**IBM – NAAN MUDHALVAN ➔ APPLIED DATA SCIENCE**

**PROJECT:COVID – 19 Vaccinations Analysis**

**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.
* The disease showed symptoms as basic as mild fever and cold but also caused life threatening symptoms like breathing problems caused by damage to the lungs.
* 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.

**Problem Statements:**

* In this project we have analyzed the top 10 fully vaccinated countries.
* We have analyzed the top 5 vaccinated countries.
* We have analyzed the top 5 daily vaccinating countries.
* We have analyzed the total number of daily vaccinations, people who have fully vaccinated, people who is vaccinated.
* We have analyzed year wise daily vaccinating details, fully vaccinated people details, vaccinated people details.
* We have analyzed the country wise vaccines and iso\_code details.

**Methodology:**

There are three steps involved in data science analysis they are

* Data Importing.
* Data Cleaning.
* Analysis.

**Step 1:**

**Data Importing:**

In power BI desktop with the help of the get data option import the CSV data which is named as country vaccinations and clicked load option.

**Step 2:**

**Data Cleaning:**

After loading the data and after analyzing the data understood that there are 86512 rows and 15 columns. And in that some of the columns contained null values I have replaced the null values by 0 with the use of replace functions and started working on the data.

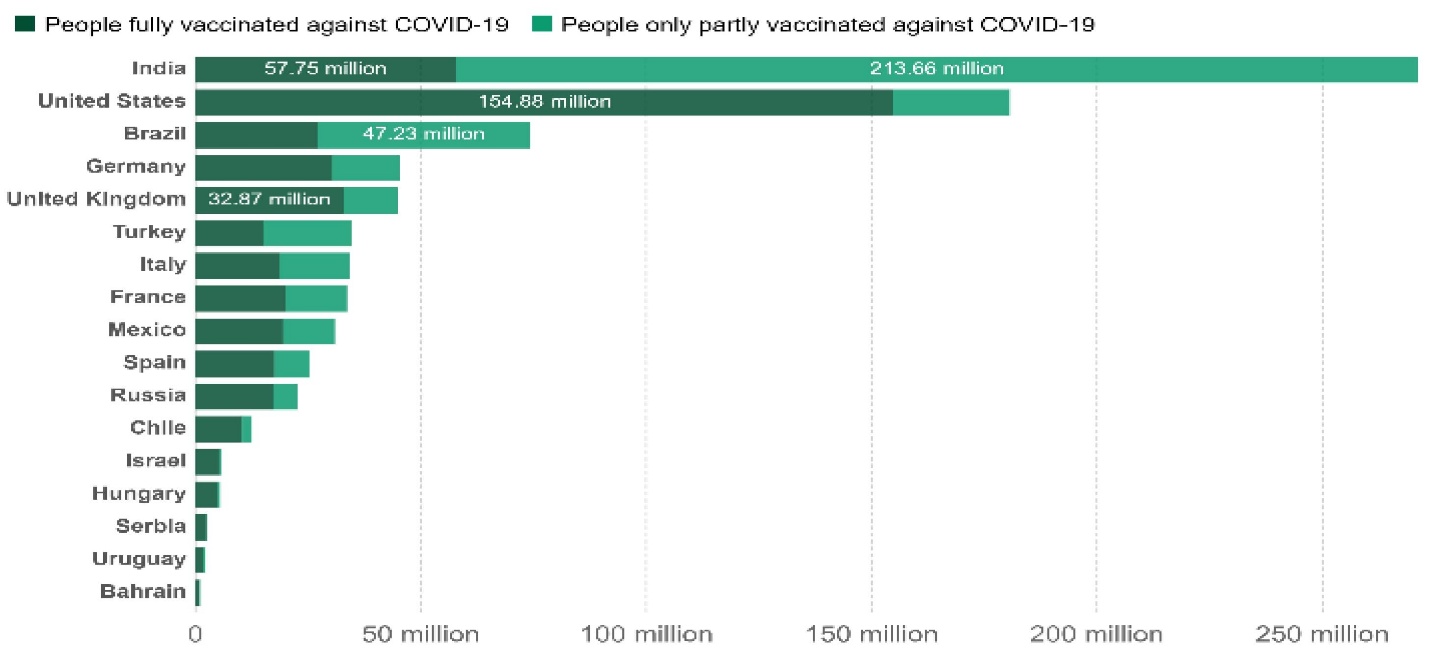
**Step 3:**

**Visualizations:**

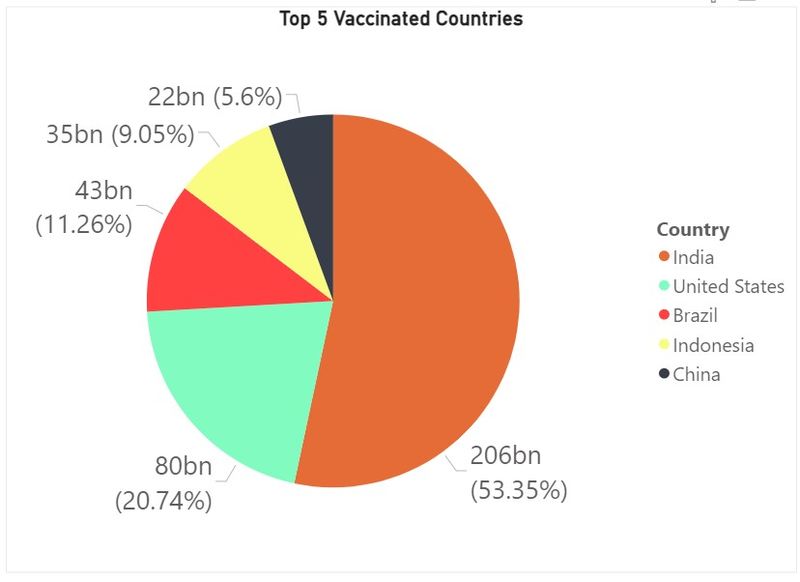
In visualization part with the help of power BI desktop software I have used different kinds of charts, graphs, cards and table to display the data in the format which will be easy to understand.

**Analysis:**

* In the analysis part first | have analyzed the top 10 fully vaccinated countries by using area chart and have used the filter option to find the top countries and the result obtained as below,
* From the below image we can able to come to know that India is the top country in terms full vaccination with 116 billon , followed by united states of America and china with 67 billion and 35 billion respectively.



* In the second analysis we have analyzed the top 5 vaccinated countries with the help of pie chart and used filter option to find the top countries and with that we came to know that India is the top country with more number of vaccinated peoples followed by United States of America and Brazil.



**Recommendations:**

* We should collect day to day reports and we should update our records daily to get more accurate details.
* So that we can move forward with more vaccination to the right country which needs the most.

**INTRODUCTION:**

Nowadays, machine learning (ML) used in every area of

computational work where algorithms are designed, and performance is increased .In the last years, learning from unbalanced data sets has become a critical problem in machine learning and is frequently found in several applications such as computer security.

Machine learning, Like some other technologies, played a crucial role in Determining the virus’s triggers and conditions . It was an Attempt to clean up the noisy data that had scattered across The world in order to educate biological areas where research Was attempting to understand how the virus resides beyond The human body and the effects of various factors such as Climate, population, and on COVID-19 spread.

Clustering is often a valuable function for Learning data. Uncontrolled clustering is known as the Segmentation of data into clusters that contain the same data, Mainly to make homogeneous groups.

**METHODOLOGY:**

Here we use two machine learning techniques they are;

* Clustering.
* Time series forecasting.

**Clustering Algorithms:**

A clustering algorithm divides a data set into many classes, With the similarity inside each group being greater than the Similarity within groups.

Clustering algorithms are commonly used for data Structure and categorization, as well as data compression and Model creation.

Now,we going to use two cluster they are;

* K-means clustering algorithm.
* Hierarchical clustering algorithm.

**K-means clustering algorithm:**

k-means algorithm is a clear partition process. It separated (N) data objects into (K) cluster sets to obtain low cross similarity and high intracluster specificity.

**Steps:**

**Step 1:** Get started: As initial centers, select k data items at

random from Data collection D. K is the number of clusters.

**Step 2:** Repetition:

a) Assume that each cluster is a centroid.

b) Evaluate the difference between all of the data points, as

well as the centroids.

c) Allocate di to the cluster that is nearest to you as a data

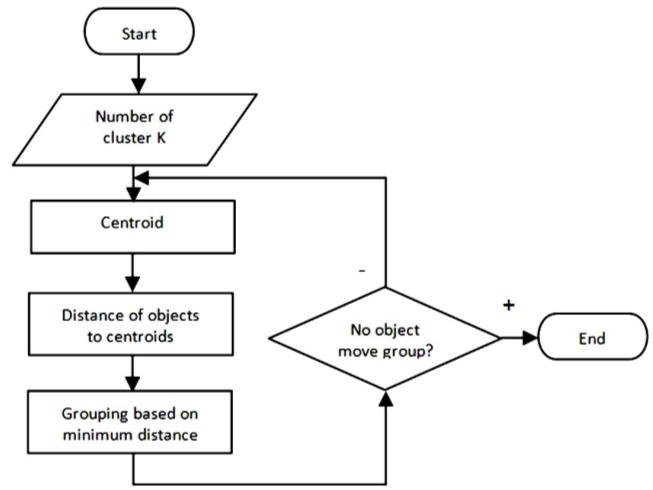
object.

