Phase 3: Development Part 1 - Loading and Preprocessing the Dataset

Introduction

In this phase of the project, we embark on building a COVID-19 cases analysis utilizing IBM Cognos for visualization. The primary focus here is on loading and preprocessing the COVID-19 dataset to ensure it is ready for analysis.

Project Objectives

The aim of this project is to enable how to utilize our Data Visualization and Data Analytics skills.

1. To identify emerging trends in the spread of Covid-19
2. To assess the impact of cases and death rates on the spread of Covid-19 on various countries across the world
3. To assess the impact of cases and death rates on the spread of Covid-19 on various days of a month in the year 2021.
4. To create predictive models to better anticipate future cases and death rates related to COVID-19.

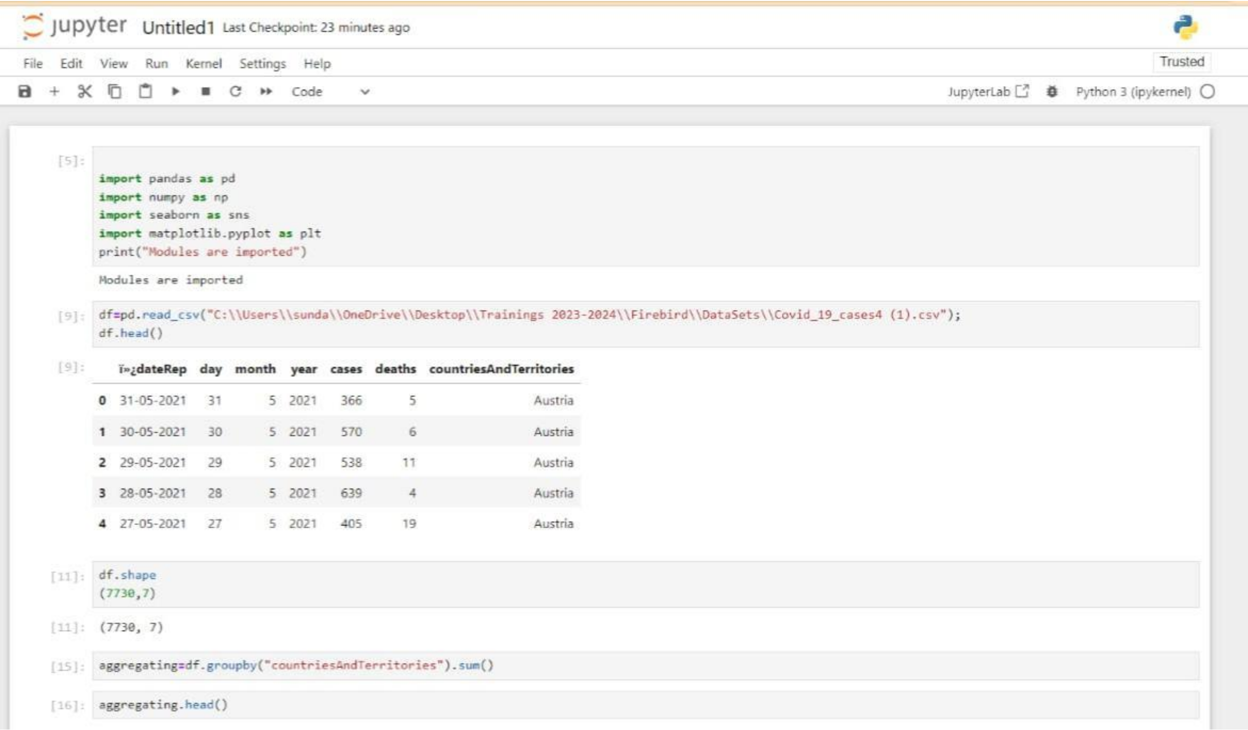
5. To provide insights and analyze demographics that could lead to decision making.

