**COVID VACCINES ANALYSIS**

**TEAM MEMBERS**

**310821243012: DHANUUSH A**

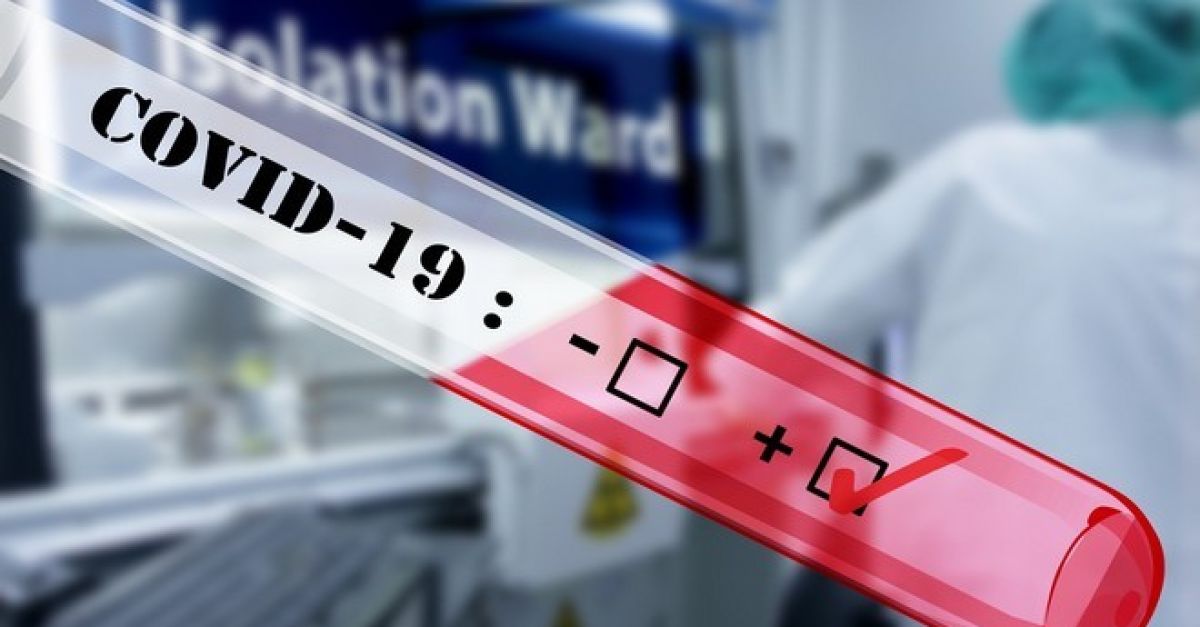
**310821243048: SAI NITHISH KUMAR S**

**310821243034: MURALI KRISHNAN M**

**310821243033: LOKESH V**

**Final Submission:** 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:

**Tools Used in COVID-19 Analysis:**

The tools and software commonly used for COVID-19 analysis include:

**1. Data Collection and Cleaning:**

- Python: Libraries like pandas for data manipulation.

- SQL databases for structured data.

- Web scraping tools to collect data from websites.

**2. Data Analysis and Visualization:**

- Python: Matplotlib, Seaborn, Plotly for data visualization.

- Jupyter Notebooks for interactive data analysis.

- R for statistical analysis.

**3. Machine Learning and Modelling:**

- Machine learning frameworks like scikit-learn for predictive modeling.

- Time-series analysis tools for forecasting.

**4. Geospatial Analysis:**

- Geographic Information System (GIS) software for mapping and spatial analysis.

**5. Reporting:**

- Microsoft Excel or Google Sheets for basic reporting.

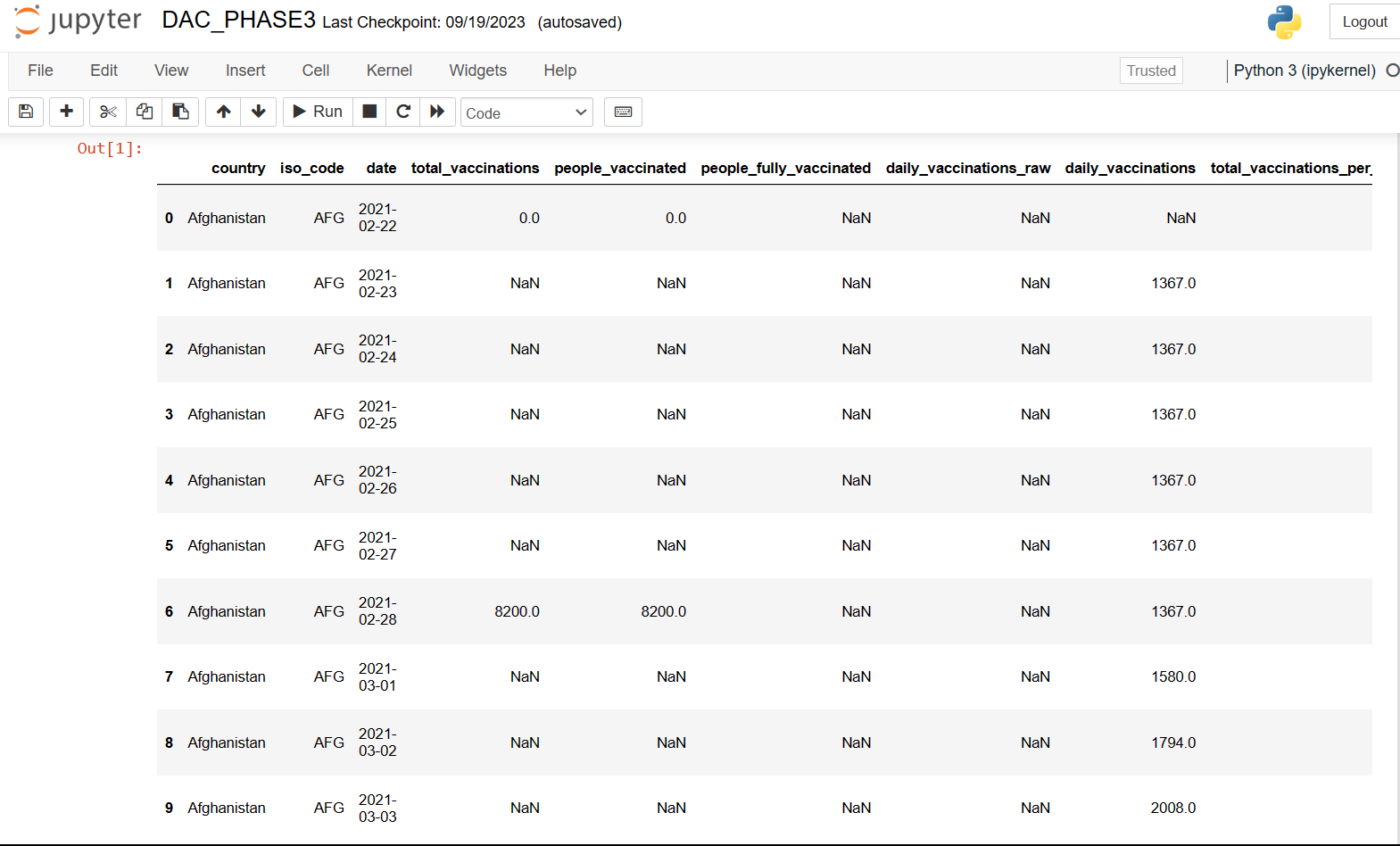
- LaTeX or Markdown for creating structured reports.