**Step 3**: By each cluster j (1=j=k), create an update. Calculate

the cluster center once more.

**Step 4:** Repeat until there is no difference in the cluster's core.

**Step 5:** Finish.



K-means Clustering Process.

**Hierarchical Clustering Algorithm:**

By creating a hierarchy of clusters, also known as a dendrogram, the hierarchical Clustering approach combined or separates identical data items . The hierarchical Clustering approach creates clusters in a step-by-step manner.

Agglomerative and Divisive algorithms are the two Kinds of hierarchical clustering algorithms. Agglomerative Hierarchical clustering algorithm: Agglomerative hierarchical Clustering is a bottom-up approach that starts from each Person within a cluster.

**Steps:**

**Step 1**: Get started: Assign a cluster number equivalent to the Number of objects.

**Step 2**: Repetition: When the number of clusters is set to 1 or

when the consumer specifies the number of clusters.

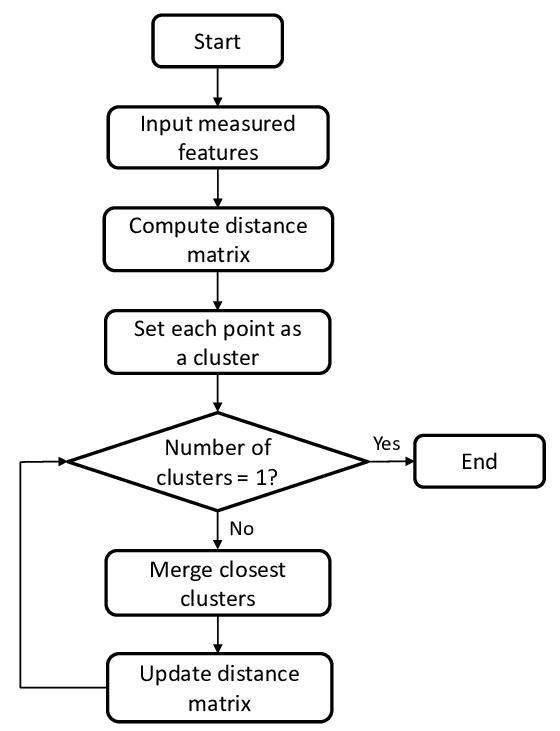
a) Calculate the shortest inters cluster size.

b) Combine the clusters with the smallest inter-cluster

distance.

**Step 3**: Finish.

Divisive hierarchical clustering algorithm: This is a hierarchical clustering algorithm, as opposed to a bottom-up approach.



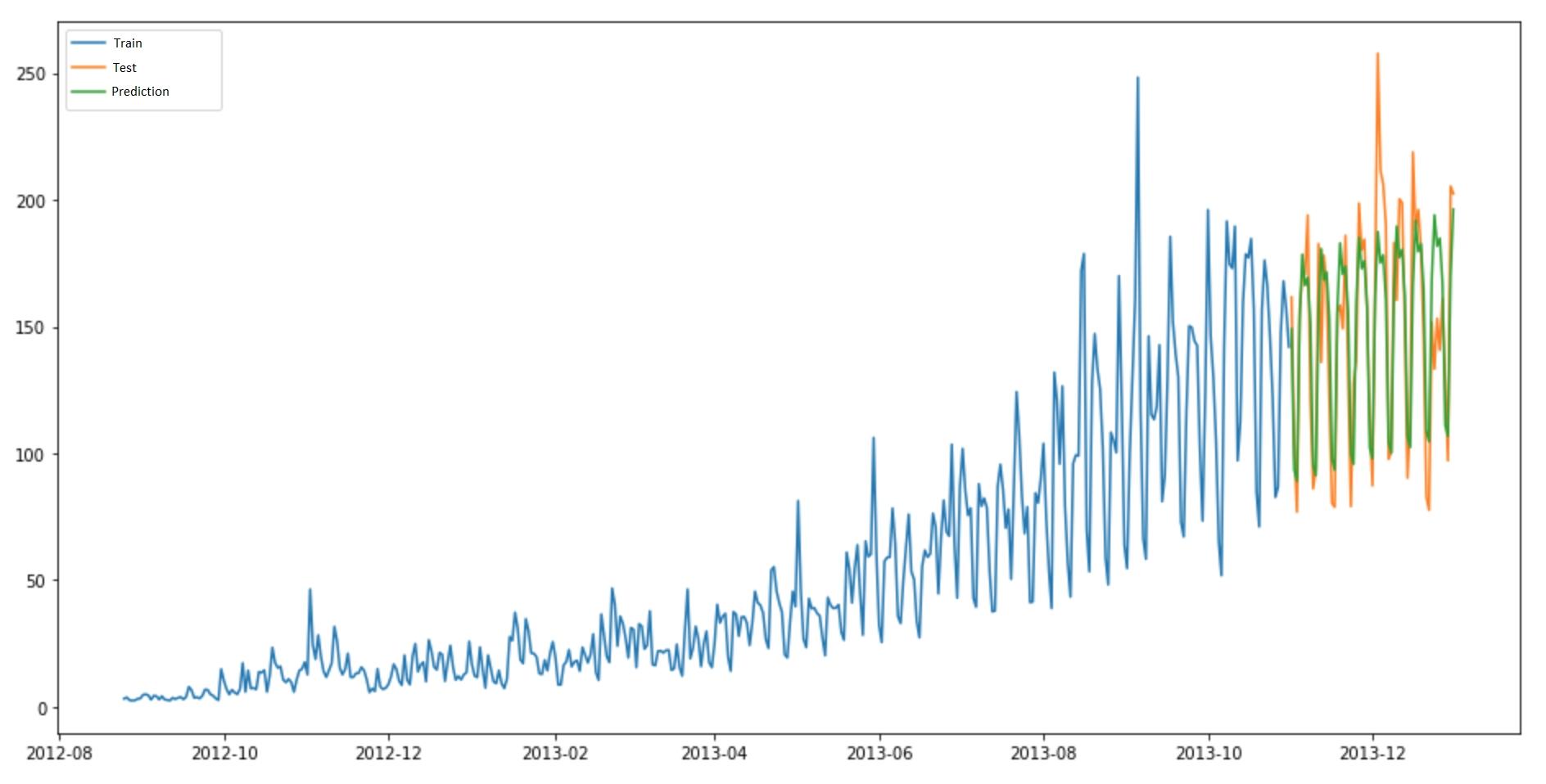
Flowchart Agglomerative Hierarchical Clustering.

**Time series forecasting:**

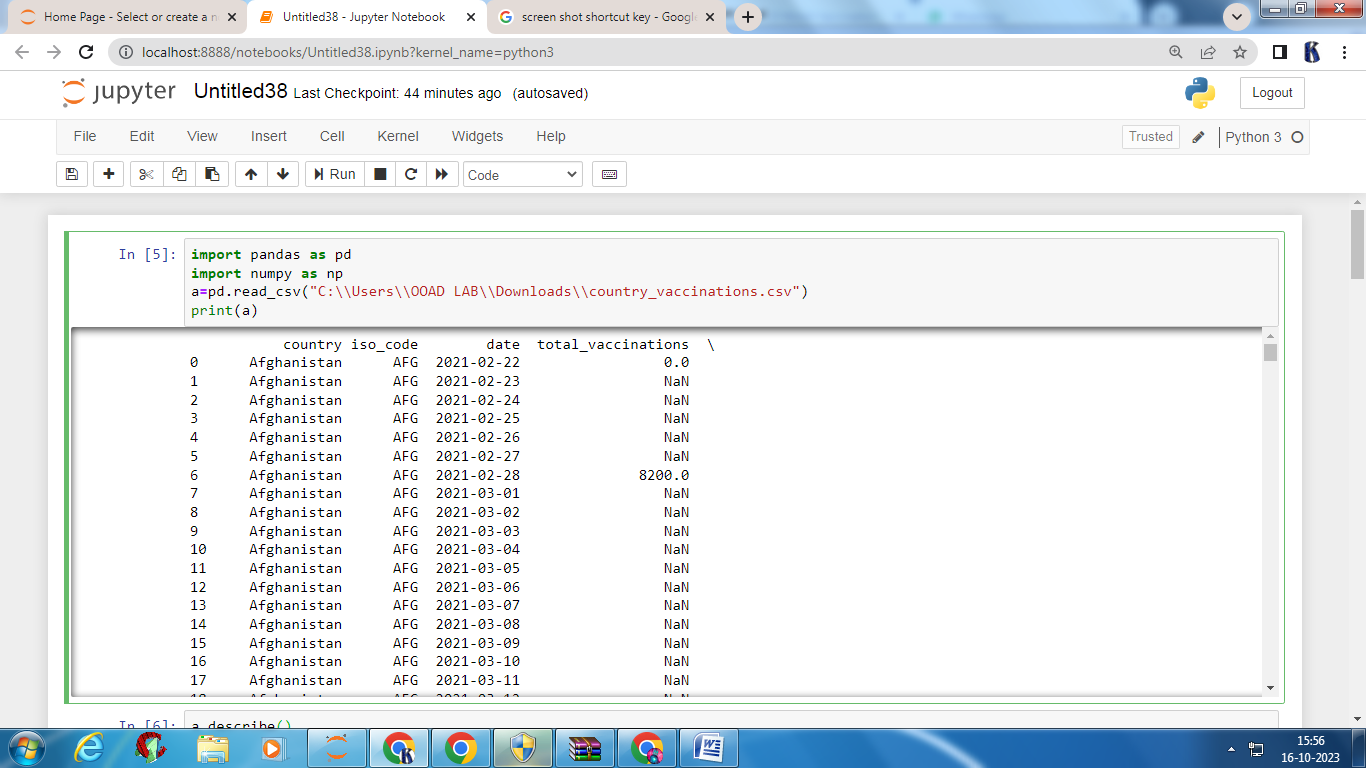
Time series forecasting is a technique used to make predictions about future data points in a time-ordered sequence. It’s widely used in various fields, such as finance, economics, and weather forecasting.

