COVID-19 DATA ANALYSIS USING PYTHON AND POWER BI

Introduction

The COVID-19 Data Analysis project provides an in-depth examination of the COVID-19 pandemic's impact up to April 29, 2020. The project focuses on various aspects such as confirmed cases, deaths, and recovered cases across different regions. The goal is to present the findings through detailed and interactive visualizations.

Python

Python is a versatile programming language widely used in data analysis projects. It allows users to manipulate, analyze, and visualize data efficiently. Python's libraries such as Pandas, NumPy, and Matplotlib provide powerful tools for data cleaning, transformation, and visualization. By leveraging Python, analysts can perform complex operations such as filtering, grouping, and aggregating data effectively.

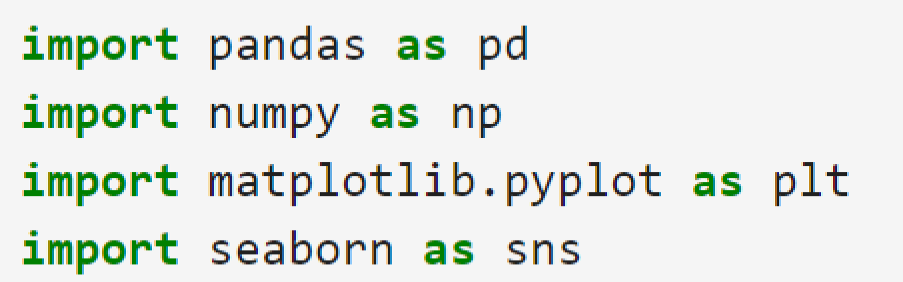
For this project, Python was used to preprocess and analyze the COVID-19 dataset. The following problem statements were addressed using Python Jupyter Notebook.

Jupyter Notebook

Jupyter Notebook is an open-source web application widely used for interactive computing, particularly in data science, machine learning, and scientific research. It supports live code execution, equations, visualizations, and narrative text in various programming languages, including Python, R, and Julia. Key features include rich text formatting with markdown, LaTeX support for mathematical equations, and integration with powerful data visualization libraries like Matplotlib and Seaborn. Its modular code organization and support for inline visualizations make it ideal for exploratory data analysis and prototyping. Additionally, Jupyter Notebooks facilitate collaboration and sharing, with tools like JupyterHub and Google Colab, enhancing productivity in data-driven projects.

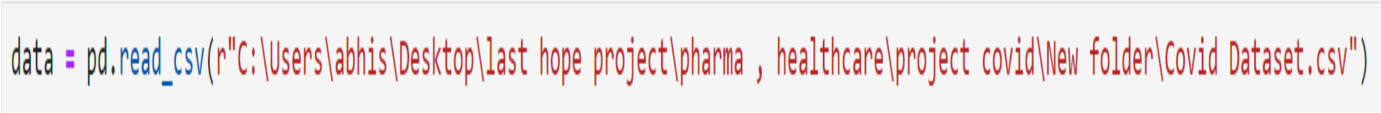
Problem Statements code and Solutions

Importing important libraries



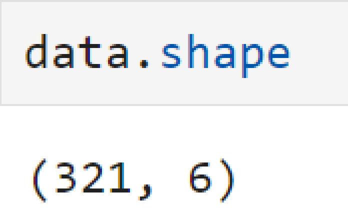
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Importing dataset :



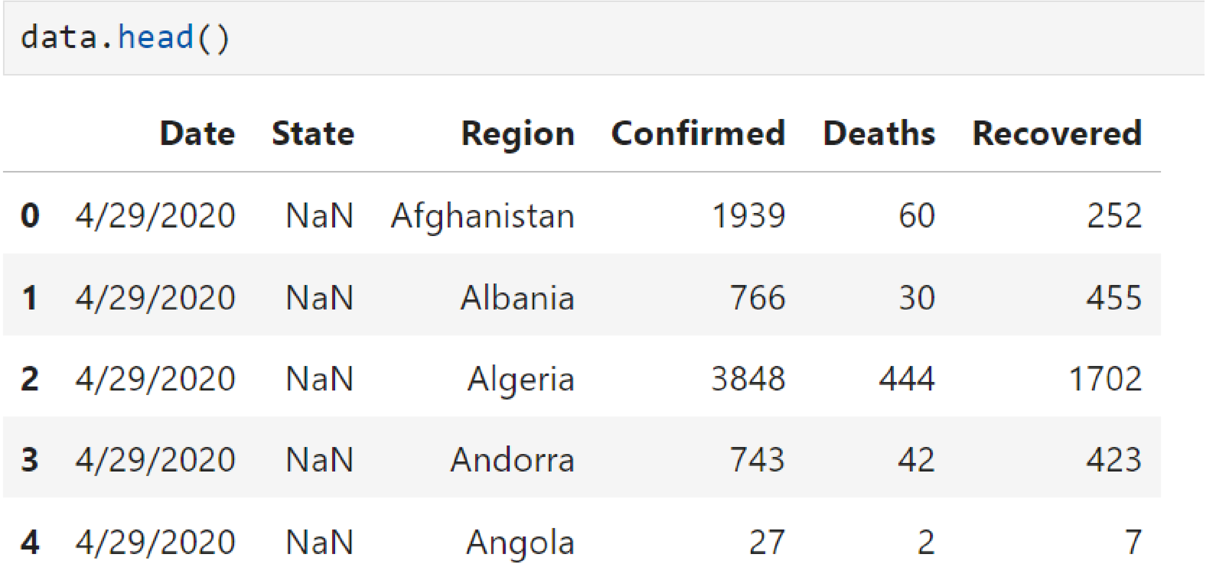
```
data = pd.read_csv(r"C:\Users\abhis\Desktop\last hope project\pharma , healthcare\project covid\New folder\Covid Dataset.csv")
```