**Step-by-Step Explanation:**

Certainly, let's dive into a more detailed step-by-step explanation of a COVID-19 analysis report, including the tools used:

**1. Data Collection:**

- **Data Sources:** Collect data from sources such as official government health departments, World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and research institutions. You can also include data from academic publications, clinical trials, and open data repositories.

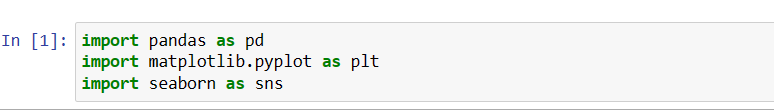


- **Tools:** For web scraping and data extraction, you might use Python libraries like Beautiful Soup and requests, or data APIs provided by official sources.

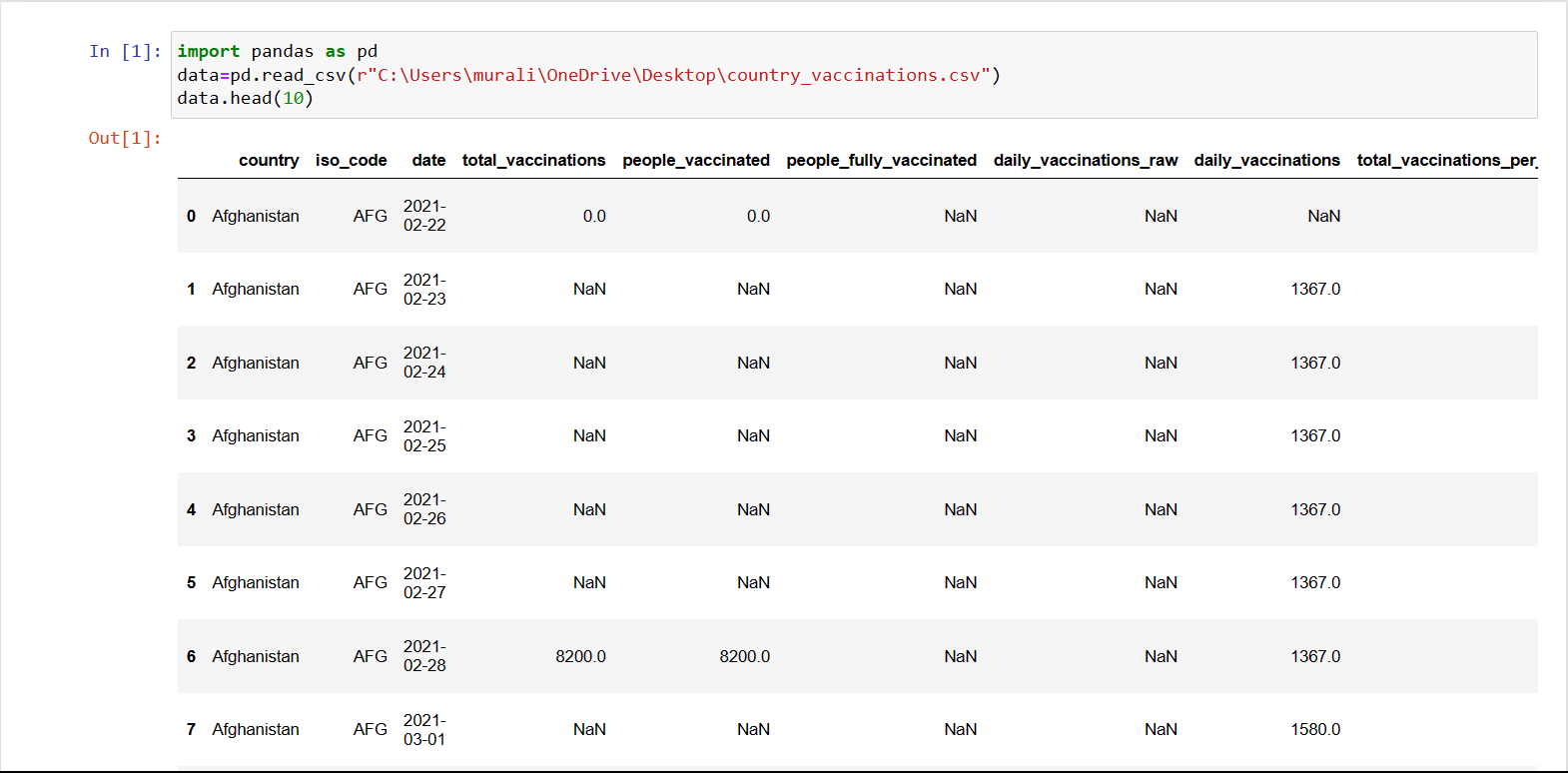
**2. Data Cleaning:**

**Necessary step to follow:**

**1.Import Libraries:**

Start by importing all the necessary libraries.

