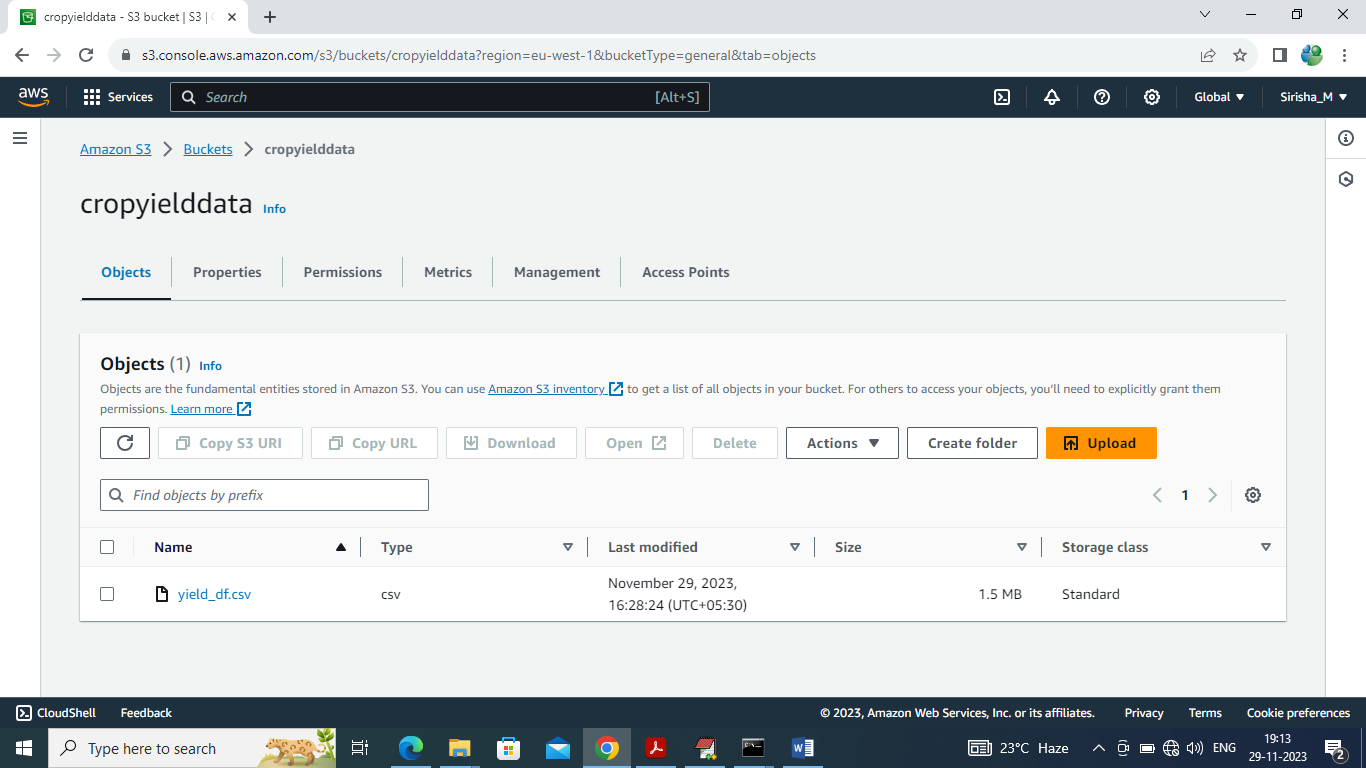
**Agroinformatics for yield prediction**

In propose work we are employing Machine Learning algorithms to predict crop yield as the crop is the main resource for any country development but this crop often get affected with varied temperature, rainfall, pesticides and many more environment factor. If farmers know future crop yield based on current environment then they can adjust their budget as per yield and based on this analysis country and farmer lost will get reduced and revenue will get increased.

To predict future crop we need to train ML algorithms but when crop data increased then traditional algorithms may not work so to work with any size of data we are employing distributed technology based framework such as HADOOP and SPARK which has inbuilt ML algorithms to train any size of data in distributed format.

Spark can be utilized to analyse and train Bigdata so we are employing SPARKML for crop yield prediction and employing SPARK SQL to analyse and visualize dataset.

To train ML algorithms we have utilized same CROP YIELD dataset given by you and this dataset can be downloaded by SPARK framework from Amazon S3. In below screen showing crop data saved in Amazon S3



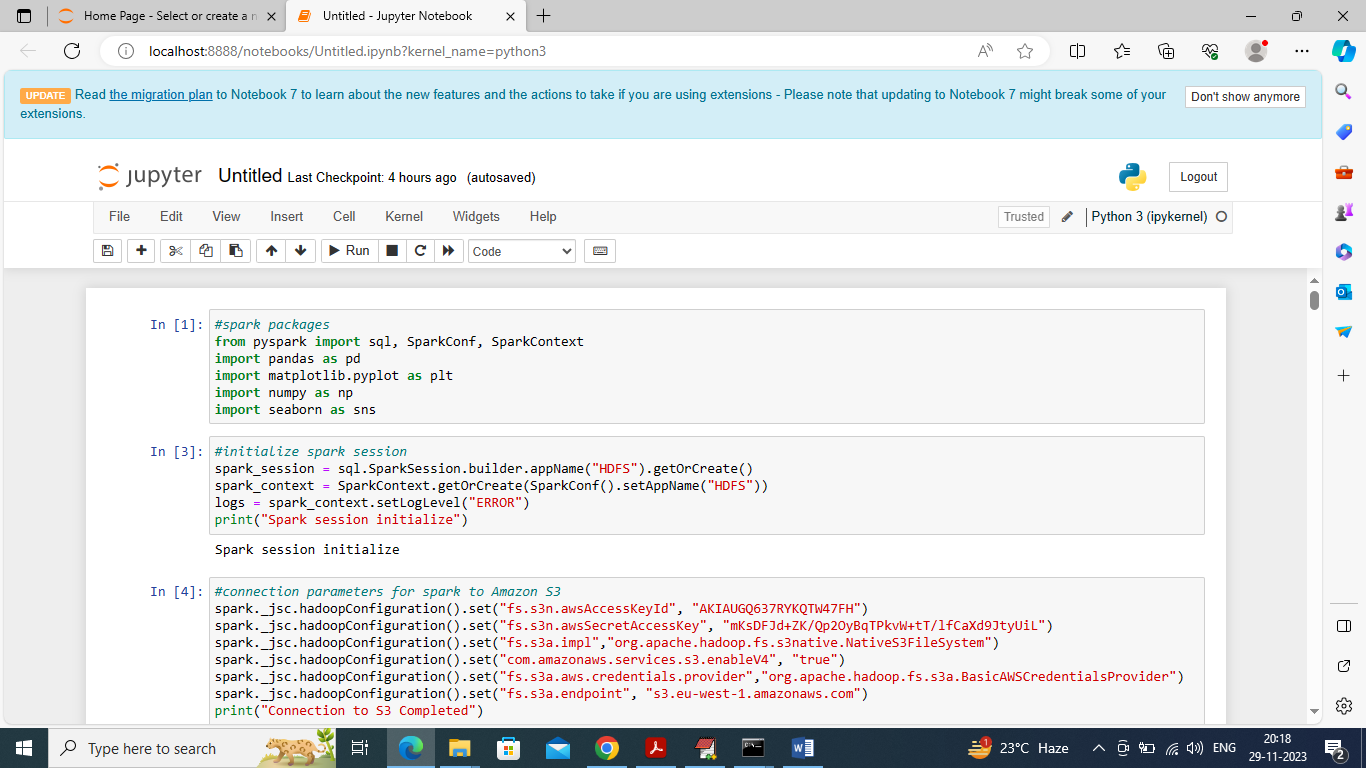
In above screen can see crop yield data available at Amazon S3 and spark can read from here and then perform analysis.

