**Project : Covid-19 Cases Analysis**

**Phase 5: Project Documentation & Submission**

**PROJECT'S OBJECTIVE :**

The objective of this project is to analyze COVID-19 cases data to gain insights into the trends and impacts of the pandemic. This analysis aims to understand patterns in the data, identify hotspots, and extract meaningful insights to aid in decision-making and public health efforts.

**DESIGN THINKING PROCESS :**

**1.Analysis Objectives:**

Define the specific objectives of analyzing COVID-19 cases and deaths data, such as comparing mean values and standard deviations.

**2.Data Collection**:

Obtain the provided data file containing COVID-19 cases and deaths information per day and by country in the EU/EEA.

**3.Visualization Strategy:**

Plan how to visualize the mean values and standard deviations using IBM Cognos to create informative charts and graphs.

**4.Insights Generation:**

Identify potential insights from the comparison of mean values and standard deviation of cases and deaths.

**DEVELOPMENT PHASE :**

In this project phase, the focus is on laying the foundation for the COVID-19 cases analysis using IBM Cognos. The process begins by defining specific analysis objectives, followed by obtaining the COVID-19 cases and deaths data file. The crucial step of preprocessing and cleaning the data ensures its accuracy and reliability, involving activities like handling missing values and structuring data appropriately. Simultaneously, IBM Cognos is introduced as the visualization tool to translate the processed data into meaningful visual representations, setting the stage for in-depth analysis and actionable insights.

**ANALYSIS OBJECTIVES :**

Analysis of COVID-19 dataset containing cases and death information

Study of the performance curves of various countries in fighting this pandemic.

Analyses the effectiveness of lockdown and to know the further implementations that can be done accordingly.

Learning the right steps to contain the infection by not repeating the mistakes of other countries.

**DATA COLLECTION :**

In this data preprocessing workflow:

**Importing Libraries:**

Necessary libraries like Pandas, NumPy, and SciPy's zscore function are imported for data manipulation and analysis.

**Loading the Dataset:**

The COVID-19 data is read from a CSV file and stored in a Pandas DataFrame called 'data'.

**Data Inspection:**

The initial rows of the dataset are displayed, providing an overview of the data. Additional information, such as data types and missing values, is also printed.

**Handling Missing Data:**

Missing values in the dataset are filled with the mean values of their respective columns, ensuring completeness of the data.

**Data Validation and Cleaning:**

Duplicate rows are removed from the dataset, ensuring each row is unique and avoiding redundant information.

**Correcting Data Types:**

The 'dateRep' column is converted to the datetime data type, enabling easier date-based operations and analysis.

**Further Analysis:**

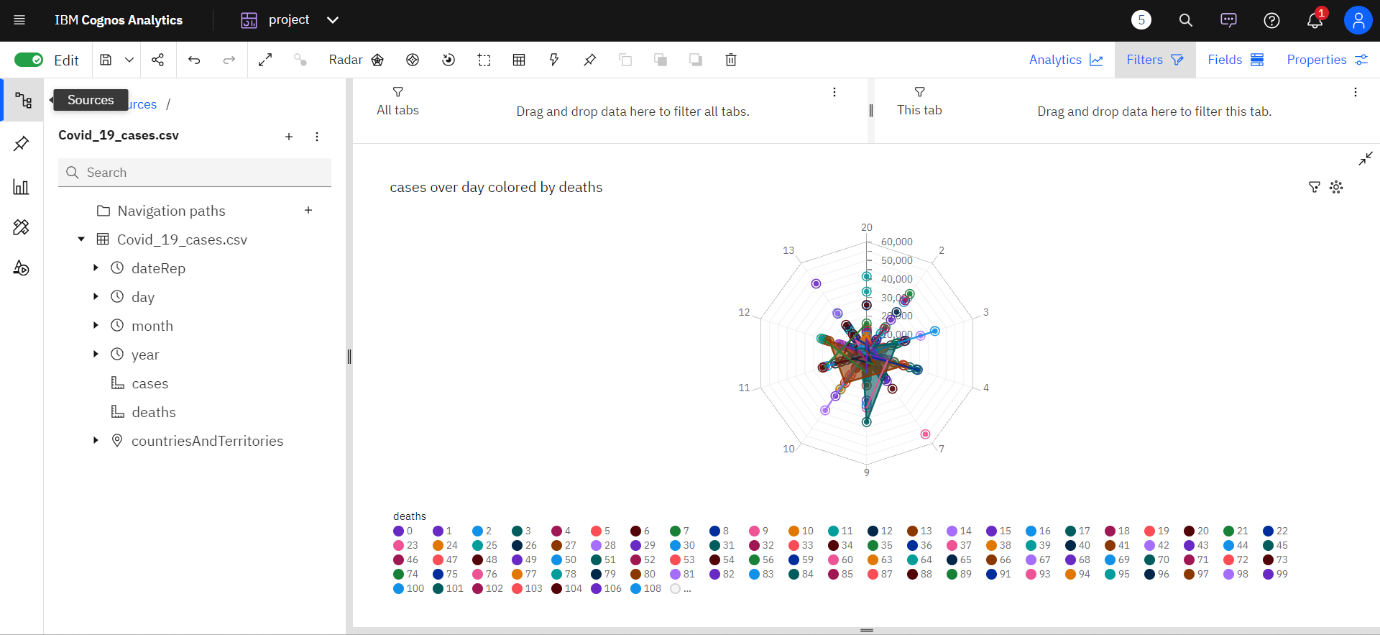
After preprocessing, the cleaned data is ready for in-depth analysis, modeling, or other specific tasks tailored to the project objectives.

