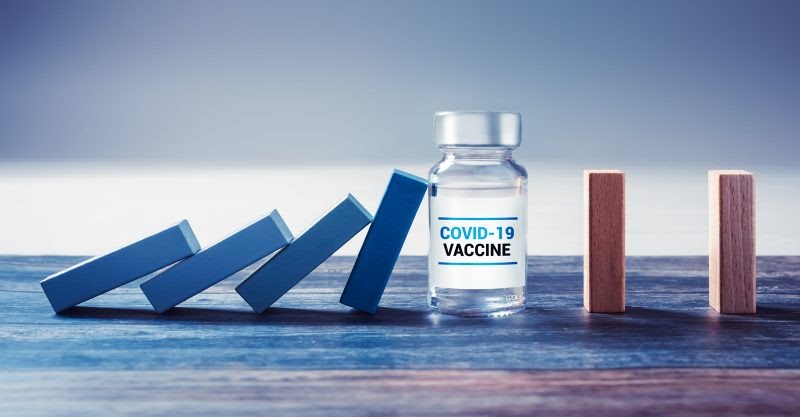
**COVID VACCINES ANALYSIS**

**PHASE-IV:** Data Analysis and Model Building

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



**AIM:**

A COVID-19 analysis report is a comprehensive document that presents data-driven insights and conclusions about the COVID-19 pandemic. It typically covers various aspects, such as the spread of the virus, its impact on public health, economic repercussions, and more. Here's a general structure of such a report:

**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.

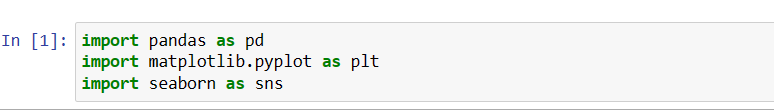
In later stages of 2020 several experimental vaccines were developed and was administered to humans. The efforts were successful as the vaccines were helpful in reducing the affects the virus and even if people were infected, they were not in any life threating situation and escaped the illness having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

# **1.Exploratory Data Analysis (EDA)**

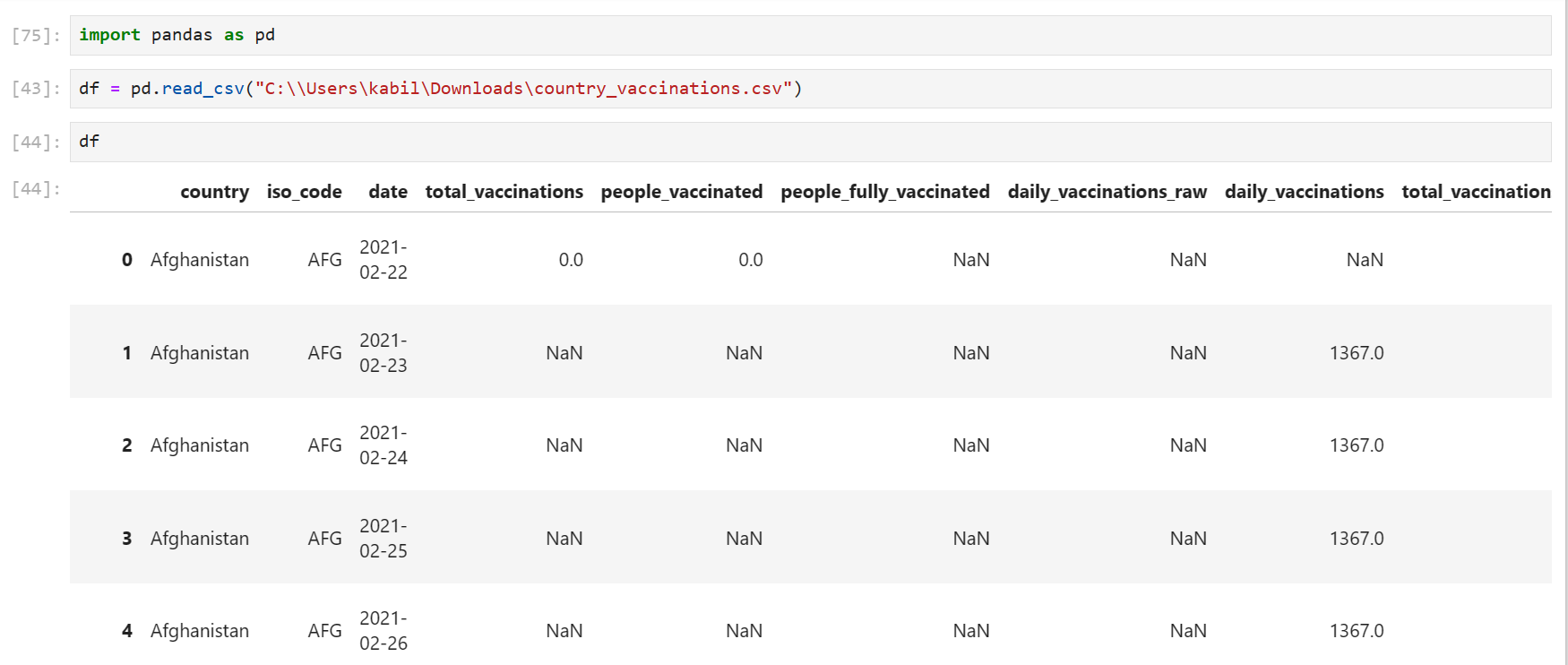
Itinvolves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

**Necessary steps to follow in EDA:**

**1.Import Libraries:**

We start by importing all the necessary libraries that are required for performing exploratory analysis.

