Phase 5

PROJECT DOCUMENTATION & SUBMISSION

Date	**1-11-2023**
Team ID	**720**
Project Name	**COVID 19 CASE ANALYSIS**

Project Title: COVID 19 Cases Analysis Using Data Analytics Tool

Problem Statement

Designing a project to analyze COVID-19 cases and deaths using IBM cognos, The objective is to compare and contrast the mean and standard deviation of cases and deaths, which is a valuable undertaking. This project will involve data analysis, visualization, and deriving insights from the data.

Problem identified:

The global COVID-19 pandemic has created an urgent need for comprehensive and data-driven analysis of cases to inform public health policies and responses. As the virus's impact evolves, accurate and timely analysis of infection rates, transmission patterns, and vaccination effectiveness is paramount. However, the vast and dynamic nature of COVID-19 data, coupled with the need to integrate diverse sources of information, presents a complex challenge. Developing an accessible, robust, and up-to-date system for COVID-19 cases analysis is crucial to facilitate evidence-based decision-making and resource allocation.

Introduction:

The COVID-19 pandemic has had a profound impact on societies, healthcare systems, and economies worldwide. As the pandemic continues to evolve, it is crucial to conduct a comprehensive case analysis to better understand, manage, and mitigate its effects. This problem statement outlines the key aspects of the analysis required to address the on going challenges posed by COVID-19

The task at hand is to design and implement a comprehensive COVID-19 case

analysis system for tracking, visualizing, and deriving insights from pandemic data

LITERATURE SURVEY

- **1. 4.“A Mathematical model for covid-19 transmission dynamics with a case study of India”, Piu Samui, Jayanta Mondal, Subhas Khajanchi[August-2020]**

As of July 15, 2020, there were 968,117 confirmed cases, 612,782 recovered cases, and 24,915 deaths in India due to the ongoing COVID-19 pandemic, which has caused a serious global crisis. Predictive mathematical models can be used to manage and control coronavirus disease in the absence of effective treatments or medications and in the absence of an established epidemiological life cycle. This study uses epidemic data up to April 30, 2020, to propose a mathematical model to predict and control the COVID-19 pandemic's transmission dynamics in India. To examine model simulations and predictions, the model computes the fundamental reproduction number, or R0. To ascertain the relative significance of model parameters to the spread of disease, sensitivity analysis is performed. According to the model, COVID-19 peak in India is expected to be higher.

2. Data Analytics for Predicting COVID-19 Cases in Top Affected Countries: Observations and Recommendations (abdelrahman EE) 2020

3.

The main objectiveof this study is to develop a prediction method that accurately forecast the number of COVID-19 cases by incorporating both historical data and external factors influencing the spread of the virus. By utilizing a Nonlinearauto regressive exogenous input (NARX) neural network-based algorithm, the study aims to improve Prediction Accuracy compared to existing methods.The methodology involves using a nonlinear autoregressive exogenous input (NARX) neural network-based algorithm. This algorithm considers both historical data of COVID-19 cases and external factors that influence the virus's spread. By training the NARX neural network on these combined datasets, it becomes capable of making more accurate predictions compared to methods that only rely on historical data. In conclusion, the study

highlights the need for accurate COVID-19 case prediction due to its global impact. By incorporating historical data and external factors through a NARX neural

network-based algorithm, the study improves prediction accuracy compared to existing methods. These predictions aid governments and affected populations in making informed decisions and resuming normal activities following the pandemic.

4. corona virus dieases (covid19)cases analysis using machine learning applications

The study aims to assess the impact of machinelearning onCOVID19research. By reviewing studies from 2020, it examines how machine learning can aid in investigating, predicting, and distinguishing COVID-19 cases. The objective is to understand its role in healthcare programs, specifically in assessing and prioritizing COVID-19 cases. It suggests using recurrent supervised learning for even better accuracy in the futureThe study's methodology involves searching for relevant articles published in 2020 using specific keywords on various platforms. After selecting 14 suitable articles, data is extracted and analyzed to understand how machine learning impacts COVID-19 research. The study compares supervised The study concludes that machine learning plays a vital role in COVID-19 research, assisting in investigation, prediction, and case differentiation. Supervised learning performs better than unsupervised learning, with potential for improved accuracy using recurrent supervised learning in the future.

4.“Covid-19 vaccines and current opinion in immunology”, Duduzile Ndwandwe and Charles S Wiysonge[July-2021]

COVID-19 is a pandemic of unprecedented proportions, with nearly 200 million confirmed cases and four million deaths worldwide. Efforts have been made to find safe and effective vaccines, with 184 candidates in pre-clinical development, 105 in clinical development, and 18 approved for emergency use. These vaccines include whole virus live attenuated or inactivated, protein-based, viral vector, and nucleic acid vaccines. By mid-2021, three billion doses of COVID-19 vaccine have been administered worldwide, primarily in high-income countries. COVID-19 vaccination provides hope for an end to the pandemic, if equal access and optimal

uptake are provided in all countries. Vaccines are biological preparations that provide active acquired immunity to a particular infectious disease by stimulating an immune response to an antigen. The pandemic has accelerated vaccine development, with 184 candidate vaccines in preclinical development and 104 in clinical stage.