For analysis we did following experiments

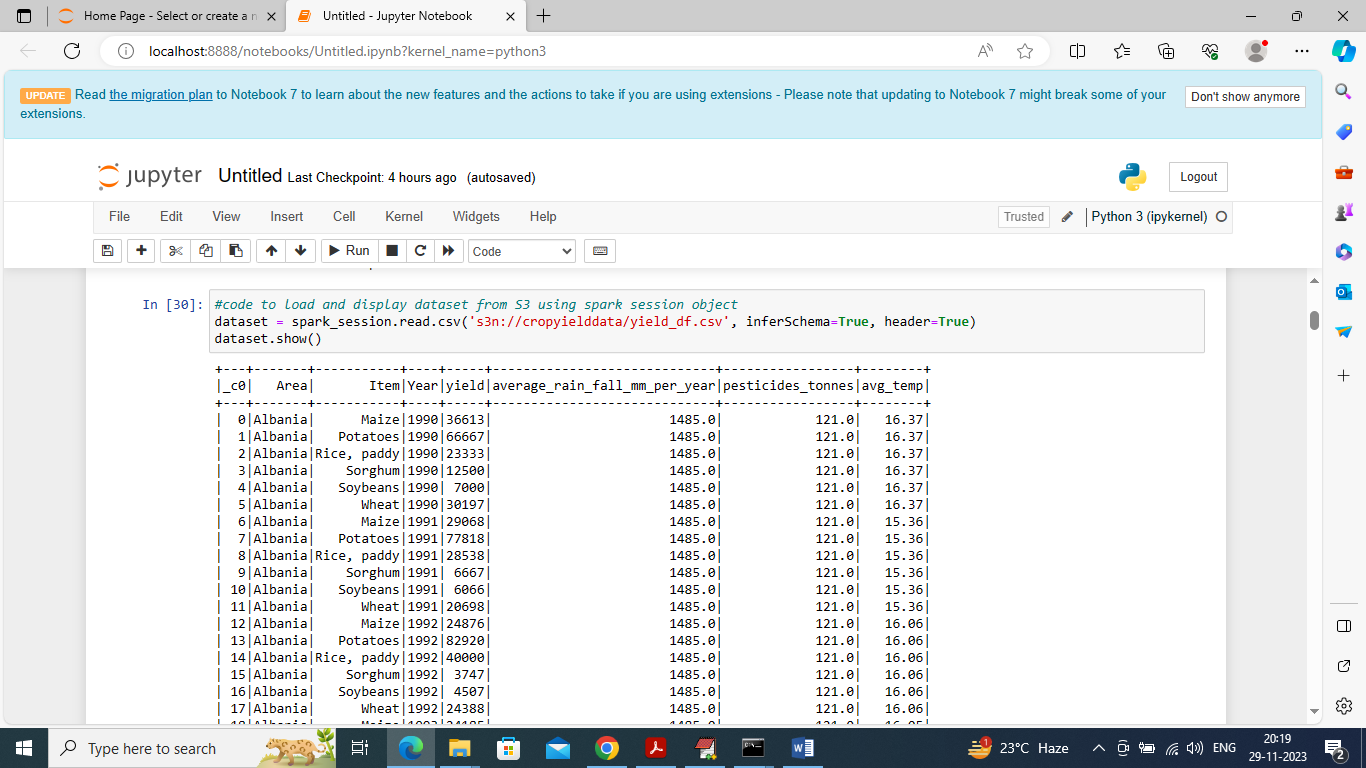
1. Describing dataset using mean, min, max, standard deviation
2. Plotting histogram graph for each column values to know how values are distributed from min to max range
3. Visualizing different areas available in dataset
4. Visualizing different crops found in dataset
5. Visualizing graph of different crop yield by different countries
6. Visualizing Top 20 Area Wise Average Rainfall Graph
7. Visualizing Top 20 Area Wise Average Pesticides Consumption Graph
8. Employing Decision Tree ML algorithm to predict future crop for select country and crop name
9. Employing Linear Regression ML algorithm to predict future crop for select country and crop name
10. Evaluating performance of decision tree and linear regression in terms of Root Mean Square Error (RMSE). RMSE refers to difference between true crop yield and predicted crop yield so the lower the difference the better is the algorithm.

SCREEN SHOTS

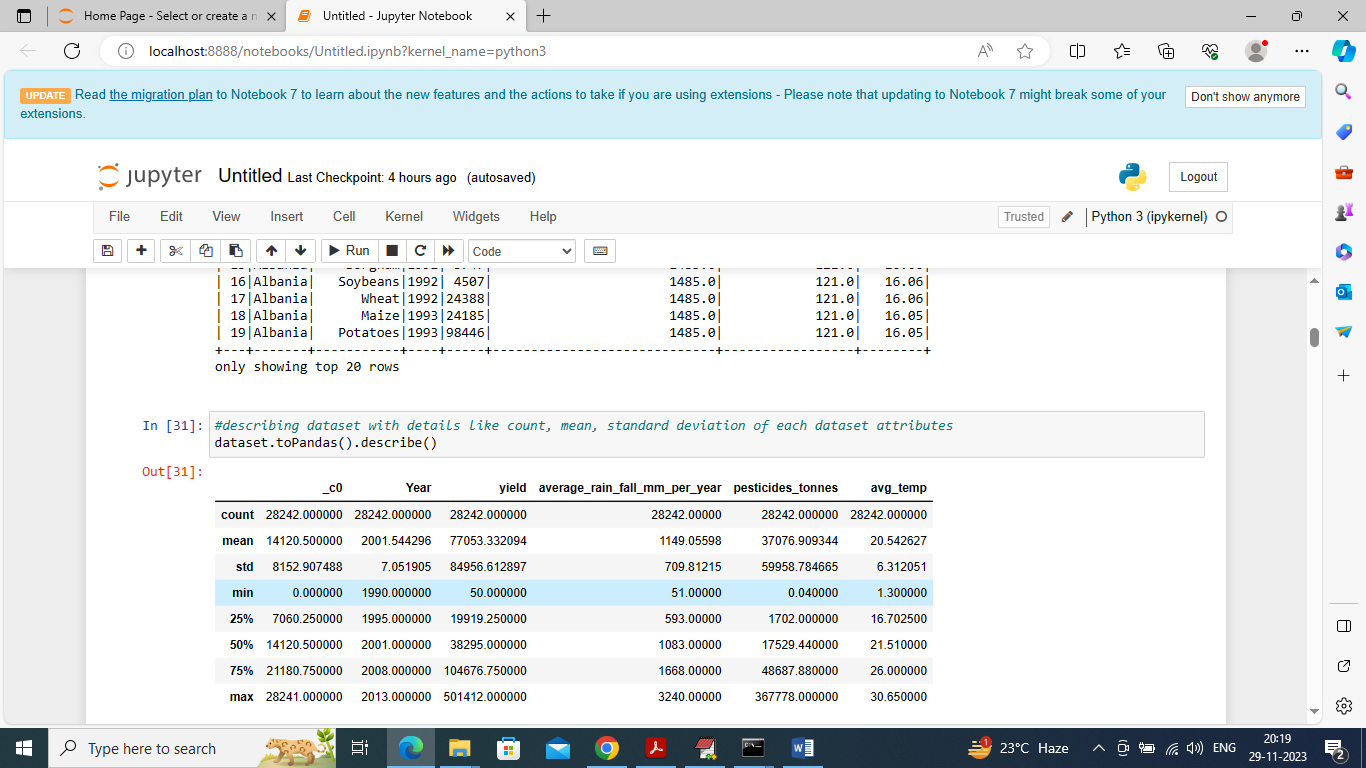
We have coded this project using JUPYTER NOTEBOOK and below are the code and output screens with blue colour comments



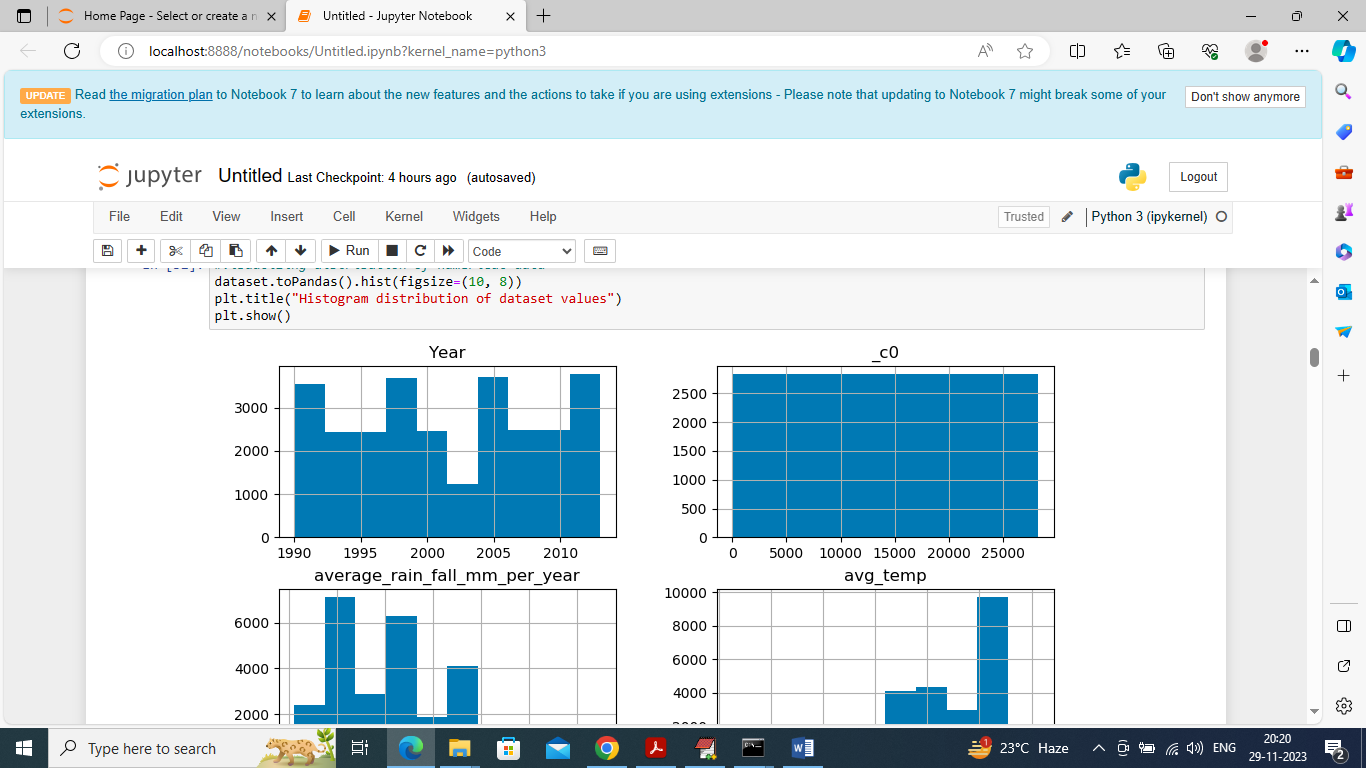
In above screen importing python classes and packages and then connecting to Amazon S3



