**COVID VACCINES ANALYSIS**

**PHASE-III:** Data preprocessing and visualization

**PROJECT:** *DATA ANALYSIS ON COVID VACCINATION DATA*

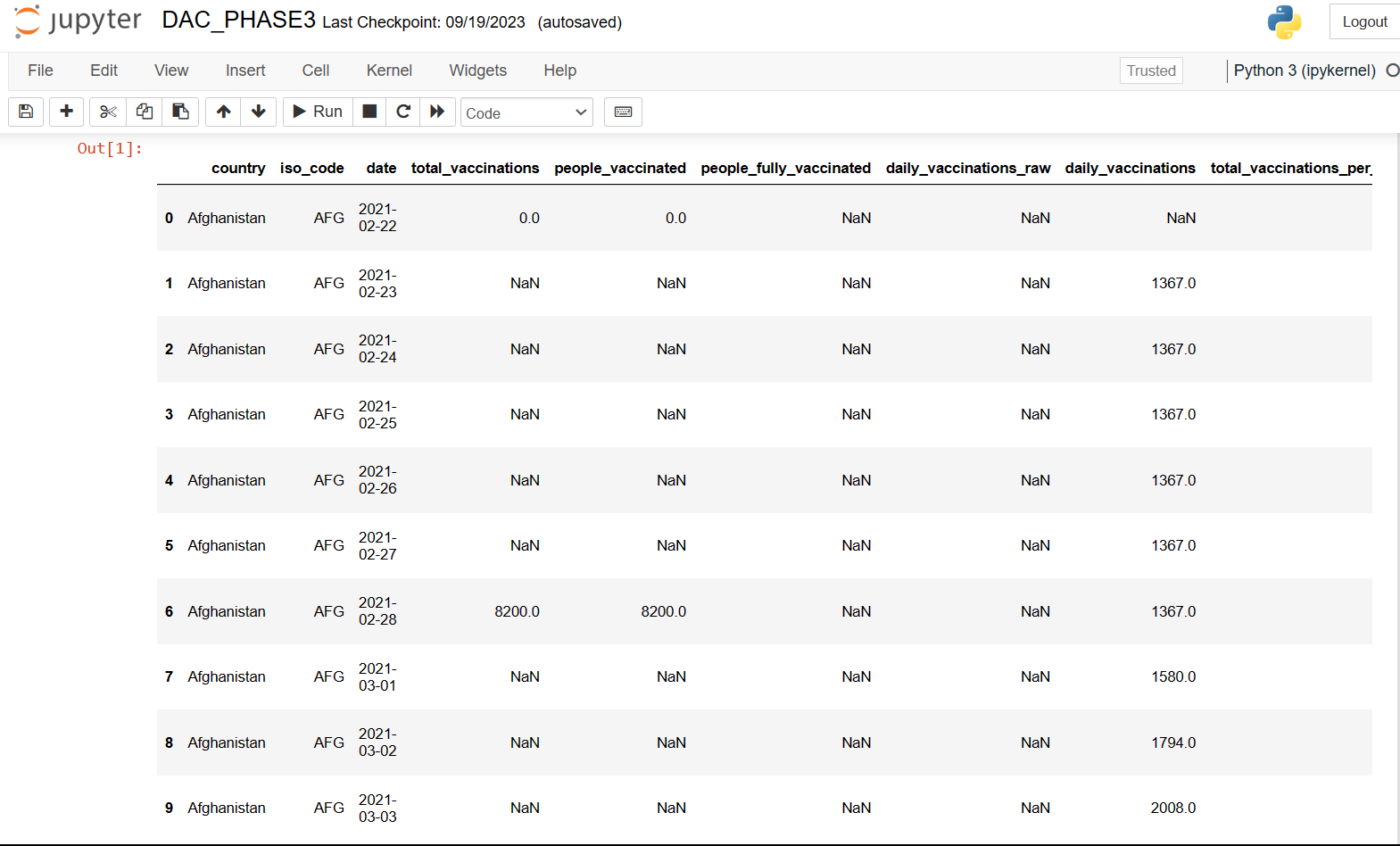


**Introduction:**

The COVID 19 pandemic caused due to the Corona virus devastated the world by causing several fatalities around the world. This virus originated in Wuhan, China in 2019 and was later spread throughout the world due to human contact in one way or the other. An effort was made to find a cure or vaccine by several health organizations to bring a stop to this pandemic.

In later stages of 2020 several experimental vaccines were developed and was administered to humans. The efforts were successful as the vaccines were helpful in reducing the affects the virus and even if people were infected, they were not in any life threating situation and escaped the illness having only minor symptoms. Many countries later developed their own vaccines and also helped other countries without the resources by providing them with vaccines developed.