Data cleaning :



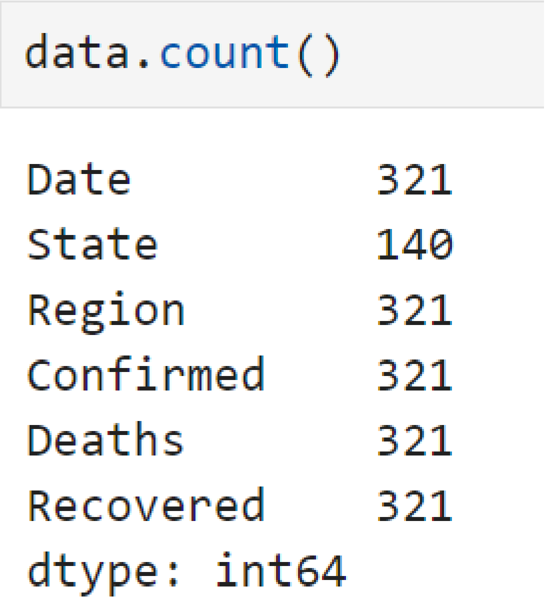
```
data.shape
```

```
(321, 6)
```



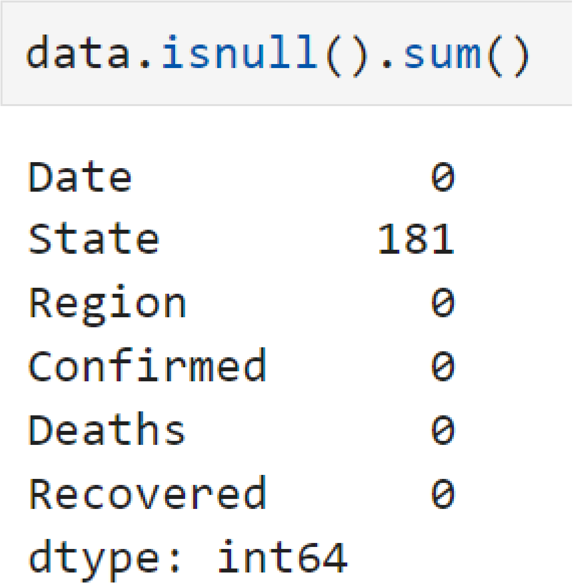
```
data.head()
```

	Date	State	Region	Confirmed	Deaths	Recovered
0	4/29/2020	NaN	Afghanistan	1939	60	252
1	4/29/2020	NaN	Albania	766	30	455
2	4/29/2020	NaN	Algeria	3848	444	1702
3	4/29/2020	NaN	Andorra	743	42	423
4	4/29/2020	NaN	Angola	27	2	7



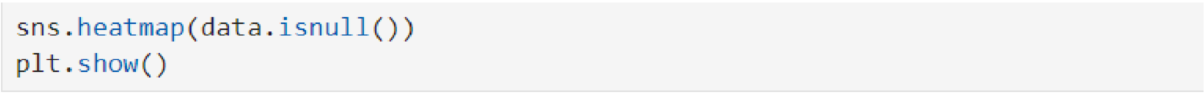
```
data.count()
```

```
Date         321
State        140
Region       321
Confirmed    321
Deaths       321
Recovered    321
dtype: int64
```

