

Pune Institute of Computer Technology, Pune

**DSBDA Mini Project**

(AY 2022-23)

**Clas**s: T.E. 2

**Batch**: K2

**Mini-Project**

On

**Analysis of Covid Pandemic Response in India**

**Group Members:**

31302: Amrit Singh

31318: Jaydeep Datre

**Guided by**

Prof. Kimaya Urane

**Title:** Analysis of Covid Pandemic Response in India

# Problem Definition:

Use the following covid\_vaccine\_statewise.csv dataset and perform the following analytics on the given dataset:

1. Describe the dataset
2. Number of persons state-wise vaccinated for the first dose in India
3. Number of persons state-wise vaccinated for the second dose
4. Number of Males vaccinated
5. Number of females vaccinated

# Objectives:

1. To understand how to load and pre-process a given dataset.
2. To understand how to remove useless, wrong, or null values.
3. To create visualizations to find relationships among the data.
4. To perform analysis using various techniques to find the relation between dependent and independent variables.

# Outcomes:

After completion of the project, students will be able to:

* 1. Load and pre-process data, and remove unwanted and null values.
  2. Categorize and rename columns, and set up the dataset.
  3. Use data visualization and find relationships in the data.
  4. Find the effectiveness of the Anti-Covid drive implemented in India.

# Requirements:

* Computer System with:
* I5 processor, 256 GB SSD, 8GB RAM. ● Jupyter Notebook
* Python with numpy, pandas, matplotlib

# Abstract:

This project aimed to analyze and visualize data from India's Covid19 vaccination drive to identify trends and inefficiencies in the program. We used various data visualization and analysis techniques, including linear regression, to explore the relationships between different variables and the number of successfully vaccinated individuals.

# Introduction:

The Covid pandemic has brought about unprecedented changes in the lives of people around the world, including India. With the sudden

emergence of the virus, the Indian government has had to take a number of measures to control the spread of the virus, including a vaccination drive to ensure that every citizen of India has access to safe and free vaccines. However, there are concerns about the efficiency of the immunization program, with questions about the proportion of people receiving vaccines properly and the people who are being treated unfairly or not given the opportunity for immunization.

The vaccination drive is a crucial aspect of India's fight against the Covid pandemic, but it is important to understand how well the program is being implemented to ensure that the maximum number of people are being vaccinated in an efficient and fair manner. The objective of our data analytics project is to analyze and visualize the available data to find out the efficiency of the immunization drive and identify state-wise trends in vaccination.

To achieve this objective, we will be exploring trends in the number of successfully vaccinated individuals based on various factors such as age, gender, state of residence, the vaccine manufacturer, and other important variables. By aggregating and properly visualizing this data, we can gain insights into whether some people have been deprived of proper healthcare due to poor infrastructure, prejudices, or other reasons.

Through our data analysis, we hope to provide valuable insights to policymakers and health professionals, enabling them to identify areas of improvement and take appropriate measures to ensure that the vaccination drive is implemented in the most efficient and equitable manner possible. Ultimately, our project aims to contribute to India's efforts to combat the Covid pandemic and protect its citizens.

# Methodology:

**Softwares Used:**

# NumPy:

NumPy is a widely used package in Python for scientific computing and data analysis. Its core feature is the N-dimensional array object, which is the basis for many numerical computing applications in Python. NumPy provides a variety of tools for performing operations on arrays, such as mathematical, logical, and shape manipulation operations. It also includes useful tools for linear algebra, Fourier analysis, and random number generation.

# Pandas:

Pandas is a powerful Python package for data manipulation and analysis. It provides two primary data structures: Series and DataFrame. A Series is a onedimensional array-like object that can hold any data type, while a data frame is a two-dimensional table-like data structure with labeled axes.

Pandas are particularly useful for handling tabular data, such as data from spreadsheets and SQL tables. It provides a wide range of tools for data cleaning, merging, reshaping, slicing, indexing, and aggregation. Pandas is also capable of handling time-series data and providing tools for working with dates and times.

# Matplotlib.pyplot:

Matplotlib.pyplot is a popular Python package for creating visualizations and plots. It provides a range of functions for creating various types of plots, including line plots, scatter plots, bar plots, histograms, and many more.

Matplotlib. pyplot is highly customizable, and users can modify almost every aspect of a plot, such as colors, labels, axis limits, and annotations. It is widely used in the scientific and engineering communities for creating high-quality visualizations of data.

# Seaborn:

Seaborn is a Python data visualization library built on top of matplotlib. It provides a high-level interface for creating statistical graphics. Seaborn is particularly useful for visualizing complex statistical relationships, such as regression models, categorical data, and time-series data. It includes many built-in themes and color palettes that make it easy to create aesthetically pleasing plots. Seaborn is widely used in data science and machine learning applications.

# Scikit-learn:

Scikit-learn is a popular open-source Python library that provides a range of tools for data analysis, machine learning, and artificial intelligence. It is built on top of other scientific computing libraries in Python, such as NumPy, SciPy, and Matplotlib.

Scikit-learn offers a wide range of machine learning algorithms, including supervised and unsupervised learning algorithms, clustering, regression, classification, and dimensionality reduction, among others. It also provides tools for data preprocessing, feature extraction, and data visualization.

# Implementation:

The actual implementation of the project involved various steps. According to the data analysis life cycle, different steps were implemented

# Data Collection:

A dataset called covid\_19\_india.csv and another dataset called covid\_vaccine\_statewise.csv was retrieved from Kaggle and used for the project.

# Data Processing:

Many pre-processing tasks were done such as renaming the columns to increase the ease of coding tasks, checking for null values, removing columns having the same value for each record, removing null values, and replacing them with appropriate values (either 0 or mean of the column). We also checked for outliers or errors in data and accordingly removed those outliers by using Z square method.

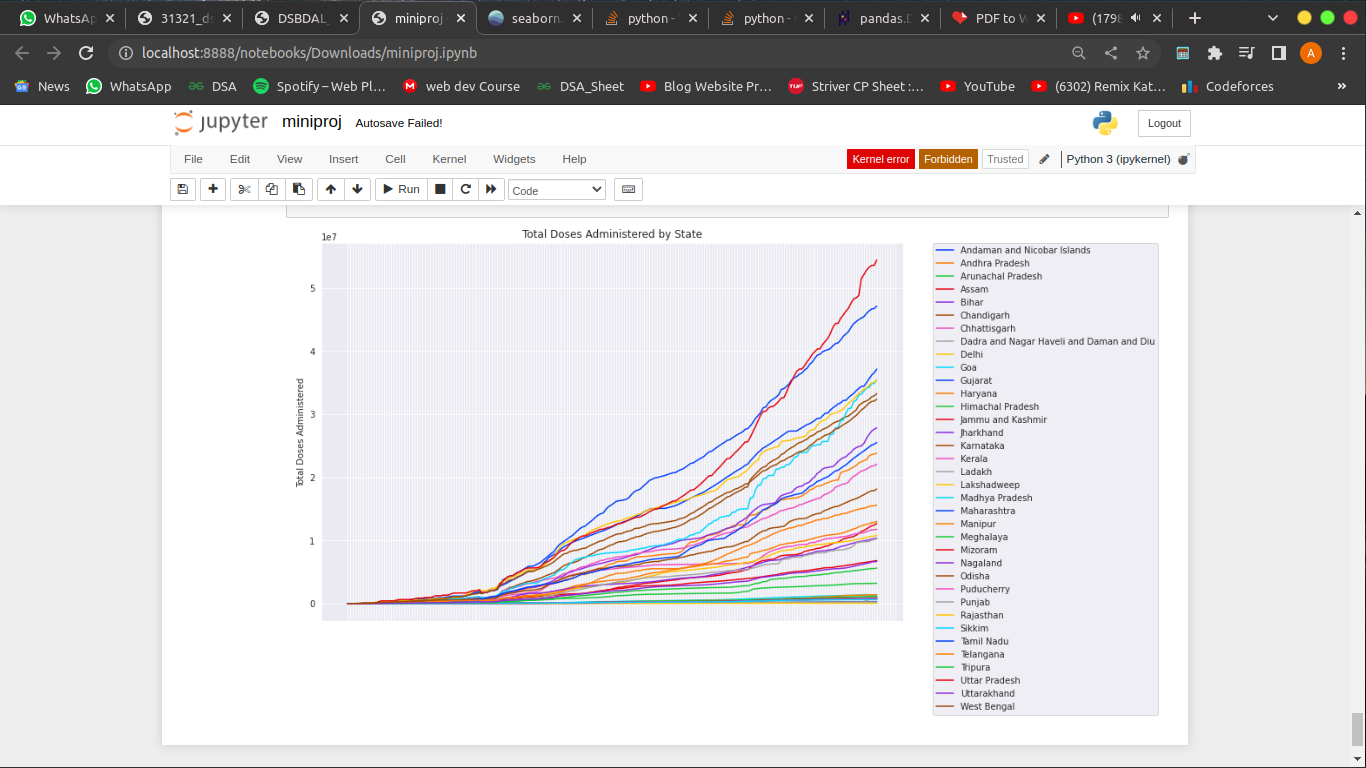
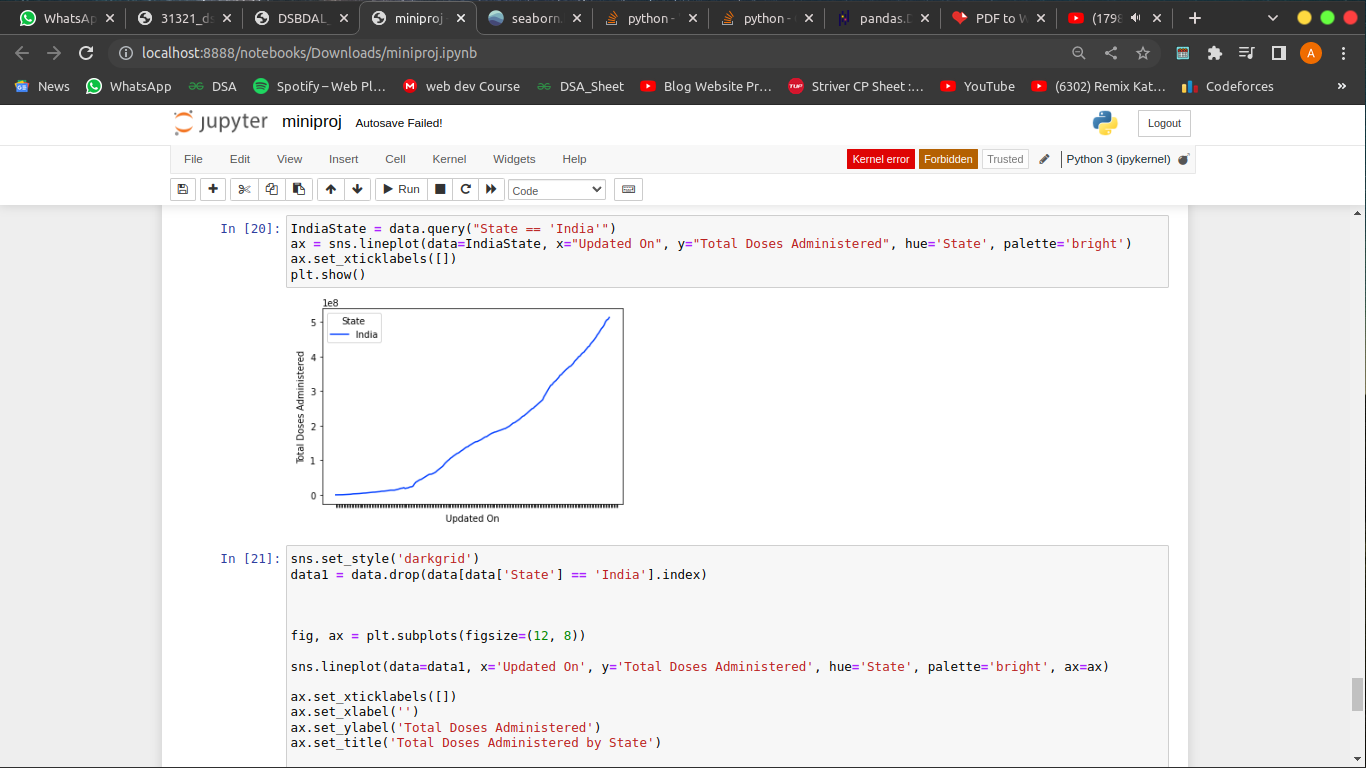
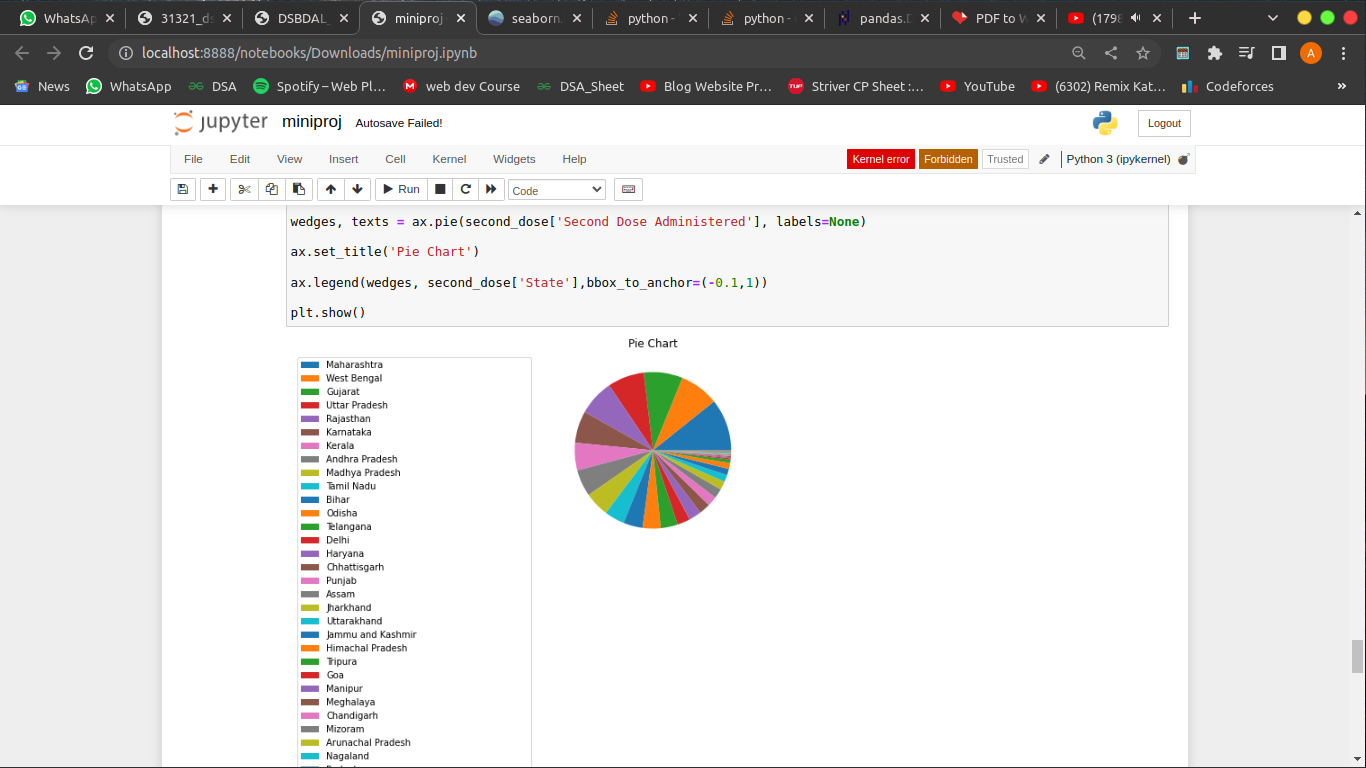
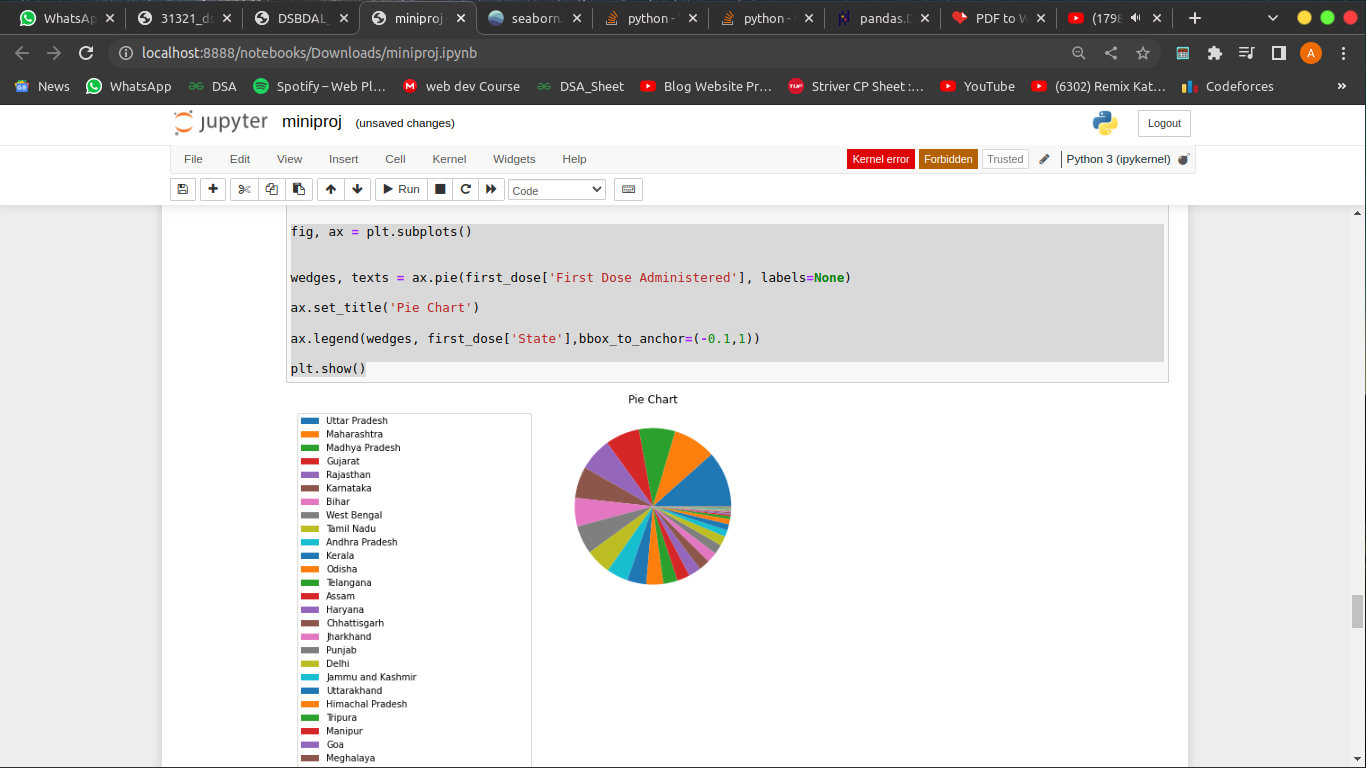
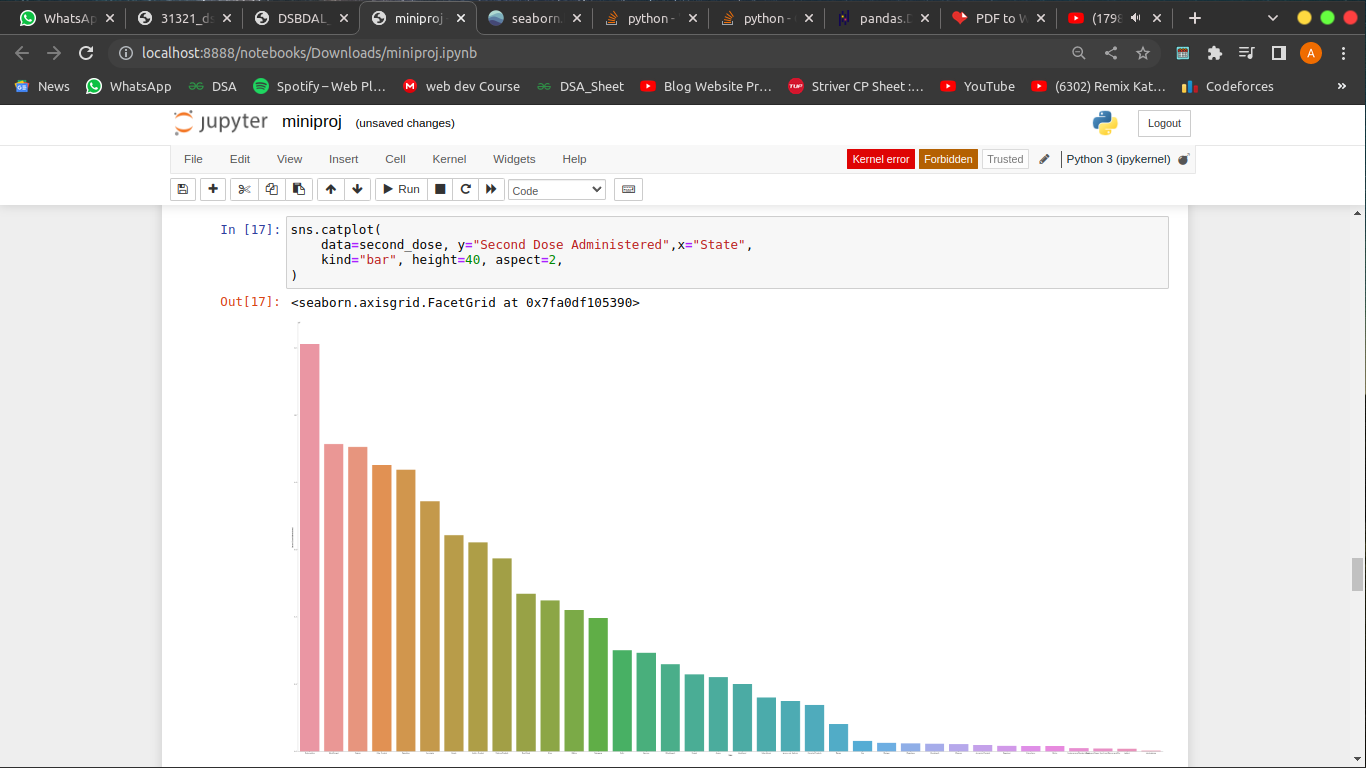
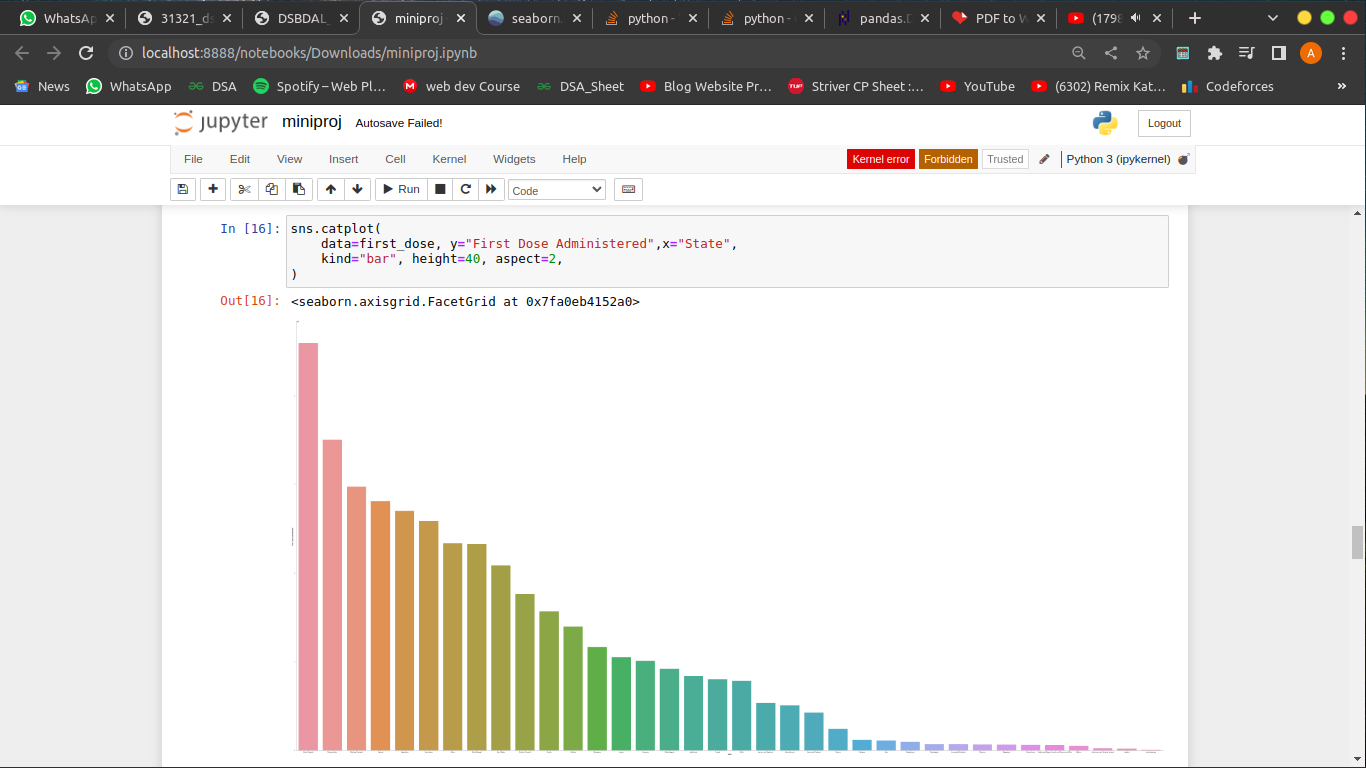
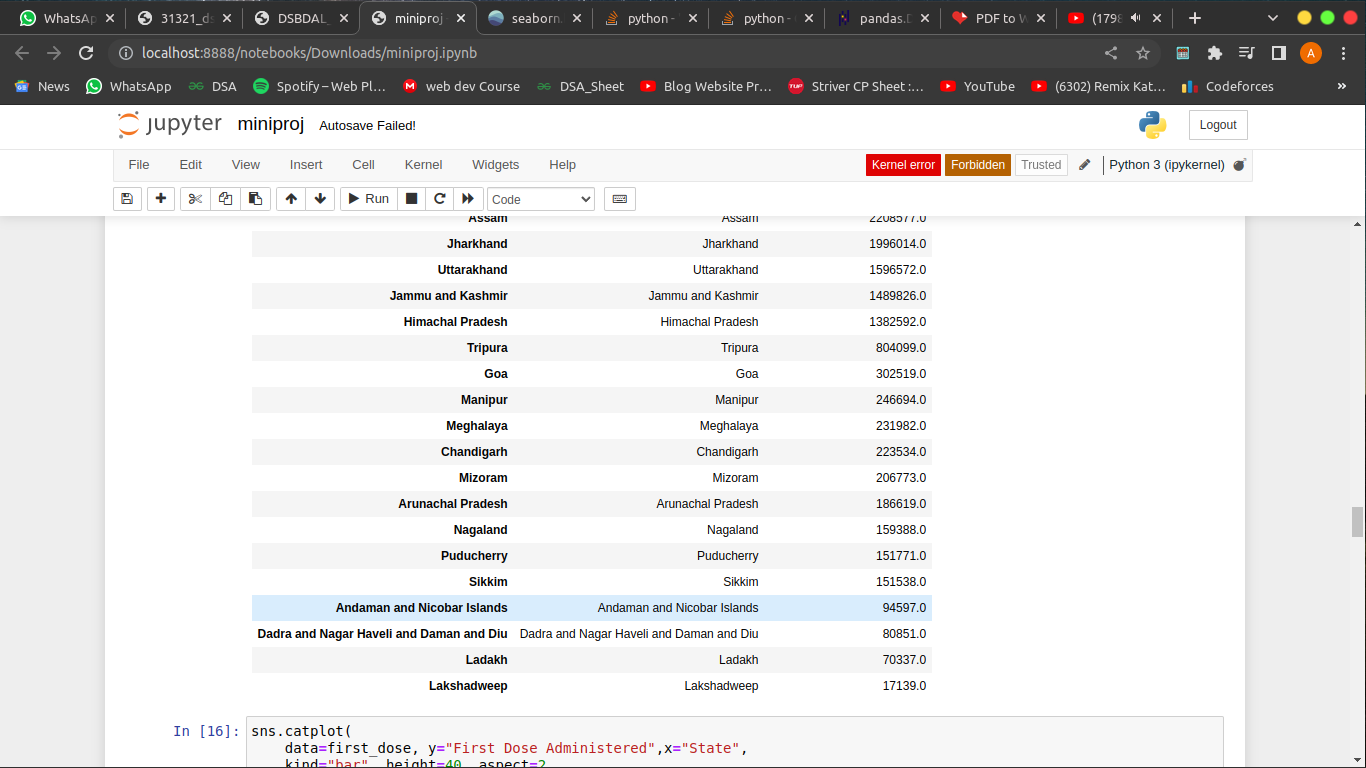
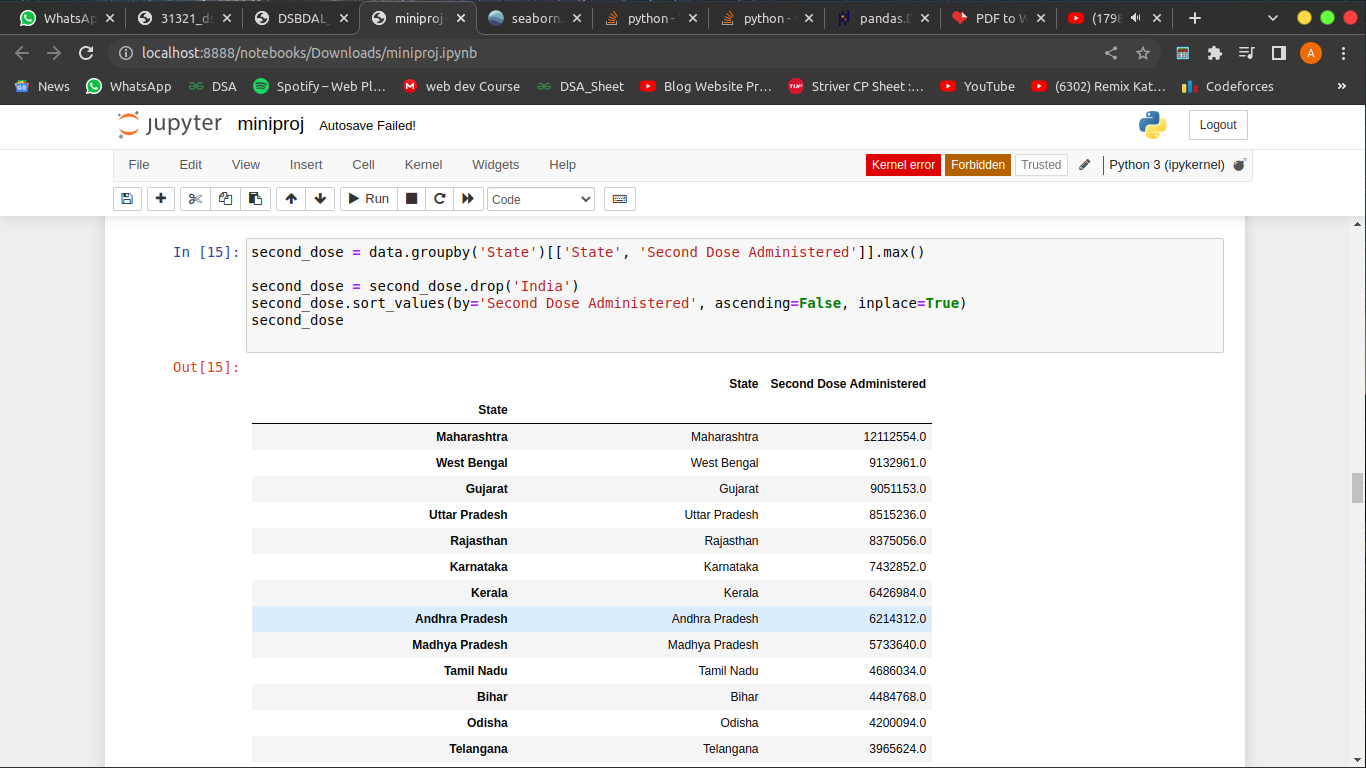
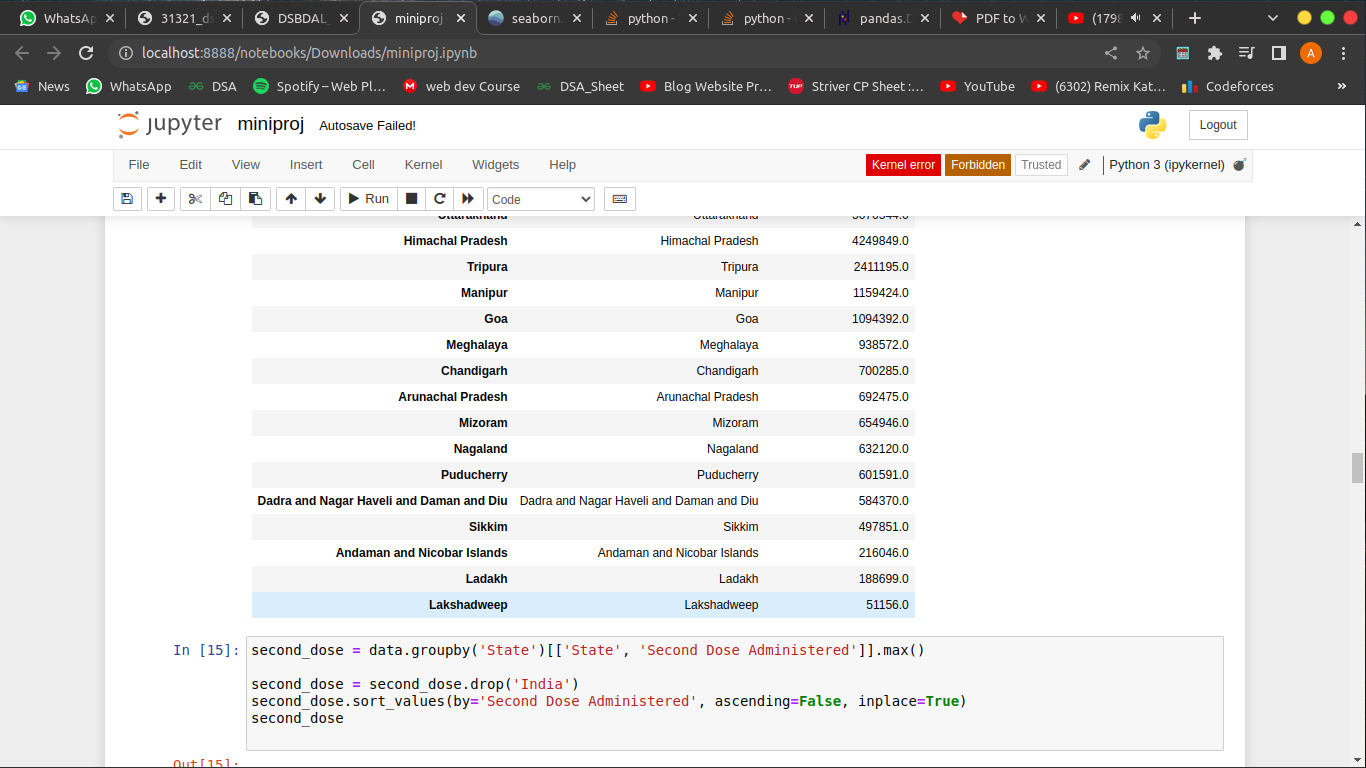
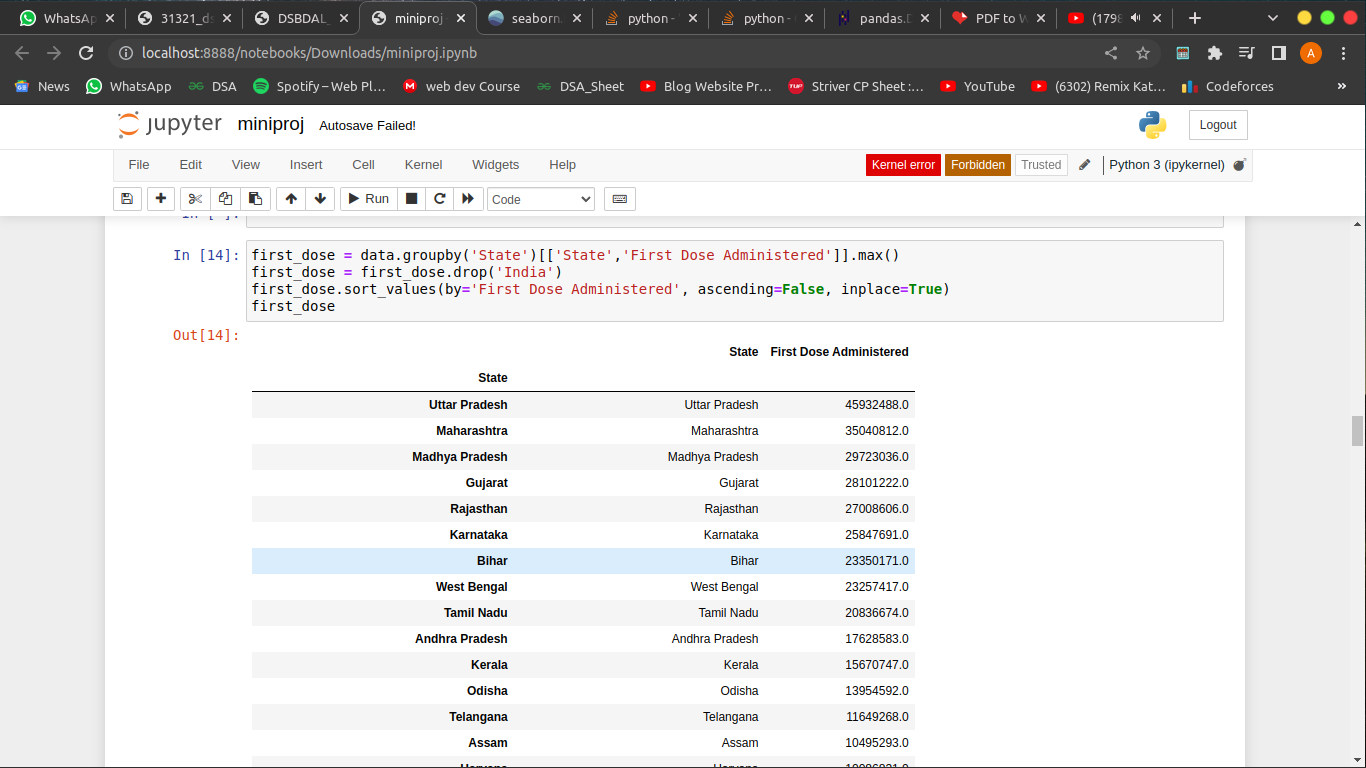
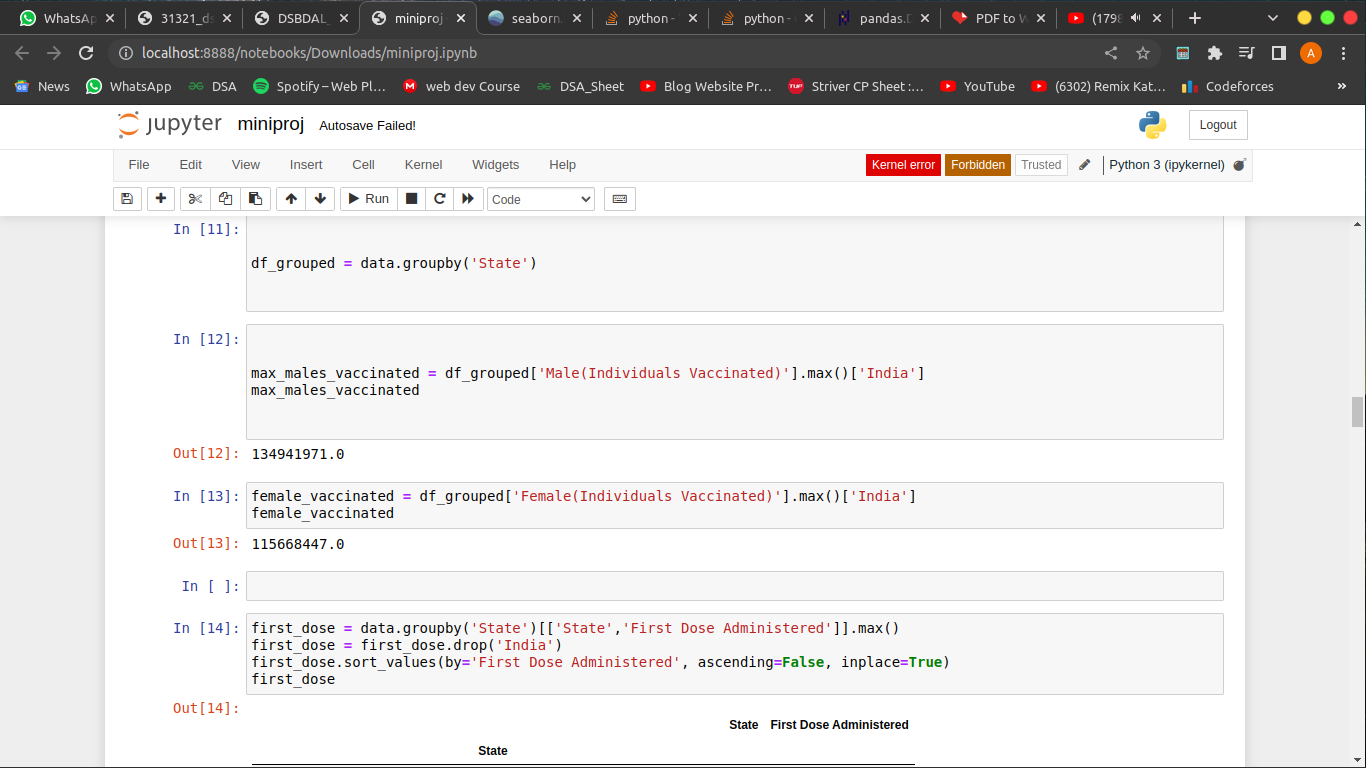
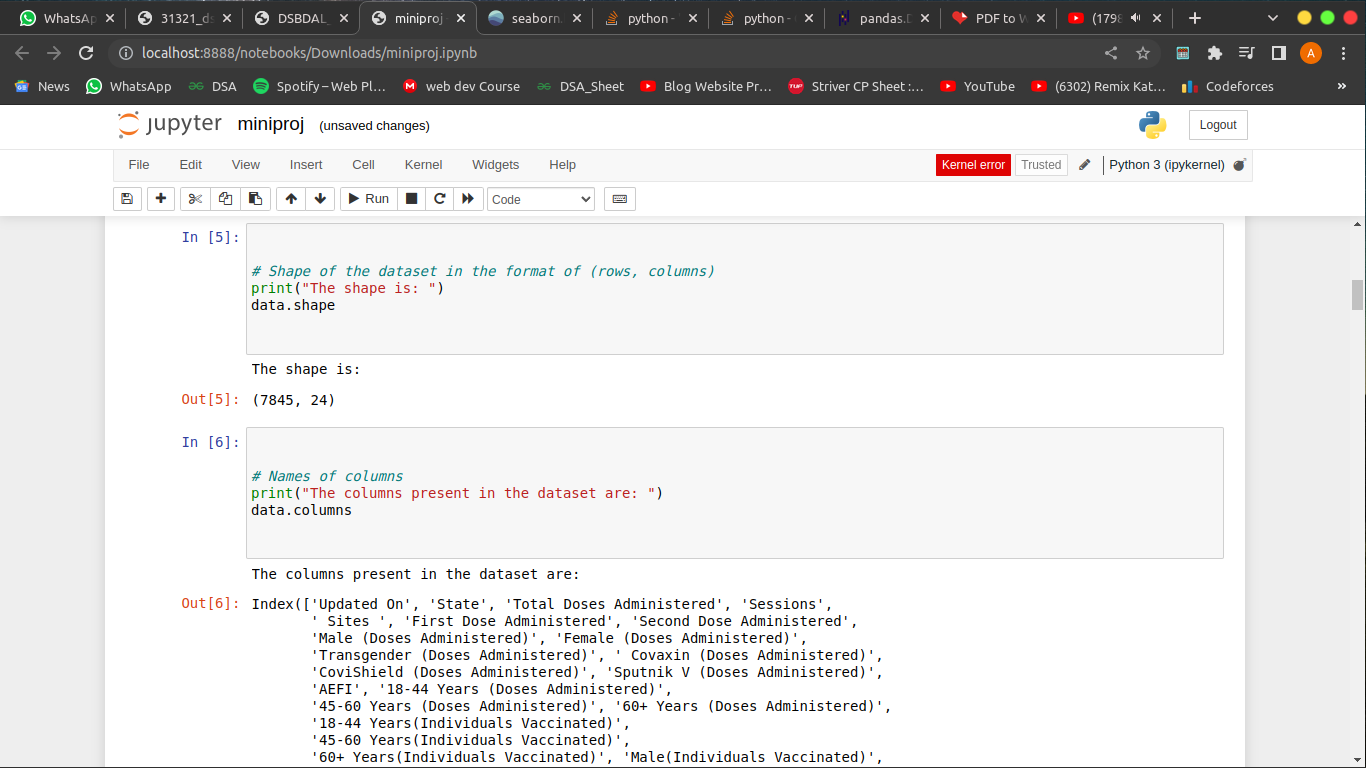
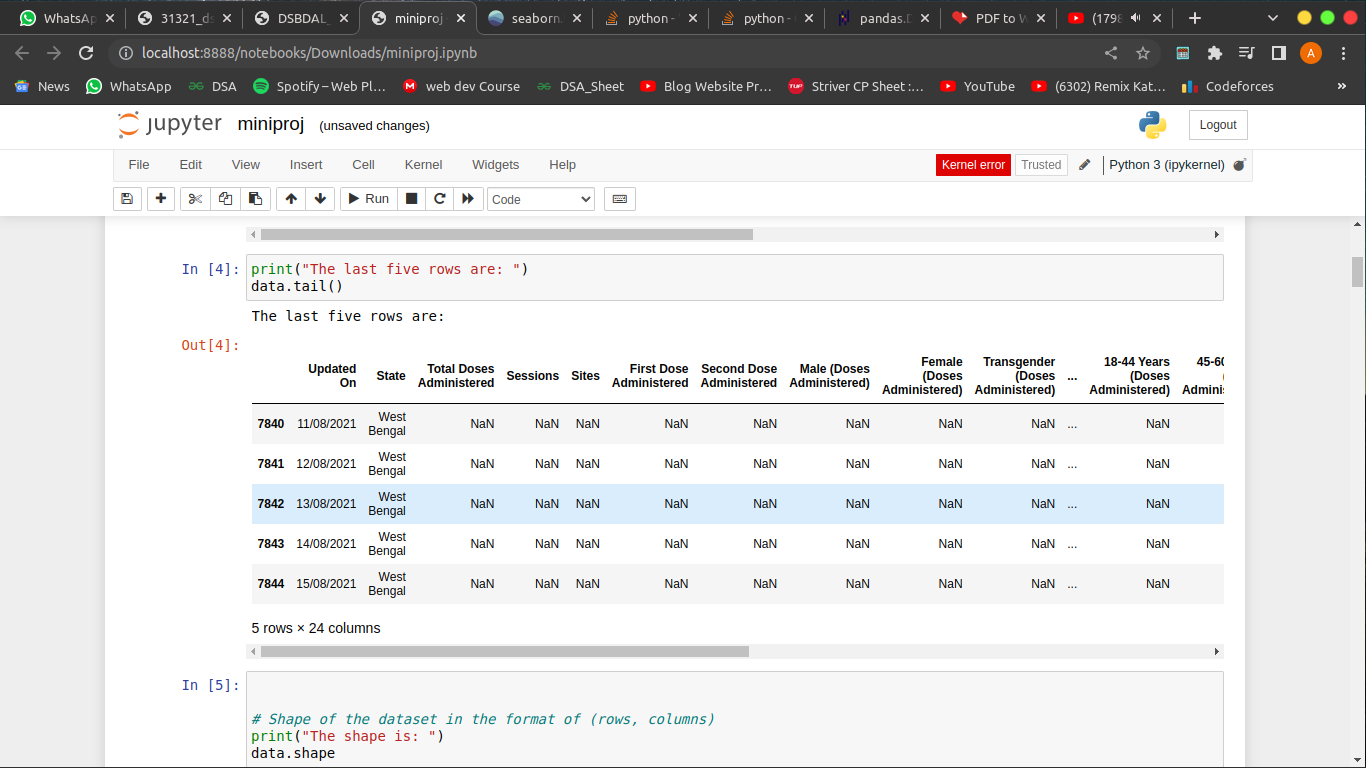
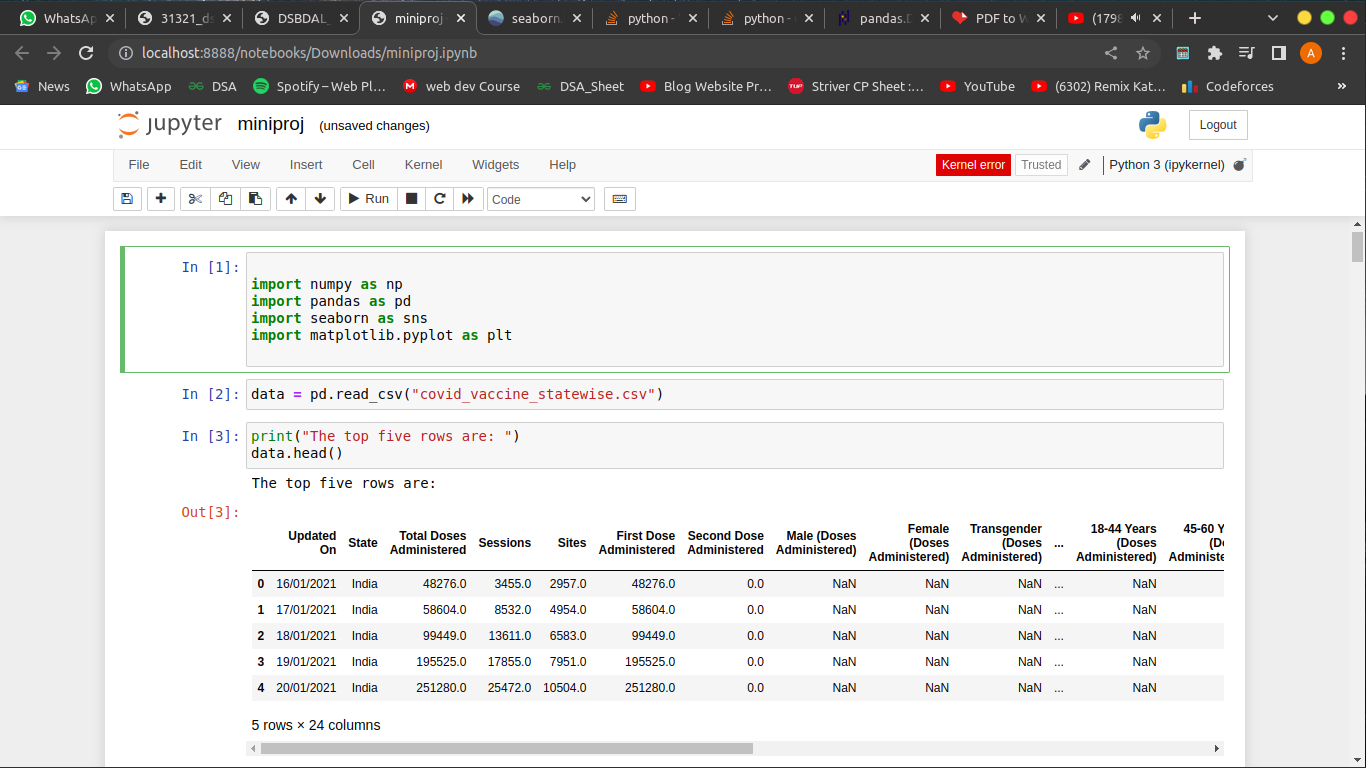
# Data visualization

After properly processing the data, we started creating visualizations to find a preliminary shape and structure as well s trends in the data. We initially created a graph of Total Doses Administered vs Date, then we created more useful and informative graphs such as Statewise doses administered (used lineplot to plot the data of all the states on the same graph for comparative analysis). We also plotted graphs to find the relation between ages and genders vs the number of vaccines administered by date.

# Data Analysis:

We used Scikit-learn’s inbuilt LinearRegression module to find relations between the independent variable (Date) and the dependent variable (Number of doses administered). We also added extra criteria such as the state, the age of the person, and the gender to find out how the analysis varied.

**Code & Output:**



# Future Scope:

In the future, we could add extra features such as analyzing the number of new cases and deaths that were incurred during this period and finding the relationship between the cases, deaths, and vaccinations to find how effective the vaccines were at preventing deadly or highly injurious cases as well as casualties

# Conclusion:

Thus, we successfully analyzed the available dataset about Covid Cases and vaccinations in India and were able to find how categorical values like State, Age, and Gender are related to the number of people vaccinated by Date.