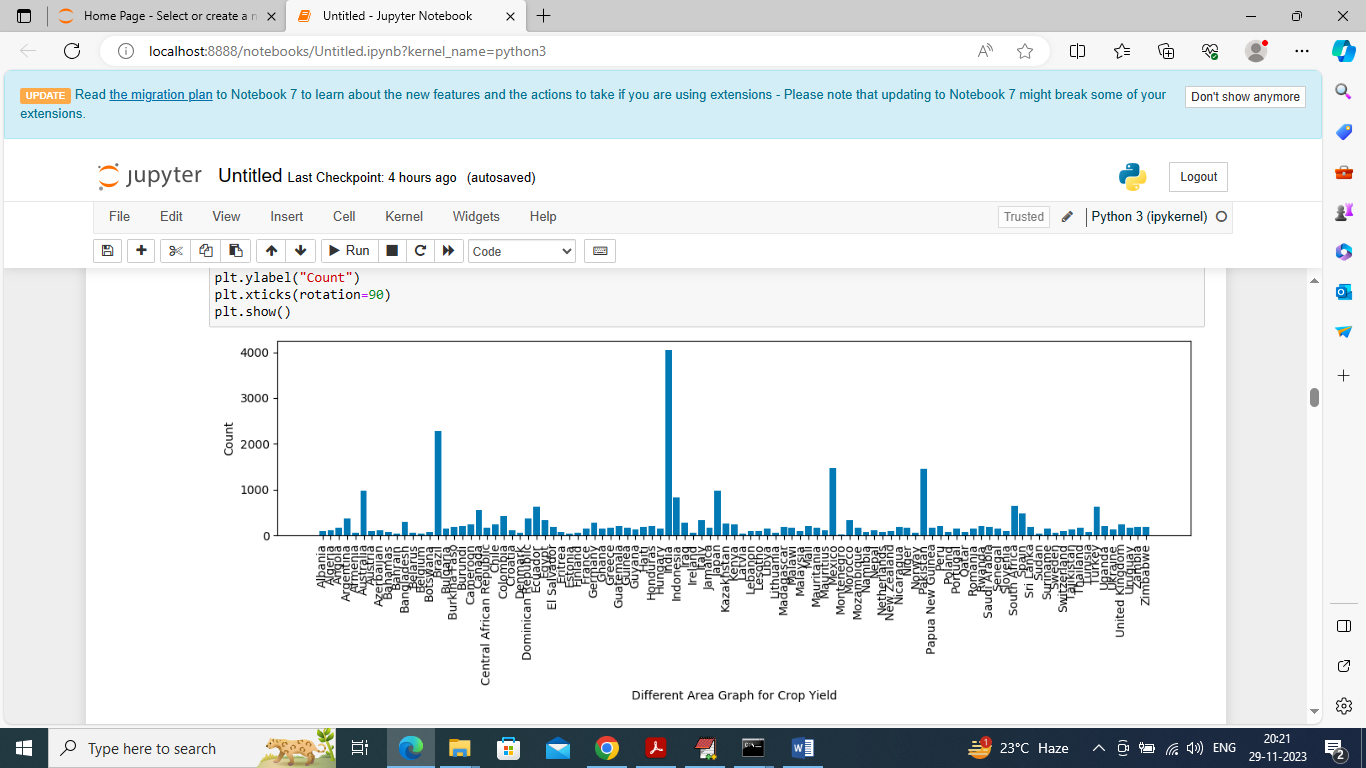
In above screen reading and displaying dataset values from Amzon S3



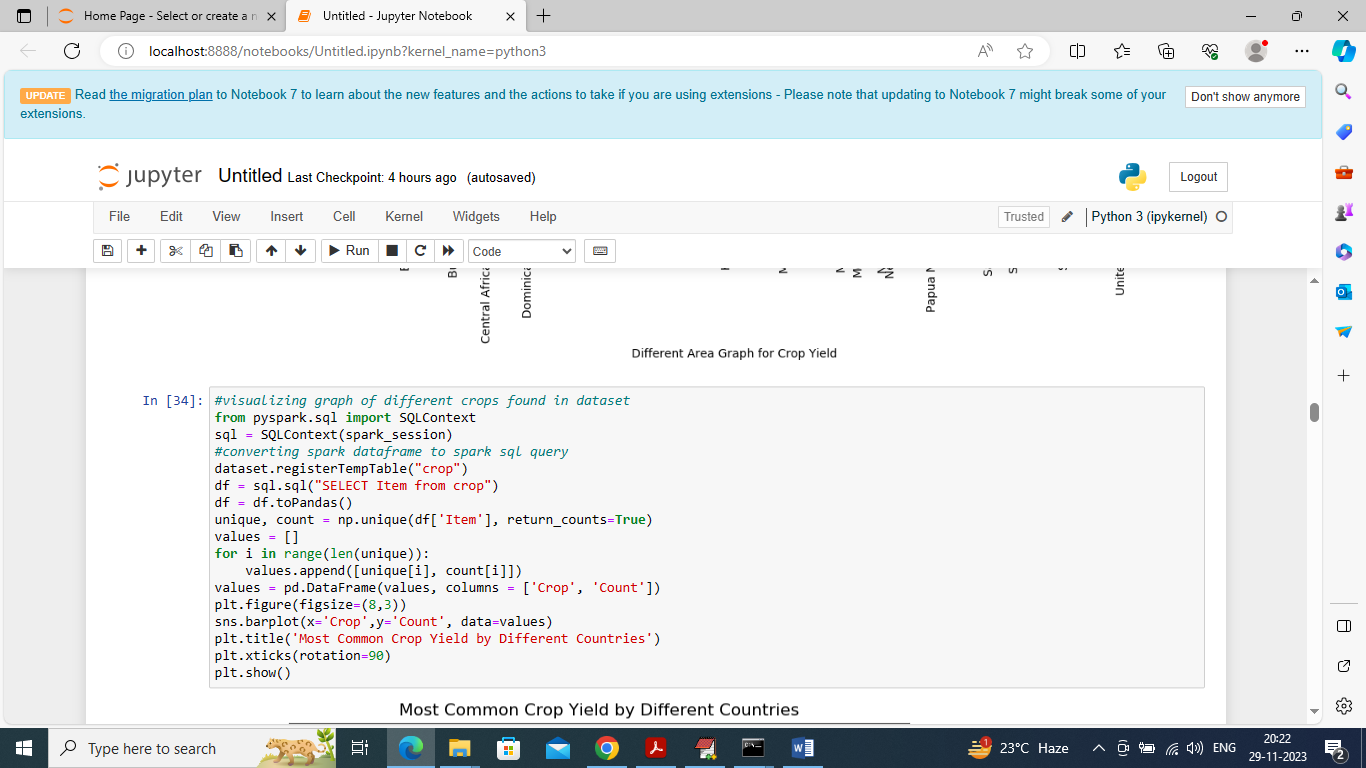
In above screen describing dataset values using metrics like min, max, standard deviation etc.



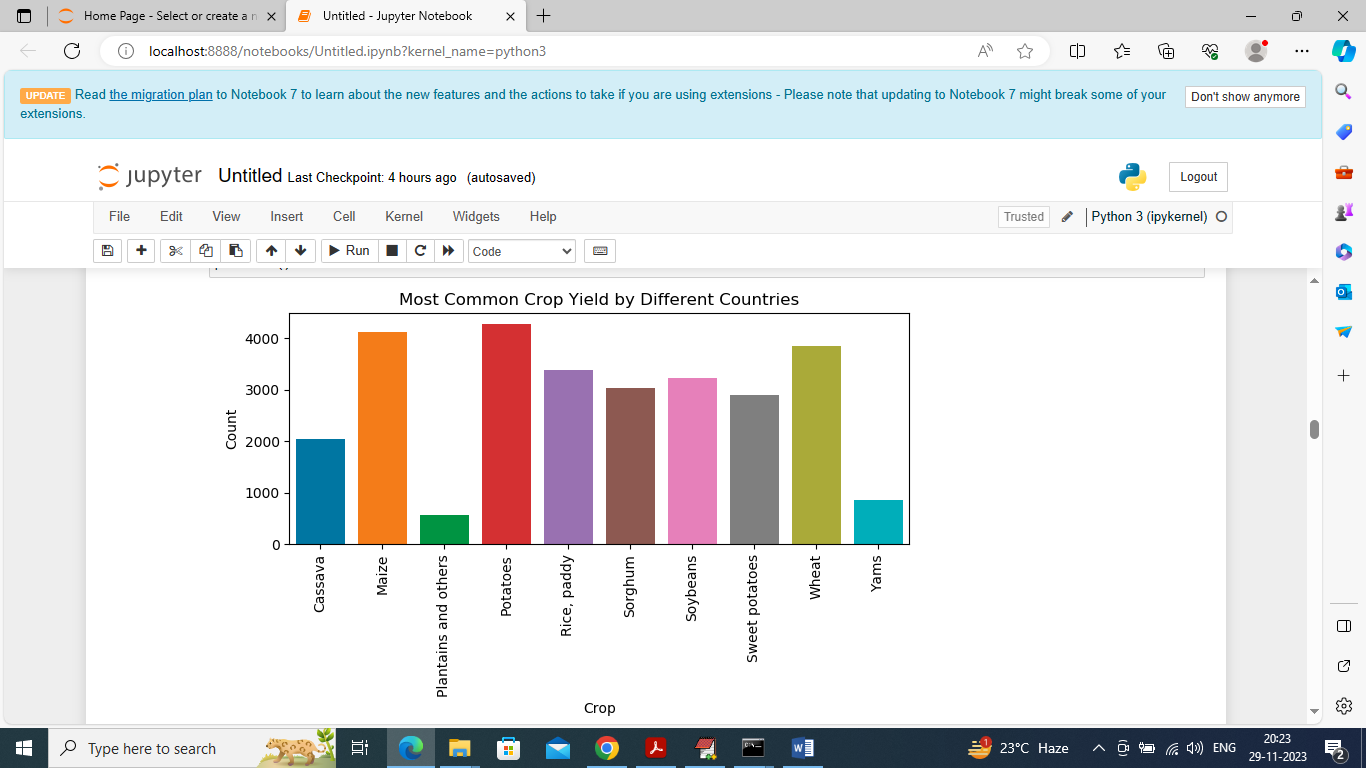
In above screen plotting histogram graph of different column values to know how values are distributed in each column