**Given data set:**

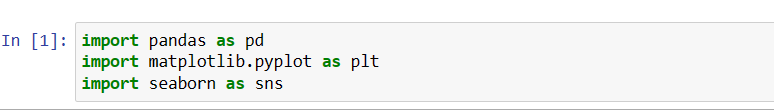
****

**Data Preprocessing:**

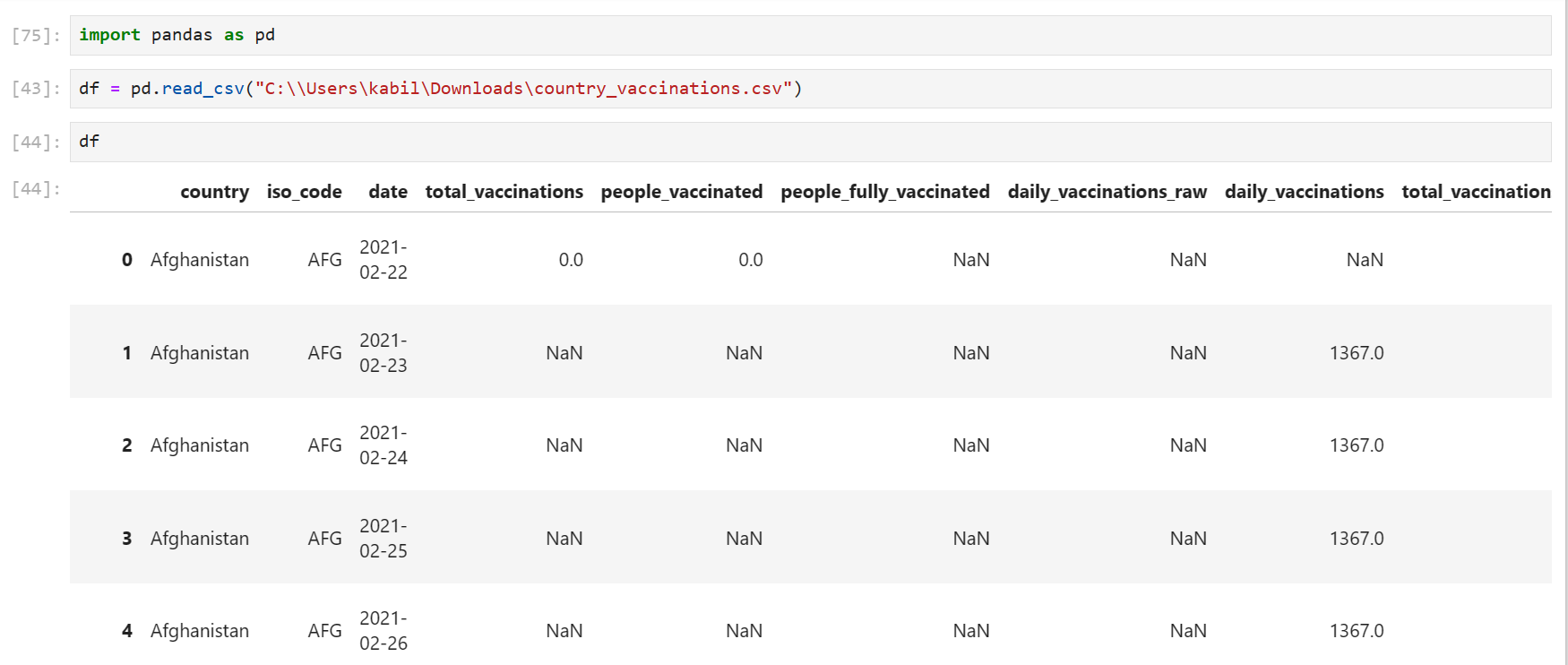
Itinvolves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

**Necessary step to follow:**

**1.Import Libraries:**

Start by importing all the necessary libraries.

**2.Loading the DataSet:**

****

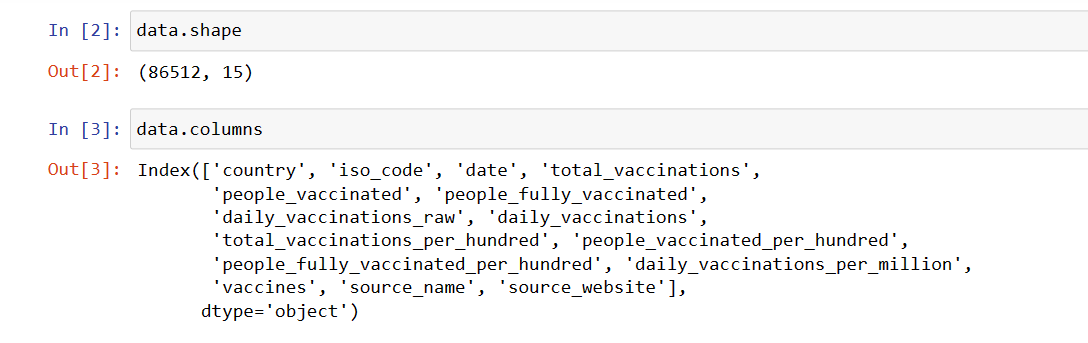
**3.Data Cleaning:**

Data **cleaning**, **also** known as data **cleansing** or data preprocessing, is a crucial step in the data science pipeline that involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability.

