**COVID-19 CASE ANALYSIS**

**Introduction**

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has profoundly impacted the world since its emergence in late 2019. This global health crisis has led to unprecedented challenges in healthcare, the economy, and everyday life. In the face of this crisis, data analysis has proven to be an essential tool for understanding the spread of the virus, evaluating public health measures, and informing policy decisions.COVID-19 case analysis involves the systematic examination of data related to the pandemic, including the number of confirmed cases, deaths, recoveries, testing, vaccination rates, and various other related variables. This analysis aims to extract insights and patterns from the data to support informed decision-making, resource allocation, and the development of effective public health strategies.

**1. Objectives**

In Phase 4, we aim to achieve the following objectives:

1. **Data Exploration and Visualization:** Data exploration and visualization are fundamental steps in data analysis. Data exploration involves examining data for patterns, anomalies, and missing values. Data visualization converts data into graphical representations to highlight insights, trends, and relationships. These processes enhance data understanding, guide modeling choices, and facilitate effective communication of findings, supporting informed decision-making.
2. **Dashboard Design:** Using IBM Cognos, we will design interactive dashboards that condense complex data into easily digestible formats. These dashboards will serve as a valuable resource for stakeholders and decision-makers to monitor and assess performance.
3. **Data Preparation and Pre-processing:** Data preparation and preprocessing are critical steps in the data analysis pipeline. Data preparation involves collecting, cleaning, and organizing raw data from various sources to ensure it is ready for analysis. This may include handling missing values, dealing with outliers, and structuring data for analysis. Data preprocessing encompasses tasks like feature engineering, scaling, and encoding to make the data suitable for machine learning algorithms. These steps are essential for generating accurate, reliable insights and models from the data.

**2.1 Data Sources**

Data sources refer to the origins of data used for analysis, research, or decision-making. These sources can be diverse and include databases, surveys, sensors, social media, web scraping, and more. Data sources provide the foundational information that organizations and individuals rely on to gain insights, make informed choices, and conduct research. Accessing, managing, and analyzing data from these sources is a fundamental aspect of data-driven activities.

**2.2 Data Cleaning**

1. Data Collection:

- Obtain COVID-19 case data from reliable sources such as government health agencies, the World Health Organization (WHO), or academic research institutions.

2. Data Format:

- Ensure that the data is in a structured format, such as CSV, Excel, or a database, to facilitate cleaning and analysis.

3. Duplicate Removal:

- Identify and remove duplicate records if they exist. Duplicates can skew analysis results.

4. Missing Data:

- Identify missing values and decide how to handle them. Options include imputation (filling in missing values with estimates), removal of incomplete records, or keeping missing values intact with proper documentation.

5. Outliers:

- Detect and address outliers that may indicate data entry errors or anomalies. You can use statistical methods like Z-scores or visualization techniques to identify outliers.

6. Data Types:

- Ensure that variables are of the correct data types (e.g., dates as dates, numbers as numbers). Convert data as needed.

7. Inconsistent Data:

- Check for inconsistent or erroneous data, such as typos or inconsistent date formats. Standardize data formats and resolve discrepancies.

8. Standardize Variable Names:

- Make sure variable names are consistent and descriptive, which makes analysis more accessible. Renaming columns if necessary can help in this regard.

9. Geospatial Data:

- If the data includes geospatial information (e.g., location of cases), ensure that the coordinates or place names are accurate and standardized.

10. Time Series Data:

- Ensure that the time series data is in chronological order, and date-related variables are correctly formatted.

11. Categorical Data:

- Clean and standardize categorical variables. This may involve consolidating categories or recoding data.

12. Data Integrity:

- Verify that there are no logical inconsistencies, such as cases with future dates or negative case counts.

13. Data Validation:

- Cross-check the data against other sources or references to validate its accuracy.

14. Data Documentation:

- Create detailed documentation that describes the data cleaning process, including the decisions made, transformations applied, and any data quality issues encountered.

15. Data Versioning:

- If you receive updated data over time, maintain version control and keep records of changes made in each version.

16. Backups:

- Create backups of the original data and cleaned data to ensure data recovery in case of errors or accidental changes.

**2.3 Data Structuring**

Data structuring involves organizing data into a logical and manageable format. It often consists of arranging data into rows and columns, where rows represent individual data points, and columns represent attributes or features of those points. This structured format makes data more accessible and facilitates various data analysis and modeling techniques, making it easier to derive insights and patterns from the data.

**2.4 Data Documentation**

Throughout the data preparation and preprocessing phase, we maintained comprehensive documentation. This documentation provides transparency and traceability, ensuring that the processes applied to the dataset are well-documented and reproducible.

The result of our data preparation and preprocessing efforts is a structured, clean, and reliable dataset ready for in-depth exploration, visualization, and regression analysis. This robust dataset will underpin our subsequent analyses, providing a solid foundation for meaningful insights and predictions.

**CODING :**

import matplotlib.pyplot as plt

# Sample COVID-19 case data by country or territory

data = {

'countriesAndTerritories': ['Austria', 'Germany', 'Italy', 'France', 'Spain'],

'cases': [5000, 8000, 6000, 7000, 7500]

}

# Create a pie chart

plt.figure(figsize=(8, 8))

plt.pie(data['cases'], labels=data['countriesAndTerritories'], autopct='%1.1f%%', startangle=140)

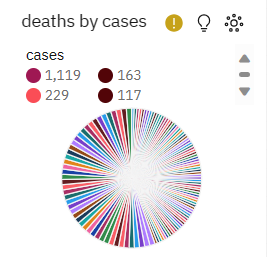
plt.title('COVID-19 Cases by Country/Territory')

# Show the pie chart

plt.show()

**OUTPUT :**

**IBM Cognos :**

****