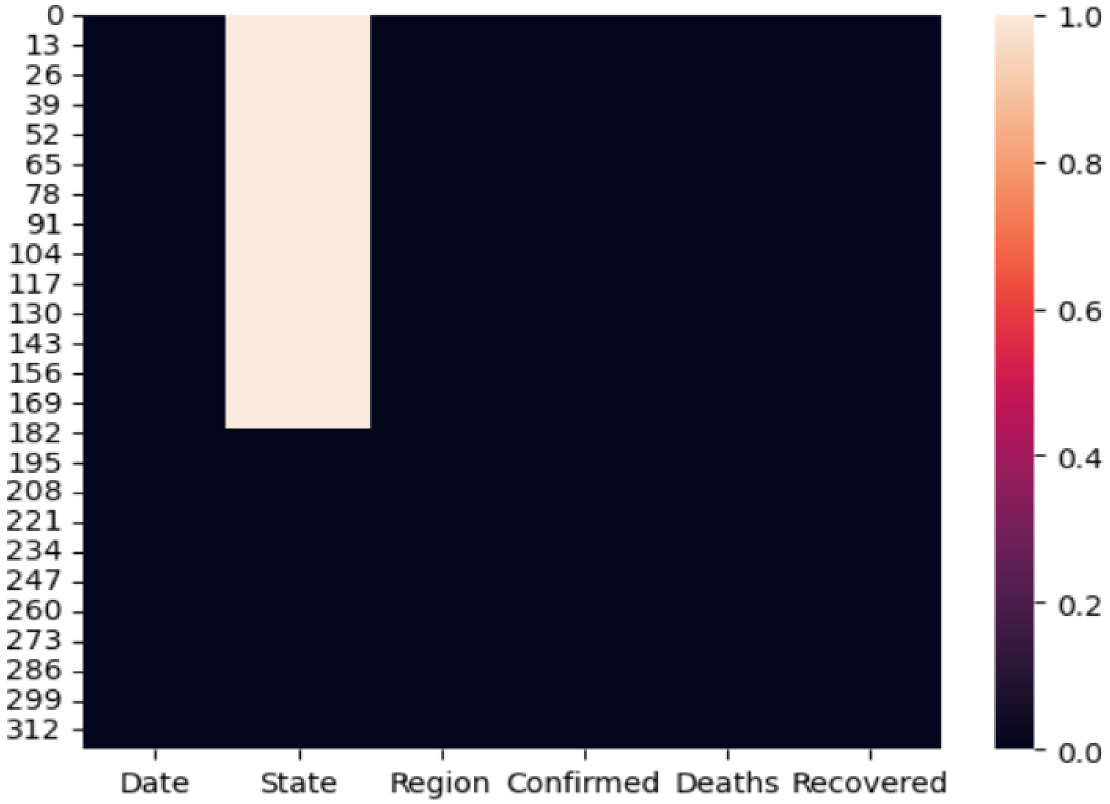


```
data.isnull().sum()
```

```
Date           0
State        181
Region         0
Confirmed      0
Deaths         0
Recovered      0
dtype: int64
```

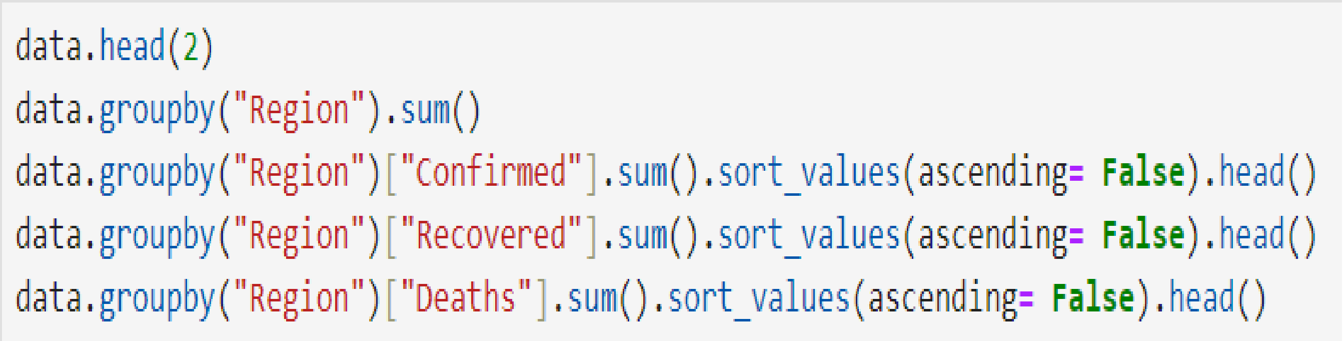


```
sns.heatmap(data.isnull())
plt.show()
```



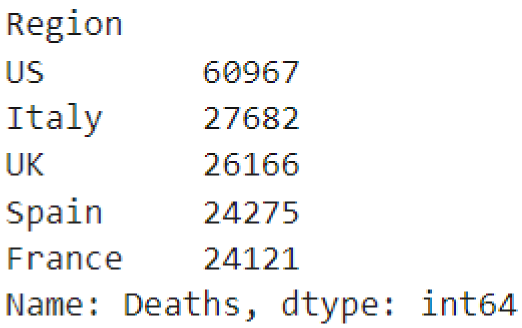
Question 1: Show the top 5 highest number of confirmed, death, and recovered cases in each region.

Code :



```
data.head(2)
data.groupby("Region").sum()
data.groupby("Region")["Confirmed"].sum().sort_values(ascending= False).head()
data.groupby("Region")["Recovered"].sum().sort_values(ascending= False).head()
data.groupby("Region")["Deaths"].sum().sort_values(ascending= False).head()
```

Output :

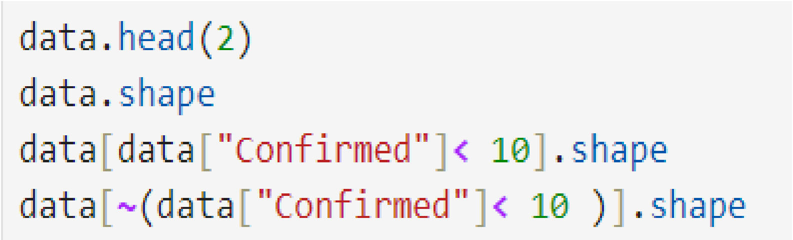


```
Region
US          60967
Italy       27682
UK          26166
Spain       24275
France      24121
Name: Deaths, dtype: int64
```

- Here are the list of top 5 results of death, confirmed and recovered cased of covid cases in each region.

Question 2: Remove all the records where confirmed cases are less than 10.