**Machine Learning:**

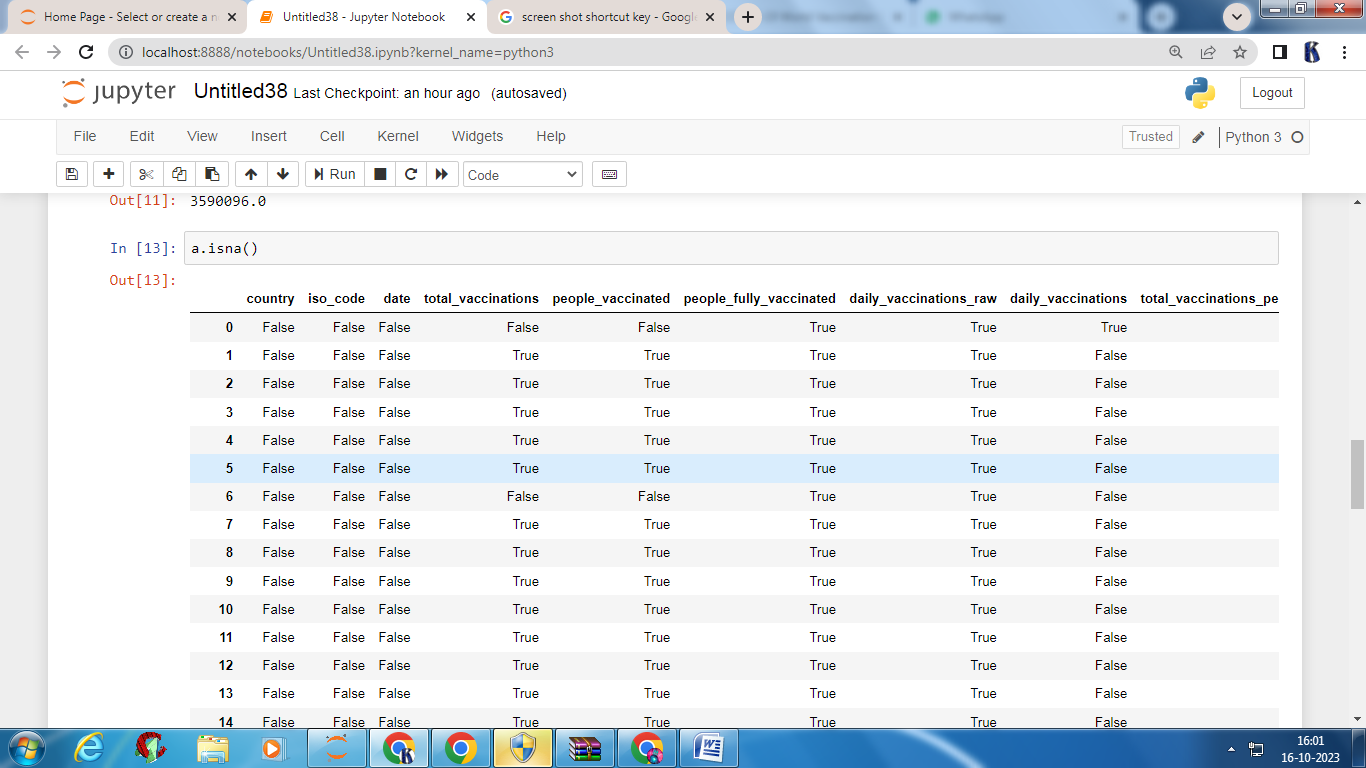
Methods like linear regression, decision trees, and neural networks can be used for time series forecasting. LSTM (Long Short-Term Memory) networks are particularly effective for sequential data.