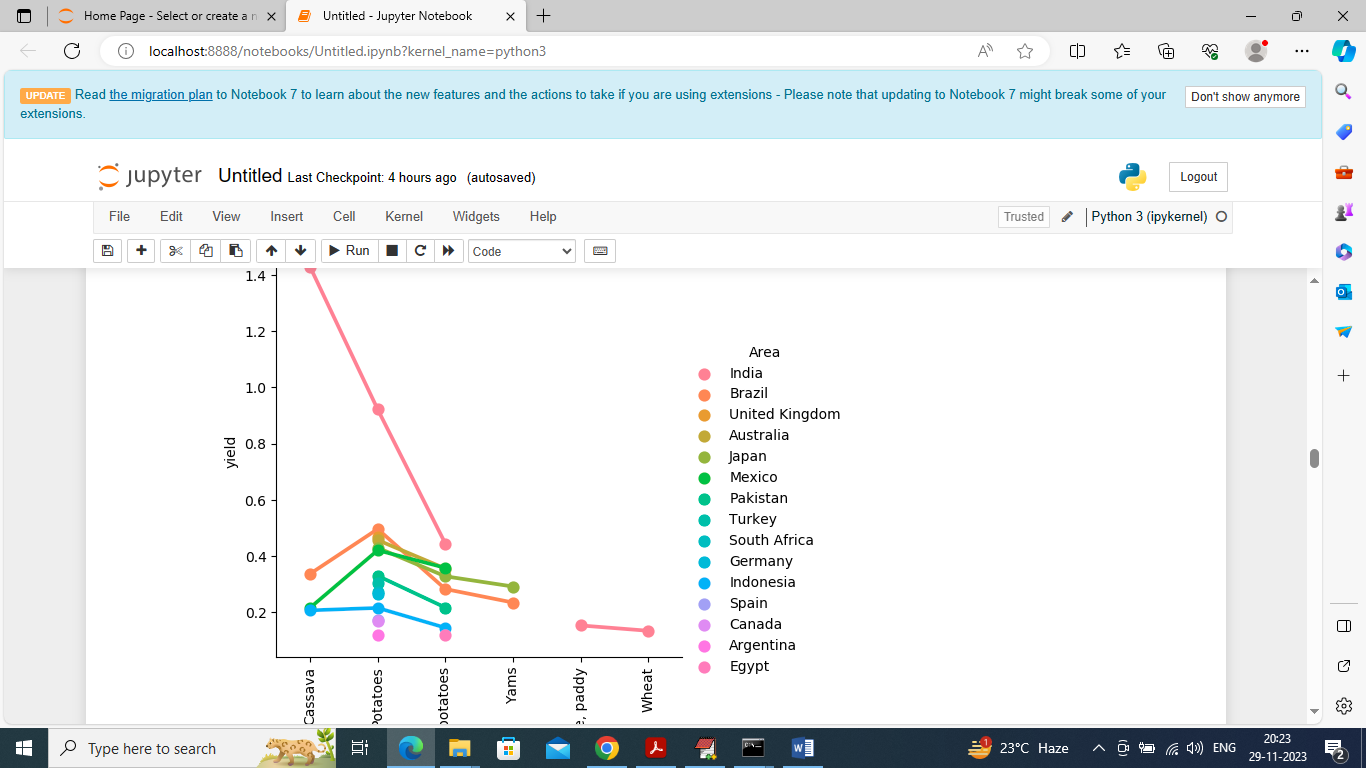
In above screen visualizing graph of Area involved in crop yield where x-axis represents Area Name and y-axis represents sum of all crops cultivation doning by that area



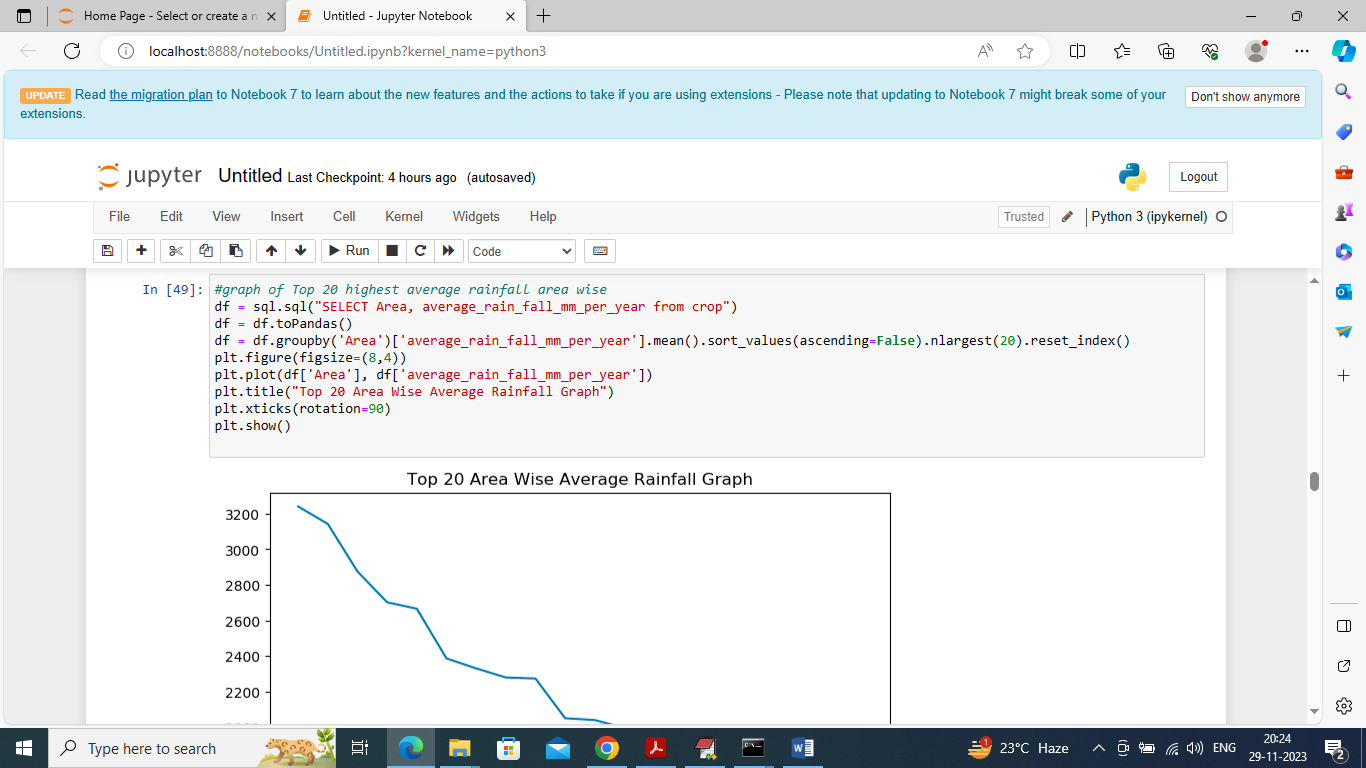
In above screen executing spark SQL query to find and visualize different crops found in dataset

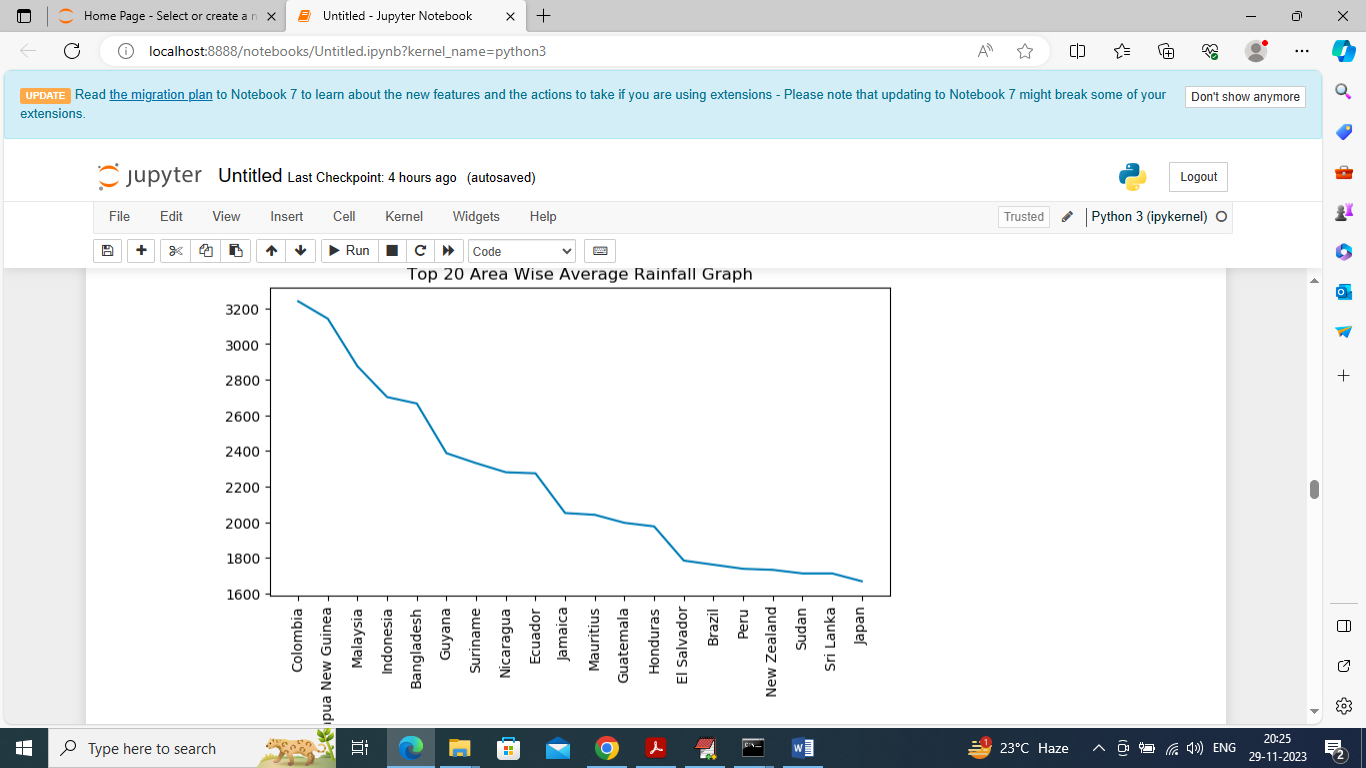


In above graph x-axis represents crop name and y-axis represents total cultivation of that crop found in dataset

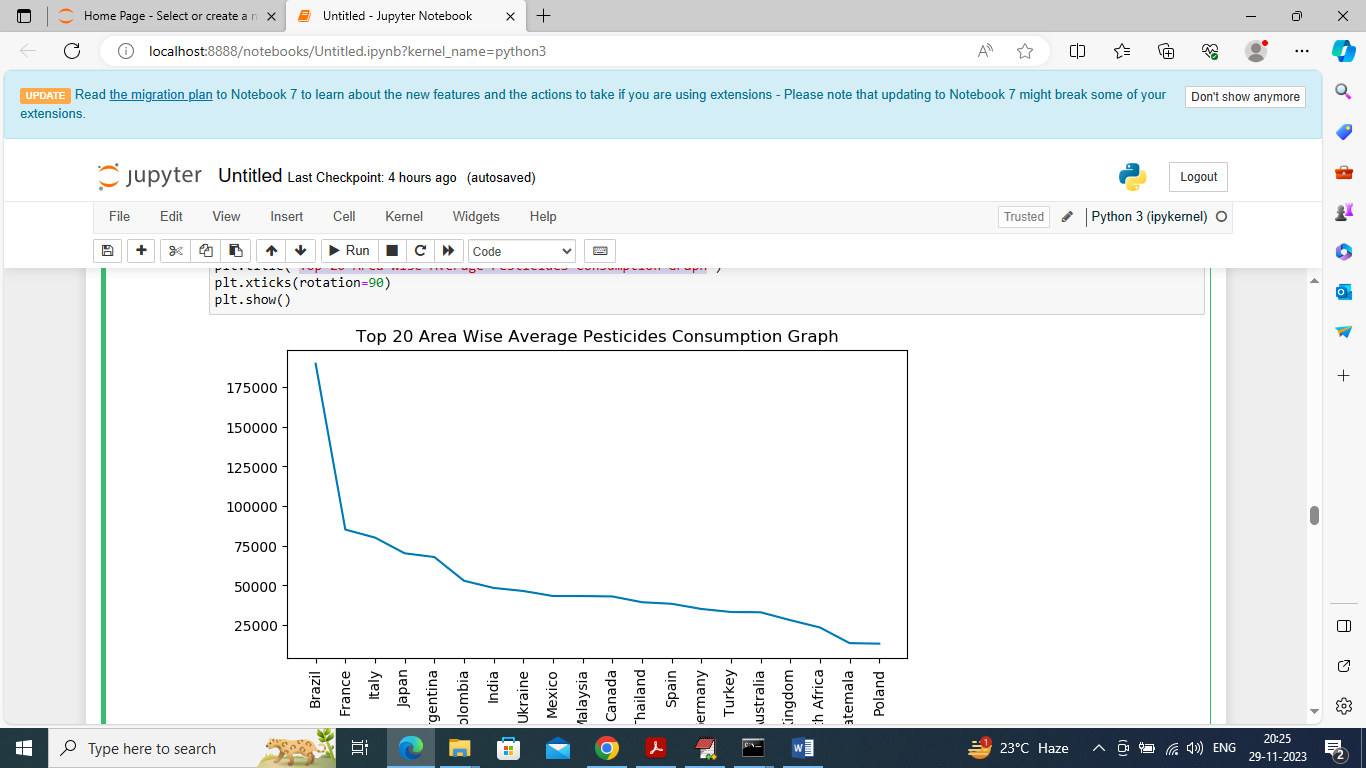


In above graph visualizing different crop cultivation doing by different countries where x-axis represents Crop Name and y-axis represents crop yield and each line showing crop yield for particular country

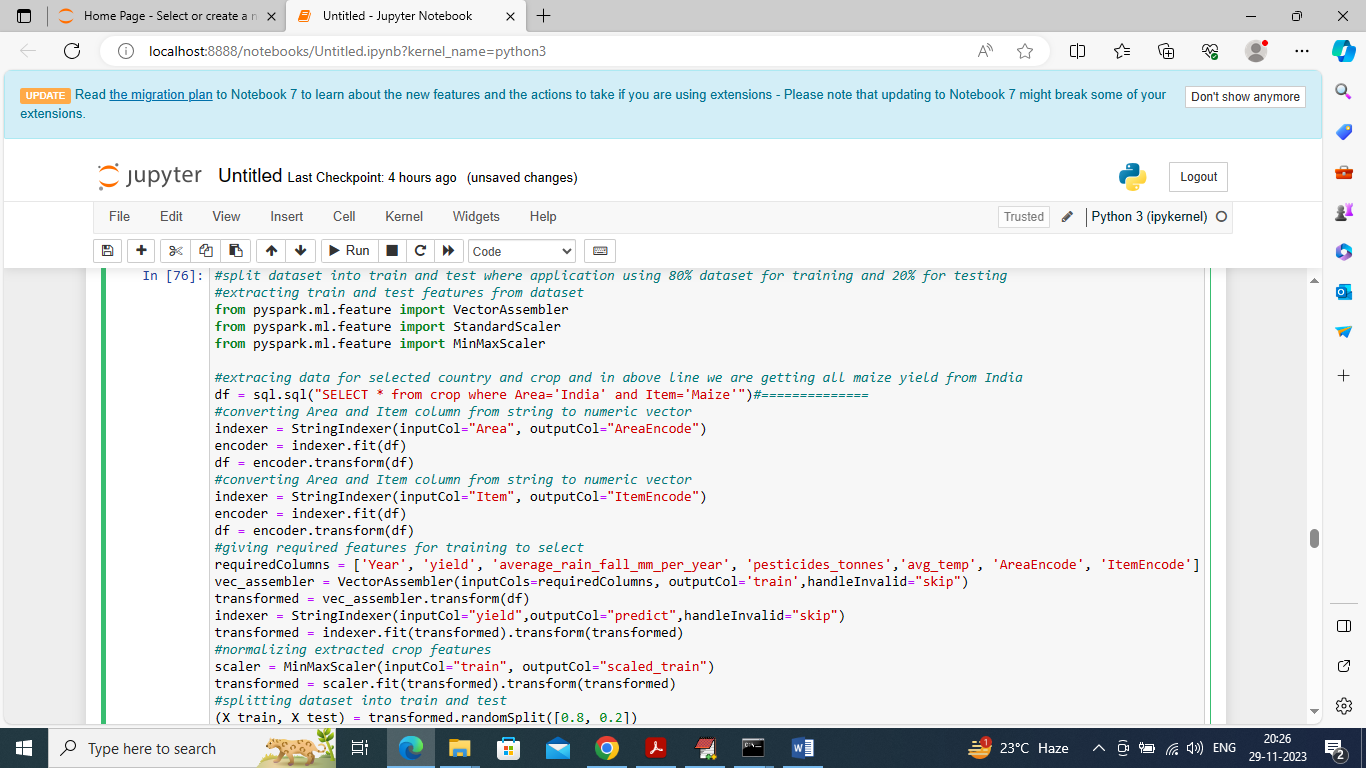




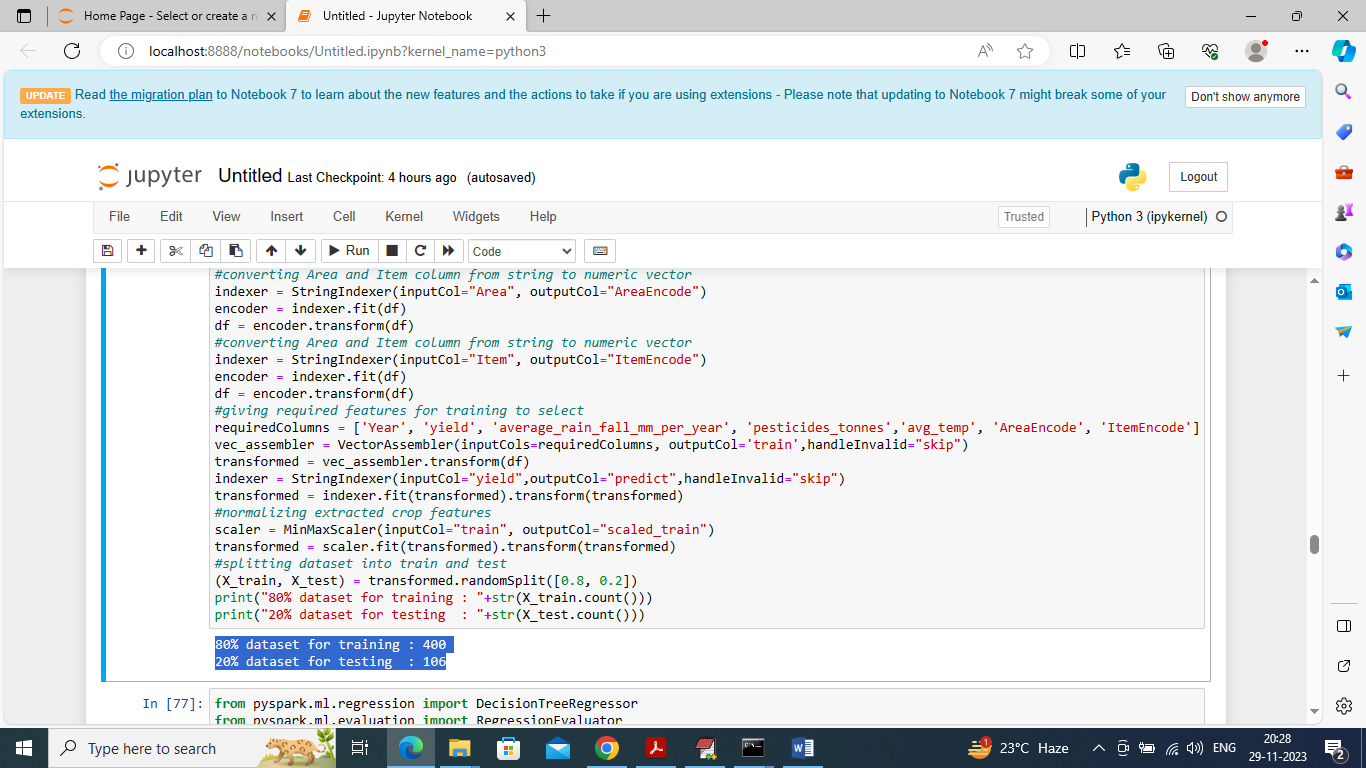
In above 2 screen executing SQL query to find top 20 countries with highest rainfall where x-axis represents country name and y-axis represents rainfall received



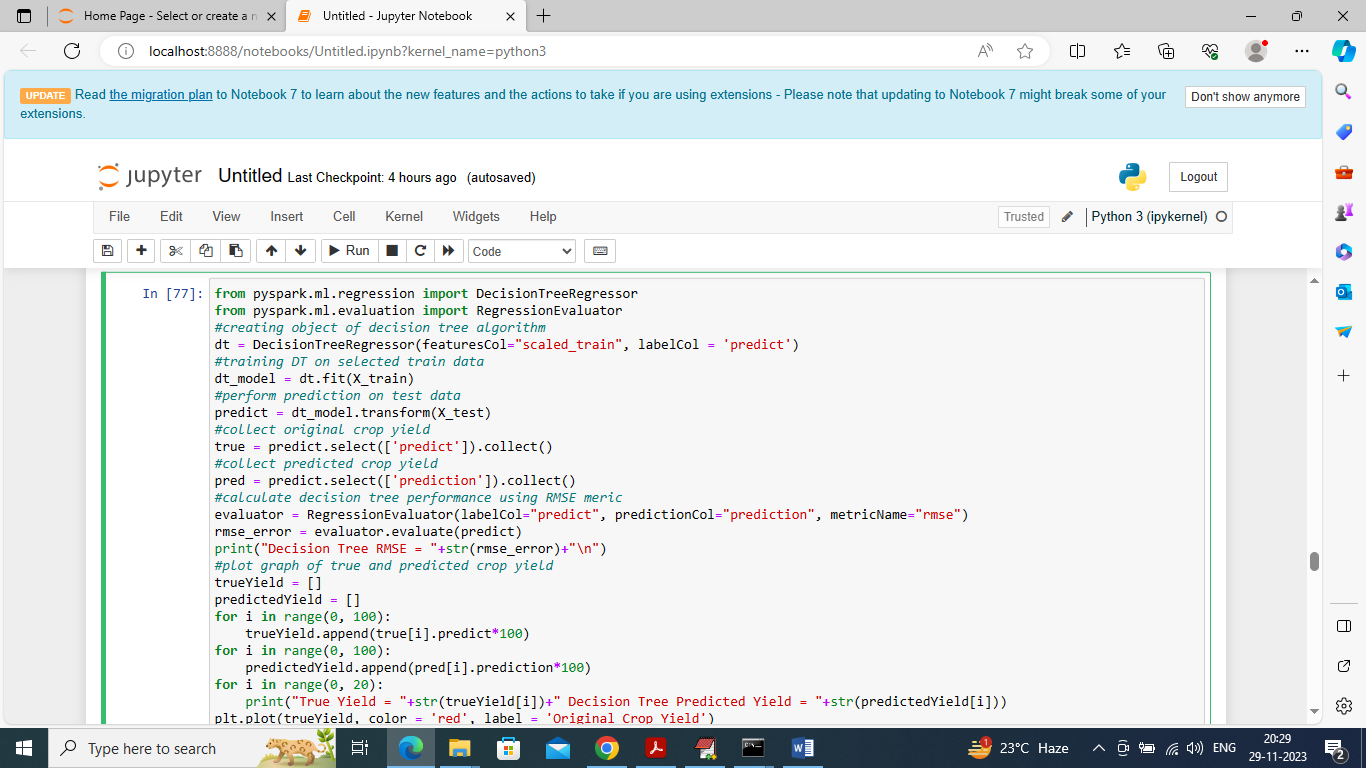
In above graph visualizing top 20 countries consuming highest pesticides



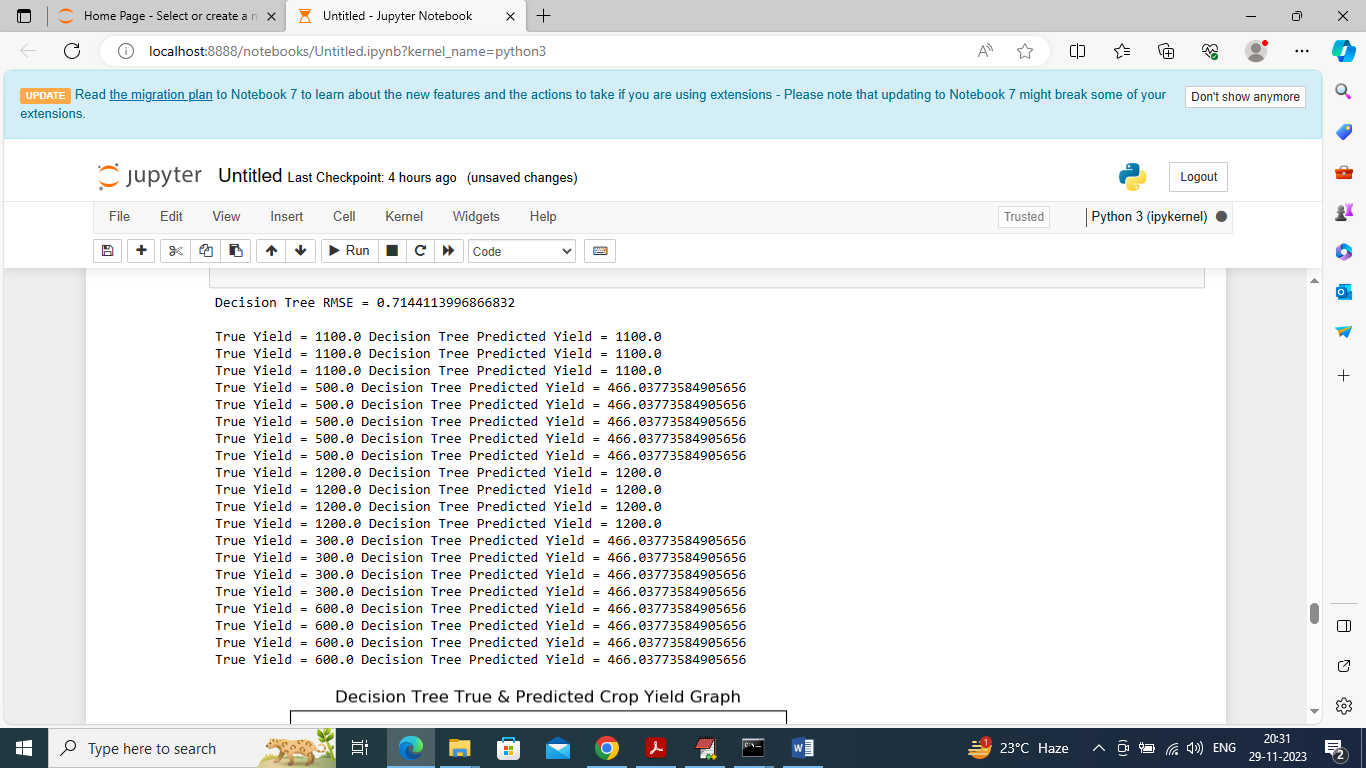
In above screen applying ML technique to process dataset and in above screen we executing SQL query to get all Maize crop cultivation by Indian country and then converting dataset to SPARL ML format by using transformation and then splitting dataset into train and test with a ratio of 80:20 and after executing above block will get below output



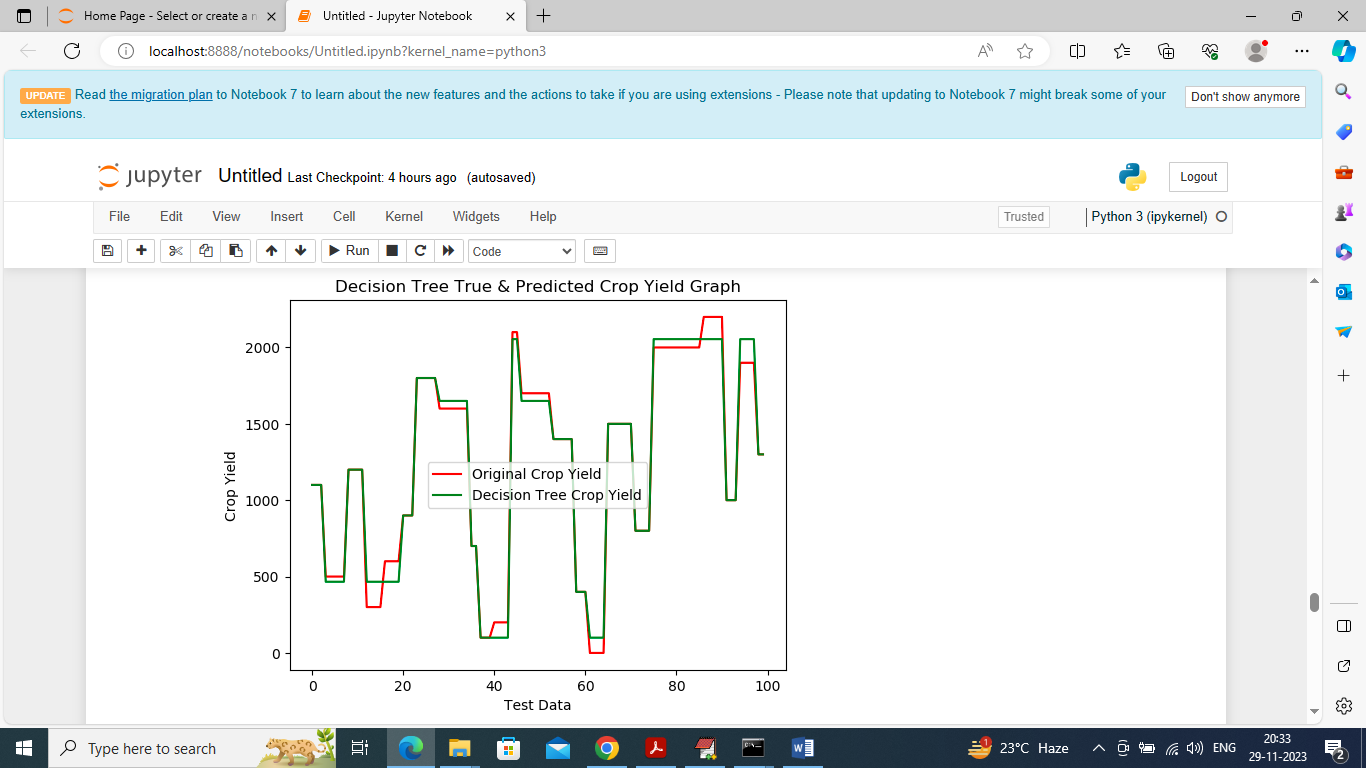
In above screen in blue colour text can see training and test size of Maize crop and by replacing country and crop name you can train ML to predict different crop yield



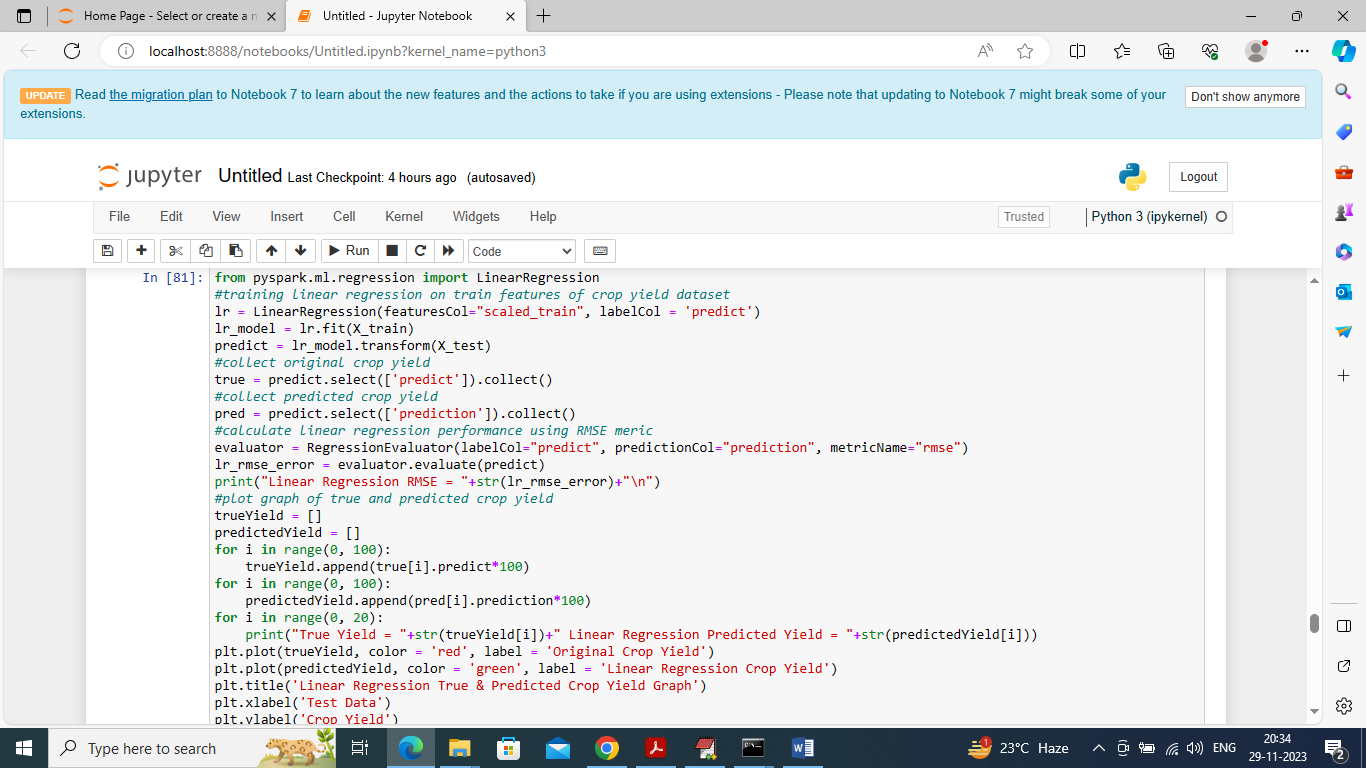
In above screen applying Decision Tree on train data and then performing prediction on test data to predict future crop