**2.Loading the DataSet:**

****

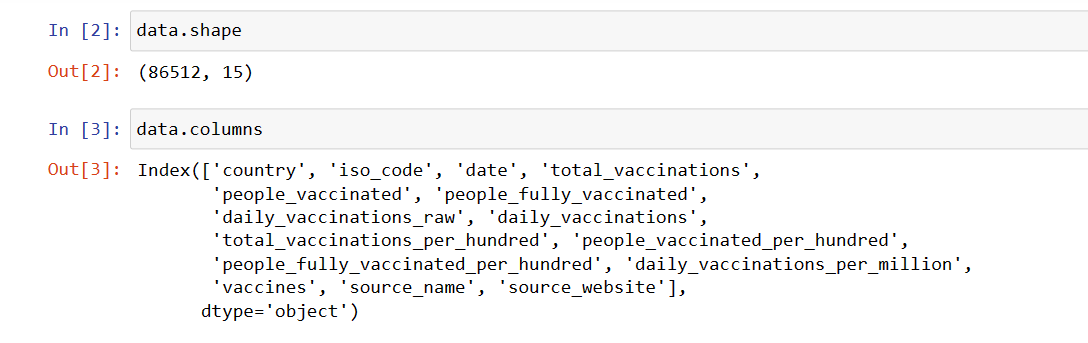
**3.Data Cleaning:**

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.

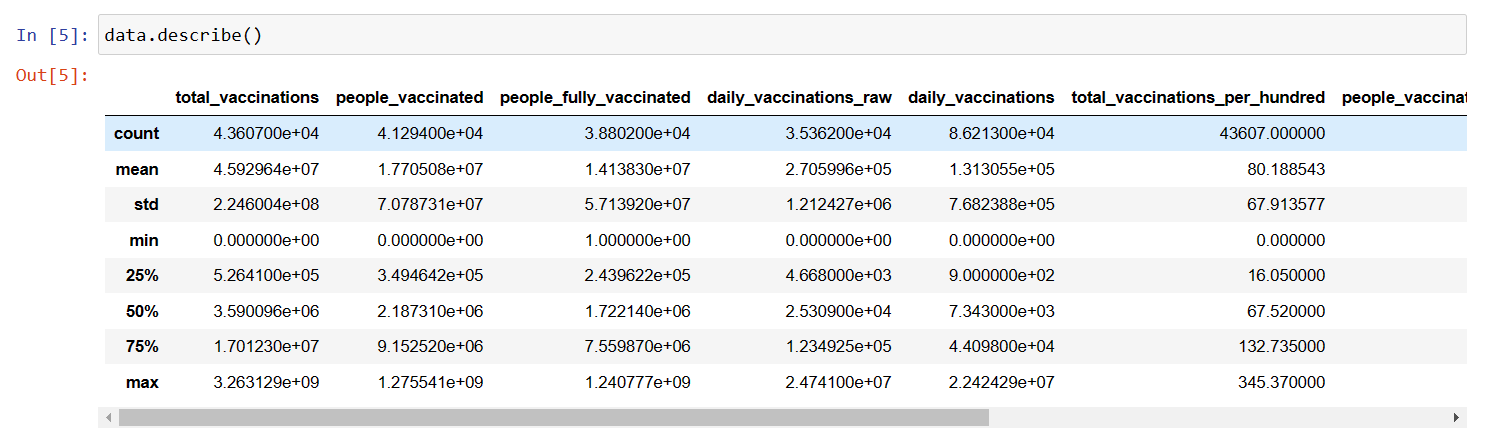
**Important steps :**

**1.Data inspection and exploration:**

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

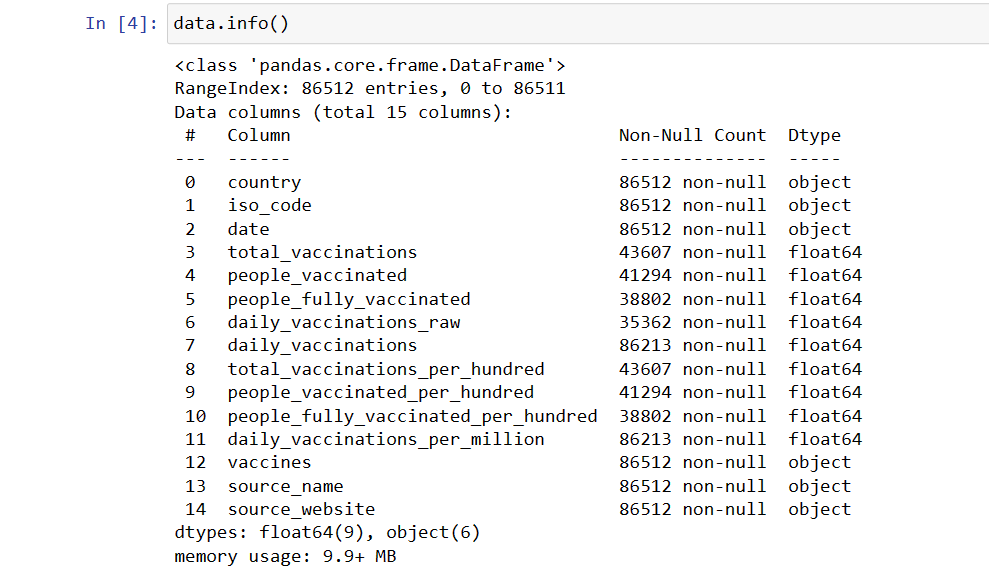


.shape function returns the total no of rows and columns present in our dataset and .columns function gives us the columns names