**Important steps :**

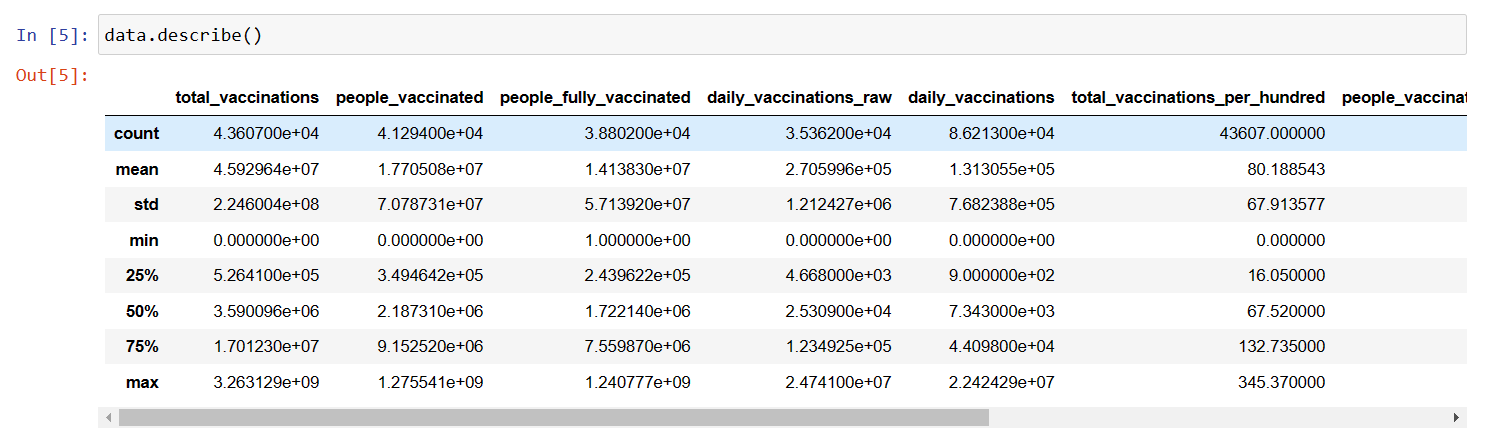
**1.Data inspection and exploration:**

This step involves understanding the data by inspecting its structure and identifying missing values, outliers, and inconsistencies.



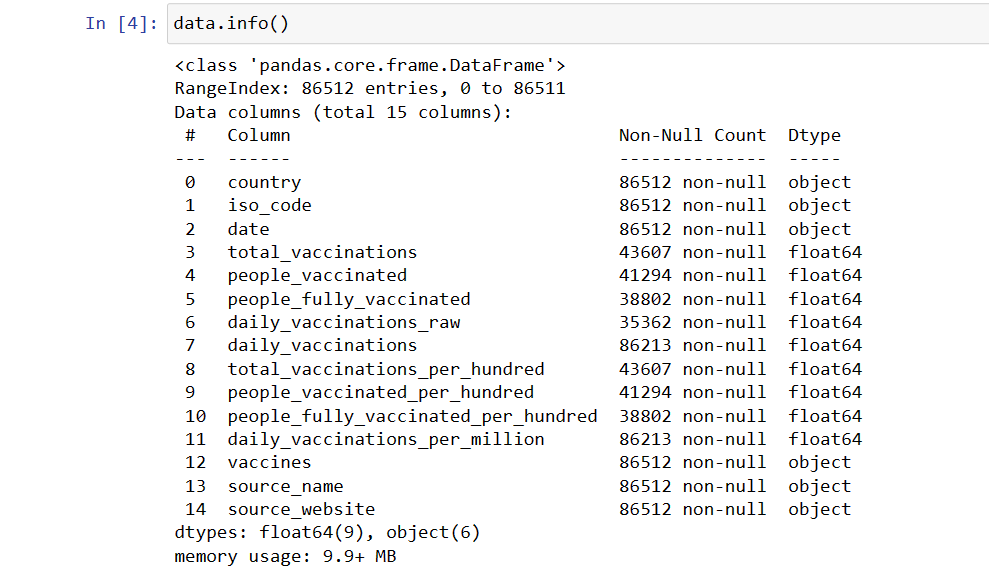
shape function returns the total no of rows and columns present in our dataset and .columns function gives us the columns names

**Let’s see the descriptive structure of the data using data.describe() and data.info()**



**Checking data Information using .info()**

From the below data info, we can see that many columns have an unequal number of counts. And some of the columns have data type objects and some are float values.