**Saving the Cleaned Dataset:**

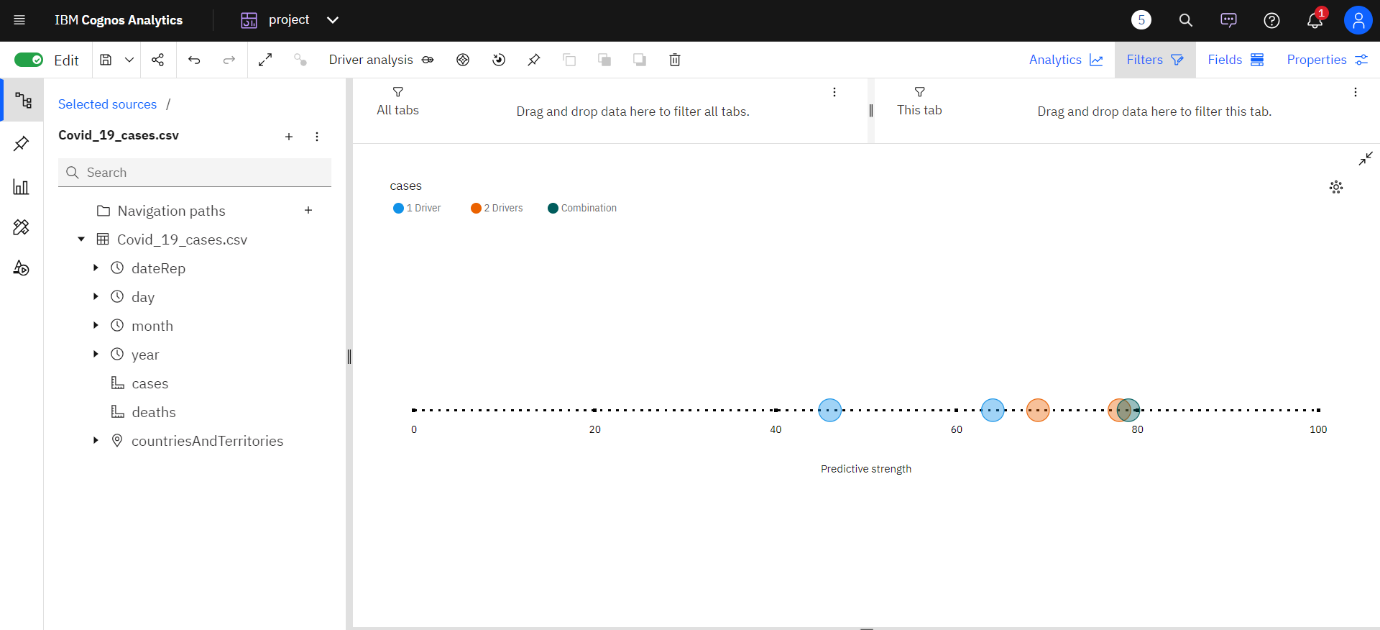
The cleaned dataset is saved to a new CSV file named "Covid\_19\_cases.csv" without including the index column, preserving the processed data for future use.