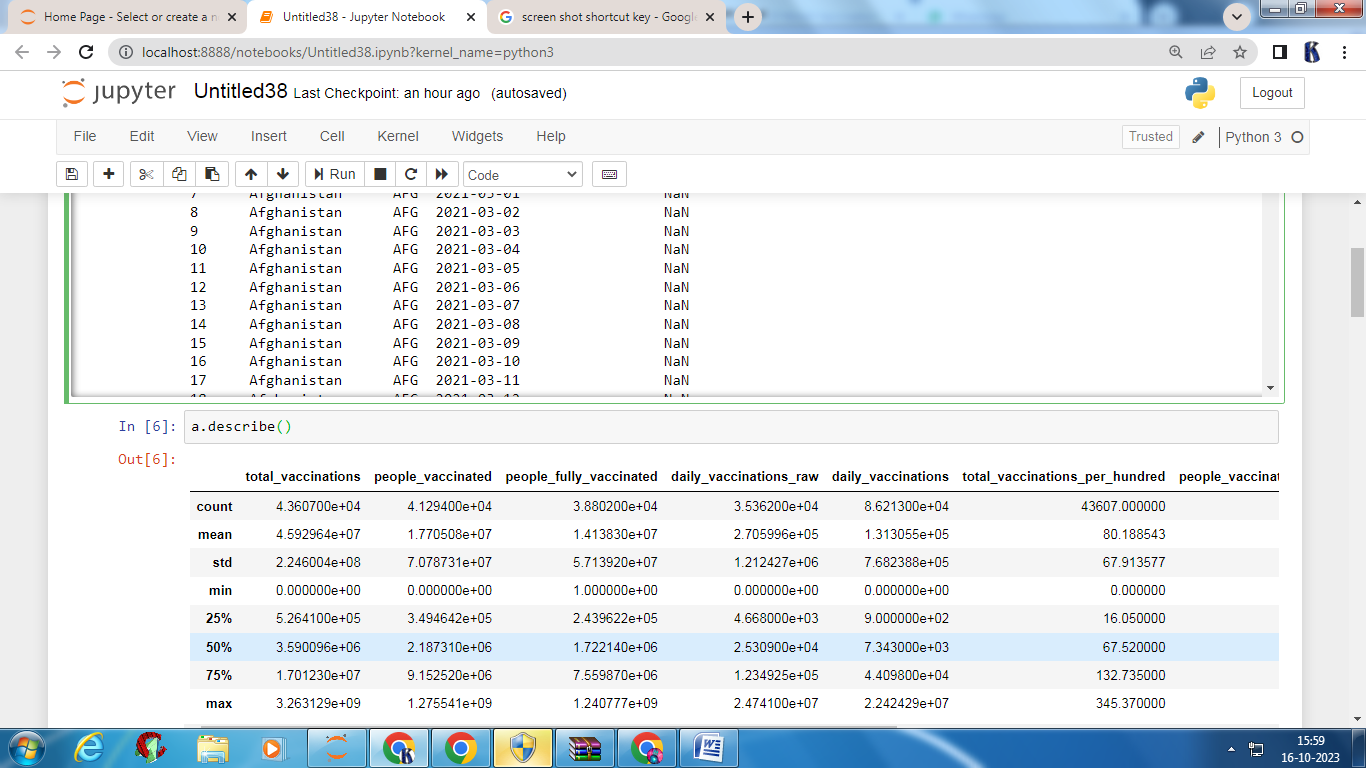
**Import CSV File:**

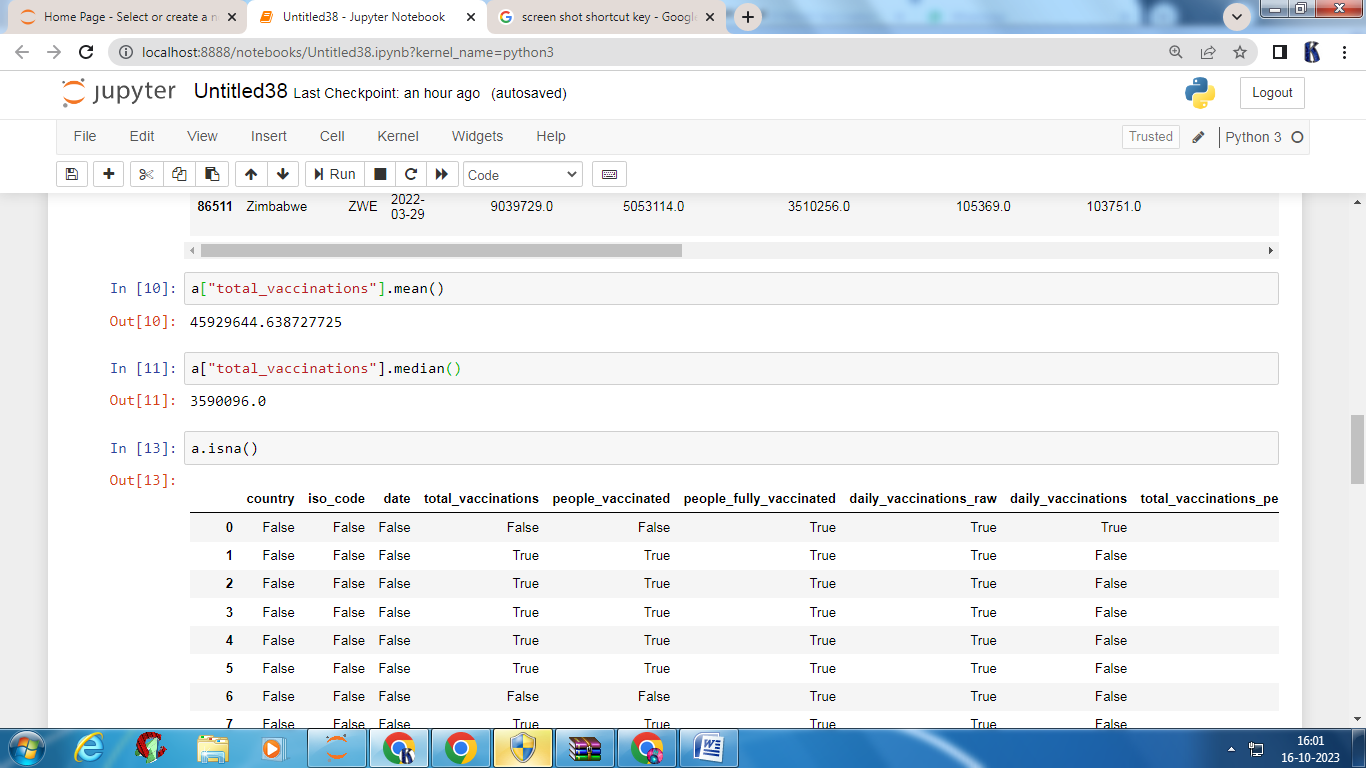


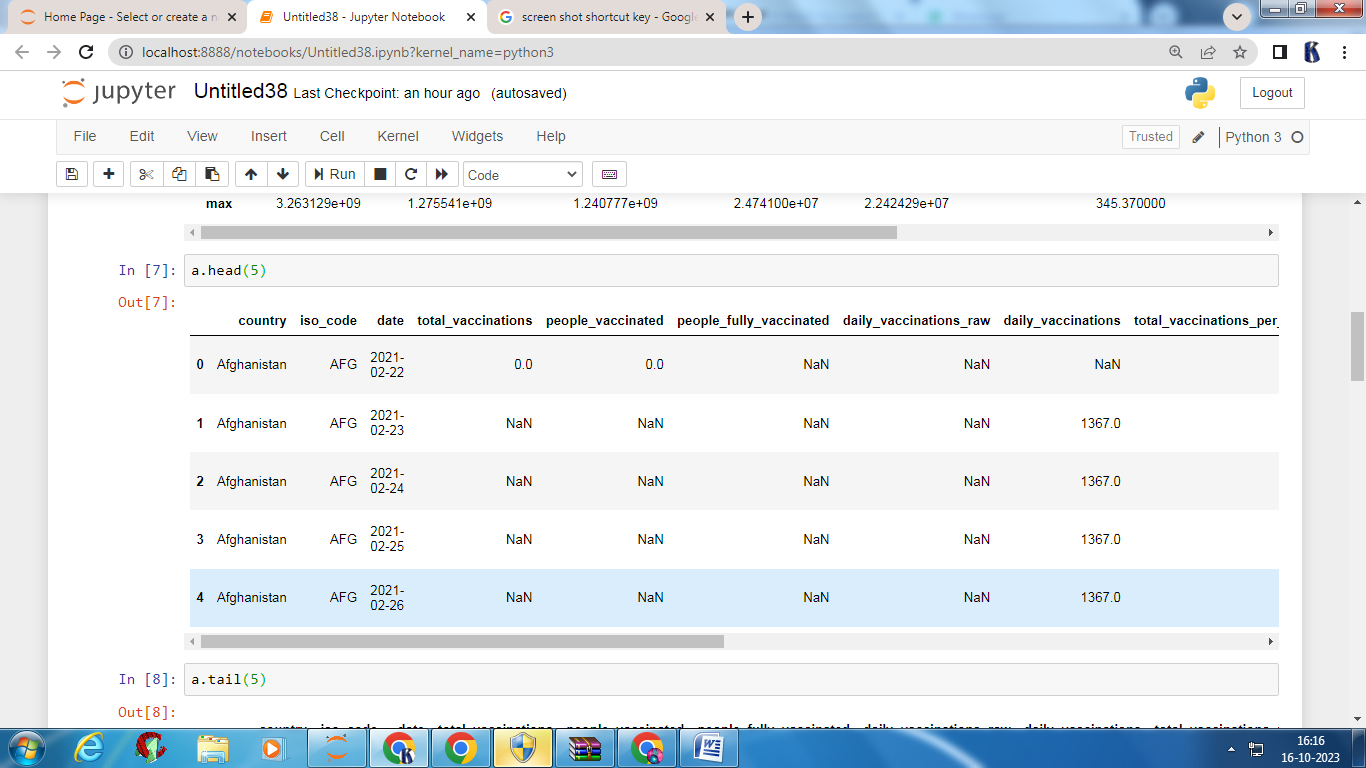
