**Machine Learning with Pyspark**

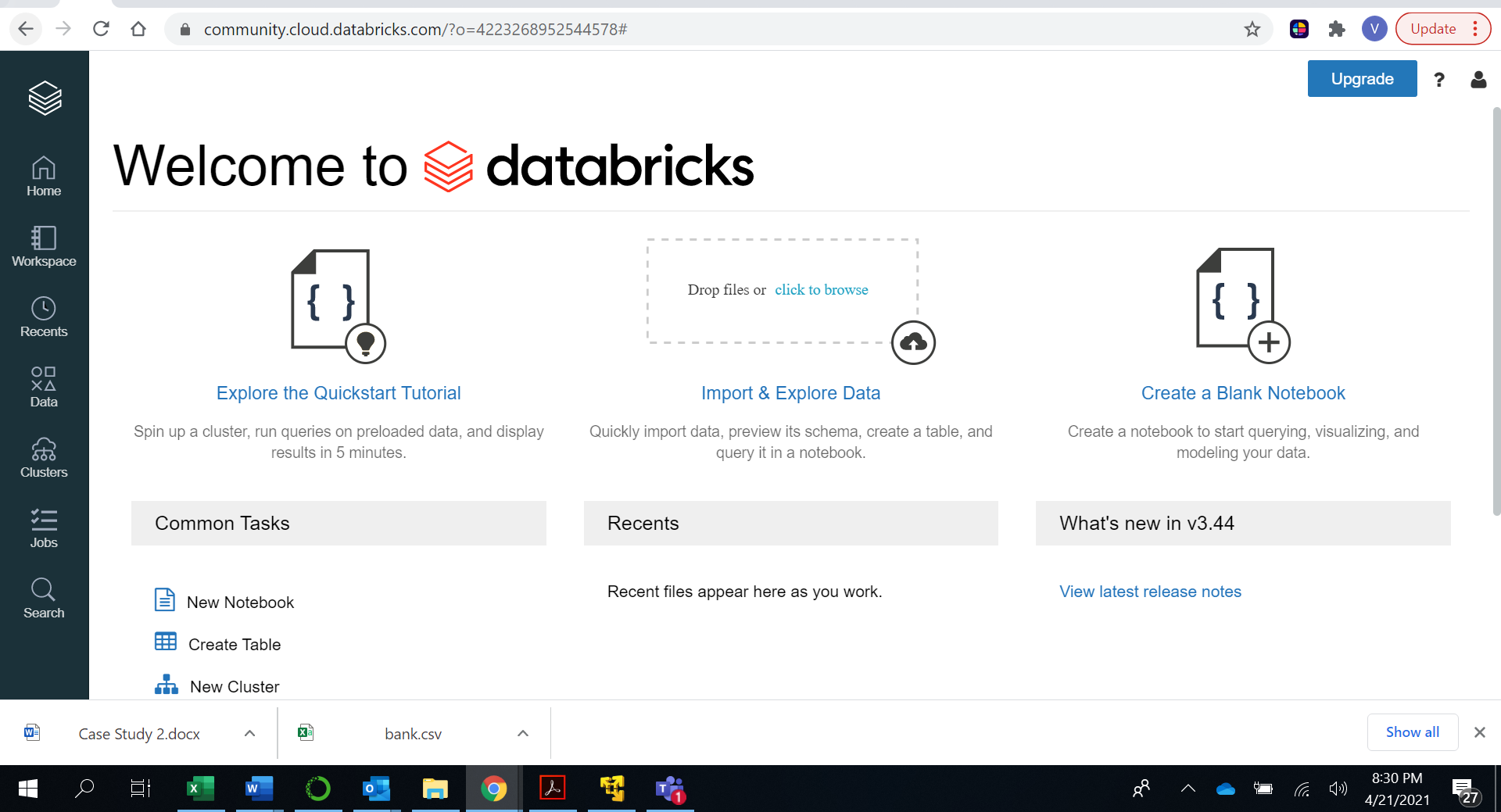
**Case Study Partner: Vivek Narasimhamurthy**

Part 1: Create a databricks account.

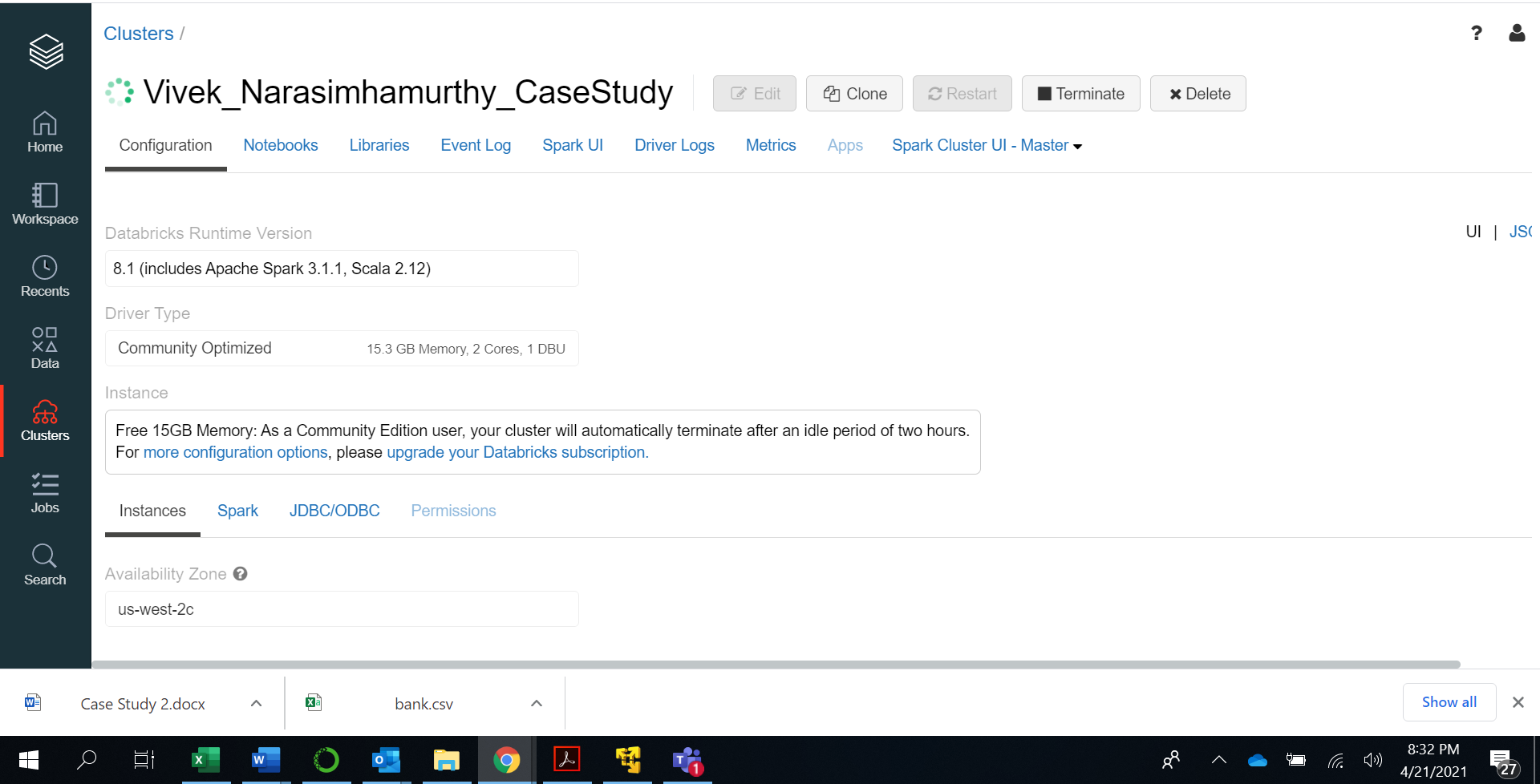
Step 1: Open the link:

Part 2: Create a cluster node.

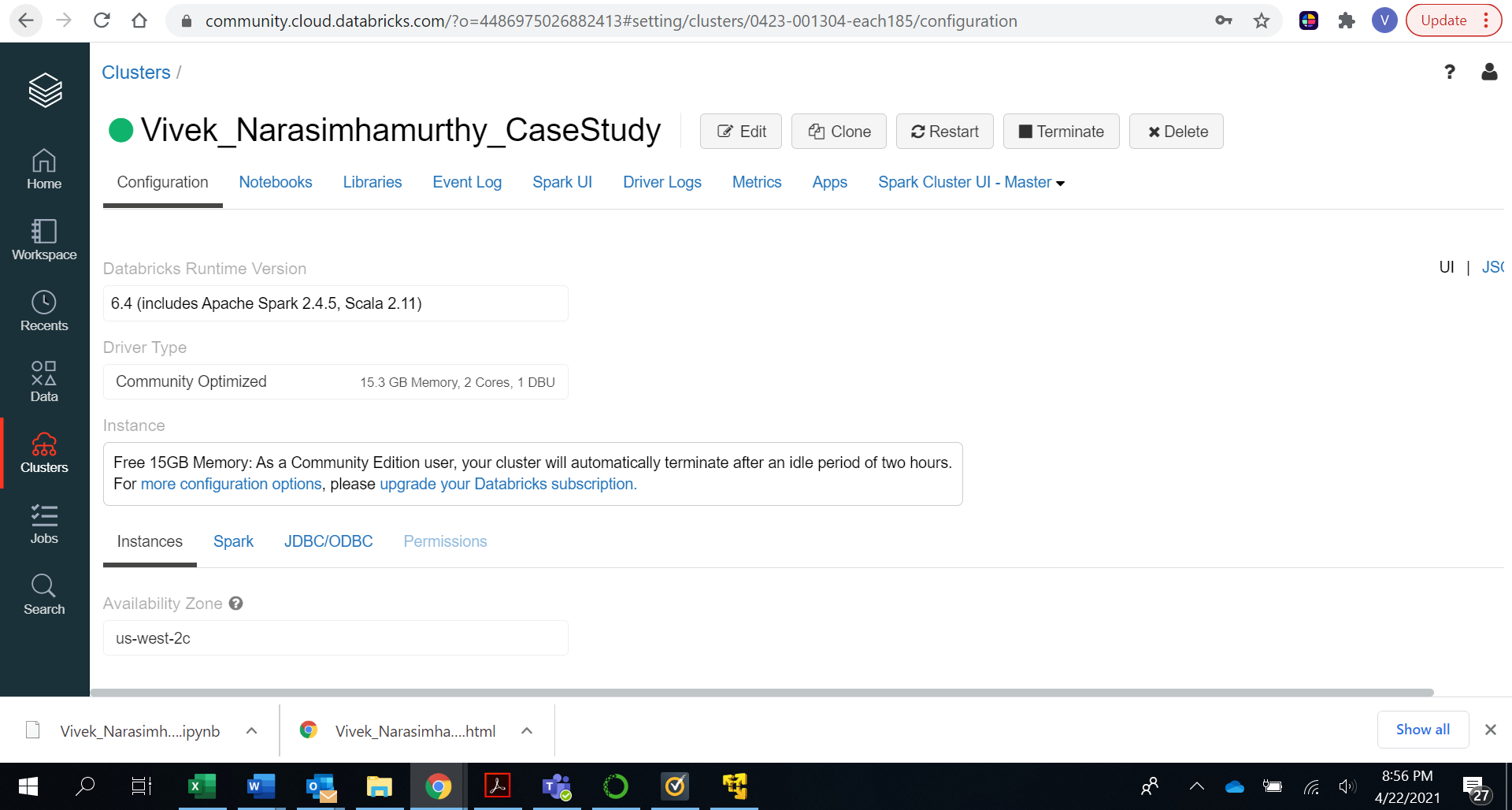
Step 3: Click on the “Clusters” icon as shown in the screenshot below:



Step 4: Name the cluster as “FirstName\_LastName\_CaseStudy”.



