Sinopharm/Beijing                   1
3           Argentina  2020-12-29           Sputnik V               20481
4           Argentina  2020-12-30             Moderna                   2
...               ...         ...                 ...                 ...
35618  European Union  2022-03-29  Oxford/AstraZeneca            67403106
35619  European Union  2022-03-29     Pfizer/BioNTech           600519998
35620  European Union  2022-03-29   Sinopharm/Beijing             2301516
35621  European Union  2022-03-29             Sinovac                1809
35622  European Union  2022-03-29           Sputnik V             1845103

[35623 rows x 4 columns]
15083574.386969093
1305506.0
51817679.1531268
```

➢ Save Processed Data:

- ❖ After preprocessing, save the clean dataset to ensure you can work with it in future analysis without repeating these steps.



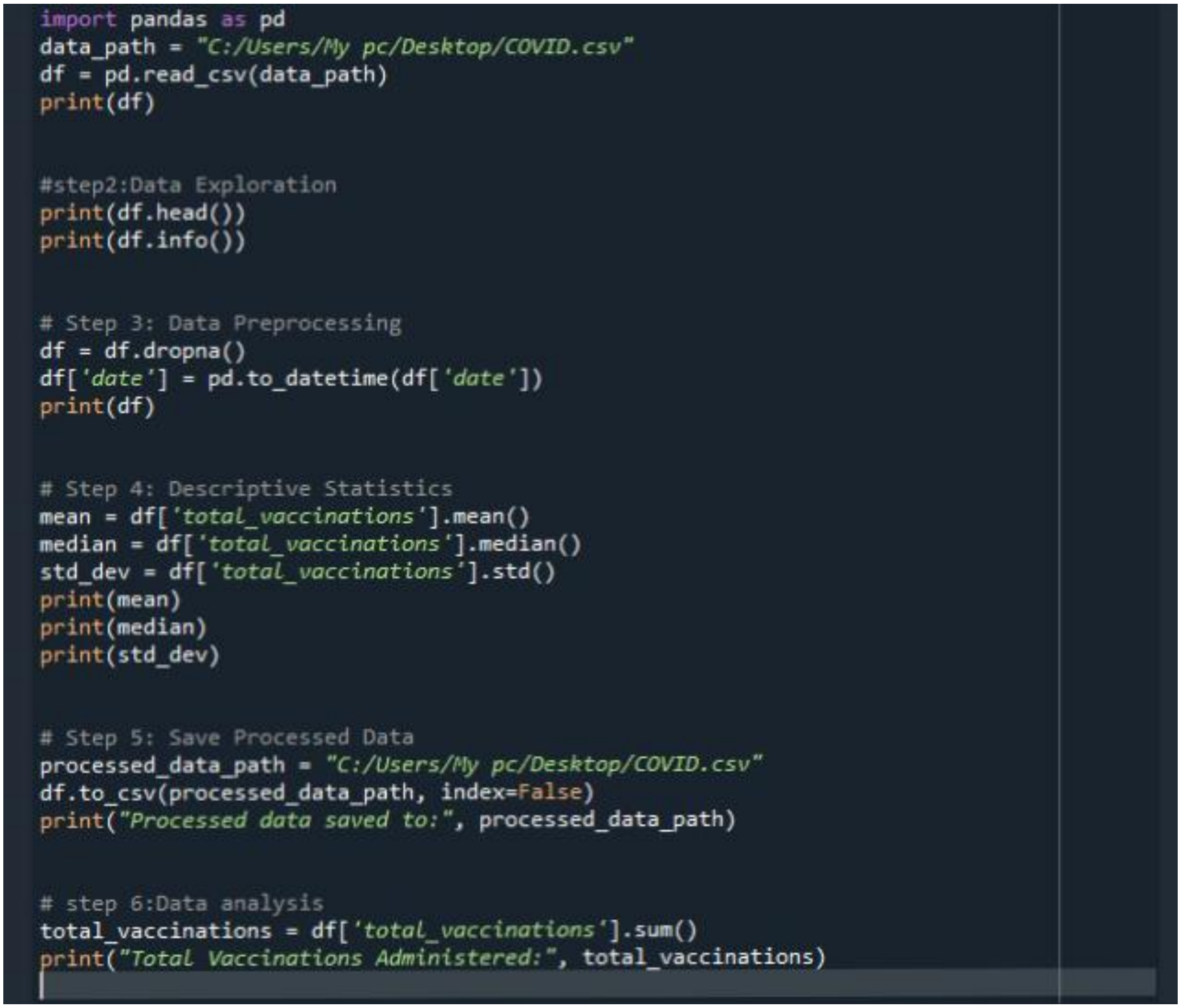
```
# Step 5: Save Processed Data
processed_data_path = "C:/Users/My pc/Desktop/COVID.csv"
df.to_csv(processed_data_path, index=False)
print("Processed data saved to:", processed_data_path)
```



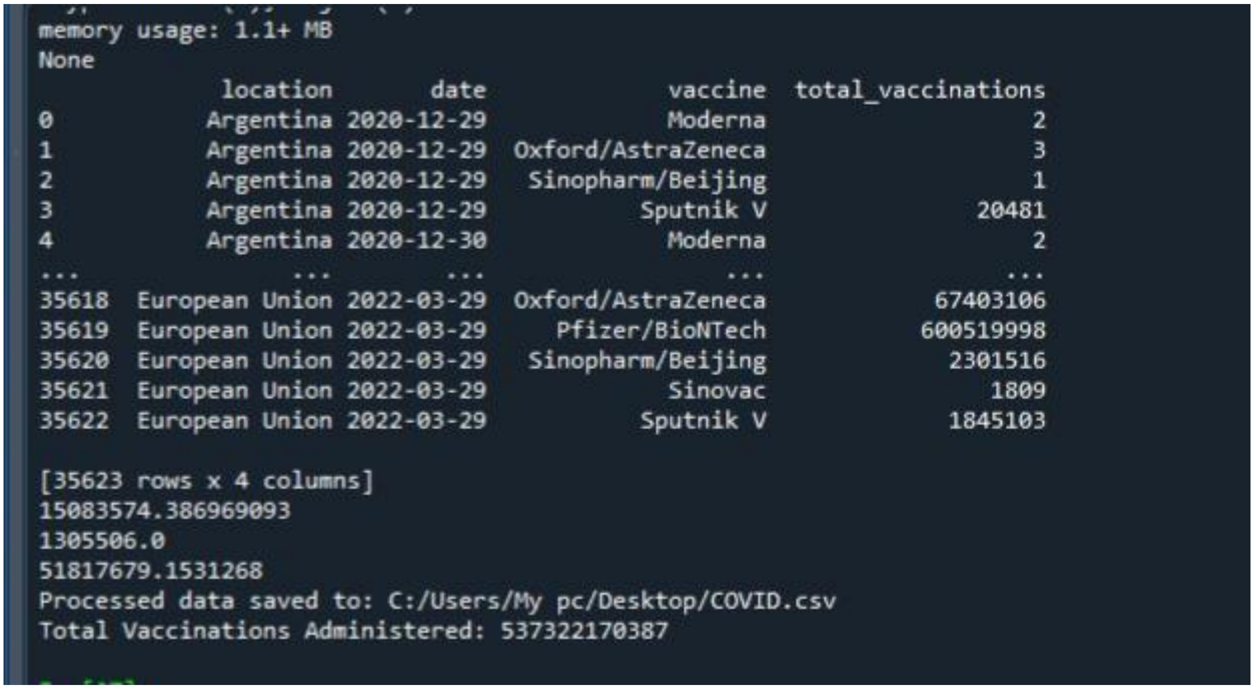
```
dtypes: int64(1), object(3)
memory usage: 1.1+ MB
None
             location        date             vaccine  total_vaccinations
0           Argentina  2020-12-29             Moderna                   2
1           Argentina  2020-12-29  Oxford/AstraZeneca                   3
2           Argentina  2020-12-29   Sinopharm/Beijing                   1
3           Argentina  2020-12-29           Sputnik V               20481
4           Argentina  2020-12-30             Moderna                   2
...               ...         ...                 ...                 ...
35618  European Union  2022-03-29  Oxford/AstraZeneca            67403106
35619  European Union  2022-03-29     Pfizer/BioNTech           600519998
35620  European Union  2022-03-29   Sinopharm/Beijing             2301516
35621  European Union  2022-03-29             Sinovac                1809
35622  European Union  2022-03-29           Sputnik V             1845103

