**Project Title:** COVID-19 using Cognos

PHASE: 5

**Project Definition and Design Thinking**

**Project Definition:** The project involves analyzing COVID-19 cases and deaths data using IBM Cognos. The objective is to compare and contrast the mean values and standard deviations of cases and associated deaths per day and by country in the EU/EEA. This project encompasses defining analysis objectives, collecting COVID-19 data, designing relevant visualizations in IBM Cognos, and deriving insights from the data.

**Design Thinking:**

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 deviations of cases and deaths.

**1.Analysis Objectives:**

When defining the specific objectives of your analysis, consider the following points:

* Comparing Mean Values: Specify what you aim to achieve by comparing mean values of COVID-19 cases and deaths per day and by country. Are you looking for trends over time, differences between countries, or something else?
* Standard Deviation Comparison: Clarify the purpose of comparing standard deviations. Are you interested in assessing the variability in cases and deaths data? Is it to identify countries or time periods with particularly high or low variability?
* Hypotheses or Research Questions: If applicable, outline any hypotheses or research questions that you intend to address through your analysis. For example, "Is there a significant difference in the standard deviation of COVID-19 cases between EU/EEA countries?"

**2.Data Collection:**

Ensure a smooth data collection process by following these steps:

* Data Source: Clearly specify where you will obtain the COVID-19 cases and deaths data. Mention the source, such as a government health agency, a reliable dataset repository, or a research organization.
* Data Scope: Define the scope of the data you'll collect, including the time period (start and end dates) and the specific countries within the EU/EEA you intend to analyze. Ensure that the data is regularly updated to maintain accuracy.
* Data Format: Describe the format in which the data is available (e.g., CSV, Excel, API). Ensure that you have the necessary permissions or licenses to access and use this data.

**3.Visualization Strategy:**

To effectively visualize mean values and standard deviations, consider these points:

* Choice of Visualization Types: Decide on the most suitable types of charts or graphs for your data. For example, line charts for time series data, bar charts for country comparisons, and error bars for standard deviations.
* Visual Design: Consider the visual design principles to make your charts informative and easy to understand. Pay attention to color choices, axis labels, legends, and titles.
* Interactivity: Explore the interactive capabilities of IBM Cognos to allow users to explore the data and gain insights interactively. This might involve drill-down features, filters, or tooltips.

**4.Insights Generation:**

Identifying potential insights is a crucial part of your project:

* Key Metrics: Outline the specific metrics or indicators you'll use to generate insights. For instance, you may focus on daily case growth rates, hotspots, or the relationship between cases and deaths.
* Trends and Anomalies: Consider what trends, patterns, or anomalies you expect to find in the data. These could include spikes in cases, regional disparities, or changes in trends over time.
* Implications: Think about the implications of your findings. How might these insights be useful for public health decision-makers or the general public? Are there any actionable recommendations that can be derived from your analysis?
* Data Storytelling: Plan how you will communicate your insights effectively. This might involve creating a narrative that guides readers or viewers through the data story, highlighting key findings along the way.

**Innovation - Data Segmentation by Time Periods or Countries**

**1.Introduction**

* In this phase of the project, we dive deep into the analysis of COVID-19 cases and deaths data in the EU/EEA region.
* The primary objective is to compare and contrast the mean values and standard deviations of cases and associated deaths per day and by country.
* By incorporating innovative approaches, data segmentation, and leveraging IBM Cognos, our aim is to derive actionable insights crucial for strategic decision-making.

# 2.Data Segmentation

To ensure a meticulous analysis, the dataset will be segmented by both time periods and countries.

Time Periods*:*

* + Segmentation into daily, weekly, and monthly intervals enables us to capture short-term fluctuations and long-term trends effectively.

Countries:

* + Individual analysis of EU/EEA countries provides a detailed understanding of regional variations, aiding in targeted interventions.

# 3.Mean and Standard Deviation Analysis

## Calculation Steps:

Mean Calculation:

For each country and time period, calculate the average number of daily COVID-19 cases and deaths. This provides a central value representing the typical daily occurrence.

Formula:

Mean = (Sum of values) / (Number of days)

Standard Deviation Calculation*:*

Determine the variation or dispersion of data points from the mean for each country and time period.

Formula:

Standard Deviation = √[(Σ(xi - μ)²) / N], where xi is each data point, μ is the mean, and N is the number of data points.

