**Machine Learning with Pyspark**

**Student Name: Prem Kumar Pulluri Student Id: 2021470494**

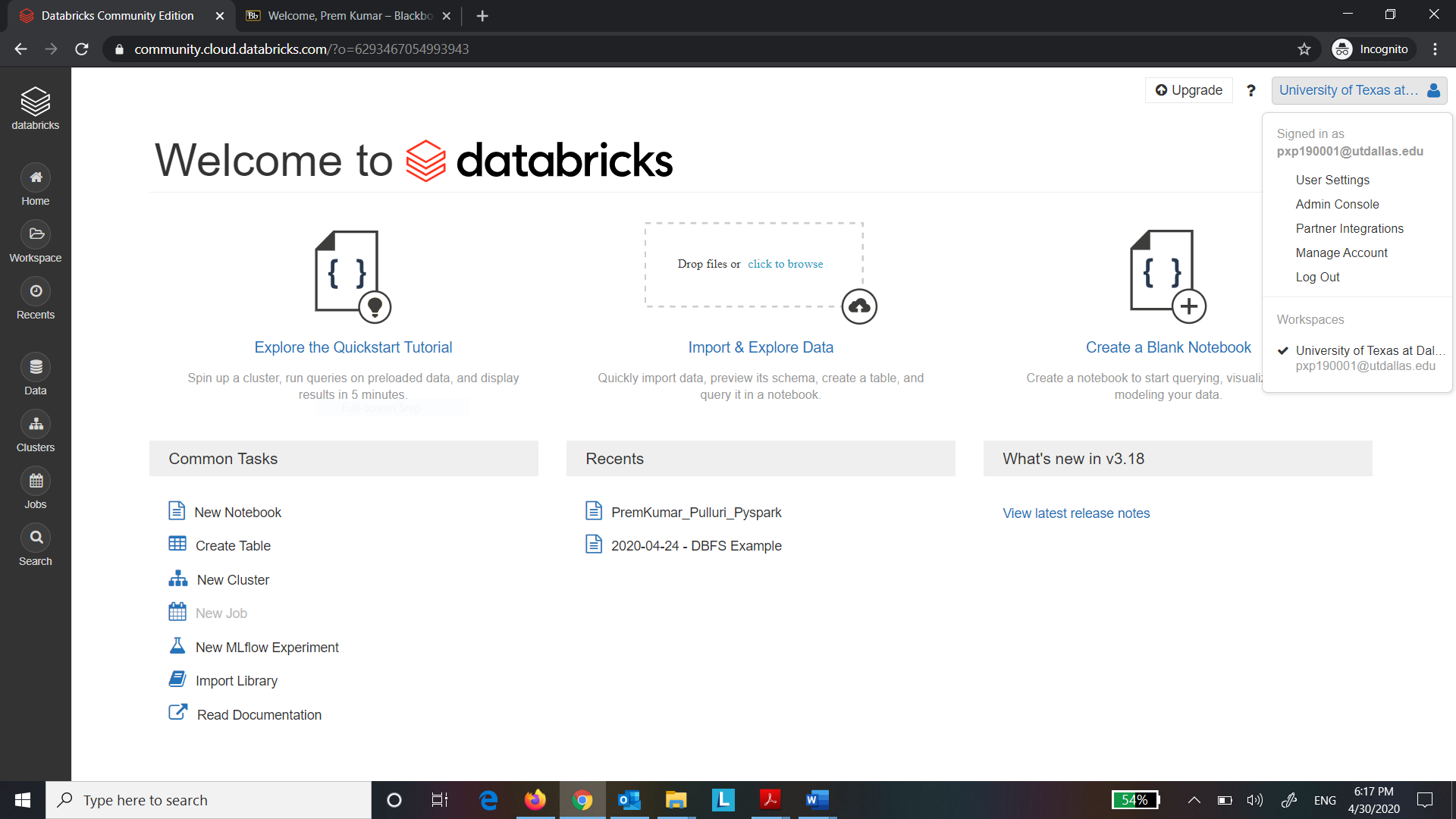
**Date: 04/30/2020**

# Objective:

PySpark is a great language for performing exploratory data analysis and building machine learning pipelines. MLlib is a Spark implementation of useful machine learning (ML) functionality. The objective of this case study is to become familiar with Pyspark. We will load the data in Pyspark and build a machine learning model using the K-Means algorithm.