Analysis Objectives

Our analysis objectives encompass understanding the patterns, trends, and insights related to COVID-19 cases and deaths. We aim to address questions such as the impact of the pandemic on different regions, the progression of cases over time, and the correlation between various factors and COVID-19 outcomes.

Data Source: We obtained our data from Kaggle, a reputable platform for datasets. The dataset link is provided here: COVID-19 Cases Dataset.

Data Loading: To begin the analysis, we loaded the COVID-19 dataset into our IBM Cognos environment. This step involved using the necessary functions and tools to import the data.


jupyter Untitled1 Last Checkpoint: 23 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)

```
[5]:
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
print("Modules are imported")

Modules are imported
```

```
[9]: df=pd.read_csv("C:\\Users\\sunda\\OneDrive\\Desktop\\Trainings 2023-2024\\Firebird\\DataSets\\Covid_19_cases4 (1).csv");
     df.head()
```

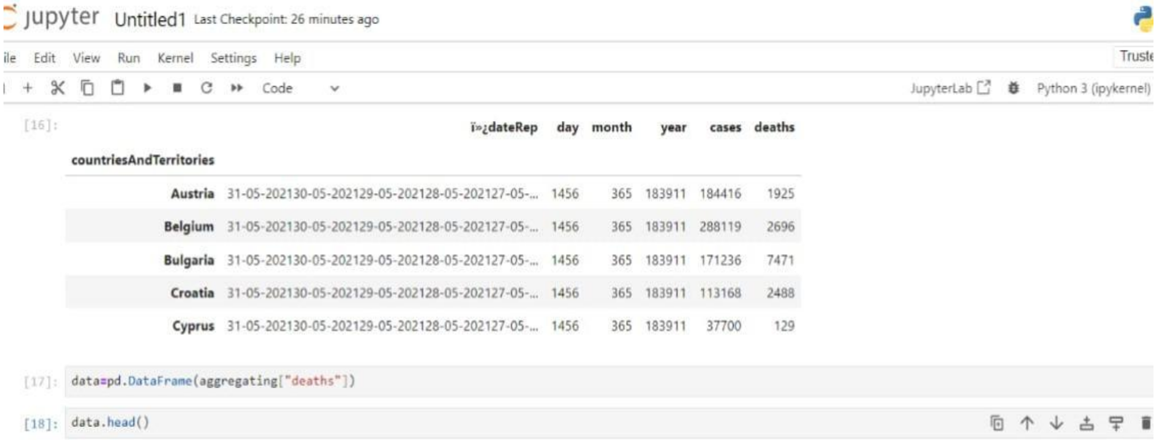
[9]:

	ï»¿dateRep	day	month	year	cases	deaths	countriesAndTerritories
0	31-05-2021	31	5	2021	366	5	Austria
1	30-05-2021	30	5	2021	570	6	Austria
2	29-05-2021	29	5	2021	538	11	Austria
3	28-05-2021	28	5	2021	639	4	Austria
4	27-05-2021	27	5	2021	405	19	Austria

```
[11]: df.shape
      (7730,7)

[11]: (7730, 7)

[15]: aggregating=df.groupby("countriesAndTerritories").sum()

[16]: aggregating.head()
```

Data Preprocessing: Data preprocessing is a critical step to ensure the data's accuracy and reliability. It involves several key aspects:

- Handling Missing Values: We meticulously dealt with missing data points. This might include imputation, removal of incomplete records, or other strategies based on the nature of the data.
- Data Transformation: Depending on the analysis objectives, we performed data transformations such as normalizing values, aggregating data over specific time periods, or converting data types to match the analytical requirements.
- Data Cleaning: Data cleaning involved removing outliers, duplicate entries, or any irrelevant information that could skew the analysis.
- Data Verification:To ensure the dataset's accuracy and reliability, we conducted verification steps, including cross-checking data with authoritative sources, validating data against known statistics, and conducting internal consistency checks.