- **5. "Clinical Microbiology and infection", Anosmia in covid –19 patients, Burghart Sniffin[May-2020]**

Burghart Sniffin' Sticks®, a popular screening tool for smelling disorders, was utilized in a study to measure the severity of SARS-CoV-2-related smelling disorder in COVID-19 patients. According to the study, anosmia was diagnosed in 18 out of 45 patients (40%) with COVID-19 patients smelling four fewer sticks on average than uninfected controls. When it came to identifying anosmia, the Sniffin' Stick test was more accurate than self-reporting or obtaining a medical history. Patients with and without anosmia or hyposmia had comparable clinical pictures, test results, and outcomes at day 15. A quantitative and objective test revealed that 40% of the patients were hyposmia and nearly half were osmic. But only 49% of patients said they can smell.

DESIGN THINKING

Empathize:

Before diving into solving the problem, it's essential to empathize with the context and potential stakeholders. In this case, the stakeholders may include public health authorities and policymakers. Understanding their concerns and objectives is crucial.

Actions:

- Research the goals and objectives of public health authorities regarding COVID-19 data analysis.

- Consider the questions policymakers might have about the spread and impact of the virus.

- Identify the key variables of interest (new cases and deaths) and any additional data that might be relevant.

Define:

Based on our understanding of the problem and stakeholders' needs, we will define clear objectives and success criteria for our analysis.

Objectives:

- Compare the mean values of new cases and deaths reported per day in EU/EEA countries.

- Contrast the standard deviations of new cases and deaths.

- Provide clear and insightful visualizations of the analysis results.

Ideate:

Brainstorm potential solutions and approaches to analyze COVID-19 data and perform statistical comparisons.

Actions:

- Explore descriptive statistics techniques to calculate means and standard deviations.

- Consider hypothesis testing to assess if there are significant differences among countries.

- Think about data visualization methods such as bar charts, box plots, or heatmaps to illustrate the findings.

Prototype:

Create a prototype of the data analysis workflow, including data preprocessing, statistical tests, and visualization.

Actions:

- Clean and preprocess the COVID-19 data to prepare it for analysis.

- Perform statistical tests (e.g., ANOVA, t-tests) to compare means and standard deviations.

- Create initial visualizations to check the feasibility and effectiveness of data presentation.

Test

Evaluate the analysis prototype, ensuring it meets the defined objectives and provides meaningful insights.

Actions:

- Review the statistical results and visualizations to ensure they address the objectives.

- Collect feedback from potential stakeholders to identify any necessary improvements.

- Refine the analysis based on feedback and iterate if needed.

Implement:

Once the analysis prototype meets the objectives and stakeholder requirements, proceed with full implementation.

- Apply the analysis workflow to the complete COVID-19 dataset for EU/EEA countries.

Key Challenges:

1. Data Quality: Ensuring the dataset is clean, complete, and free of errors.

2. Data Preparation: Aggregating and preprocessing the data for meaningful analysis.

3. Statistical Analysis: Conducting appropriate statistical tests to compare mean values and standard deviations.

4. Visualization: Creating visualizations to present the analysis results effectively.

5. Interpretation: Drawing meaningful insights from the analysis and making recommendations if necessary.

Actions:

- Generate final visualizations and statistical reports.

- Document the analysis process and results for transparency.

Iterate

Continuous improvement is vital. Collect feedback and iterate on the analysis if new data or questions arise.

Actions:

- Monitor updates to COVID-19 data and rerun the analysis periodically.

- Update visualizations and insights to reflect the most recent data.

- Be prepared to provide additional analyses or insights as requested by stakeholders.

Design and Innovation Strategies

Data Collection and Feature Engineering

Innovation: Comprehensive Data Compilation

Utilize data cleaning techniques to ensure data accuracy. Integrate data from diverse sources, including official health organizations and government databases, to create a robust dataset for analysis.

Statistical Analysis

Innovation: Advanced Statistical Methods

Apply robust statistical methods, including ANOVA (Analysis of Variance) and Tukey's test, to compare means and identify significant differences in Covid-19 cases and deaths among different countries.

Implement bootstrapping techniques to calculate reliable standard deviations and confidence intervals for the mean values.

Visualization and Interpretation.

Innovation: Interactive Data Visualization

Linear regression is used in COVID-19 analysis to model relationships between variables, such as infection rates and factors like population density or public health measures. By fitting a linear equation to the data, researchers can identify trends and make predictions, aiding in understanding the impact of various factors on the spread and severity of the virus.

Develop interactive visualizations such as heat maps, box plots, and trend graphs to present the comparative analysis of mean values and standard deviations.

Implement tooltips and interactive elements to allow users to explore specific data points and gain detailed insights.

Temporal Analysis.

Innovation: Time Series Forecasting

Apply time series forecasting techniques, such as ARIMA (AutoRegressive Integrated Moving Average) and Prophet, to predict future trends in Covid-19 cases and deaths.