**Preprocessing:**

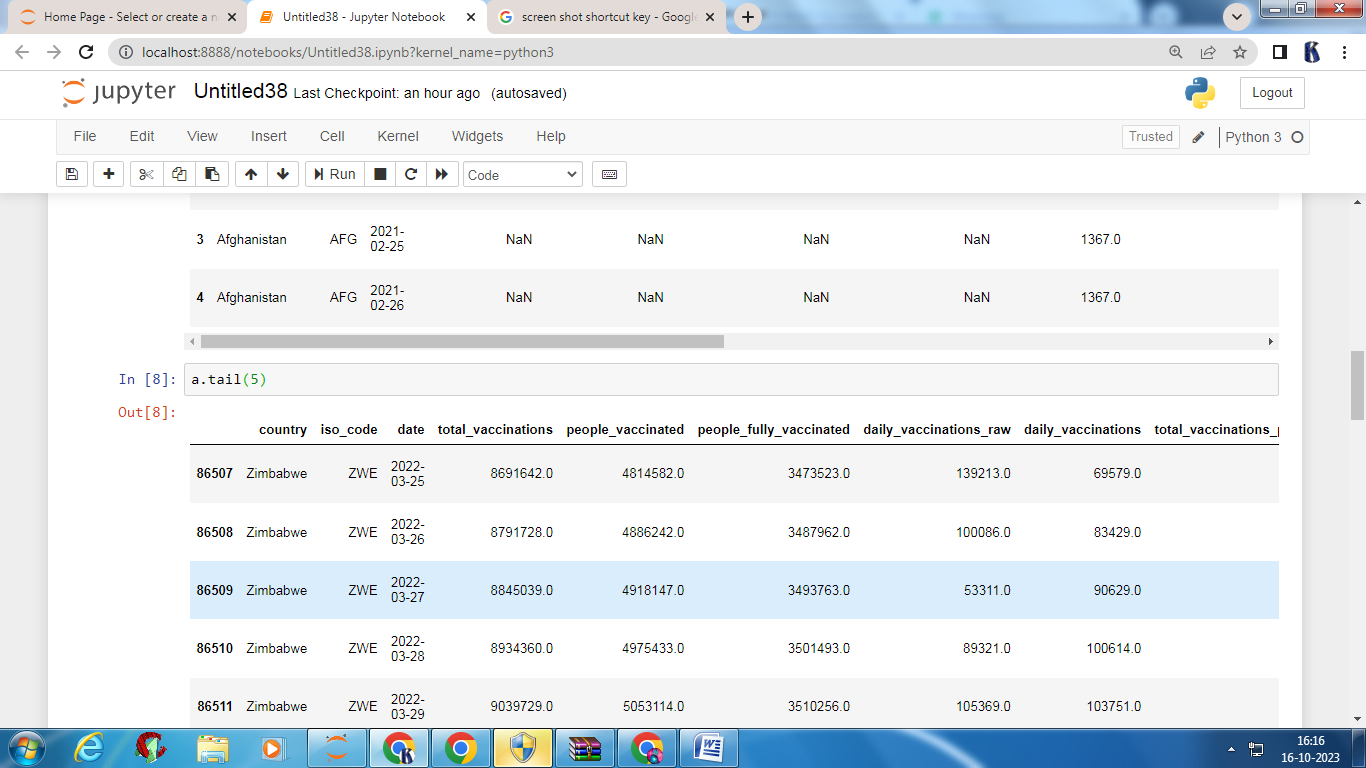


**Preforming various operation:**

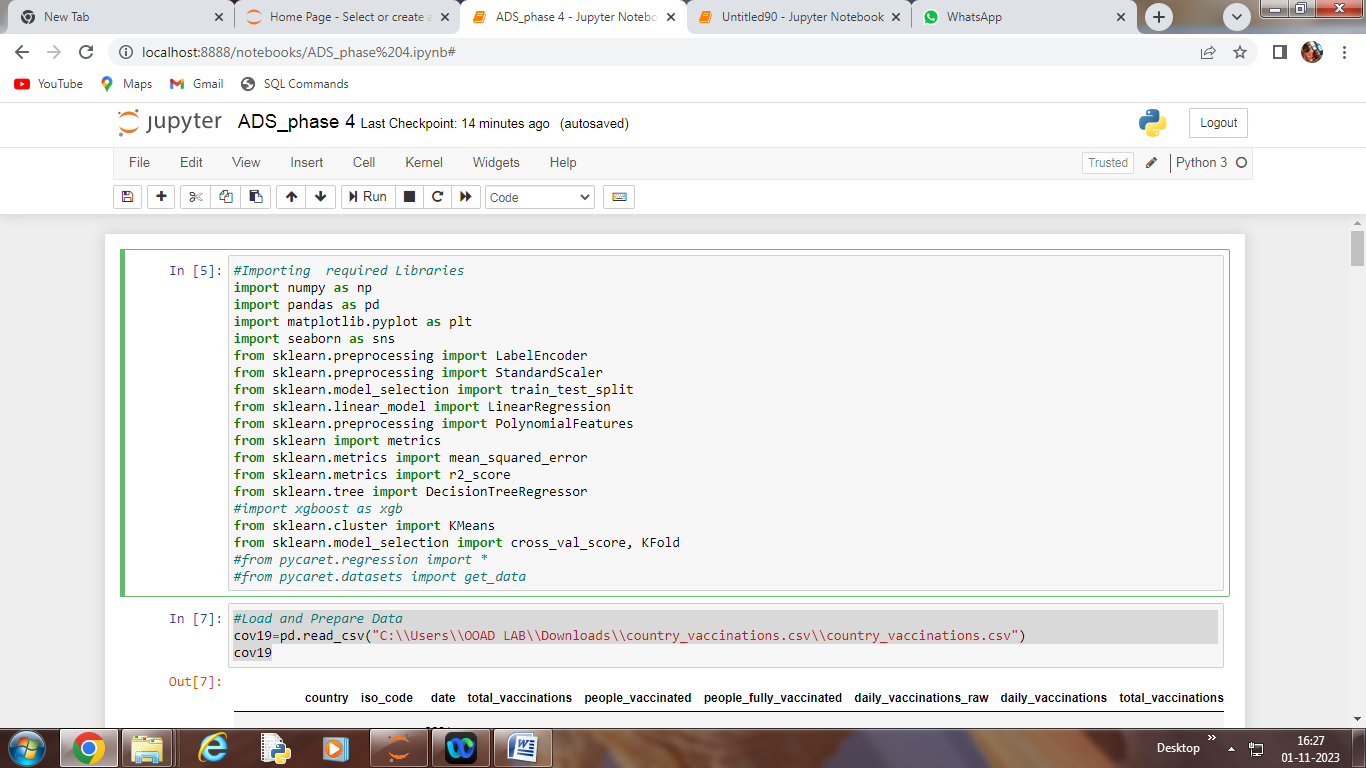




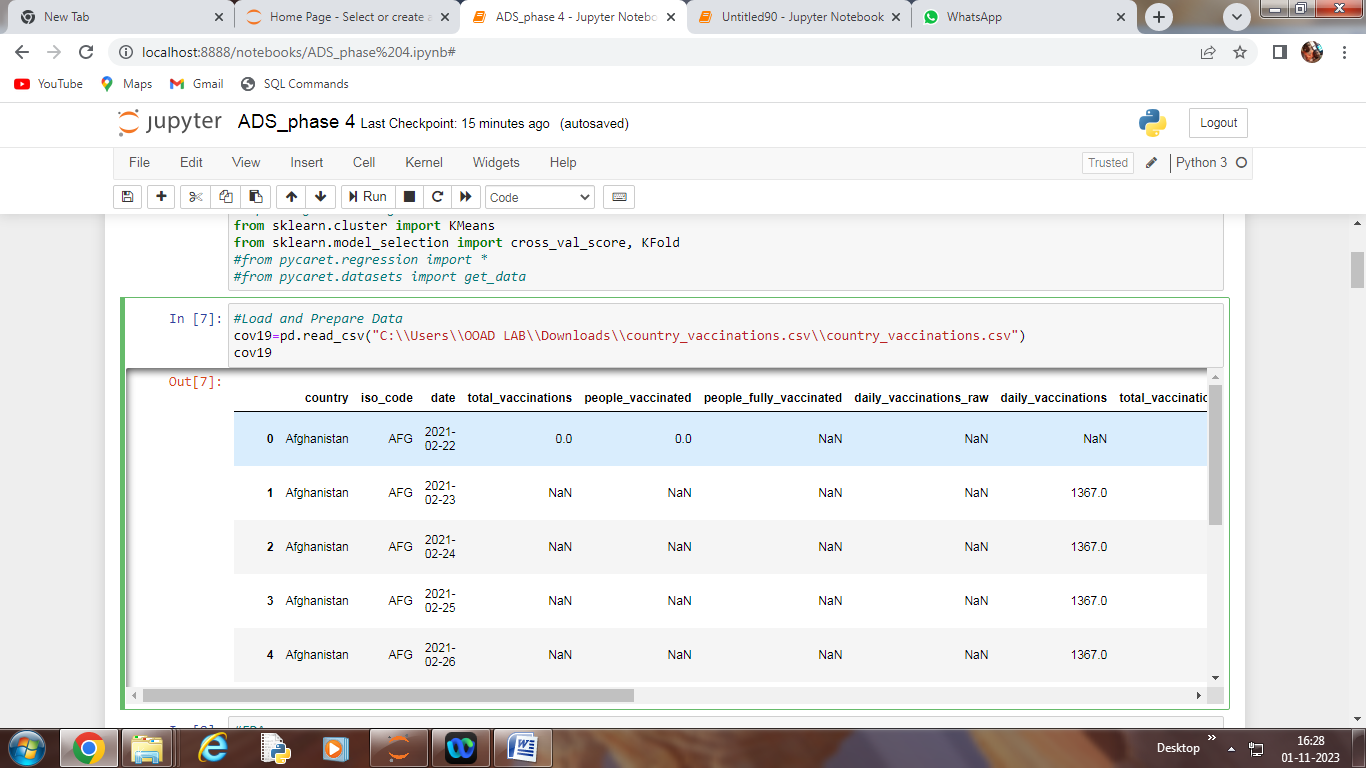