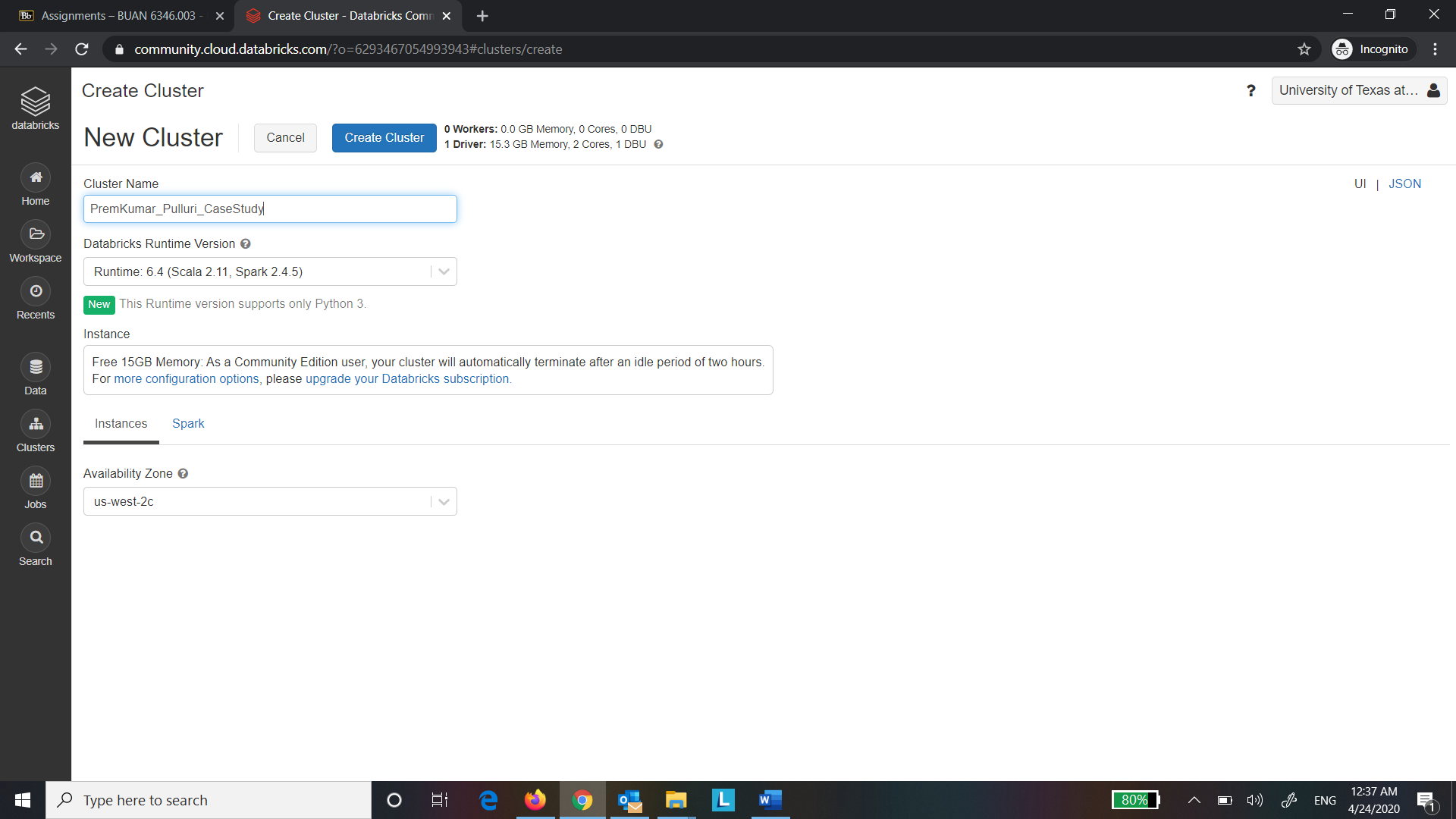
Part 1: Create a databricks account.