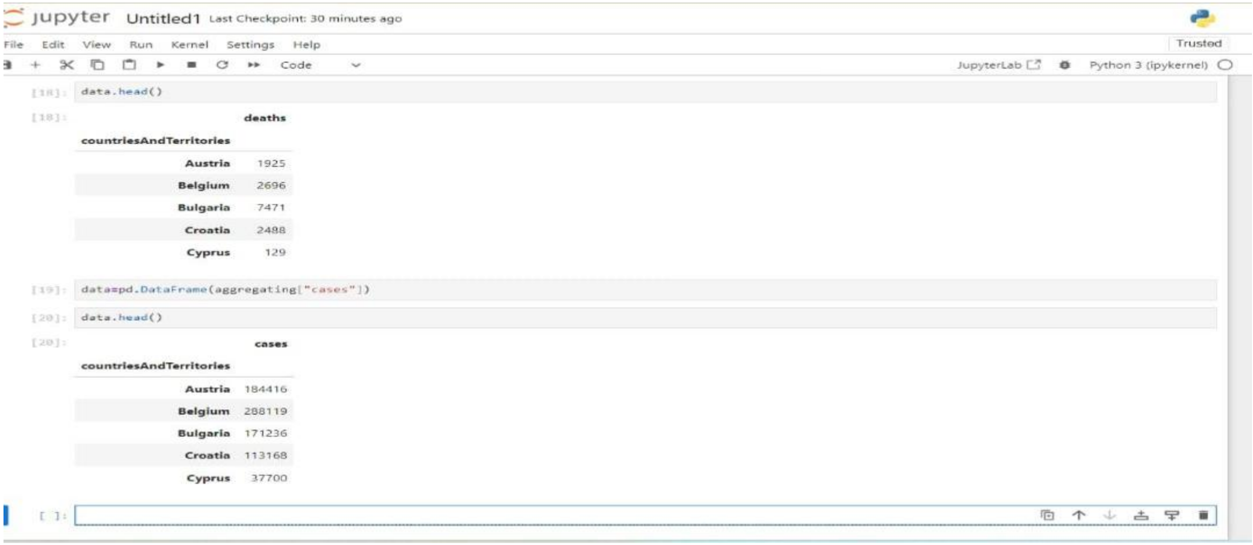


jupyter Untitled1 Last Checkpoint: 26 minutes ago
ile Edit View Run Kernel Settings Help
Truste
Code
JupyterLab
Python 3 (ipykernel)

```
[16]:
                                                                  ï»¿dateRep  day  month    year   cases  deaths
countriesAndTerritories
                Austria  31-05-202130-05-202129-05-202128-05-202127-05-...  1456    365  183911  184416    1925
                Belgium  31-05-202130-05-202129-05-202128-05-202127-05-...  1456    365  183911  288119    2696
               Bulgaria  31-05-202130-05-202129-05-202128-05-202127-05-...  1456    365  183911  171236    7471
                Croatia  31-05-202130-05-202129-05-202128-05-202127-05-...  1456    365  183911  113168    2488
                 Cyprus  31-05-202130-05-202129-05-202128-05-202127-05-...  1456    365  183911   37700     129
```

```
[17]: data=pd.DataFrame(aggregating["deaths"])
```

```
[18]: data.head()
```



jupyter Untitled1 Last Checkpoint: 30 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)

```
[18]: data.head()
```

```
[18]:
                          deaths
countriesAndTerritories
                Austria     1925
                Belgium     2696
               Bulgaria     7471
                Croatia     2488
                 Cyprus      129
```

```
[19]: data=pd.DataFrame(aggregating["cases"])
```

```
[20]: data.head()
```

```
[20]:
                           cases
countriesAndTerritories
                Austria   184416
                Belgium   288119
               Bulgaria   171236
                Croatia   113168
                 Cyprus    37700
```

```
[ ]:
```

Conclusion

This part of the development phase has focused on laying the foundation for the COVID-19 analysis. The dataset has been successfully loaded into IBM Cognos, meticulously preprocessed, and verified for accuracy. We are now ready to move on to the exciting phase of data visualization and analysis using IBM Cognos.