**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.

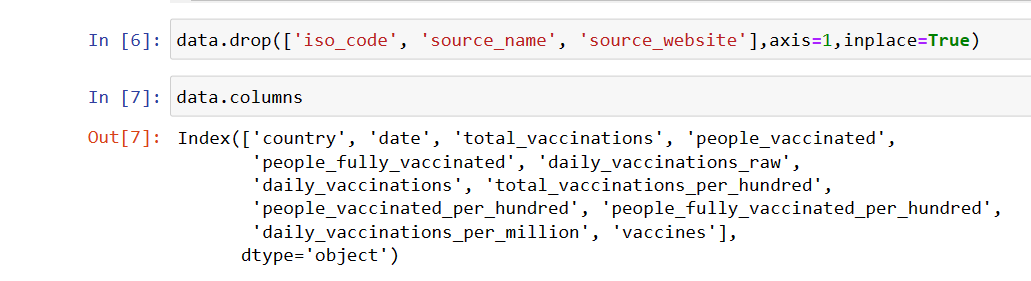
****

### **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

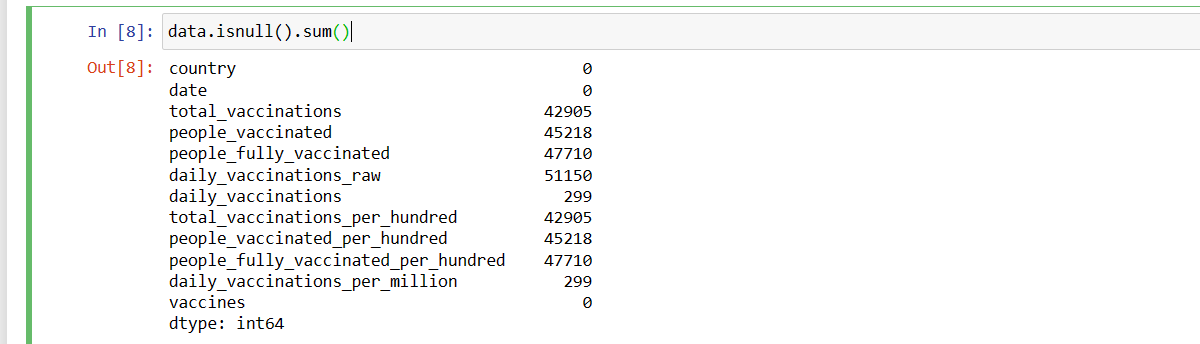
As we know our machines don’t understand the text data. So, we have to either drop or convert the categorical column values into numerical types.Here we are dropping the columns because it hasn’t a great influence on target variables

****

### **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.

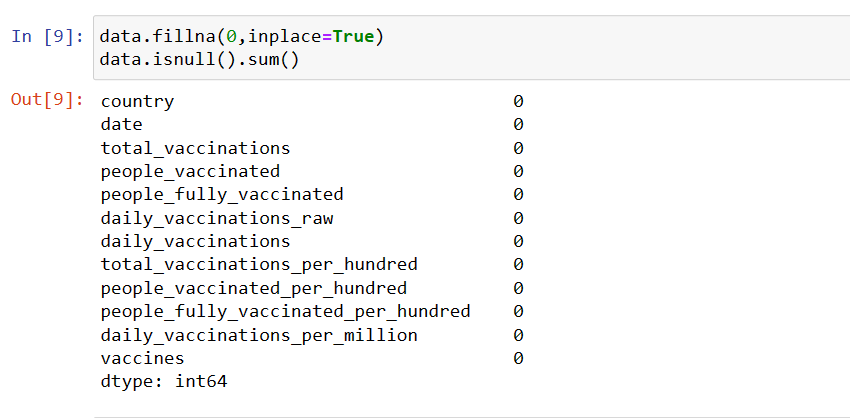
Let’s check the missing values columns-wise for each row using data.isnull() it checks whether the values are null or not and gives returns boolean values and .sum() will sum the total number of null values rows



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

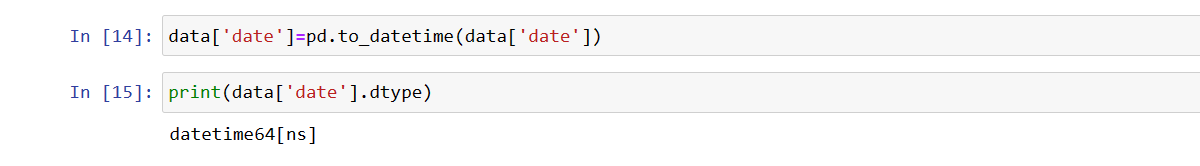
1.Dropping observations with missing values.