**Data visualization using IBM Cognos:**

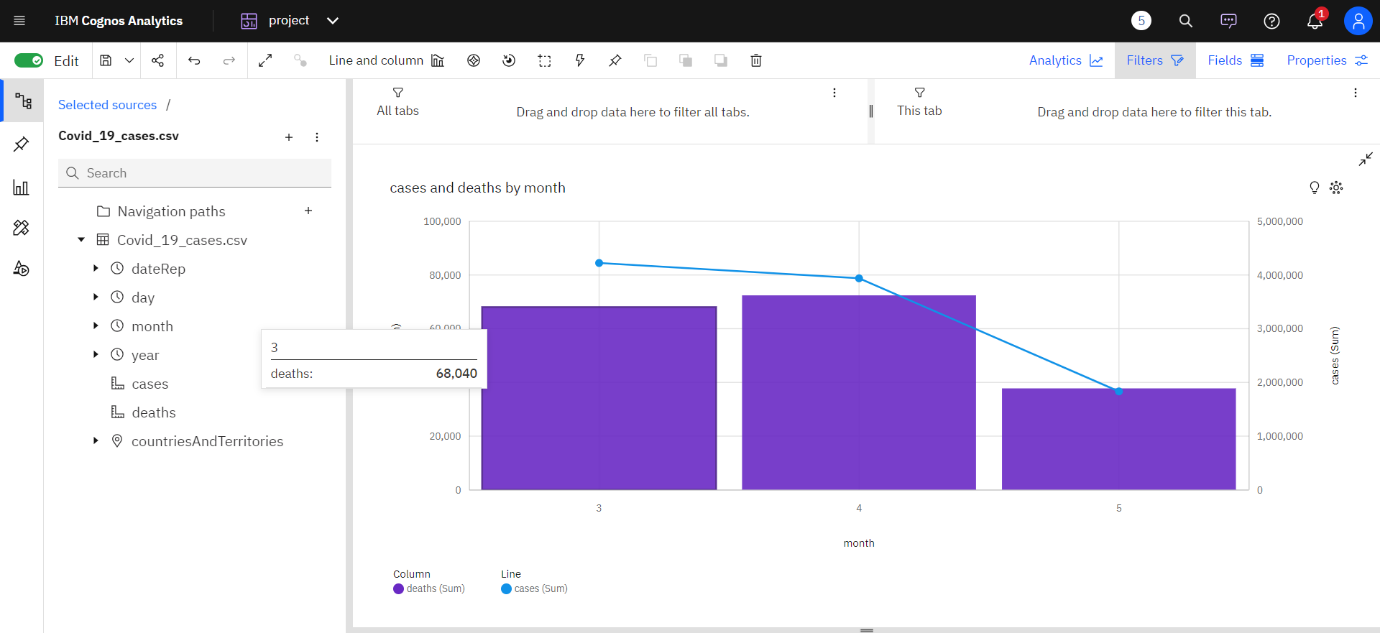
**Visualization of cases over day colored by deaths**