## Visualization Techniques:

Bar Charts with Error Bars:

* + - Utilize bar charts to represent mean values for cases and deaths, and overlay them with error bars representing the standard deviations.
    - This visual representation provides a clear comparison across countries and time periods.

Scatter Plots:

* + - * Use scatter plots to display individual data points against the mean.
      * This visualization helps identify outliers and understand the data distribution around the mean.

# 4.Innovative Approaches

## Advanced Visualization Techniques:

Dynamic Comparison Charts*:*

* + - Interactive charts enabling users to compare mean values and standard deviations dynamically.
    - Users can select specific countries or time periods for focused analysis.

Regression Analysis:

* + - Implement regression models to identify trends over time, providing insights into the trajectory of COVID-19 cases and deaths.
    - This aids in forecasting and trend prediction.

## Predictive Analytics:

Machine Learning Models:

* + - Utilize machine learning algorithms for predictive modeling.
    - Algorithms like Random Forest or LSTM can forecast future mean values and standard deviations based on historical data, providing valuable insights for proactive planning.

# 5.Insights and Decision Support

Through meticulous mean and standard deviation analysis, coupled with innovative techniques, we aim to extract actionable insights:

Identification of High-Risk Periods:

* + Pinpoint specific days or weeks with unusually high standard deviations, indicating periods of significant volatility in COVID-19 cases.
  + This insight aids in resource allocation and preparedness.

Comparative Analysis:

* + Compare mean values and standard deviations between countries to identify variations in pandemic management strategies.
  + Understanding the differences can provide valuable lessons for other regions.

Predictive Insights:

* Utilize machine learning predictions to foresee potential spikes in cases or deaths. Timely intervention during these periods can significantly impact the outcome of the pandemic.

**Additionally we included ML algorithm**

# 6.Machine learning algorithm for predictive analysis

Linear regression is a suitable algorithm for predicting COVID-19 cases and deaths when you want to understand the relationship between one or more independent variables (features) and the number of cases or deaths. In your case, you can use it to predict the number of cases or deaths based on specific variables from the provided dataset. Here's how you can implement linear regression using IBM Cognos and the given dataset:

Steps for Implementing Linear Regression in IBM Cognos:

## Data Loading:

Import the COVID-19 cases and deaths dataset into IBM Cognos. Ensure the dataset is cleaned, and missing values are handled appropriately.

## Data Exploration:

* + Explore the dataset to understand the available variables and their distributions.
  + Identify potential independent variables (features) that could influence the number of cases or deaths. These variables could include factors like population density, healthcare facilities, government policies, etc.

## Data Preparation:

* + Select the relevant independent variable(s) (features) and the target variable (number of cases or deaths).
  + Split the data into training and testing sets. A common split ratio is 80% for training and 20% for testing.

## Model Building:

* + In IBM Cognos, navigate to the modeling section and choose Linear Regression as the algorithm.
  + Select the independent variable(s) as input features and the number of cases or deaths as the target variable.
  + Train the model using the training dataset.

## Model Evaluation:

* + Evaluate the model using the testing dataset to assess its performance.
  + Common evaluation metrics for regression models include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (coefficient of determination).

## Interpretation and Visualization:

* + Interpret the coefficients of the linear regression equation to understand the impact of each independent variable on the number of cases or deaths.
  + Visualize the regression line along with the actual data points to observe the fit of the model.

## Iterative Refinement:

Based on the model performance, you might need to iterate: include more features, handle outliers, or try other regression algorithms if linear regression does not provide satisfactory results.

## Prediction:

Once you are satisfied with the model, you can use it for predicting the number of COVID-19 cases or deaths based on new data or future time periods.

**PROGRAM**

import pandas as pd

# Read the dataset

df = pd.read\_csv('Covid\_19\_cases4.csv')

print(df)

df = pd.read\_csv('Covid\_19\_cases4.csv')

# Check for missing values

print(df.isnull().sum())

# Drop rows with missing values

df = df.dropna()

# Remove duplicate rows

df = df.drop\_duplicates()

# Save the cleaned dataset

df.to\_csv('Tiger.csv', index=False)

OUTPUT

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

... ... ... ... ... ... ... ...