Q. Login to your databricks account and take a screenshot of the Home page output and paste it below.



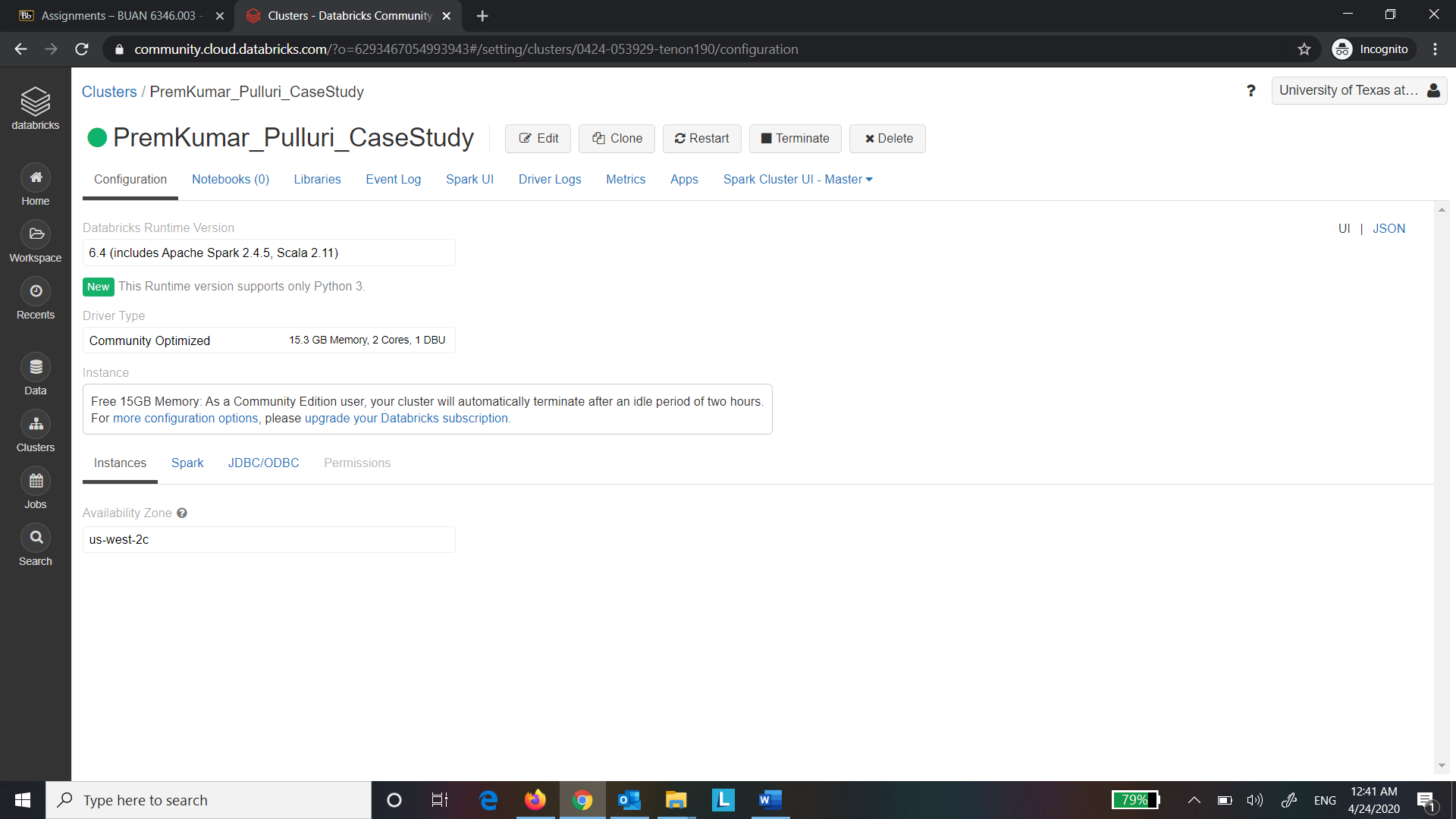
Part 2: Create a cluster node.