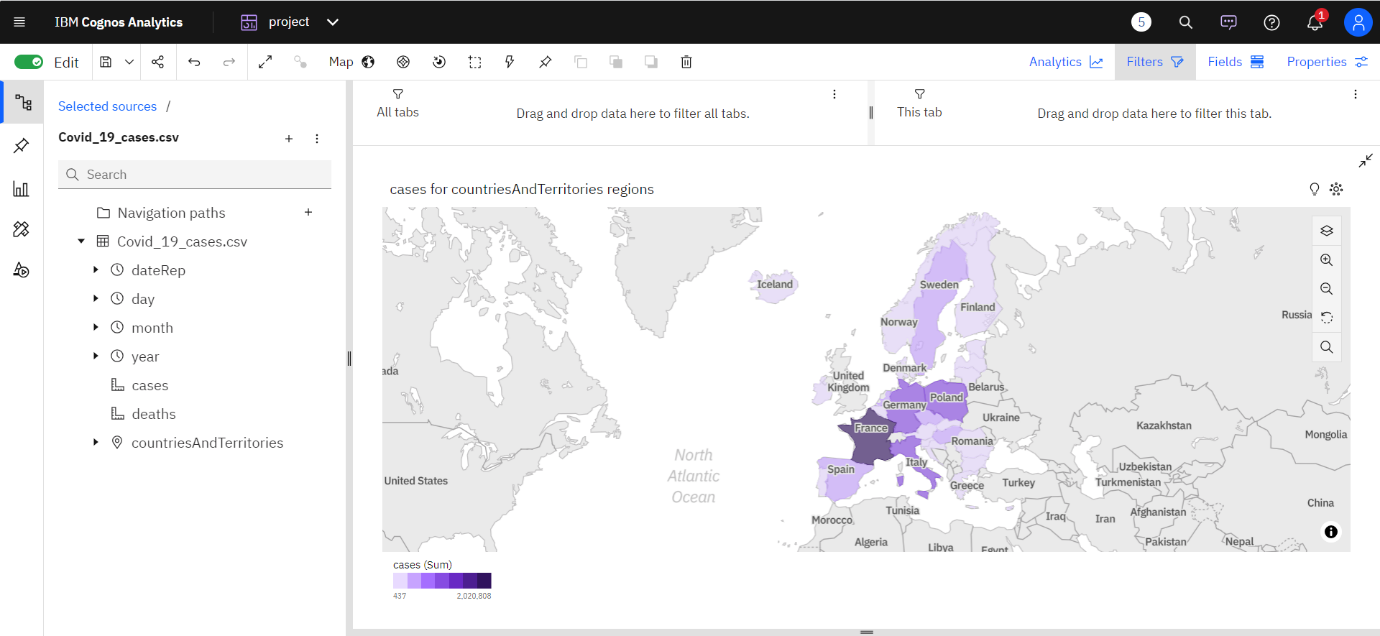
**Visualization by cases**



**Visualization cases and deaths by month**



**visualization of cases for countries and territories regions**



**INSIGHTS GENERATED :**

**Identifying Trends:**

**Cases:**

Top Peak Dates for Cases:

**1. Timestamp('2021-12-05 00:00:00')**

**2. Timestamp('2021-12-05 00:00:00')**

**3. Timestamp('2021-12-04 00:00:00')**

**Deaths:**

Top Peak Dates for Deaths:

**1. Timestamp('2021-12-05 00:00:00')**

**2. Timestamp('2021-12-05 00:00:00')**

**3. Timestamp('2021-12-05 00:00:00')**

**Analyzing Variations:**

**Cases:**

Top Sudden Rise Dates for Cases:

**839 2021-11-05**

**846 2021-04-05**

**849 2021-01-05**

**850 2021-04-30**

**851 2021-04-29**

**...**

**2167 2021-03-18**

**2171 2021-03-14**

**2173 2021-12-03**

**2590 2021-04-19**

**2597 2021-12-04**

**Deaths:**

**Top Sudden Rise Dates for Deaths:**

**853 2021-04-27**

**859 2021-04-21**

**860 2021-04-20**

**867 2021-04-13**

**872 2021-08-04**

**...**

**2174 2021-11-03**

**2614 2021-03-26**

**2618 2021-03-22**

**2635 2021-05-03**

**2637 2021-03-03**

**CORRELATION ANALYSIS :**

Correlation Coefficient between Cases and Deaths is **0.766308878657635**

0.766308878657635 (indicating a strong positive correlation).

**SUBMISSION PART :**

**CREATING MEAN VISUALIZATION :**

**Create a Query:**

Start by creating a query in IBM Cognos based on the imported dataset, selecting specific columns: date, cases, and deaths.

**Aggregate the Data:**

Use aggregation functions in IBM Cognos to calculate mean values for cases and deaths.

For cases, apply the "mean" or "average" aggregation function to the "cases" column.

For deaths, use the same aggregation function on the "deaths" column.

Group the data by the "date" column to calculate mean values for each date.