2725 06-03-2021 6 3 2021 3455 17 Sweden

2726 05-03-2021 5 3 2021 4069 12 Sweden

2727 04-03-2021 4 3 2021 4884 14 Sweden

2728 03-03-2021 3 3 2021 4876 19 Sweden

2729 02-03-2021 2 3 2021 6191 19 Sweden

dateRep 0

day 0

month 0

year 0

cases 0

deaths 0

countriesAndTerritories 0

dtype: int64

**Data visualization for Covid\_19\_cases analysis**

**Program:**

import pandas as pd

import matplotlib.pyplot as plt

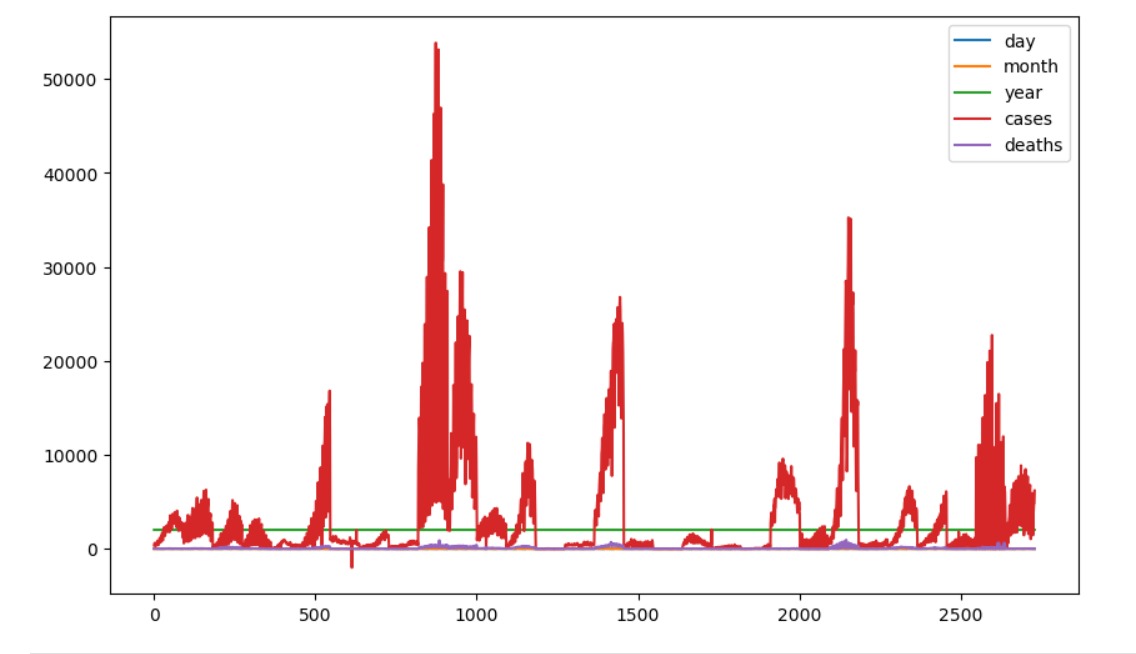
df=pd.read\_csv("Covid\_19\_cases4.csv")

df.head()

df.info()

df.plot(figsize=(10,6))

**OUTPUT:**

****

**PROGRAM**

import pandas as pd

import matplotlib.pyplot as plt

from statsmodels.tsa.seasonal import seasonal\_decompose

from statsmodels.graphics.tsaplots import plot\_pacf

from statsmodels.tsa.arima\_model import ARIMA

import statsmodels.api as sm

data = pd.read\_csv("Covid\_19\_cases4.csv")

print(data.head())

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

print(data.info())

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2730 entries, 0 to 2729

Data columns (total 7 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 dateRep 2730 non-null object

1 day 2730 non-null int64

2 month 2730 non-null int64

3 year 2730 non-null int64

4 cases 2730 non-null int64

5 deaths 2730 non-null int64

6 countriesAndTerritories 2730 non-null object

dtypes: int64(5), object(2)