2.Inputing the missing values from past observation

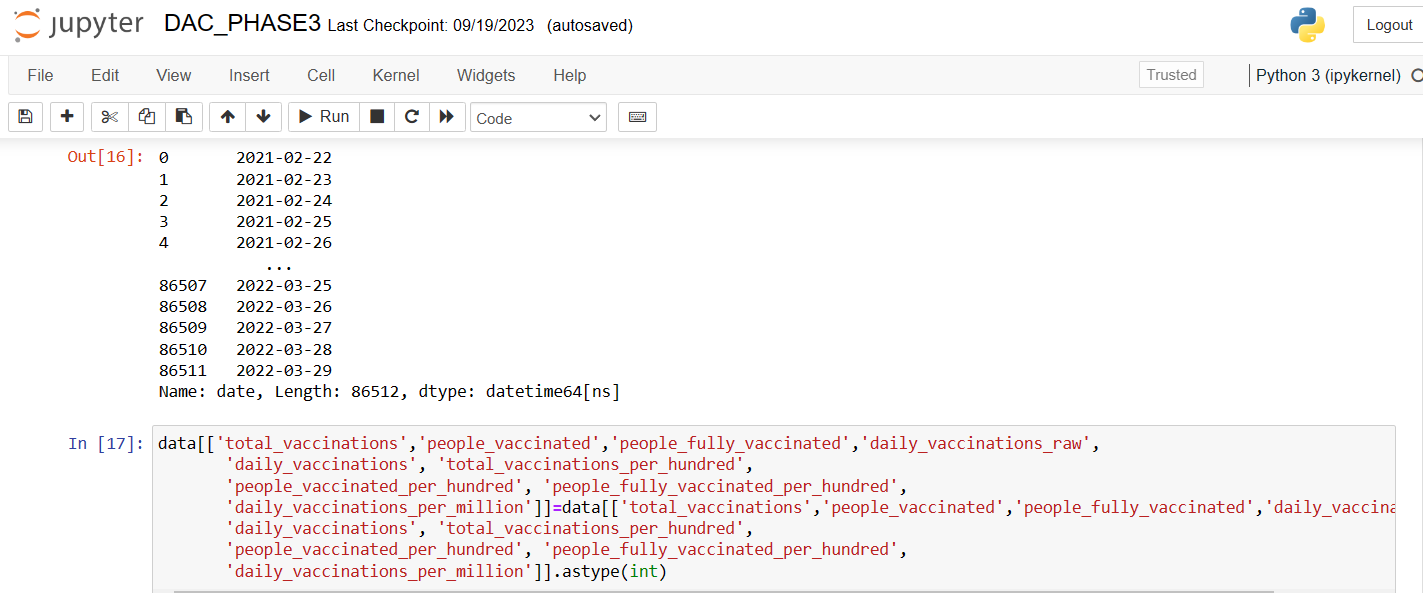


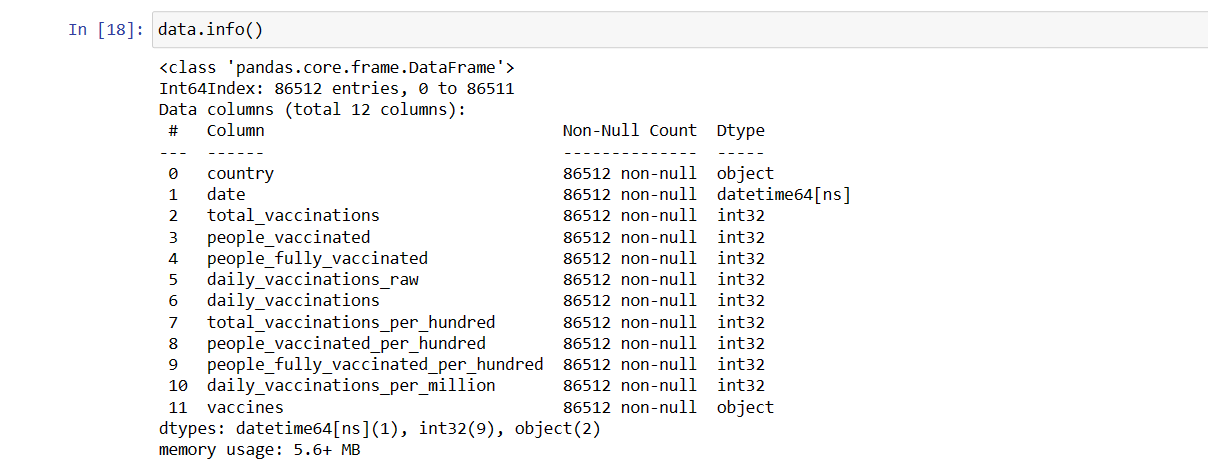
### **4. Data transformation:**

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.





**5.Model Building:**

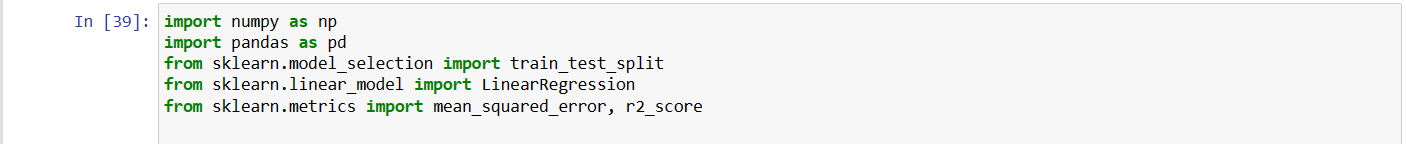
Model building is an essential part of data analysis and is used to extract insights and knowledge from the data to make business decisions and strategies. In this phase of the project data science team needs to develop data sets for training, testing, and production purposes. Model building in data 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**