**2.Loading the Dataset:**

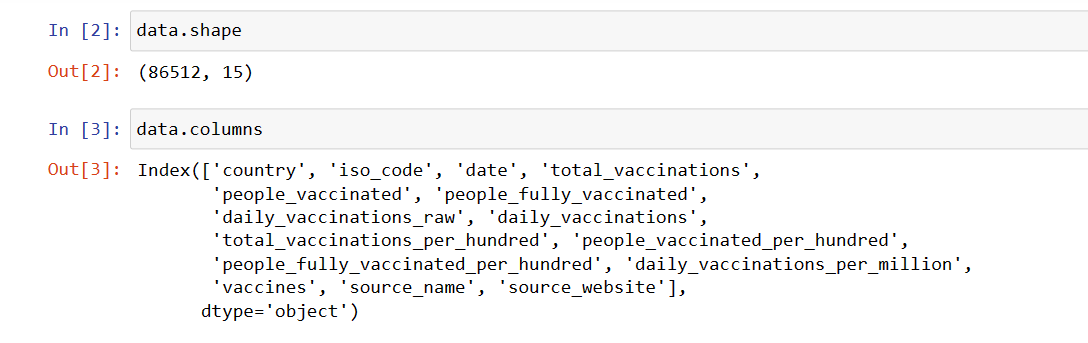
****

#### **Analysing the Data:**

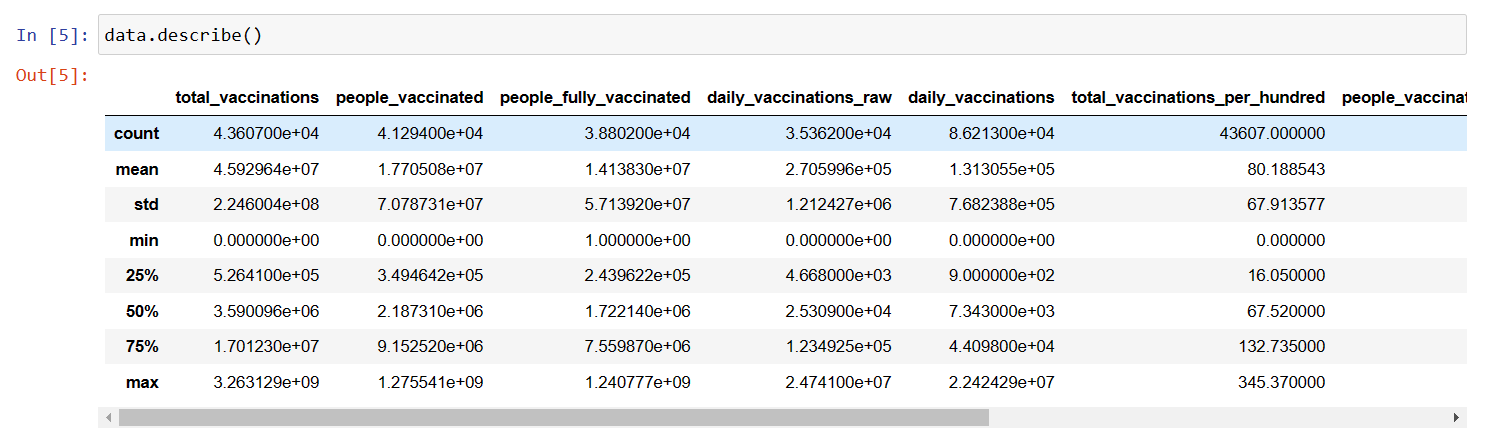
Before we make any inferences, we listen to our data by examining all variables in the data. The main goal of data understanding is to gain general insights about the data, which covers the number of rows and columns, values in the data, datatypes, and Missing values in the dataset.

**Important Steps:**

**1.Data inspection and exploration:**

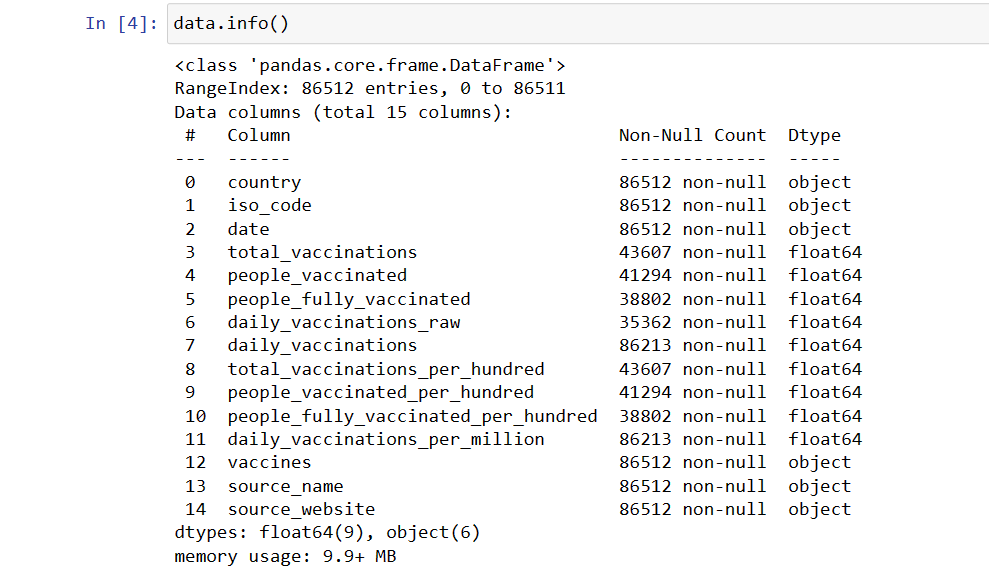
This step involves understanding the data by inspecting its structure and identifying missing values, outliers, and inconsistencies.

.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()**

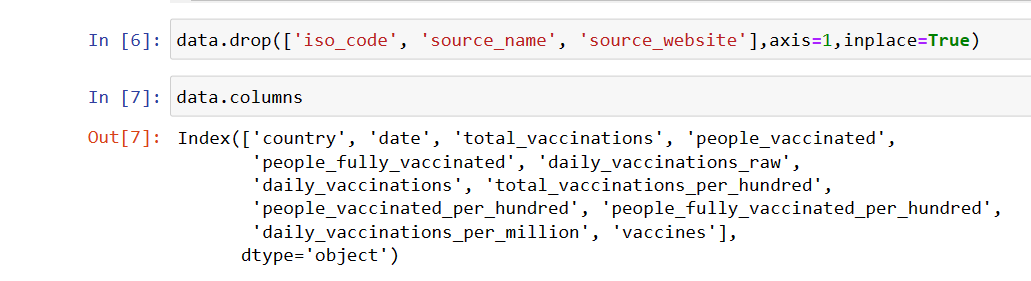
From the below data info, we can see that many columns have an unequal number of counts. And some of the columns have data type objects and some are float values.

****

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

### 

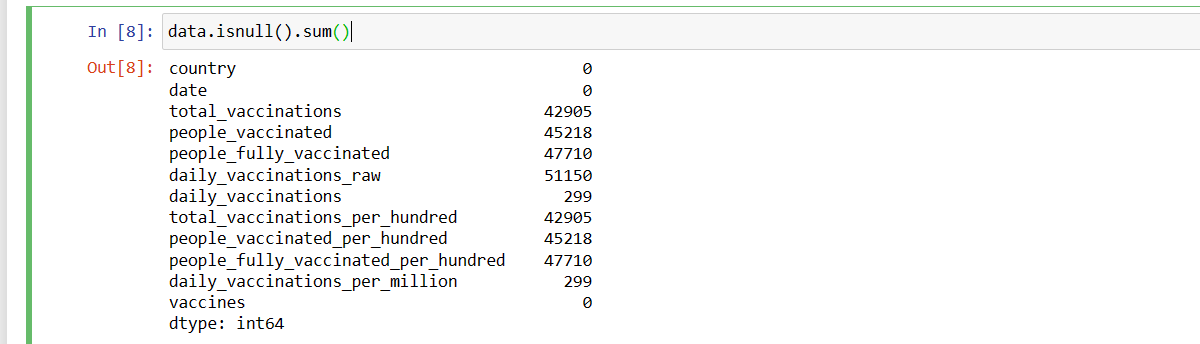
This includes deleting duplicate/ redundant or irrelevant values from our dataset. Duplicate observations most frequently arise during data collection and Irrelevant observations are those that don’t actually fit the specific problem