memory usage: 128.0+ KB

None

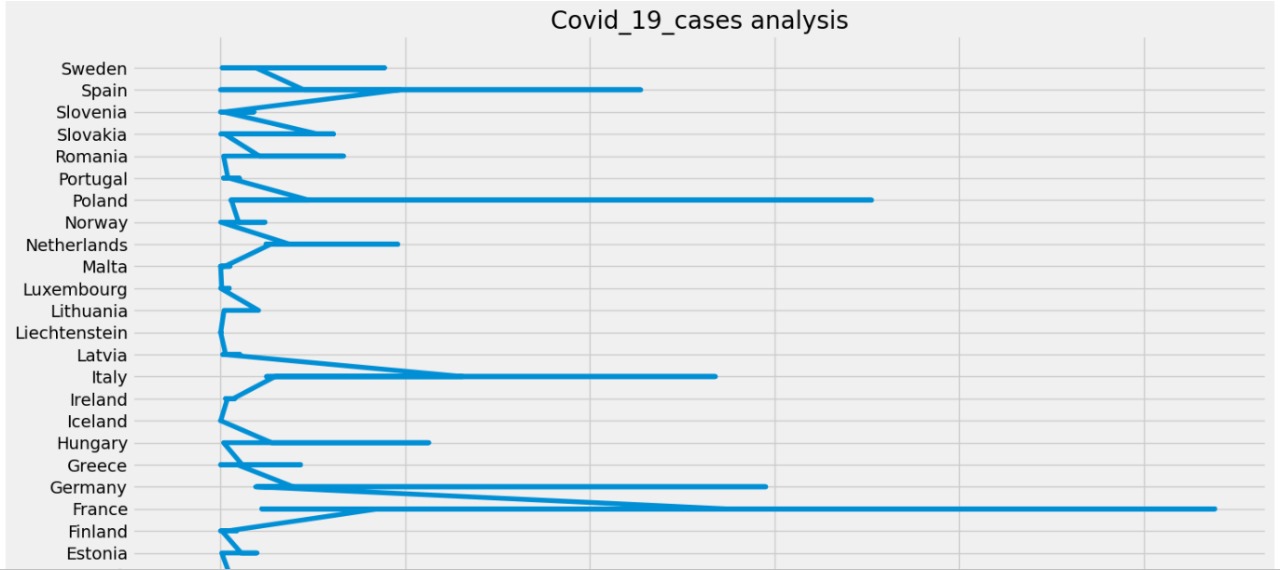
plt.style.use('fivethirtyeight')

plt.figure(figsize**=**(15, 10))

plt.plot(data["cases"], data["countriesAndTerritories"])

plt.title("Covid\_19\_cases analysis")

plt.show()



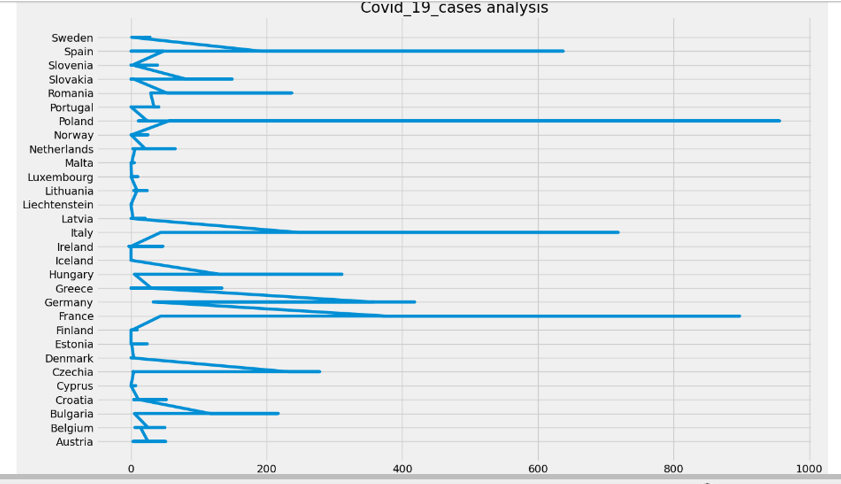
plt.style.use('fivethirtyeight')

plt.figure(figsize**=**(15, 10))

plt.plot(data["deaths"], data["countriesAndTerritories"])

plt.title("Covid\_19\_cases analysis")

plt.show()



**TO visualize and compare the mean values and standard deviations of COVID-19 cases associated with deaths**

import matplotlib.pyplot as plt

import numpy as np

# Sample COVID-19 data (replace this with your actual data)

cases = [100, 200, 300, 400, 500, 600]

deaths = [10, 20, 30, 40, 50, 60]

# Calculate mean and standard deviation

cases\_mean = np.mean(cases)

deaths\_mean = np.mean(deaths)

cases\_std = np.std(cases)

deaths\_std = np.std(deaths)

# Create bar plots for means

plt.bar(['cases', 'deaths'], [cases\_mean, deaths\_mean], color=['blue', 'red'])

plt.title('Mean Values of COVID-19 Cases and Deaths')

plt.ylabel('Mean Value')

plt.show()