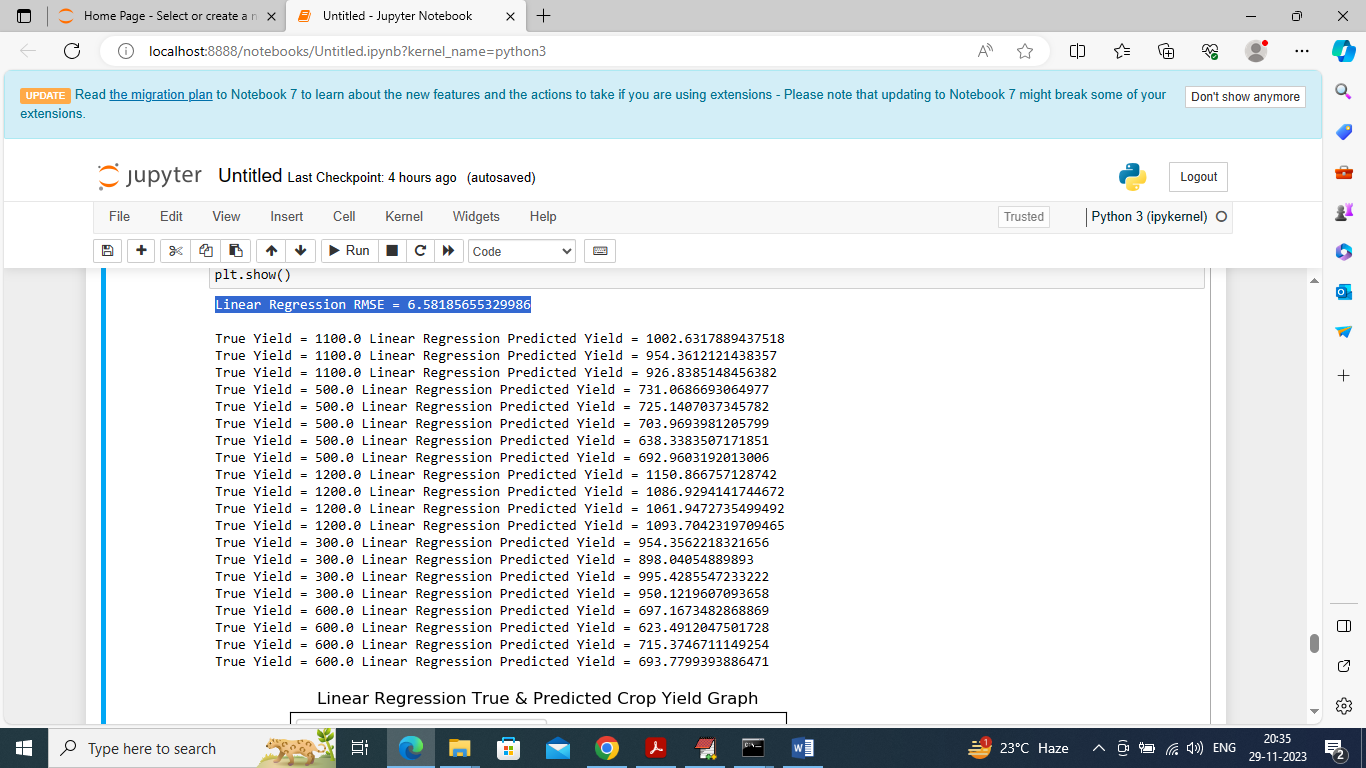
In above screen can see Decision Tree RMSE error rate as 0.71% and then can see true and predicted crop yield from decision tree and can see both true yield and predicted yield values are too close. Below is the decision tree predicted values comparison graph



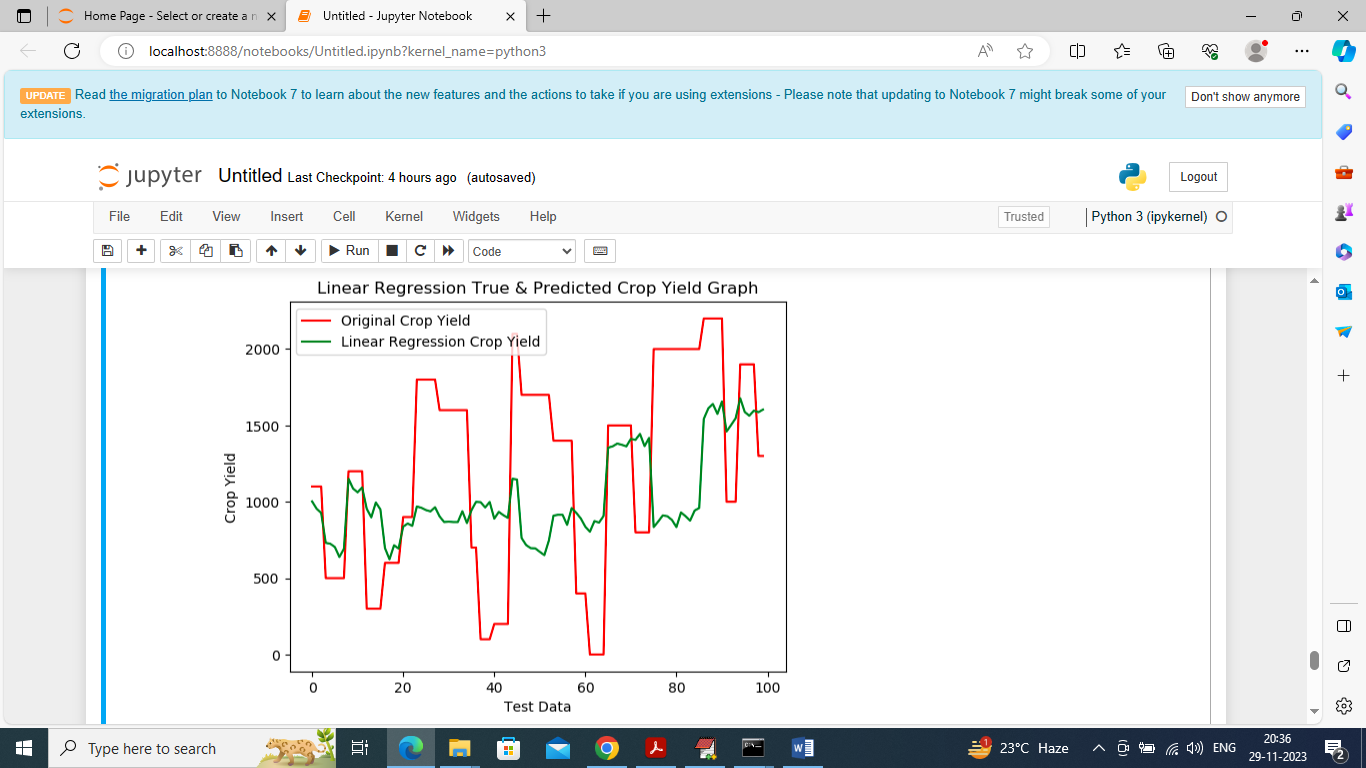
In above graph x-axis represents test data count and y-axis represents crop yield and then red line represents True Yield and green line represents Decision Tree crop yield and can see both lines are fully overlapping so predicted and true test data crop yield are too close



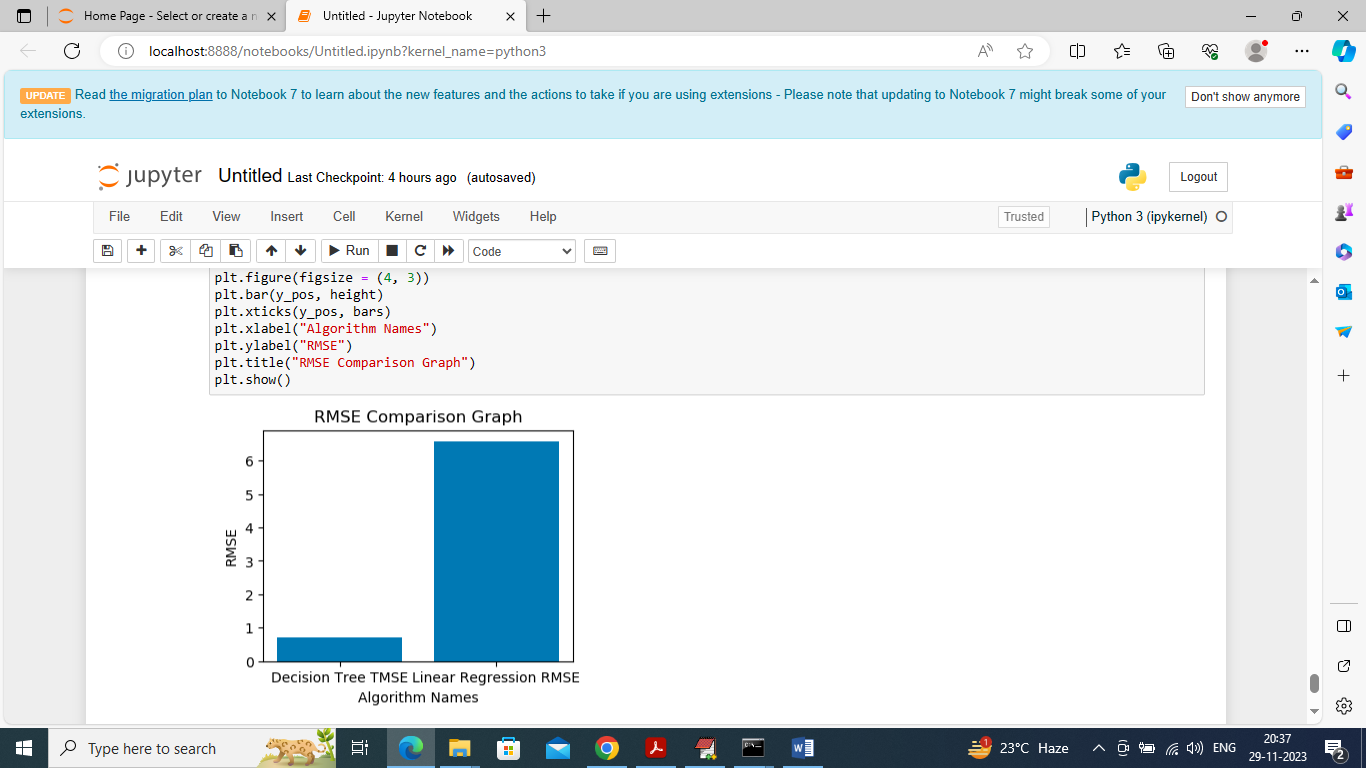
In above screen training Linear Regression algorithm and after executing above block will get below output



In above screen Linear Regression got 6.50% RMSE error and then can see test data original crop yield and linear regression predicted crop yield and below is the linear regression comparison graph



In above graph can see there is huge difference between true test data yield and green colour linear regression predicted yield and can say Linear Regression is not accurate



In above screen displaying comparison graph between Decision Tree and linear regression RMSE error where x-axis represents algorithm names and y-axis represents RMSE and in both algorithms Decision Tree got less RMSE error.

So by using above code we can see predict accurate crop yield for any country and crop using Decision Tree Algorithm