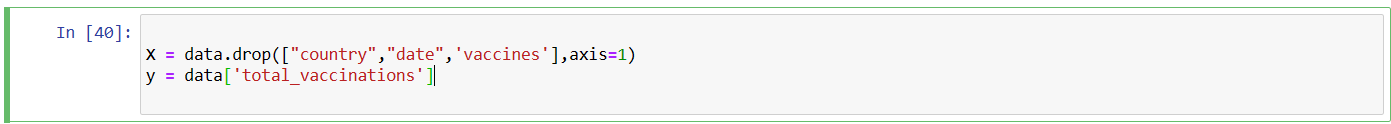
**2.Testing Data**

**Importing libraries**

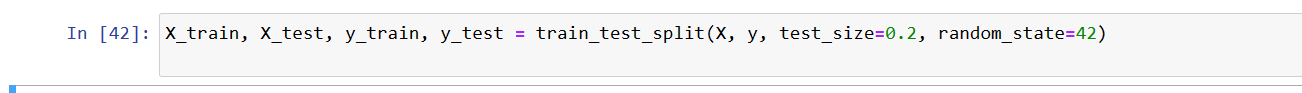


**Splitting Data:**

We need split our dataset into training and testing dataset

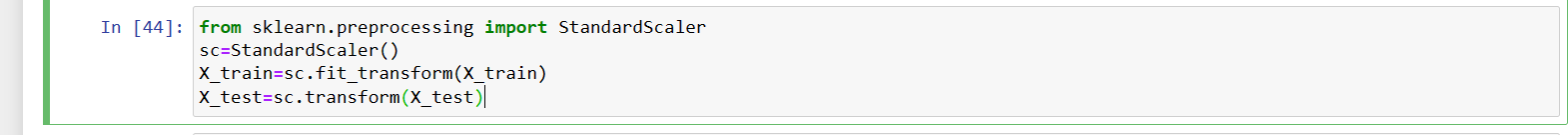


Next we need to train and test those divided dataset. This can be done by using the sklearn library.



**Scaling Data:**

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

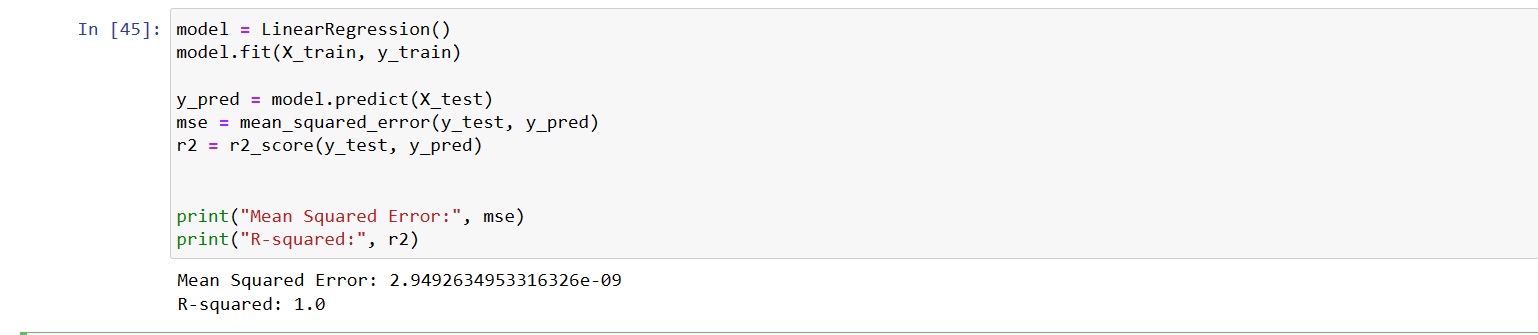
****

**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

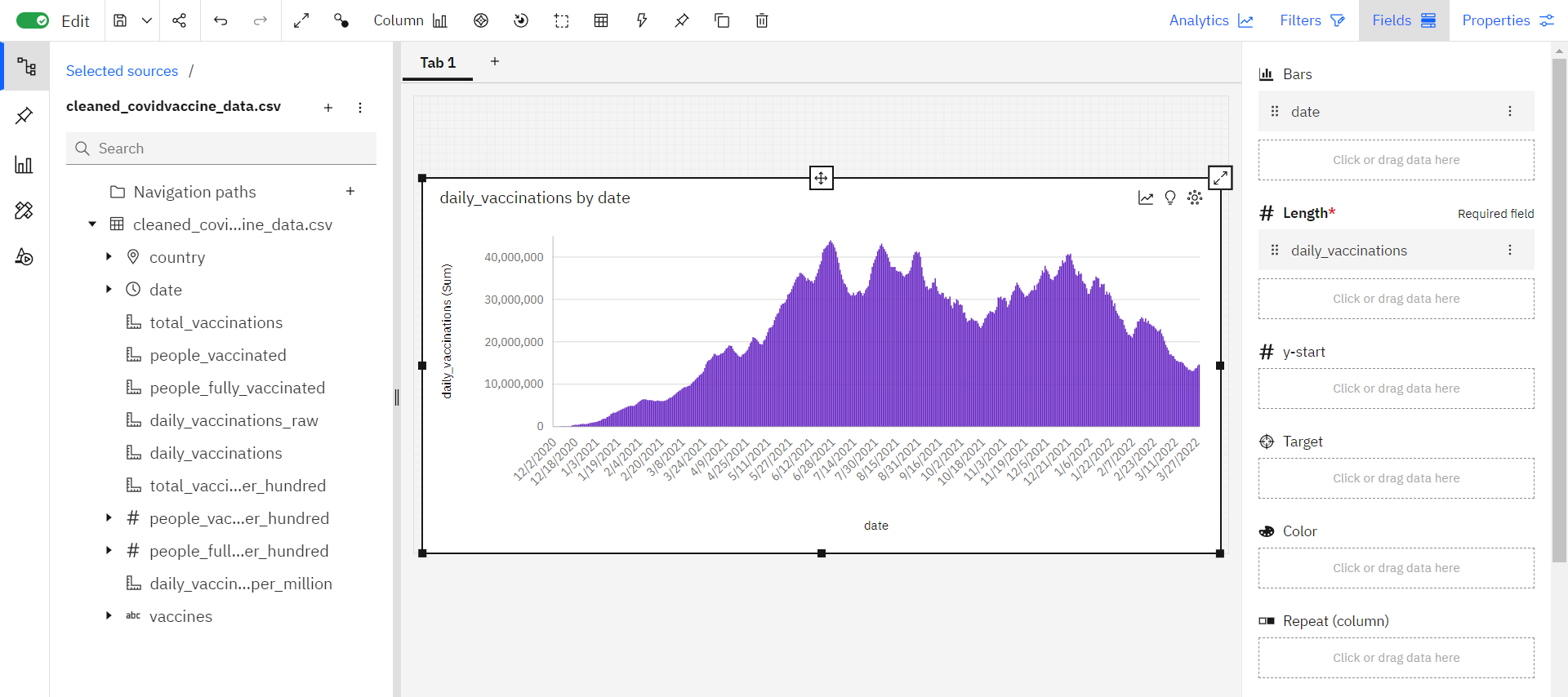
**6. Data Visualization:**

- **Charts and Graphs:** Create various types of charts, such as line graphs for time-series data, bar charts to compare regions, and heatmaps to show the geographic distribution of cases.

