**Project 7: COVID-19 using Cognos**

**Phase 5: Project Documentation & Submission**

**By Badri N**

**Project Definition:** In this part you will document your project and prepare it for submission.

Document the COVID-19 cases analysis project and prepare it for submission.

**Design Thinking:**

* 1. **Analysis Objectives**: Define the specific objectives of analysing 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.
  5. **Analysis Objectives:**
* To identify and understand the different trends in COVID-19 cases and deaths.
* Calculate the Mean and Standard Deviation of Daily, Monthly and Yearly cases and deaths per Country/Territory.
* Compare the Mean and Standard Deviation values of EU region to calculate the mortality rate.
* To visualize the data in a graphical representation.

**Introduction**:

This project is designed to analyze the impact of COVID-19 on various countries in the European region. The main aim is to explore how different nations are responding to the outbreak. This project will help us to list out the cases and deaths rate on these various countries across EU region spread across different days in the year 2021.

**Project Objectives**:

The aim of this project is to enable how to utilize our Data Visualization and Data Analytical skills.

* To identify emerging trends in the spread of COVID-19.
* To assess the impact of cases and death rates on the spread of COVID-19.
* To assess the impact of cases and death rates on the spread of COVID-19 on various days of a month in the year 2021.
* To create predictive models to better anticipate future cases and death rates related to COVID-19.
* To provide insights and analyse demographics that could lead to decision making.

**Data Collection:**

* The dataset used here is the COVID 19 dataset.
* It contains information about the cases and deaths per country affected by COVID-19.
* The dataset contains 7 fields: DateReported, day, month, year, cases, deaths, countries and 2730 records but we choose a sample data of 31 records only.

**Visualization Strategy:**

* To first calculate the Mean and Standard Deviation for both Cases and Death individually, we modify the dataset by adding calculation to existing table using “Calculations” option and calculate Standard Deviation, Average for each attribute deaths and cases individually.
* Then, we create a new Report and select COVID-19 data module as Source
* Now to create the graph, we select the “Visualization” option and select Clustered Column graph and substitute Average cum Standard Deviation values of deaths and cases on X and Y comparatively to produce 2 graphs.

**Insights Generation:**

|  |  |  |
| --- | --- | --- |
|  | Cases | Deaths |
| Mean | 896.7742 | 13 |
| Standard Deviation | 475.9697 | 6.801961 |

* Description/Comparison of Means:

The mean values of the cases are far higher than that of the deaths. The mean value of Cases is 896.7742 and the Deaths having 13. This explains that people were infected in large numbers, but only smallest number of people lost their lives.

* Description/Comparison of Standard Deviation:

The stand deviation values of the Cases are also far higher than that of Deaths. The Cases have an S.D. value of 475.9697 and the Deaths have an S.D. value of 6.801961.

This shows that the average number of cases are more spread out than the death’s average data

**Analysis Objectives:**

* Load and preprocess the dataset.
* Clean the dataset and ensure its accuracy and reliability.
* Export the dataset and import to Cognos for Visualization.

**Data Preprocessing:**

* Data preprocessing is a crucial step within the statistics analysis and gadget gaining knowledge of pipeline.
* It includes a sequence of strategies and operations finished on uncooked statistics to clean, organize, and transform it right into a layout that is suitable for analysis or device mastering version schooling.
* Data preprocessing goals to enhance the first-class of the records, making it greater reliable and conducive to generating accurate consequences.
* Data preprocessing is an iterative process that may involve several of these steps in various orders, depending on the specific dataset and the analysis goals. Proper data preprocessing is essential for improving the accuracy and effectiveness of machine learning models, as well as for making data more accessible for traditional statistical analysis.

**Data Cleaning:**

* Handling missing values: Deciding how to deal with missing data, whether by imputing values or removing incomplete records.
* Outlier detection and treatment: Identifying and handling data points that significantly deviate from the norm.

**Data Transformation:**

* **Data normalization:** Scaling numerical features to a standard range (e.g., between 0 and 1) to ensure that they have similar influence in the analysis.
* **Encoding categorical variables:** Converting categorical data into numerical format, such as one-hot encoding or label encoding.
* **Feature engineering:** Creating new features or modifying existing ones to capture more meaningful information from the data.
* **Dimensionality reduction:** Reducing the number of features while retaining essential information, using methods like Principal Component Analysis (PCA).

**Data Integration:**

* **Merging or joining datasets:** Combining data from multiple sources into a single dataset for analysis.

**Aggregation:** Summarizing data at a higher level of granularity, such as aggregating daily sales into monthly totals.

**Data Reduction:**

* **Sampling:** Reducing the size of a large dataset by randomly selecting a representative subset.
* **Binning:** Grouping continuous data into discrete bins to simplify analysis.
* **Filtering:** Selecting a subset of data based on specific criteria.

