**PHASE III : Development Part 1**

**COVID-19 Cases and Deaths Data in the EU/EEA**

Introduction:

The COVID-19 analysis project is aimed at extracting, processing, and analyzing COVID-19 cases and death data to gain insights into the pandemic's impact. The project uses a dataset from Kaggle, which contains information related to COVID-19 cases and deaths. The primary objectives of this project are to understand the trends, patterns, and geographical distribution of COVID-19 cases and deaths. By extracting and cleaning the data, the project ensures accuracy and reliability in the analysis. The insights generated from this analysis can be valuable for public health officials, researchers, and policymakers to make informed decisions.

Analysis Overview:

Data Extraction:

The project begins by downloading the COVID-19 dataset from the provided Kaggle link. It then uses Python and the pandas library to load the dataset into a data frame.

Data Cleaning:

After loading the data, the project focuses on data cleaning. This includes:

Removing irrelevant columns that do not contribute to the analysis.

Addressing missing data by either imputing values or removing rows with missing information.

Ensuring data consistency, such as data types, date formats, and handling outliers.

Correcting data anomalies and errors.

Converting data types to the appropriate format, for example, converting date columns to datetime data type.

Data Analysis: With the cleaned data, the project prepares the dataset for analysis. This step includes creating subsets of data based on specific attributes, such as location or time, and performing various analyses, such as:

Time-series analysis to understand the progression of COVID-19 cases and deaths over time.

Geographic analysis to examine the distribution of cases and deaths across regions.

Demographic analysis, if available in the dataset.

Exploring correlations and patterns in the data.

This analysis project serves as a valuable tool for understanding the COVID-19 pandemic's impact, and the insights generated can be used to inform public health strategies and policies to combat the virus effectively.

Python Code

import pandas as pd

# Load the dataset

dataset\_path = 'path\_to\_your\_dataset.csv'

# Replace with the actual path to your CSV file

df = pd.read\_csv(dataset\_path)

# Extract relevant columns (e.g., date, location, cases, deaths)

relevant\_columns = ['date', 'location', 'cases', 'deaths']

df = df[relevant\_columns]

# Data cleaning

# Remove rows with missing values

df.dropna(subset=['date', 'location', 'cases', 'deaths'], inplace=True)

# Data type conversions if needed (e.g., date columns)

df['date'] = pd.to\_datetime(df['date'])

# Save the cleaned data to a new CSV file

cleaned\_dataset\_path = 'cleaned\_covid\_data.csv'

df.to\_csv(cleaned\_dataset\_path, index=False)

This code will help in extracting the new data set, cleaning it and saving it in new file.

The link for the new dataset File is below.

<https://drive.google.com/file/d/1HSIE6rawcZVbl7bDPRO4VxAouPw7VoJy/view?usp=sharing>

Team members of the project:

**Ajit Ashok – 2021103001**

**Mohammed Faizan – 2021103023**

**Ragul Raj – 2021103035**

**Rushil A Krishanan – 2021103041**

**Pal Dhinakaran P - 2021103736**