Utilize historical data to validate the accuracy of the forecasting models and provide reliable predictions for decision-makers.

Communication of Results

1. **Finalize Report Layout:**

Review and finalize the layout of your report. Ensure that all elements, such as tables, charts, and text, are properly arranged and formatted.

2. **Formatting:**

Apply formatting options to enhance the report's appearance. You can format fonts, colors, borders, and backgrounds to make the report visually appealing and easy to read.

3. Report Styling:

Apply a consistent styling theme to your report to maintain a unified look and feel. Cognos Analytics provides predefined styles that you can choose from or customize.

4. Testing and Validation:

Test your report to ensure that all data elements are displayed correctly. Verify that data calculations, aggregations, and filters are working as expected.

5. Save and Publish:

Save your report in Cognos Analytics and publish it to the appropriate location, such as a shared folder, report server, or email distribution list.

6. Security and Permissions:

Set up security and permissions for the report to control who can access, view, or modify it. Define user roles and access levels as needed.

7. Documentation:

Document the report creation process, including data sources, calculations, and any customizations.

5. ANALYSIS OBJECTIVE

The analysis objective appears to be centered on calculating the mean (average) and the corresponding standard deviation for a dataset. The standard deviation measures the spread or dispersion of the data points around the mean. In the context of this analysis, it's important to determine the level of variability in the data.

Calculate the Mean: This involves adding up all the data points and dividing by the total number of data points. The mean provides a measure of the central tendency of the data.

Calculate the Standard Deviation: The standard deviation quantifies how individual data points deviate from the mean. A higher standard deviation indicates greater data variability, while a lower standard deviation suggests data points are closer to the mean.

Interpret the Results: Once you have the mean and standard deviation, you can analyze the data distribution. If the data follows a normal distribution, the mean and standard deviation provide essential information about the data's characteristics.

In addition to the mean and standard deviation, you might also want to consider other statistics, depending on the specific objectives of your analysis, such as median, quartiles, skewness, and kurtosis. These additional statistics provide a more comprehensive understanding of the data's distribution and shape.

DATA COLLECTION

- The data collection process for COVID-19 typically involves the following steps:
- Source Identification: Identify sources such as health authorities, hospitals, and laboratories that report COVID-19 data.
- Data Variables: Define the specific data variables to collect, including cases, deaths, recoveries, testing, and demographics.
- Data Points: Collect daily or periodic data points, including location, date, and case status.
- Reporting Methods: Establish reporting methods, which can be manual, electronic, or through online portals.
- Standardization: Ensure data is reported in a standardized format, following national or international guidelines.
- Privacy Considerations: Protect individuals' privacy and adhere to data protection regulations.
- Quality Control: Implement data validation and quality control measures to minimize errors.
- Data Aggregation: Aggregate data at various levels (local, regional, national) for analysis.
- Data Transparency: Publish data in a transparent and accessible manner for the public and researchers.
- Data Sharing: Collaborate with other organizations to share data for a comprehensive picture.
- Historical Data: Maintain historical data for trend analysis and research.
- Data Security: Secure data to prevent breaches and maintain public trust.
- Data Analytics: Employ data analytics tools to identify patterns, hotspots, and potential outbreaks.

- Public Communication: Disseminate findings and updates to the public through official channels.
- Research Use: Encourage researchers to utilize the data for studies and modelling.
- Feedback Loops: Establish feedback mechanisms with data providers to improve data collection.
- International Collaboration: Collaborate with other countries to monitor global trends and travel-related cases.
- Data Storage: Store data securely for long-term research and policymaking.
- Policy Adaptation: Use data insights to adapt public health policies and interventions.
- Continuous Improvement: Continuously refine the data collection process to respond to evolving needs and challenges.
- This process helps in tracking and understanding the COVID-19 pandemic's dynamics and informs public health responses.

Project Objectives:

➢ **Data Collection**: Gather covid19 data from from reliable sources across the globe. This data will include parameters such as cases, deaths, countries/Territories, and other relevant variables.

➢ **Data Analysis:** Perform exploratory data analysis (EDA) to understand the distribution of covid cases, detect outliers and identify trends and patterns.

➢ **Visualization:** Utilize data visualization techniques to represent covid19 data geospatially and temporally. This will help identify hotspots areas and understand trends over time.

➢ **Identification of Highly-Affected Areas:** Determine areas with consistently high case levels and investigate the factors contributing to this covid trend.

- **Predictive Model:** Develop a predictive model, likely using machine learning techniques, to estimate case and death rate . This model will be valuable for forecasting covid19 spreads, fatalities and identifying hotspot areas that require immediate attention.

Steps Involved in Model Evaluation:

Data Collection:

First, ensure you have access to COVID 19 case data ,Gather data from reliable sources such as government health agencies, the World Health Organization (WHO), and reputable research institutions. Collect data on the number of cases, deaths, recoveries, vaccination rates, and other relevant variables .

import Libraries:

Start by importing the necessary libraries such as numpy, pandas for data manipulations, matplotlib and seaborn for visualisations etc

IMPORT LIBRARIES

```
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import plotly as py
import plotly.graph_objects as go
import plotly.io as pio
import plotly.express as px
```

Load the Dataset:

- This step involves loading your COVID 19 dataset into your Python environment.