**Data Standardization:**

* Ensuring that data follows a consistent format and structure.
* Date and time format conversion: Converting date and time data into a uniform format.
* Currency conversion: Converting monetary values into a common currency.

**Data Scaling:**

* Scaling numerical data to a common range to prevent some features from dominating the analysis.

**Importing the libraries:**

Import the below two libraries for data preprocessing.

* **Matplotlib:** this library helps in plotting graphs and charts, which are very useful while showing the result of your model
* **Pandas:** pandas allow us to import our dataset and also creates a matrix of features containing the dependent and independent variable.
* **NumPy:** it is a library that allows us to work with arrays and as most machine learning models work on arrays NumPy makes it easier

Using the above three libraries, we shall now preprocess the COVID-19 Cases CSV data file.

**Steps:**

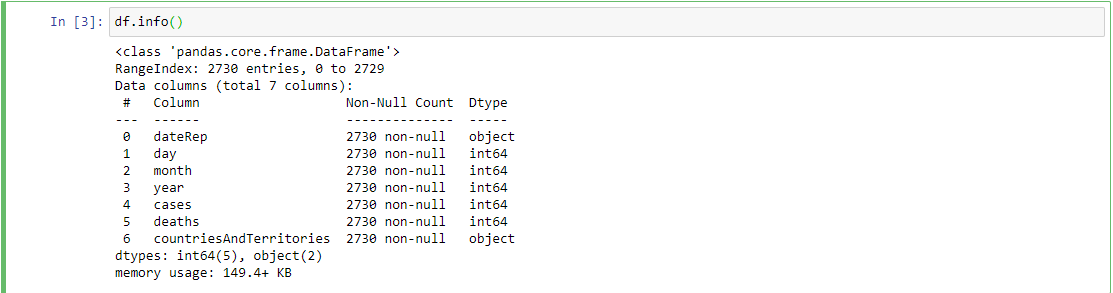
1. Import matplotlib, numpy and pandas package into jupyter notebook.

****

1. Load the CSV file into a dataframe.

****

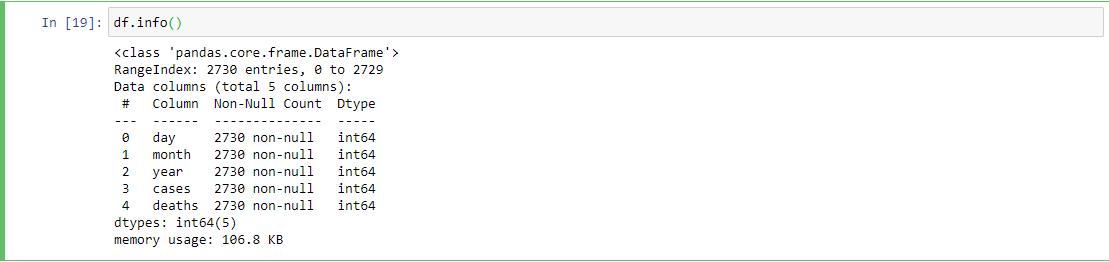
1. Use info() function to retrieve information about the dataframe.

****

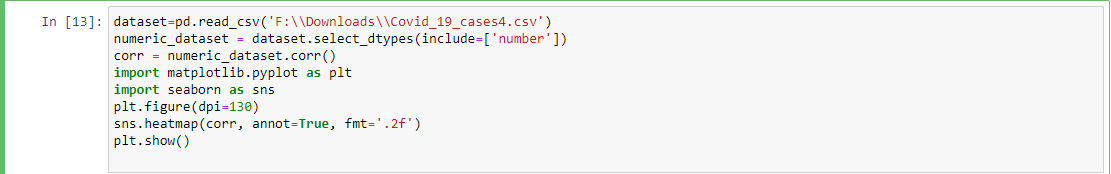
1. Since we only need cases and deaths column, drop the columns that have non-integer values(i.e. dateRep and countriesandTerritories)

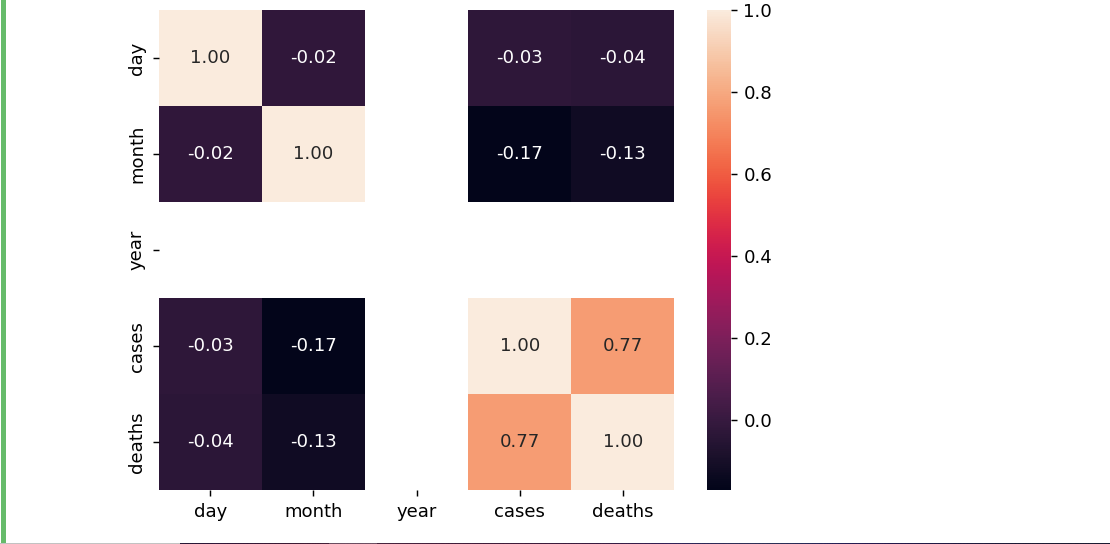
****

1. Retrieve information about the table using info() again.

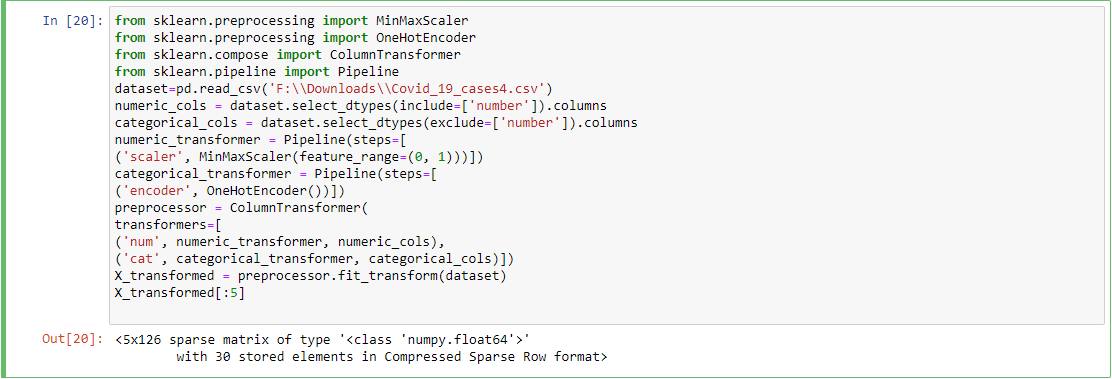
****

1. Now since we have the required columns, let’s find the correlation for the current dataframe by writing a program code for visualization.

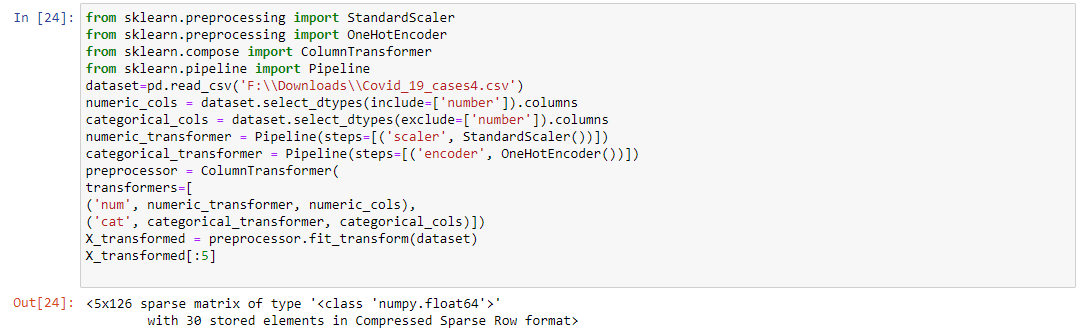
****

****

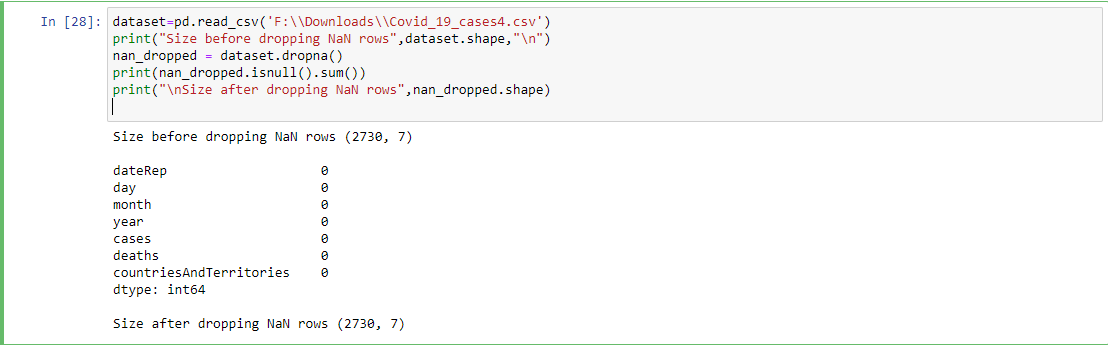
1. Let’s rescale the dataset by performing Normalization.

****

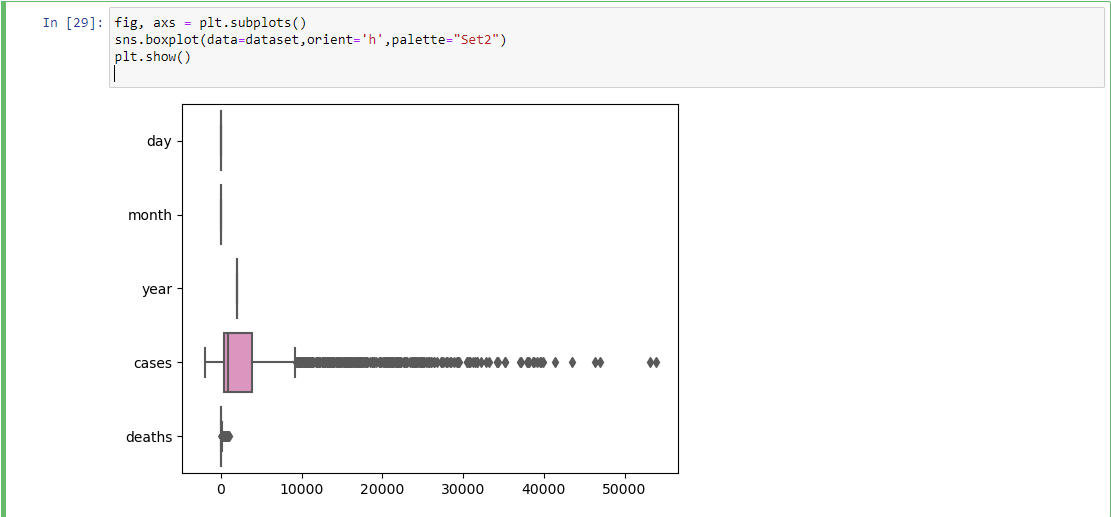
1. Perform Standardization operation.

****

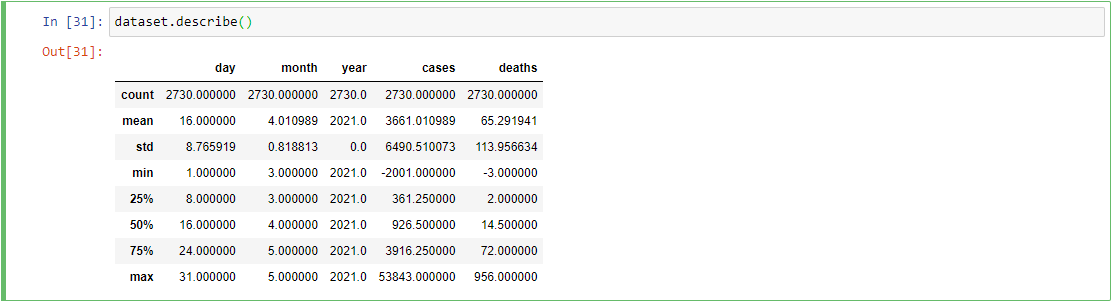
1. Dropping null columns.

****

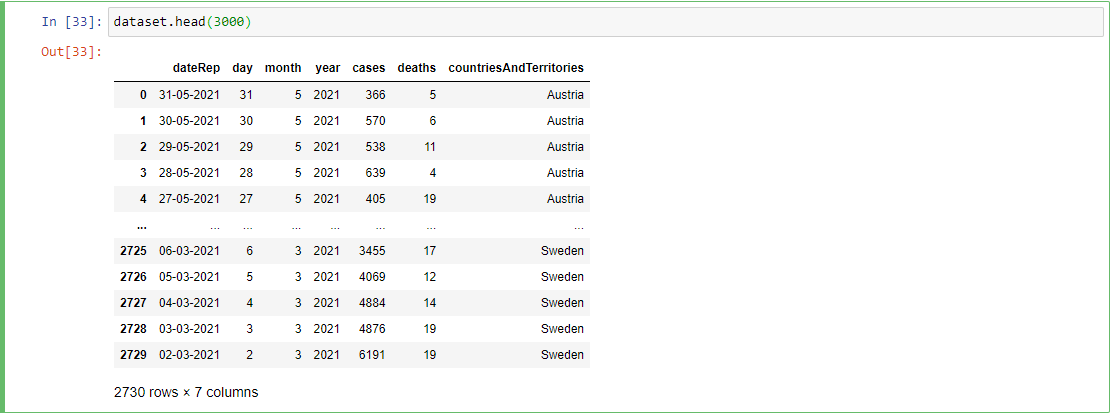
1. Use subplot function for data visualization

****

1. Use of describe() function.

****

1. Use of head() function.

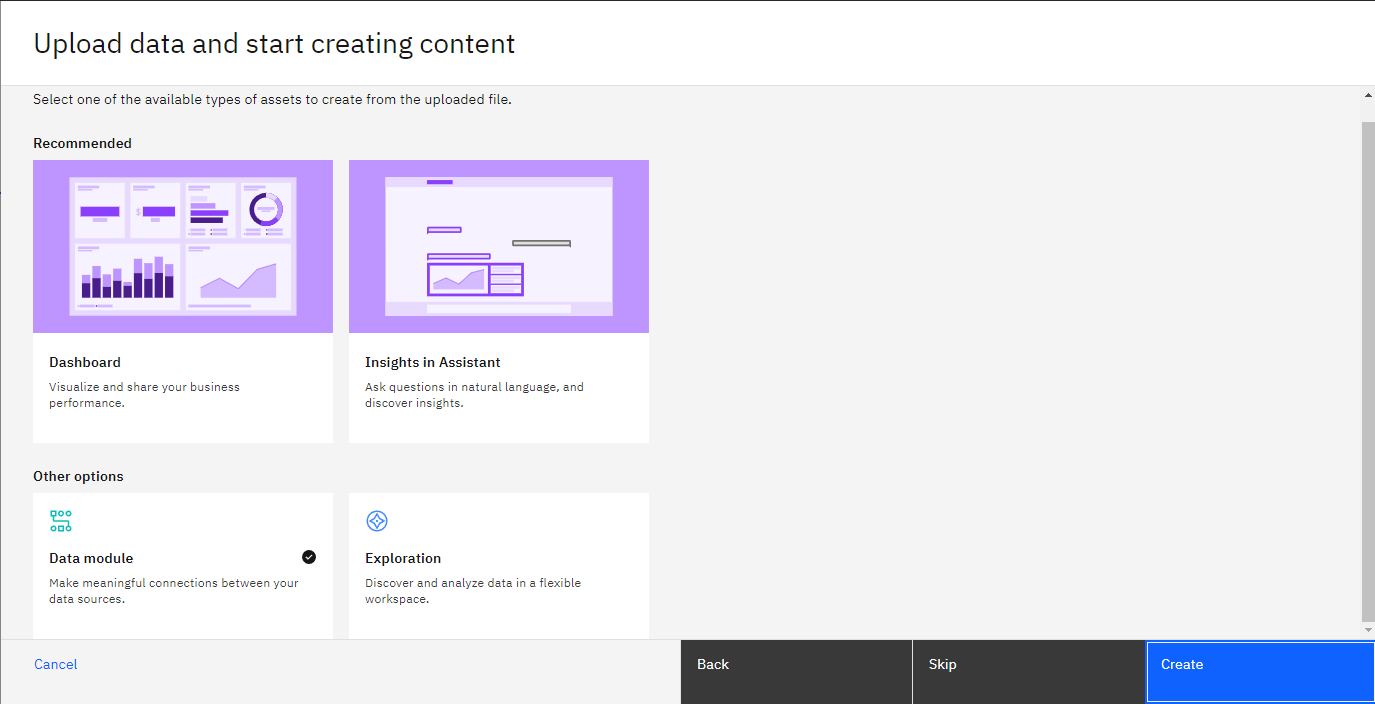
****

1. The data has been completely preprocessed. Now export the dataset to an csv file using to\_csv() function.

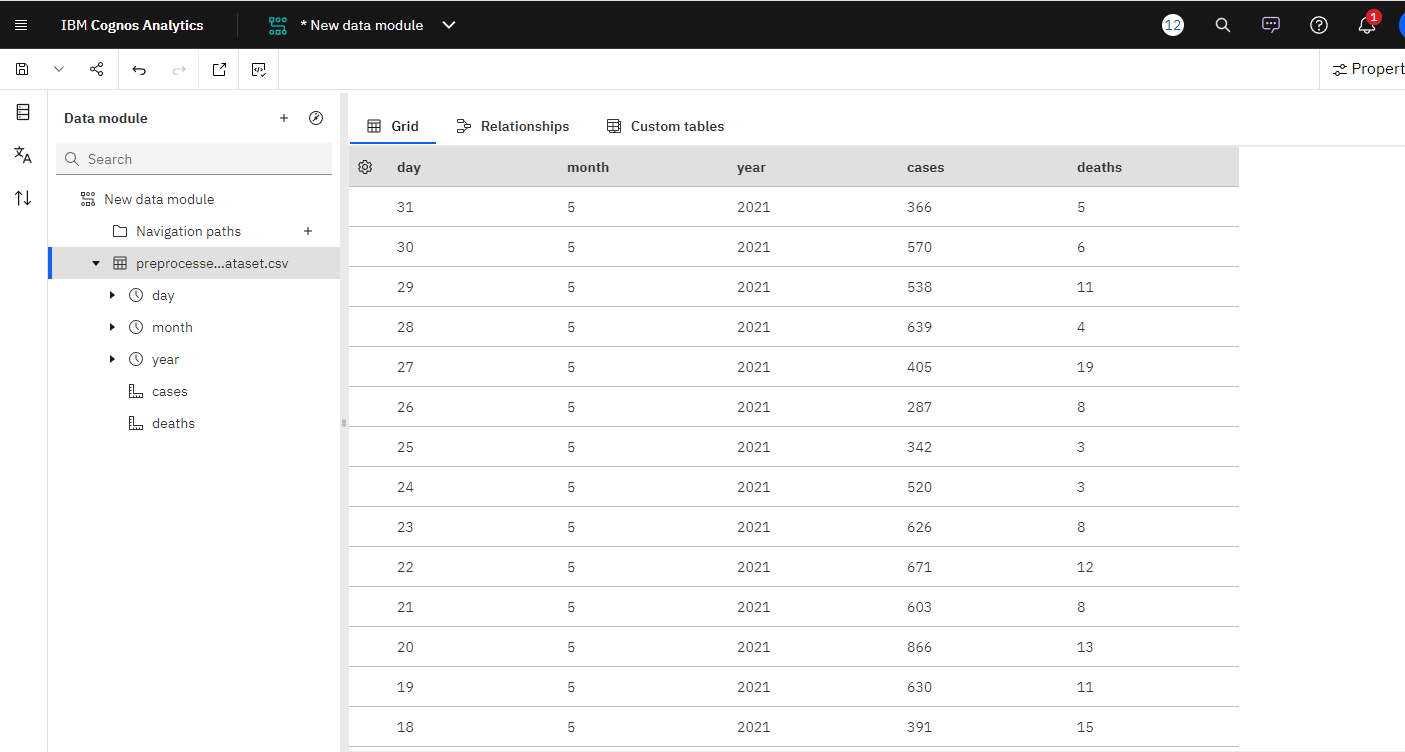
****

****

1. Download the CSV file and navigate to Cognos Analytics website and upload the file to “My Content” and create as “Data Module”

****

1. The Data Module is successfully created and can be imported into “Reports”, “Dashboard” or a “Story” for visualization.

****

**Applying Innovation:**

In this stage, from the previous Development Part 1, we are going to import the operated dataset into IBM Cognos and create a dashboard analyzing the visualizations of deaths, cases and find trends, correlation between them.

**Ideology:**

We are going to create a dashboard in IBM Cognos Analytics. Then, import the refined COVID data module. Import datasets onto the dashboard and visualizations to view cases and deaths within each country individually and find correlations, trends.

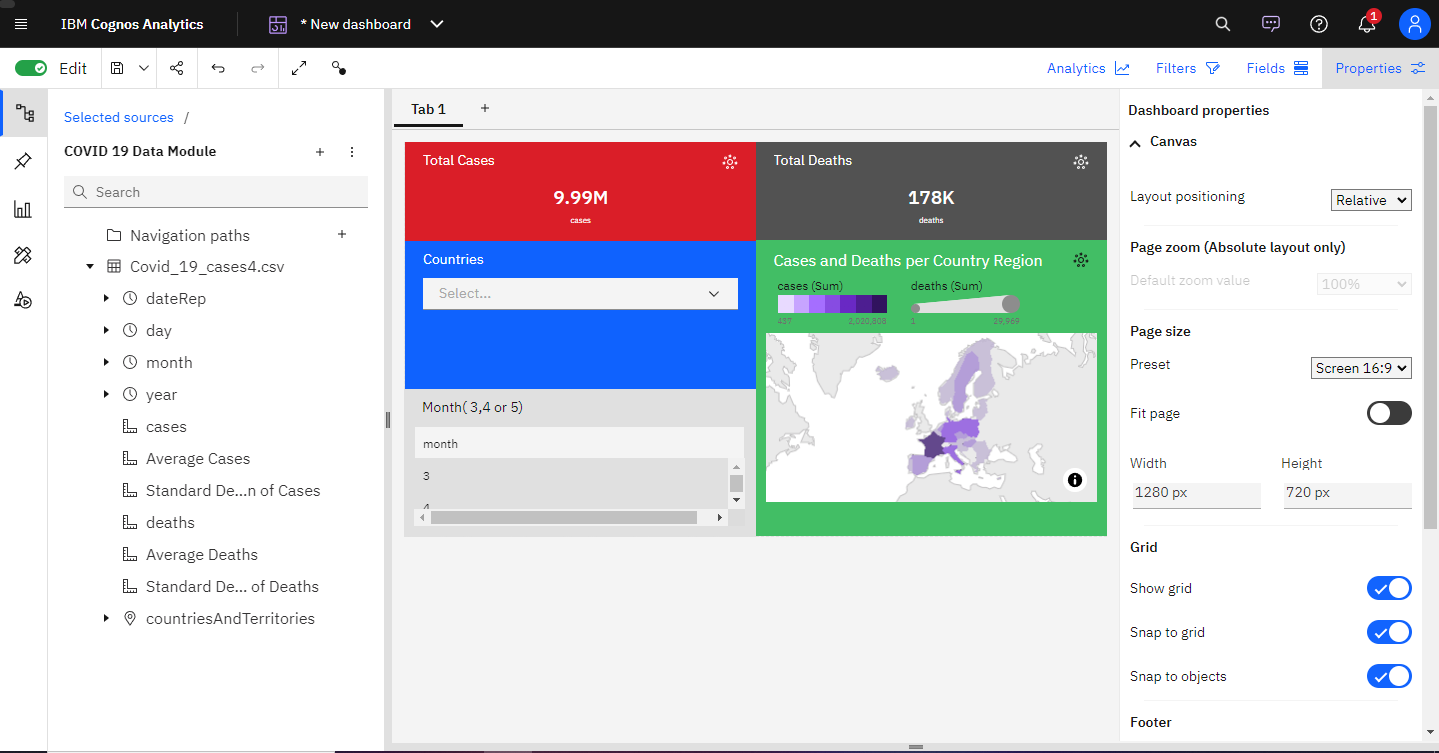
**Steps:**

* To begin, create a new dashboard(Open menu > New > Under Present tab > Dashboard).
* Choose a format(Tabbed or Infographic).
* Once the empty dashboard is visible, import “COVID-19 Data Module” from “My Content”.
* Drag and Drop “cases”, “month” and “deaths” column data into empty blocks each having its individual block.
* Create a “Map” visualization by going into the Visualization tab. Add “countriesandTerritories”, “cases” and “deaths” as parameters.
* Add a “Drop down” from the “Visualization” tab and drag-drop “countriesandTerritories” data onto it.

Dashboard is now completed. It can be modified in future if there’s any updates in datasets from the provided module.

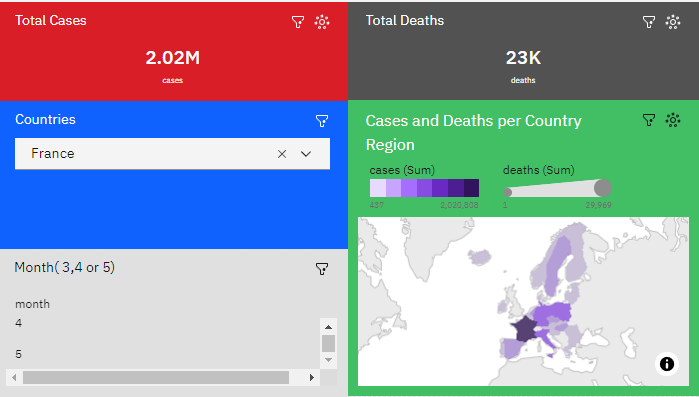
* The new values can be compared and updated to view from different aspects as per need.

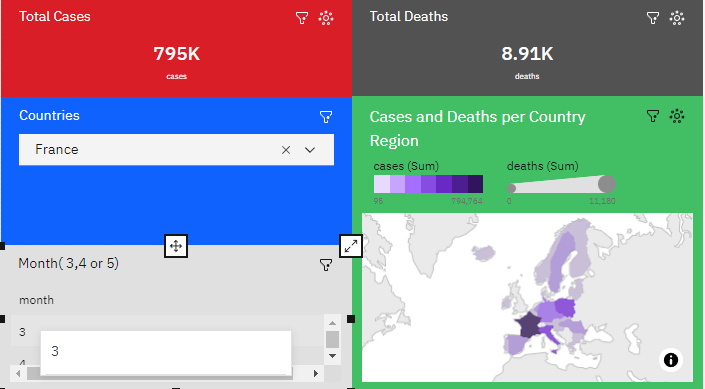
**Dashboard Visualization:**



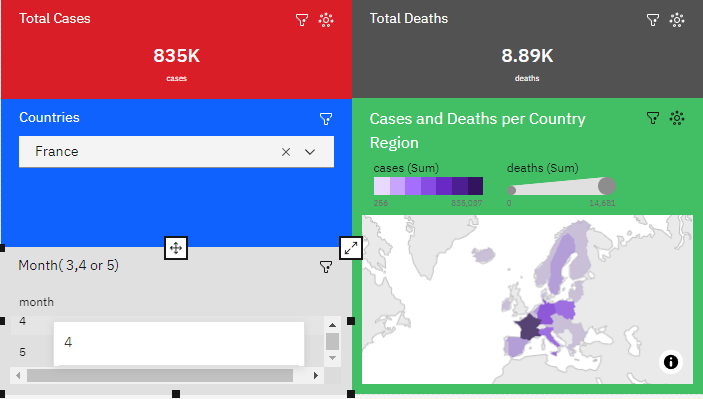
**Working with Example:**

Below listed are individual countries with their Total Cases and Total Deaths represented in the Dashboard.

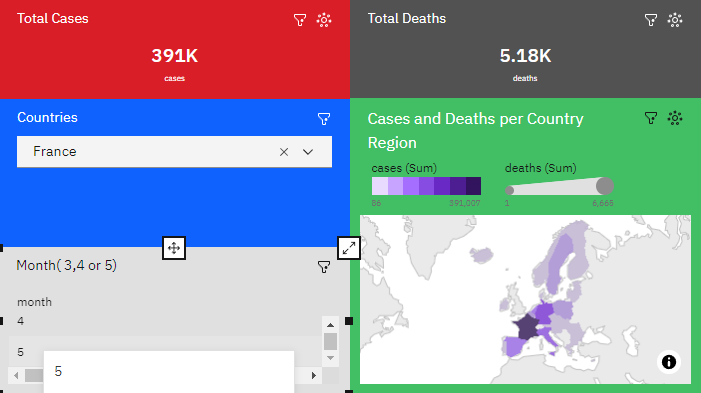
* France(No months specified)
* France(3rd Month i.e. March)



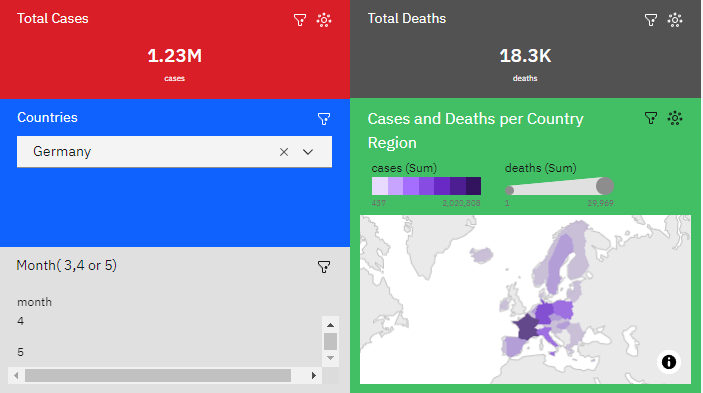
* France(4th Month i.e. April)



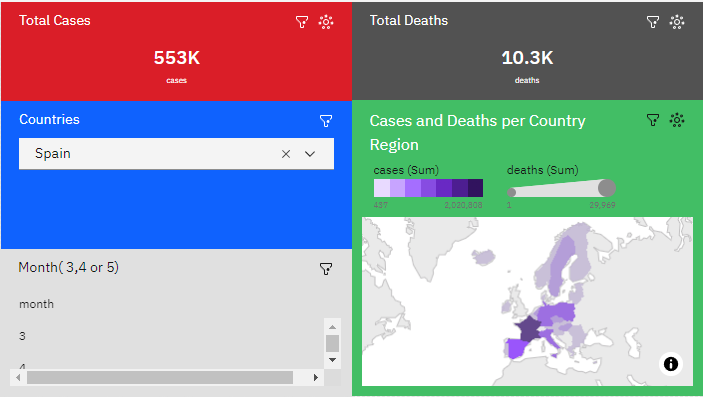
France(5th Month i.e. May)



* Germany(No months specified)



* Spain(No month specified)



**Conclusion:**

From the previous development, we have imported the dataset to IBM Cognos and successfully created a Data Module. Next, we created a dashboard and imported the COVID Data Module and created visualizations. Through the IBM Cognos Analytics tool, we are able to drill down through the COVID-19 Data Module to view the dataset from various individual factors and aspects such as viewing individual country’s total cases and deaths per month. We are also able to find correlation and trends between cases and deaths. The Cognos Analytical Tool has brought innovation to our previous design.

From the above insights, we can see that Germany has the highest number of cases and as well as deaths followed by Spain and then France. These insights help in providing in-depth analysis about the widespread of COVID-19 virus over the European region.

The insights helps in analysis of cases and deaths which in turns gives us information about which countries have been hit with the most casualties. Using the data, we can then determine which country needs to speed up their vaccination procedure and tighten the steps/measurements to avoid further widespread of the virus which in turn leads to the graphical increase in cases and deaths.

This whole project is based on analyzing COVID-19 Cases over the European region. The data is helpful for the Data Analysts to use the data in order to gain information and insights from it which can help in largely aiding the affected countries. Industries such as Medicine, Vaccine and Hospital can use the data to formulate strategic plans for producing health immunity medicines, doses of vaccines and managing hospital space wards for the oncoming patients.

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