Output :

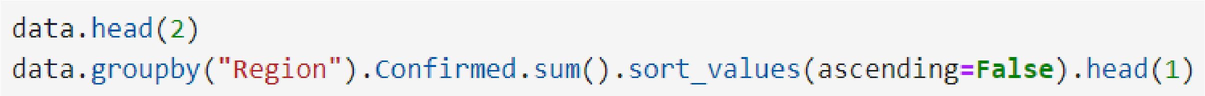


```
data.head(2)
data.shape
data[data["Confirmed"]< 10].shape
data[~(data["Confirmed"]< 10 )].shape
```

- The following cases are removed where the confirmed cases are less then 10

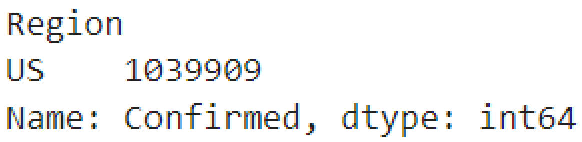
Question 3: In which region were the maximum number of confirmed cases recorded?

Code :



```
data.head(2)
data.groupby("Region").Confirmed.sum().sort_values(ascending=False).head(1)
```

Output :

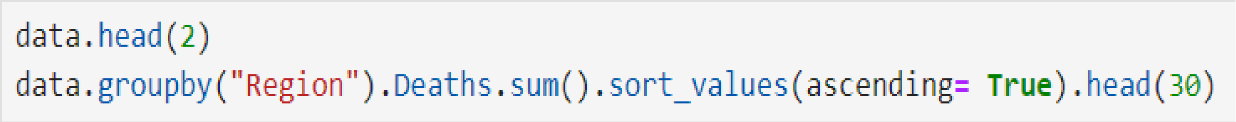


```
Region
US     1039909
Name: Confirmed, dtype: int64
```

- The US is the region where the maximum number of confirmed cases are recorded which is 1039909.

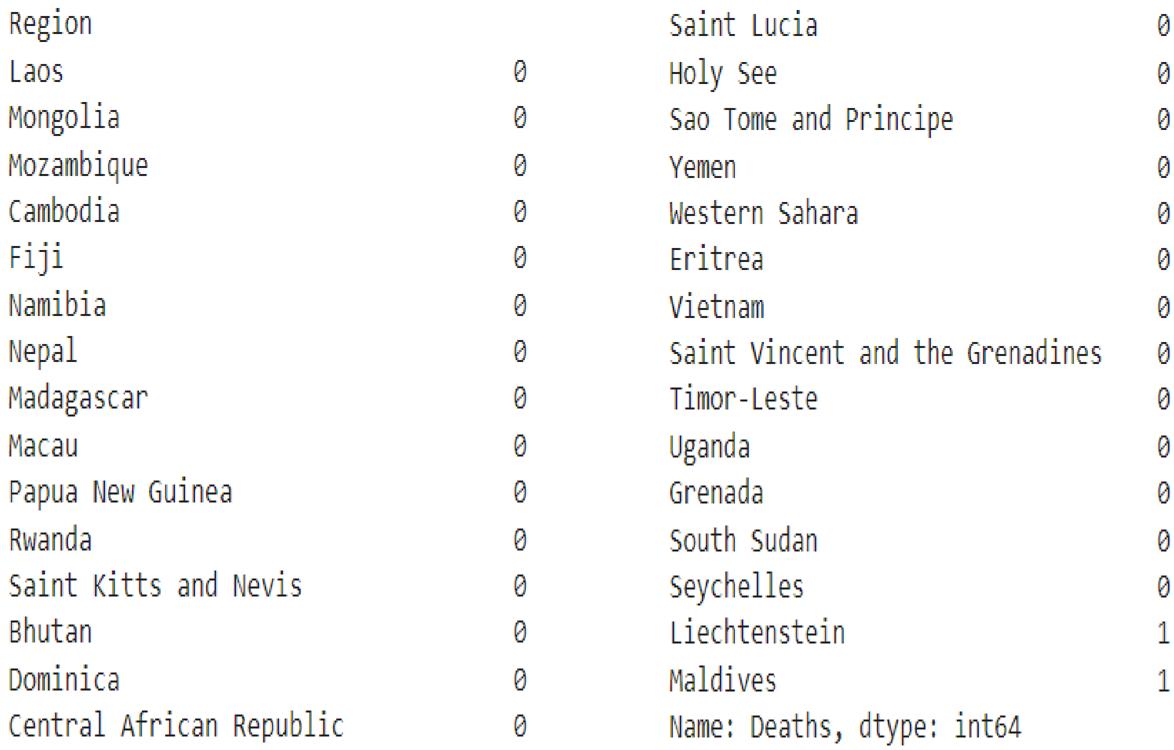
Question 4: In which region were the minimum number of death cases recorded?