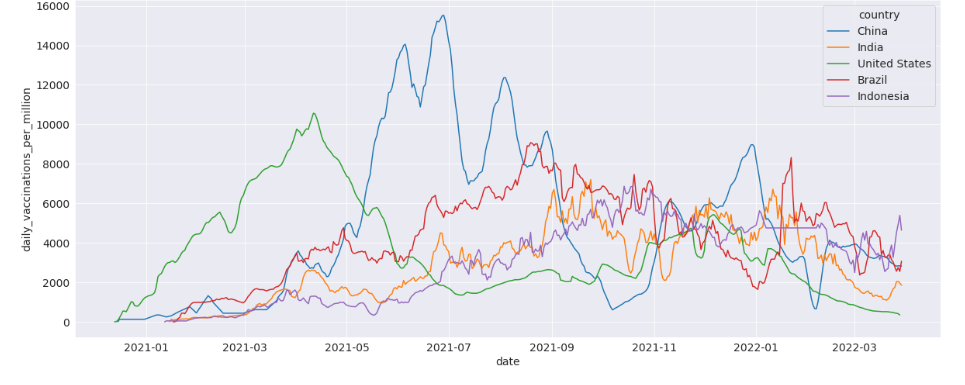
- **Geospatial Visualization:** Use Geographic Information System (GIS) tools like ArcGIS or open-source options like QGIS to create maps that display the spatial distribution of cases.

- **Tools:** Python libraries like Matplotlib, Seaborn, and Plotly for data visualization. For geospatial analysis and mapping, consider ArcGIS, QGIS, or Python libraries like Folium and Geopandas.

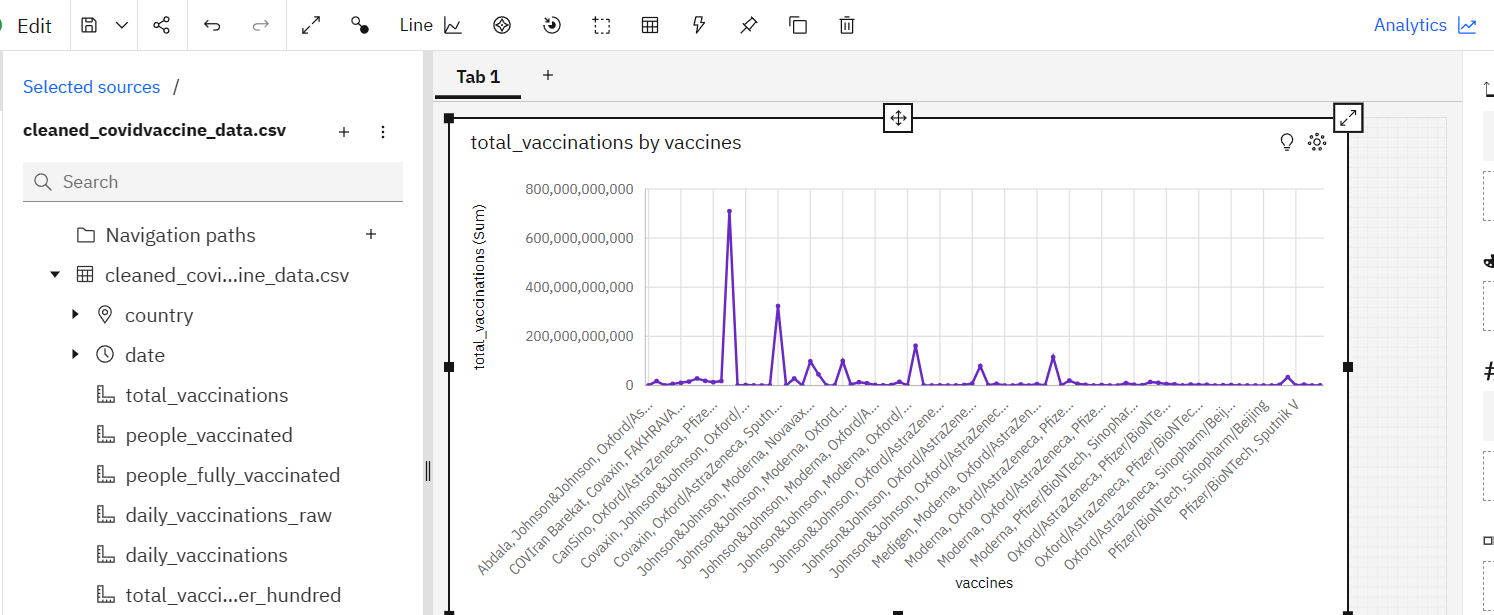
**a.The number of daily vaccinations by date :**