Q. Take a screenshot of the Create Cluster page and paste it below.



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

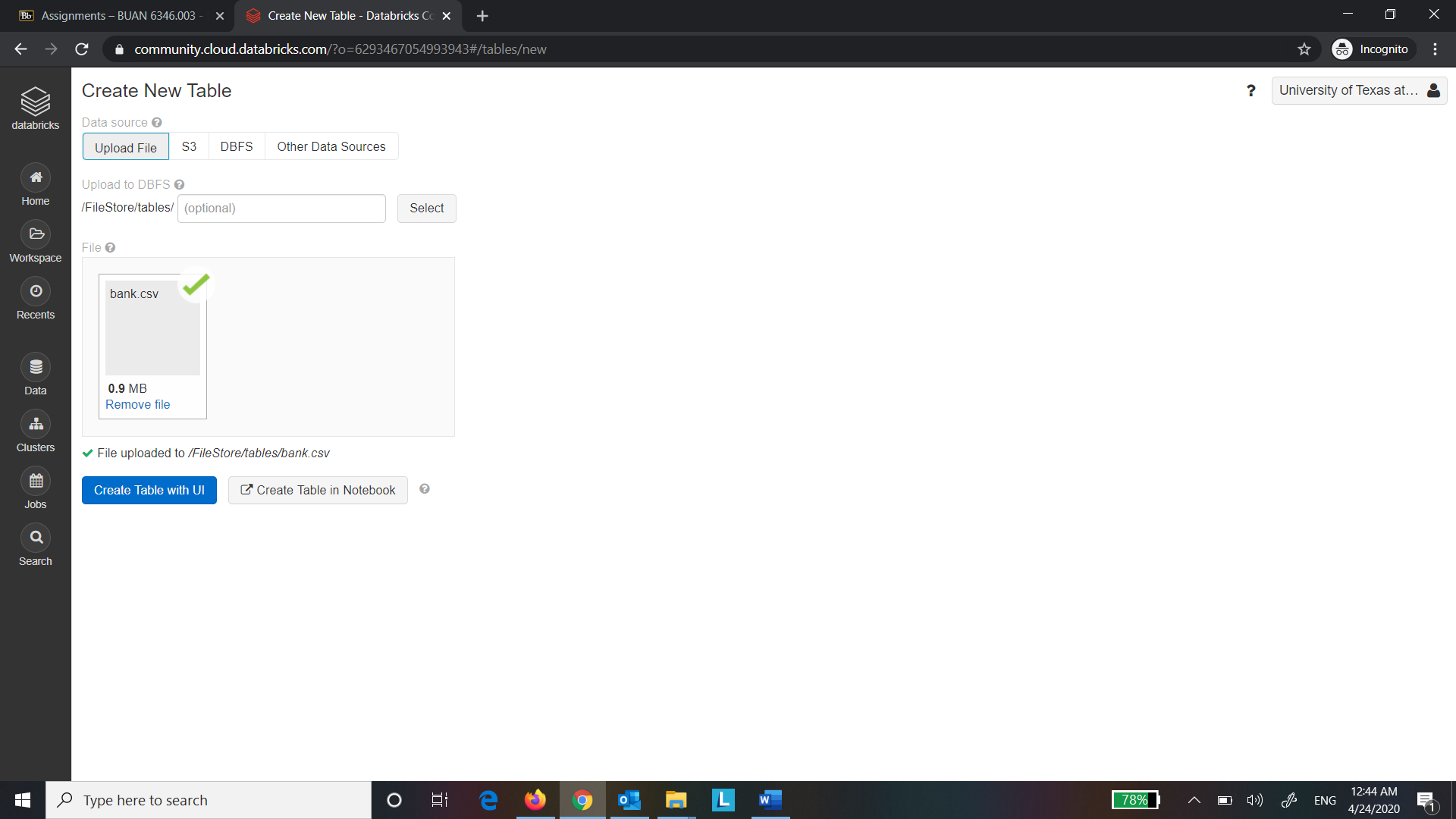
Q. Take a screenshot of the running Cluster Page and paste it below.



Part 3: Create Datatable.

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

Take a screenshot of the uploaded file and paste it below.



Step 9: The file location can now be seen.

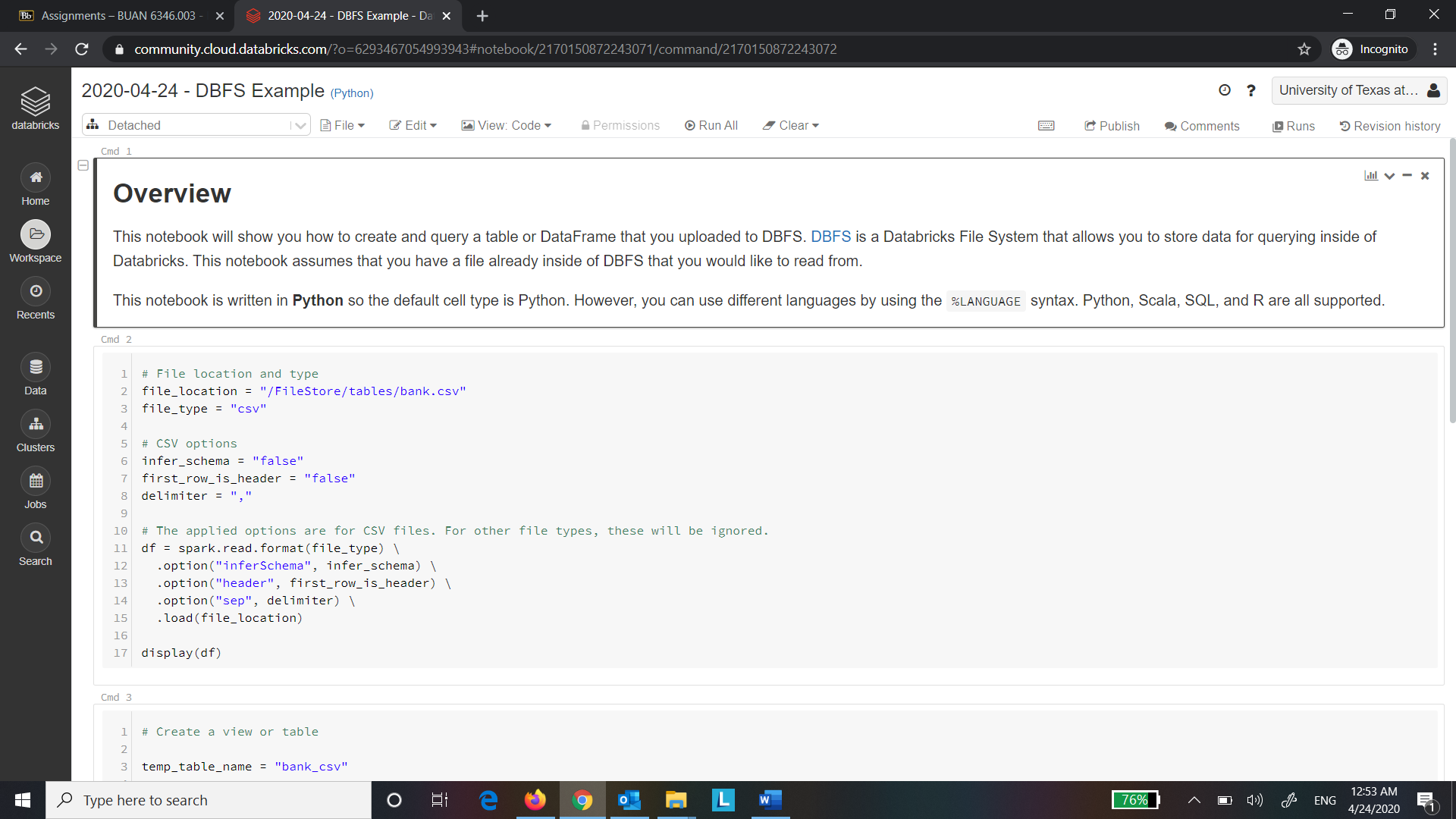
Q. Take a screenshot of the table created and paste it below.