****

### **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.

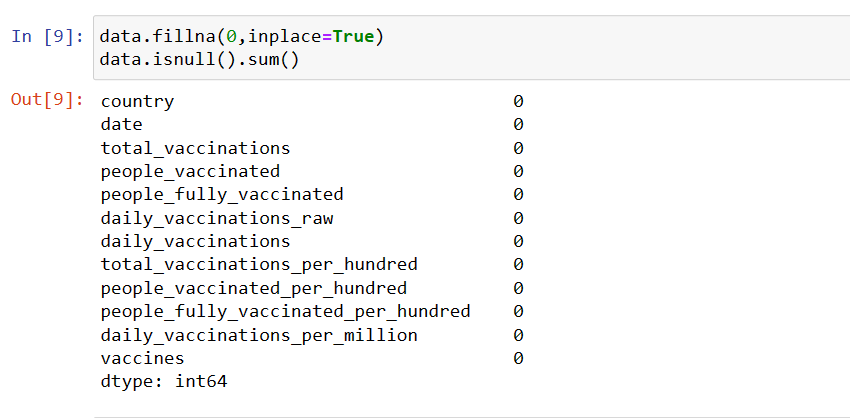
Let’s check the missing values columns-wise for each row using data.isnull() it checks whether the values are null or not and gives returns boolean values and .sum() will sum the total number of null values rows



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

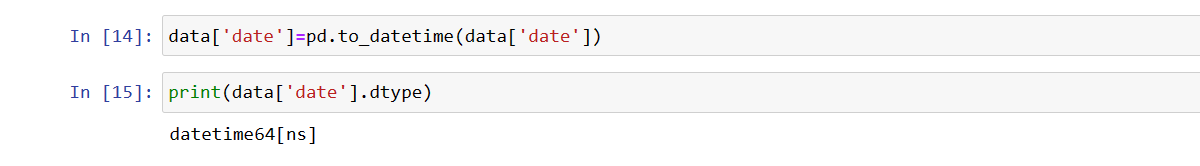
1.Dropping observations with missing values.

2.Inputing the missing values from past observation



**Data transformation:**

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



### **EDA Exploratory Data Analysis:**

Exploratory Data Analysis refers to the crucial process of performing initial investigations on data to discover patterns to check assumptions with the help of summary statistics and graphical representations.