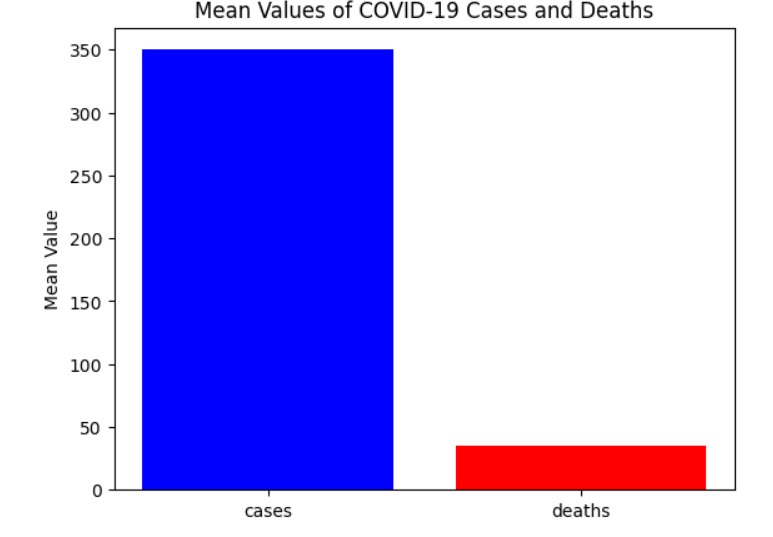
# Create bar plots for standard deviations

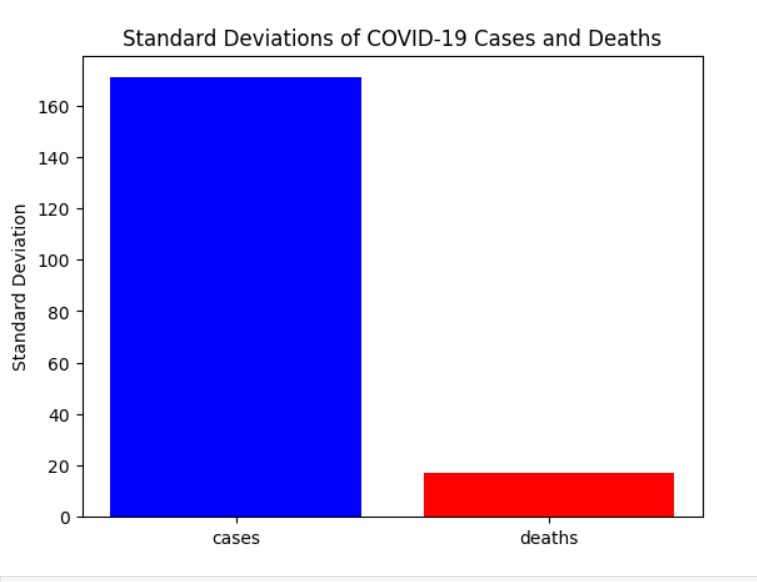
plt.bar(['cases', 'deaths'], [cases\_std, deaths\_std], color=['blue', 'red'])

plt.title('Standard Deviations of COVID-19 Cases and Deaths')

plt.ylabel('Standard Deviation')