Step 5: Green sign shows that a cluster is created and running successfully.

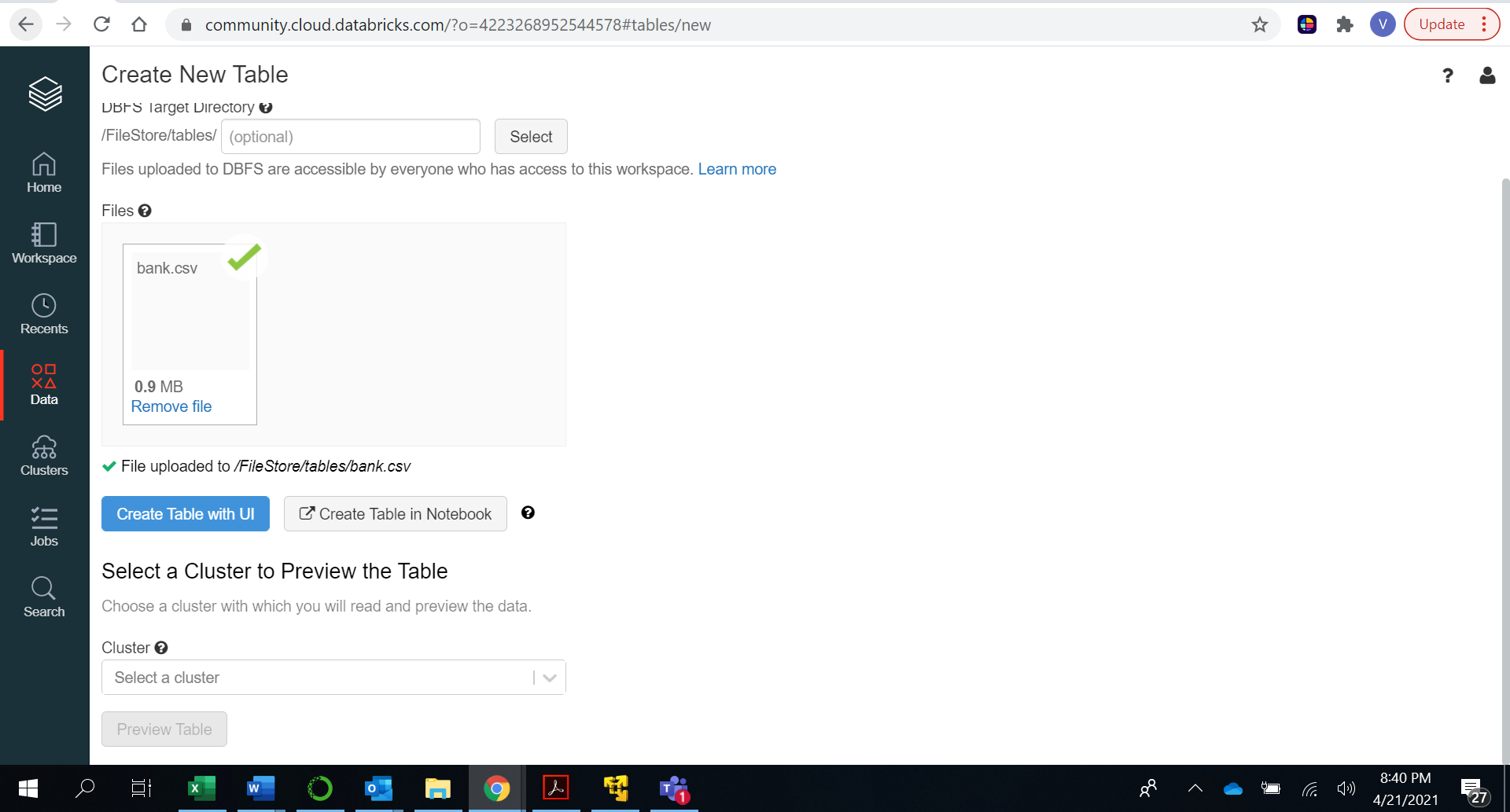


Part 3: Create Datatable.

Step 6: Download the “Bank.csv” file from eLearning.

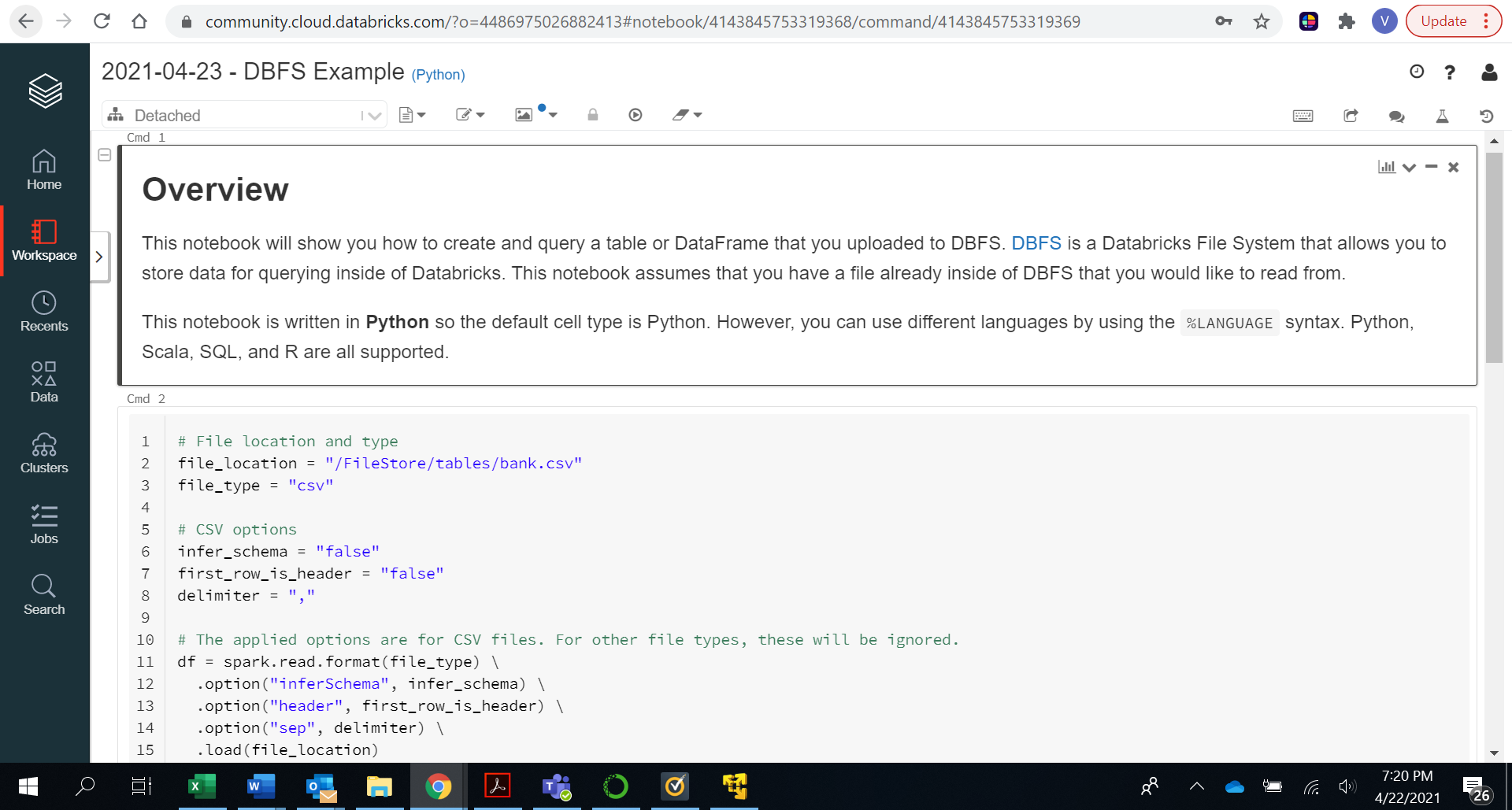
Select the “Data” icon from the left panel and click on “Add Data”.

Step 7: Click on browse and select the “bank.csv” file that you downloaded from eLearning.



Step 8: Click on “Create Table in Notebook”.

Step 9: The file location can now be seen.