Code :



```
data.head(2)
data.groupby("Region").Deaths.sum().sort_values(ascending= True).head(30)
```

Output :

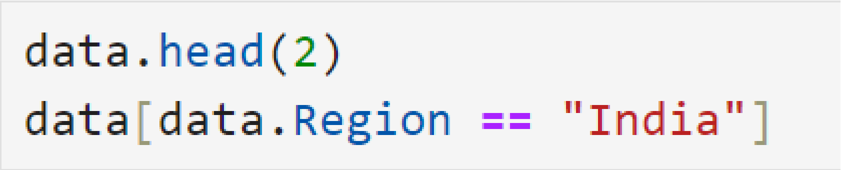


```
Region                                    Saint Lucia                         0
Laos                              0       Holy See                            0
Mongolia                          0       Sao Tome and Principe               0
Mozambique                        0       Yemen                               0
Cambodia                          0       Western Sahara                      0
Fiji                              0       Eritrea                             0
Namibia                           0       Vietnam                             0
Nepal                             0       Saint Vincent and the Grenadines    0
Madagascar                        0       Timor-Leste                         0
Macau                             0       Uganda                              0
Papua New Guinea                  0       Grenada                             0
Rwanda                            0       South Sudan                         0
Saint Kitts and Nevis             0       Seychelles                          0
Bhutan                            0       Liechtenstein                       1
Dominica                          0       Maldives                            1
Central African Republic          0       Name: Deaths, dtype: int64
```

- Here we can see that the minimum number of the death cases are zeros and 1 here is the list of regions where the condition is being fully satisfied.

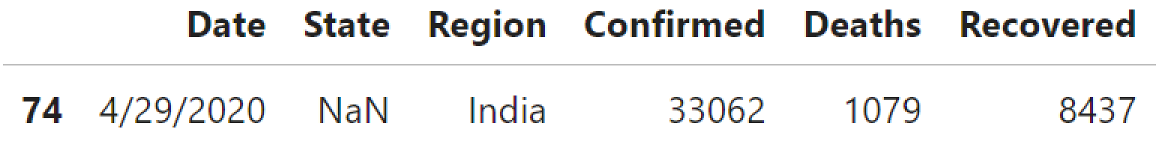
Question 5: How many confirmed, death, and recovered cases were reported from India till April 29, 2020?

Code :



```
data.head(2)
data[data.Region == "India"]
```

Output :

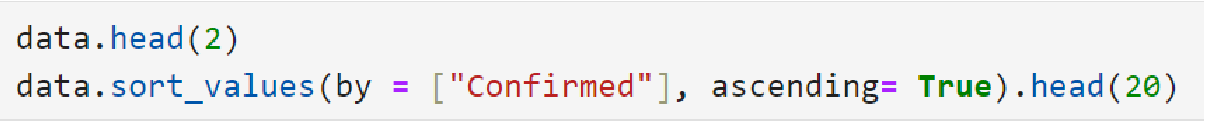


	Date	State	Region	Confirmed	Deaths	Recovered
74	4/29/2020	NaN	India	33062	1079	8437

- Here we can clearly see the number of confirmed cases are 33062 , the number of deaths are 1079 and the number of recovered cases are 8437.

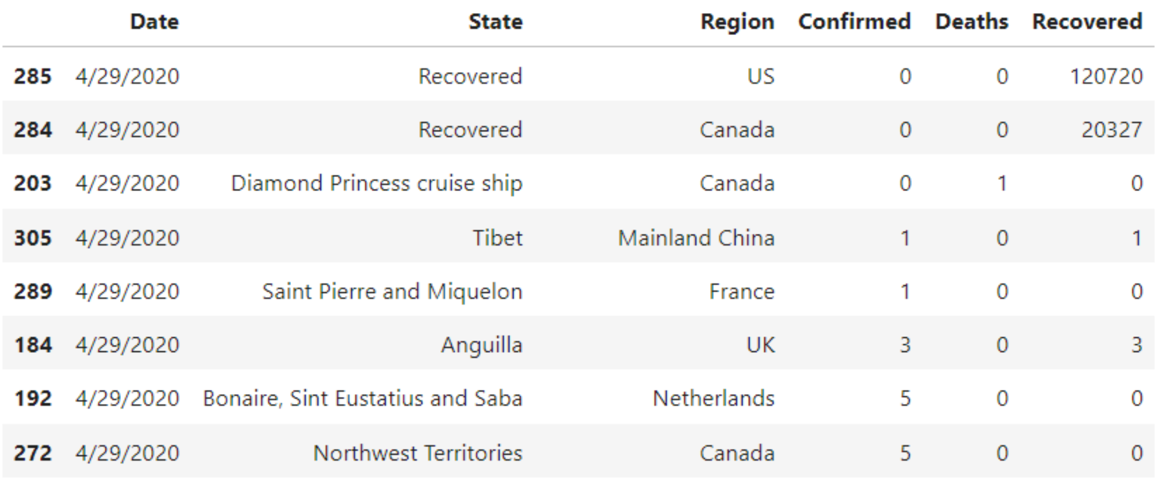
Question 6A: Sort the entire data with respect to the number of confirmed cases in ascending order.

