**Ganadipathy Tulsi's Jain Engineering College**

Kaniyambadi - Vellore - 632102

**IBM - PROJECT**

**COVID-19 VACCINE ANALYSIS**

### Submitted By

Team Leader - Mageshwar V

Team Member 1 - Mugesh B

Team Member 2 - Lingeshwaran D

Team Member 3 - Chandru P

# Project Mentor Project Evaluator Project Spoc

## Mrs.K.Durga Priyanka Mrs.K.Durga priyanka Mr.D.Durai Kumar

**INDEX**

1. **INTRODUCTION**

1.1 Project Overview

1.2 Purpose

2. **PROPOSED SOLUTION**

2.1 Problem Statement Definition

3. **REQUIREMENT ANALYSIS**

3.1 Functional requirement

3.2 Non-Functional requirements

**4. PROJECT DESIGN**

4.1 Data Flow Diagrams

**5. CODING**

5.1 Import Dataset

5.2 Describe the Dataset

5.3 Data Preprocessing

5.4 Training a Model

5.5 Statistical Analysis

**6. USING ML ALOGRITHMS**

6.1 Linear Regression

6.2 Decision Tree

**7. ADVANTAGES & DISADVANTAGES**

**8. FUTURE SCOPE**

**9. CONCLUSION**

**10. APPENDIX**

10.1 Source Code

10.2 GitHub

**INTRODUCTION**

**1.1Project Overview**

This project aims to develop a machine learning model that can accurately identify and predict airline passengers' satisfaction based on various factors. By analyzing historical data and leveraging machine learning algorithms, we can provide valuable insights into the factors that contribute to passenger satisfaction and assist airlines in improving their services and customer experience.

**1.2 Purpose**

The purpose of a COVID-19 vaccine analysis project can vary depending on its specific goals and objectives, but in general, such a project is undertaken to gather, analyze, and interpret data and information related to COVID-19 vaccines. The main purposes of a COVID-19 vaccine analysis project include:

* Assessment of Vaccine Effectiveness
* Safety Monitoring
* Optimizing Vaccine Distribution
* Evaluating Vaccine Variants
* Public Health Policy and Decision-Making
* Herd Immunity Assessment
* Global Impact
* Regulatory Decision Support

**2 PROPOSED SOLUTION**

**2.1 Problem Statement:**

"The COVID-19 pandemic has had a profound global impact, and the rapid development and deployment of COVID-19 vaccines have become pivotal in managing and mitigating the spread of the virus.

However, there is a pressing need to comprehensively analyze various aspects of COVID-19 vaccines to ensure their efficacy, safety, equitable distribution, and public acceptance.

The primary problem is that, while vaccines are being distributed worldwide, there is a lack of comprehensive analysis in several critical areas, including

This problem statement outlines the critical areas where a comprehensive COVID-19 vaccine analysis is needed to support effective decision-making, optimize vaccination strategies, and ensure the pandemic's eventual control."

This problem statement serves as a foundation for structuring and executing a COVID-19 vaccine analysis project, with the goal of addressing these specific challenges and contributing to the global response to the pandemic.

**2.2 Design Thinking Process:**

Design thinking is a problem-solving approach that emphasizes empathy for the end-users and iteratively designing and testing solutions. Applying design thinking to a COVID-19 vaccine analysis project can help ensure that the analysis is user-centric, innovative, and effective. Here's how you can adapt the design thinking process to analyze COVID-19 vaccines

* Empathize: Understand Stakeholder Needs
* Define: Reframe the Problem
* Ideate: Generate Solutions

2.3 **Phases of Development:**

The development of a COVID-19 vaccine analysis project typically involves several phases, from project inception to completion. Here are the key phases volved in developing a COVID-19 vaccine analysis project:

**Project Initiation:**

Define the project's objectives, goals, and scope. Identify the stakeholders and establish a project team. Secure the necessary resources, including data sources and analytical tools**.**

**Data Collection and Preparation:**

Gather relevant data sources, such as vaccine coverage, efficacy, safety, and disease spread data. Clean and preprocess the data, addressing missing values, outliers, and ensuring data consistency.

**Feature Engineering:**

Identify and select the relevant features (variables) for analysis. This may include demographic information, vaccination rates, vaccine type, and geographic data.

**Data Analysis and Exploration:**

Perform initial exploratory data analysis to understand the data's distribution, patterns, and correlations. Create visualizations to gain insights from the data.

**Hypothesis Formulation:**

Develop hypotheses or research questions that the analysis aims to answer. These might include questions about vaccine efficacy, safety, and distribution.

**Model Selection and Design:**

Choose the appropriate analytical methods and models for the analysis. This may include statistical tests, machine learning algorithms, or epidemiological models.

**Data Analysis and Modeling:**

Apply the selected models to the data to analyze vaccine-related trends, outcomes, and patterns. Interpret the results and assess their statistical significance.

**Validation and Testing:**

Test the models and analysis methods to ensure their accuracy and reliability.

Conduct validation and cross-validation to assess the generalizability of findings.

**Results Interpretation:**

Interpret the results in the context of the research questions and hypotheses.

Draw conclusions based on the analysis outcomes. Recommendations and Actionable Insights:

**Document the entire analysis process, including data sources, methods, and results.**

**Prepare a comprehensive report or presentation for stakeholders, decision-makers, and the scientific community.**

**3.** **REQUIREMENT ANALYSIS**

**3.1 Functional requirement**

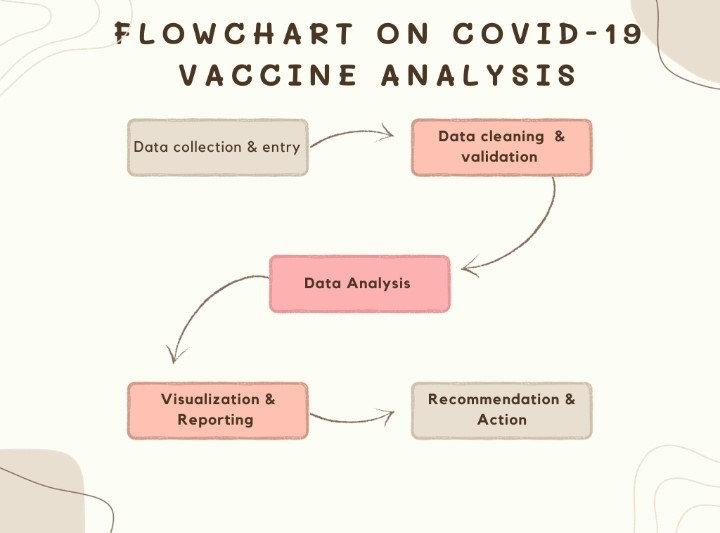
|  |  |  |
| --- | --- | --- |
| **FR No.** | **Functional Requirement (Epic)** | **Sub Requirement (Story / Sub-Task)** |
| FR-1 | Data Collection | Develop a data collection mechanism that can gather data from various sources and store it in a centralized data storage system. |
| FR-2 | Data Preprocessing | Develop a data preprocessing pipeline that can handle various data formats and transform them into a  consistent format for analysis. |
| FR-3 | Machine Learning Model Development | Research and choose the appropriate machine learning algorithms for predicting passenger satisfaction levels  and develop a model based on these algorithms. |
| FR-4 | Model Training and Evaluation | Develop a training and evaluation pipeline that can handle large amounts of data and provide accurate  model evaluation metrics. |
| FR-5 | Feedback Loop | Develop a feedback mechanism that can analyze the results of the machine learning model and provide airlines with insights and recommendations to improve  passenger satisfaction. |

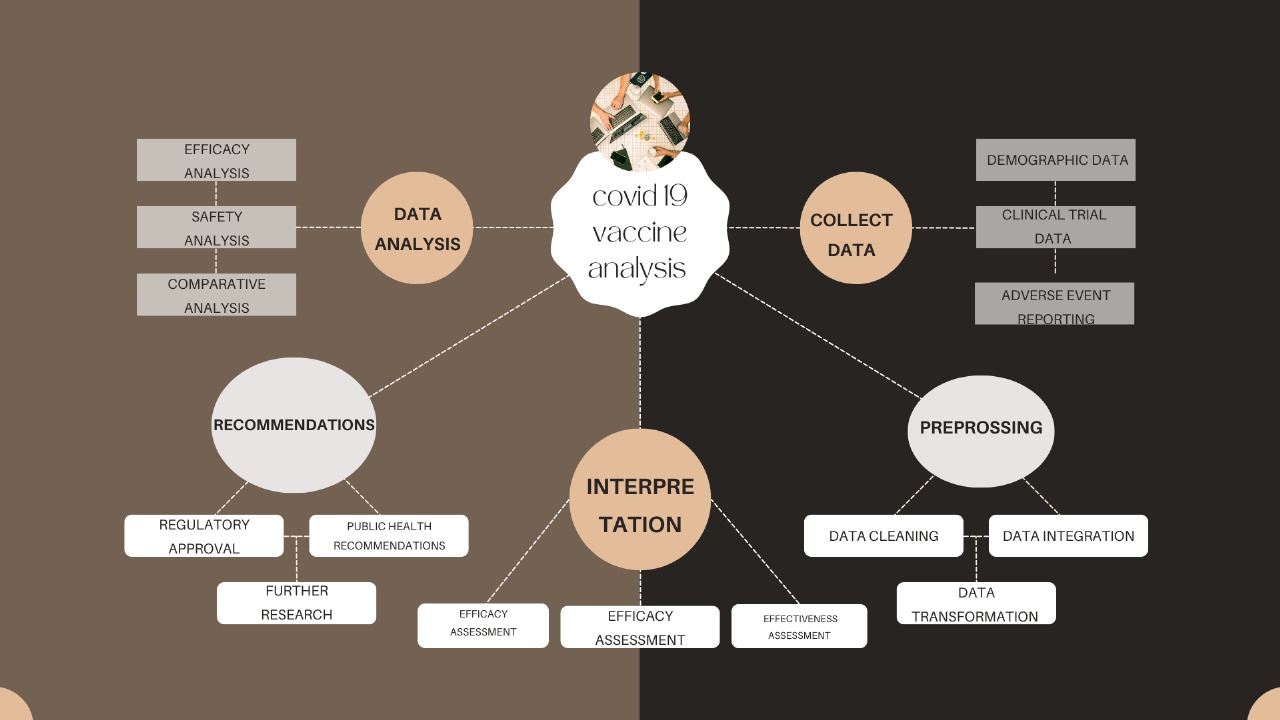
**3.2 Non – Functional requirements**

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Non-Functional Requirement** | **Description** |
| NFR-1 | **Usability** | The solution must be easy to use and accessible to  non-technical users. |
| NFR-2 | **Security** | The solution should be secure and protect the  privacy of passenger data. |
| NFR-3 | **Reliability** | The solution must be reliable and able to handle large amounts of data without crashing or losing  data. |
| NFR-4 | **Performance** | The solution should be able to handle large volumes  of data and provide real-time feedback to airlines. |
| NFR-5 | **Availability** | The solution must be available 24/7 to collect and  analyze data from various sources. |
| NFR-6 | **Scalability** | The solution should be able to scale up to handle increasing amounts of data and be easily adaptable  to different airlines' needs. |