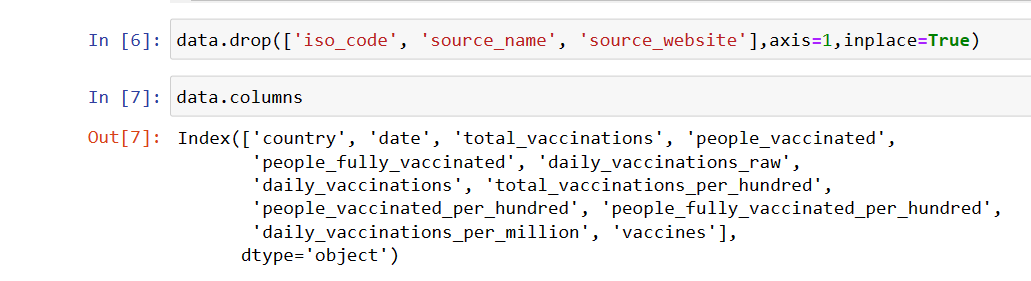
****

### **2. Removal of unwanted observation:**

### 

This includes deleting duplicate/ redundant or irrelevant values from our dataset. Duplicate observations most frequently arise during data collection and Irrelevant observations are those that don’t actually fit the specific problem

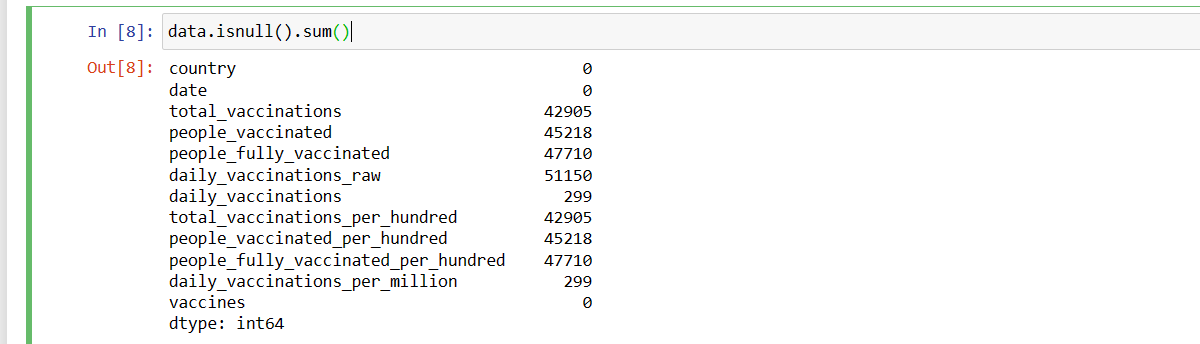
As we know our machines don’t understand the text data. So, we have to either drop or convert the categorical column values into numerical types.Here we are dropping the columns because it hasn’t a great influence on target variables

****

### **3. Handling missing data:**

Missing data is a common issue in real-world datasets, and it can occur due to various reasons such as human errors, system failures, or data collection issues. Various techniques can be used to handle missing data, such as imputation, deletion, or substitution.

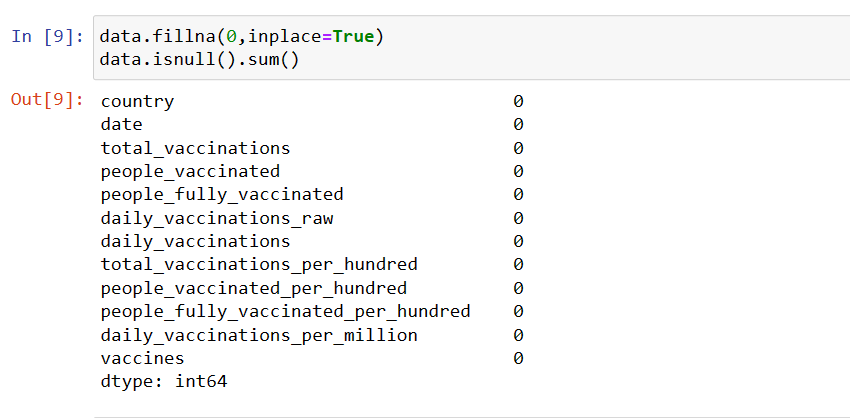
Let’s check the missing values columns-wise for each row using data.isnull() it checks whether the values are null or not and gives returns boolean values and .sum() will sum the total number of null values rows



We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important.

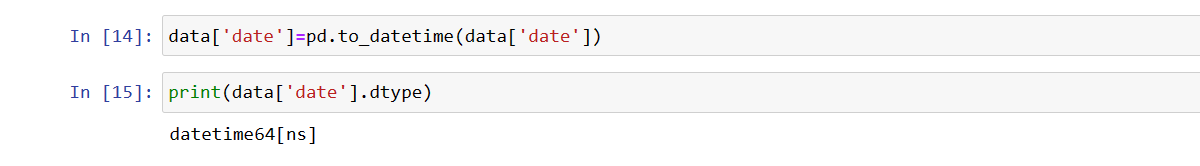
1.Dropping observations with missing values.

2.Inputing the missing values from past observation

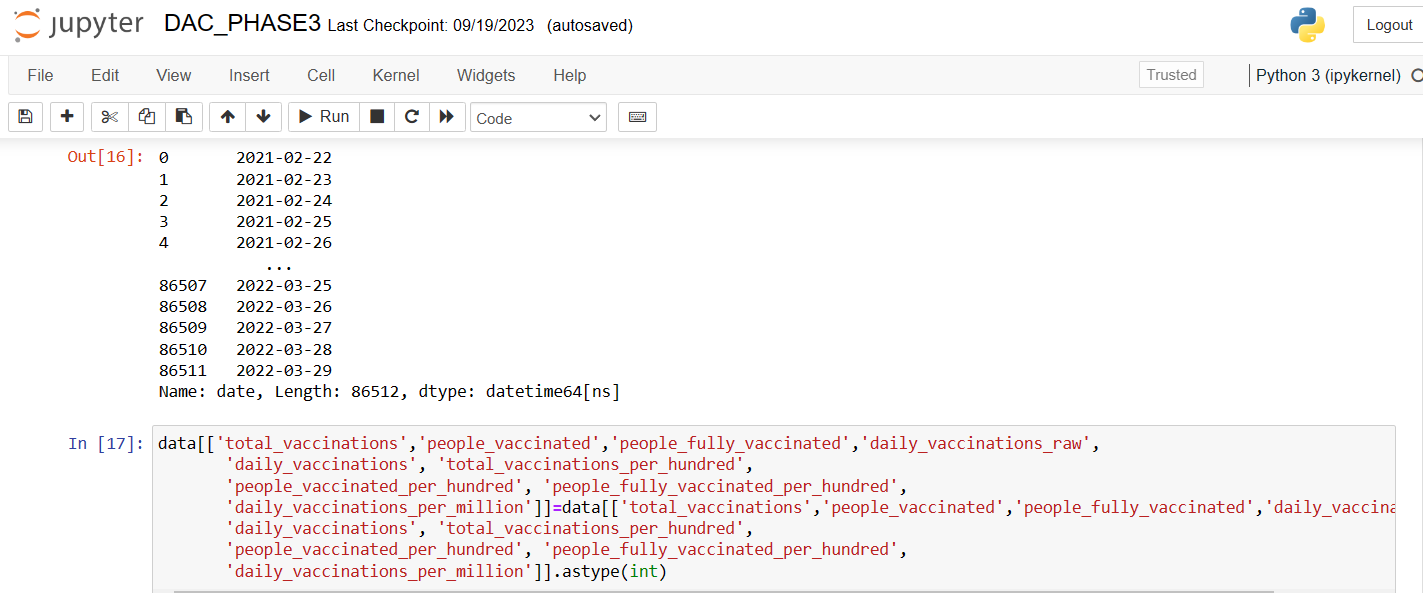


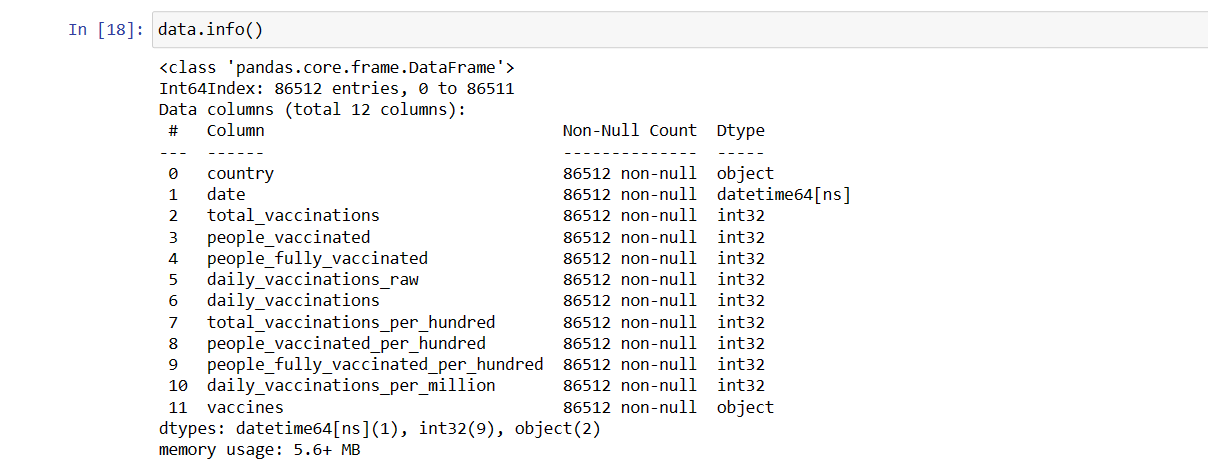
### **4. Data transformation:**

Data transformation involves converting the data from one form to another to make it more suitable for analysis.



Since some of the columns in our dataset contains float datatype we converted them into integer datatype.

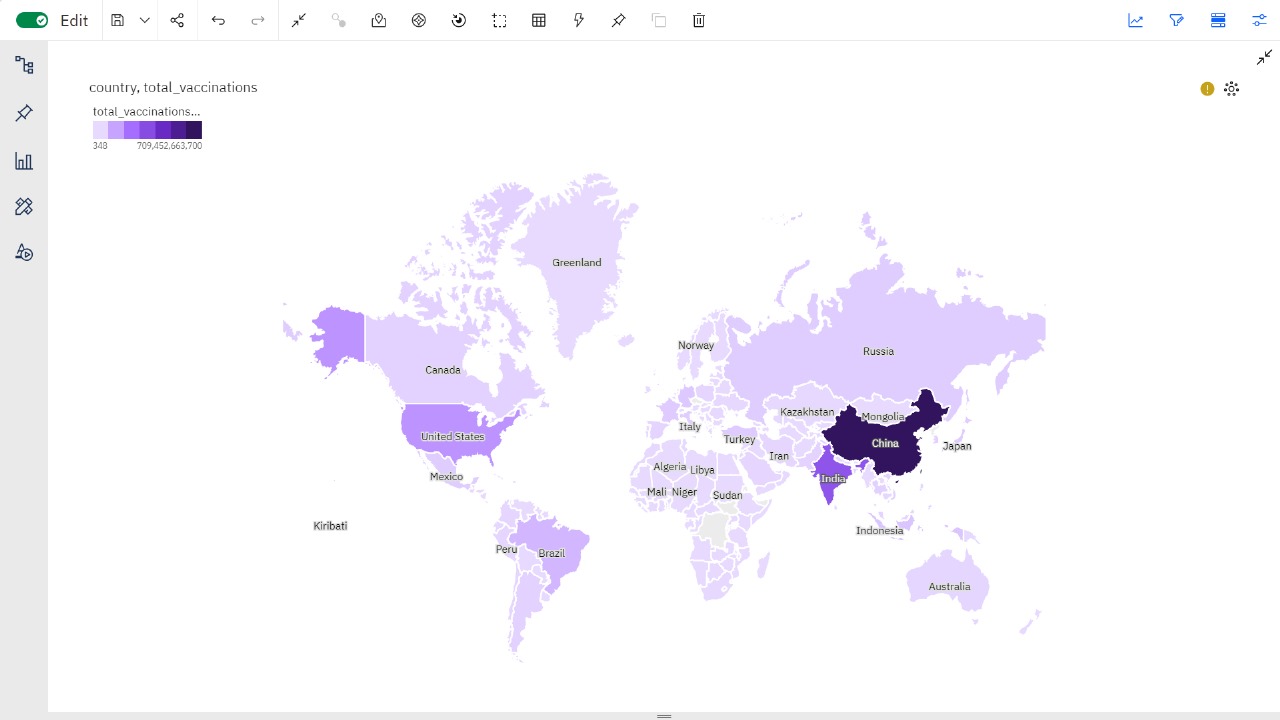




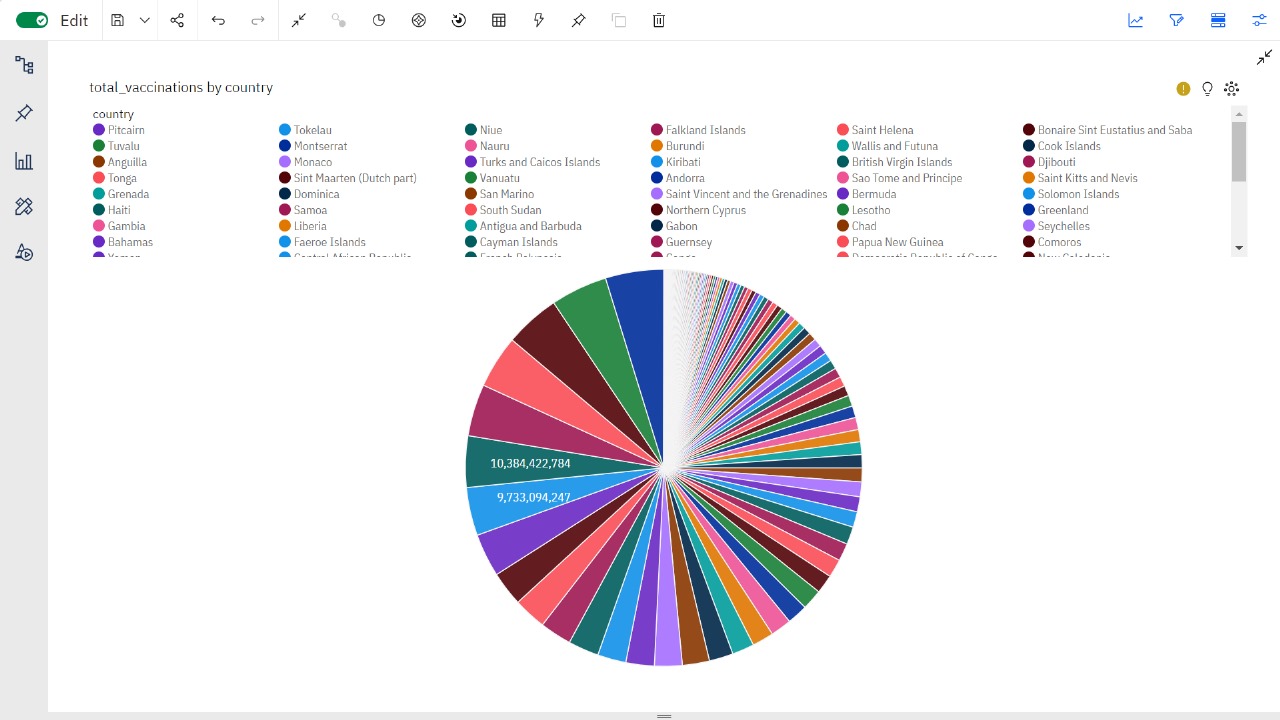
**4.Visualization:**

After loading the cleaned dataset into IBM Cognos , we visualized the relations between some of the columns using IBM Cognos analytics.