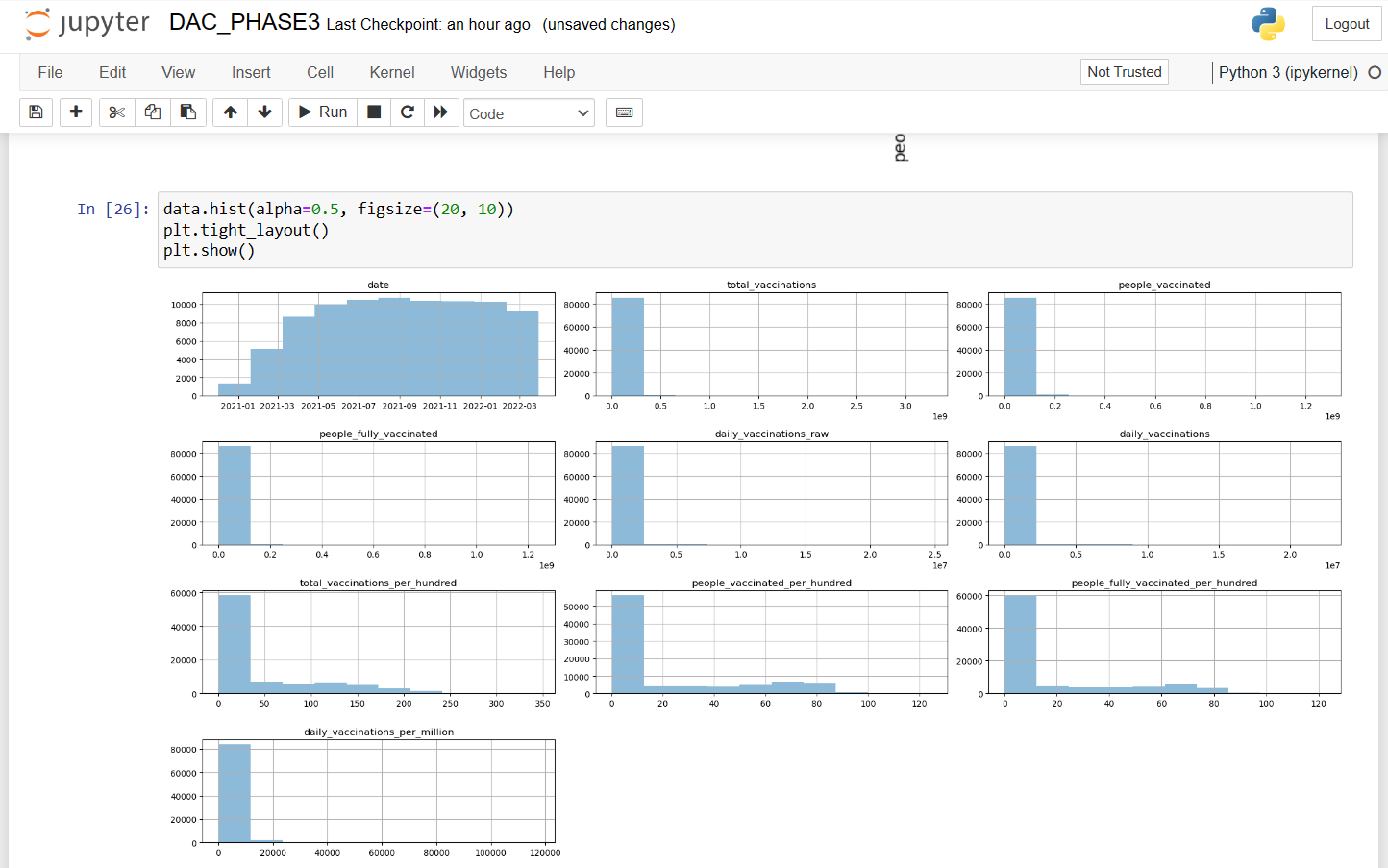
1.EDA Univariate Analysis

2.EDA Bivariate Analysis

3.EDA Multivariate Analysis

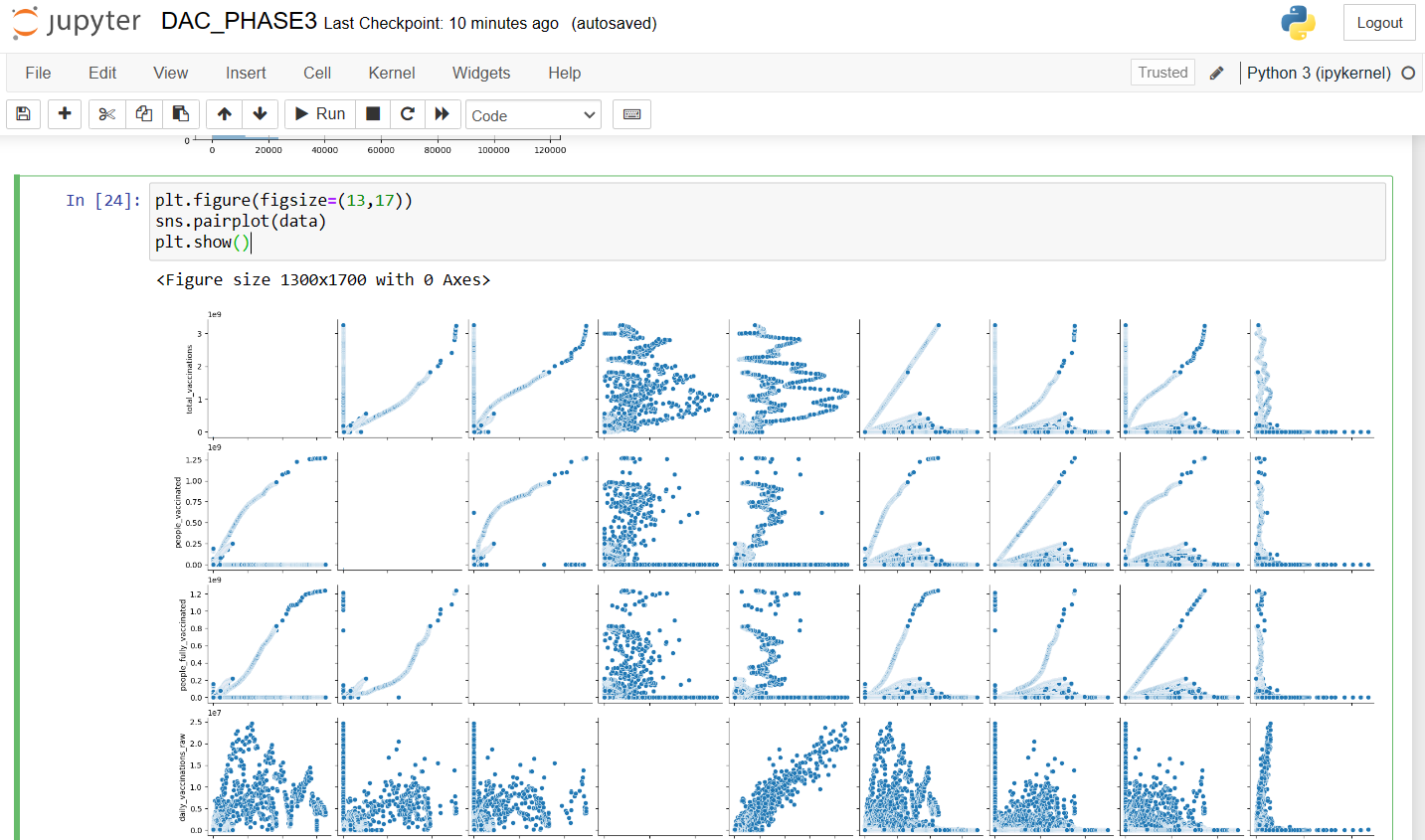
### **EDA Univariate Analysis:**

Univariate analysis can be done for both Categorical and Numerical variables. Categorical variables can be visualized using a Count plot, Bar Chart, Pie Plot, etc. Numerical Variables can be visualized using Histogram, Box Plot, Density Plot, etc.



### **EDA Bivariate Analysis:**

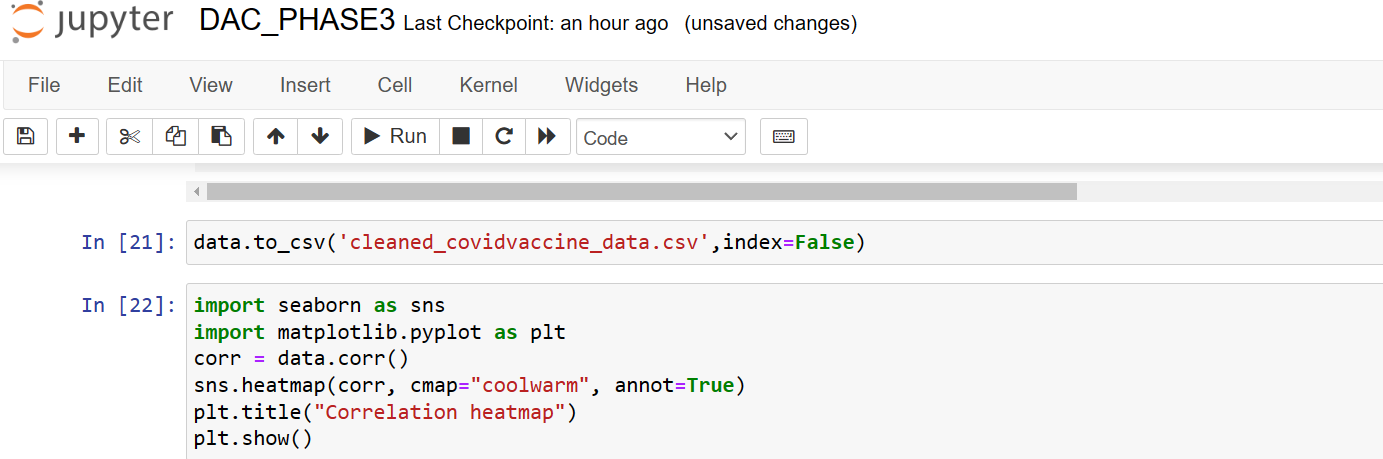
Bivariate Analysis helps to understand how variables are related to each other and the relationship between dependent and independent variables present in the dataset