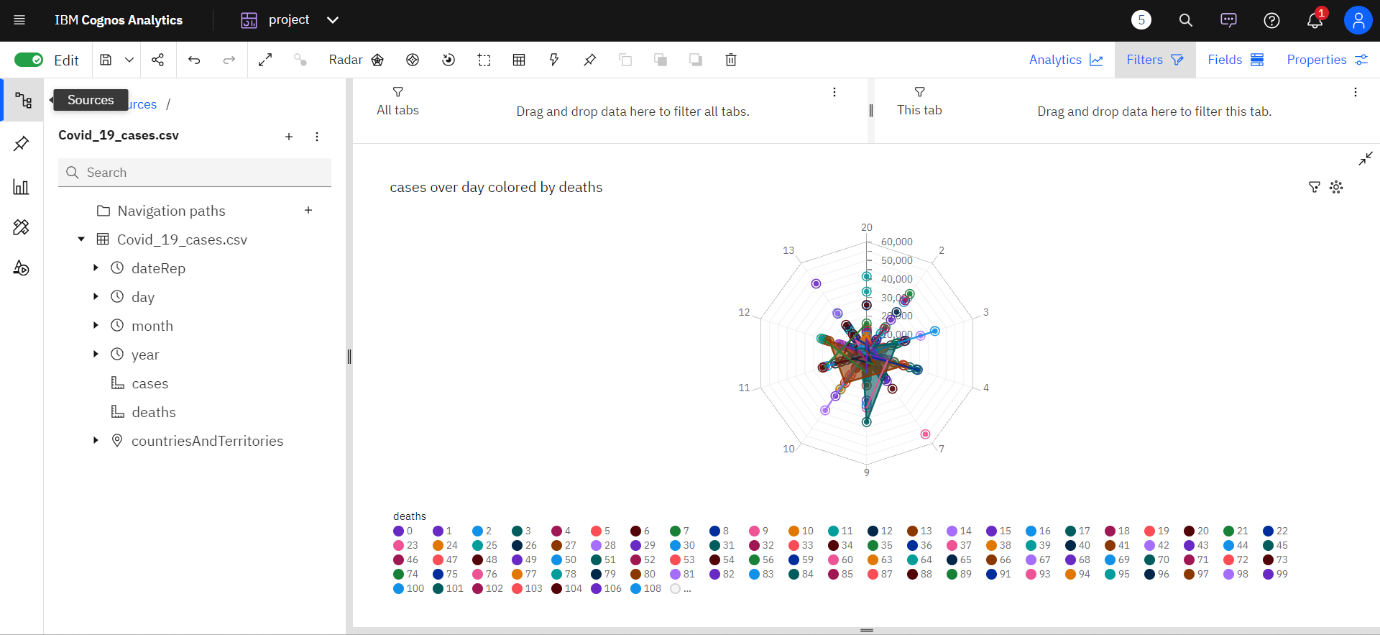
plt.show()