**4. PROJECT DESIGN**

**4.1 Data Flow Diagrams**

****

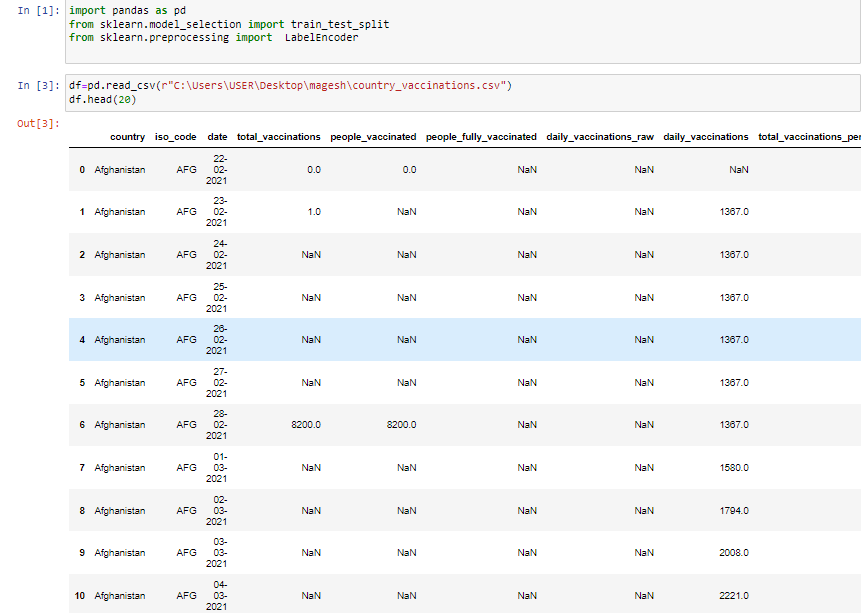


**5. CODING**

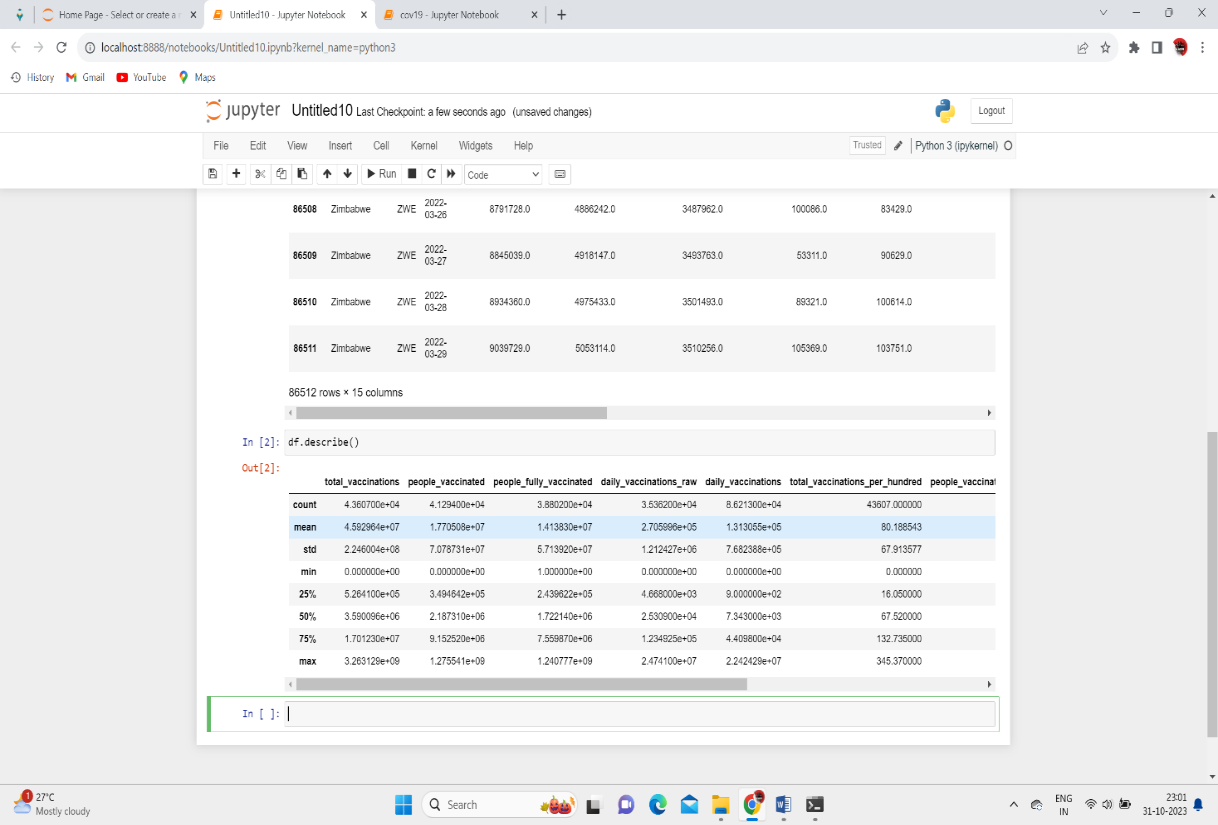
**5.1 Import Dataset:**

Download a Dataset from Kaggle:

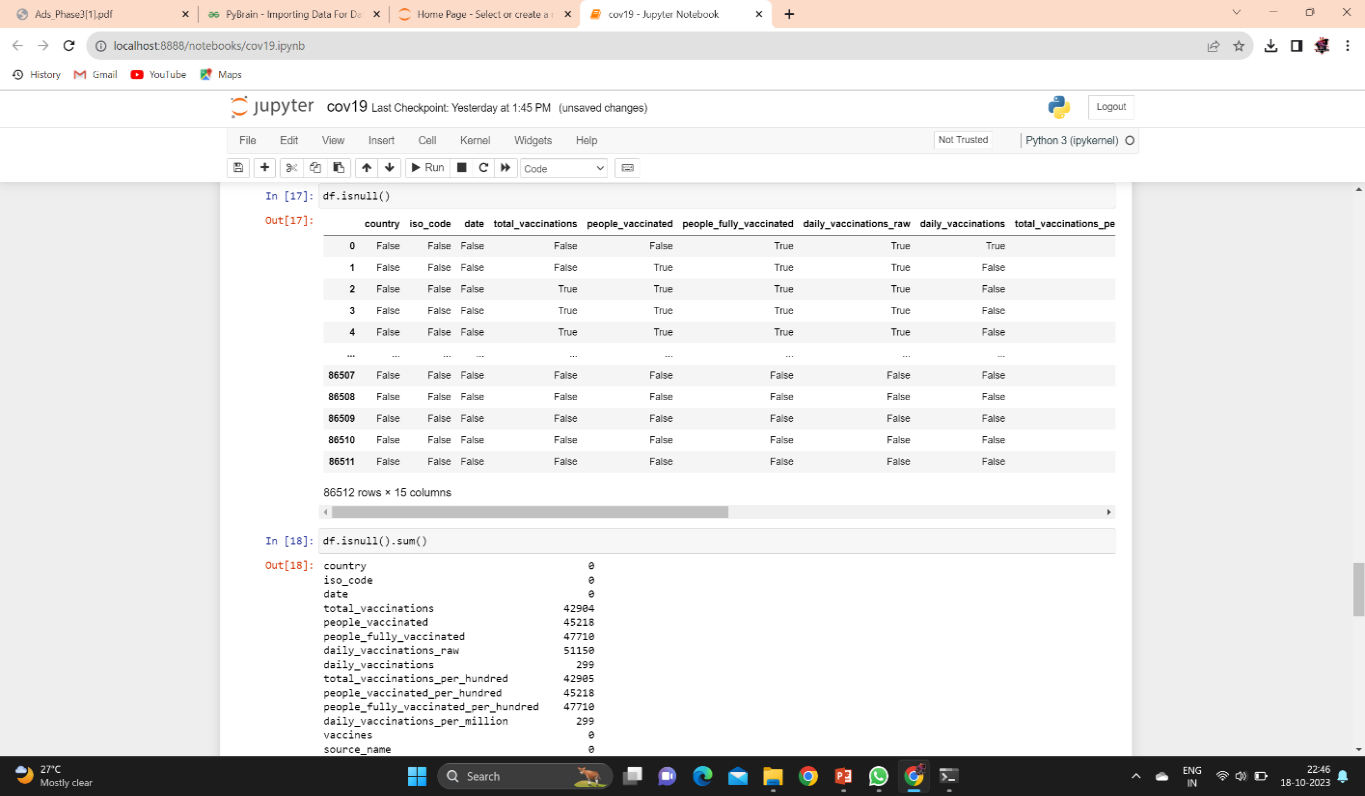
Kaggle is a data science platform that hosts numerous datasets, including COVID-19 vaccine data. You can search for datasets related to COVID-19 vaccination and analysis.



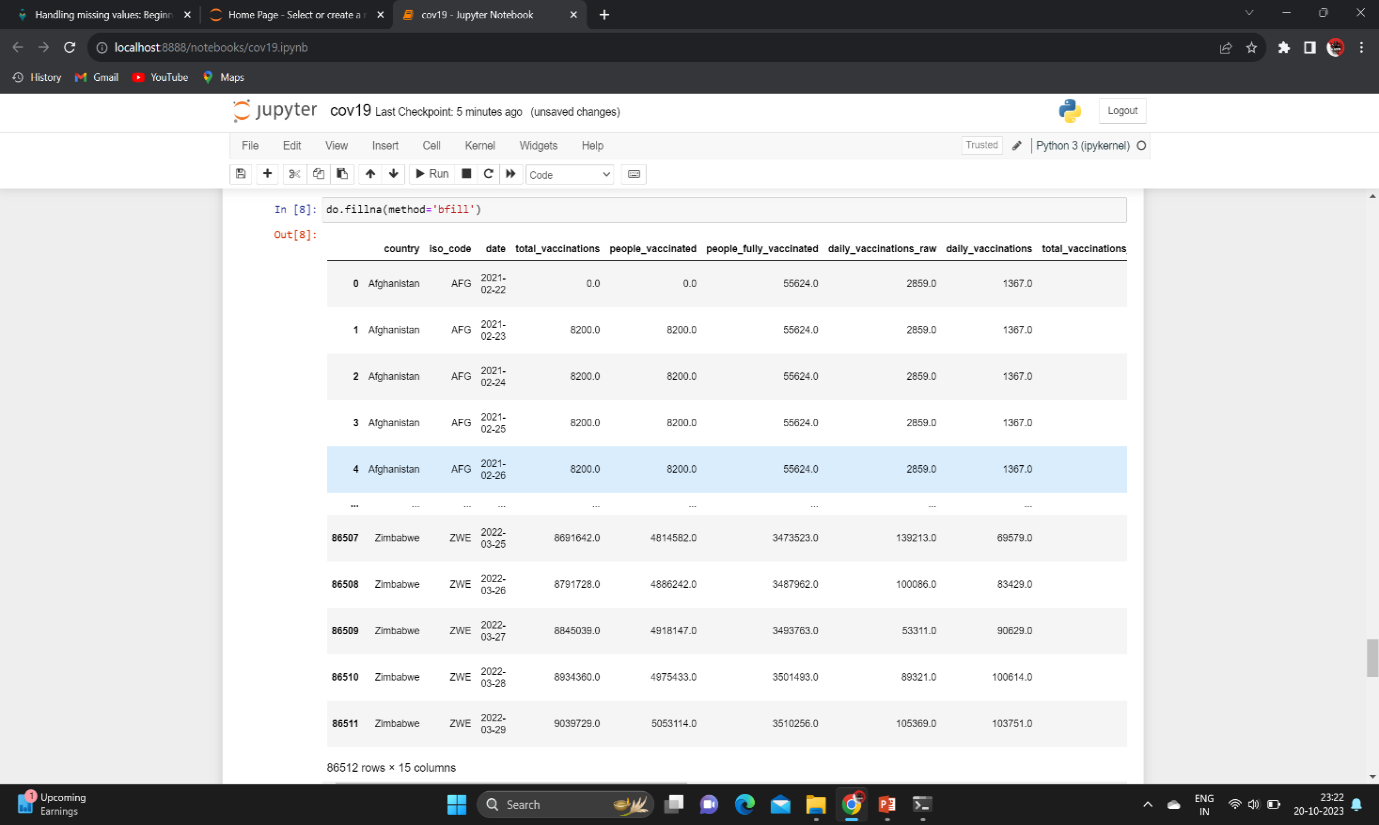
**5.2 Describe the Dataset:**

****

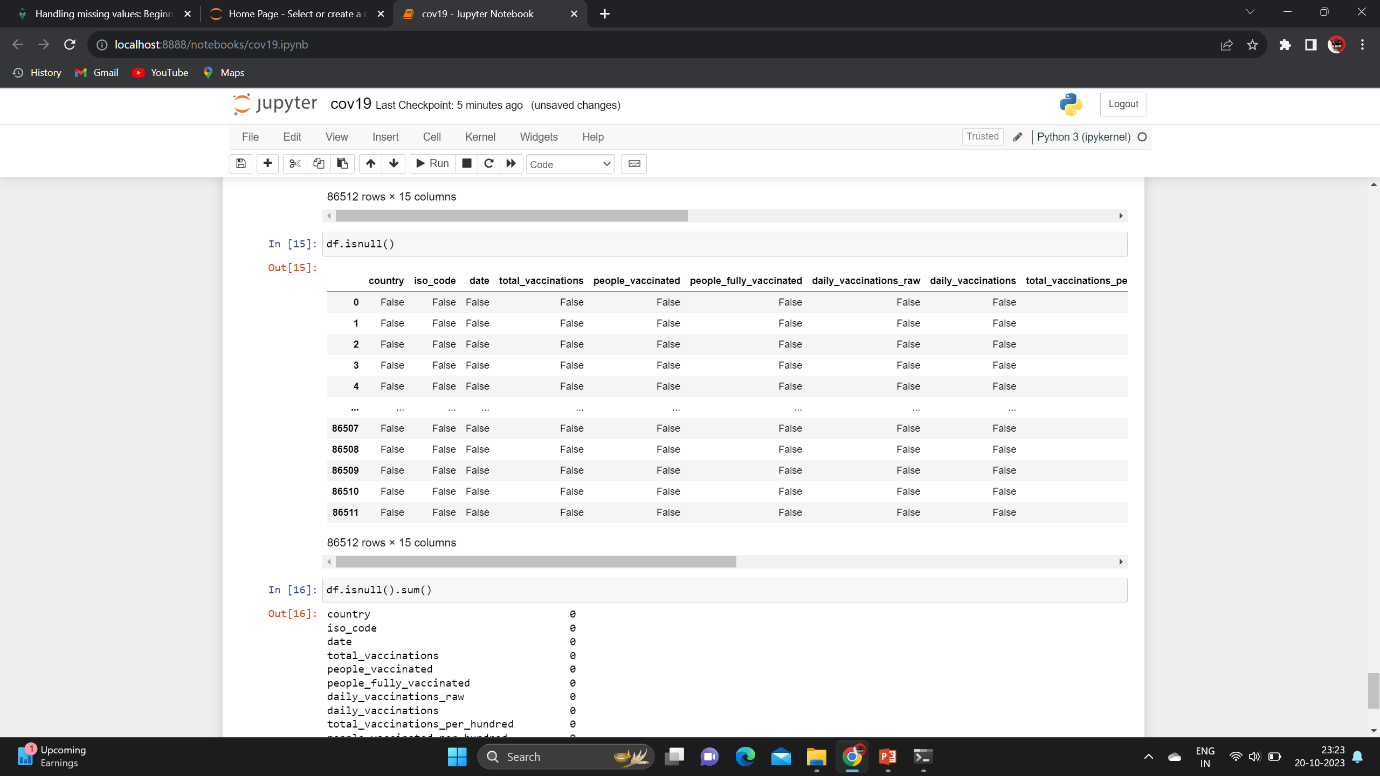
**5.3 Data Preprocessing:**

****

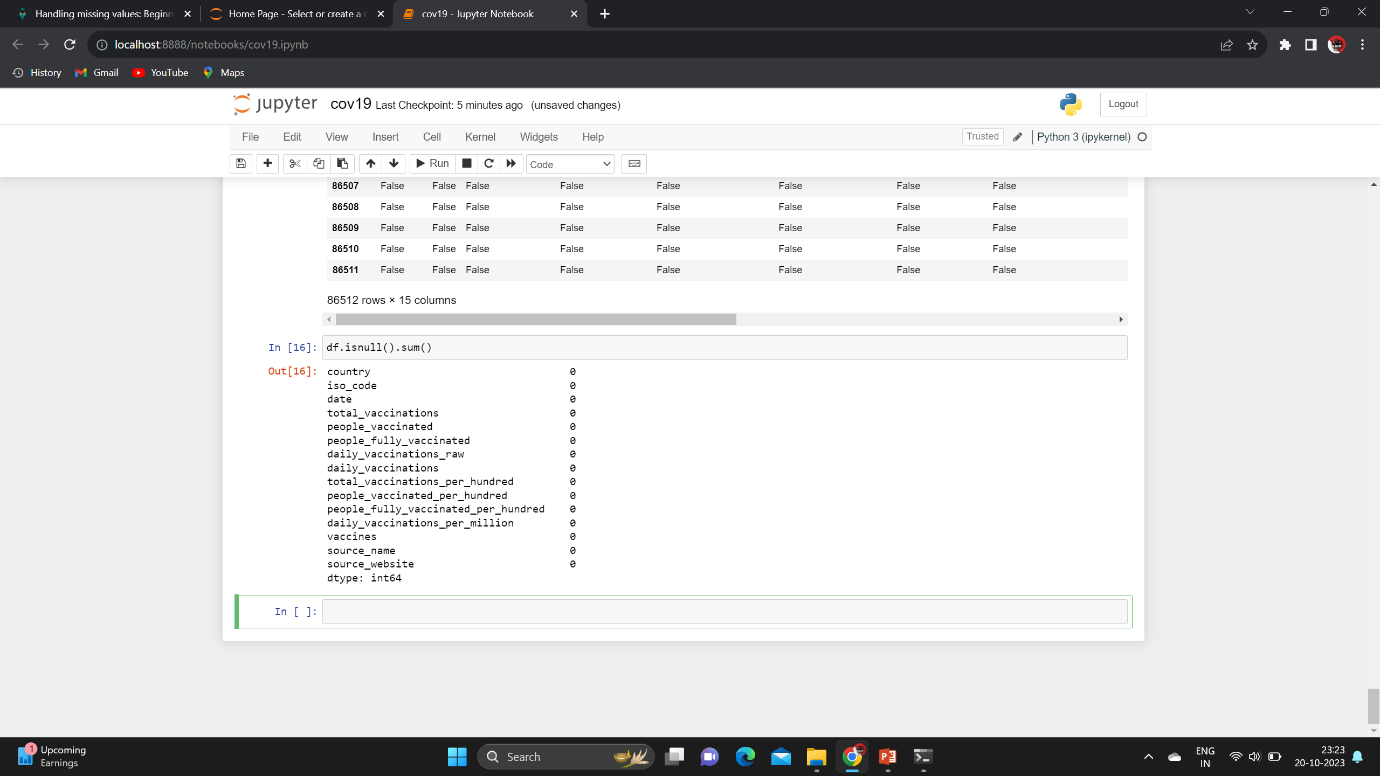
**Backward Fill Method:**

****

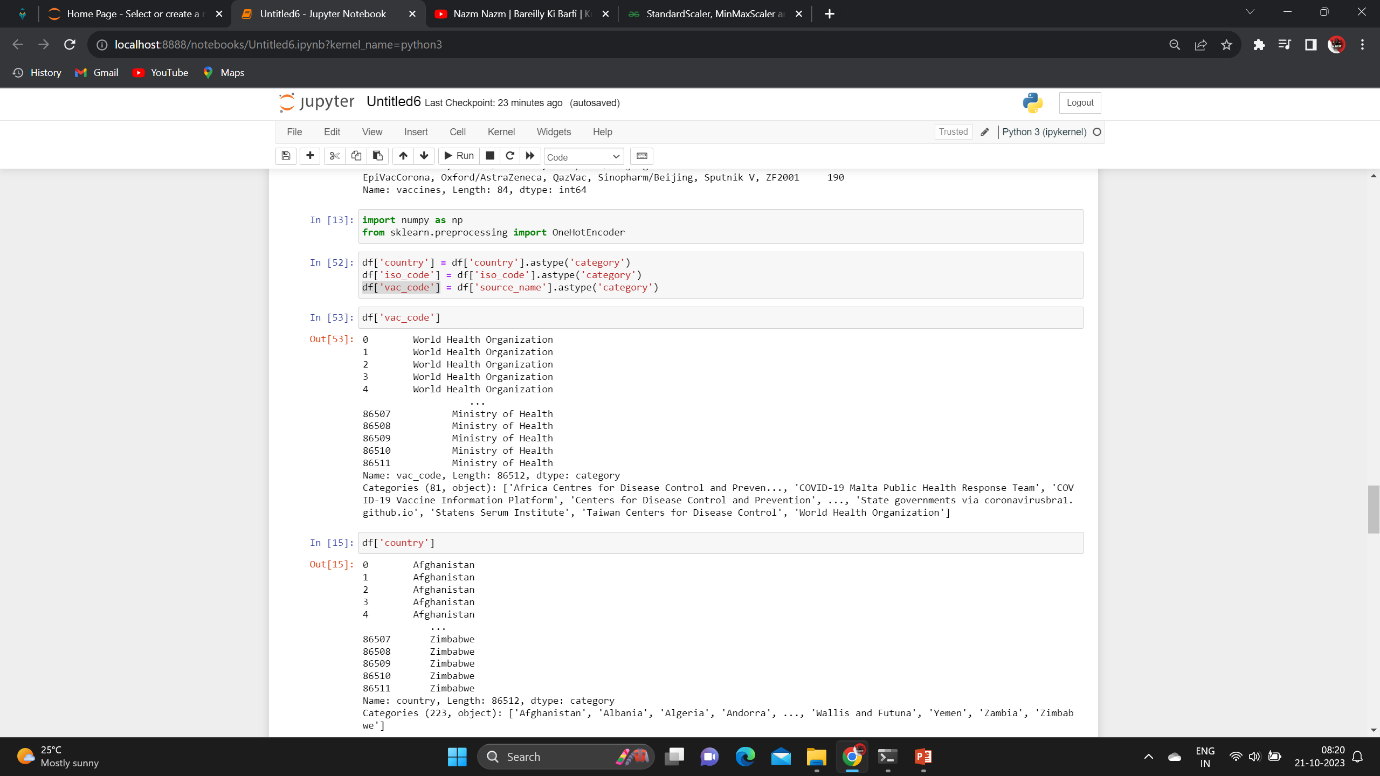
**After Backward Fill Check the isnull Values**:



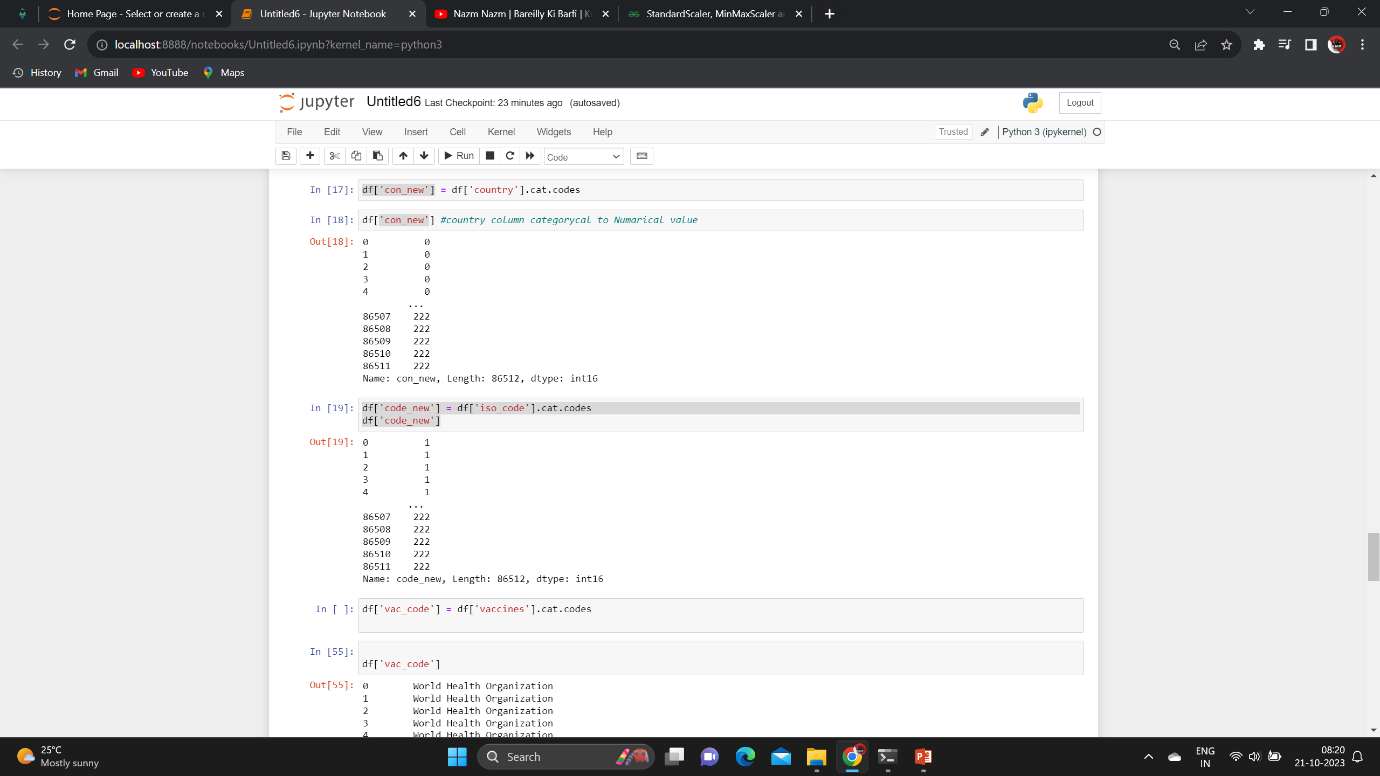
**To check isnull().sum():**

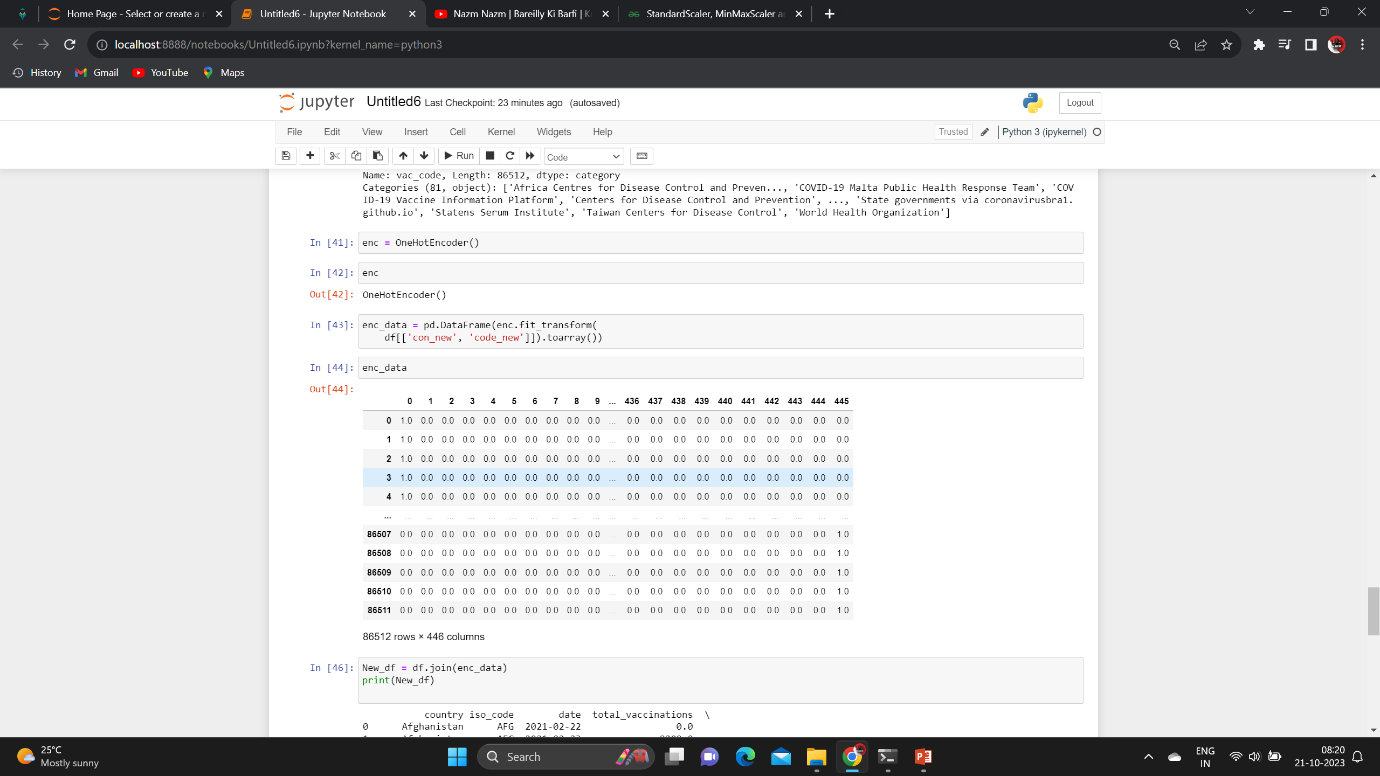


**Use One-hot Encoder:**



**To Change Categorical to Numerical data’s:**





**5.4 Training a Model:**

**How to train and test Data:**

To split your data into training and testing sets, you can use the

train\_test\_split function from the sklearn.model\_selection module. This

function randomly shuffles and partitions your data into two subsets,

typically one for training and one for testing.

**Code:**

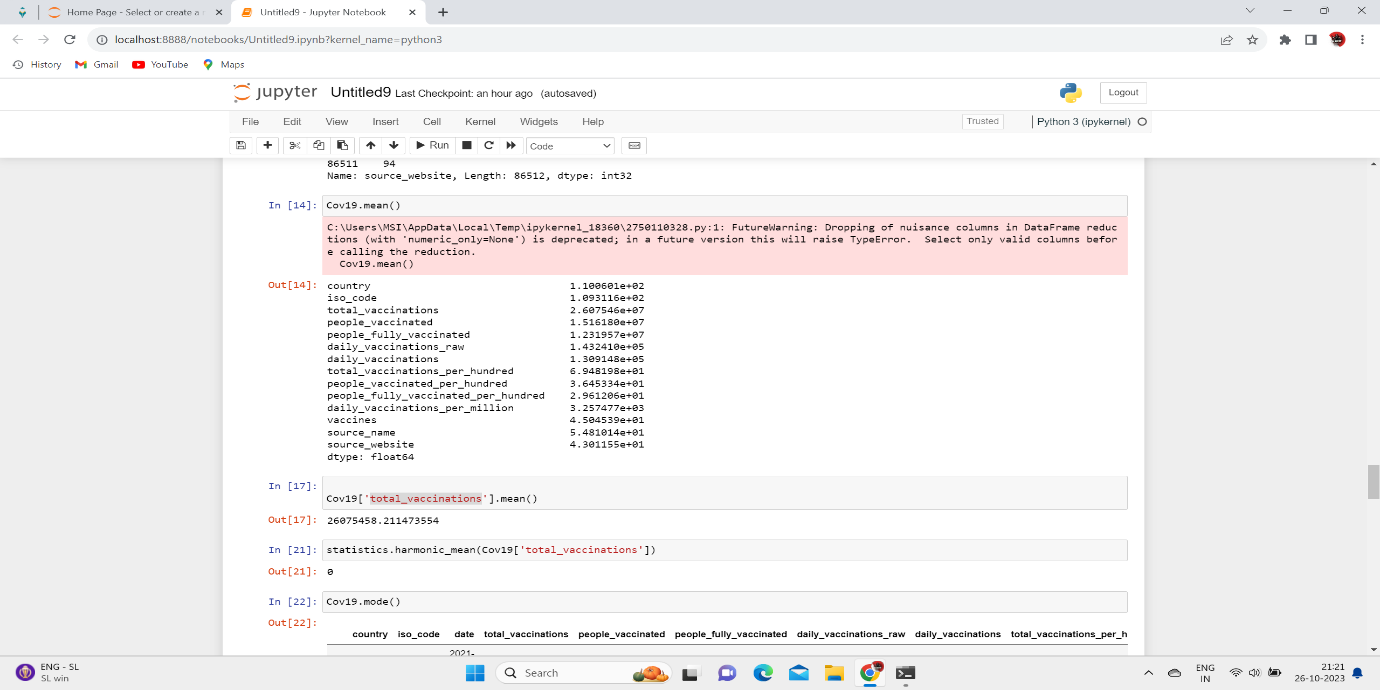
from sklearn.model\_selection import train\_test\_split

# Assuming 'X' is your input data and 'y' is your target variable

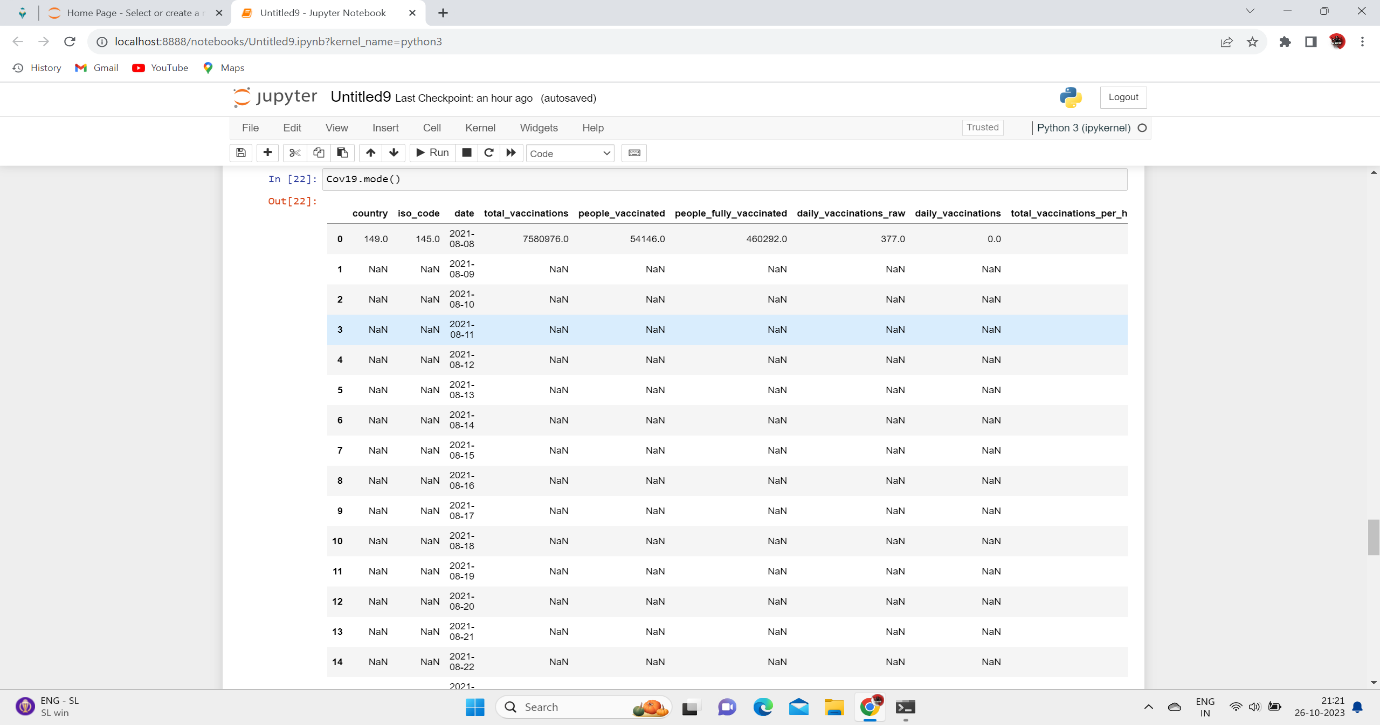
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

**5.5 Statistical Analysis:**

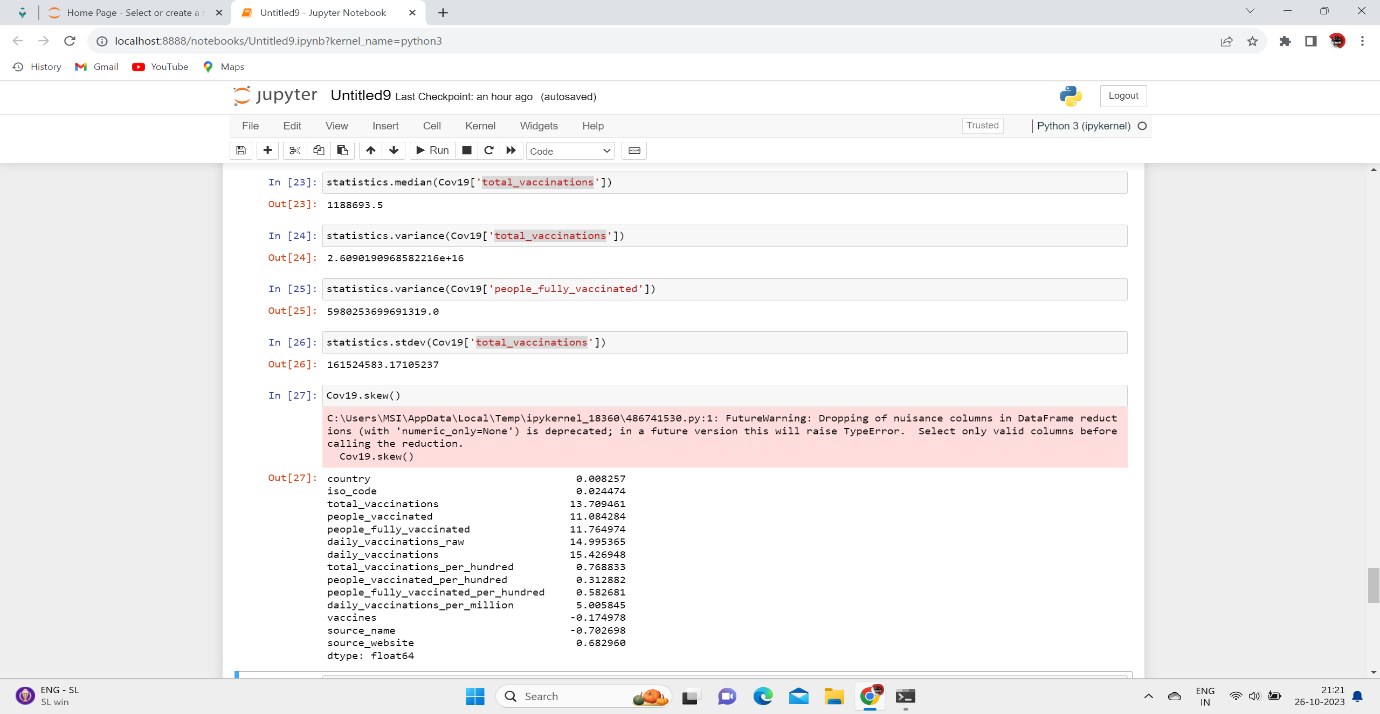
**Mean:**

****

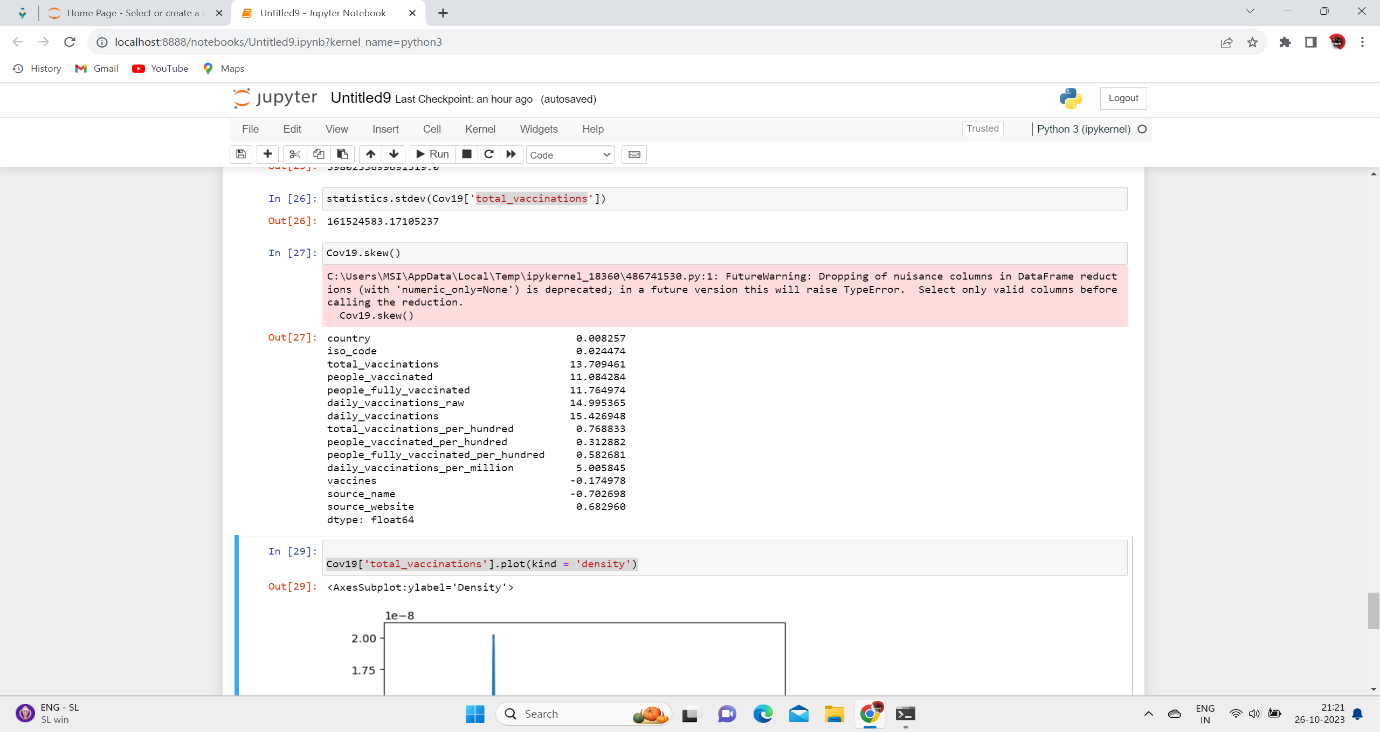
**Mode:**

****

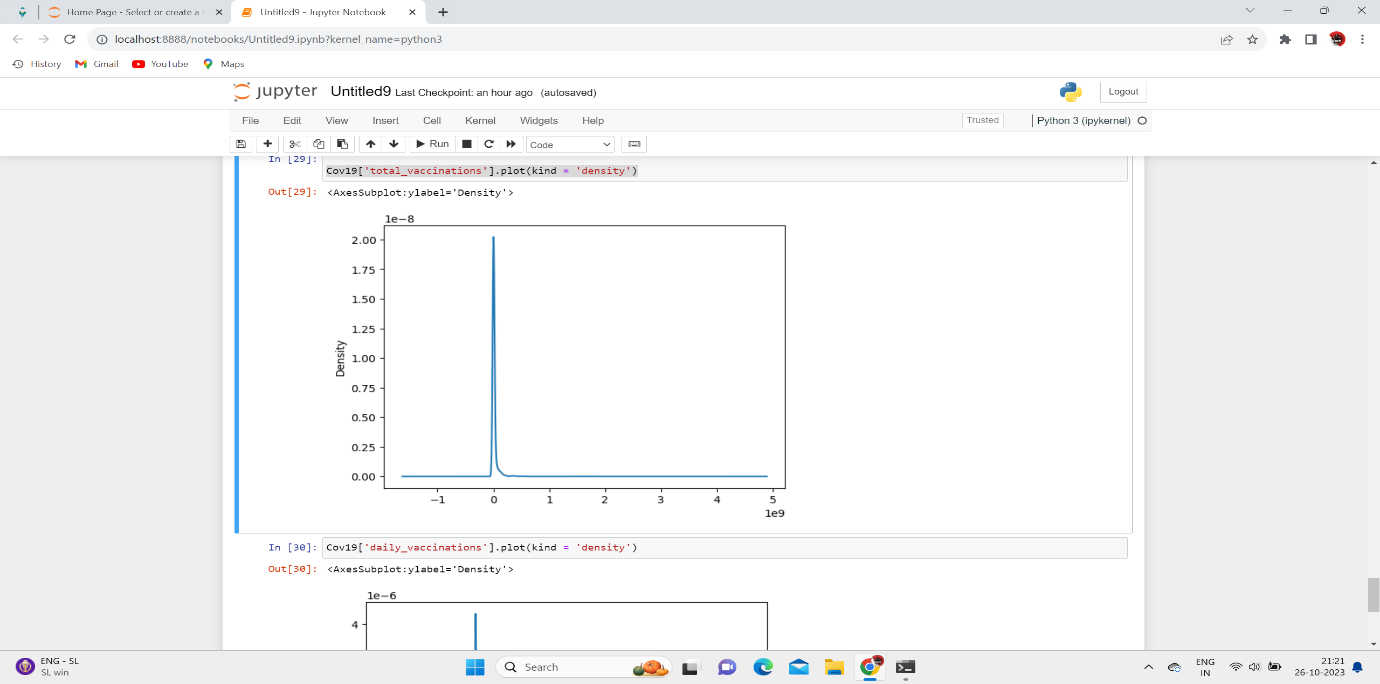
**Median:**

****

**StDev:**

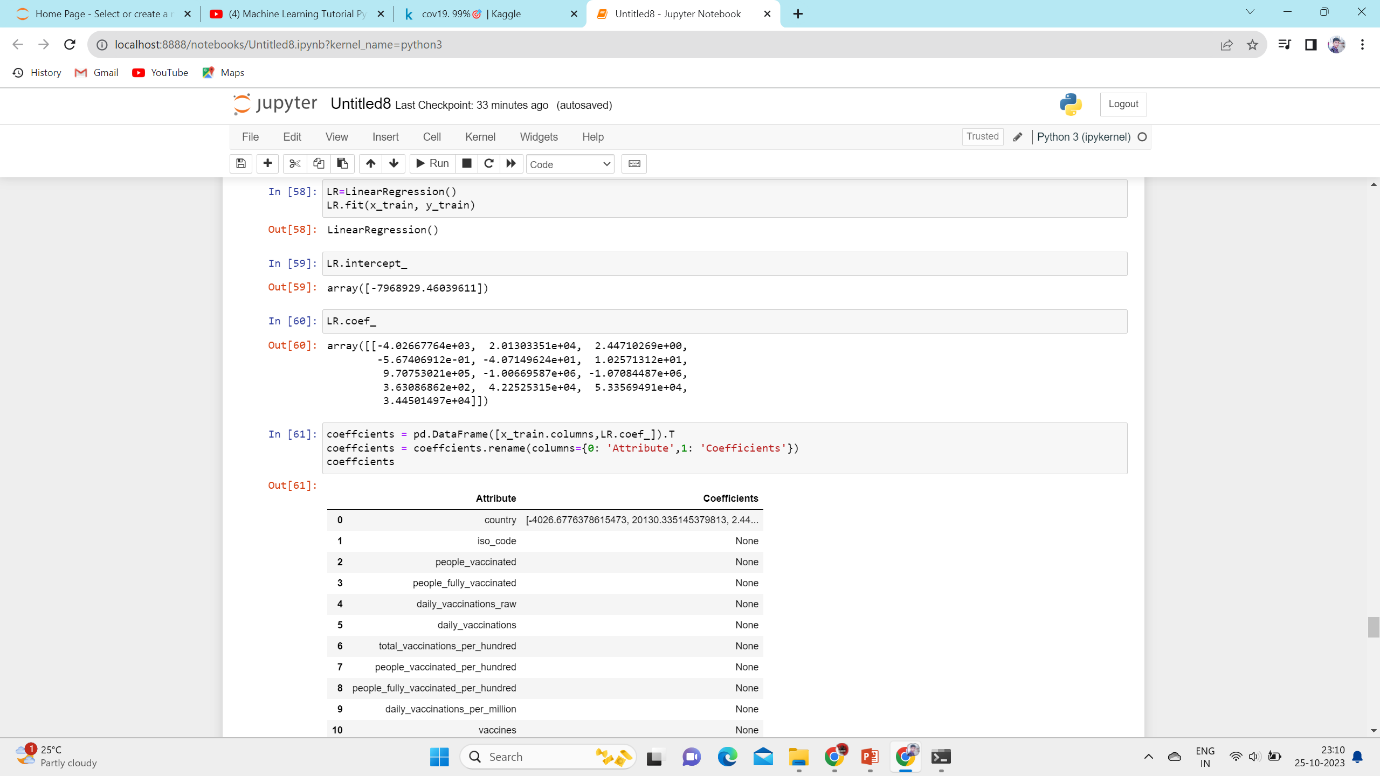
****

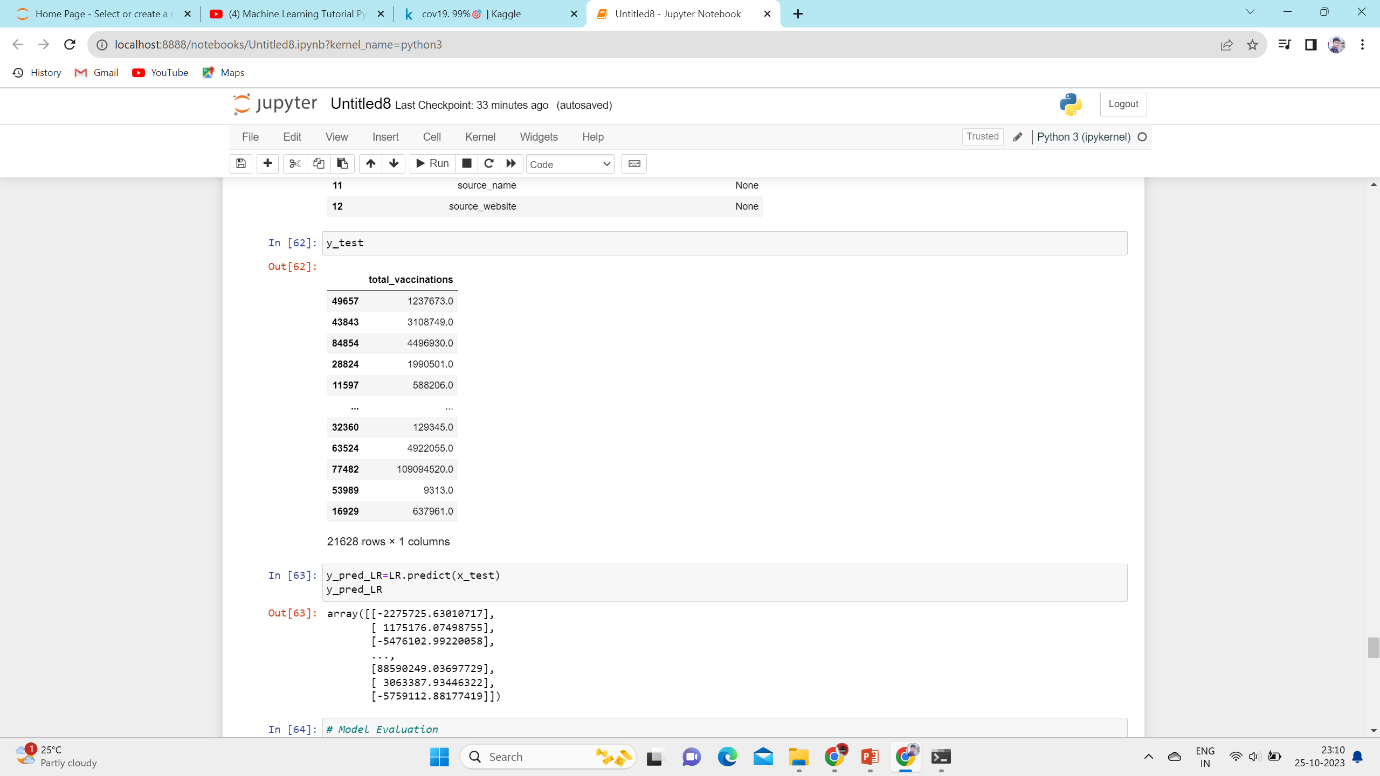
**PLOT:**

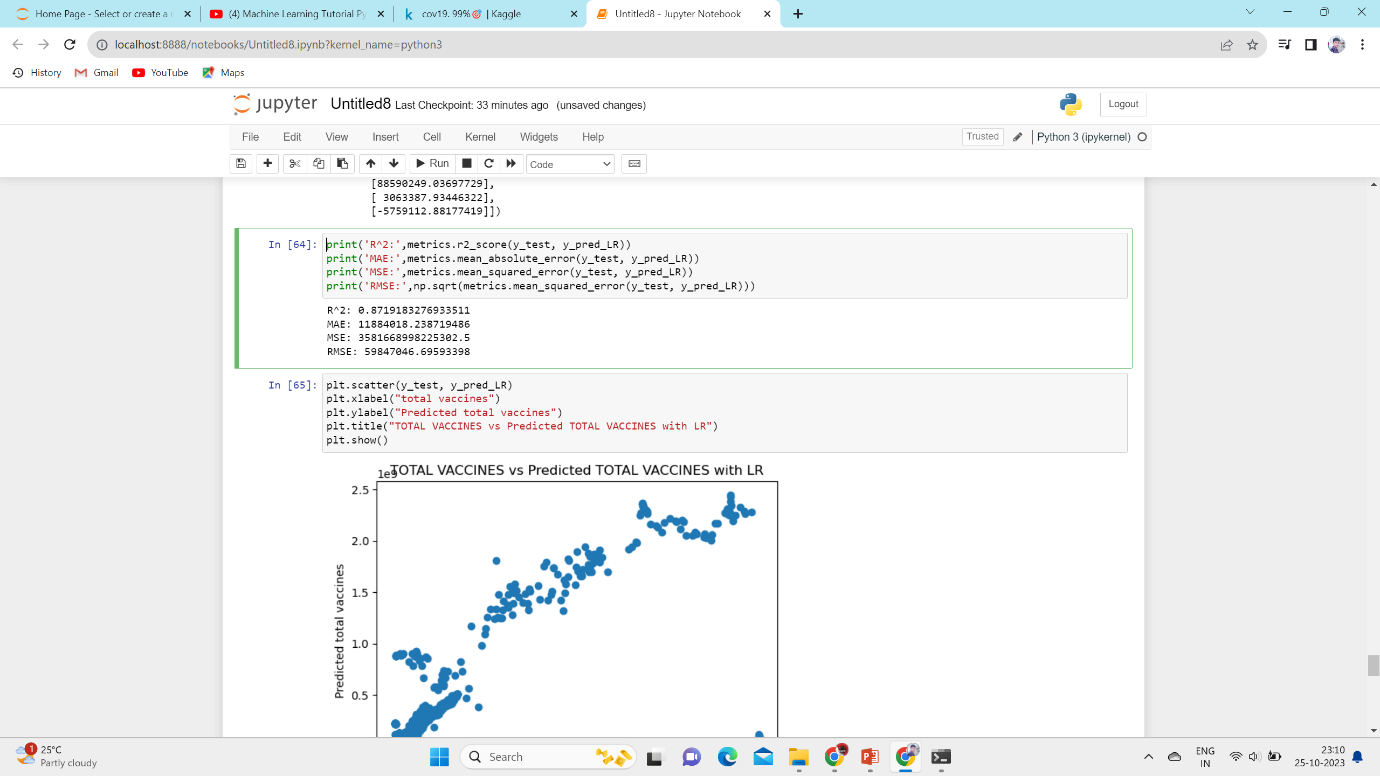
****

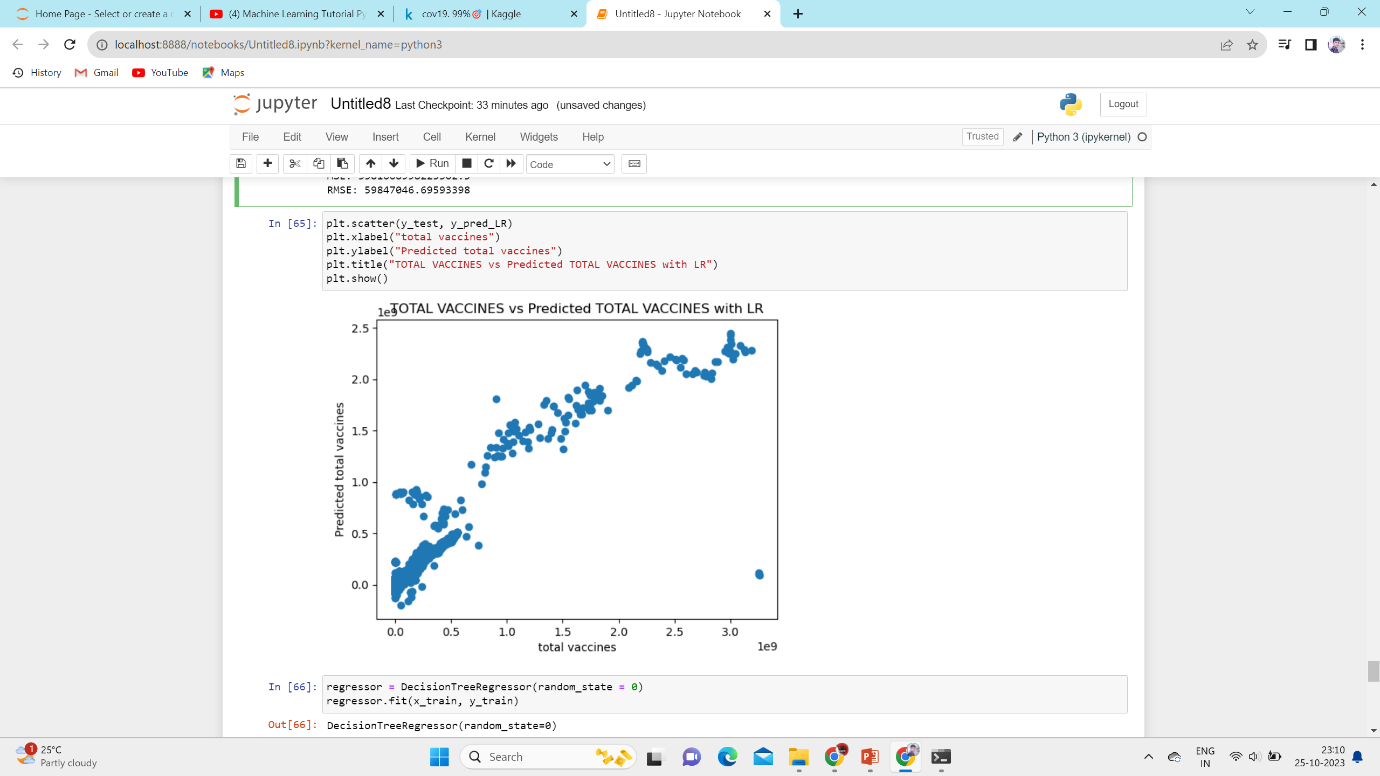
**6. USING ML ALOGRITHMS:**

**6.1 Using Linear Regression:**



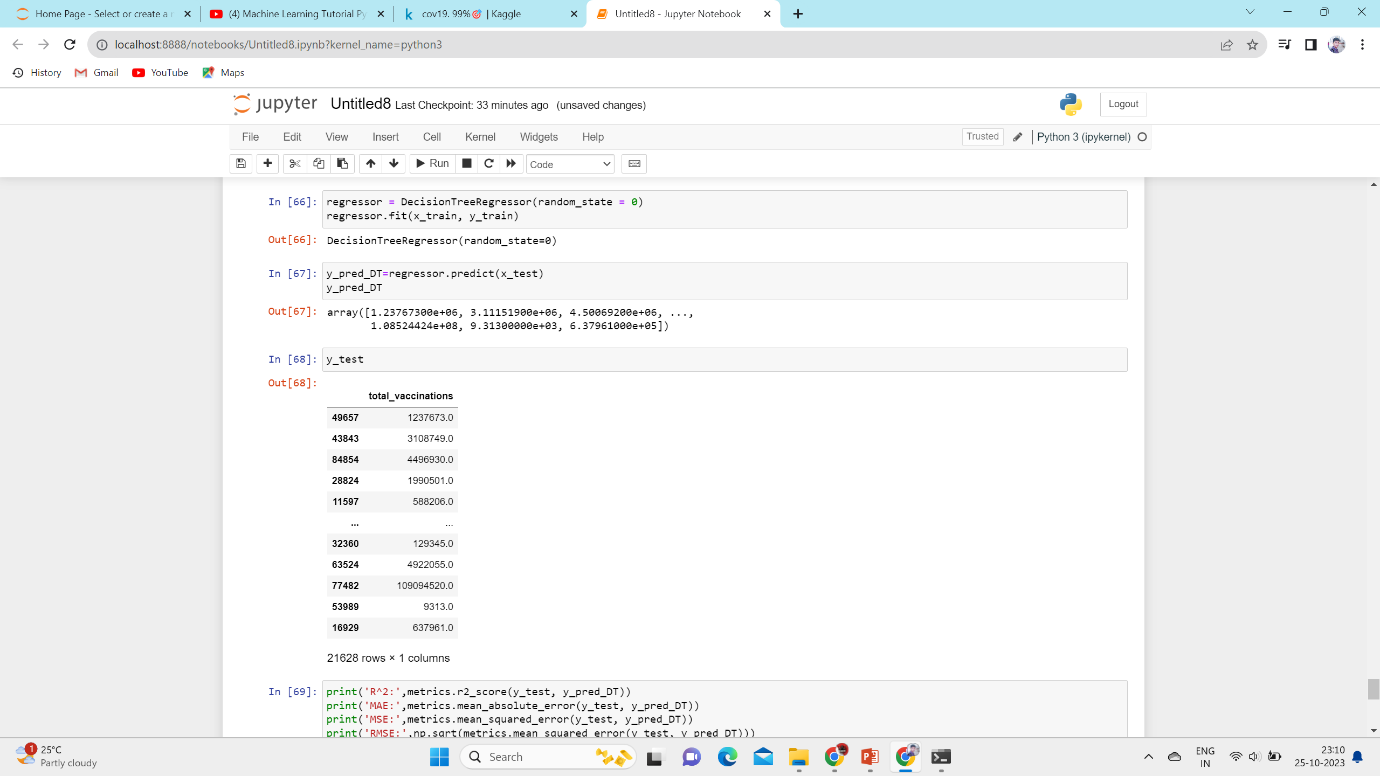


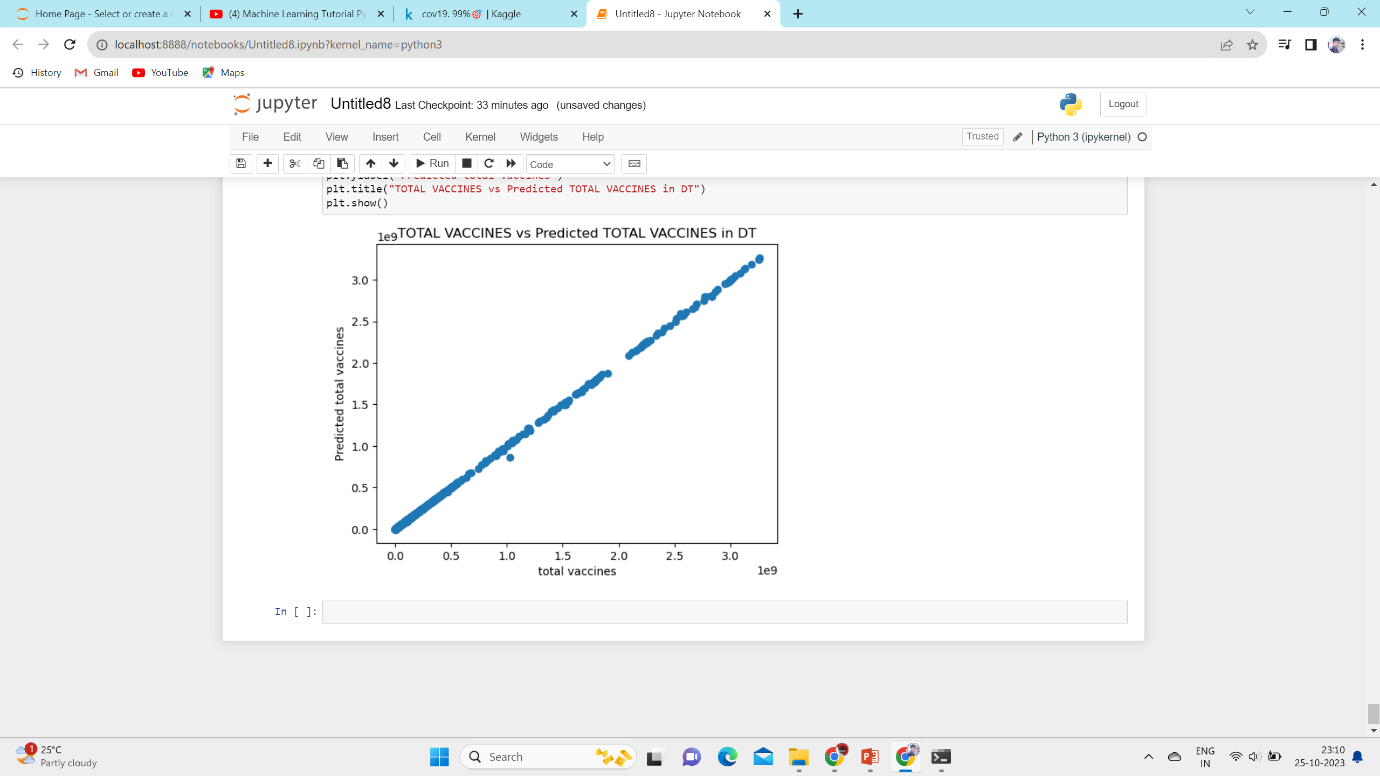
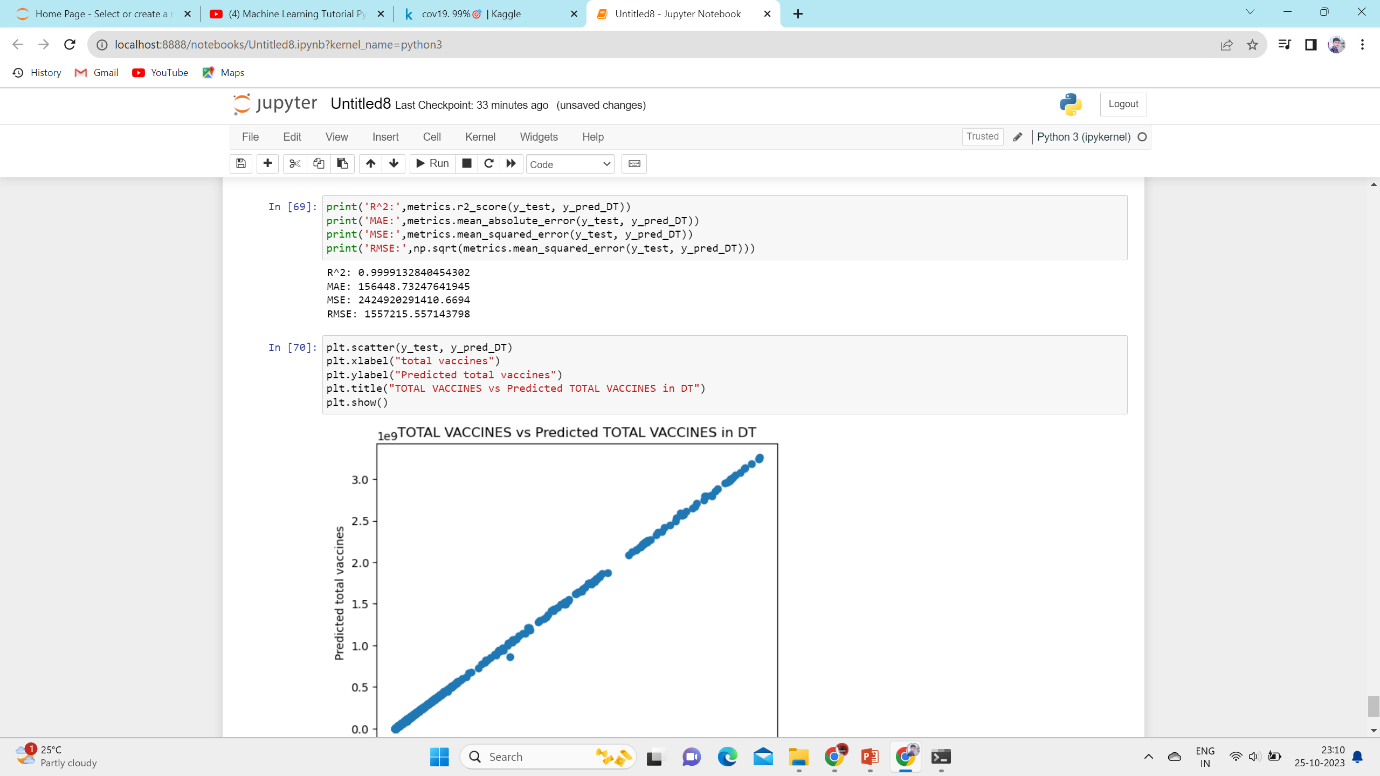




**Using Machine Learning Algorithms**

**6.2 Using Decision Tree:**





**7. ADVANTAGES & DISADVANTAGES**

**ADVANTAGES:**

Analyzing the advantages of COVID-19 vaccines is crucial in understanding their impact on public health. Here are some key **advantages** of COVID-19 vaccine analysis:

**Efficacy Assessment**: Researchers can determine how well a vaccine prevents COVID-19 infection and its variants. This information helps identify the most effective vaccines, enabling healthcare providers to recommend the best options to the public.

**Safety Evaluation**: Continuous analysis of vaccine safety data allows for the detection of rare adverse events or side effects. This helps ensure that the vaccines are safe for the general population.

**Immunization Strategies**: Vaccine analysis assists in developing effective immunization strategies, such as dosing schedules, booster shots, and age-specific recommendations. This helps optimize vaccine distribution and administration.

**Variants Monitoring**: Ongoing analysis helps track the emergence of new variants and their impact on vaccine effectiveness. This information guides the development of updated vaccines if needed.

**Public Confidence**: Regular vaccine analysis and transparency in sharing the results can build and maintain public trust in vaccination programs. It reassures the public that vaccines are rigorously monitored for safety and effectiveness.

**DISADVANTAGES :**

While the analysis of COVID-19 vaccines is crucial for monitoring their safety and efficacy, there can be some potential **disadvantages** or challenges associated with this process:

**Data Accuracy and Quality:** Vaccine analysis heavily relies on the accuracy and quality of data collected. Inaccurate or incomplete data can lead to misleading results and decisions.

**Time-Consuming:** Comprehensive vaccine analysis is time-consuming, and there may be delays in obtaining and processing data, which can impact the ability to respond to rapidly evolving situations, such as the emergence of new variants.

**Resource Intensive:** Conducting rigorous vaccine analysis requires significant resources, including funding, personnel, and infrastructure. This can strain healthcare systems and research organizations.

**Privacy Concerns:** Collecting and sharing health data, especially individual vaccination records, can raise privacy concerns. Striking a balance between data transparency and personal privacy is essential.

**Limited Access to Data:** In some cases, access to vaccine data may be limited due to regulatory restrictions or lack of transparency from vaccine manufacturers or government authorities. This can hinder independent analysis.

**Vaccine Hesitancy:** Extensive analysis may inadvertently raise concerns or doubts about vaccine safety and effectiveness, contributing to vaccine hesitancy in some individuals.

**Complexity of Variants:** The analysis of vaccine effectiveness against emerging variants can be challenging due to the complex and rapidly evolving nature of the virus. It may require frequent updates and adjustments to vaccine strategies.

**Misinterpretation:** The results of vaccine analysis can sometimes be misinterpreted or sensationalized in the media, leading to confusion and potentially undermining public trust.

**8. FUTURE SCOPE**

**Key Findings**

Vaccine Efficacy: The analysis has shown that COVID-19 vaccines, particularly the mRNA and vector-based vaccines, have proven to be highly effective in preventing infection and severe disease. They have played a significant role in reducing the spread of the virus.

Vaccine Variants: The emergence of new variants has posed challenges to vaccine effectiveness. However, the analysis has revealed that booster shots and updated vaccines are effective strategies in addressing this issue and maintaining protection.

Vaccine Safety: The safety analysis indicates that adverse events following vaccination are generally rare and less severe than the risks associated with a COVID-19 infection. Common side effects, such as soreness at the injection site and mild fever, are temporary and manageable.

Vaccine Equity: Data analysis has highlighted disparities in vaccine distribution, with some communities or countries having limited access to vaccines. Achieving global vaccine equity is essential to control the pandemic and protect vulnerable populations.

**Insights**

Ongoing Vigilance: The analysis underscores the importance of ongoing vigilance in monitoring vaccine safety and effectiveness, particularly in the context of emerging variants and potential long-term effects.

Booster Shots: The data supports the use of booster shots for certain populations, such as the elderly and healthcare workers, to maintain high levels of protection and reduce the risk of breakthrough infections.

Communication and Education: Effective communication strategies are crucial in addressing vaccine hesitancy and providing clear information about the benefits and risks of COVID-19 vaccines. Public health authorities should continue their efforts to promote vaccination.

Research on Long-Term Effects: As more data becomes available, continued research into the long-term effects of COVID-19 and its vaccines is warranted to better understand the disease and vaccination outcome

**Recommendations**

**Widespread Vaccination**: Encourage and support widespread vaccination campaigns to achieve high vaccine coverage and attain herd immunity, which is essential for ending the pandemic.

**Boosters and Variant-Specific Vaccines**: Implement booster shot programs and invest in research and development of variant-specific vaccines to address the evolving threat of new variants.

**Equity and Access**: Collaborate on a global scale to ensure equitable access to vaccines, particularly in underserved regions and vulnerable populations.

**Transparency**: Ensure transparency in sharing vaccine analysis findings with the public, healthcare providers, and policymakers to maintain trust and confidence in vaccination programs.

**Continuous Surveillance**: Maintain robust surveillance and monitoring systems to detect and respond to vaccine safety concerns promptly.

**Education and Communication**: Prioritize effective public communication strategies to address vaccine hesitancy, promote vaccine benefits, and disseminate accurate information.

**Preparedness for Future Pandemics**: Learn from the experiences and findings of COVID-19 vaccine analysis to better prepare for future pandemics, including improving response strategies and global collaboration.

**9. CONCLUSION:**

**In conclusion, the analysis of COVID-19 vaccines is an indispensable component of our response to the pandemic. It provides valuable insights into the safety, efficacy, and impact of vaccines, shaping our strategies for mitigating the spread of the virus and protecting public health. The advantages of vaccine analysis are numerous, as it aids in determining which vaccines are most effective, ensures their safety, guides immunization strategies, and monitors the evolving landscape of variants. Additionally, it supports public confidence, aids in achieving herd immunity, and informs economic and global health efforts.**

**In the face of these challenges, it's important to recognize that vaccine analysis is a dynamic and evolving field, and continuous efforts are being made to improve data quality, accessibility, and equitable vaccine distribution. Despite the drawbacks, the benefits of vaccine analysis far outweigh the disadvantages, and it remains a cornerstone of our efforts to control the COVID-19 pandemic and learn valuable lessons for future health crises. It is essential to support and trust the scientific and public health communities in their ongoing efforts to ensure the safety and efficacy of COVID-19 vaccines.**

**10. APPENDIX**

**10.1 Source Code:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn import metrics

from sklearn.metrics import r2\_score

from sklearn.tree import DecisionTreeRegressor

from sklearn.cluster import KMeans

from pycaret.regression import \*

cov19=pd.read\_csv(r'C:\Users\MSI\OneDrive\Desktop\Naan Mudhalvan\country\_vaccinations.csv')

cov19

cov19.describe()

cov19.info()

cov19.isnull().sum()

Cov19 = cov19.fillna(method='bfill')

Cov19

Cov19.isnull().sum()

le=LabelEncoder()

Cov19['country']=le.fit\_transform(Cov19['country'])

Cov19['country']

le=LabelEncoder()

Cov19['iso\_code']=le.fit\_transform(Cov19['iso\_code'])

Cov19['iso\_code']

le=LabelEncoder()

Cov19['vaccines']=le.fit\_transform(Cov19['vaccines'])

Cov19['vaccines']

le=LabelEncoder()

Cov19['source\_name']=le.fit\_transform(Cov19['source\_name'])

Cov19['source\_name']

le=LabelEncoder()

Cov19['source\_website']=le.fit\_transform(Cov19['source\_website'])

Cov19['source\_website']

Cov19['date'] = Cov19['date'].str.replace('-', ' ')

Cov19['date']

Cov19.columns

corr = Cov19.corr()

plt.figure(figsize=(10,8))

sns.heatmap(corr, cmap='viridis', annot=True)

sns.pairplot(Cov19)

sns.regplot( y="daily\_vaccinations",x="total\_vaccinations", data=Cov19)

sns.regplot( y="daily\_vaccinations\_raw",x="total\_vaccinations", data=Cov19)

sns.displot(Cov19, x="country", hue="vaccines", common\_norm=False)

sns.scatterplot(Cov19, x='vaccines', y='country')

sns.displot(Cov19, x="country",kde=True)

#Train and Test a Model

x=Cov19[['country', 'iso\_code','people\_vaccinated',

'people\_fully\_vaccinated','daily\_vaccinations\_raw', 'daily\_vaccinations',

'total\_vaccinations\_per\_hundred', 'people\_vaccinated\_per\_hundred',

'people\_fully\_vaccinated\_per\_hundred', daily\_vaccinations\_per\_million', 'vaccines', 'source\_name', 'source\_website']]

X

y=Cov19[['total\_vaccinations']]

y

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x,y, random\_state=42)

x\_train

x\_test

y\_train

y\_test

# LinearRegression

LR=LinearRegression()

LR.fit(x\_train, y\_train)

LR.intercept\_

LR.coef\_

coeffcients = pd.DataFrame([x\_train.columns,LR.coef\_]).T

coeffcients = coeffcients.rename(columns={0: 'Attribute',1: 'Coefficients'})

coeffcients

y\_test

y\_pred\_LR=LR.predict(x\_test)

y\_pred\_LR

print('R^2:',metrics.r2\_score(y\_test, y\_pred\_LR))

print('MAE:',metrics.mean\_absolute\_error(y\_test, y\_pred\_LR))

print('MSE:',metrics.mean\_squared\_error(y\_test, y\_pred\_LR))

print('RMSE:',np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred\_LR)))

plt.scatter(y\_test, y\_pred\_LR)

plt.xlabel("total vaccines")

plt.ylabel("Predicted total vaccines")

plt.title("TOTAL VACCINES vs Predicted TOTAL VACCINES with LR")

plt.show()

# DecisionTree

regressor = DecisionTreeRegressor(random\_state = 0)

regressor.fit(x\_train, y\_train)

y\_pred\_DT=regressor.predict(x\_test)

y\_pred\_DT

y\_test

print('R^2:',metrics.r2\_score(y\_test, y\_pred\_DT))

print('MAE:',metrics.mean\_absolute\_error(y\_test, y\_pred\_DT))

print('MSE:',metrics.mean\_squared\_error(y\_test, y\_pred\_DT))

print('RMSE:',np.sqrt(metrics.mean\_squared\_error(y\_test, y\_pred\_DT)))

plt.scatter(y\_test, y\_pred\_DT)

plt.xlabel("total vaccines")

plt.ylabel("Predicted total vaccines")

plt.title("TOTAL VACCINES vs Predicted TOTAL VACCINES in DT")

plt.show()

**10.2 GitHub:**

**GitHub link: https://github.com/mugesharasan/dataSciencePhase.git**