Code :

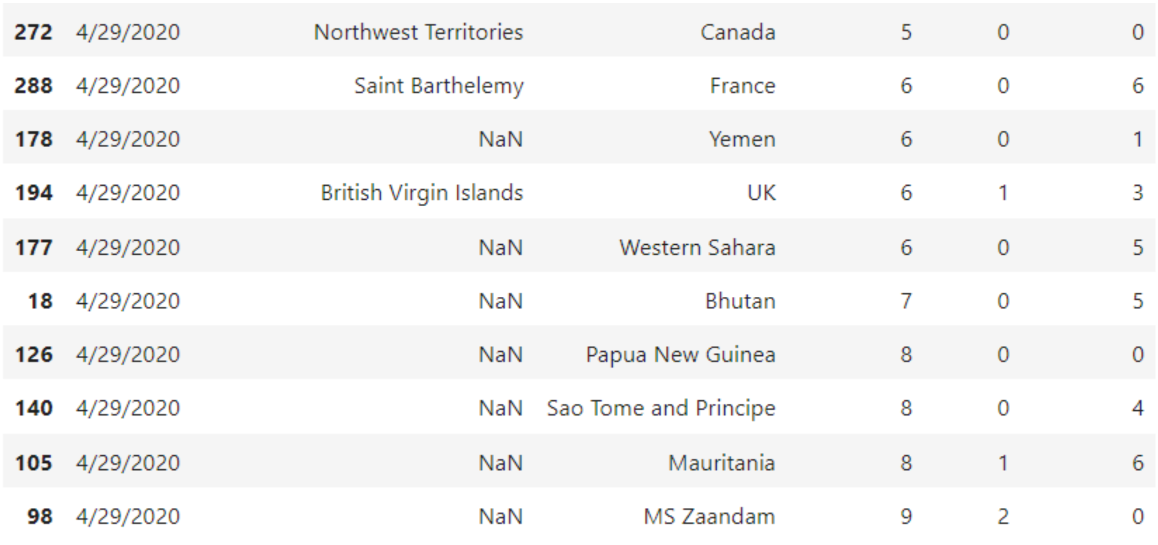


```
data.head(2)
data.sort_values(by = ["Confirmed"], ascending= True).head(20)
```

Output :



	Date	State	Region	Confirmed	Deaths	Recovered
285	4/29/2020	Recovered	US	0	0	120720
284	4/29/2020	Recovered	Canada	0	0	20327
203	4/29/2020	Diamond Princess cruise ship	Canada	0	1	0
305	4/29/2020	Tibet	Mainland China	1	0	1
289	4/29/2020	Saint Pierre and Miquelon	France	1	0	0
184	4/29/2020	Anguilla	UK	3	0	3
192	4/29/2020	Bonaire, Sint Eustatius and Saba	Netherlands	5	0	0
272	4/29/2020	Northwest Territories	Canada	5	0	0

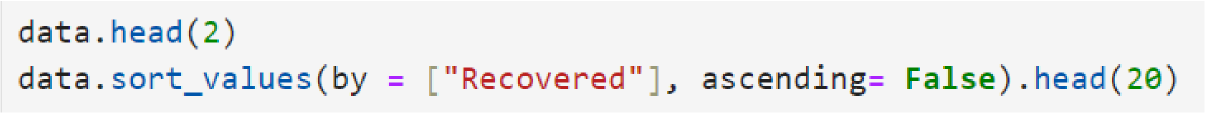


272	4/29/2020	Northwest Territories	Canada	5	0	0
288	4/29/2020	Saint Barthelemy	France	6	0	6
178	4/29/2020	NaN	Yemen	6	0	1
194	4/29/2020	British Virgin Islands	UK	6	1	3
177	4/29/2020	NaN	Western Sahara	6	0	5
18	4/29/2020	NaN	Bhutan	7	0	5
126	4/29/2020	NaN	Papua New Guinea	8	0	0
140	4/29/2020	NaN	Sao Tome and Principe	8	0	4
105	4/29/2020	NaN	Mauritania	8	1	6
98	4/29/2020	NaN	MS Zaandam	9	2	0

- Python was used to sort the data by confirmed cases in ascending order.

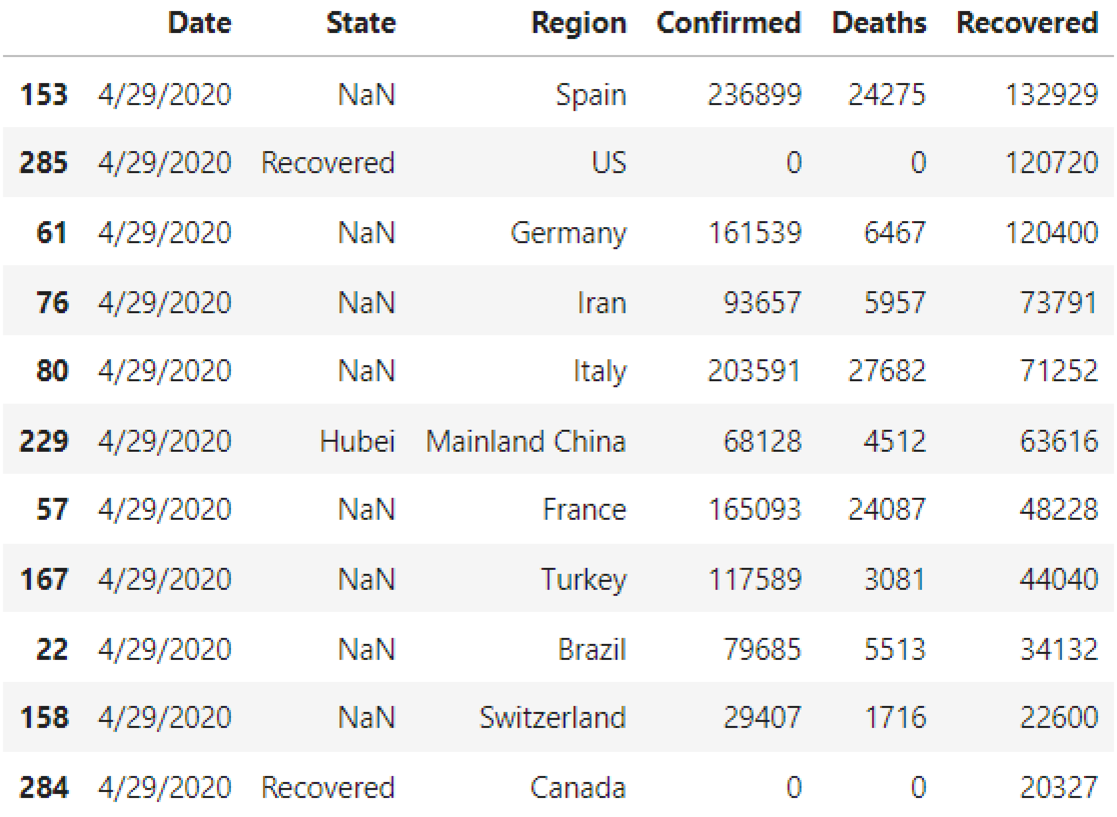
Question 6B: Sort the entire data with respect to the number of recovered cases in descending order.