**Answer : /FileStore/tables/bank.csv**

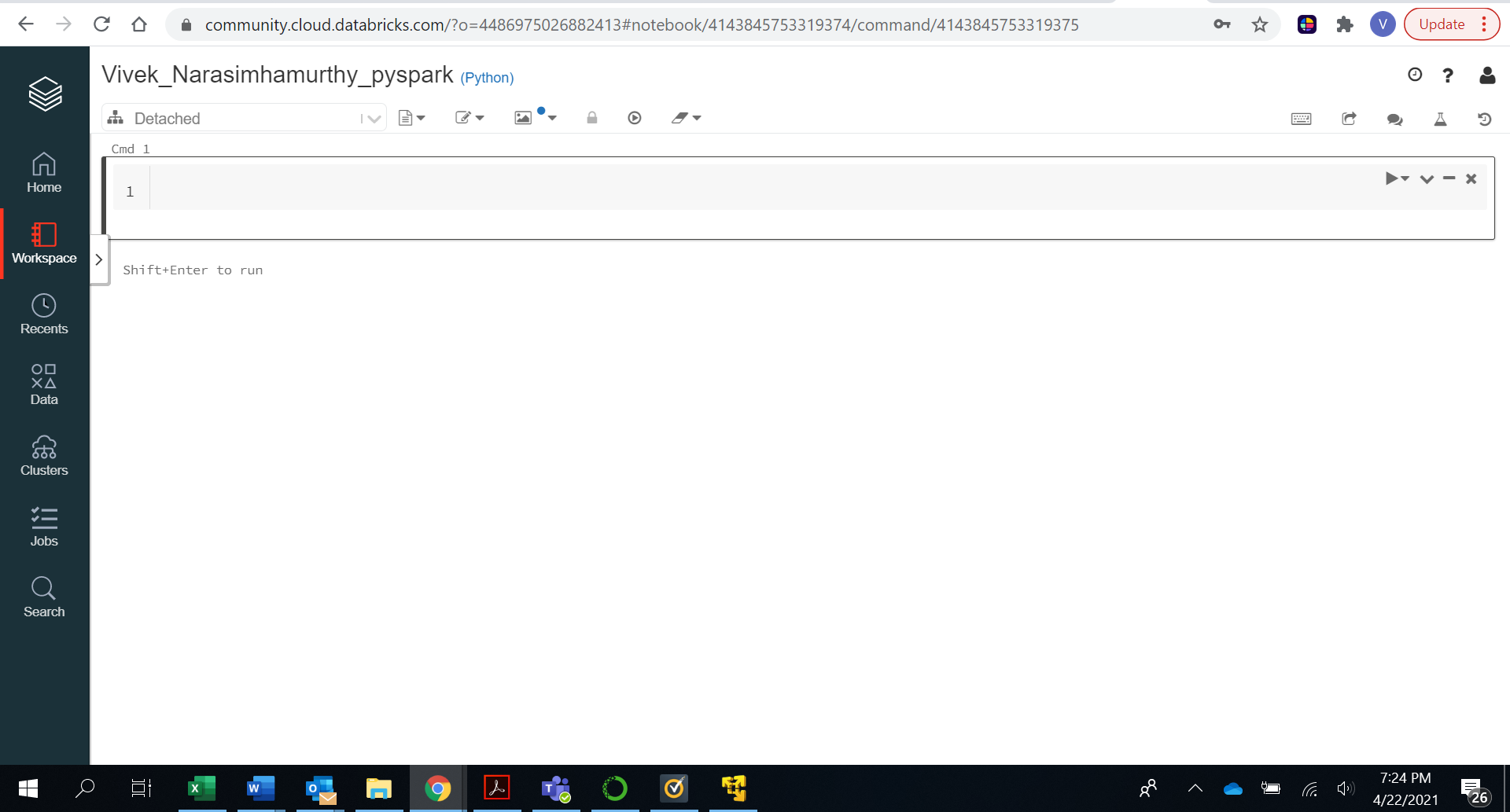
Part 4 : DataFrame Operations.

Step 10: Create new Notebook. Follow the steps below.

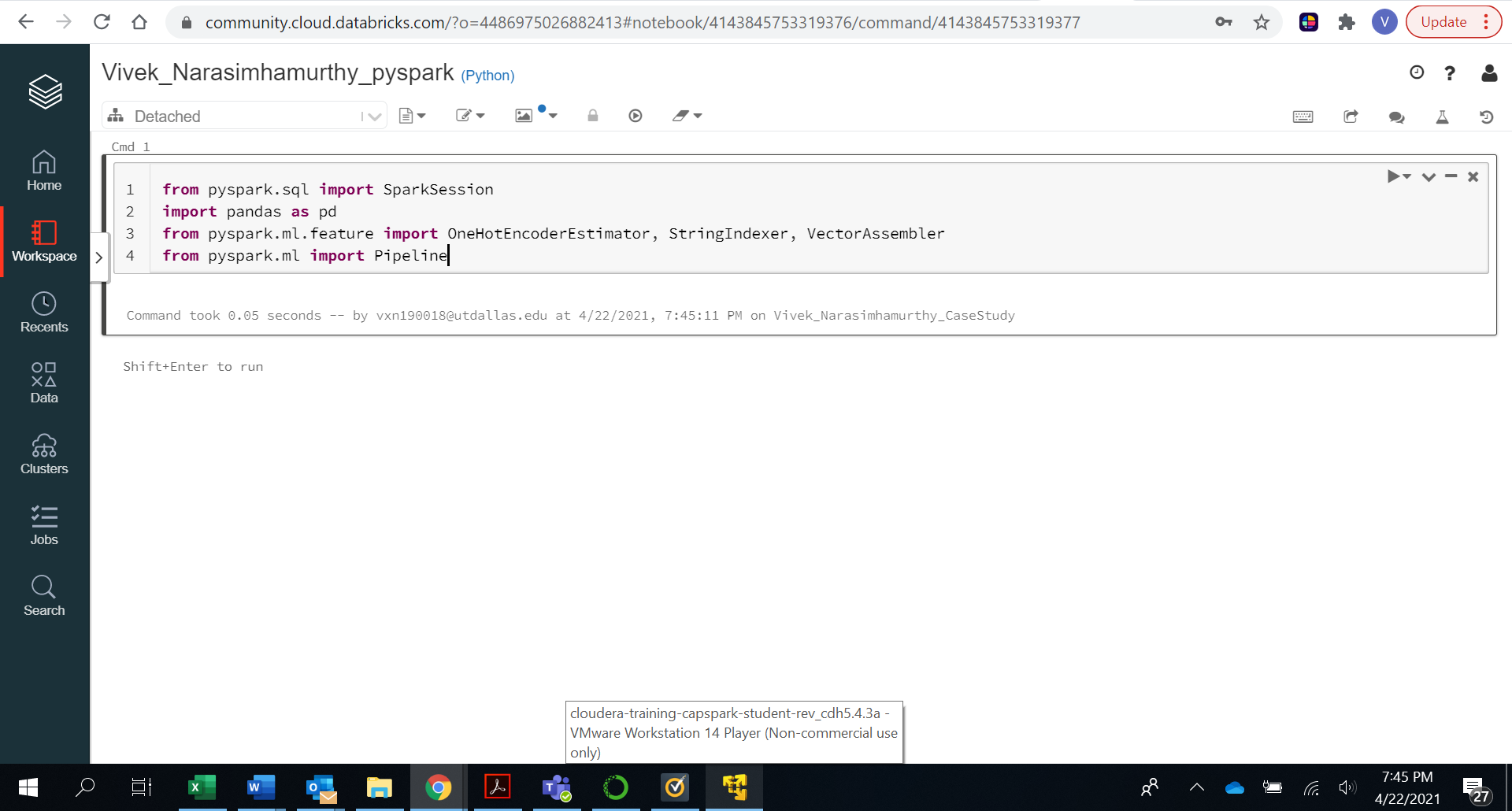
Step 11: Create the notebook with “FirstName\_LastName\_Pyspark”.

Language: “Python”

Cluster: Select the cluster we have created in part 1.

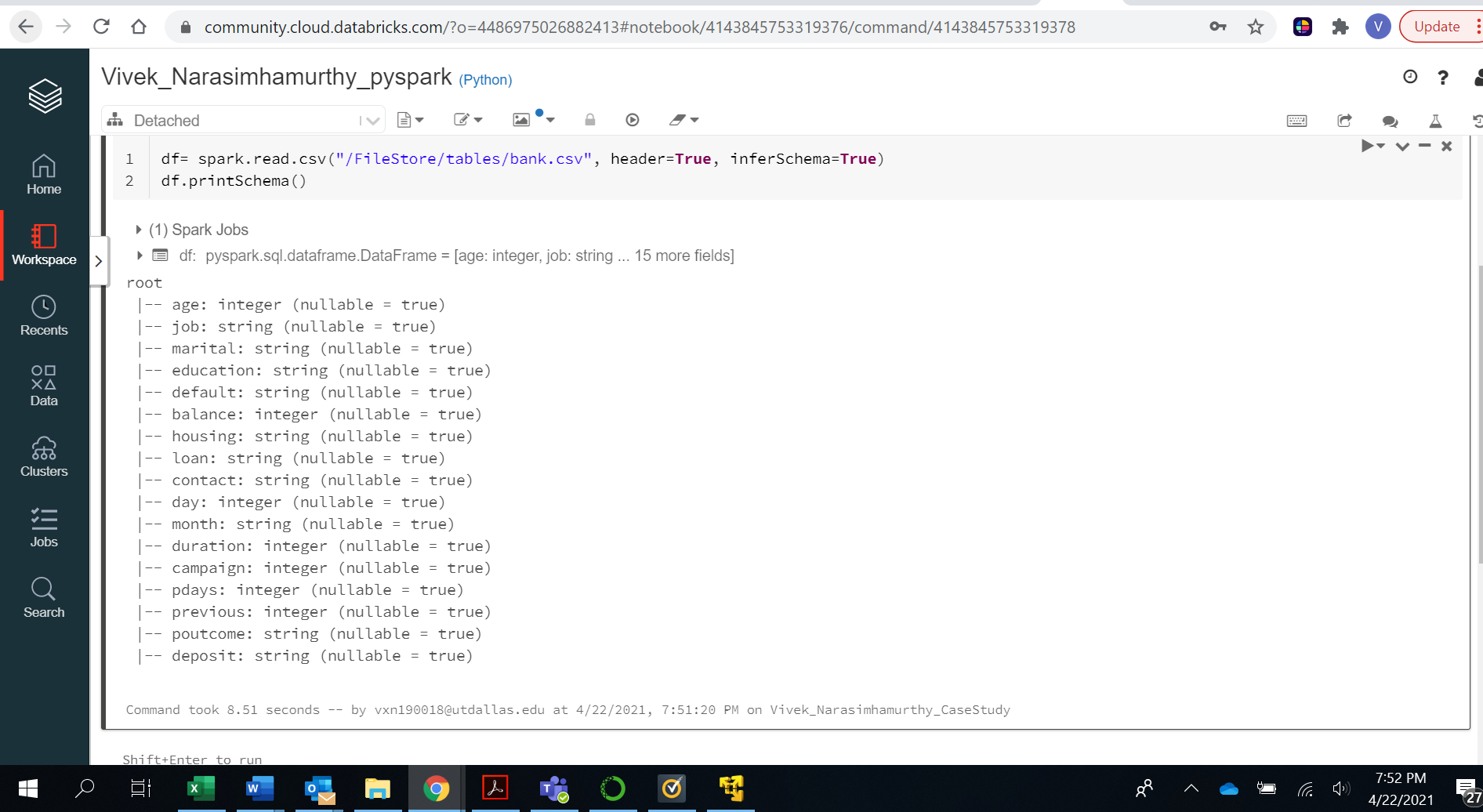


Step 11: Import the required libraries from Spark.



Step 12:

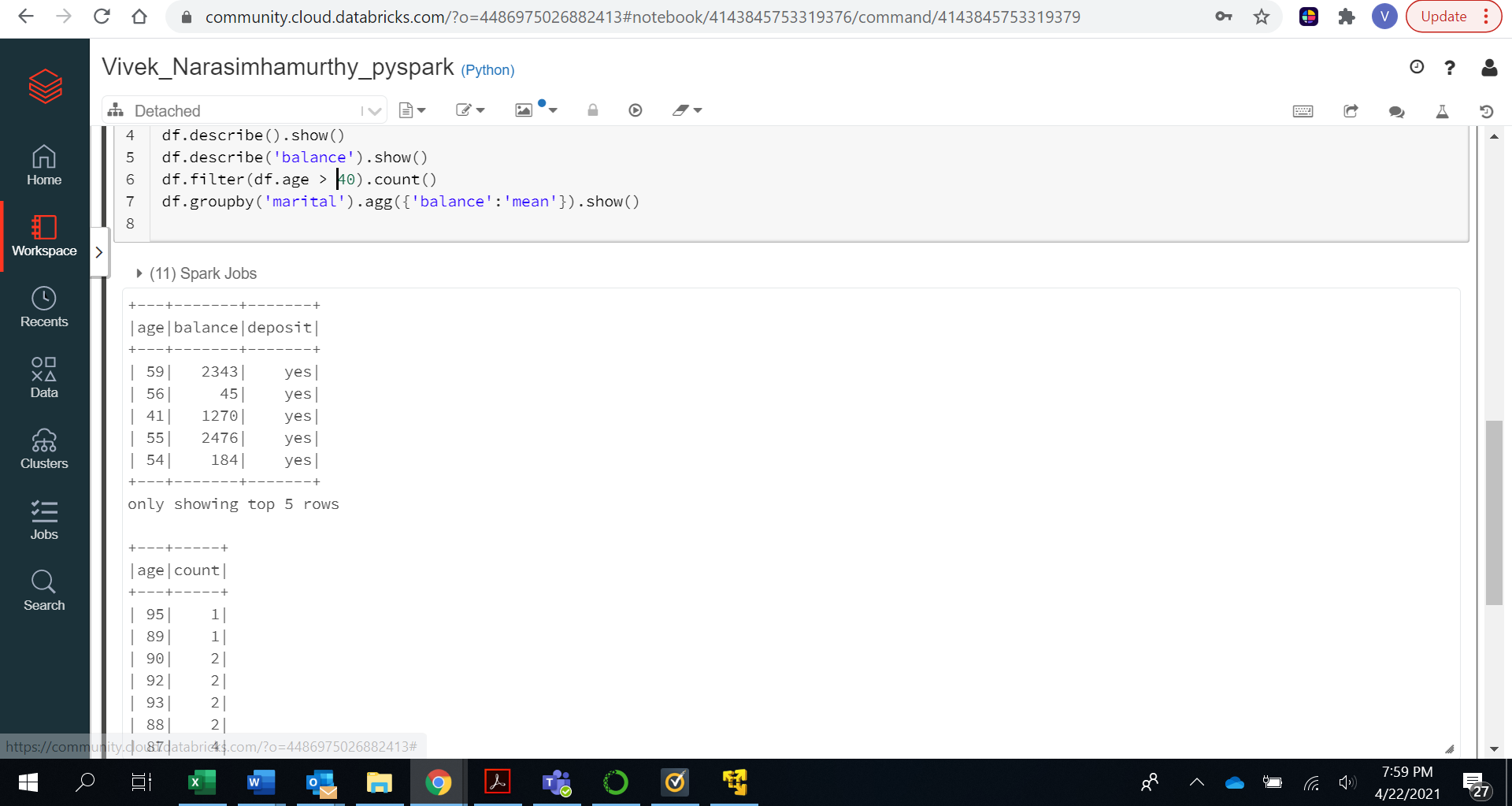
Click on “+” sign at the end of the cell.



**Answer : Input variables are:**

Age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome.

Step 13: Type the code below in the next cell and Run the cell.

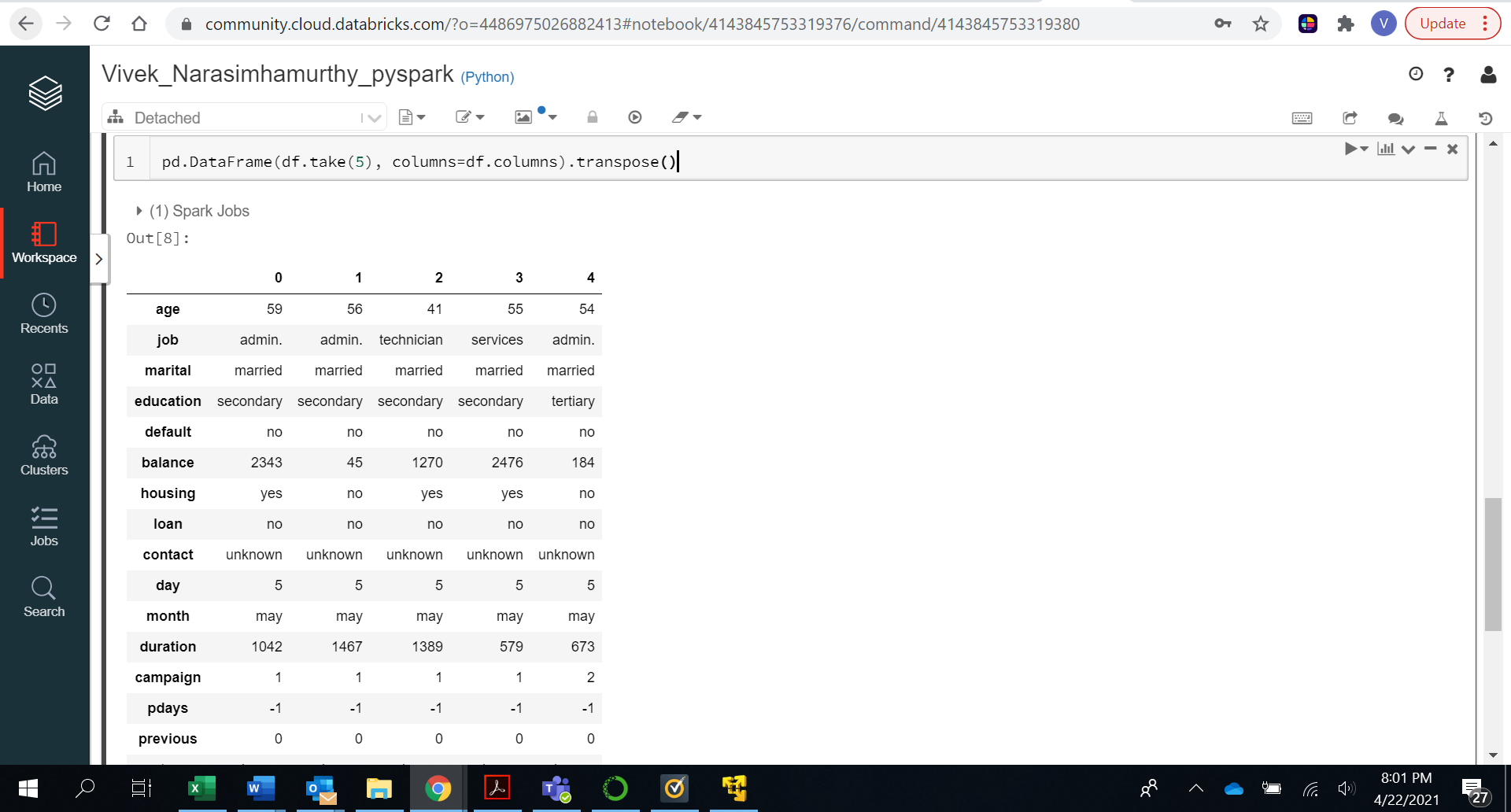


Answers : Select statement selects input based on criteria. Groupby groups rows based on a column/s.

There are 11162 people with age greater than 40.

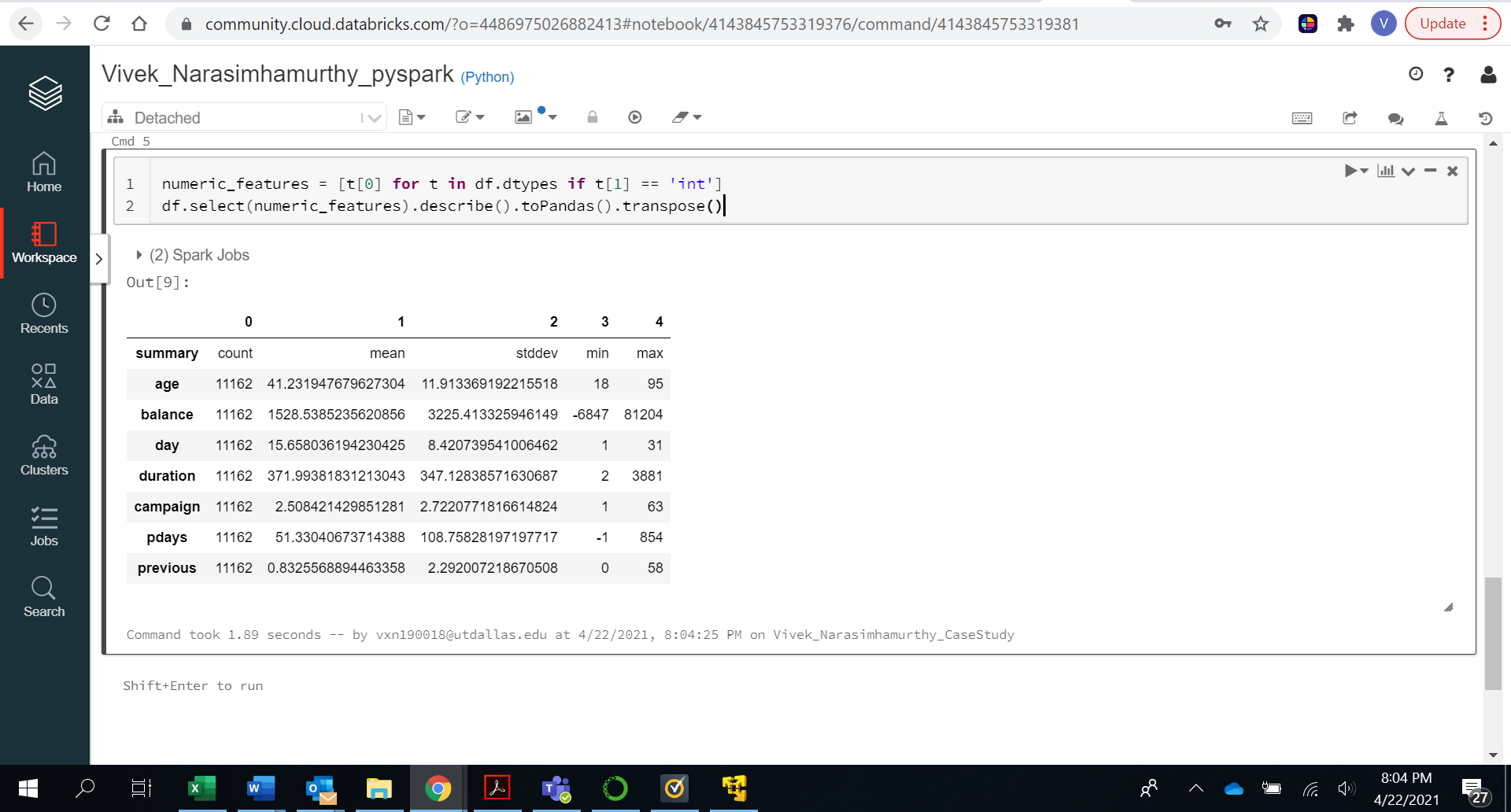
We have specified to show only 5 rows inside the parenthesis.

Step 14: Execute the following code:



Answer : All the rows have become columns and vice-versa.

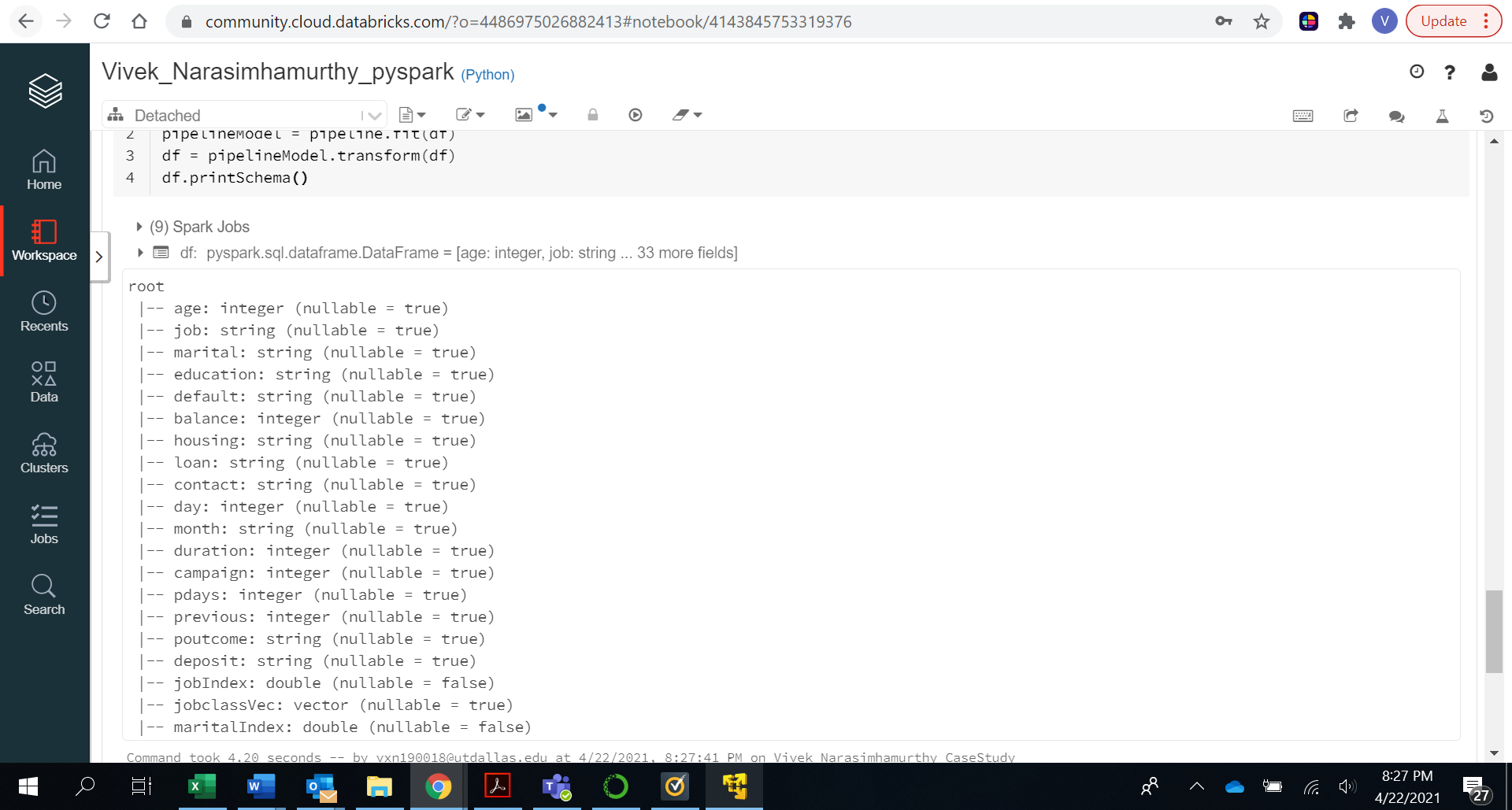
Step 15: Execute the following code:



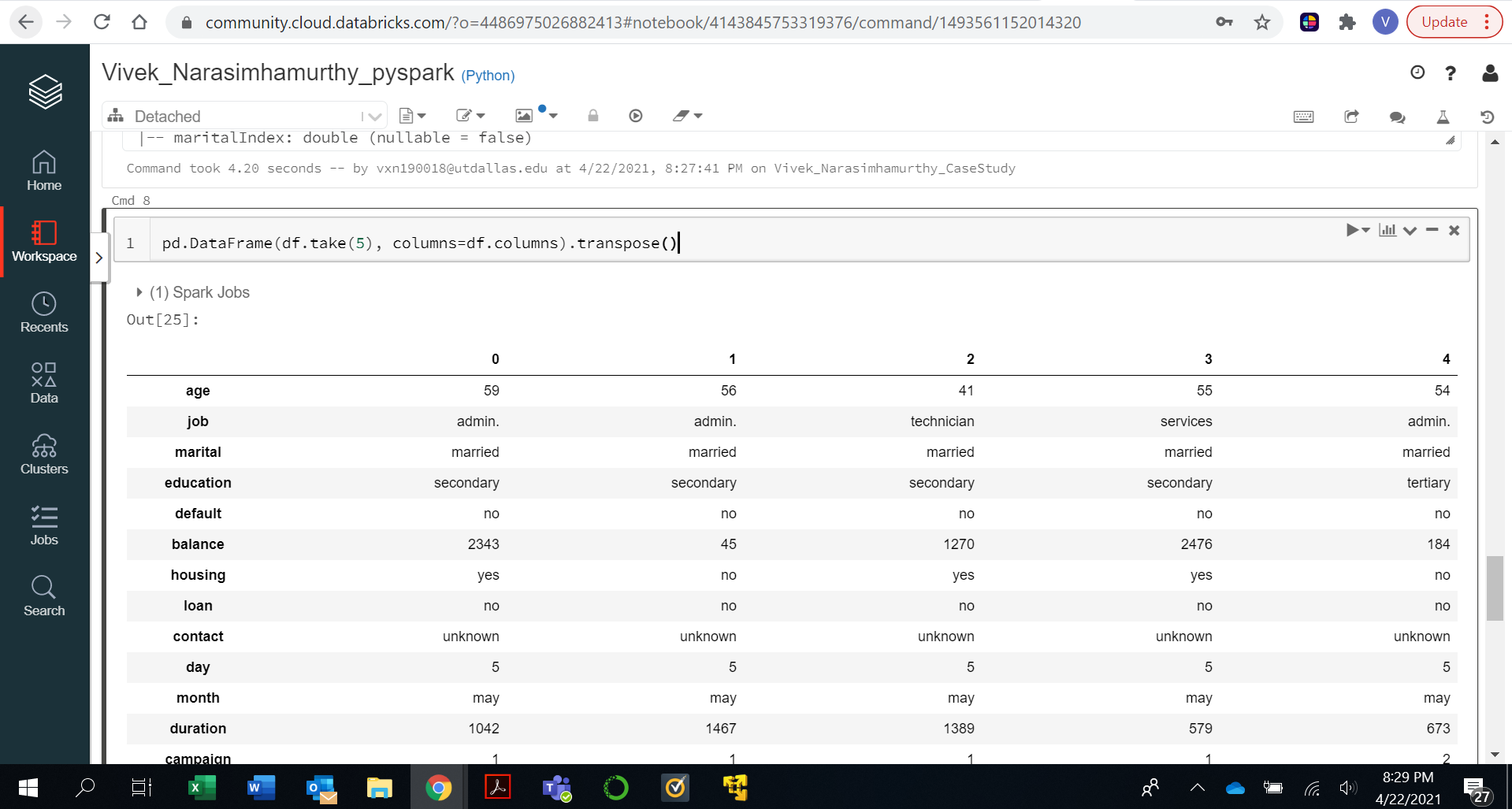
Step 16: Execute the following code:



Step 17: Execute the following code:

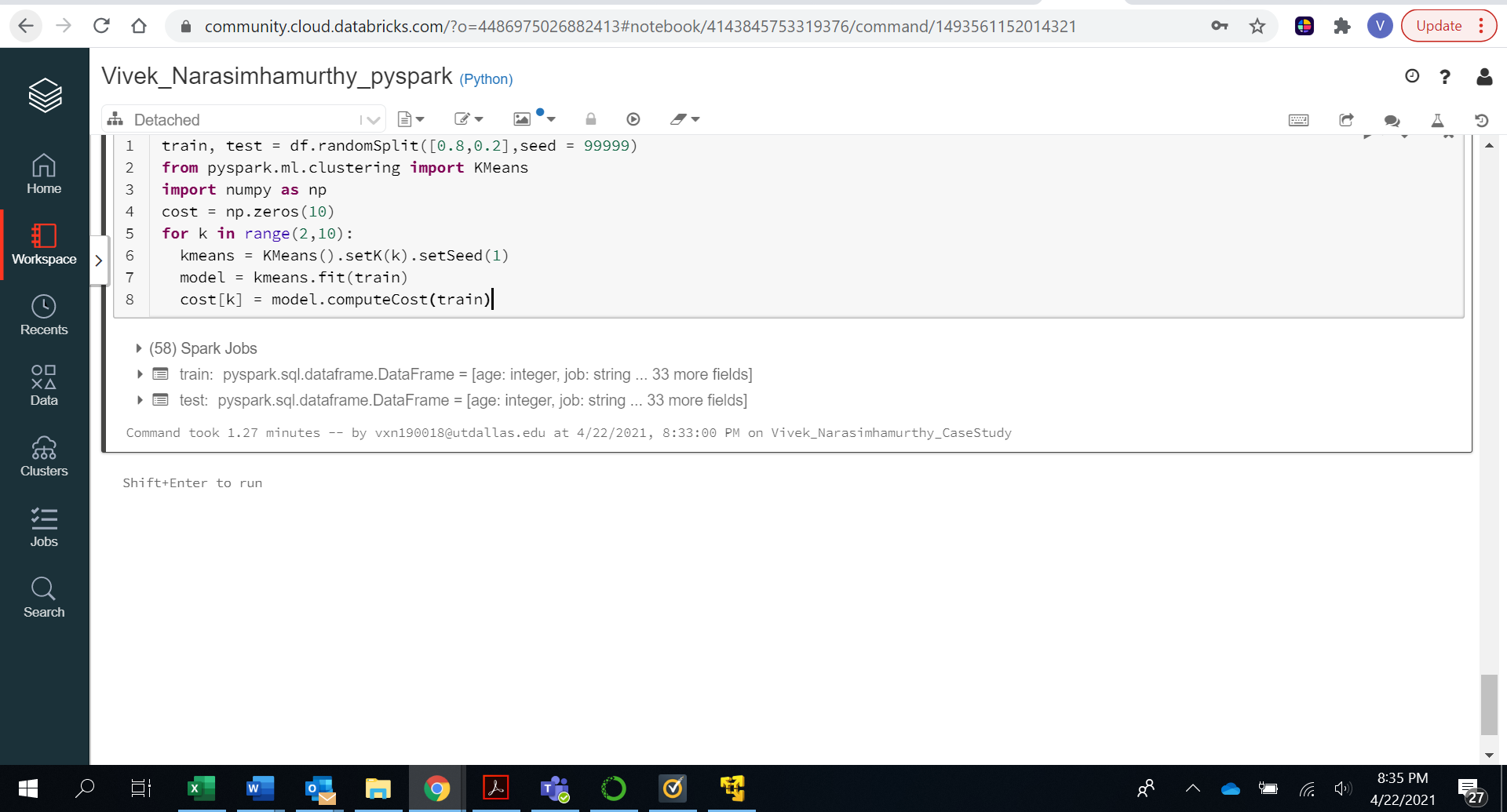


Step 18: Execute the following command:

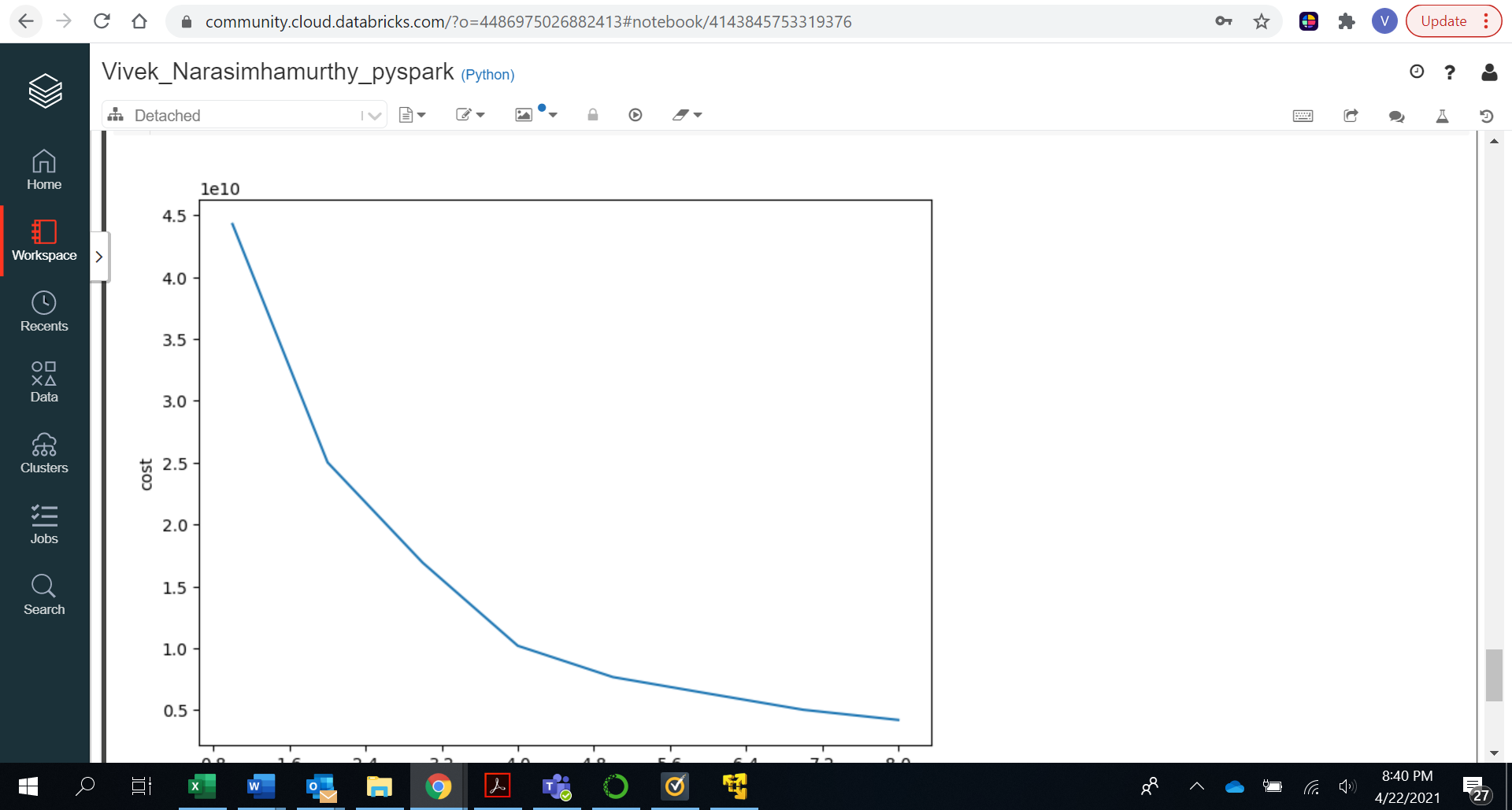


Answer : To convert column into rows or vice versa we use transpose function.

Step 19: Execute the following code:

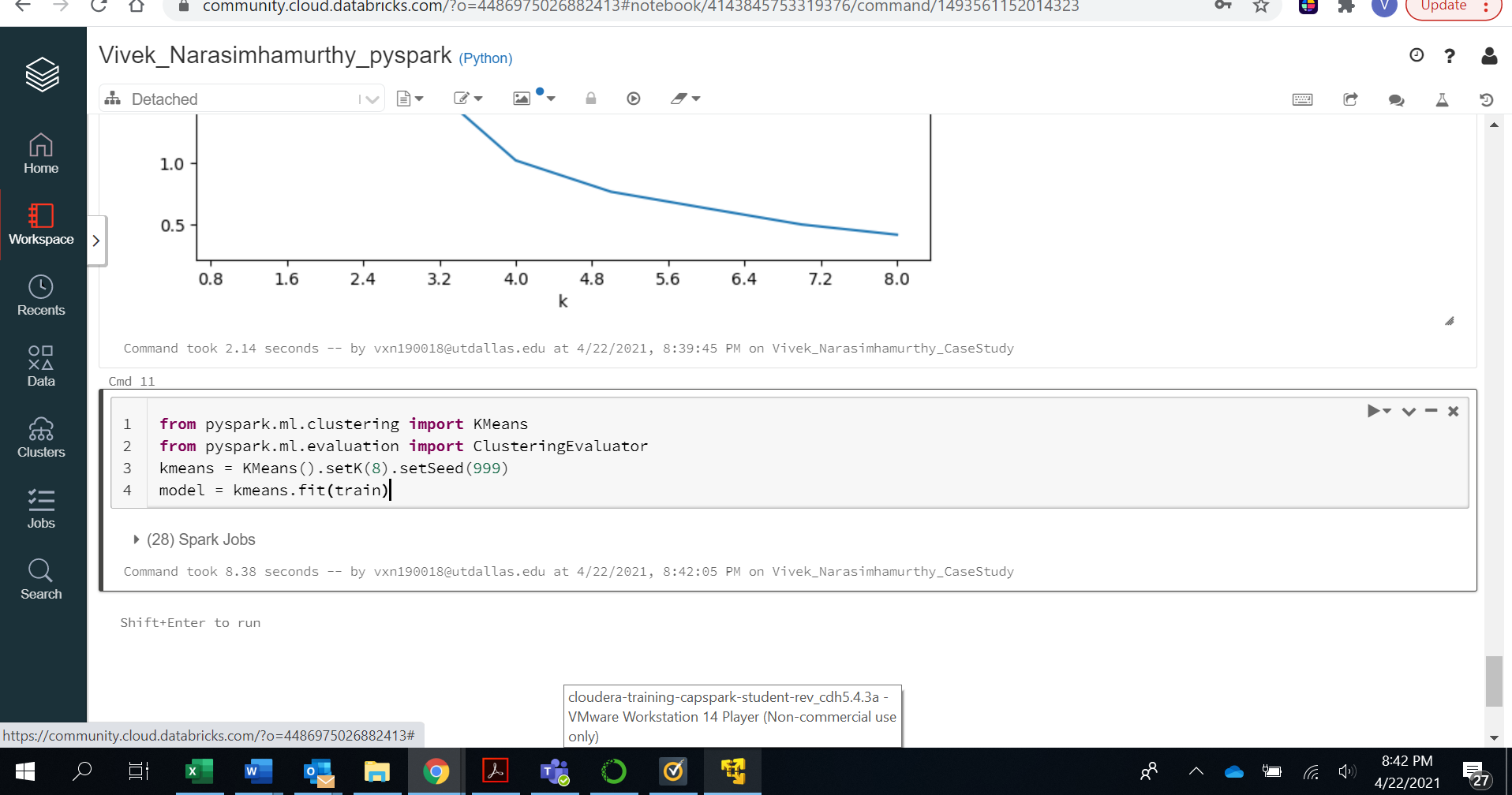


Step 20: Execute the following code:

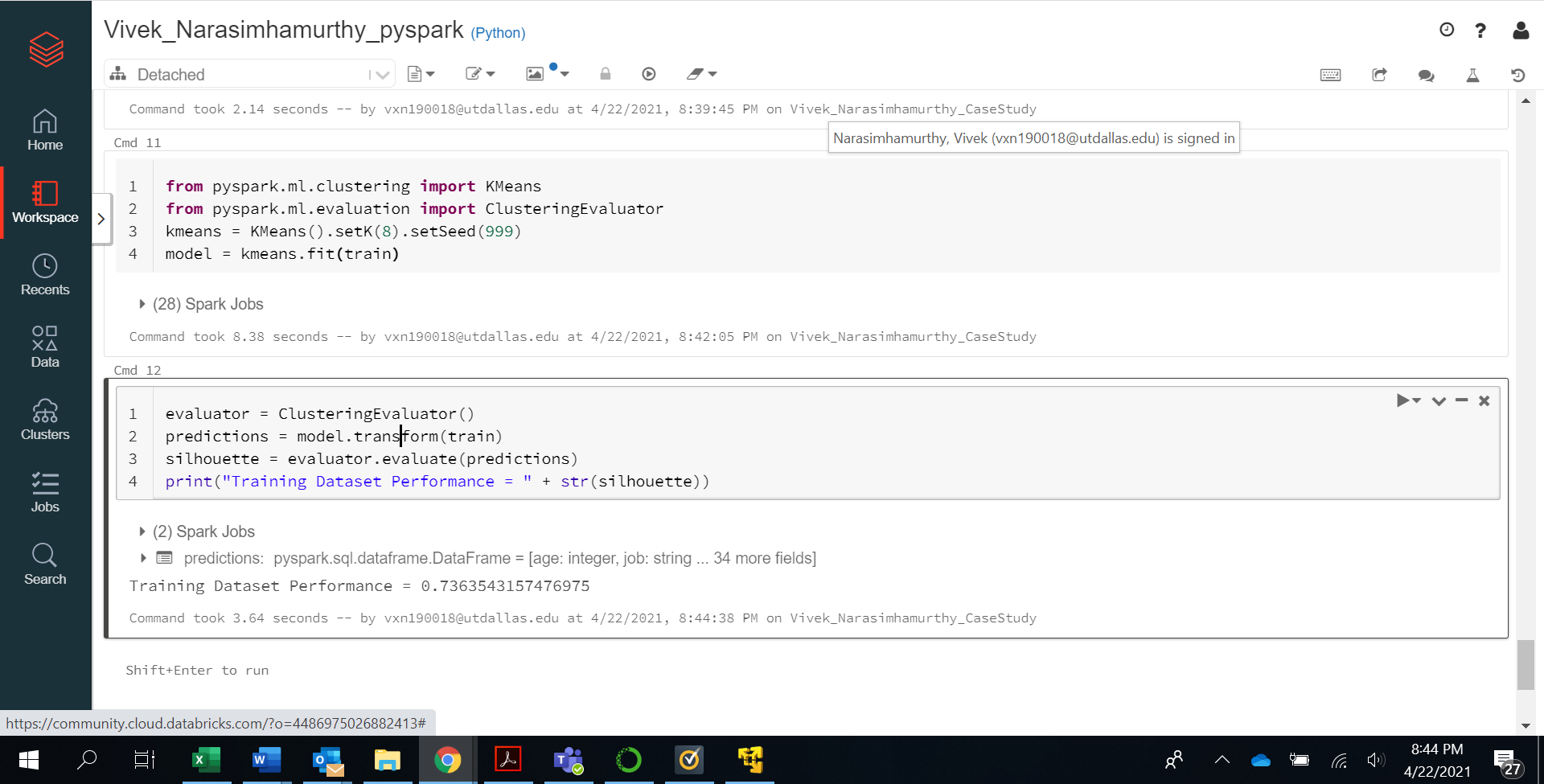


Answer : As the number of clusters increases cost/error of misclassifying to a wrong cluster decreases but it attains saturation after some point.

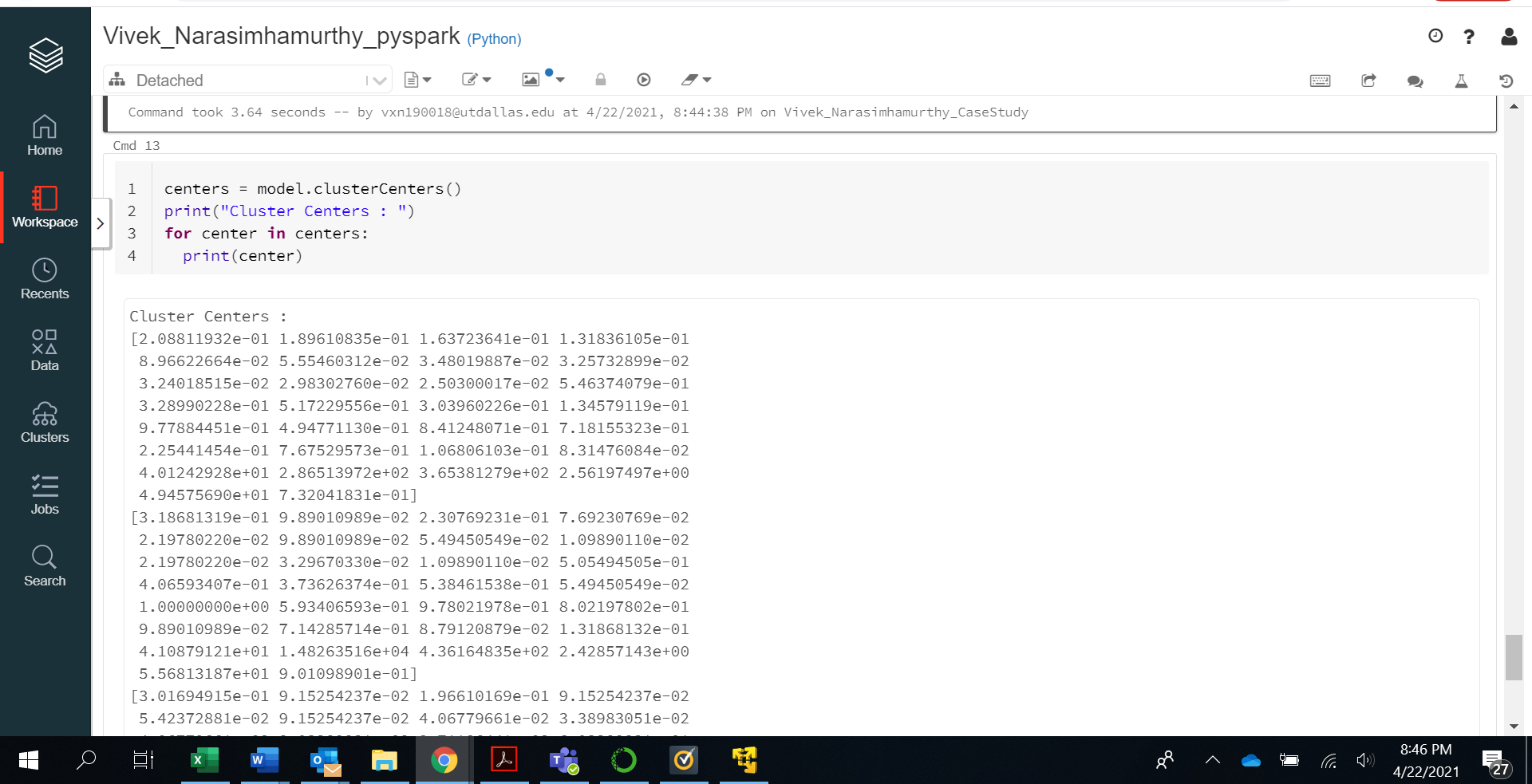
Step 21: Execute the following code:



Step 22: Execute the following code:



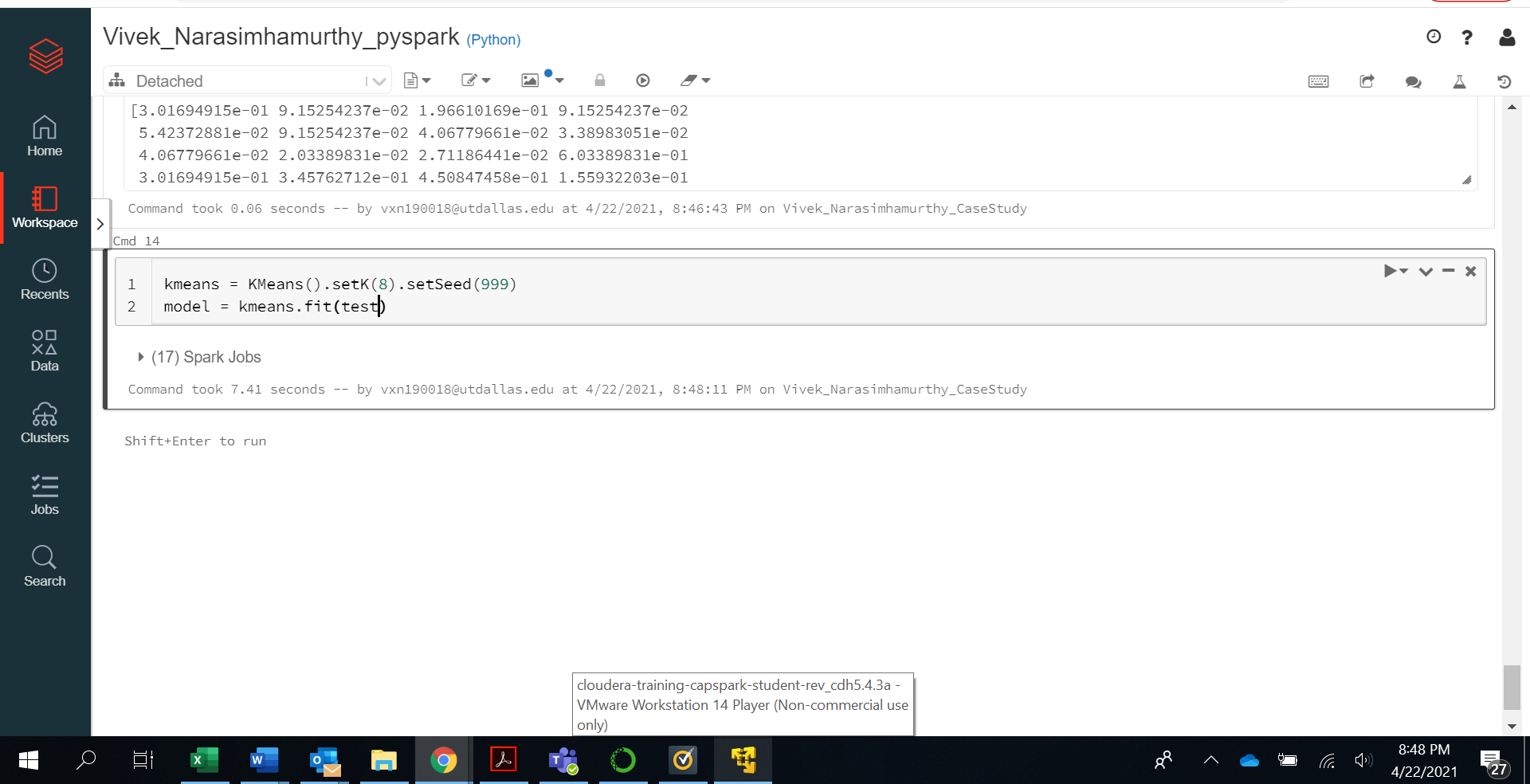
Step 23: Execute the following code:



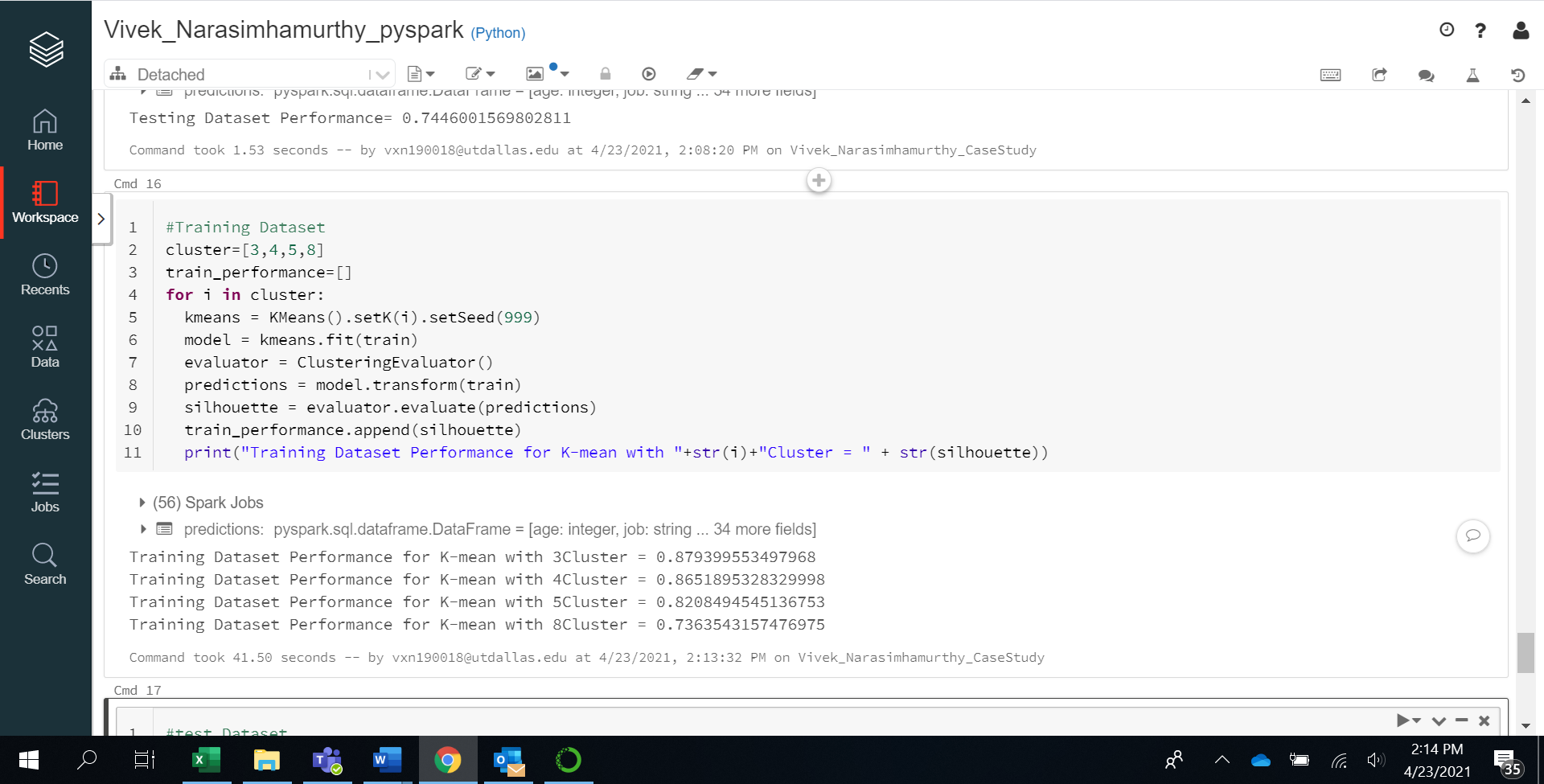
Answer : The centroid is the point representing the center of the cluster. The average of all the points/members of the cluster is taken as centroid of the cluster.

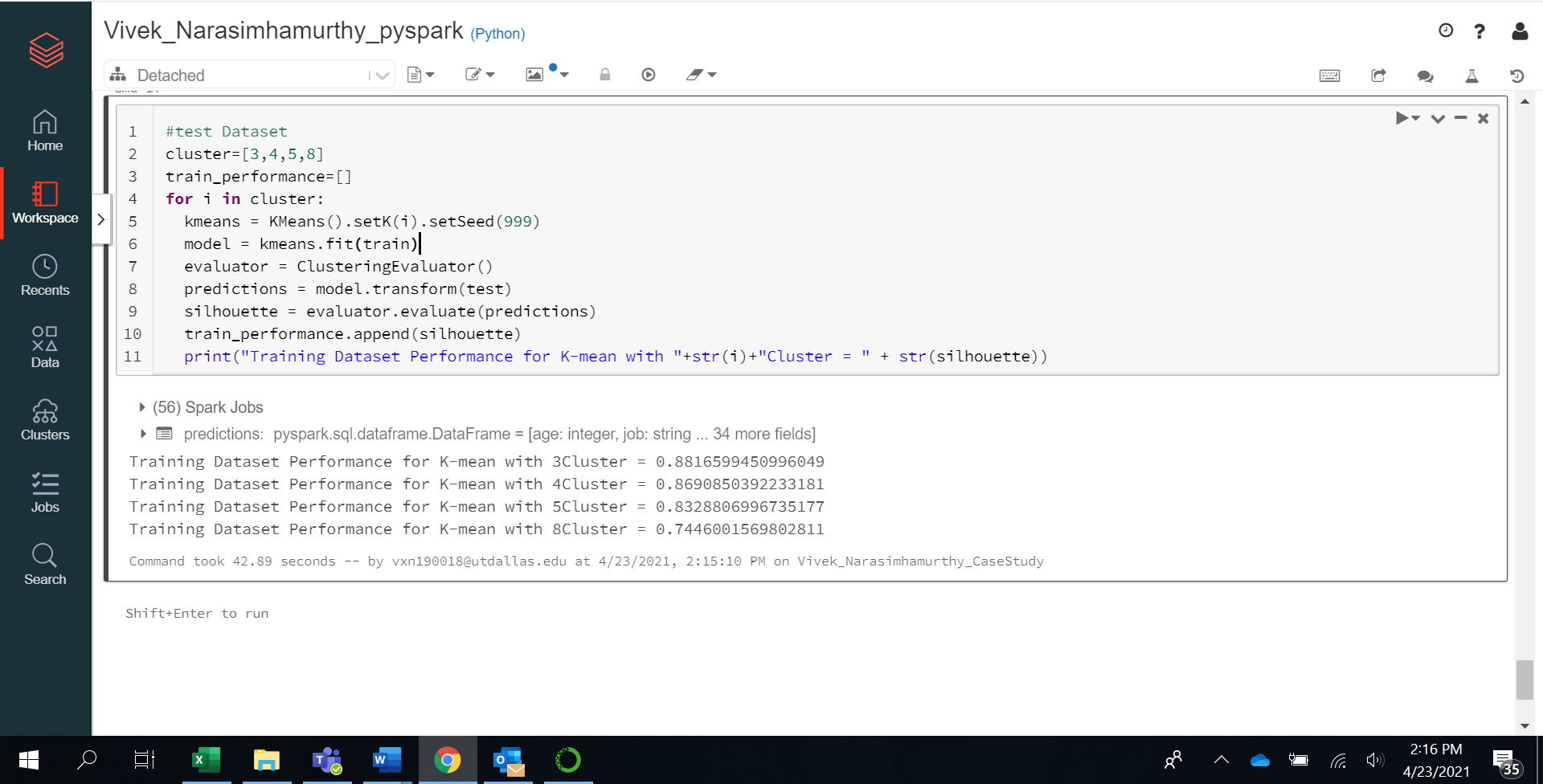
Step 24:

Execute the following code:



Step 25: Execute the following code:





Answers :

As expected Clustering as 3 segments gives us better performance in both training and test data sets.

The performance indicate the extent to which misclassification happens. That is wrongly assigning a point to a cluster. Ex – point 1 might belong to Cluster A but it might be wrongly assigned to Cluster B.

With 3-Mean Clusters a performance of 87.93% on train dataset and 88.16% on test dataset is the best we can achieve.

Step 26:

Export the file as “HTML” and save that with your “Firstname\_Lastname” naming convention.