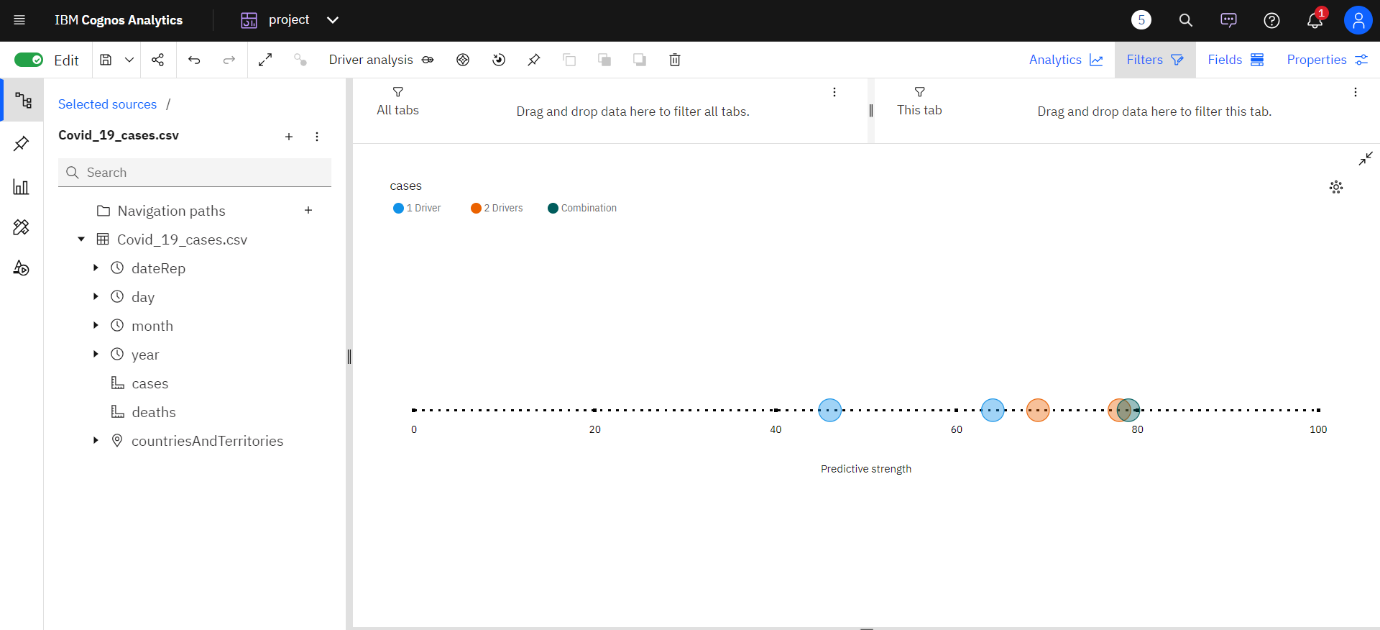


**Data visualization using IBM Cognos:**

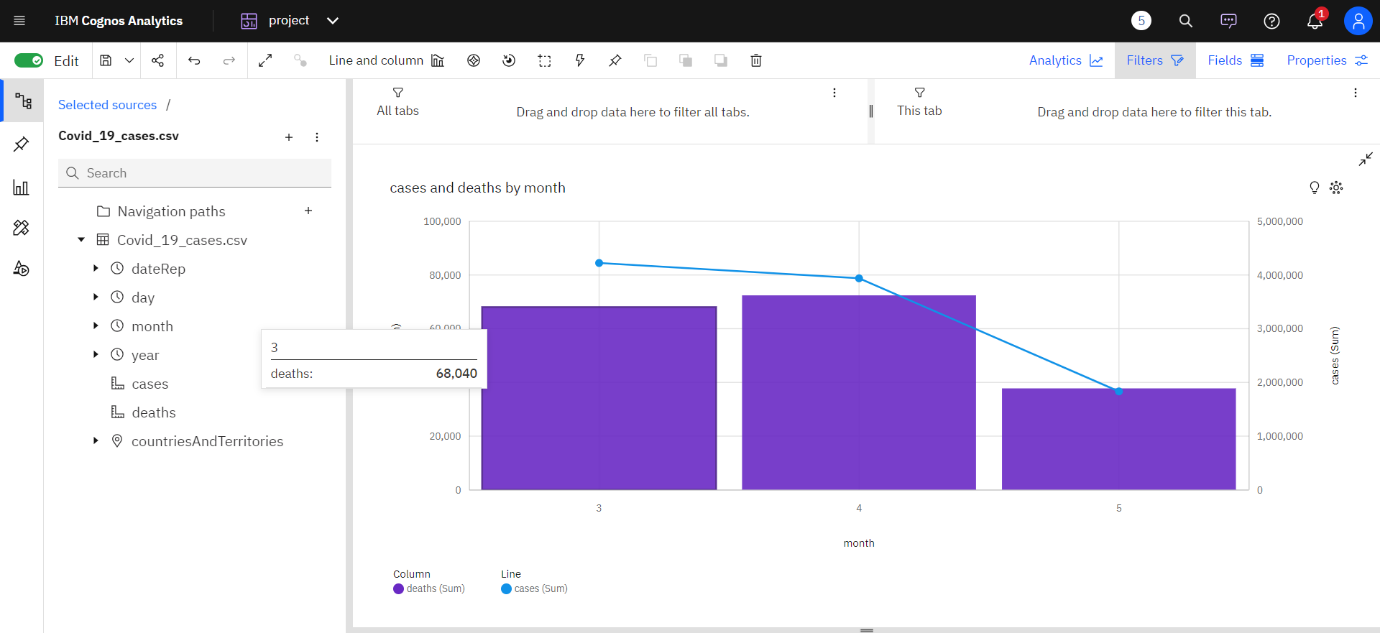
**Visualization of cases over day colored by deaths**