Code :

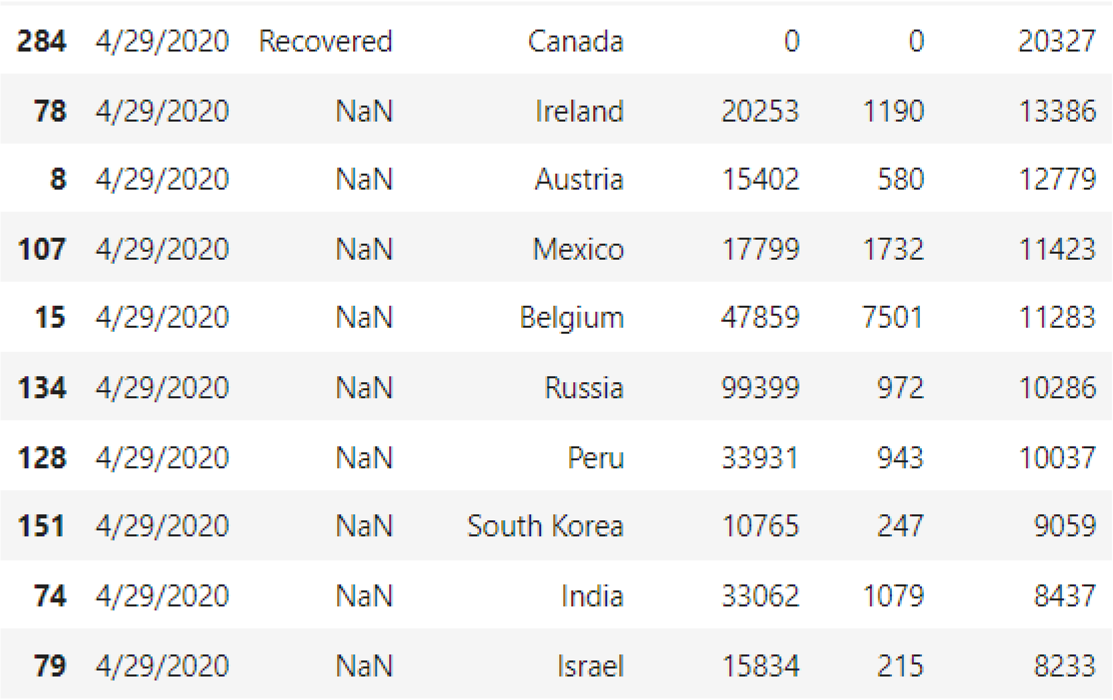


```
data.head(2)
data.sort_values(by = ["Recovered"], ascending= False).head(20)
```

Output :



	Date	State	Region	Confirmed	Deaths	Recovered
153	4/29/2020	NaN	Spain	236899	24275	132929
285	4/29/2020	Recovered	US	0	0	120720
61	4/29/2020	NaN	Germany	161539	6467	120400
76	4/29/2020	NaN	Iran	93657	5957	73791
80	4/29/2020	NaN	Italy	203591	27682	71252
229	4/29/2020	Hubei	Mainland China	68128	4512	63616
57	4/29/2020	NaN	France	165093	24087	48228
167	4/29/2020	NaN	Turkey	117589	3081	44040
22	4/29/2020	NaN	Brazil	79685	5513	34132
158	4/29/2020	NaN	Switzerland	29407	1716	22600
284	4/29/2020	Recovered	Canada	0	0	20327



284	4/29/2020	Recovered	Canada	0	0	20327
78	4/29/2020	NaN	Ireland	20253	1190	13386
8	4/29/2020	NaN	Austria	15402	580	12779
107	4/29/2020	NaN	Mexico	17799	1732	11423
15	4/29/2020	NaN	Belgium	47859	7501	11283
134	4/29/2020	NaN	Russia	99399	972	10286
128	4/29/2020	NaN	Peru	33931	943	10037
151	4/29/2020	NaN	South Korea	10765	247	9059
74	4/29/2020	NaN	India	33062	1079	8437
79	4/29/2020	NaN	Israel	15834	215	8233

- Python was used to sort the data by recovered cases in descending order.