The dataset should be in a format that Pandas can easily handle, such as a CSV file.

LOADING DATASET

```
covid_data=pd.read_csv('Covid_19_cases4 (1).csv')
```

- The read_csv () function is used to load a CSV (Comma-Separated Values) file into a Pandas DataFrame. You specify the file path within the parentheses.

- The result of this operation is a DataFrame, which is a tabular data structure that's similar to a spreadsheet. It allows you to work with your data in a structured and flexible way.

Explore the Dataset

Before diving into data preprocessing, it's important to understand your dataset. You can use various Pandas functions to explore it:

data.head ():

- This function displays the first few rows of your dataset, giving you a glimpse of its structure.

```
covid_data.head()
```

Data.describe():

- It provides basic statistical information about your data, including measures like mean, standard deviation, and quartiles for numerical columns.

```
covid_data.describe()
```

data.columns():

- This helps you see the names of all the columns in your dataset

```
covid_data.columns
```

Data.info():

- This method prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.

Data Pre-proccessing:

- Data preprocessing is crucial for ensuring the quality and usability of your data:

```
covid_data.info()
```

Handle Missing Values:

• Check for missing values in your dataset and decide on an appropriate strategy to handle them. You can fill missing values using methods like forward-fill, backward fill, mean, median, or simply remove rows with missing values.

```
covid_data.isnull().sum()
```

```
covid_data.isnull().any()
```

Data Transformation:

• If your dataset contains date or time columns, convert them to the datetime data type for time-based analysis.

```
"# Example: Convert a date column to datetime data['Date'] = pd.to_datetime (data['Date'])"
```

Data Cleaning:

• Inspect your data for inconsistencies, outliers, or irregularities. Ensure that the data is clean and standardized. This may include dealing with irregular units, correcting typos, or removing duplicates.

Predictive Model training:

- Choose Support Vector Machine (SVM) for regression and classification tasks, handling complex data relationships.

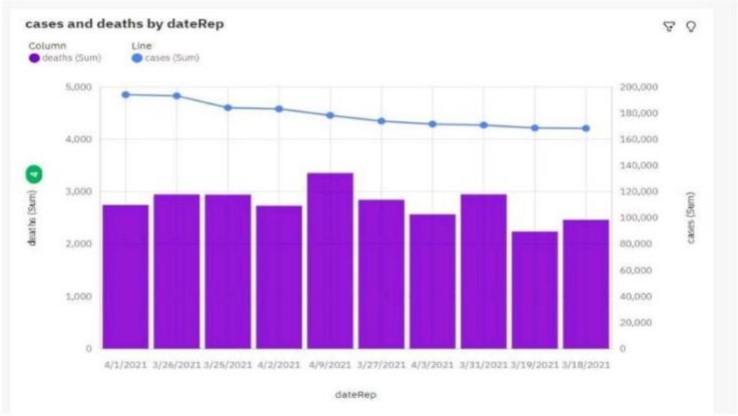
- Split the data into training and testing sets

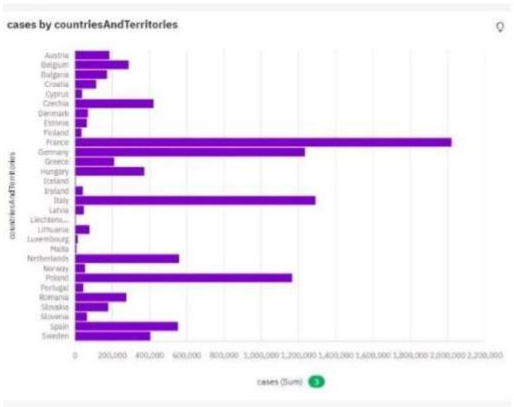
.

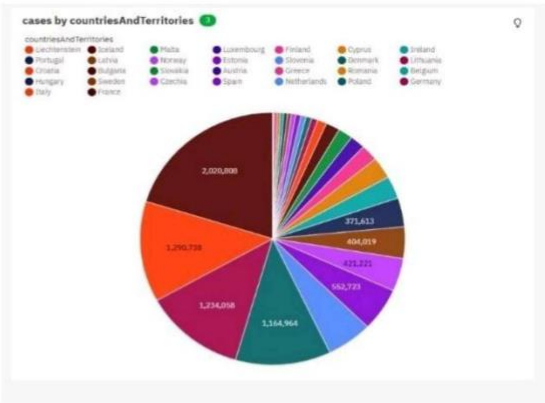
- Train the model using preprocessed dataset and target variables.

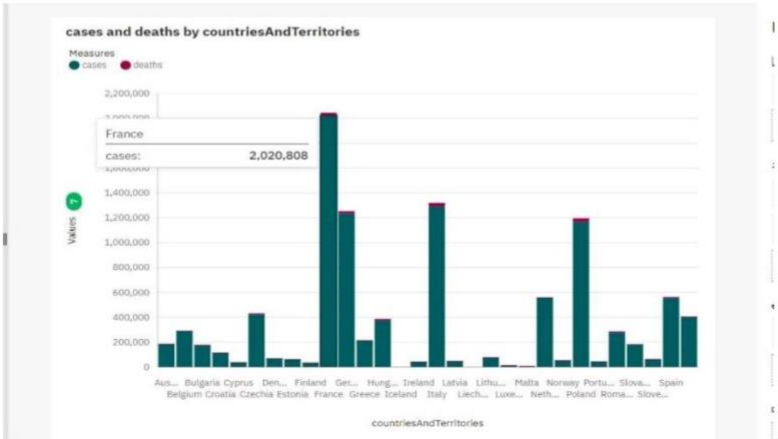
- Train the model on the training data and evaluate its performance on the test data using relevant metrics (e.g., Mean Absolute Error, Root Mean Squared Error)

Visualization using IBM Cognos:





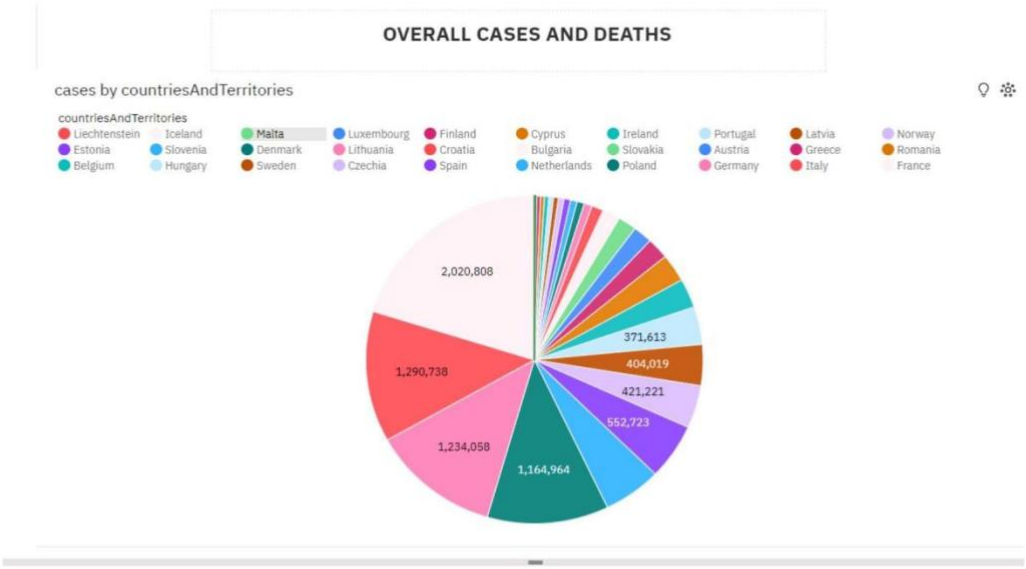




Visualisation using IBM Cognos and insights:

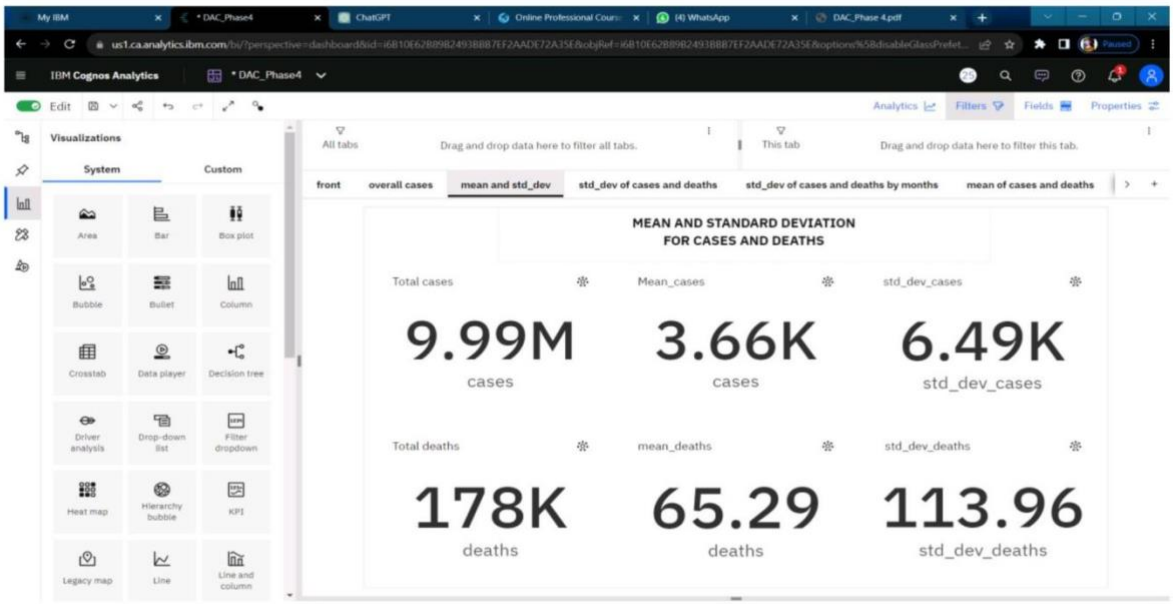
3.1.Home

3.2.Visualisation for overall cases and analyses:



In this pie-chart, various countries and territories, the total number of cases of COVID-19 is significantly high, with France having the highest number of cases, exceeding 2.0 million, while Liechtenstein has the lowest, with 437 cases. Notably, on 2021-03-29 to 2021-03-30, France experienced a staggering 937% increase in cases, indicating a significant surge in infections within a short timeframe. It is projected that by 2021-06-19, France will surpass Germany in cases by more than 14 thousand, highlighting the severity of the situation in France. Overall, the cumulative cases across all countries and territories are nearing 10.0 million, underscoring the global impact of the pandemic.

3.3.Visualisation for total Mean and Standard Deviation to analysis cases and deaths:



For Cases:

For cases, there is a noticeable strong weekly trend, with the highest values typically occurring on Thursdays and the lowest on Mondays. Additionally, there is a moderate downward trend in the number of cases. Notably, on 2021-04-06 and 2021-04-07, there were unusual spikes in cases, with a 69% increase in just one day. The lowest average cases were reported on 2021-05-25 at 953.87 and 2021-05-26 at 989.0, while the highest average cases were observed on 2021-04-01 at 6467.87 and 2021-03-26 at 6438.93. According to current forecasts, cases are expected to reach almost 1500 by 2021-06-19, and the dataset contains over 2500 results for cases.

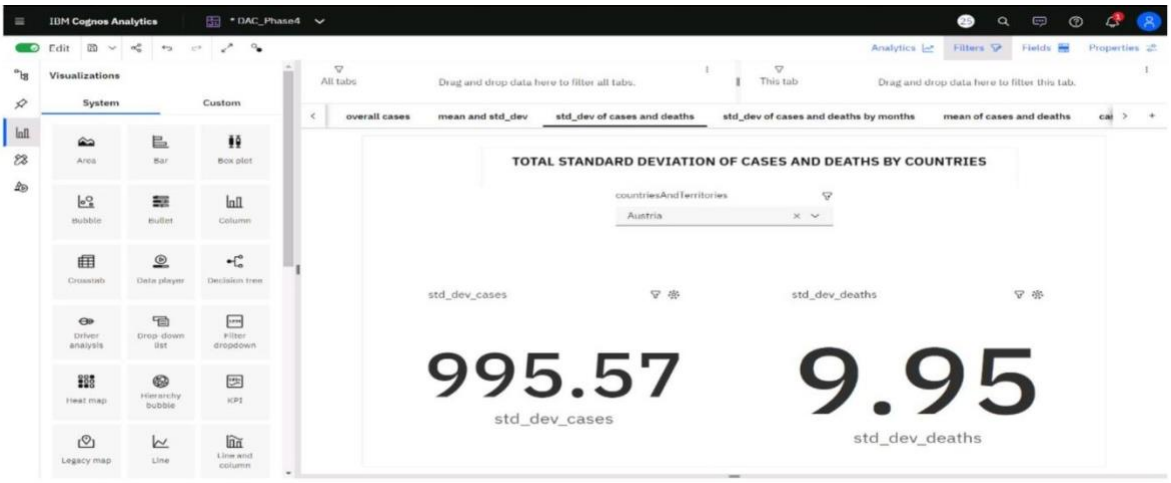
For Deaths:

As for deaths, there is also a strong weekly trend, with the highest values tending to occur on Wednesdays and the lowest on Mondays. However, there is a weak downward trend in the number of deaths. On 2021-04-08, an unusually high value was reported. The lowest average deaths were recorded on 2021-05-31 at 11.87 and 2021-05-30 at 18.07, while the highest average deaths were observed on 2021-04-09 at 111.83 and 2021-04-08 at 109.77. Current forecasts suggest that deaths may reach 13.27 by 2021-06-19, and the dataset contains over 2500 results for deaths.

For Standard Deviation:

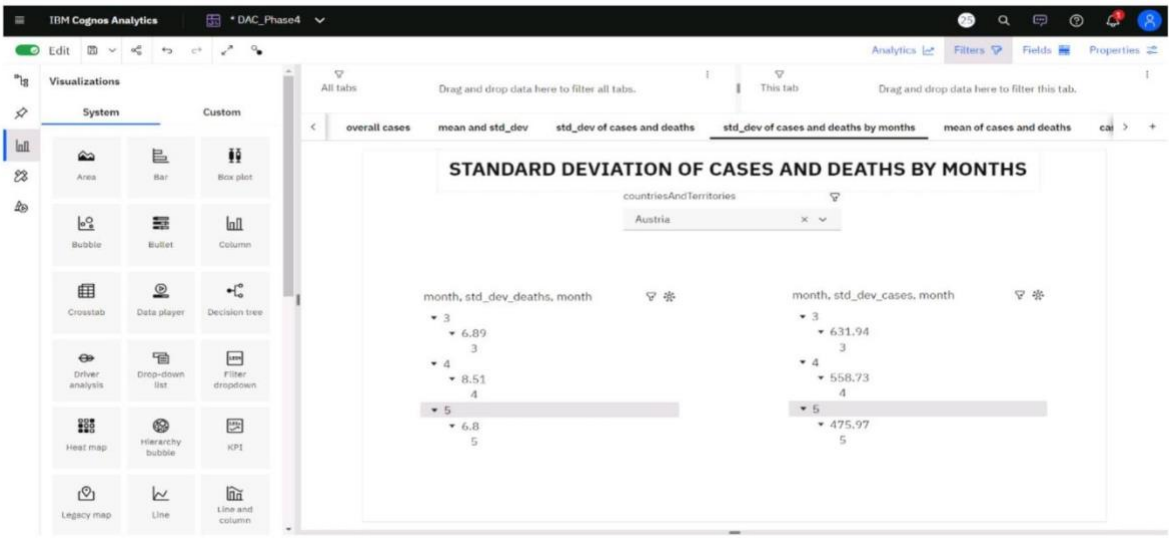
In the context of standard deviation, the number of cases has a deviation of 6.49k from the mean value, which is 3.66k. This indicates a relatively high variability in the number of cases. Similarly, the number of deaths shows a deviation of 113.96 from the mean of 65.29, suggesting a significant spread in the data. These values suggest that there is considerable variation in both cases and deaths, which may have implications for analyzing and managing the situation they represent.

Visualisation for Standard Deviation of cases and deaths by months:



In Australia, the standard deviation for COVID-19 cases is 995.97, indicating a relativelyhigh degree of variability in case numbers. In contrast, the standard deviation for deaths inAustralia is much lower at 9.95, suggesting less variability in mortality figures. Similarly, inFrance, there were 13.1k cases with a standard deviation of 995.97 and 122.02 deaths with a\standard deviation of 9.95, reflecting differing patterns in the spread of the virus and itsimpact on these two countries. Utilizing IBM Cognos, we can create a comprehensive\visualization to explore and compare these statistics across various countries and territories,offering valuable insights into the COVID-19 situation worldwide.

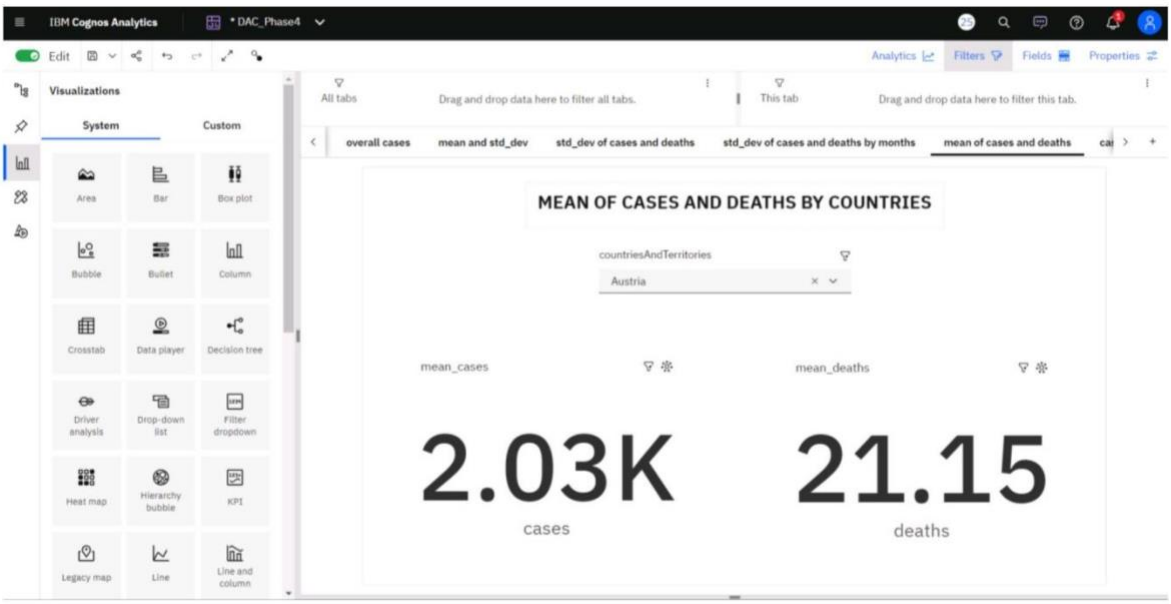
Visualisation for standard Deviation of cases and deaths by months



In this visualization, we can leverage IBM Cognos to conduct a detailed analysis of the standarddeviation for both COVID-19 cases and deaths on a monthly basis for every country and territory. By breaking down the data into monthly increments, we gain a more granular understanding of the

fluctuations in case numbers and mortality rates over time. This approach allows for the identificationof trends, anomalies, and patterns that might not be apparent when examining aggregate data. Such insights can be invaluable for policymakers, healthcare professionals, and researchers in tailoring strategies and responses to the evolving dynamics of the pandemic worldwide.

Visualisation for Mean of cases and deaths by countries:



In the IBM Cognos visualization, a comprehensive analysis of COVID-19 cases and deaths
reveals intriguing insights. Cases exhibit a strong weekly trend, with the highest values
typically observed on Tuesdays and the lowest on Mondays. Additionally, there's a weak
downward trend in cases. Notably, the dateRep for May 25, 2021, records the lowest average
cases at just over 2,000, while April 7, 2021, has the highest average cases at almost 54,000.

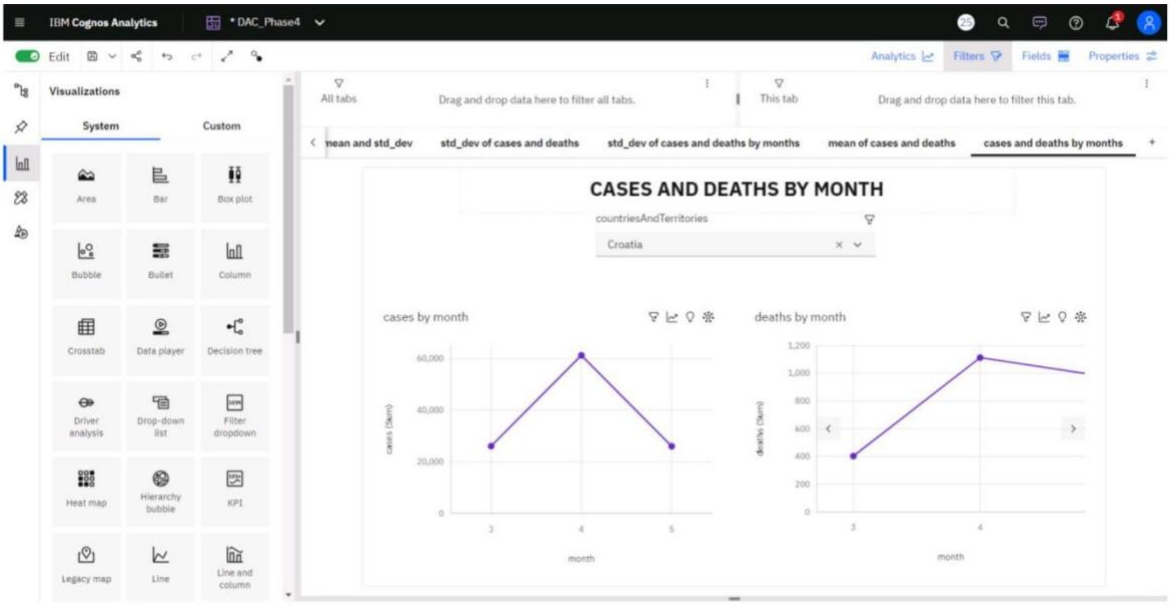
For deaths, a moderate weekly trend is observed, with the largest values occurring on

Wednesdays and the smallest on Mondays. Similar to cases, deaths also exhibit a weak
downward trend. Remarkably, March 27, 2021, stands out with the highest average deaths at
897, and April 20, 2021, follows with an average of 447.

The forecasting data suggests that both cases and deaths may increase by June 19, 2021, with
cases reaching over 18,000 and deaths potentially rising to 68.47. Several unusual data pointsare noted, particularly on March 27, 2021, for deaths and May 25, 2021, for cases. Moreover,during the period from March 27 to March 28, deaths dropped by 79%.

Overall, the IBM Cognos visualization provides comprehensive insights based on 91 data
points for both cases and deaths, offering a valuable tool for understanding the dynamics of
the COVID-19 pandemic.

Visualisation for cases and deaths by months:



In the IBM Cognos analysis, intriguing patterns emerge when comparing monthly data for
COVID-19 cases and deaths. Month 4 holds the highest total deaths but is ranked second in
total cases, indicating a discrepancy between case severity and mortality. Similarly, month 3
boasts the highest total cases but is ranked second in total deaths, highlighting variations in
the impact of the virus.
Across all months, the cumulative sum of cases exceeds 288,000, underscoring the
significance of the pandemic's reach. Cases fluctuate, with the lowest occurring in month 5 at
over 69,000, and the highest in month 3 at over 114,000. Months 3 and 4 stand out
significantly in terms of cases, contributing to nearly 76% of the total.
For deaths, the collective sum across all months surpasses 2,500. Deaths range from 708 i
month 5 to over a thousand in month 4, revealing variations in mortality trends. This
comprehensive analysis provides valuable insights into the dynamics of COVID-19 cases and
deaths across different month

CONCLUSION

In conclusion, The analysis of COVID-19 cases and deaths data for months highlights the dynamic nature of the pandemic. March and April saw surges in cases and deaths, with clear weekly trends. Countries with effective containment measures and healthcare systems reported lower case and death counts. The data emphasizes the ongoing challenges in managing the virus and the importance of public health measures, vaccination campaigns, and data-driven decision-making. These insights can guide evidence-based responses and preparedness for future waves of the pandemic.