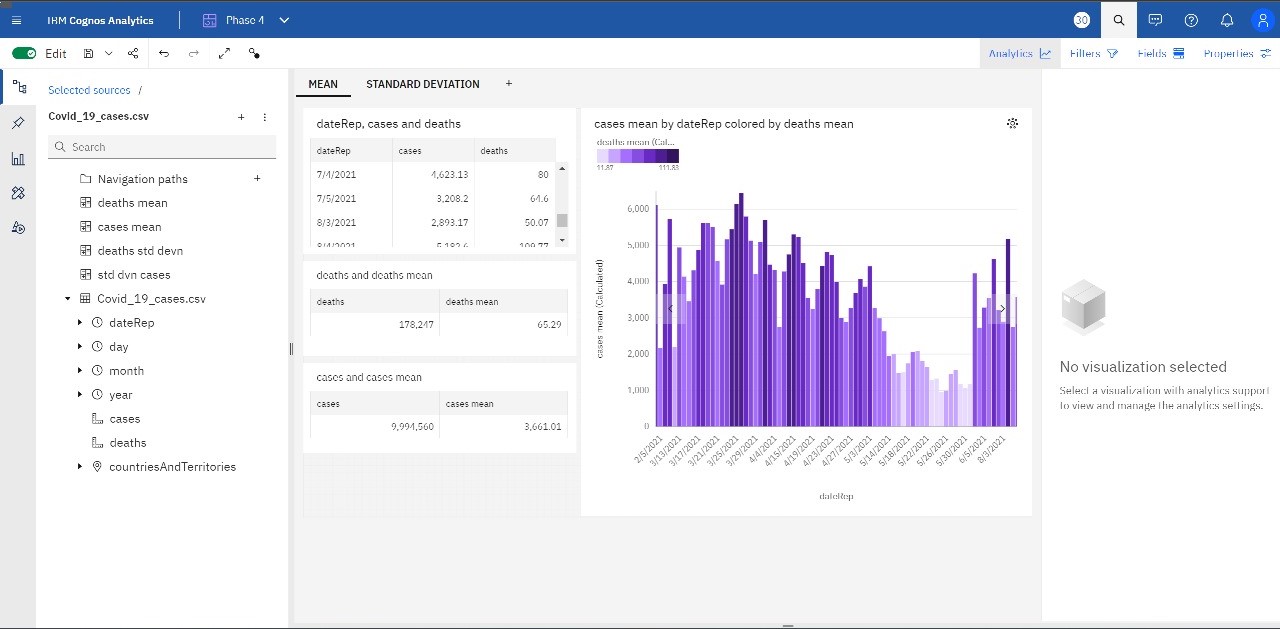
**Run the Query:**

Execute the query to obtain aggregated data, resulting in mean values for COVID-19 cases and deaths for each date in the dataset.

**Create Visualizations:**

Utilize the aggregated data to create visualizations such as line charts or bar charts. Dates are plotted on the X-axis, and mean values for cases and deaths are represented on the Y-axis. These visualizations provide clear insights into the trends of COVID-19 cases and deaths over time.

**Visualization by comparing the mean values**



**CREATING STANDARD DEVIATION VISUALIZATION :**

**Create a Query:**

Begin by creating a query in IBM Cognos based on the imported dataset, selecting specific columns: Date, Cases, and Deaths.

**Aggregate the Data (Standard Deviation):**

Utilize aggregation functions in IBM Cognos to calculate standard deviation values for Cases and Deaths.

For Cases, apply the "Standard Deviation" aggregation function to the "Cases" column.

For Deaths, use the same aggregation function on the "Deaths" column.

Group the data by the "Date" column to calculate standard deviation values for each date.

**Run the Query:**

Execute the query to obtain aggregated data, resulting in standard deviation values for COVID-19 cases and deaths for each date in the dataset.

**Create Visualizations:**

Utilize the aggregated data to create visualizations such as bar charts or line charts. Dates are plotted on the X-axis, and standard deviation values for cases and deaths are represented on the Y-axis. These visualizations provide a clear overview of the variability in COVID-19 cases and deaths over time.

**Visualization by comparing the standard deviation values**