Copy the file location and paste it here.

# File location and type

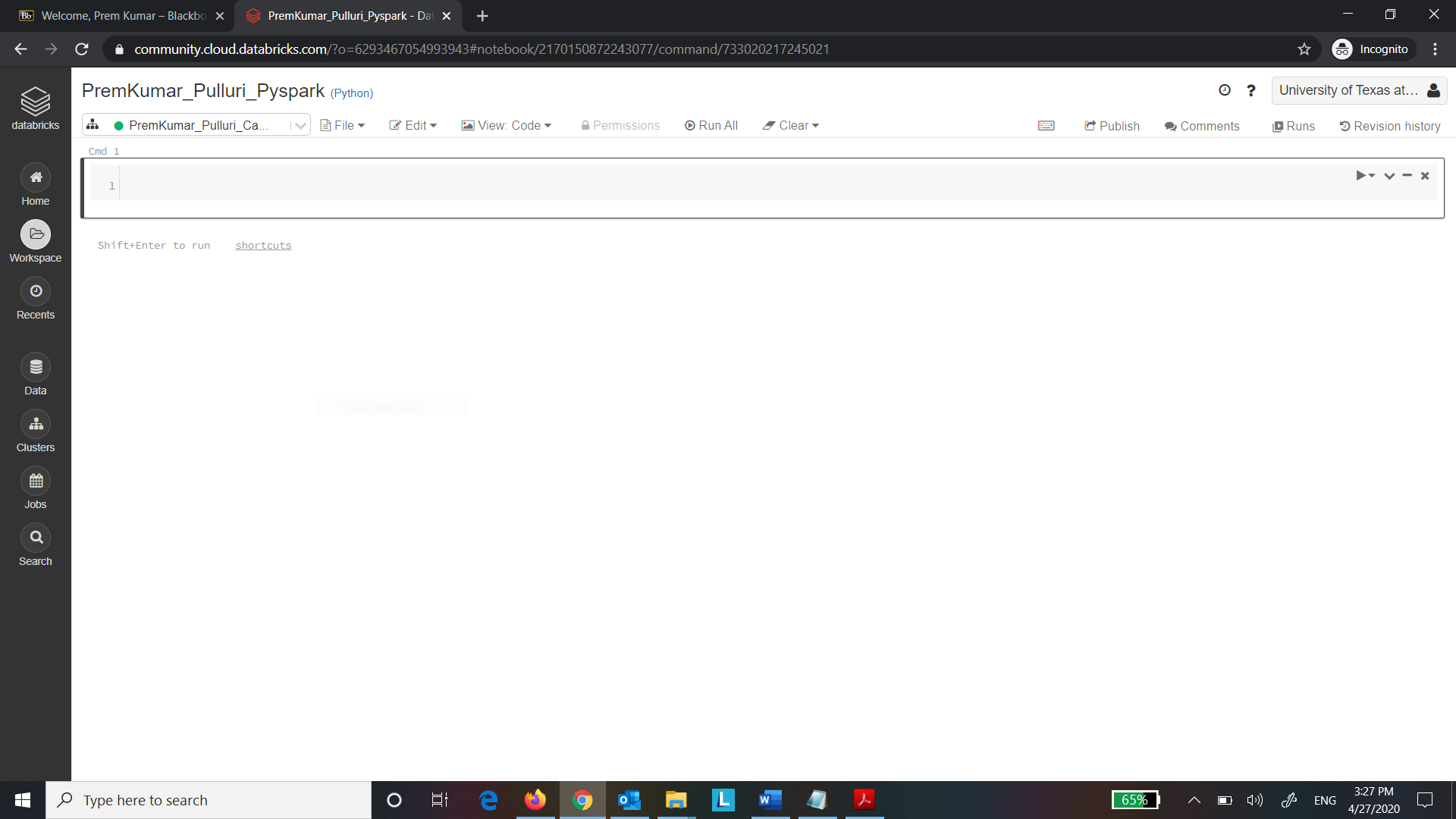
file\_location = "/FileStore/tables/bank.csv"

file\_type = "csv"