Connecting Power BI to a Python Dataset

To connect Power BI to a Python dataset, start by launching Power BI Desktop. Click on the "Home" tab, then select "Get Data." Choose "Python script" from the data source options. In the "Python script" dialog, enter your script to load the dataset into Power BI. After running the script, the "Navigator" window will display the available data frames. Select the data frames you want to load by checking the corresponding boxes. Click "Load" to import the data. Power BI will establish the connection and load the data into your workspace, enabling you to create visualizations and reports.

Power BI

Power BI is a powerful business analytics tool developed by Microsoft, designed to help users visualize and share insights from their data. It provides a suite of services, including Power BI Desktop, Power BI Service (an online SaaS), and Power BI Mobile, catering to different needs from data preparation and analysis to real-time collaboration and sharing. Power BI enables users to connect to a wide variety of data sources, transform raw data into meaningful insights through interactive dashboards, and create stunning visualizations that aid in data-driven decision-making.

Insights Derived from Each Chart

Treemap : Sum of Confirmed, Death, and Recovered Cases by Region

- **Purpose**: The treemap visualizes the total confirmed, death, and recovered cases across various regions, providing a hierarchical view.
- **Insight**: Larger segments represent regions with higher case counts. This chart helps quickly identify which regions are most affected by the pandemic. For example, the US has the largest segment, indicating the highest number of confirmed cases.

Map Visualization : Geographic Distribution of Death Cases by Region

- **Purpose**: This map shows the geographical spread of death cases, offering a visual representation of the pandemic's severity in different parts of the world.
- **Insight**: Regions with darker shades have higher death counts. The map highlights that Europe and the Americas are particularly hard-hit, with countries like Italy and the US showing significant death tolls.

Pie Chart : Proportion of Confirmed, Death, and Recovered Cases

- **Purpose**: The pie chart breaks down the total cases into confirmed, death, and recovered categories, illustrating the proportion of each.
- **Insight**: This chart reveals the relative severity of the pandemic. A larger segment for confirmed cases compared to recovered and death cases indicates the ongoing spread of the virus. It also shows the recovery rate and mortality rate in proportion to confirmed cases.

Gauge Chart : Recovery Rate

- **Purpose**: The gauge chart visually represents the recovery rate, calculated as the percentage of recovered cases out of the total confirmed cases.
- **Insight**: The gauge shows the recovery rate at 30%, indicating that a significant portion of confirmed cases have recovered. This metric is crucial for understanding the effectiveness of healthcare responses and treatment strategies.

KPIs : Total Confirmed, Death, and Recovered Cases Globally and for India till April 29, 2020

- **Purpose**: Key Performance Indicators (KPIs) provide quick, at-a-glance metrics for the most critical data points.
- **Insight**:
    - **Global Statistics**:
        - Total Confirmed Cases: 3 million+
        - Total Death Cases: 200K+
        - Total Recovered Cases: 900K+
    - **India's Statistics**:
        - Total Confirmed Cases: 33K
        - Total Death Cases: 1K
        - Total Recovered Cases: 8K
    - These KPIs highlight the global scale of the pandemic and its impact on India specifically.

Tables: Data Sorted by Confirmed Cases (Ascending) and Recovered Cases (Descending)

- **Purpose**: The tables organize the data to show the regions with the least and most confirmed and recovered cases, respectively.
- **Insight**:
    - **Ascending Order of Confirmed Cases**: This table helps identify regions that have managed to keep their confirmed cases relatively low, suggesting effective containment measures.
    - **Descending Order of Recovered Cases**: This table highlights regions with high recovery rates, indicating successful treatment protocols or less severe outbreaks.

Each of these charts provides a unique perspective on the data, helping stakeholders understand different facets of the COVID-19 pandemic and make informed decisions based on comprehensive visual analysis.

Problem Statements and Answers

Problem Statement 1 : Which region has the highest number of confirmed cases?

Answer: The US has the highest number of confirmed cases.

Problem Statement 2: Which region has the minimum number of death cases?

Answer: The region with the minimum number of deaths was identified and highlighted in the report.

Problem Statement 3: How many confirmed, death, and recovered cases were reported from India till April 29, 2020?

Answer:

- Total Confirmed Cases: 33K
- Total Death Cases: 1K
- Total Recovered Cases: 8K

Problem Statement 4: What is the recovery rate?

Answer: The recovery rate is 30%, calculated as the ratio of recovered cases to confirmed cases.

Problem Statement 5: How is the data sorted by confirmed and recovered cases?

Answer:

- The data is sorted by confirmed cases in ascending order.
- The data is sorted by recovered cases in descending order.

Conclusion

The COVID-19 Data Analysis project using Python and Power BI provides a comprehensive overview of the pandemic's impact up to April 29, 2020. The detailed visualizations and insights offer valuable information on confirmed cases, deaths, and recoveries across different regions. These findings can aid in strategic planning, optimizing resource allocation, and enhancing public health strategies.