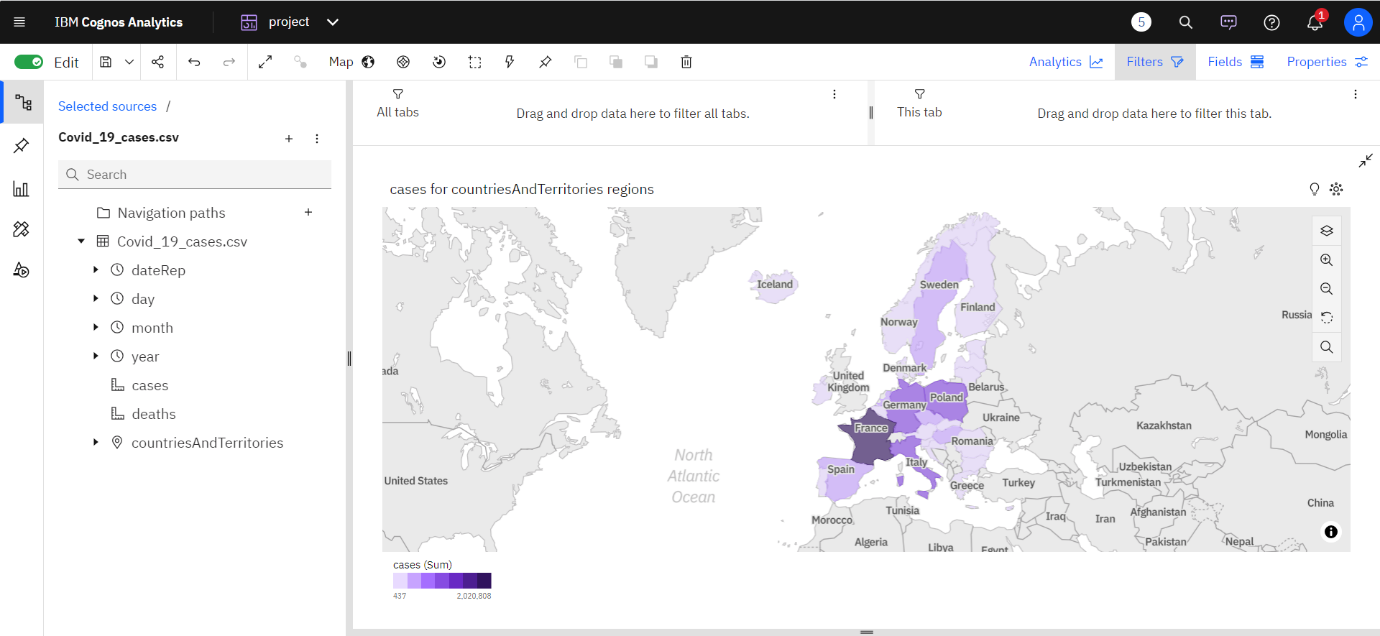
**Visualization by cases**



**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.

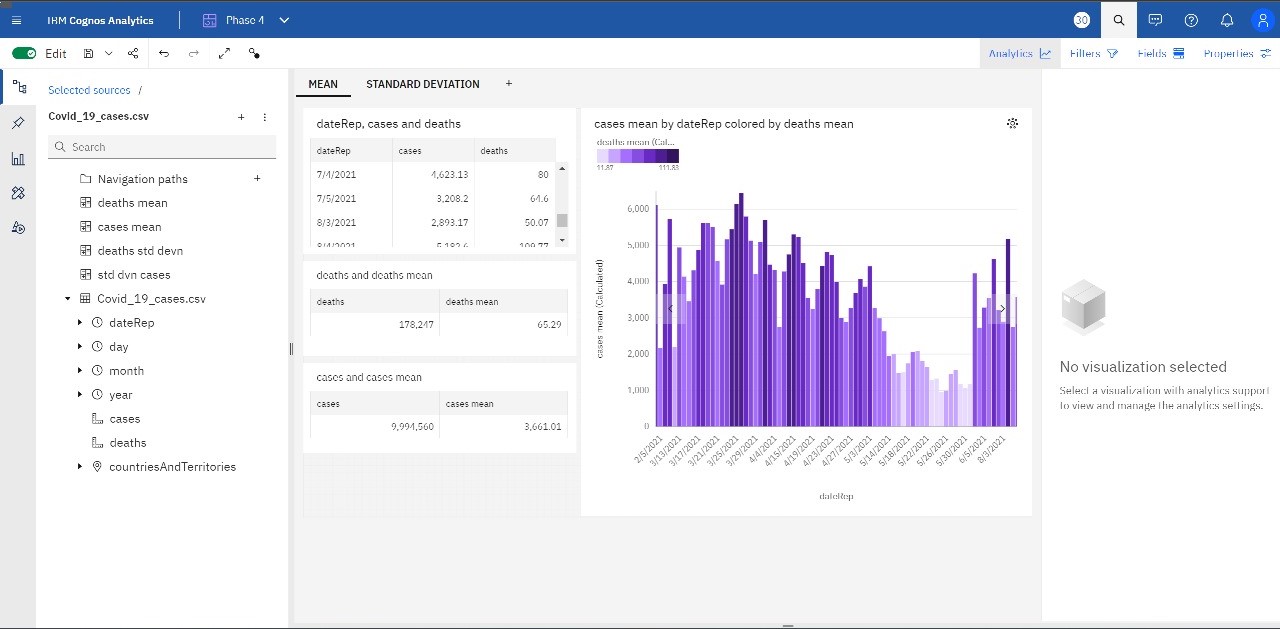
**Run the Query:**

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**