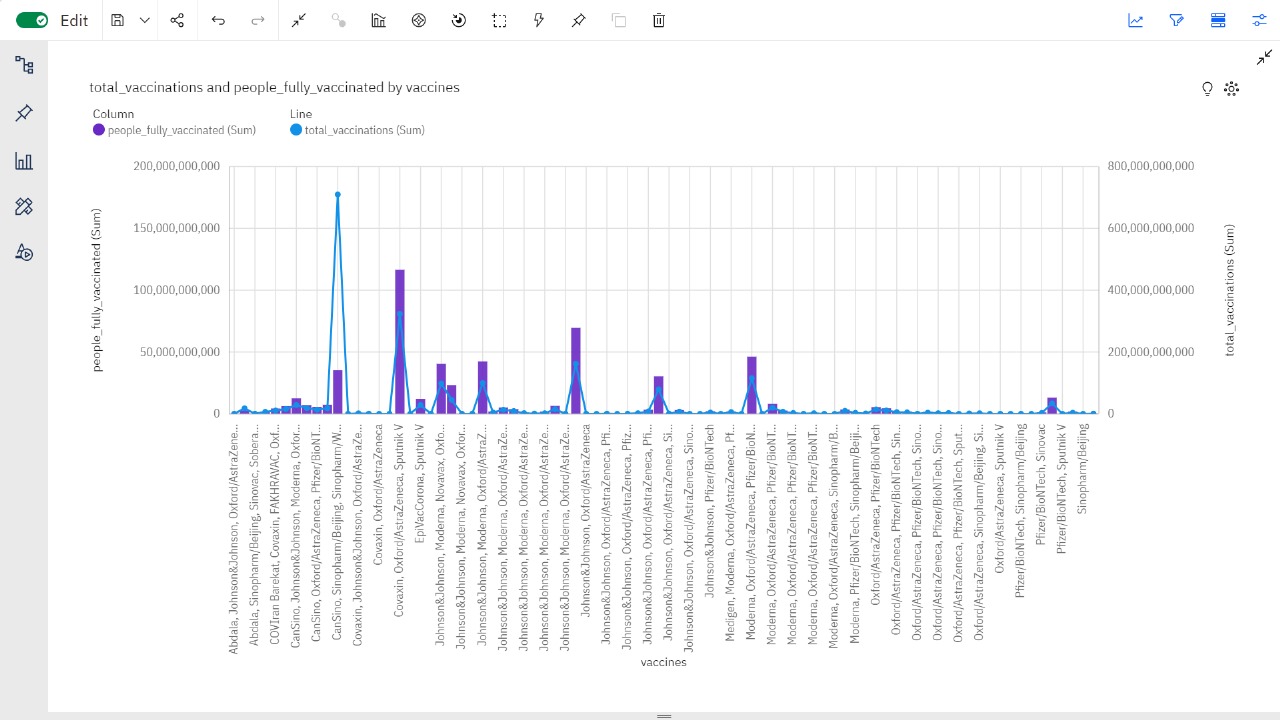
**i)Total no.of vaccination per country:**