[35623 rows x 4 columns]
15083574.386969093
1305506.0
51817679.1531268
Processed data saved to: C:/Users/My pc/Desktop/COVID.csv
```

➢ Data Analysis:

- ❖ Once your data is preprocessed, you can start your analysis, which could include trends, correlations, and more, depending on your specific research questions.



```
import pandas as pd
data_path = "C:/Users/My pc/Desktop/COVID.csv"
df = pd.read_csv(data_path)
print(df)


#step2:Data Exploration
print(df.head())
print(df.info())


# Step 3: Data Preprocessing
df = df.dropna()
df['date'] = pd.to_datetime(df['date'])
print(df)


# Step 4: Descriptive Statistics
mean = df['total_vaccinations'].mean()
median = df['total_vaccinations'].median()
std_dev = df['total_vaccinations'].std()
print(mean)
print(median)
print(std_dev)


# Step 5: Save Processed Data
processed_data_path = "C:/Users/My pc/Desktop/COVID.csv"
df.to_csv(processed_data_path, index=False)
print("Processed data saved to:", processed_data_path)


# step 6:Data analysis
total_vaccinations = df['total_vaccinations'].sum()
print("Total Vaccinations Administered:", total_vaccinations)
```



```
memory usage: 1.1+ MB
None
            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
...              ...         ...                 ...                 ...
35618  European Union 2022-03-29  Oxford/AstraZeneca            67403106
35619  European Union 2022-03-29     Pfizer/BioNTech           600519998
35620  European Union 2022-03-29   Sinopharm/Beijing             2301516
35621  European Union 2022-03-29             Sinovac                1809
35622  European Union 2022-03-29           Sputnik V             1845103

[35623 rows x 4 columns]
15083574.386969093
1305506.0
51817679.1531268
Processed data saved to: C:/Users/My pc/Desktop/COVID.csv
Total Vaccinations Administered: 537322170387
```