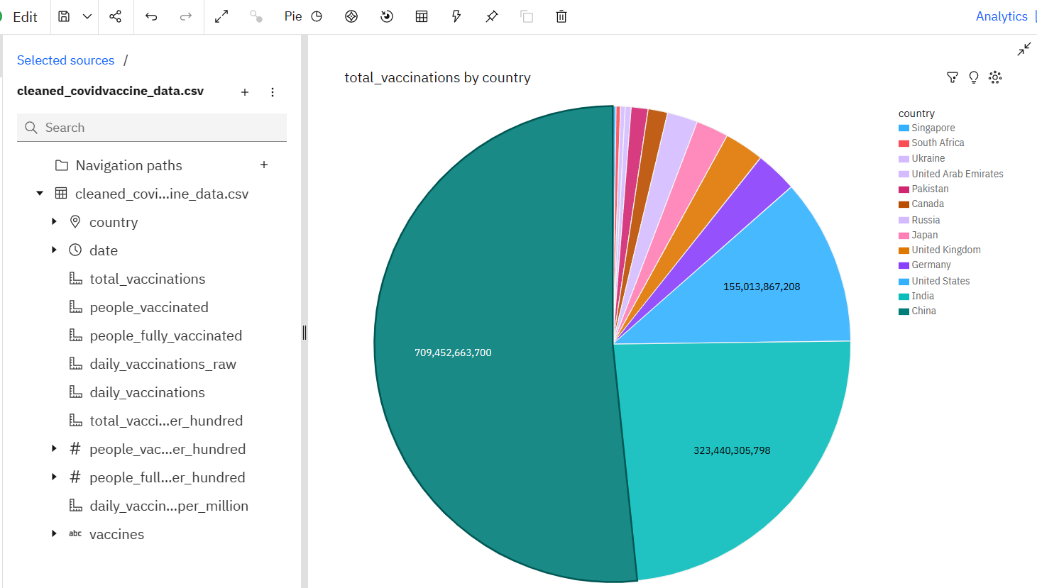
**b.Some countries vaccination progresses:**

****

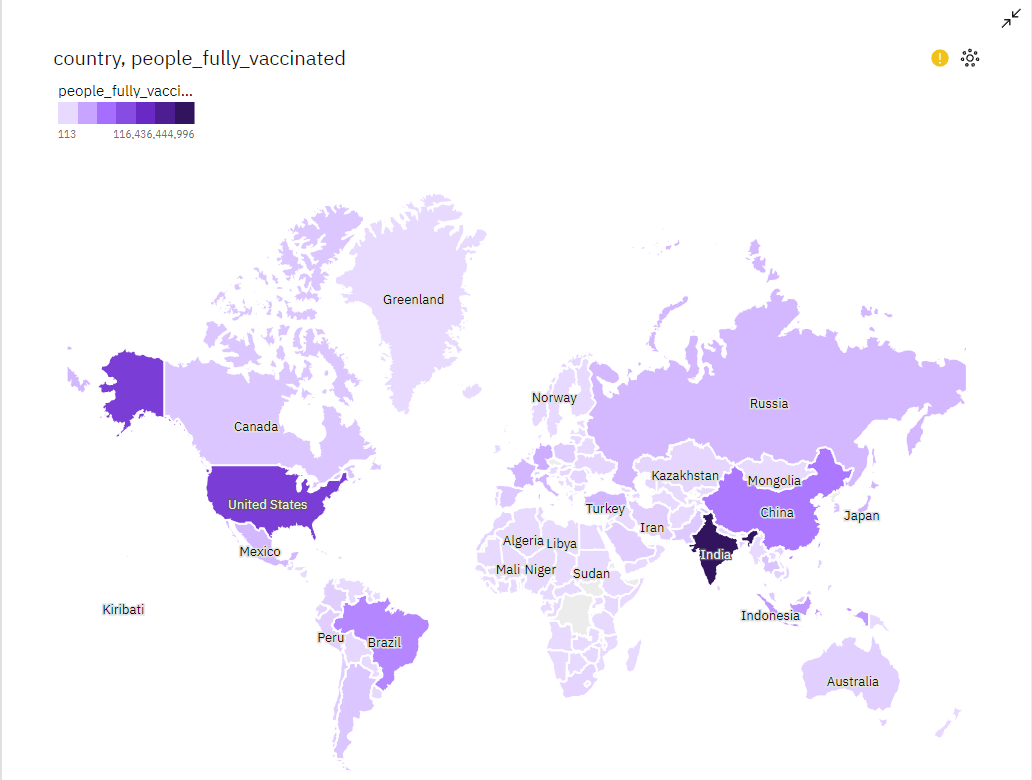
**c.Total vaccinations by vaccines**

****

**d.Total Vaccinations by country:**

****

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

****

**7. Recommendations:**

- **Policy Implications:** Provide actionable recommendations for public health policies, vaccination strategies, social distancing measures, and economic support.

- **Risk Assessment:** Evaluate the risks associated with different scenarios and interventions.

**8. Report Compilation:**

- **Software:** Use document preparation tools like LaTeX, Microsoft Word, Google Docs, or Markdown to compile the report.

- **Structured Format:** Organize the report with clear sections: executive summary, introduction, methodology, findings, discussion, recommendations, and conclusion.

**9.** **Peer Review:**

- **Expert Review:** Have experts in relevant fields review the report for accuracy, methodology, and potential biases.

- **Feedback Incorporation:** Incorporate feedback from the peer review process to improve the report's quality and objectivity.

**10. Publication:**

- **Distribution:** Share the report with relevant stakeholders, government agencies, public health organizations, and the general public through official websites, press releases, and scientific journals.

Remember that the specific tools, techniques, and depth of analysis can vary based on the scope and objectives of the report, as well as the available resources and expertise of the analysts. It's crucial to maintain transparency in your methods and sources to ensure the credibility of the analysis.

**Conclusion:**

our COVID-19 vaccination data analytics project has revealed critical insights. Targeted campaigns and socioeconomic factors significantly influence vaccination rates. To combat the pandemic effectively, we must continue using data-driven strategies and adapt to changing conditions. This analysis underscores the importance of data-driven decision-making in public health and paves the way for more informed vaccination strategies in the future.