****

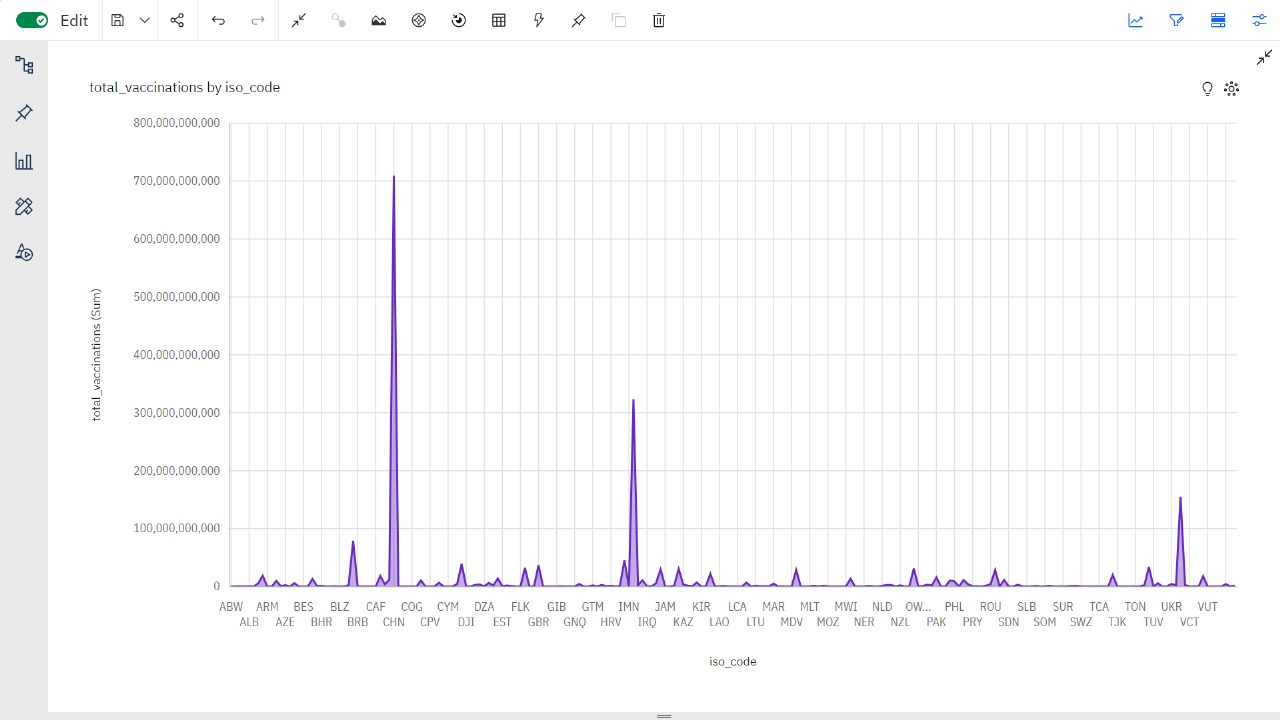
**ii)Total Vaccination per Country**

****

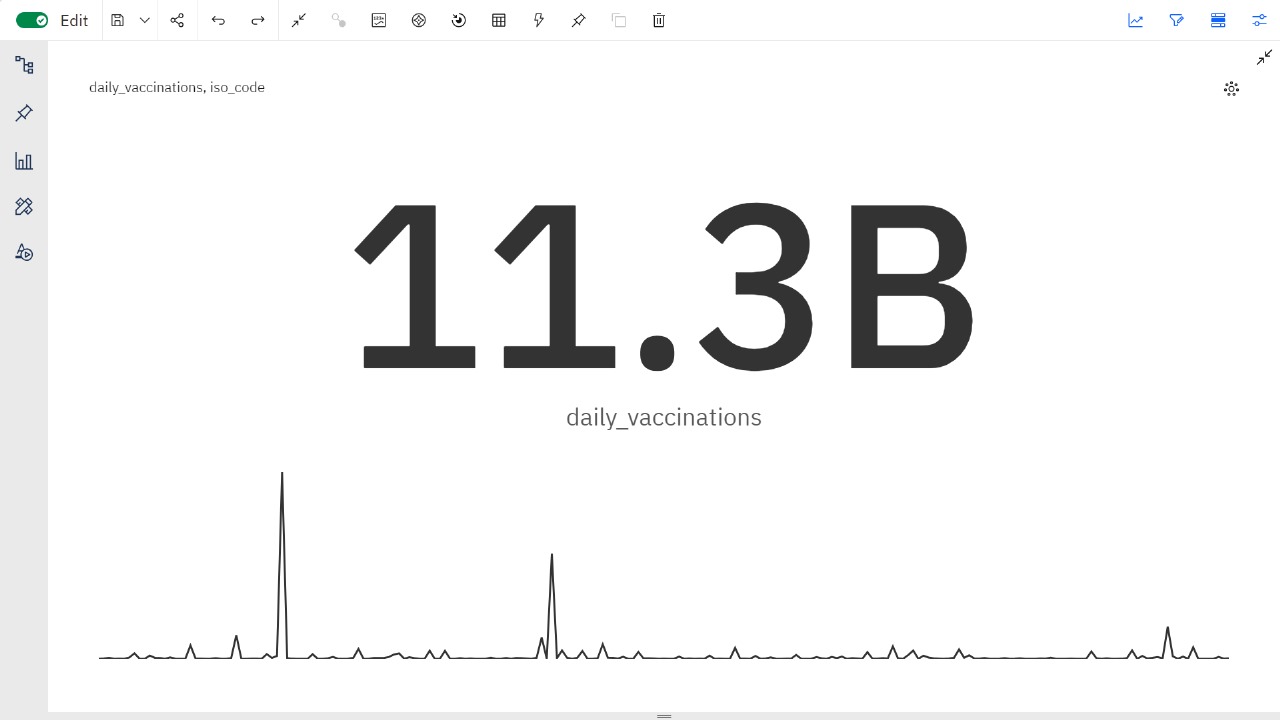
**iii)Total vaccinations and people vaccinated:**

****

**iv)Total vaccinations by ISO code:**

****

**v)Daily vaccinations by ISO code:**

****

**Conclusion:**

The preprocessing of our dataset on COVID vaccine analysis has been successfully completed, ensuring the integrity, reliability, and consistency of our data.

By managing missing values, outliers, and any anomalies present, the data is now primed for more intricate analytical and modelling tasks. Basic visualization is provided along with the trends and patterns associated with the vaccine data. While these initial visual findings have paved the way for deeper understanding, subsequent analyses will be pivotal in drawing more concrete and actionable conclusions. As we continue, it will be crucial to validate our hypotheses and findings, further utilizing this cleaned and structured dataset to its fullest potential.