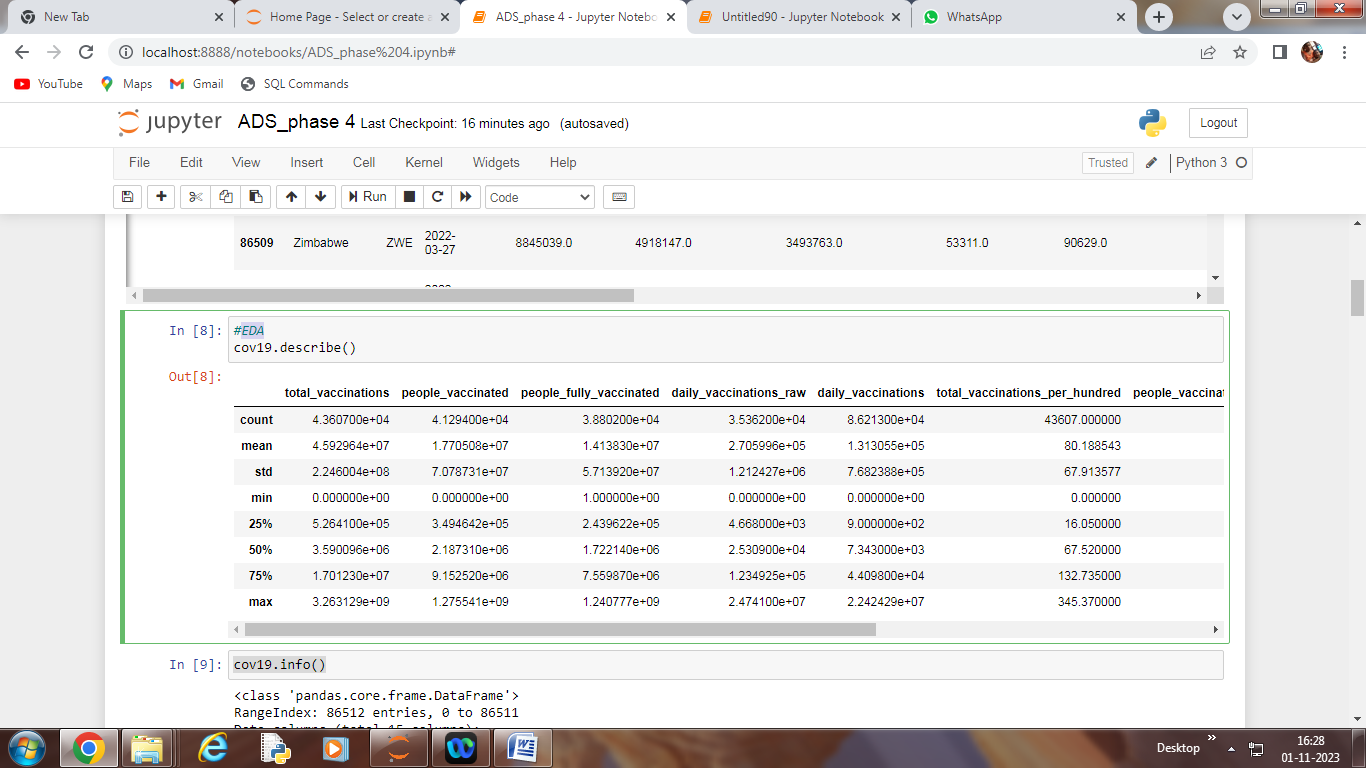
**Importing Required Libraries:**

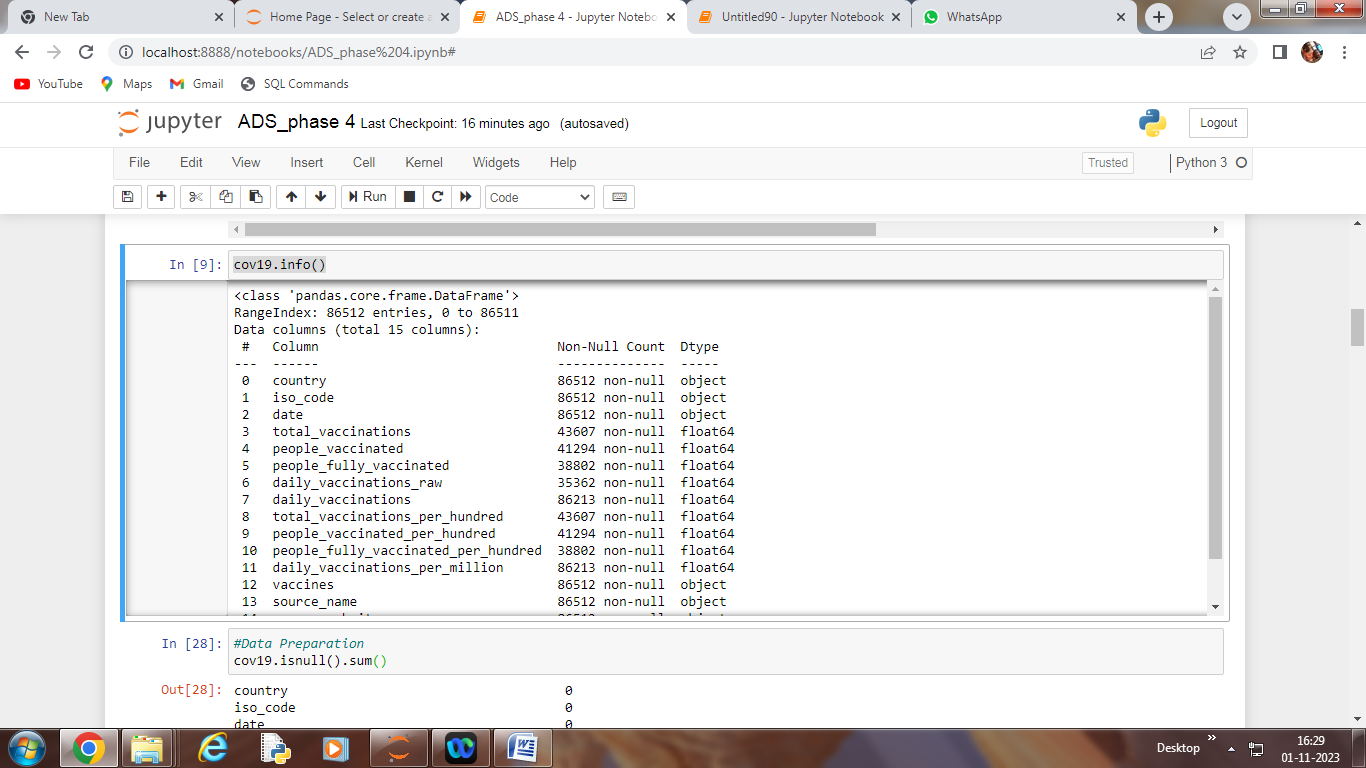


Load and Prepare Data:

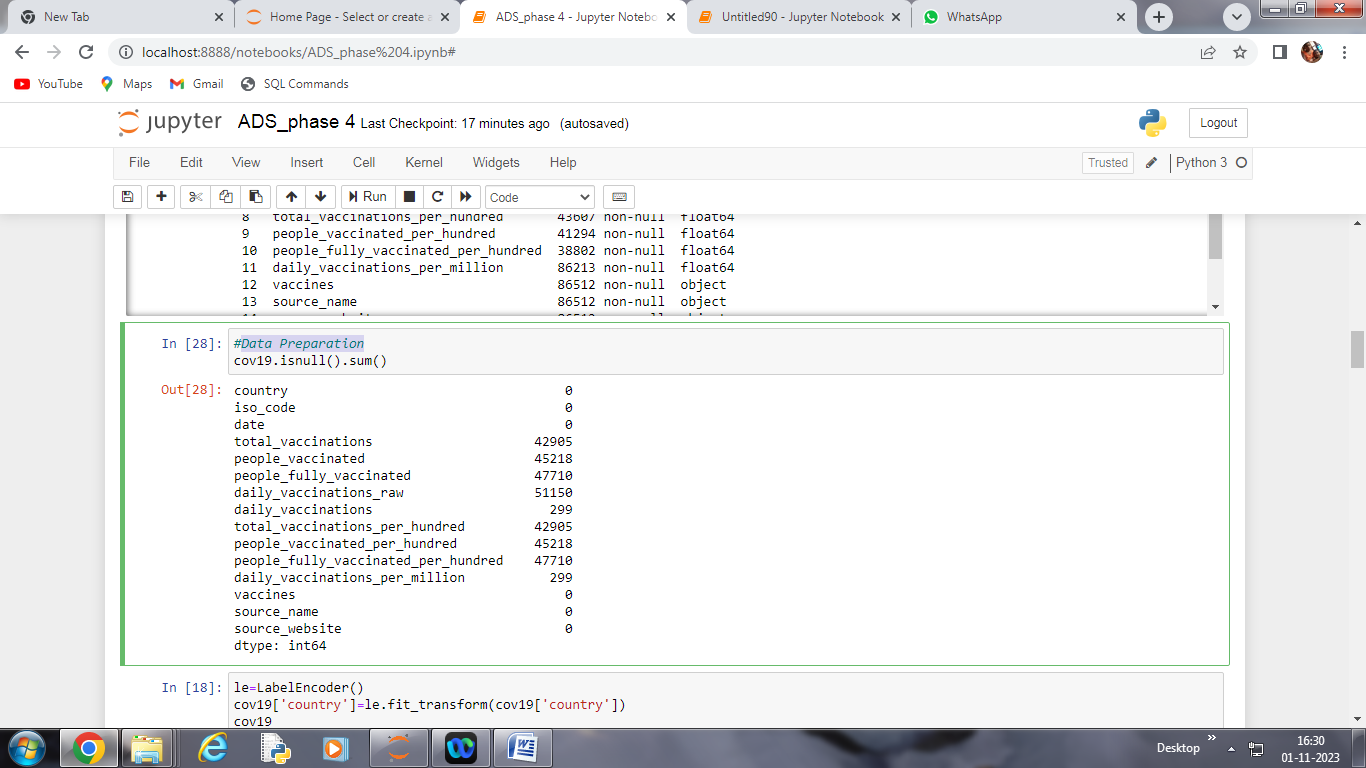


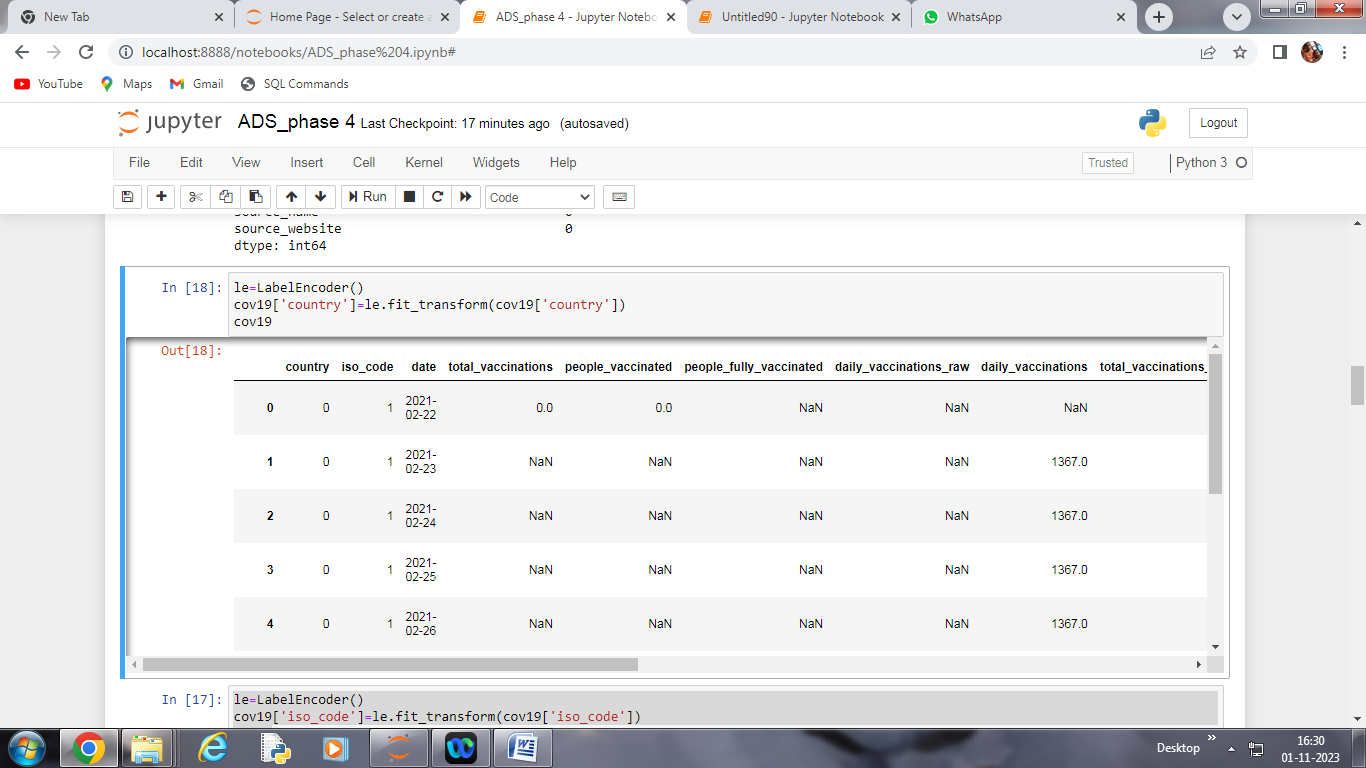
EDA:

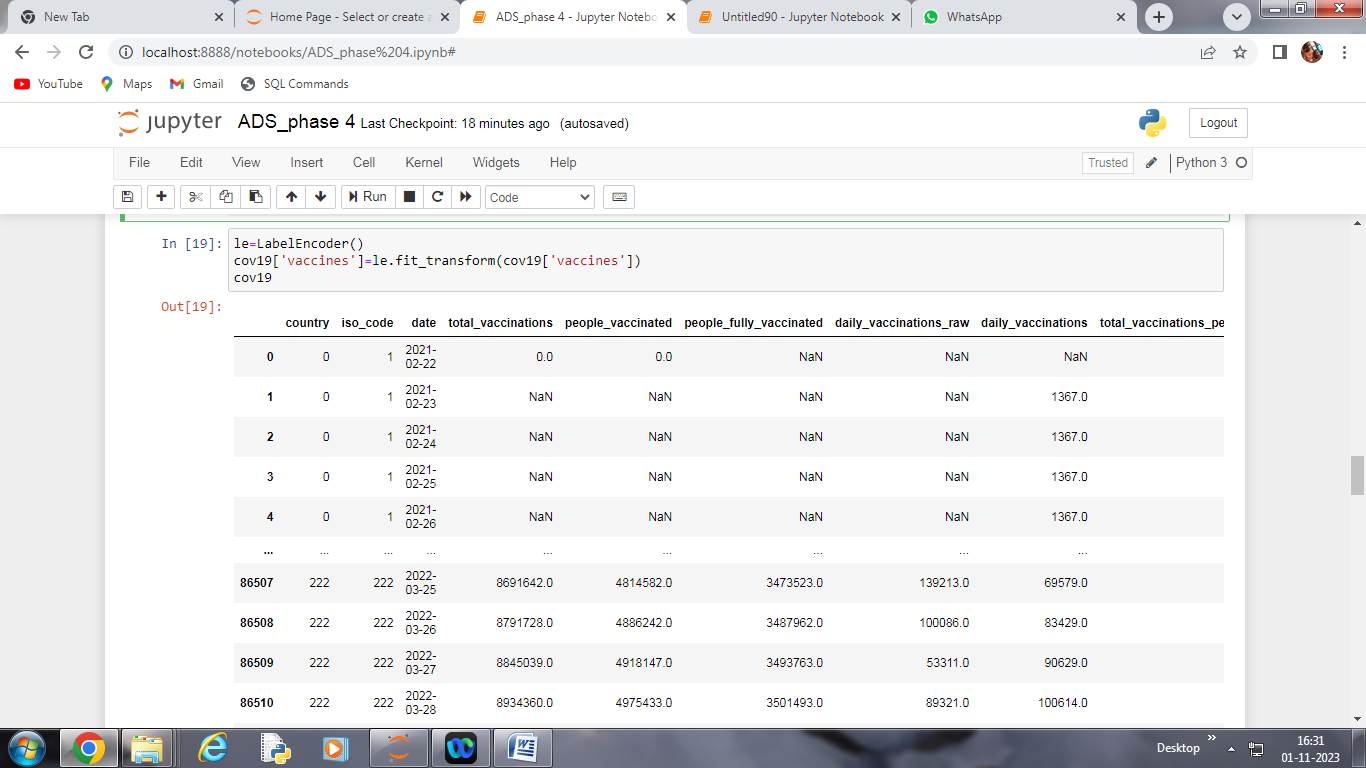




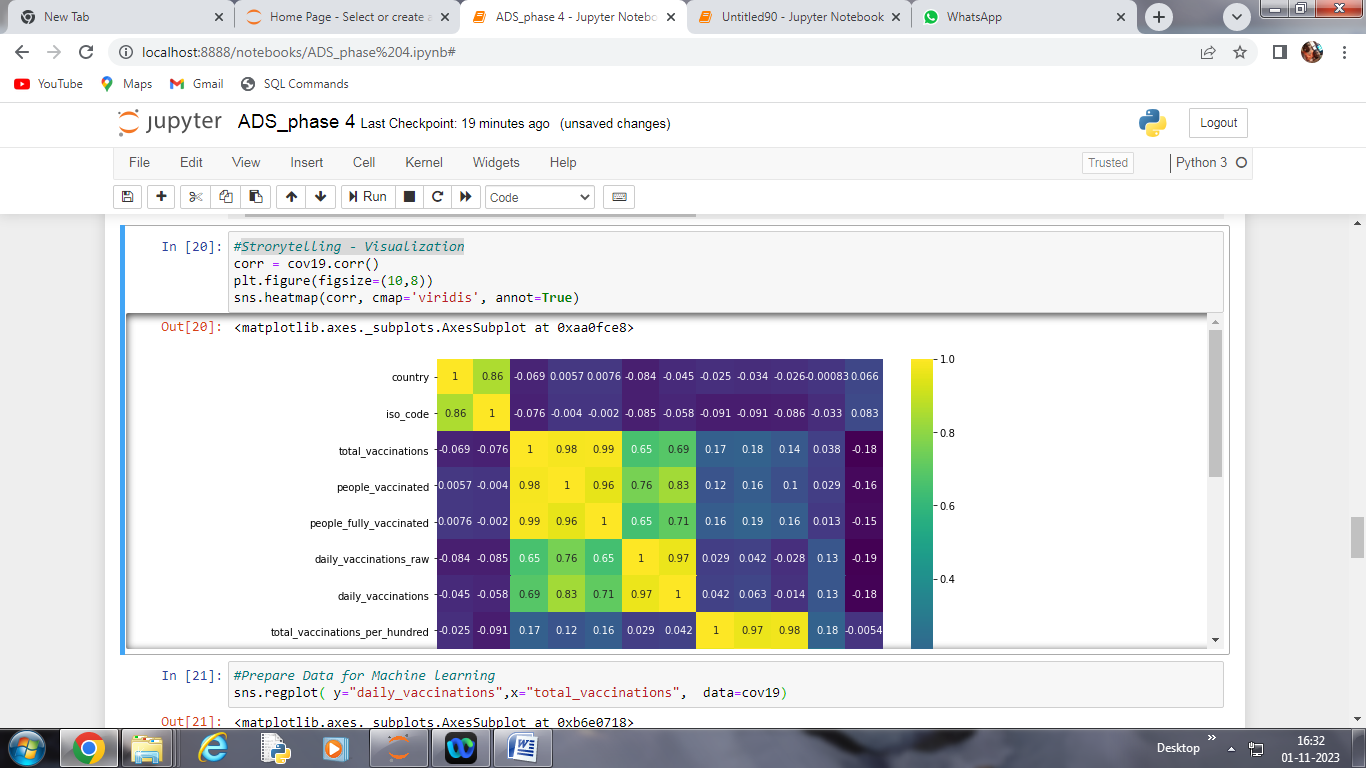
Data Preparation:



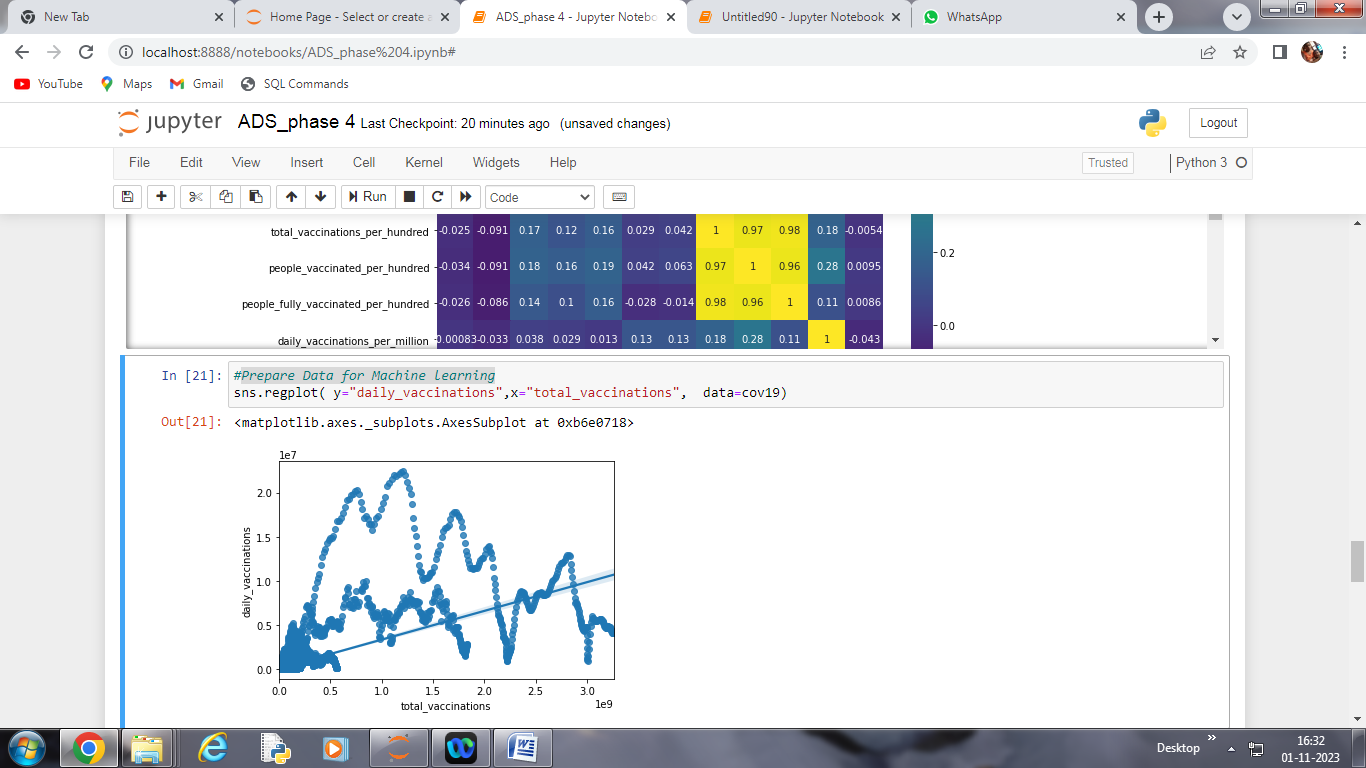


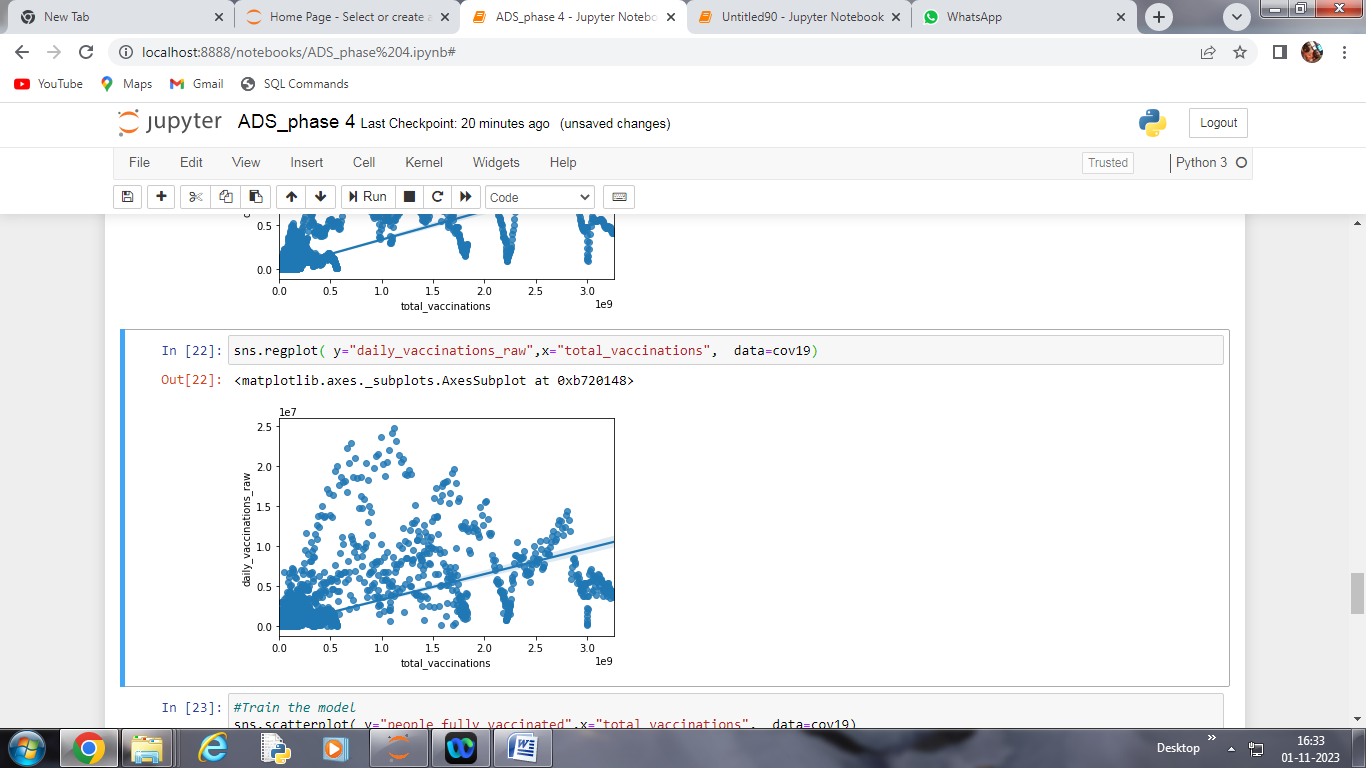


Strorytelling – Visualization:



Prepare Data for Machine learning:





Train the model:

