**COVID-19 DATA ANALYSIS**

*Submitted in partial fulfillment of the requirements*

*for the award of the degree of*

**Bachelor of Computer Applications**

To

Guru Gobind Singh Indraprastha University, Delhi

**Submitted TO Submitted BY**

**Ms. Kavita Shrivastava Swati Taheem 07213702020**

Associate Professor (IT)  **Parshav Kumar 04813702020**

****

Institute of Information Technology & Management,

New Delhi – 110058 Batch

(2020-2023)

**Certificate**

I,(Name & Roll No) certify that the Summer Training Project Report (BCA-355) entitled “ ” is done by me and it is an authentic work carried out by me at (Name of the organization or of the Institute). The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

Signature of the Student Date:

Certified that the Project Report (BCA-355) entitled “ ” done by the above student is completed under my guidance.

Signature of the Guide: Date:

Name of the Guide: Designation:

Countersign Head of the Department Countersign Director

**Table Of Contents**

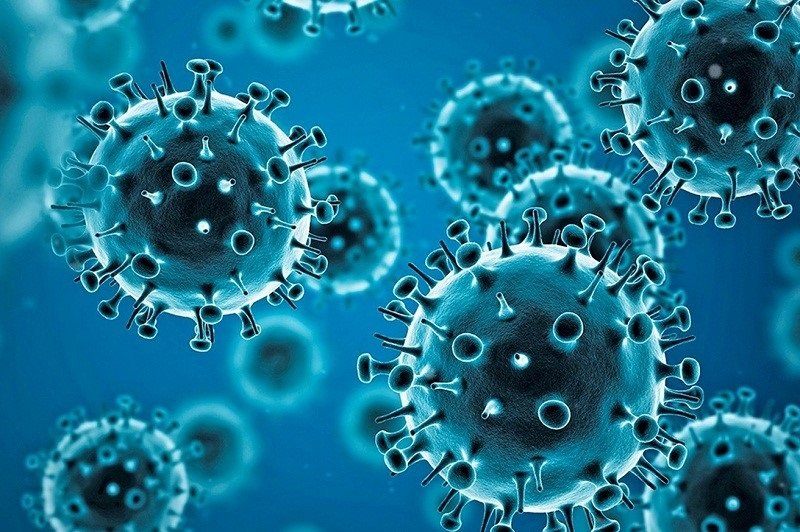
|  |  |  |
| --- | --- | --- |
| **S No** | **Topic** | **Page No** |
| 1. | Certificate | - |
| 2. | Acknowledgements | - |
| 3. | List of Tables/Figures/Symbols | - |
| 4. | Synopsis | I-IV |
| 5. | Chapter-1: Software Project Planning | 1-10 |
| 6. | 1.1 Description of the Software System under Study | 1 |
| 7. | 1.2 Data Collection | 2 |
| 8. | 1.3 Tools and Platform | 4 |
| 9. | 1.4 Project Planning | 4 |
| 10. | 1.5 Methodology | 6 |
| 11. | 1.5.1 SDLC Model | 6 |
| 12. | 1.5.2Justification for the Selection of Model | 7 |
| 13. | Chapter-2: Software Requirement Specification |  |
| 14. | 2.1 Description of Information System (Block Diagram) |  |
| 15. | 2.1.1 Product Features |  |
| 16. | 2.1.2 Input/ Output Data elements |  |
| 17. | 2.1.3 Procedures/ rules/ mathematical relationships |  |
| 18. | 2.1.4 Decision Table |  |
| 19. | 2.2 Use Case Diagram |  |
| 20. | Chapter-3: Software Project Analysis |  |
| 21. | 3.1 DFD |  |

**SYNOPSIS**

* **TITLE OF PROJECT**

**COVID-19 DATA ANALYSIS**

* **INTRODUCTION**

****

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus.

Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. Anyone can get sick with COVID-19 and become seriously ill or die at any age.

The best way to prevent and slow down transmission is to be well informed about the disease and how the virus spreads. Protect yourself and others from infection by staying at least 1 metre apart from others, wearing a properly fitted mask, and washing your hands or using an alcohol-based rub frequently. Get vaccinated when it’s your turn and follow local guidance.

The virus can spread from an infected person’s mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe. These particles range from larger respiratory droplets to smaller aerosols. It is important to practice respiratory etiquette, for example by coughing into a flexed elbow, and to stay home and self-isolate until you recover if you feel unwell.

* **OBJECTIVE**

Most of the prior research, media and projects coverage focused on the number of infections in the entire country. However, given the size and diversity of India, it is important to look at the spread of the disease in each state separately, wherein the situations are quite different.

In this project, we aim to analyse data on the number of infected people in each Indian state (restricted to only those states with enough data for prediction) and predict the number of infections for that state in the next 30 days. We hope that such statewise predictions would help the state governments better channelize their limited health care resources.

* **SCOPE**

The future scope of this analysis is that it will help in future that how accurate analysis of this kind of disease should be done. As there is huge data of the whole countries so how the predictions should be made and visualizations of the data should be done so that it will be understand by the citizens too. Project will also beneficial for the government that how in other ways data should be managed.

* **TOOLS AND PLATFORMS**
* **Hardware Components :**

Hard Disk: - 14.90GB

Processor: INTEL® Core™ i3-1005G1 CPU@ 1.20GHz….

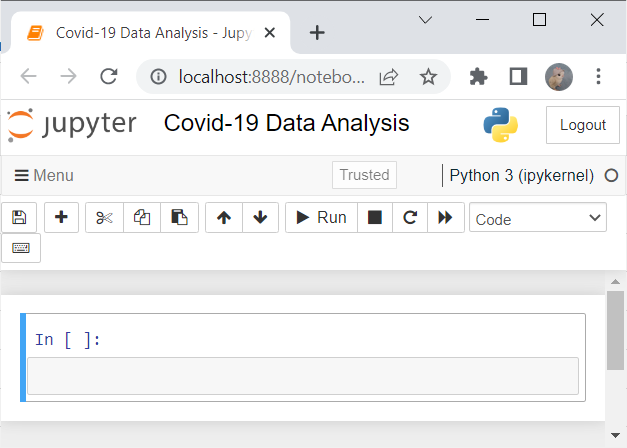
RAM: - 8 GB

Back end: - Python, Machine Learning

Front end: - ##

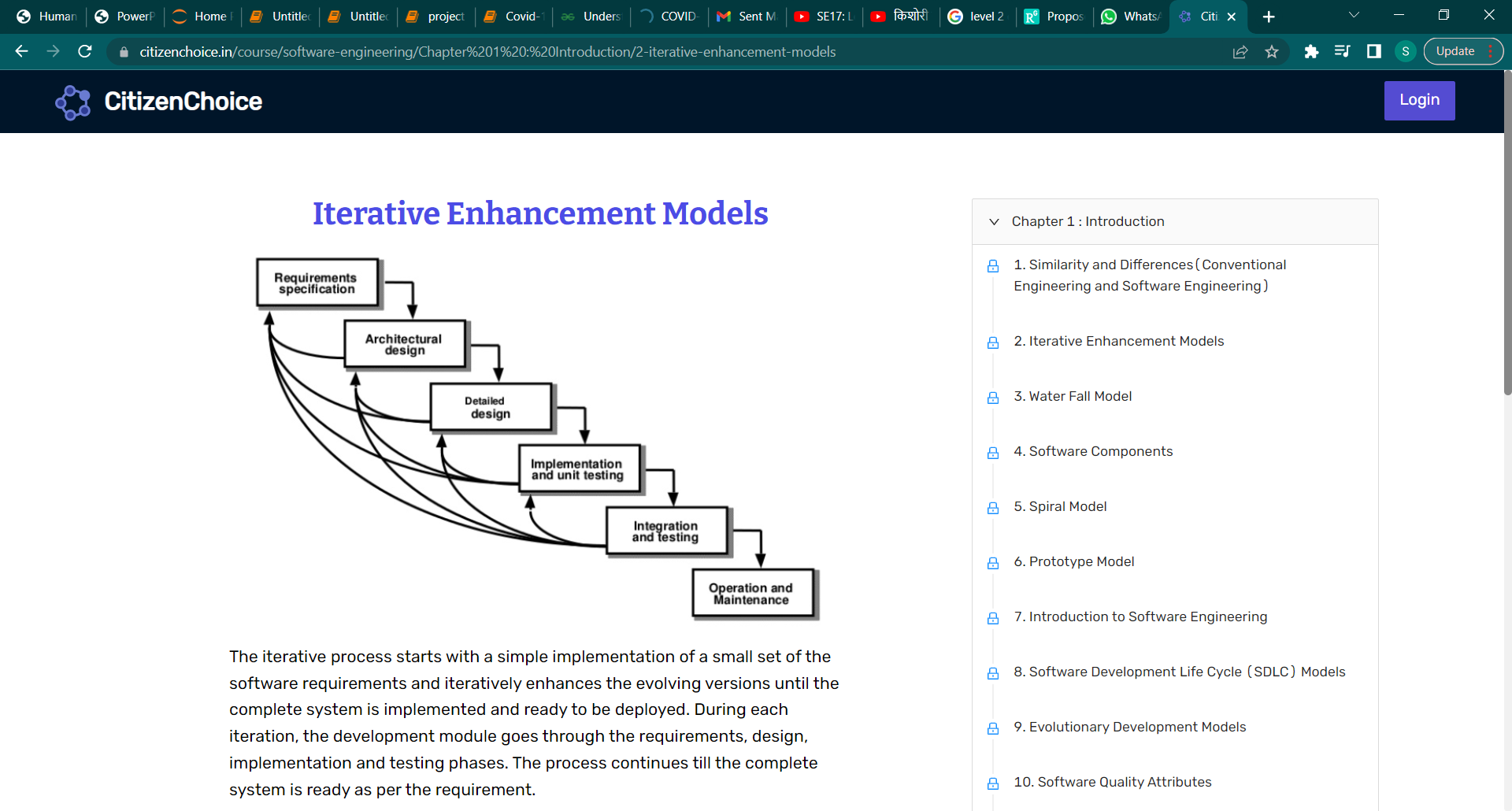
* **Software Specifications (Platform):**

Jupiter Notebook



* **METHODOLOGY:**
* Importing COVID19 dataset and preparing it for the analysis by dropping columns and aggregating rows.
* Deciding on and calculating a good measure for our analysis.
* Taking two datasets and finding correlations among our data and do visualization
* Visualizing our analysis results using Seaborn and Matplotlib.
* Then taking another dataset containing symptoms of covid-19 virus and using machine learning models to predict whether the person is corona positive or not.

**ITERATIVE ENHANCEMENT MODEL:**



The iterative process starts with a simple implementation of a small set of the software requirements and iteratively enhances the evolving versions until the complete system is implemented and ready to be deployed.

During each iteration, the development module goes through the requirements, design, implementation and testing phases. The process continues till the complete system is ready as per the requirement.

* **Main reason for using this model:**

Iterative enhancement model considers the changes made to the data are iterative in nature. This model incorporates changes in the project based on the analysis of the existing data. It assumes complete documentation of the project is available in the beginning.

* **ADVANTAGES FOR CHOOSING ITERATIVE ENHANCEMENT MODEL:**
* Results are obtained early and periodically.
* Parallel development can be planned.
* Progress can be measured.
* Less costly to change the scope/requirements.
* Testing and debugging during smaller iteration is easy.
* Risks are identified and resolved during iteration; and each iteration is an easily managed milestone.
* Easier to manage risk - High risk part is done first.
* With every increment, operational product is delivered.
* Issues, challenges and risks identified from each increment can be utilized/applied to the next increment.
* Risk analysis is better.

**CHAPTER-1**

**PROJECT DETAILS**

**(Covid-19 Data Analysis)**

1. **Description of the Project under Study:**
2. **Introduction**

COVID-19 is not just a name now. It has become a deadly widespread virus that has affected tens of thousands of people all over the world. Its origin was Wuhan City, China in Dec. 2019. When people were unaware of the virus, COVID-19 started to spread from one person to another; it has slowly reached almost all countries and has become a pandemic .

The project is related to covid-19 pandemic and includes the analysis of gathered data in which we are studying the analysis and visualisation of covid-19 datasets. Datasets are divided among the team members and then work accordingly.

1. **Division of Datasets**

* Country wise Covid cases till June 2022

-having attributes country ,total\_confirmed, total\_deaths, total\_recovery ,active cases, serious\_or\_critical

* State wise Covid cases of India till June 2022

-having attributes such as Name of state/UTs, Total Cases, Active Cases, Discharged and Deaths.

* Dataset for the prediction of Covid-19 based on symptoms

1. **Stages**
2. **First stage**

At first stage we took the dataset of the country wise covid cases having 227 countries then countries are arranged in alphabetical order so that we can easily visualize our data using python in a sorted manner.

After applying the code of python we are able to view our dataset in a alphabetical manner in which graphs are showing the country data with total\_confirmed cases and also with other attributes of the dataset.

For example: countries whose name starts with letter ‘A’ displays the graph with all other attributes of the dataset .

Similar with all alphabets

1. **Second Stage**

At this stage, we took the dataset for all the States/UTs In India having 29 states. Then all the states were organised in alphabetical order so that we can visualize the data effortlessly. After running our python code we were able to view our dataset in a graphical representation which includes :- “Number of Death in eaech State/UTs”, “Top 10 States with most active cases”, “Top 10 states with most Death cases”, Etc.

1. **Third Stage**

At this stage we are using the dataset having the symptoms for covid -19 and analyse the dataset and displays weather the person is having covid or not.(using machine learning models)

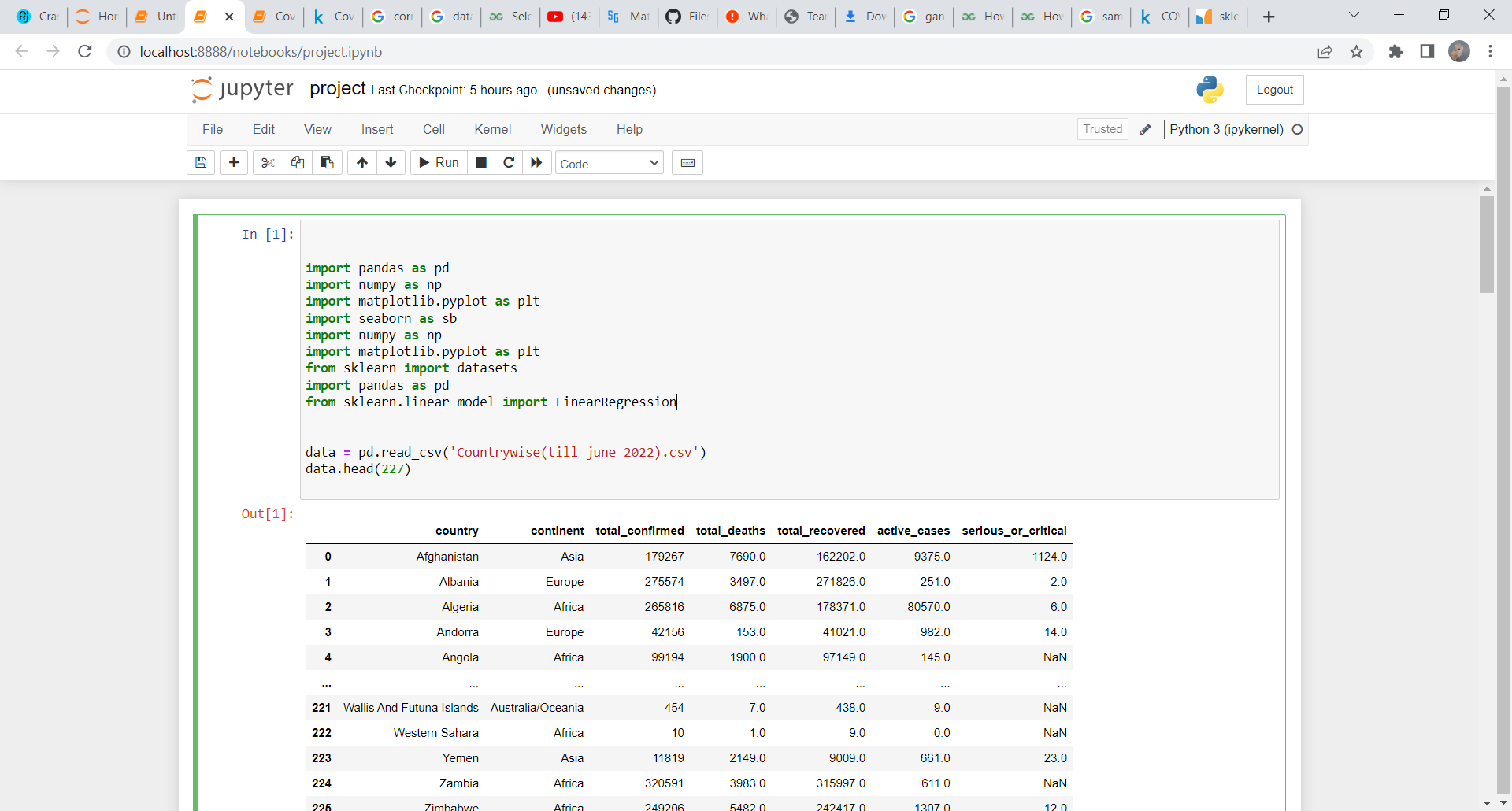
And apply the best prediction model which give accurate results.

1. **Data Collection And Modules**

Datasets are collected from the various sources on the internet most accurate data is provided by the WHO(World Health Organization) and other health government sites.

* 1. **Modules**

Various modules for data visualisation ,calculations and prediction model



1. **TOOLS AND PLATFORMS**
2. **Hardware Components :**

Hard Disk: - 14.90GB

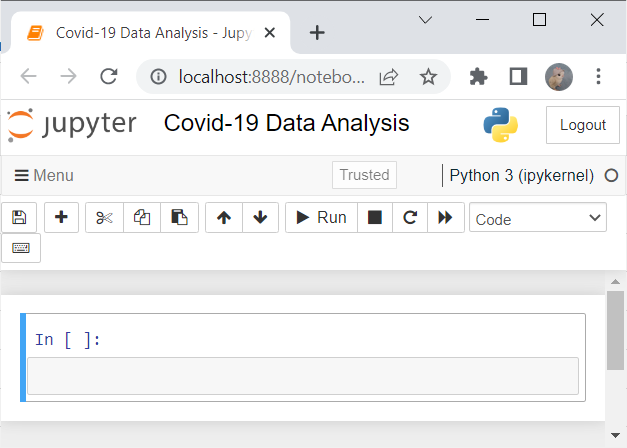
RAM: - 8 GB

Back end: - Python, Machine Learning

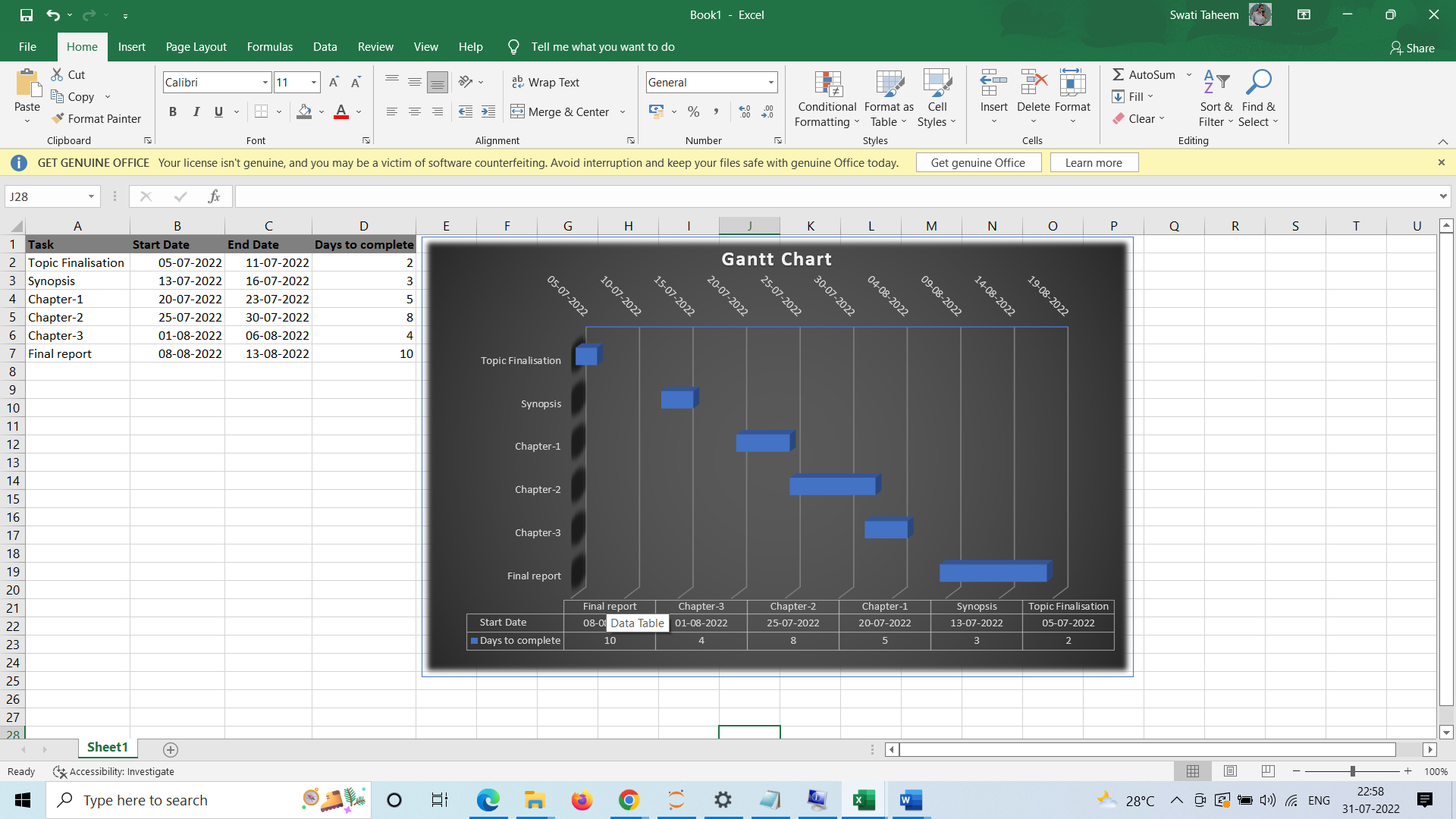
Front end: - ##

1. **Software Specifications (Platform):**

Jupiter Notebook



1. **Project Planning**



**CHAPTER-2**

**Software Requirement Specification (SRS)**

**(Covid-19 Data Analysis)**

1. **Description of Information System (Block Diagram):**

**At Visualization Phase:**

Covid-19 Datasets

Visualization

On Country Basis

On World Basis

Fig. 1.1

Above Fig. 1.1 is just a simple block diagram to show that what is in the visualization phase of the project we first take the DataSets , First DataSet is on the Country basis which means the Dataset containing the covid related information of our Country **INDIA** having attributes States/UTs, Total Cases, Active, Discharged, Deaths.

No. of Total Rows=36

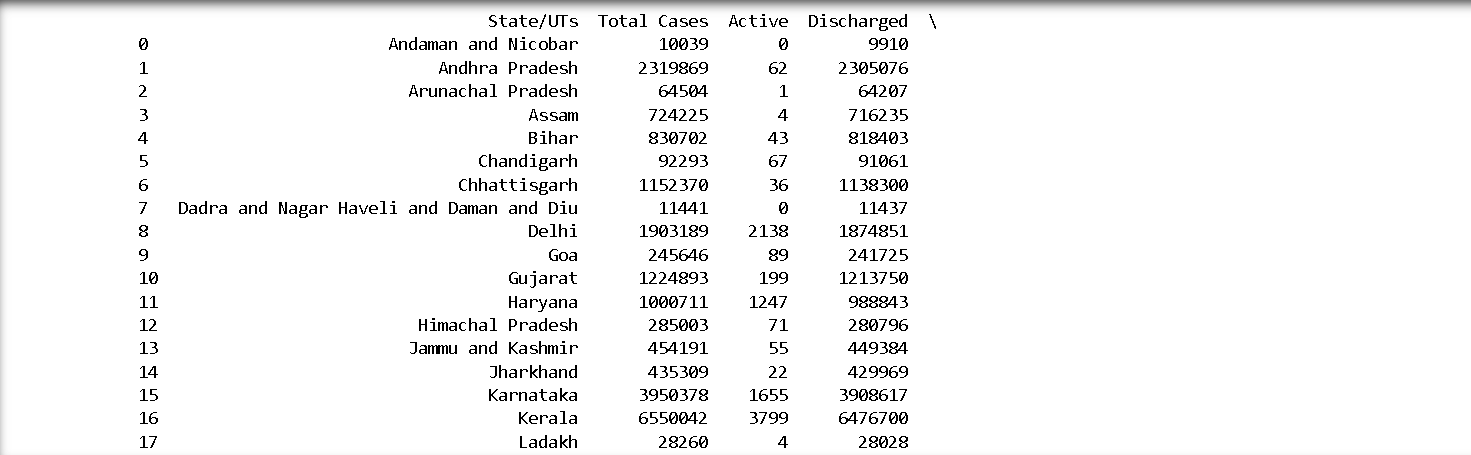
Second Dataset is on the World Basis means this dataset is containing the covid related information of all countries in the world having attributes as country, total\_confirmed, total\_deaths, total\_recovered, active cases, serious\_or\_critical.

No. of Total Rows =226

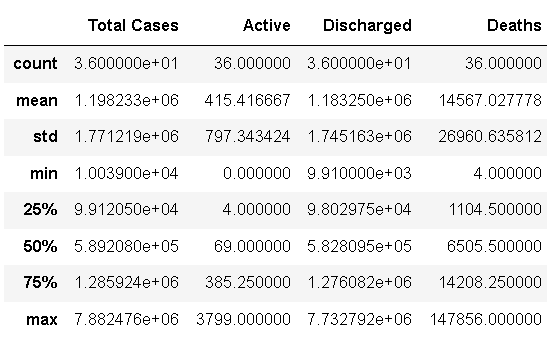
As the representation and visualization of 226 products (countries) with any of the products becomes chaotic which is not clearly readable at all. So, to improve the visualization and for better understanding of the data we represent the data in Alphabetically manner. By using this trick, we better visualize the data and also readability increases.

* **On Country Basis**

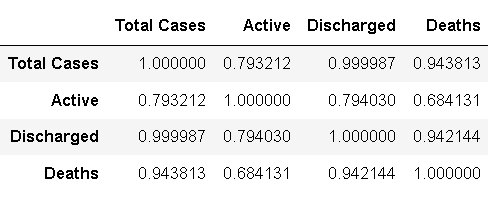
**Product features:-**In this module attributes used were:-States/UT, Active, Deaths, Discharged and Total cases. In the image given below, displays that dataset has been successfully read and the libraries used to import csv file are **“numpy”**, **“matplotlib”, “seaborn ” and “pandas”**

 **IMAGE-1**

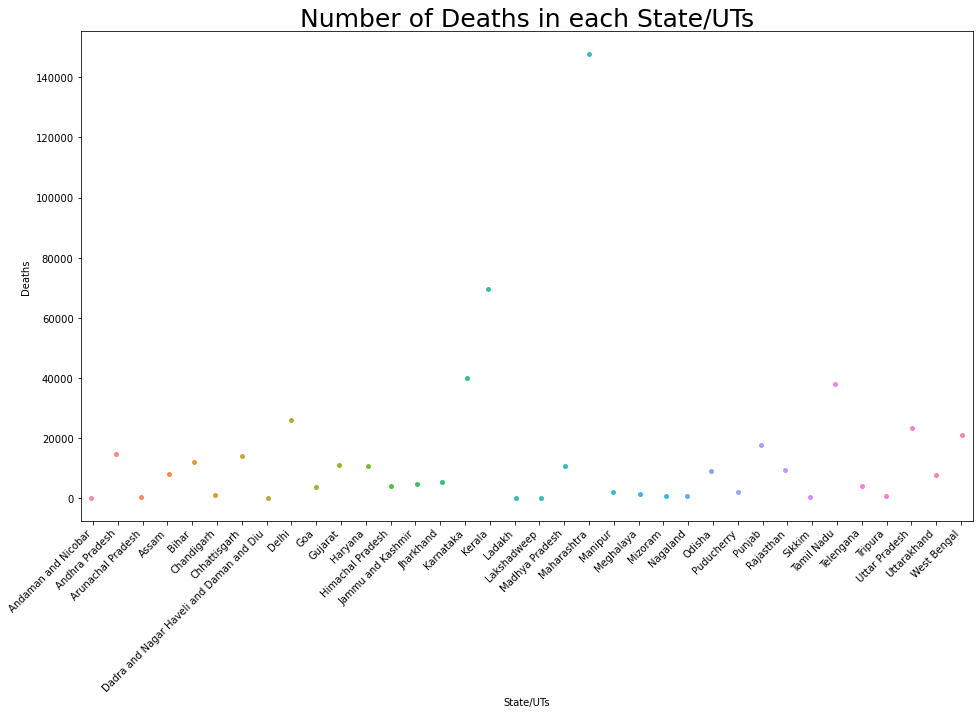
Now, from the loaded data set we will visualize the data in form of graphs and tables :-

Firstly, we have plotted a table describing the data frame with statistical manner i.e. displaying dataset through “mean”, “std”, “max”, etc.

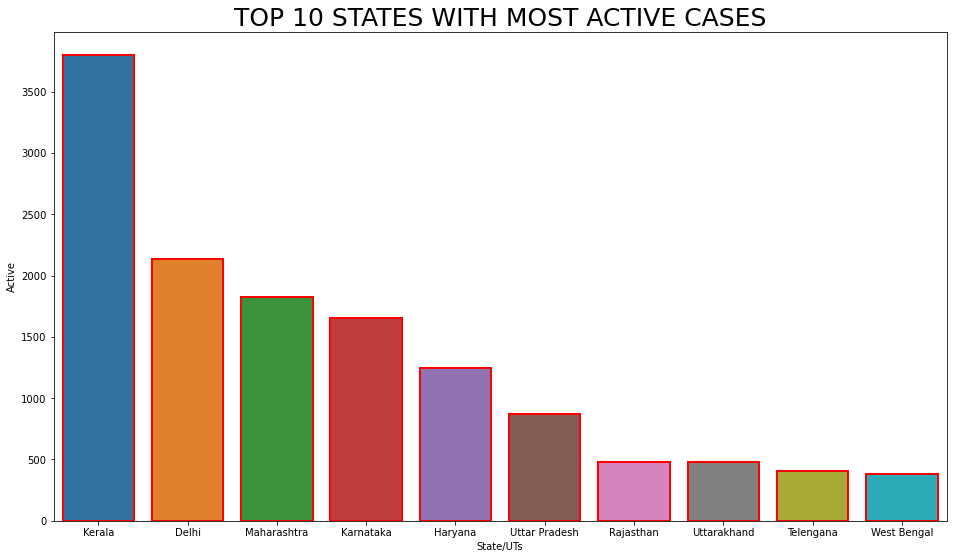
**IMAGE-2**

At second place a table has been plotted which computes pairwise correlation of columns. The function used to plot correlation is “ourdata.corr”.

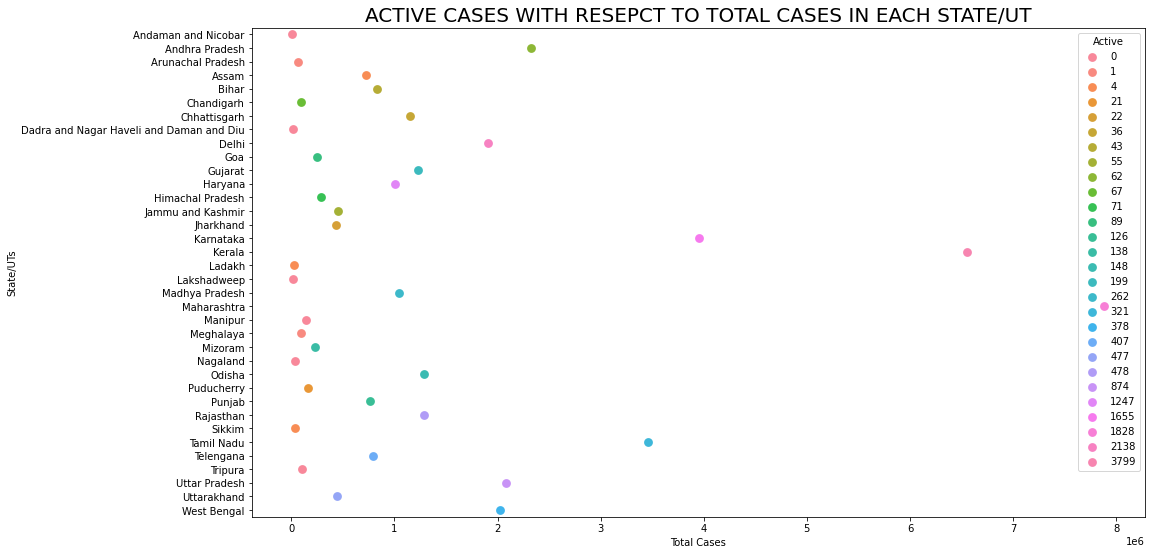
**IMAGE-3**

At third place a strip plot has been plotted displaying the “NUMBER OF DEATHS IN EACH STATE/UTs”. Strip plot is statistical representation which is easy to understand and makes our data colourful and attractive. From the given graph we can exclaim that highest number of people deaceased were in Maharashtra.

**IMAGE-4**

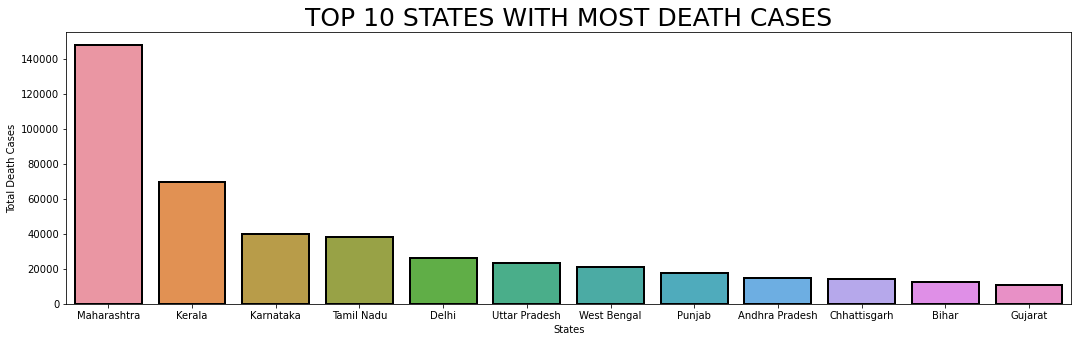
Next, the bar graph shown below depicts the top 10 states with most active cases. Function used for creating bar graph is “ourdata.barplot()”. From the given graph we can say that most active cases are present in “Kerela” and least number of active cases are available in “West Bengal”.

**IMAGE-5**

The given bar graph displays the “active cases with respect to total cases In each state/UTs”. Graphical representation used here is point plot which is used to show a data set in the form of points plotted on a graph and the function used to plot the graph is “ourdata.pointplot()”. Through the given graph we can exclaim that the most number of active cases are present in Maharashtra.

**IMAGE-6**

Lastly a bar graph has been plotted displaying the “Top 10 States/UTs with most number of Deaths”.

The bar graph had been created using “ourdata.barplot()” function. Through the given bar graph we can exclaim that the most number of deaths tool place in Maharashtra and least number of deaths took place in Gujarat.

**IMAGE-7**

* **On World Basis:**

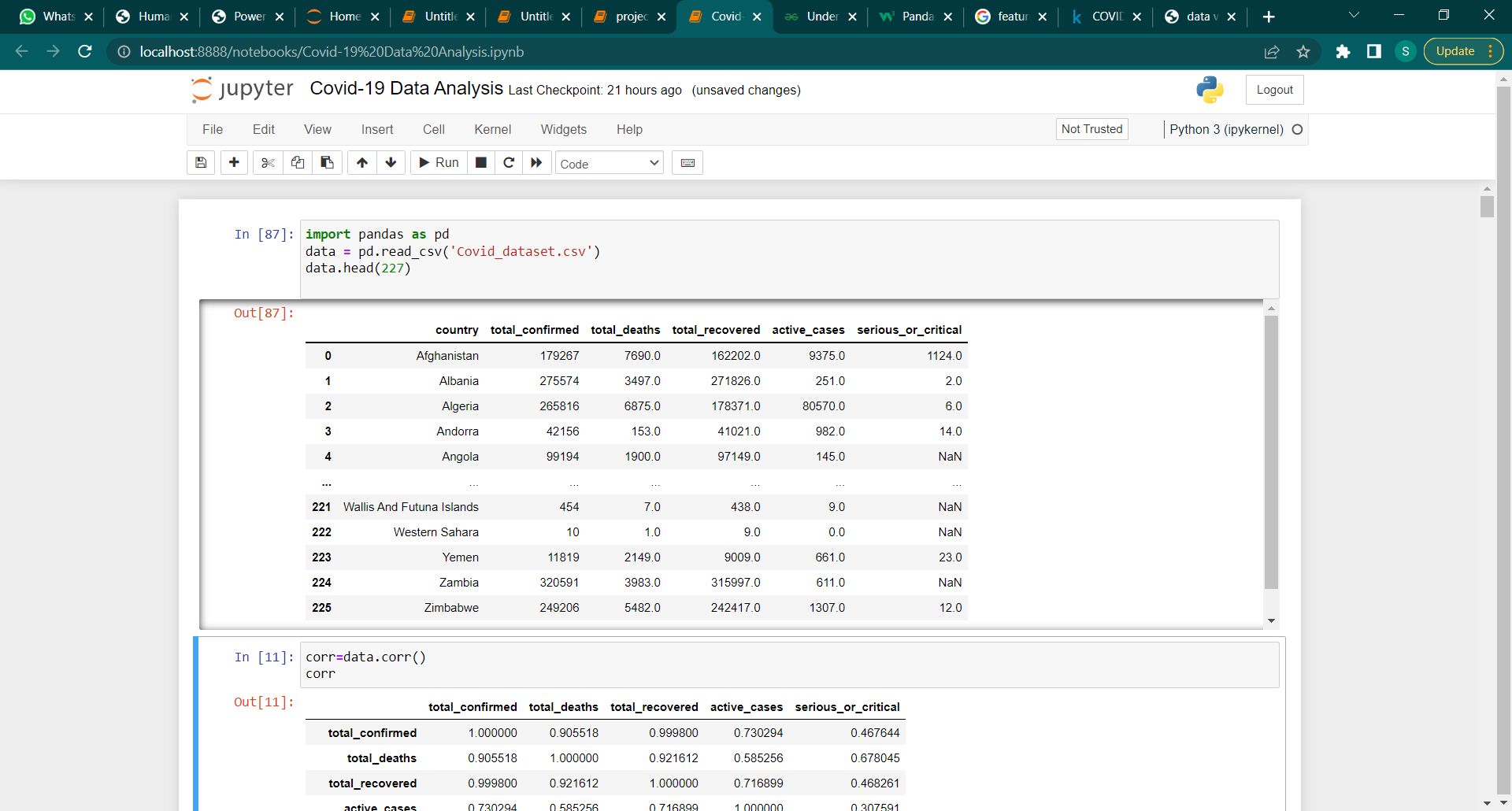
Dataset (in csv file) contains following data which is displayed as:

At very first step we need to analyse the dataset then to import all necessary libraries and modules.

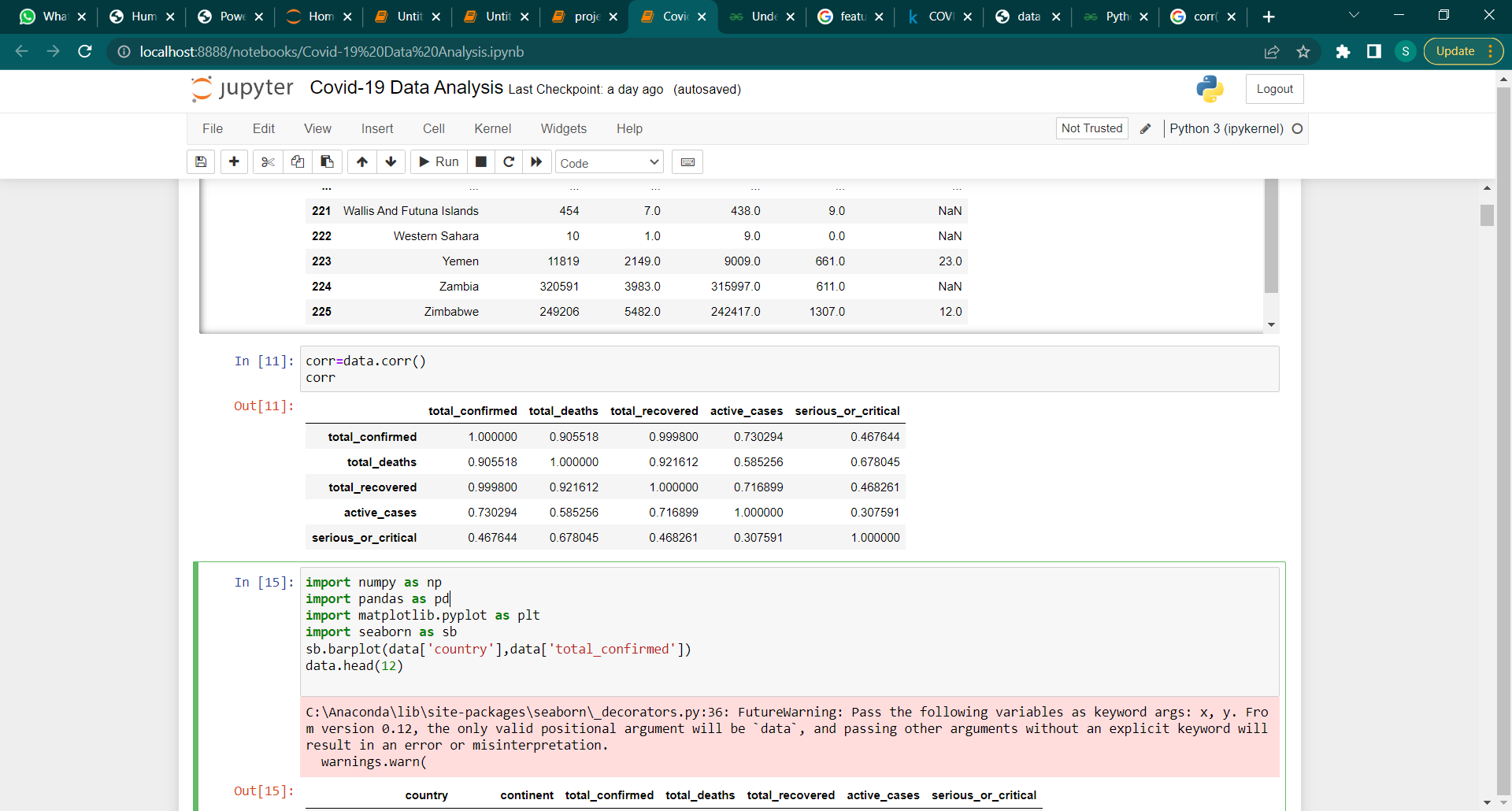
Below is just the representation, for this we import pandas for loading our dataset.

The read\_csv is used to load a CSV file as a pandas dataframe.

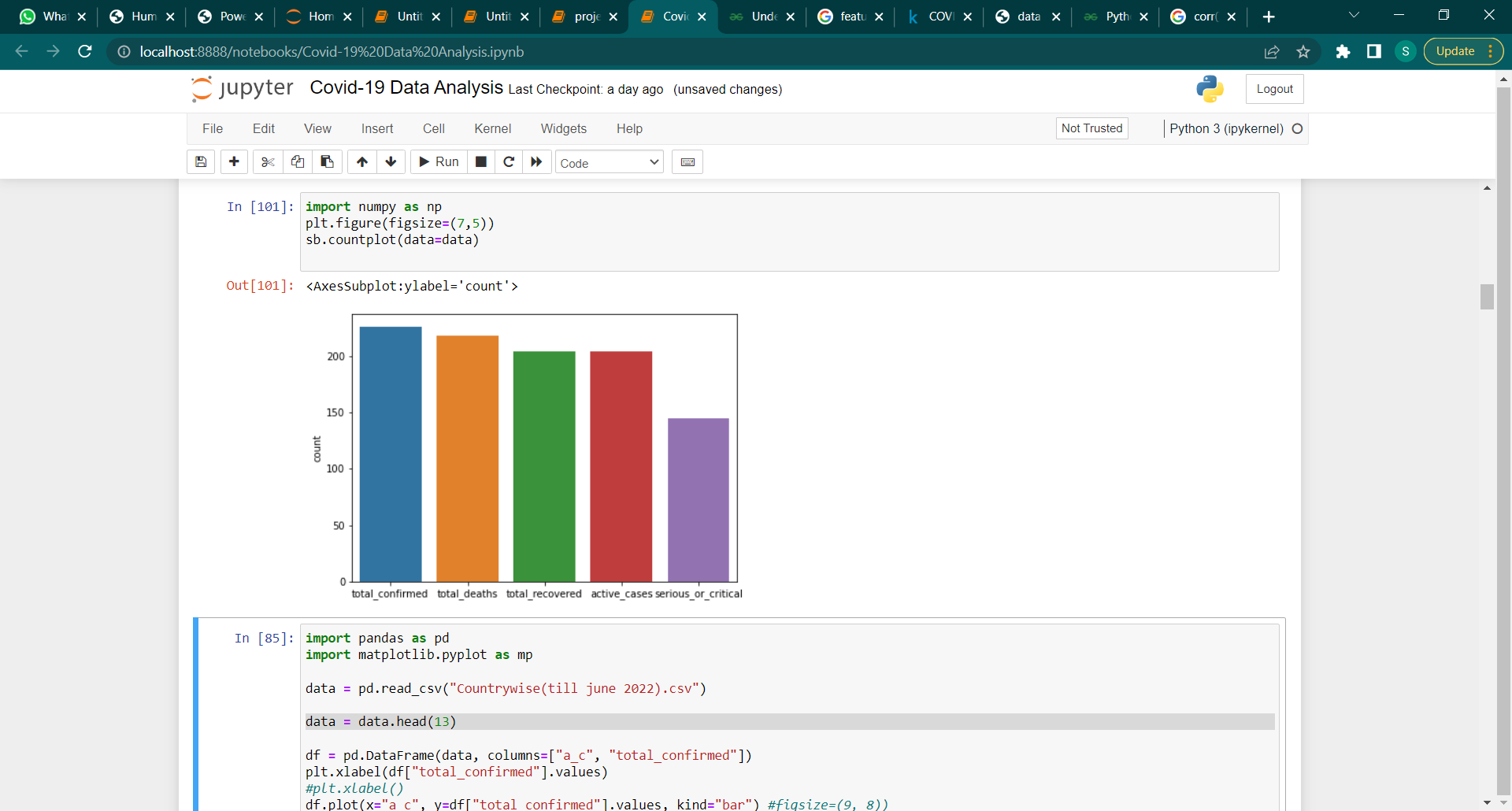
Data.head () function is used to access the first n rows of a dataframe.



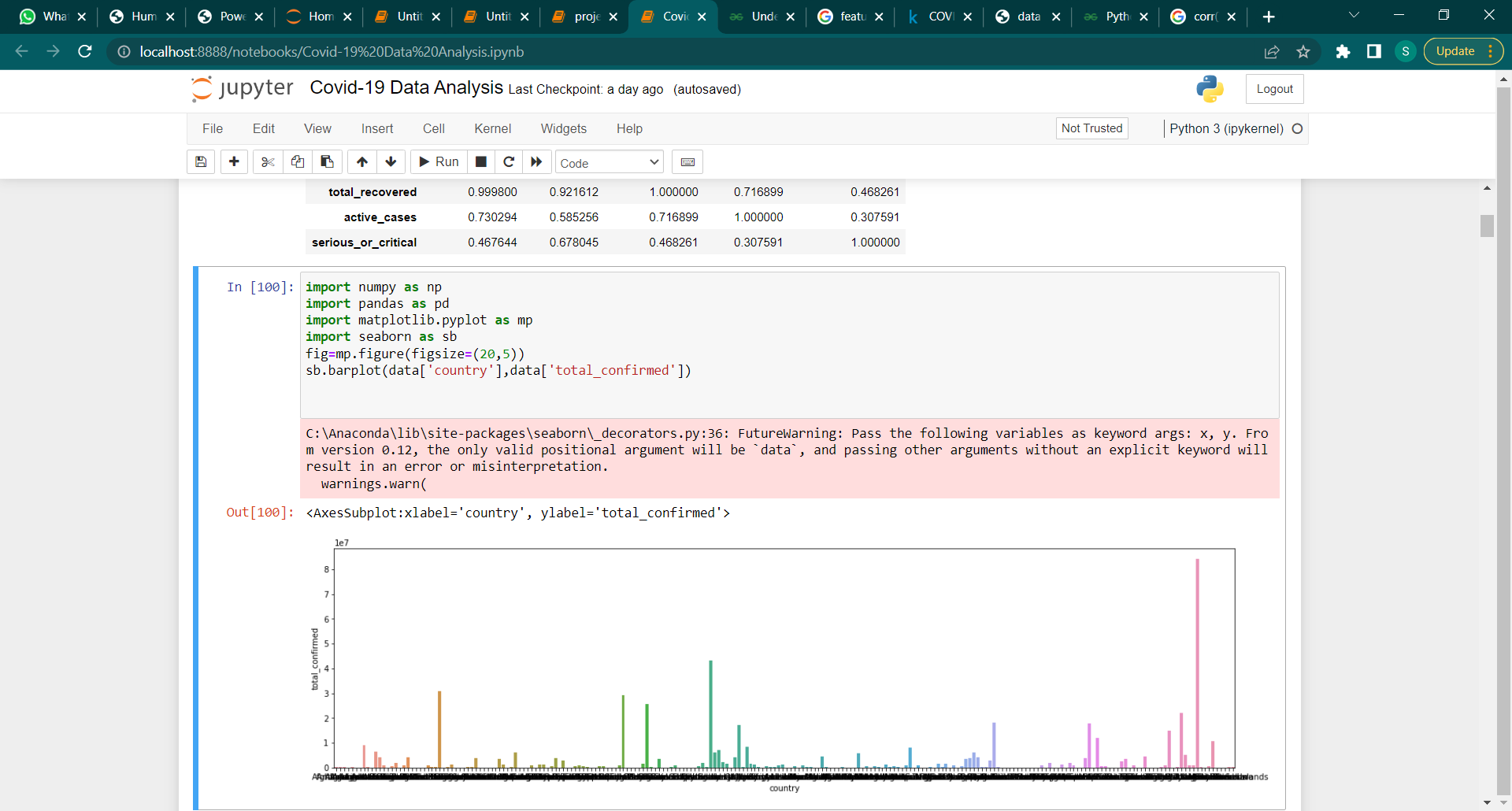
Following is the correlation of the dataframe:



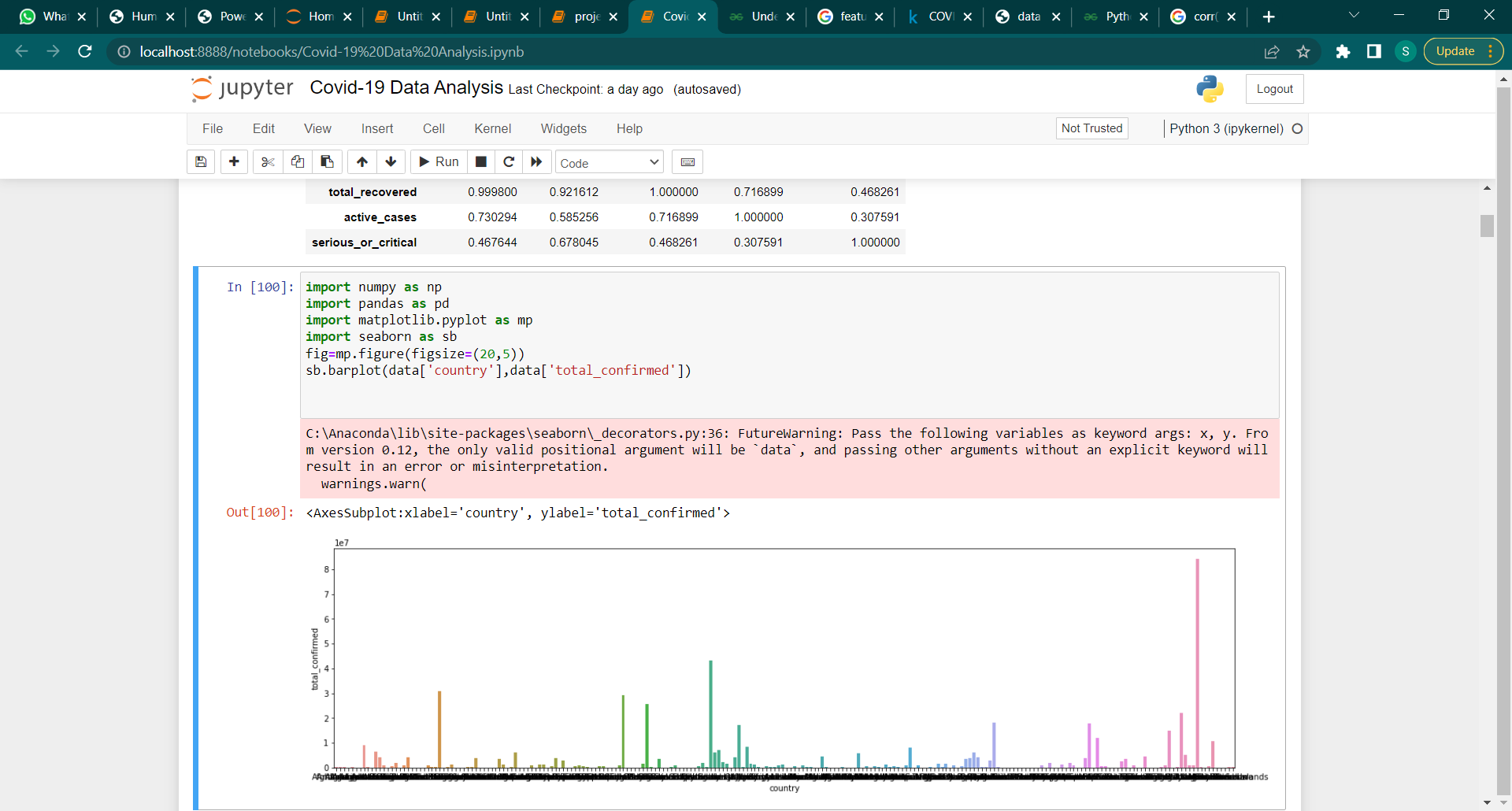
Below is the count plot of the dataset:



Now we try to visualize the above data using barpplot:



As we can see that visualization is not clear and understandable at all so we use python loops (conditional statements), Pandas, NumPy, Matplotlib and Seaborn to achieve the desired visualization.



Basically the main product in this is the country from which we extract the data that we want.

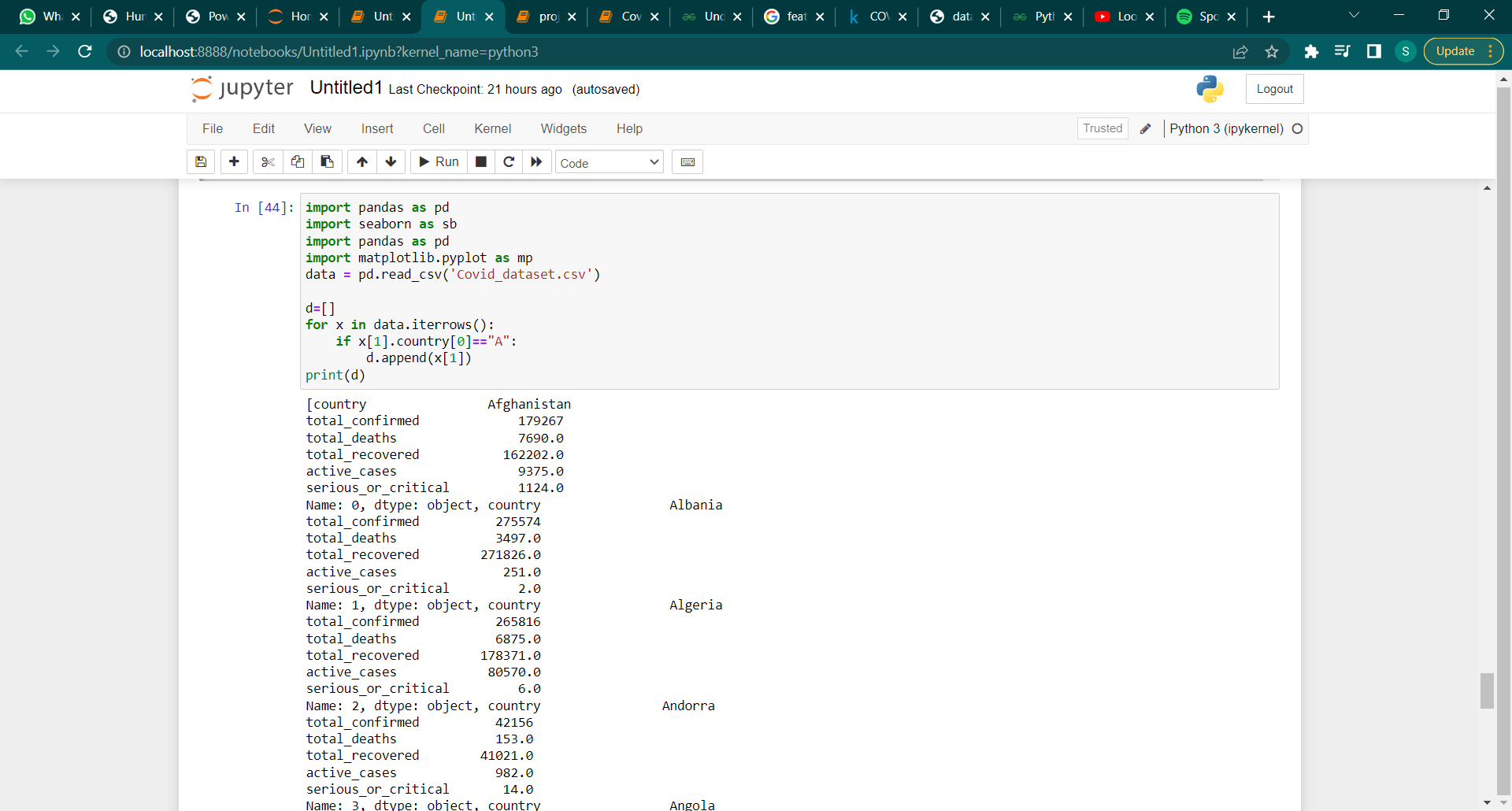
So we want to extract the country names based on alphabetical order for example, we want to visualize only in between the label country with the total\_confirmed cases where country name starts only with particular alphabet (“A”, “B”, “C”, ….. , “Z”).

At first step we write the code that gives us the country name starting with given alphabet.

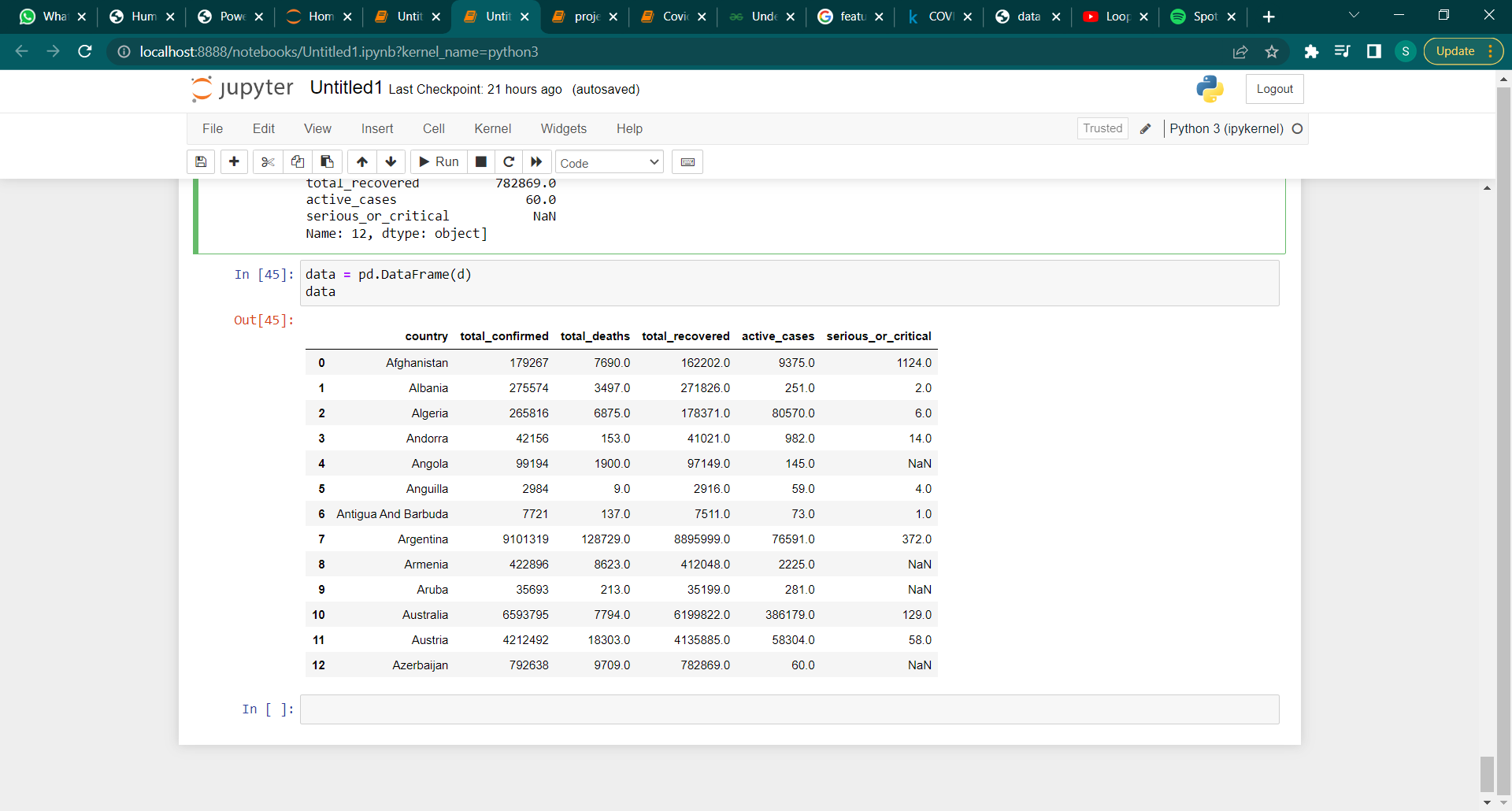
For this we first read the csv file then take the empty list for storing the values and then using for loop and iterrow() method (allowing us to iterate each row in the DataFrame) we proceed, then print.

Now we see that it gives the country names whose name only starts with given alphabet = “A”.

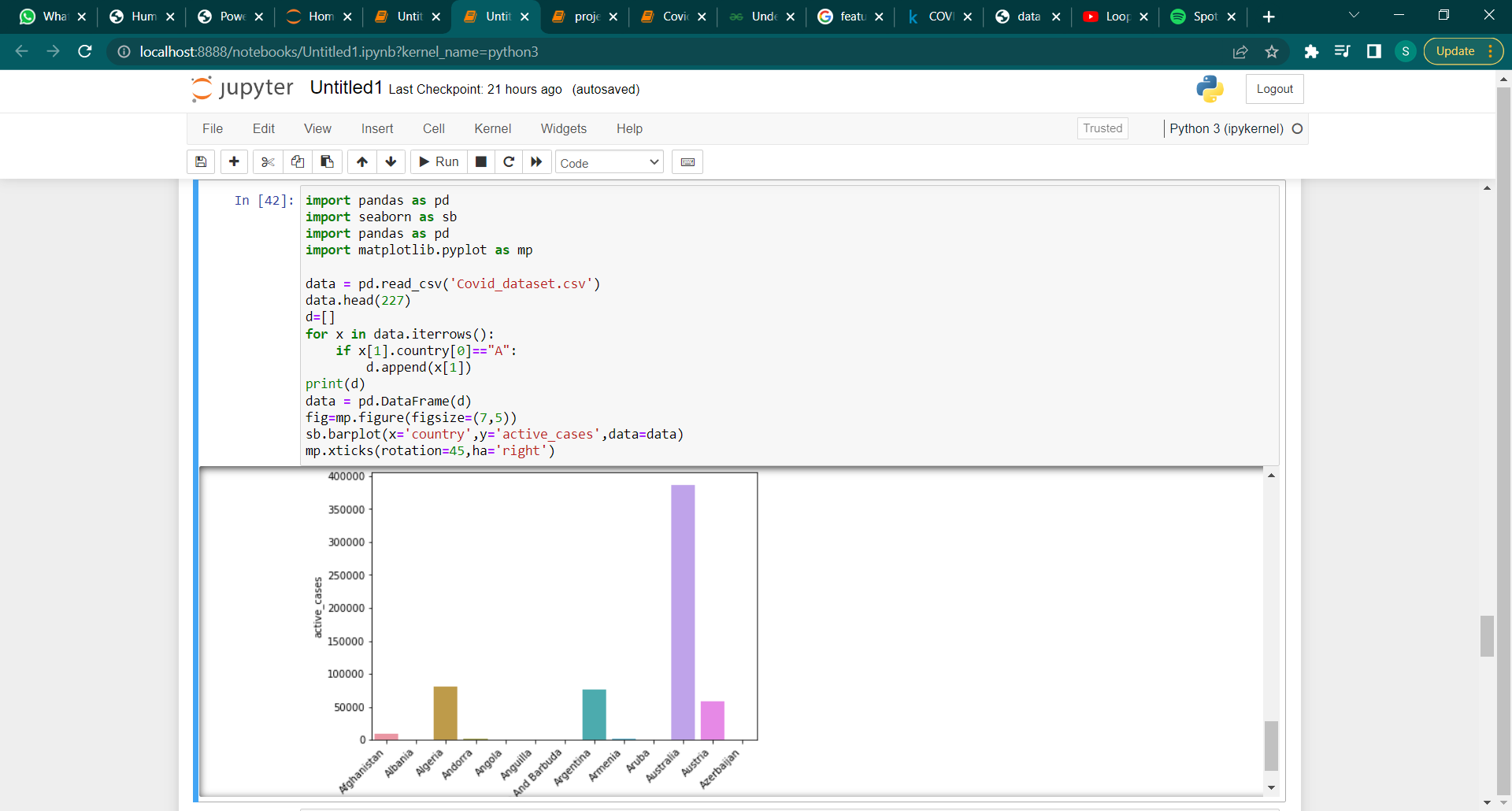
Similarly, we do the same thing for extracting other country names begin with remaining Alphabets.



Below is the dataframe which only contains the data of the country names whose name starts with alphabet A

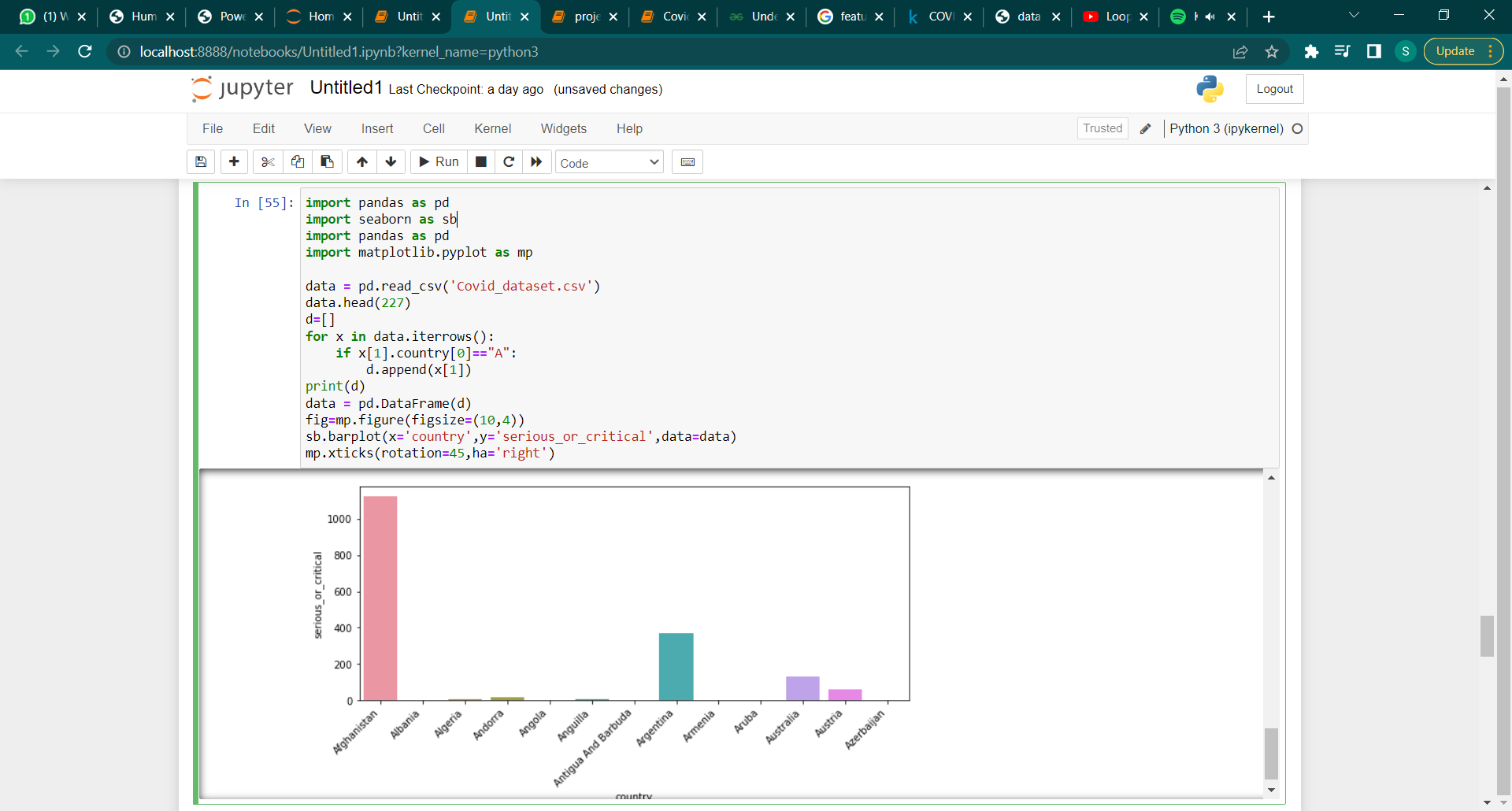


Now we visualize the barplot between the country and active\_cases of the above dataframe.



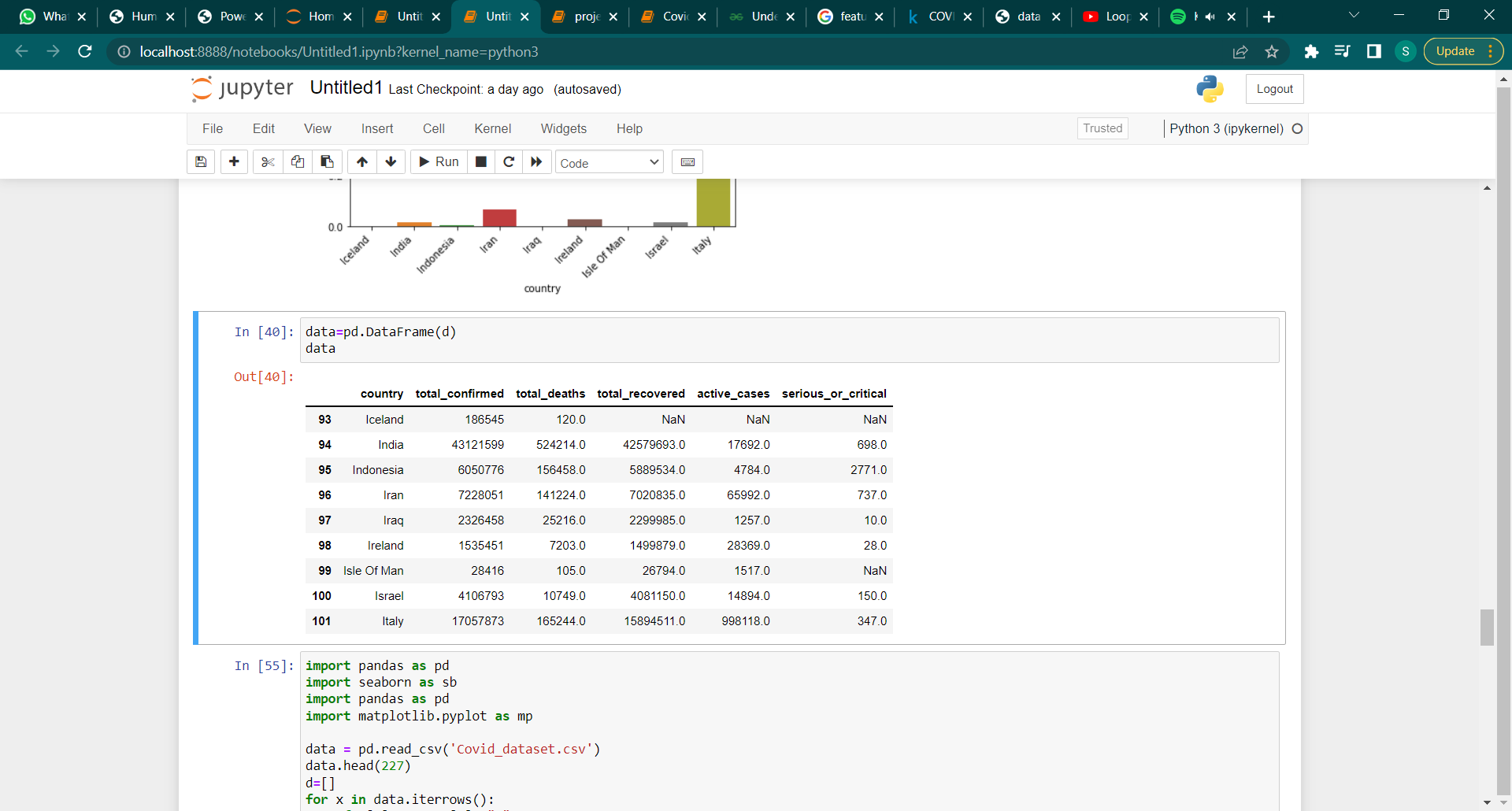
Hence we get our desired output now we can easily read and understand the visualization

Second we visualize the barplot between the country and seious\_or\_critical of the above dataframe.

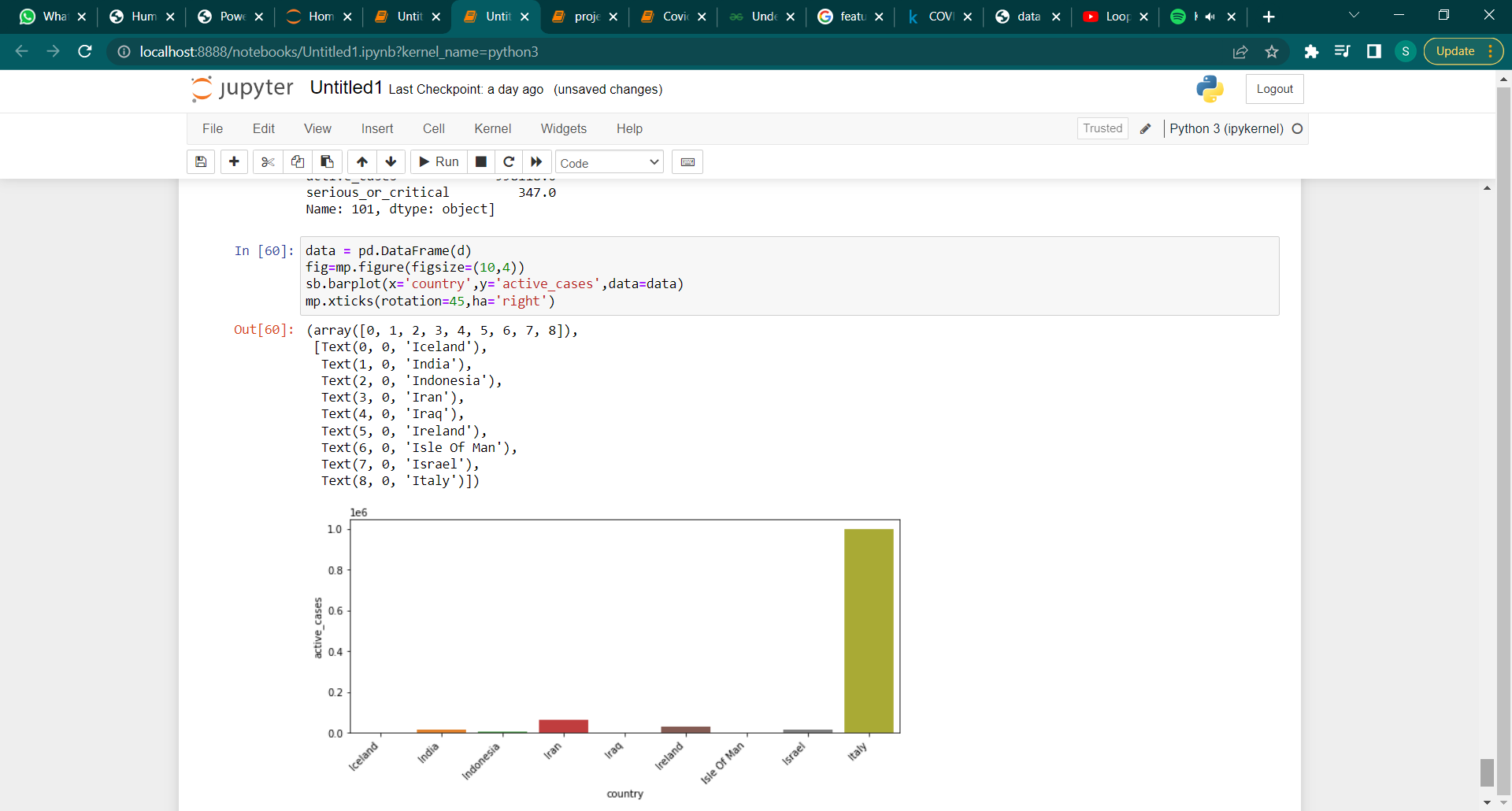


Let’s do the same for the alphabet = “I”

Below is the dataframe which only contains the data of the country names whose name starts with alphabet I



Now we visualize the barplot between the country and active\_cases of the above dataframe.



In the above ways we can extract the particular countries and visualize their correlated data of the dataframe.

**At Prediction Phase:**

Data Collection

Spliting data into training and testing

Data Pre-processing

Logistic Regression

Performance Evaluation

Classification

Statistical Model

Machine Learning

Multivariable Analysis

* **TRAINING AND TESTING PART**

After the extraction of features and labels then its, time to train the machine. So we have just divided the datasets into two parts, Training and Testing Data.

So for dividing the datasets into train and test data we used the module “sklearn.model\_selection” from this module we have imported the “train\_test\_split”.

The syntax that we used for this spliting are given here:

**x\_train, x\_test, y\_train, y\_test = train\_test\_split(x\_data, y\_data, test\_size=0.2, random\_state=7)**

So, here x\_train, x\_test, y\_train , y\_test are the variable that we used for taking some portions of datasets. Here test\_size denotes the amount of data that need to be stored in x\_test and y\_test.

After split up the data we use training data i.e x\_train and y\_train variables in the model and then we test the model by using x\_test and y\_test variables.

Applying different models to find out the best suitable model:

**Random Forest Model:**

***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

**Reason not to use this model for this:**

Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

**KNN( K Nearest Neighbour) Model:**

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

It can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

**Reason not to use this model for this:**

Always needs to determine the value of K which may be complex some time.

The computation cost is high because of calculating the distance between the data points for all the training samples.

**Support Vector Machine Model:**

It is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

**Reason not to use this model for this:**

SVM is a more complex model (non-linear model) than logistic regression (linear model). It may provide more accuracy, but may suffer from overfitting.

SVM will not work well for large amount of data (say a million data points).

**Logistic Regression Model:**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

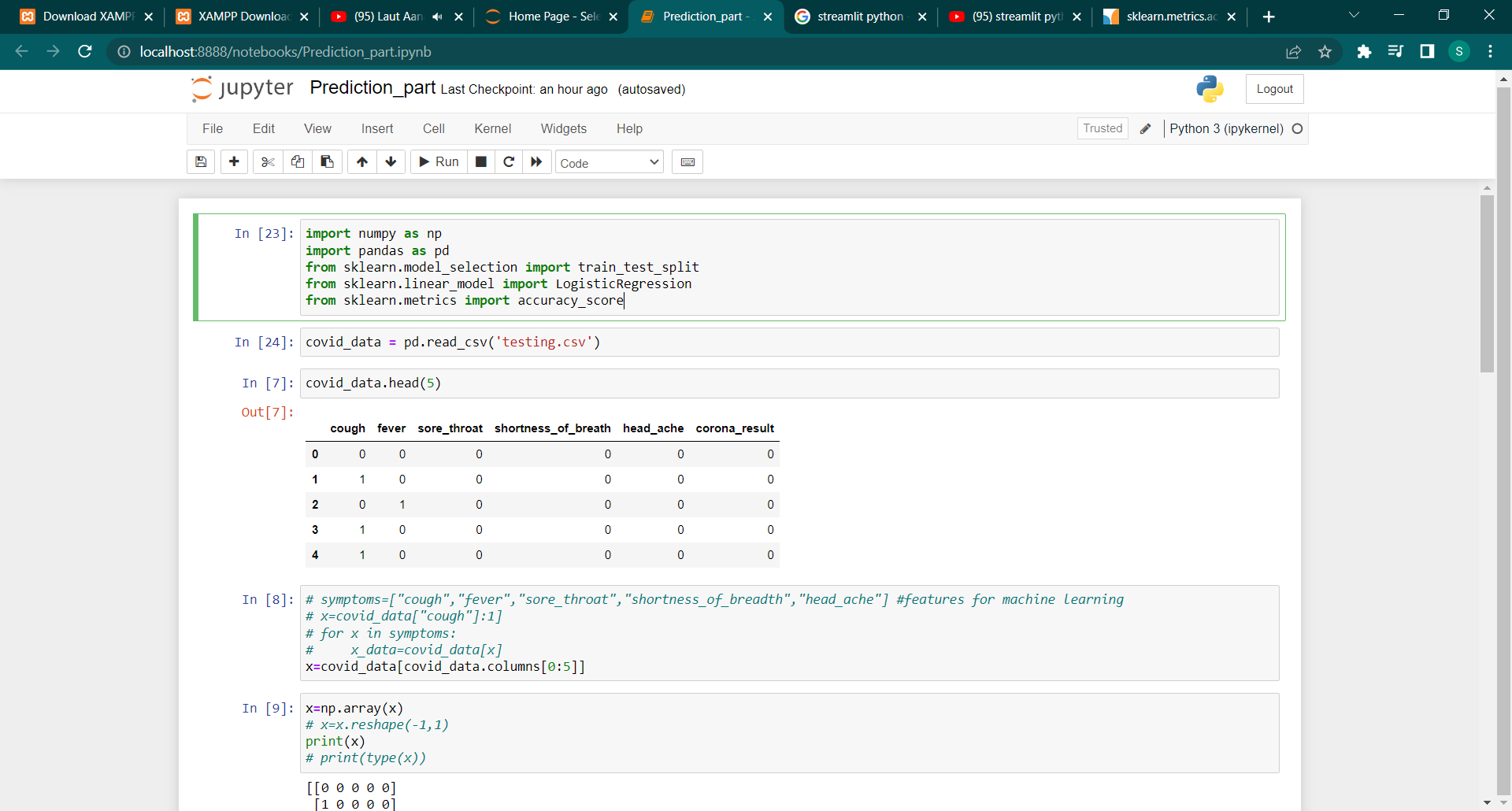
Now if compare the all the models with their measurements than we will find out that Logistic Regression model is better than other all models.

NOTE: As you can choose any one of them both the models (SVM and LOGISTIC REGRESSION) work fine and giving you the better output.

Therefore we choosing the Logistic Regression for further using this model.

Now discuss how to apply the model and predict that whether the person is suffering from covid or not.

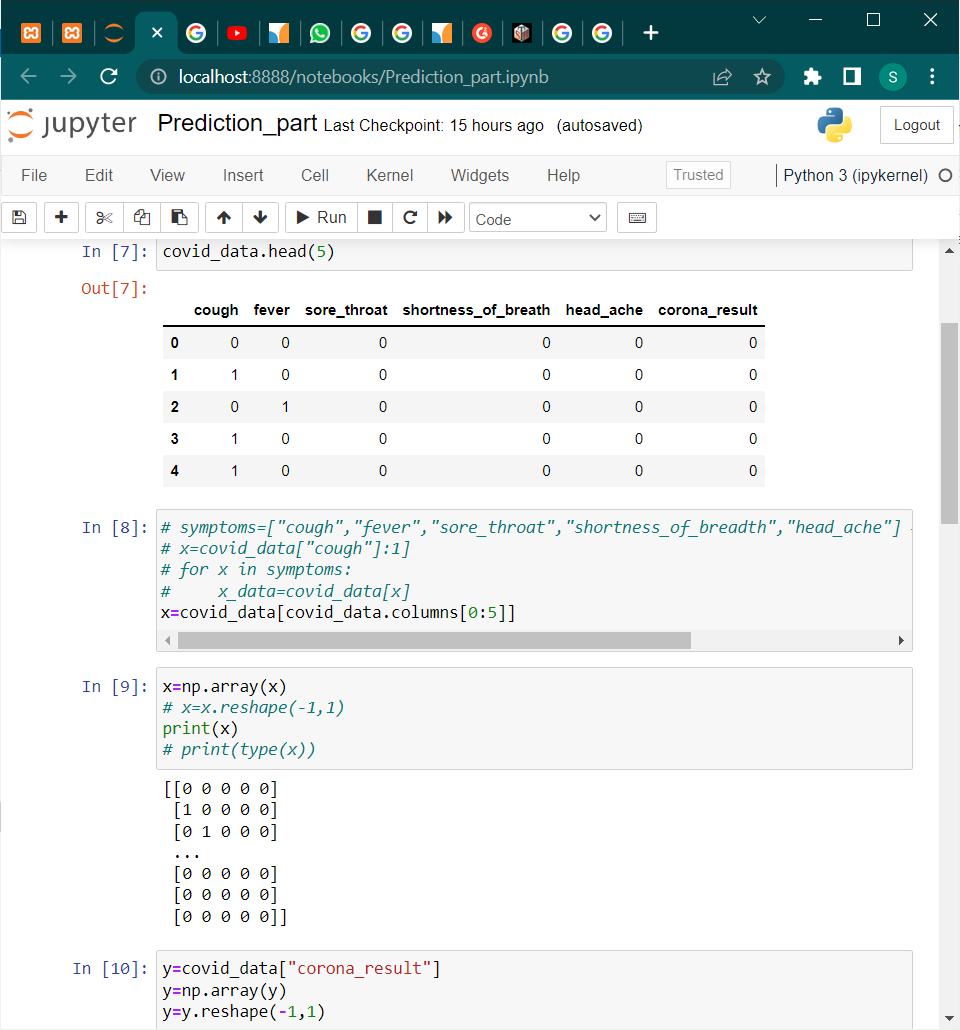
At first we need to import the necessary libraries, modules ,models as shown below:



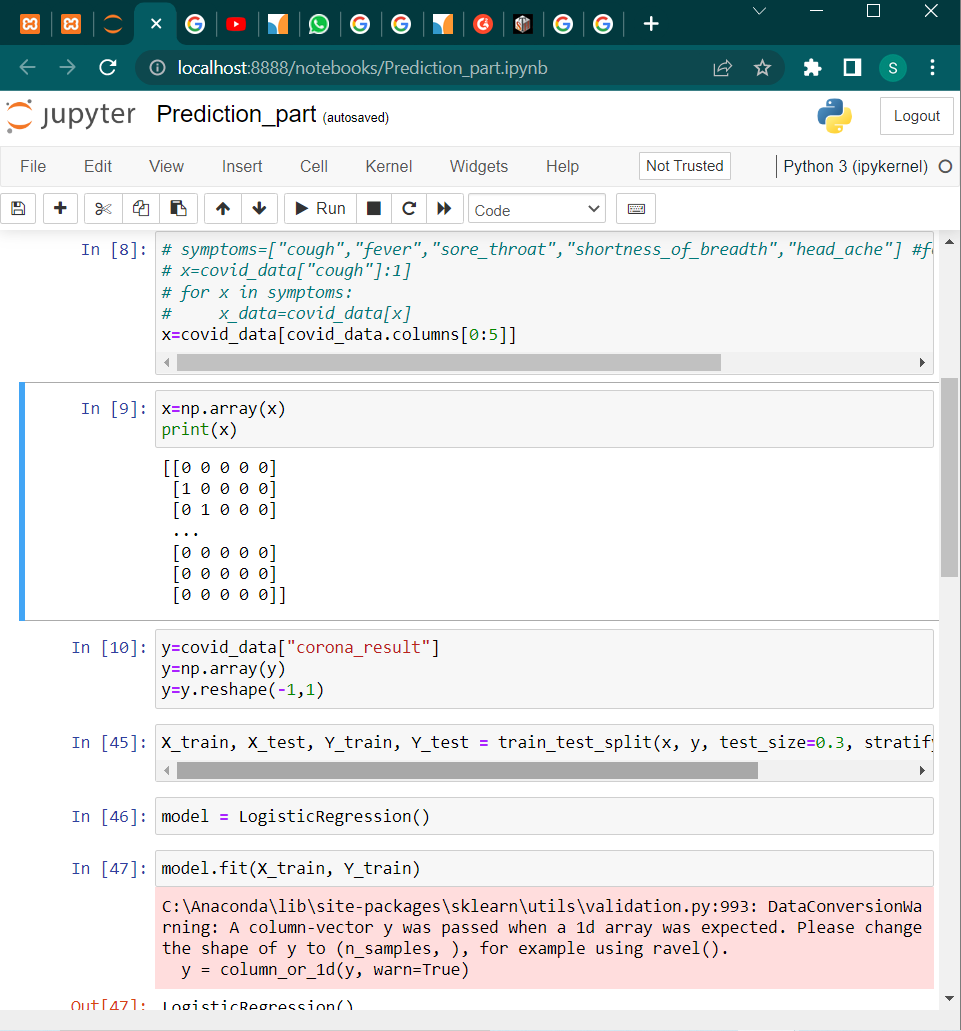
Now read the csv file and display it.



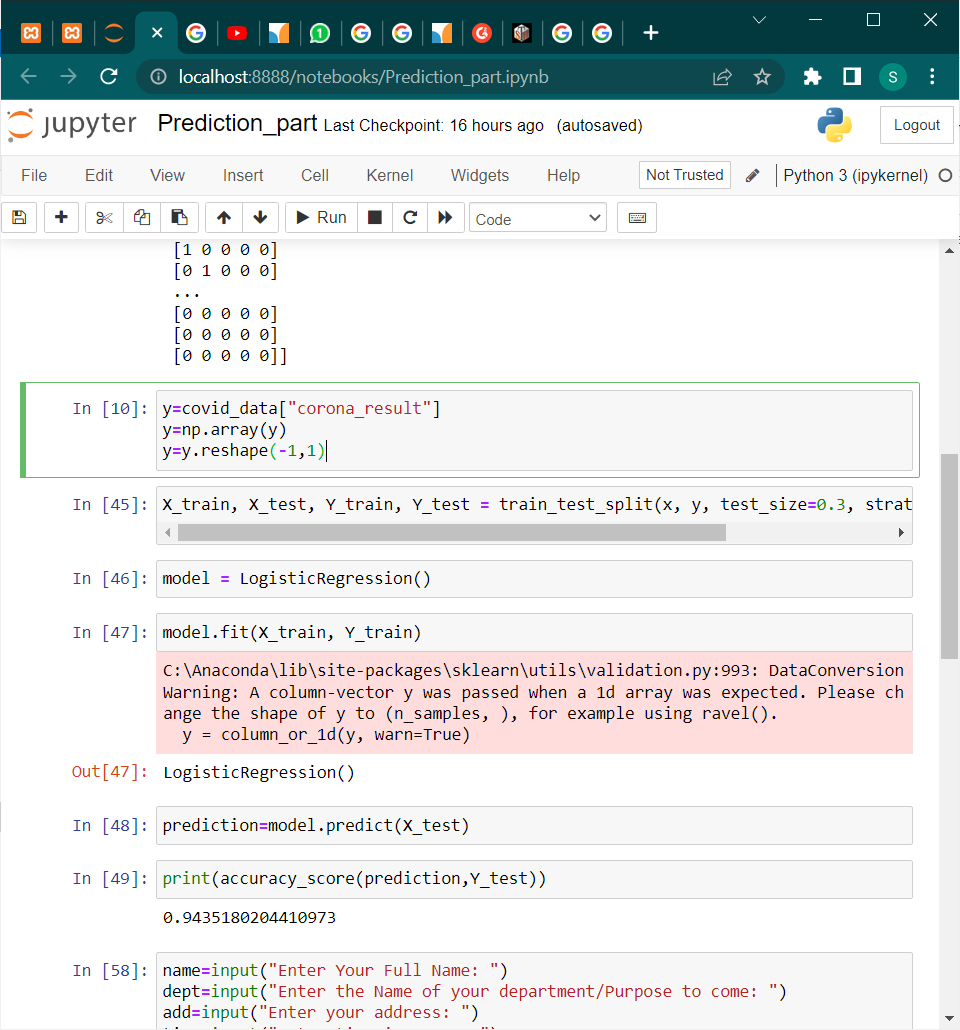
Now time to train our dataset by taking all the columns of symptoms as FEATURES and the “corona\_result” column as “LABEL”.



After this we take the values of the symptoms which are in binary in a variable in an array, as machine learning works on 2-D array and print the variable as shown:



Now we take the another variable to store our LABEL i.e. corona\_result in an array and also using reshape function to convert it into 2-D array.



Now using train\_test\_split function

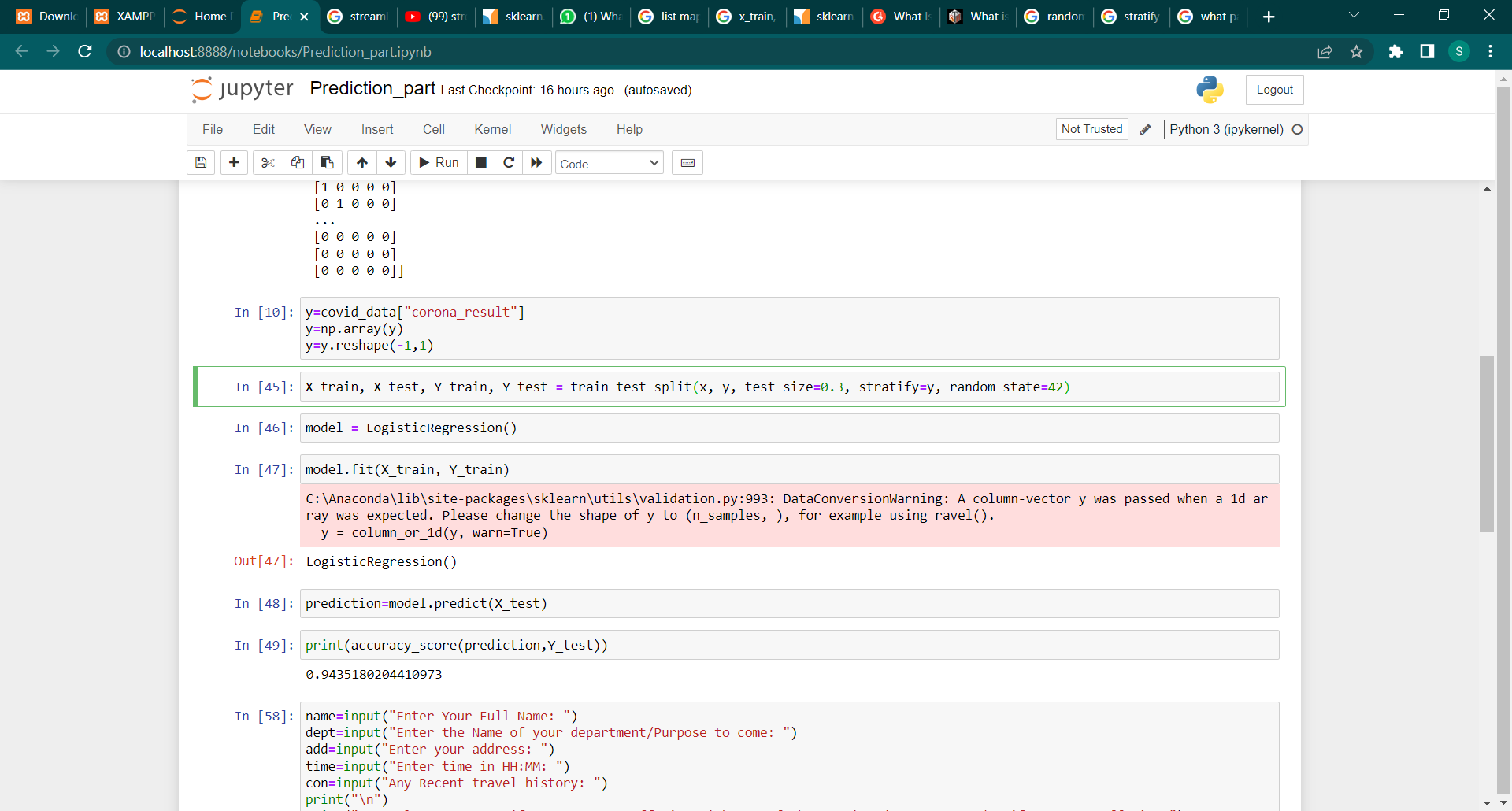


Here X\_train, X\_test, Y\_train , Y\_test are the variable that we used for taking some portions of datasets. Here test\_size denotes the amount of data that need to be stored in X\_test and Y\_test.

After split up the data we use training data i.e X\_train and Y\_train variables in the model and then we test the model by using X\_test and Y\_test variables.

Now we take the model LOGISTIC REGRESSION and Fit the model by giving training parameters.

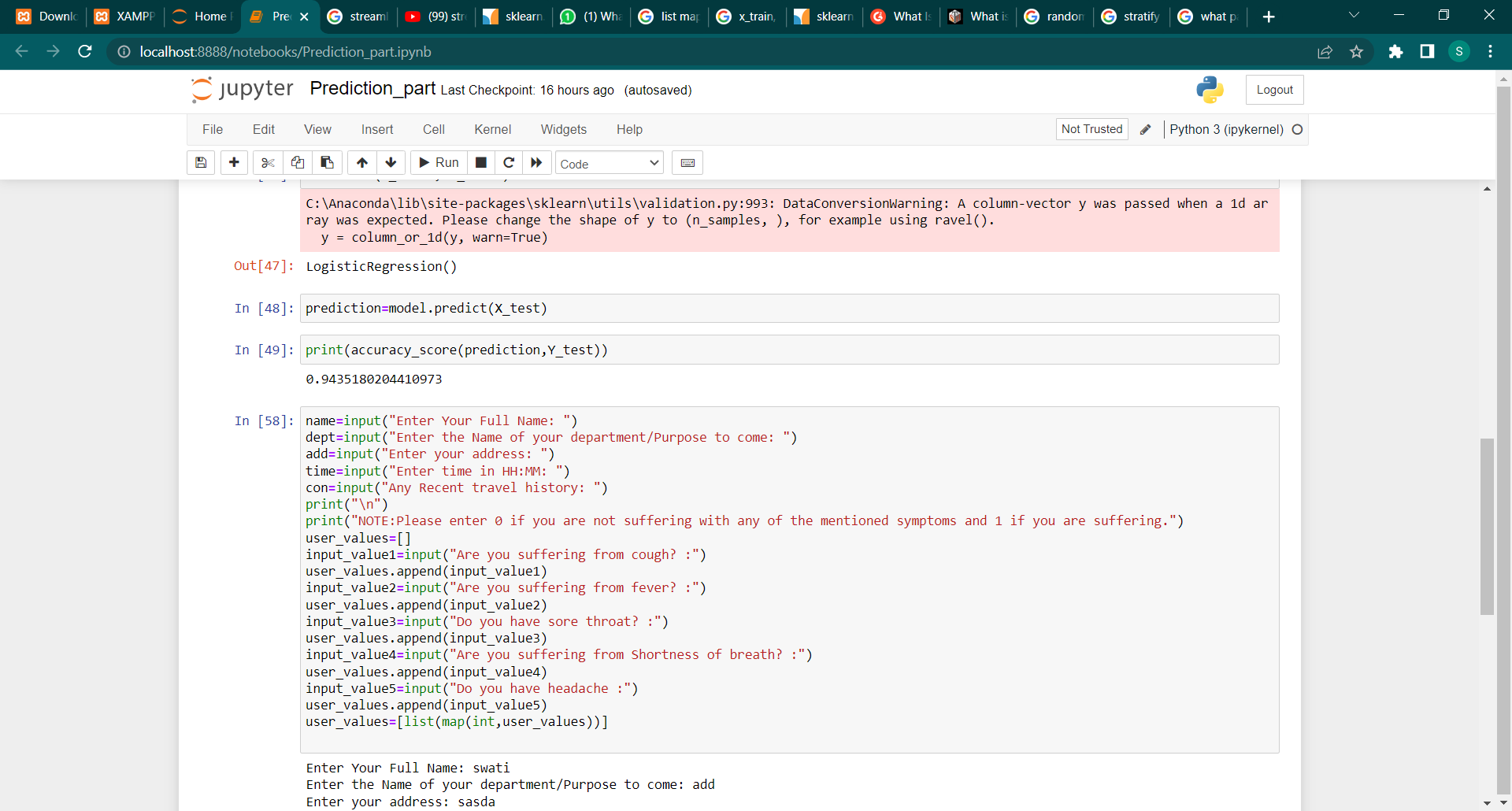
Model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes.



Now CHECKING THE ACCURACY OF THE MODEL BY USING METRICS

After training the model and then We test the model, so during the test ,we need to calculate the accuracy of that model so that we can rely on that particular model for future used, so for this we need check the “measurement metrics”. Hence here we used the accuracy metrics for checking the accuracy of the model.

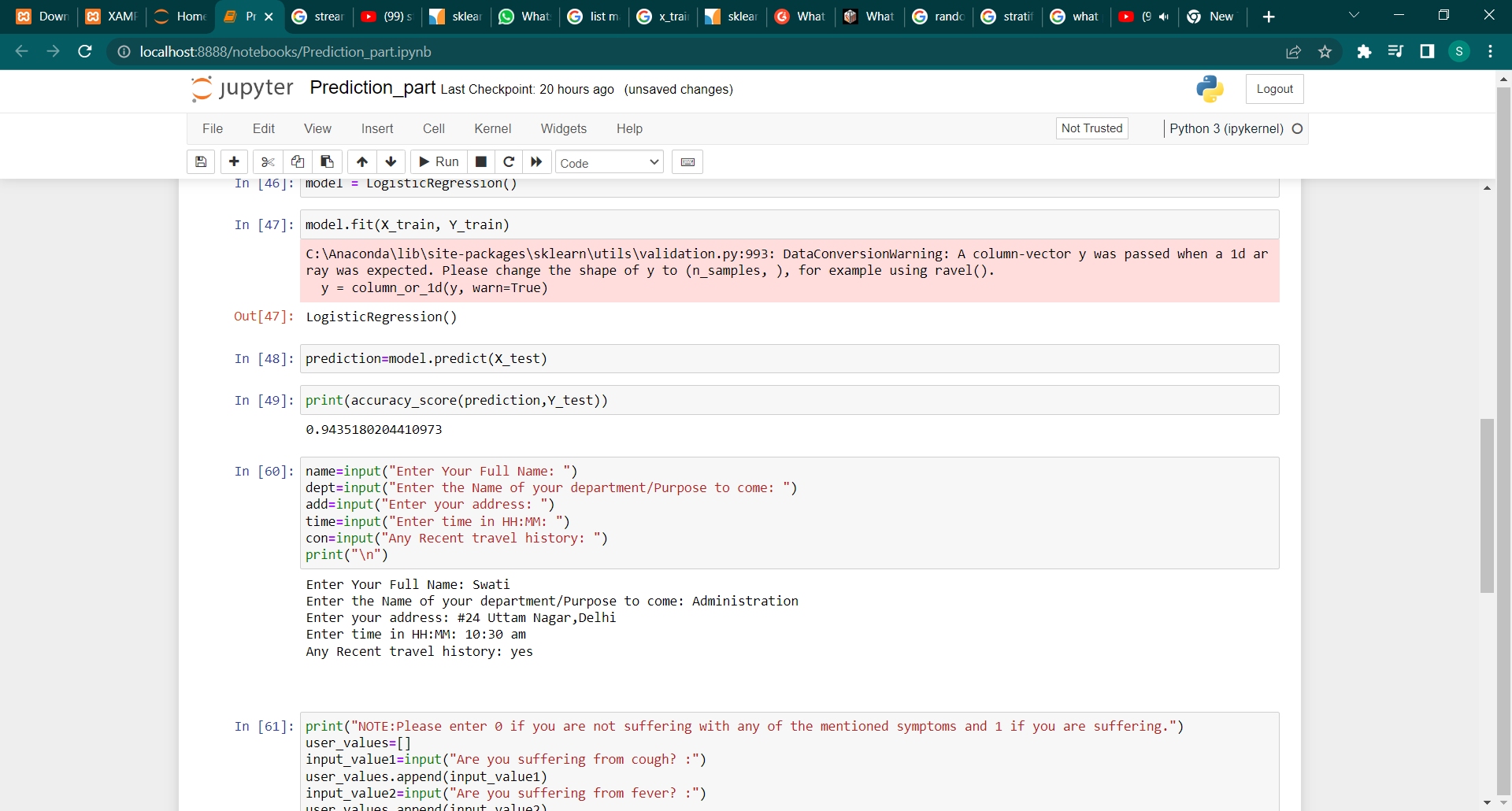
Accuracy: Accuracy can be measure by analyzing the True state or events out of total possible state or events.

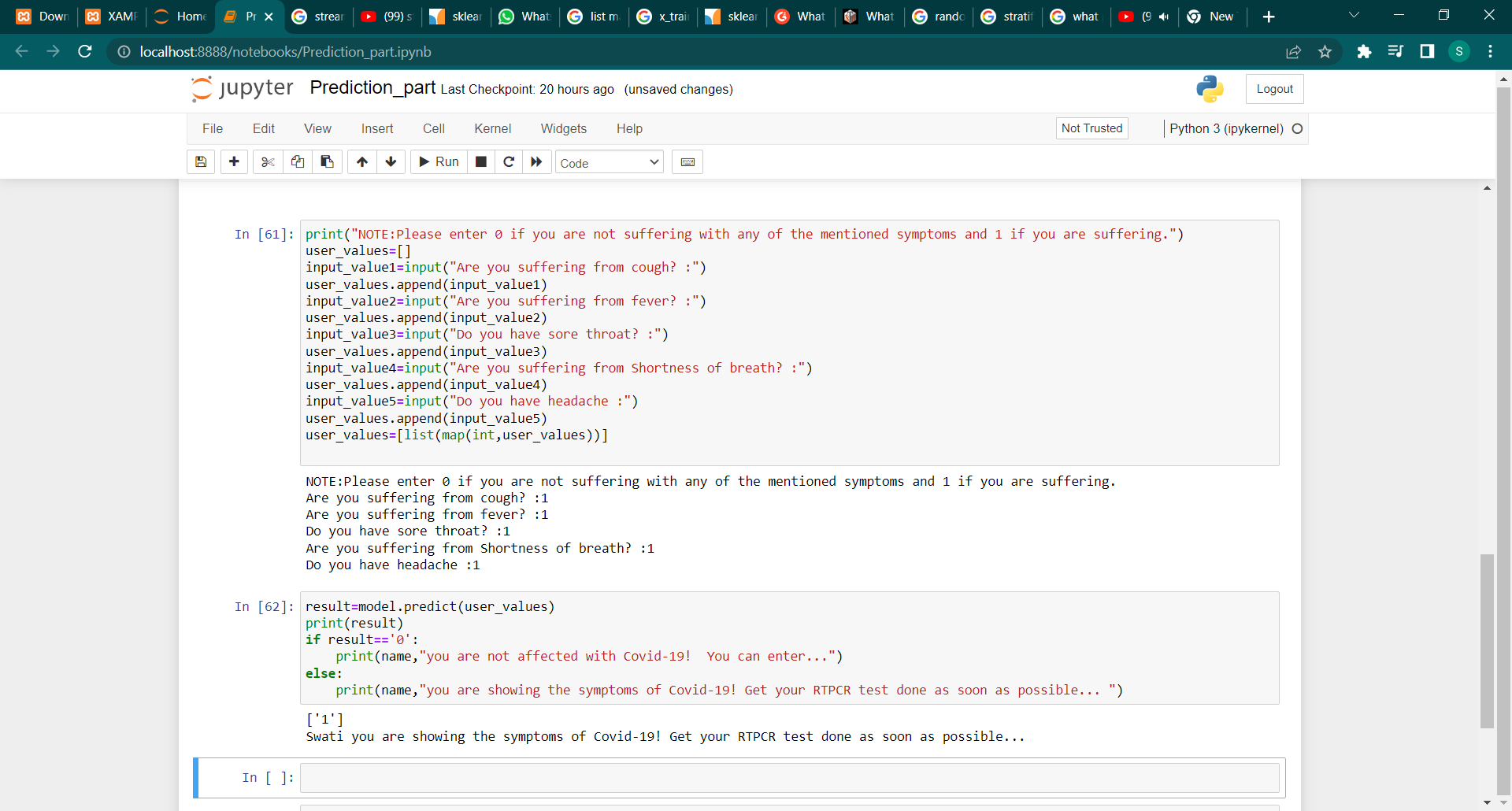


Now we can see that the accuracy of our model is near to 1 which is good so we can proceed further.

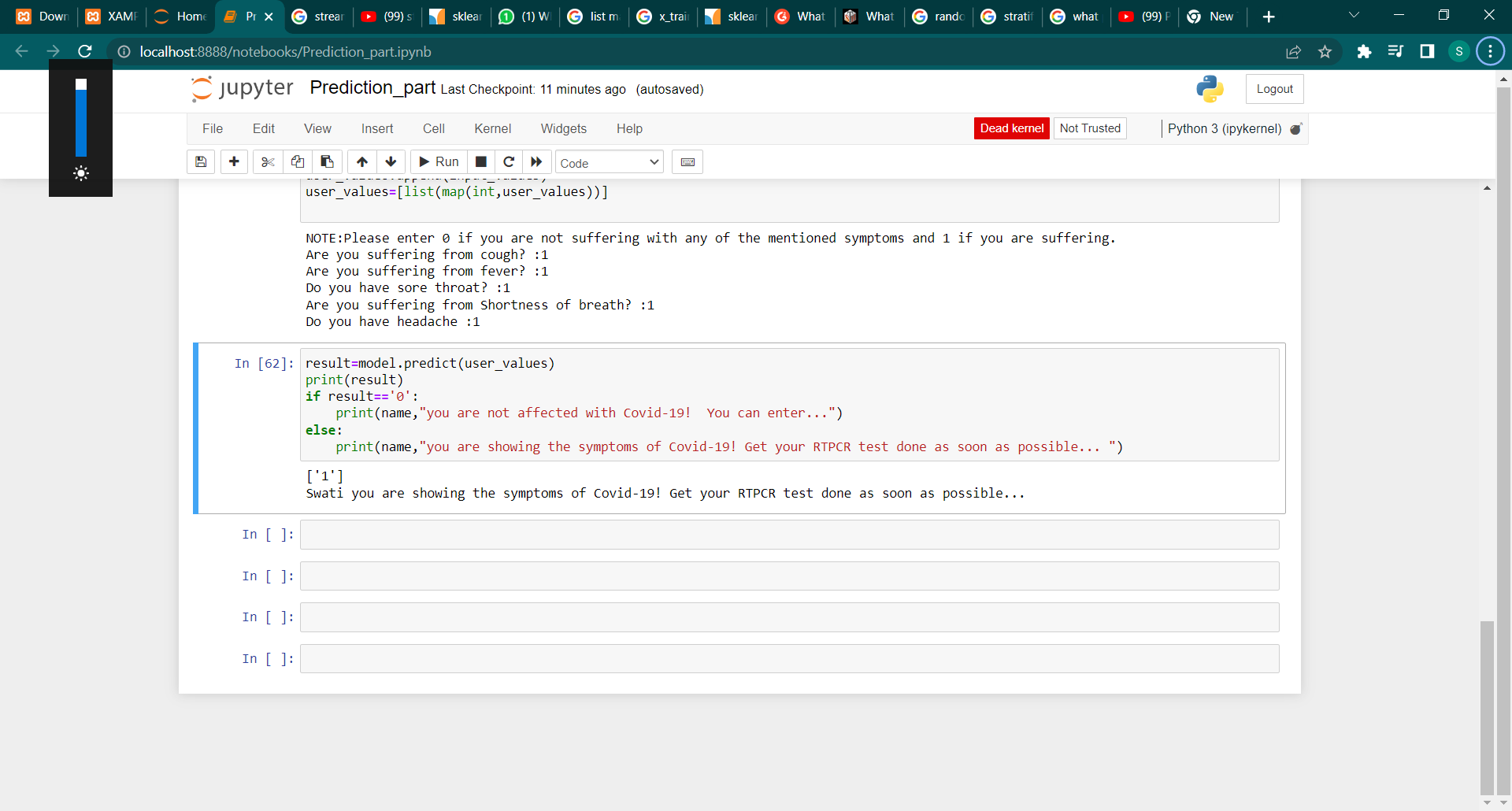
After training ,testing and accuracy part now we will ask some general information from the user and then about symptoms that they will answer and accordingly our model will predict that whether the person is covid positive or negative.

Below are the some general questions and then list of symptoms that we asked from the user:



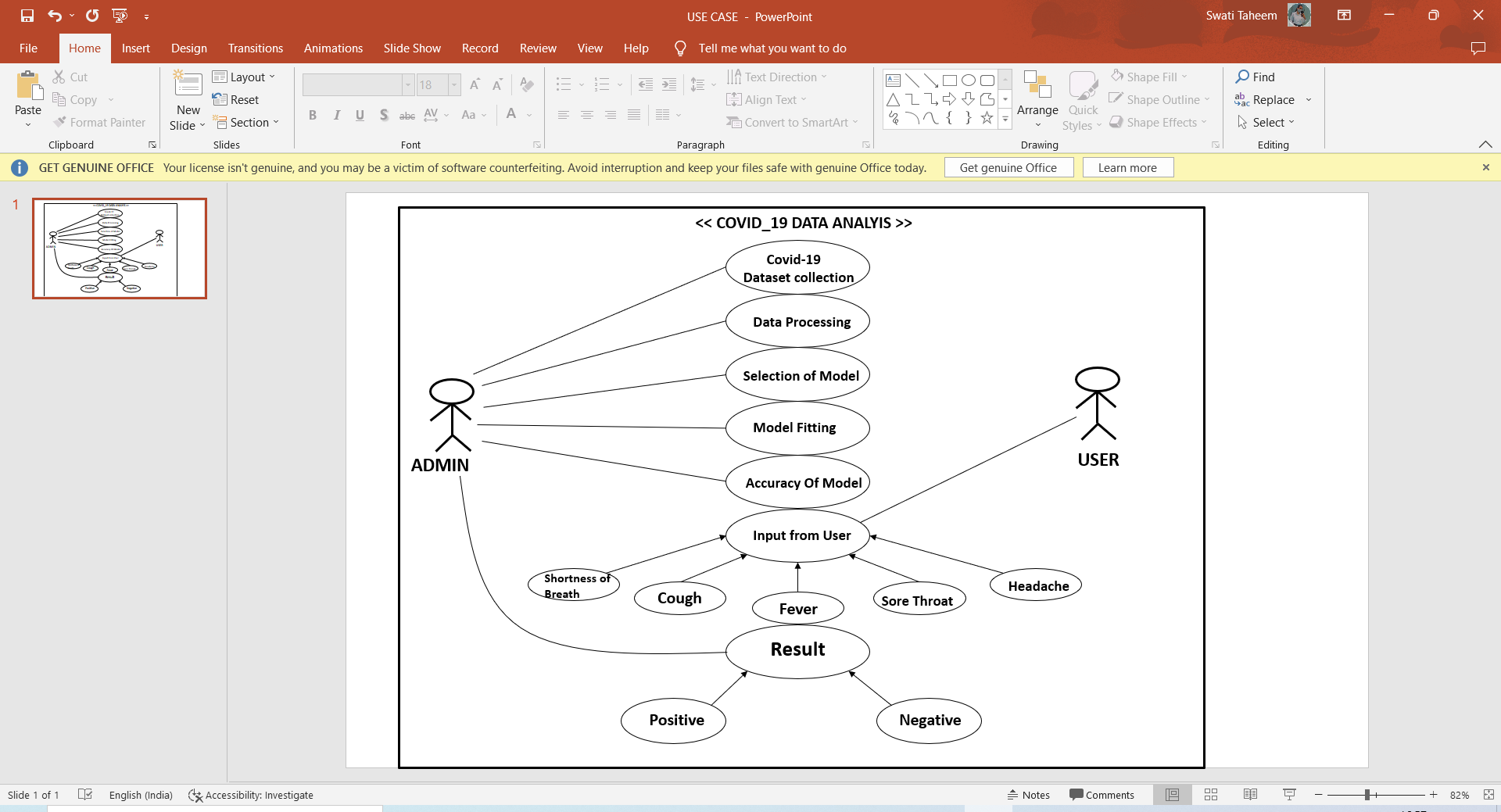


Now we store our final result in the variable called ‘result’ and if its predicts the value ‘0’ then person is not affected with covid-19 virus and if the predicted value is ‘1’ then person is affected with the covid-19 virus and needs a RTPCR test ASAP.



Framework used:

* **Use Case Diagram:**



**CHAPTER-3**

**Software Project Analysis**

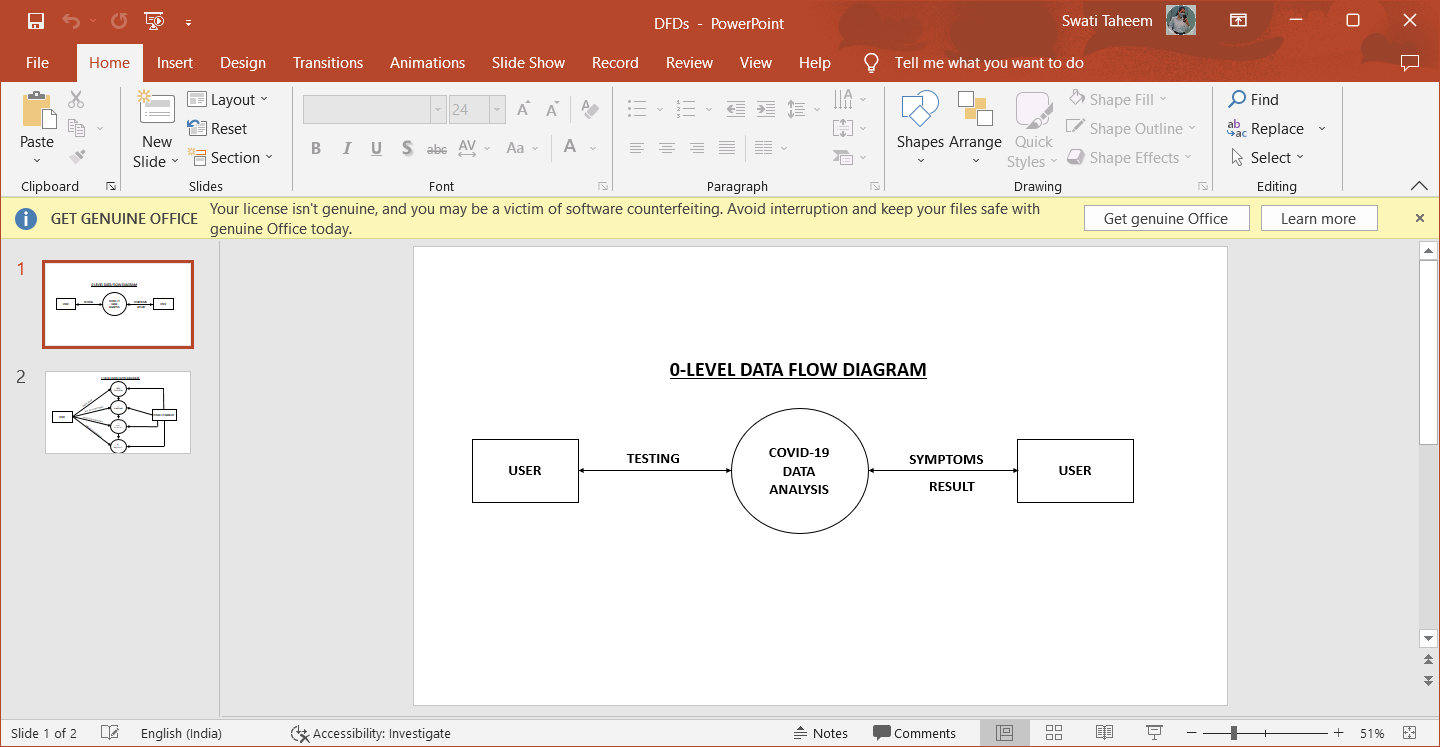
**(Covid-19 Data Analysis)**

* **DFD’s :-** Data Flow Diagram

The flow of data of a system or a process is represented by DFD. It also gives insight into the inputs and outputs of each entity and the process itself. DFD does not have control flow and no loops or decision rules are present. Specific operations depending on the type of data can be explained by a flowchart. Data Flow Diagram can be represented in several ways. The DFD belongs to structured-analysis modeling tools.

Data Flow diagrams are very popular because they help us to visualize the major steps and data involved in software-system processes.

Below is the 0-Level Data Flow Diagram of prediction phase:



Below is the 1-Level Data Flow Diagram of prediction phase:

