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

**Phase 3: Development Part 1**

**By Badri N**

**Project Definition:** In this part you will begin building your project by loading and preprocessing the dataset.

Start building the COVID-19 cases analysis using IBM Cognos for visualization.

Define the analysis objectives and obtain the COVID-19 cases and deaths data file.

Process and clean the data to ensure its accuracy and reliability.

**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:**

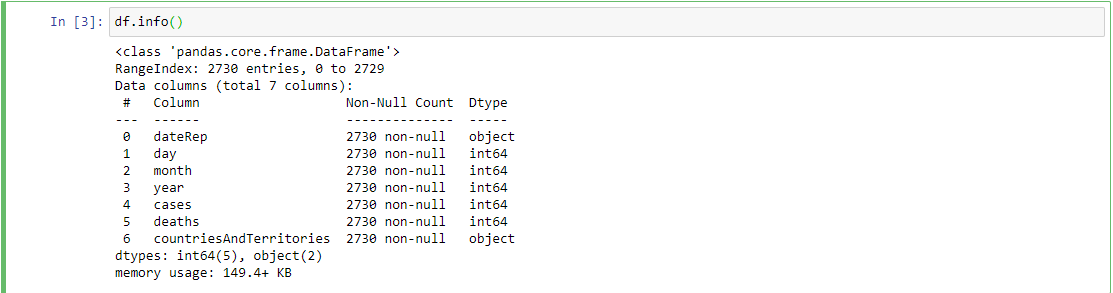
1. Import matplotlib, numpy and pandas package into jupyter notebook.

****

1. Load the CSV file into a dataframe.

****

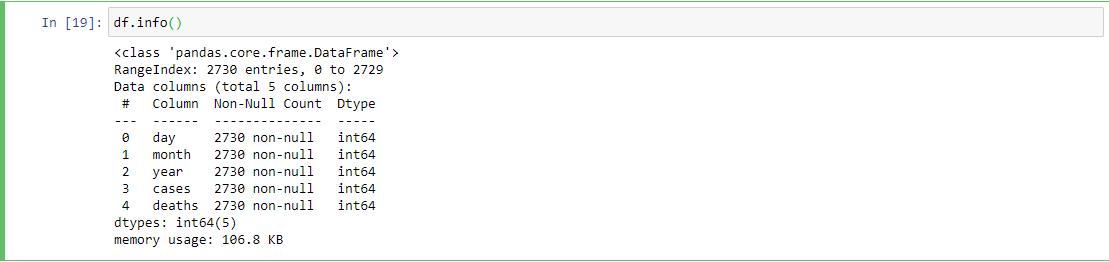
1. Use info() function to retrieve information about the dataframe.

****

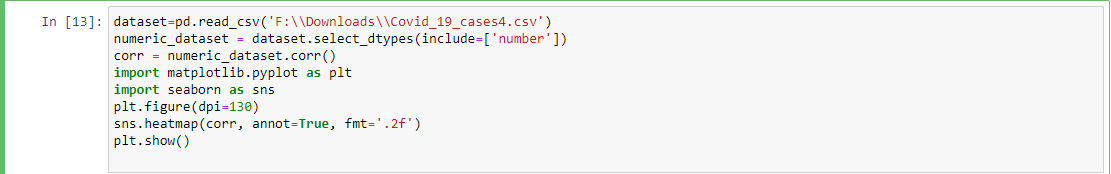
1. Since we only need cases and deaths column, drop the columns that have non-integer values(i.e. dateRep and countriesandTerritories)

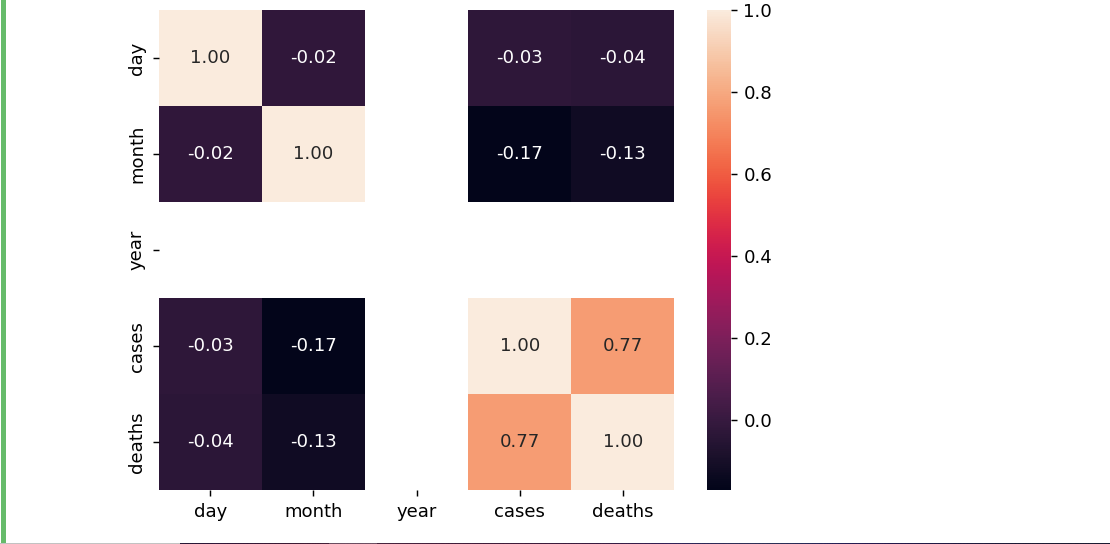
****

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

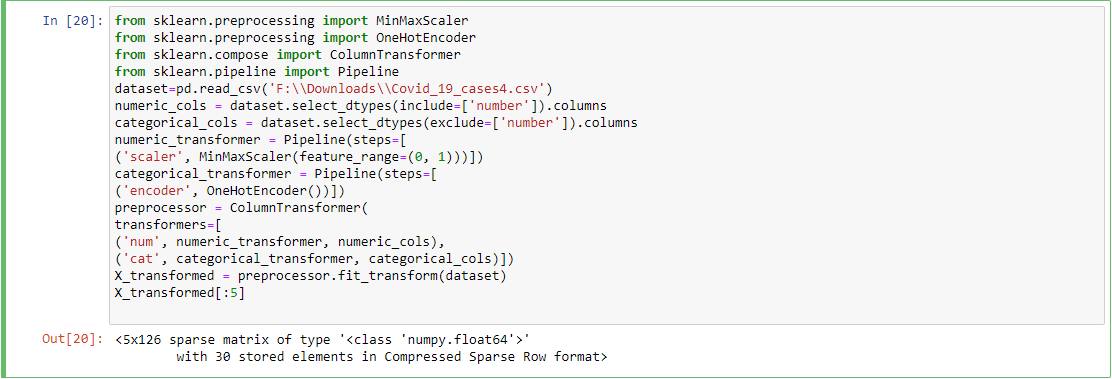
****

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

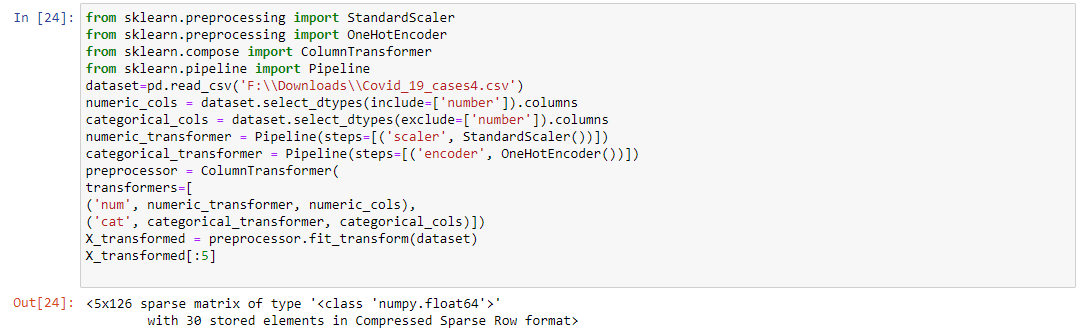
****

****

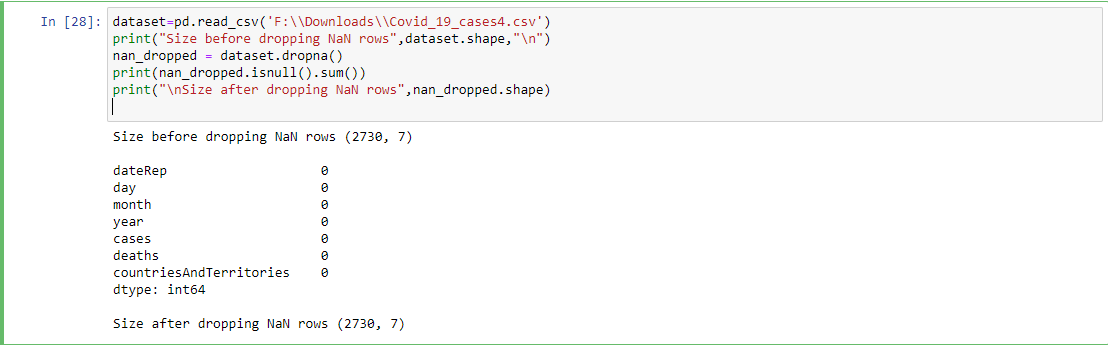
1. Let’s rescale the dataset by performing Normalization.

****

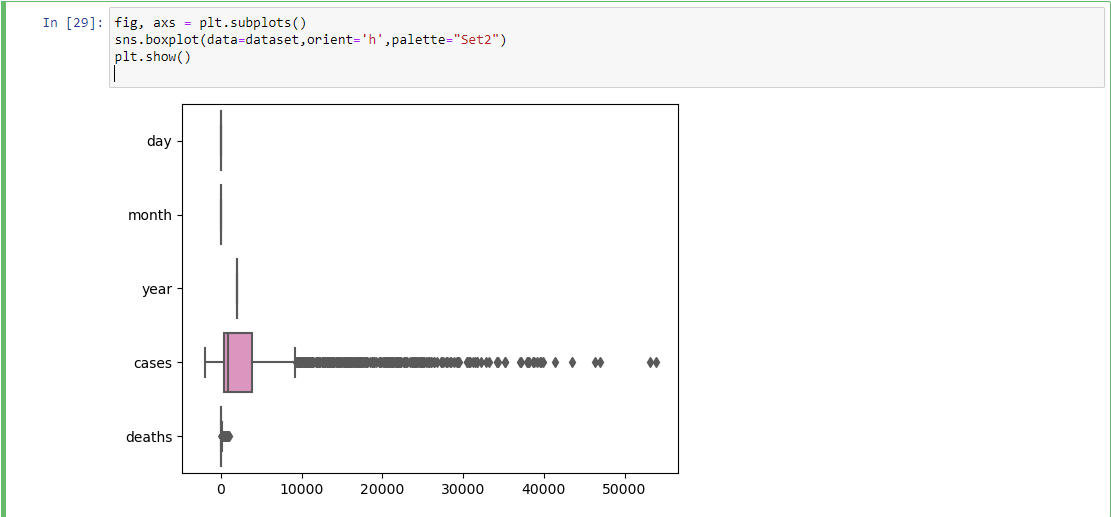
1. Perform Standardization operation.

****

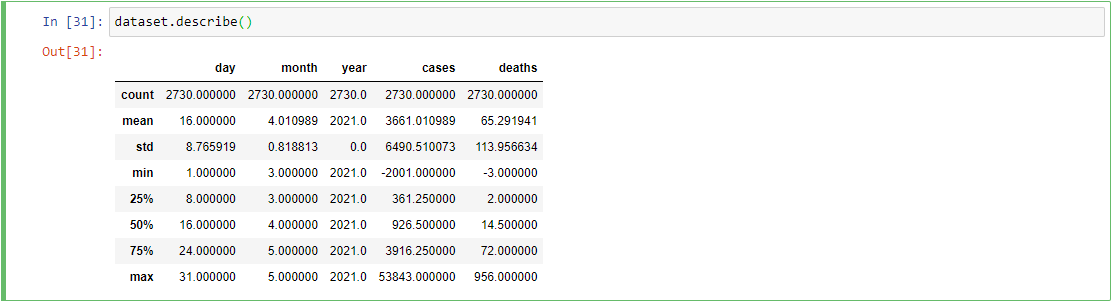
1. Dropping null columns.

****

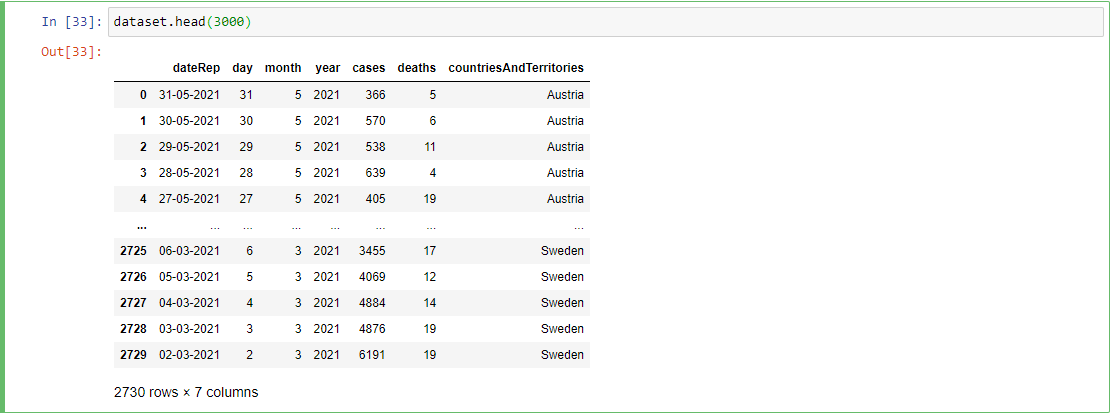
1. Use subplot function for data visualization

****

1. Use of describe() function.

****

1. Use of head() function.

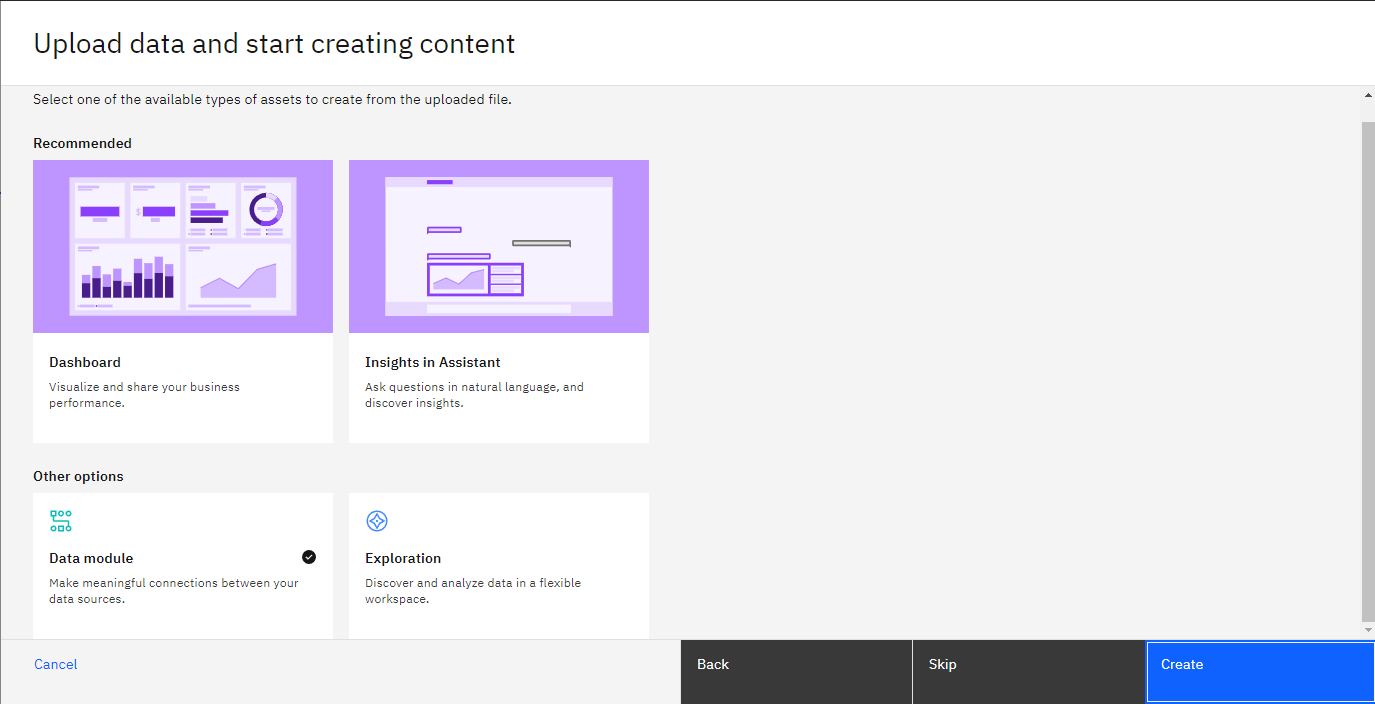
****

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

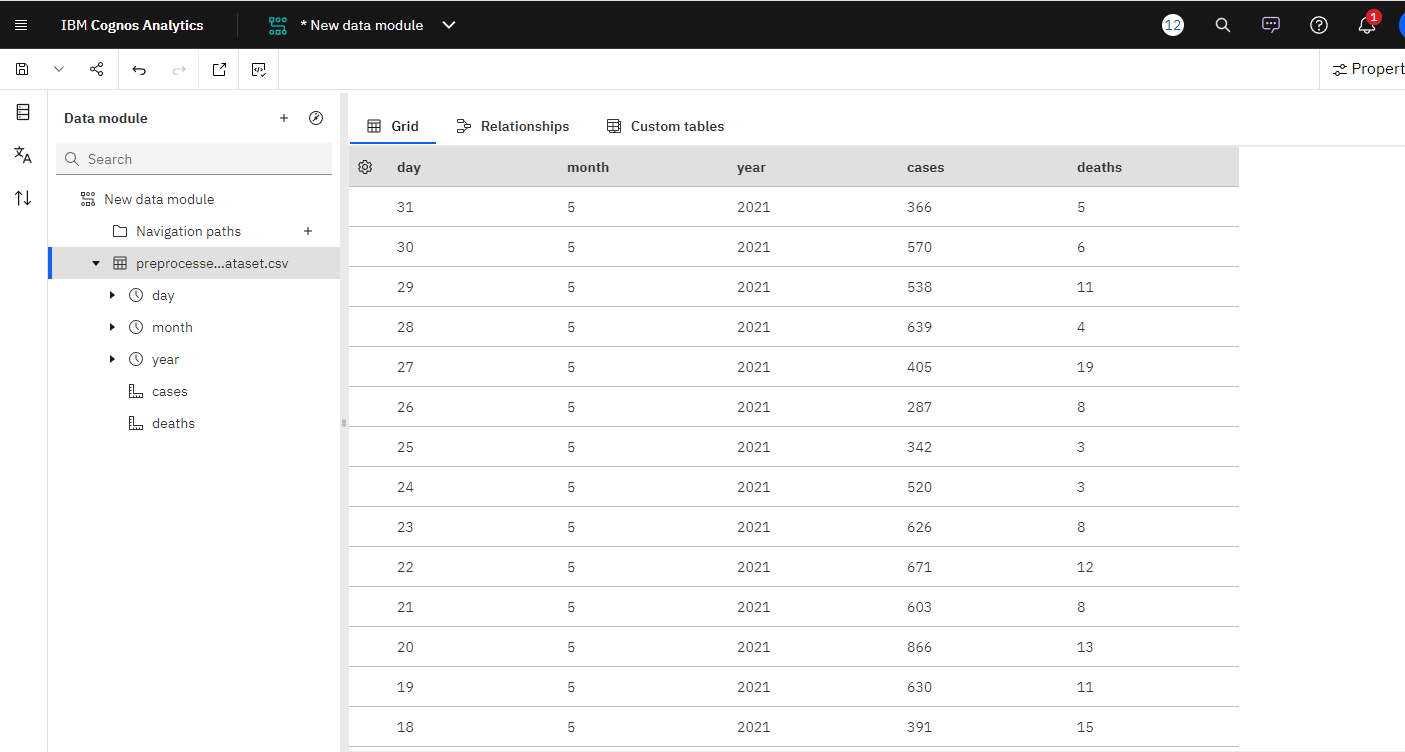
****

****

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.

****

**Conclusion:**

The COVID-19 Excel data was imported to python using Pandas library. Operations such as standardization, normalization, correlation and many others were performed on the dataset. After preprocessing the data in python, it was exported as an CSV file and uploaded to Cognos Analytics online platform.

The pre-processed data was then converted to Data Module and can now be used to create reports, dashboards or stories based on the new data.