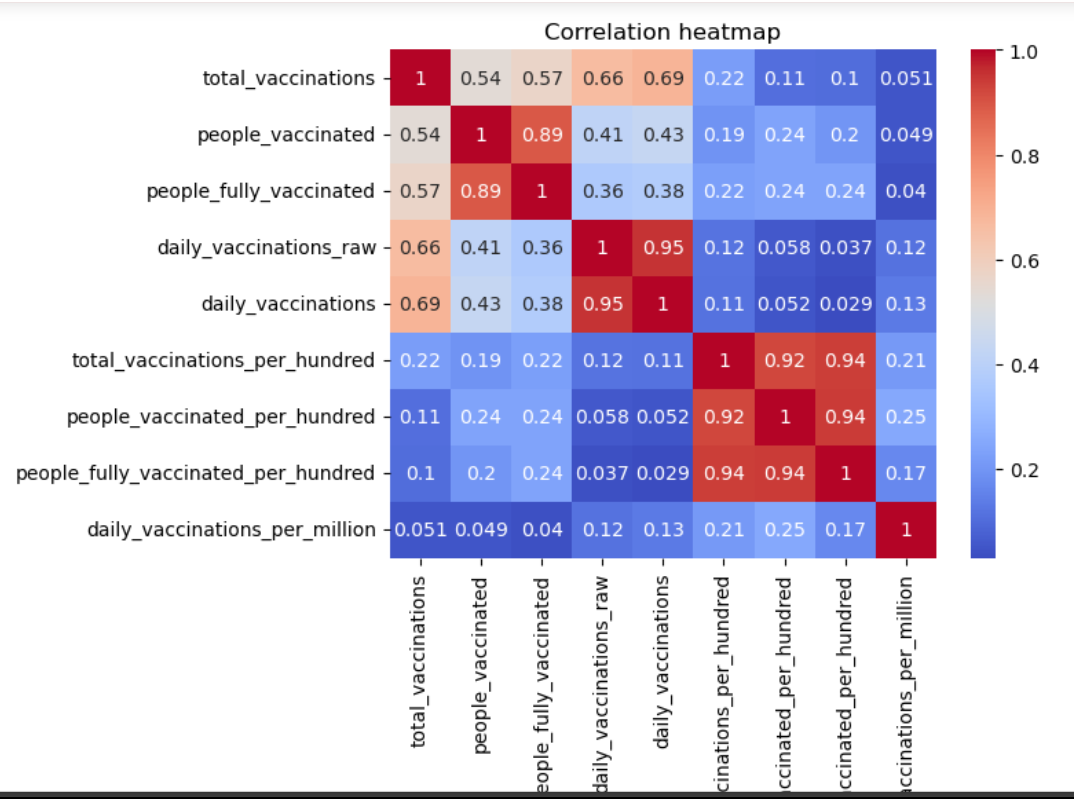


### **EDA Multivariate Analysis:**

Multivariate analysis looks at more than two variables. Multivariate analysis is one of the most useful methods to determine relationships and analyze patterns for any dataset.

A heat map is widely been used for Multivariate Analysis. Heat Map gives the correlation between the variables, whether it has a positive or negative correlation.

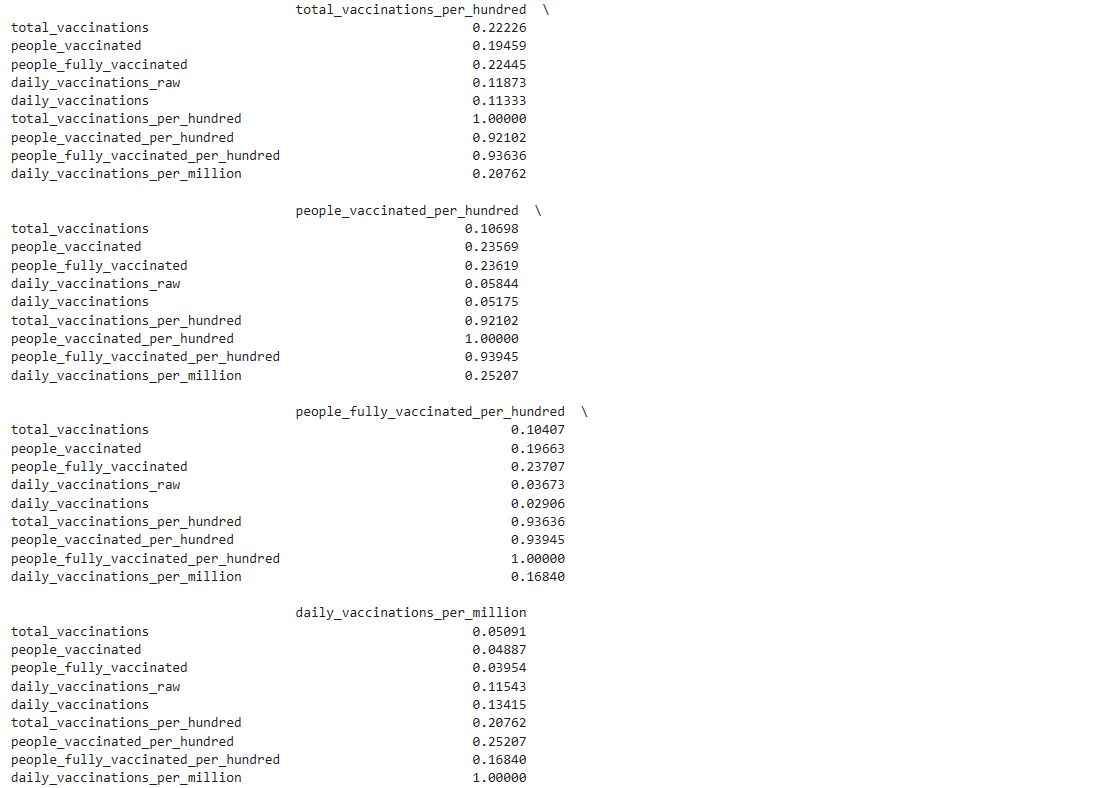




**2.Statistical Analysis:**

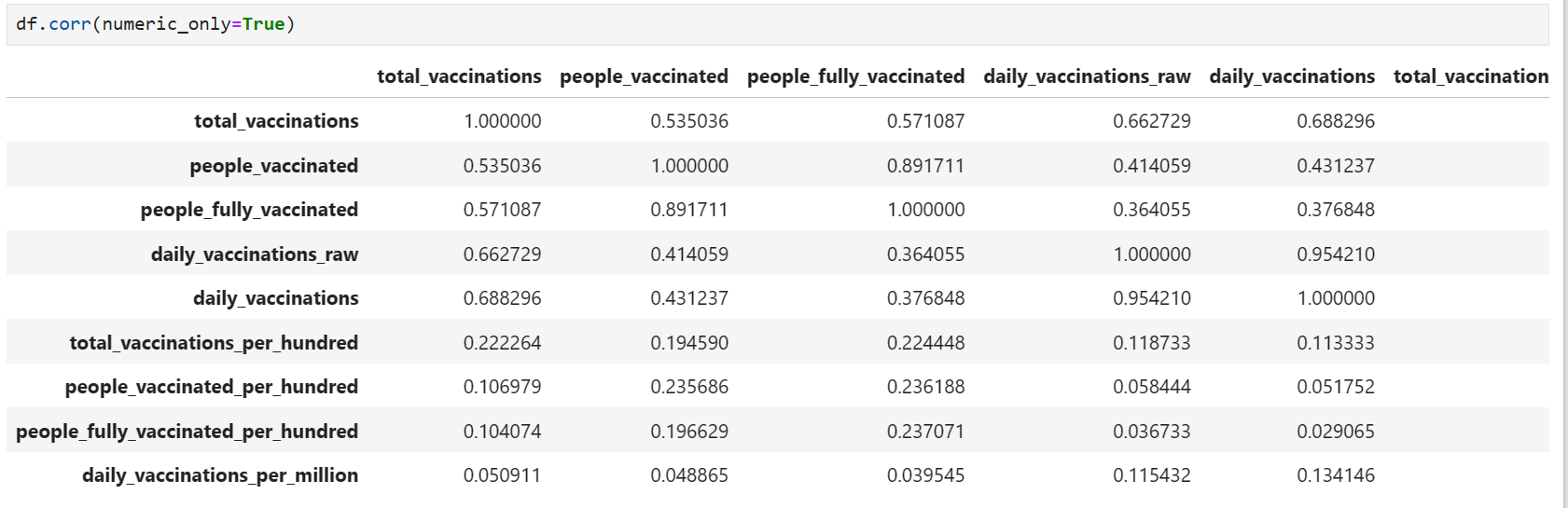
Statistical analysis can cover a wide range of techniques and tests depending on the specific questions you want to answer. Here's a general template for conducting statistical analysis on the COVID-19 World Vaccination Progress dataset using Python and some common statistical tests:





**3.Correlation:**

Correlation is a statistical measure that describes the extent to which two variables change in relation to each other. In other words, it quantifies the degree to which there is a linear relationship between two variables. Correlation does not imply causation but helps us understand how changes in one variable are associated with changes in another.

****

**4.Model Building:**

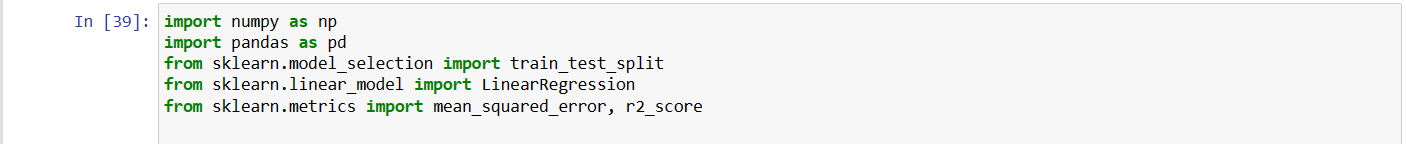
Model building is an essential part of data analysis and is used to extract insights and knowledge from the data to make business decisions and strategies. In this phase of the project data science team needs to develop data sets for training, testing, and production purposes. Model building in data analytics is aimed at achieving not only high accuracy on the training data but also the ability to generalize and perform well on new, unseen data. Therefore, the focus is on creating a model that can capture the underlying patterns and relationships in the data, rather than simply memorizing the training data.

**To do this we divide our dataset into two parts:**

**1.Training Data**

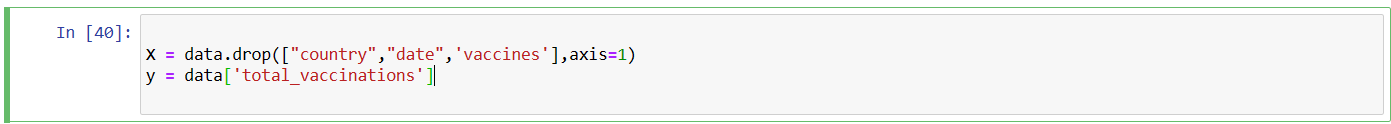
**2.Testing Data**

**Importing libraries**

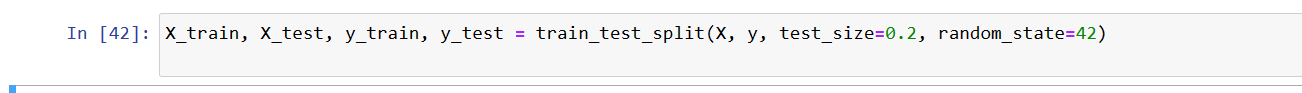


**Splitting Data:**

We need split our dataset into training and testing dataset

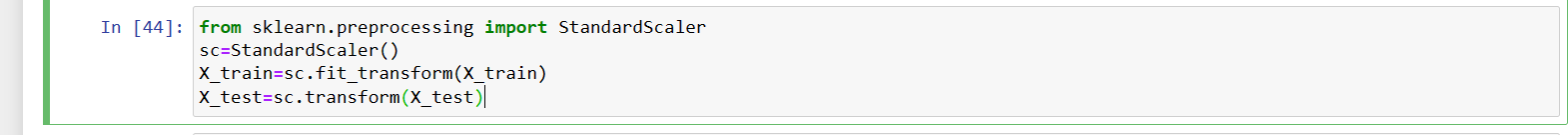


Next we need to train and test those divided dataset. This can be done by using the sklearn library.



**Scaling Data:**

Scaling the dataset is an important preprocessing step before feeding the to the outliers. there are several benefits of scaling the data.

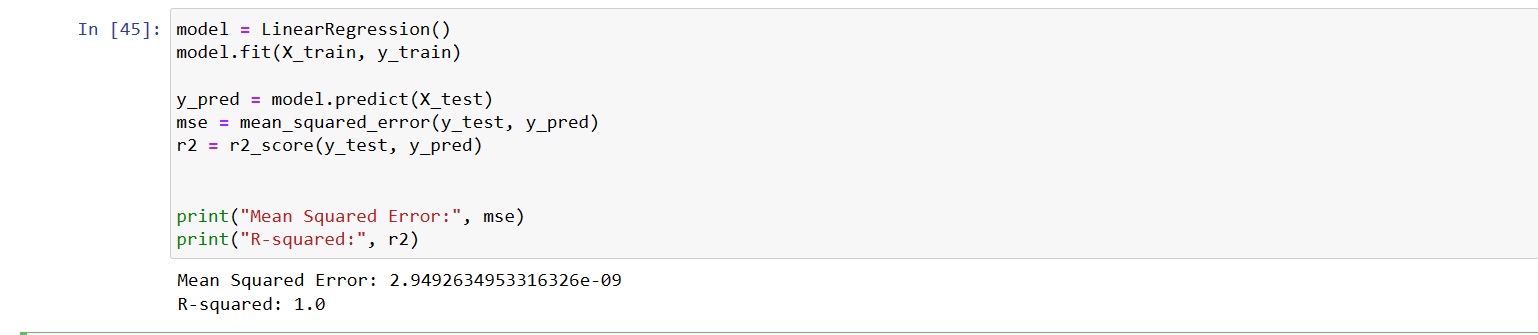
****

**Performing Regression:**

After scaling and splitting the data it has now become ready for fitting to the model. The choice of choosing model totally depends on our problem formulation. There are a variety of models present that we can choose from. However, before choosing the model first, we should identify these points in the data

1. Whether our problem is a regression problem or a classification problem
2. Whether we want a model which is more explainable or we want a model which has a higher accuracy

Here we are performing simple **linear regression** in our dataset

****

## **Evaluating the Model Performance:**

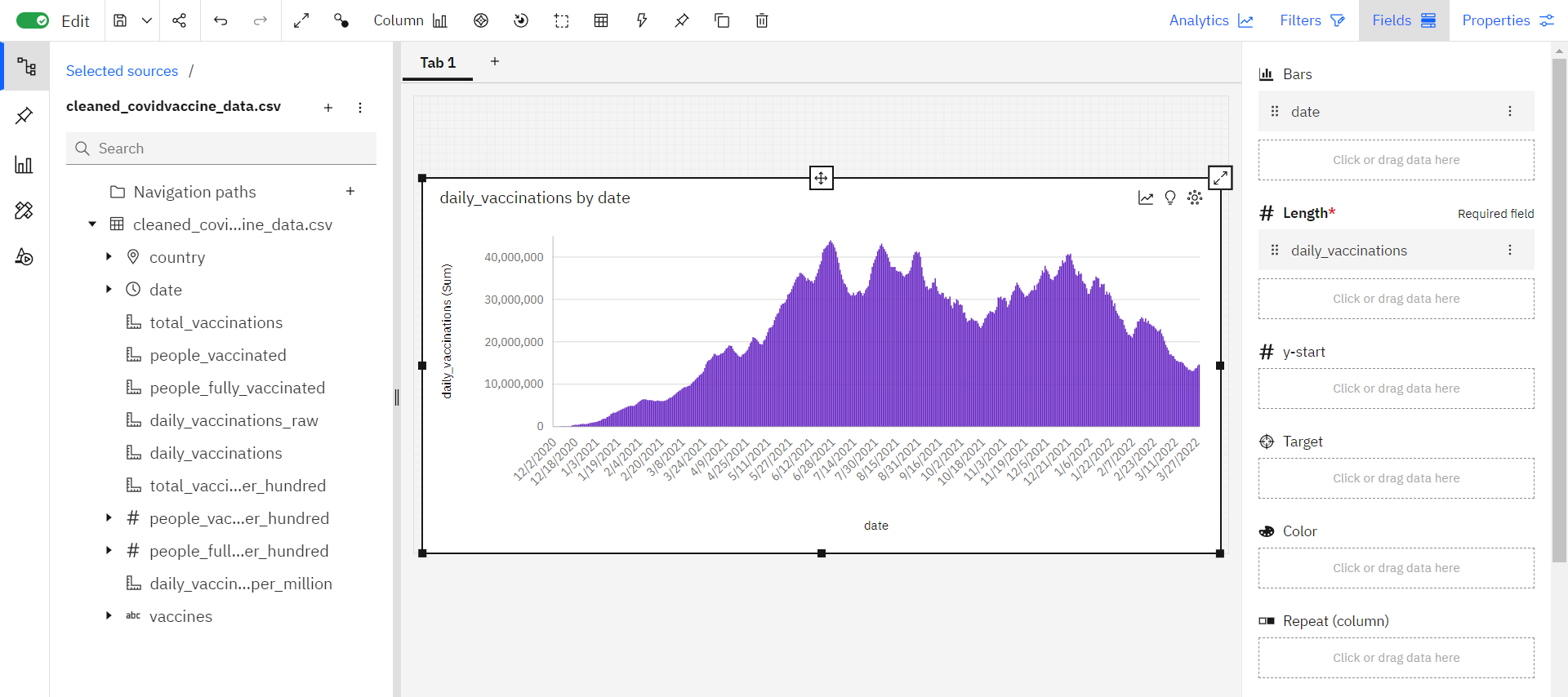
We obtain an r squared of 1.0, meaning a fit of 100%. An R 2 = 1 means that the data is **perfectly correlated**. This is reflected in your standard errors being 0 0. When performing linear regression, you want the value of R2 R 2 to be as close to 1 1 as possible.

We can be sure that this is the model that will perform better in case you want to make a prediction

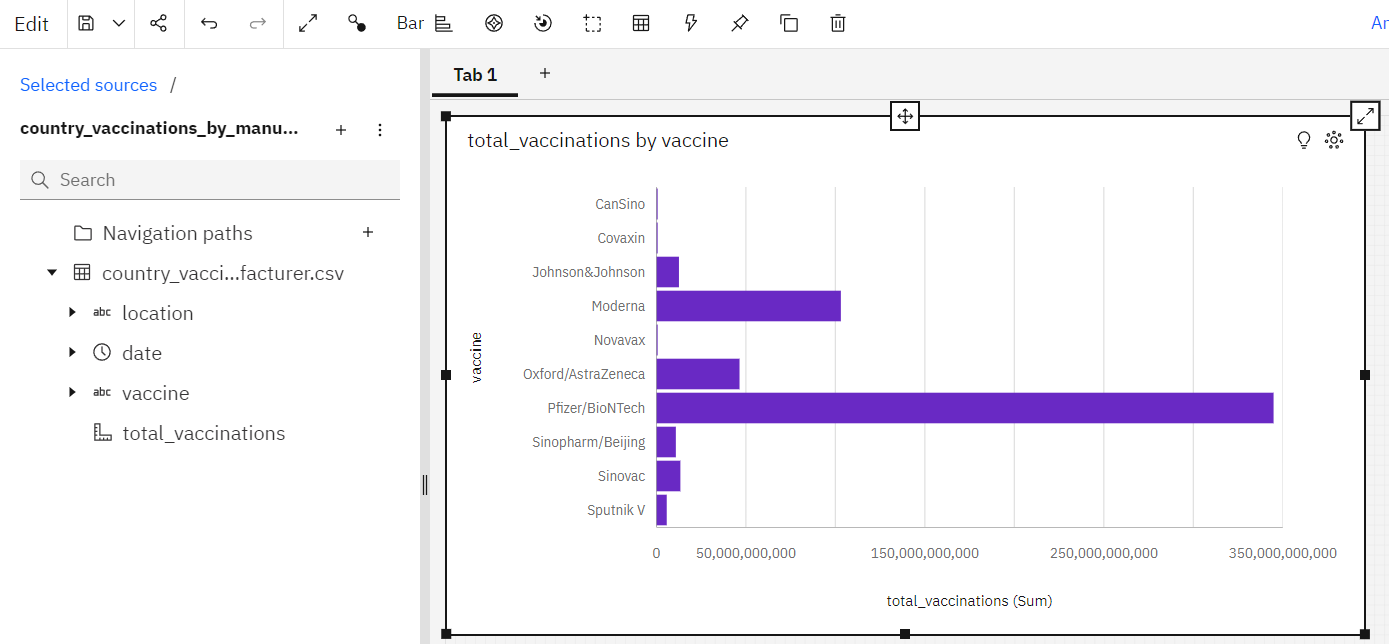
**5.Visualization:**

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

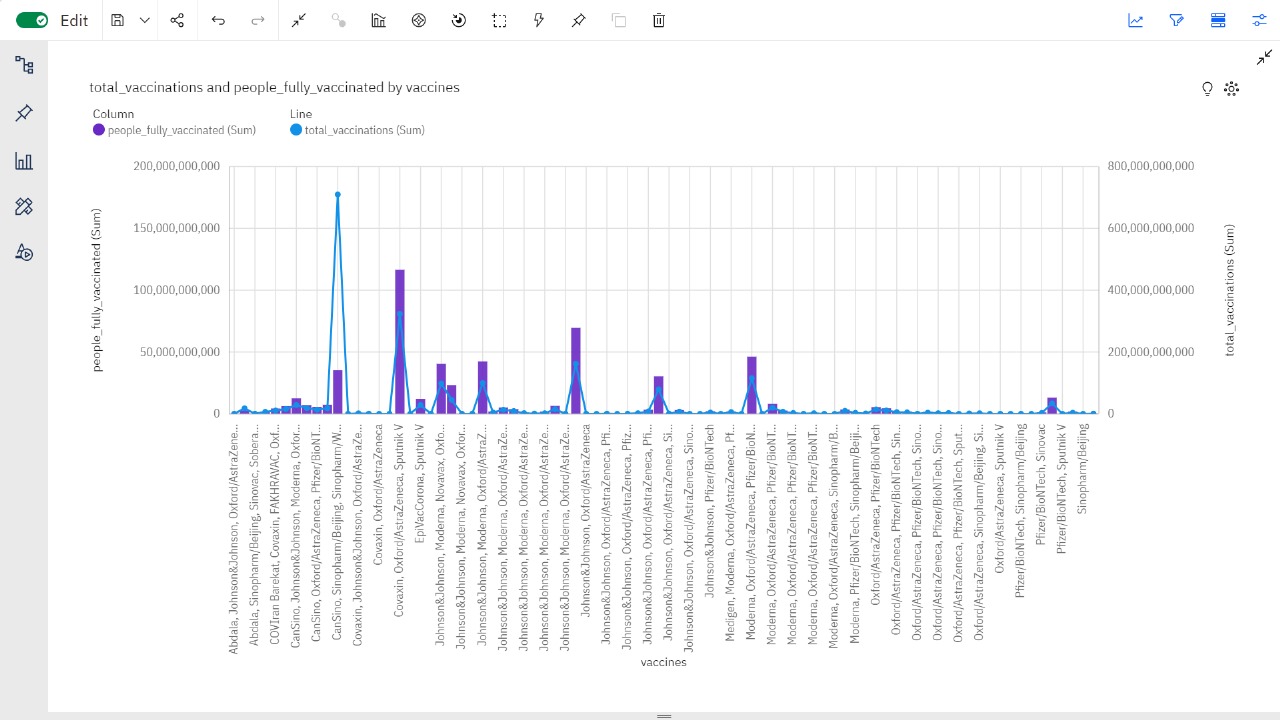
**1.The number of daily vaccinations by date:**



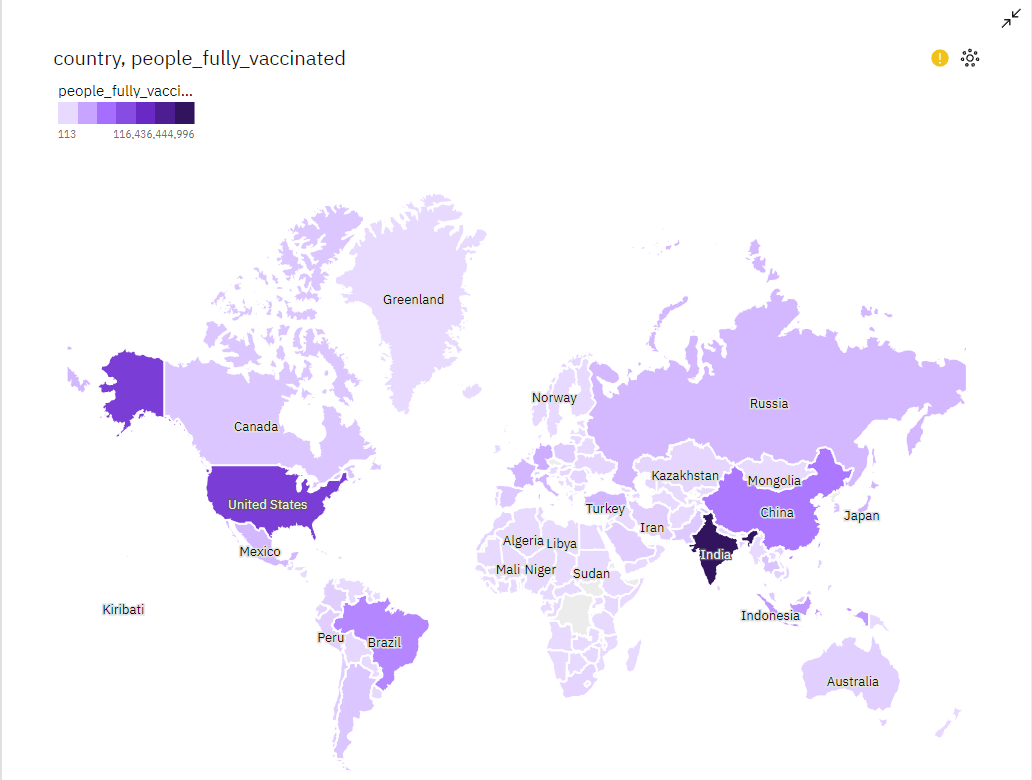
**2.Total Vaccinations by vaccine:**

****

**3.Total vaccinations and people vaccinated:**



**4.Country vs people fully vaccinated:**

****

**Conclusion:**

After an extensive journey of exploratory data analysis and model building in our COVID vaccine analysis, we have gained valuable insights and made significant progress. Our comprehensive examination of the data has illuminated crucial patterns and trends, enabling us to better understand the factors influencing vaccine effectiveness. The models we've developed have shown promise in predicting outcomes and optimizing vaccine distribution strategies.