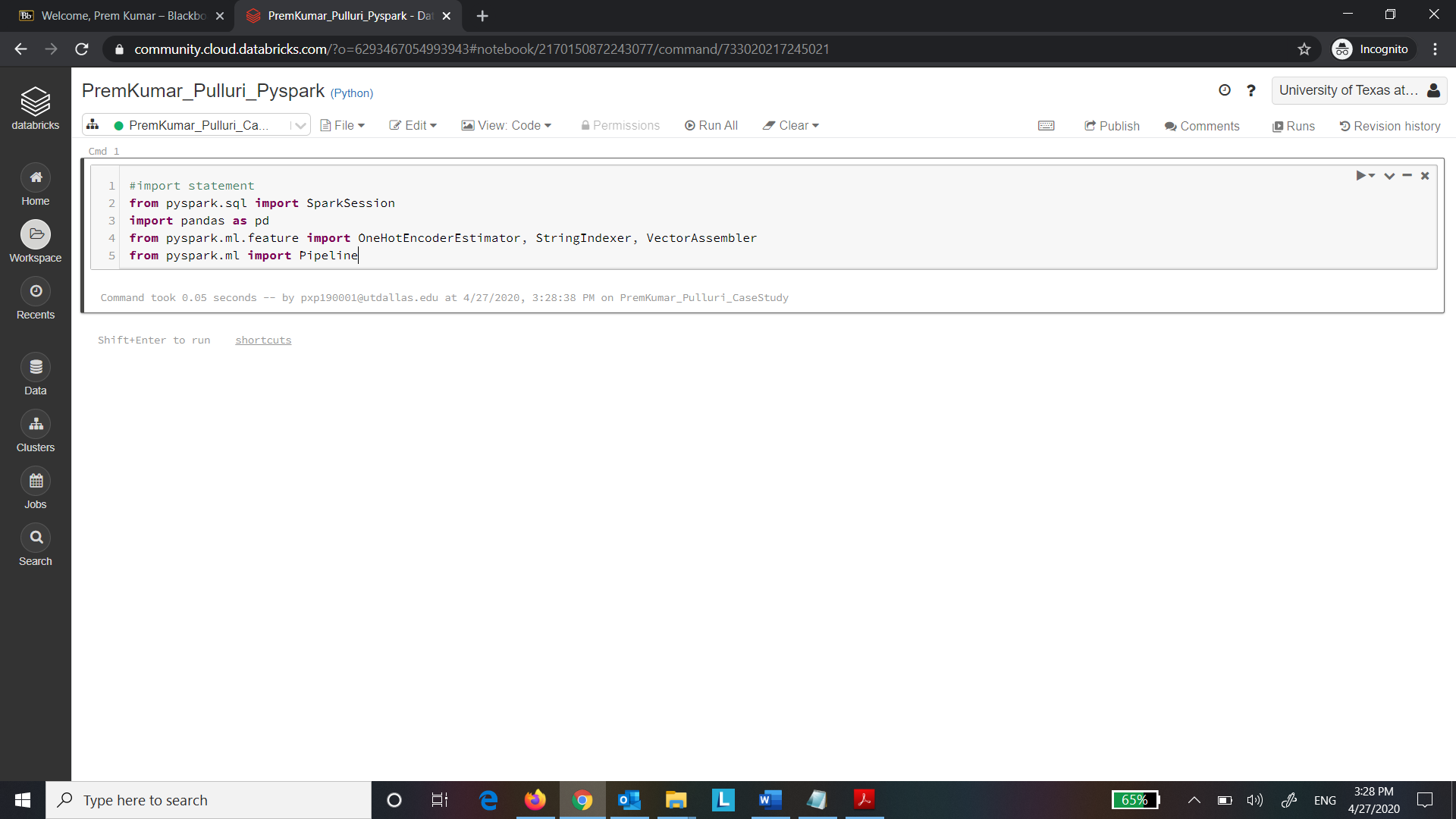
Part 4: DataFrame Operations.

Q. Take a screenshot of the blank notebook and paste it below.



Step 11: Import the required libraries from Spark.

Q. Take a screenshot of the output of the code and paste it below.

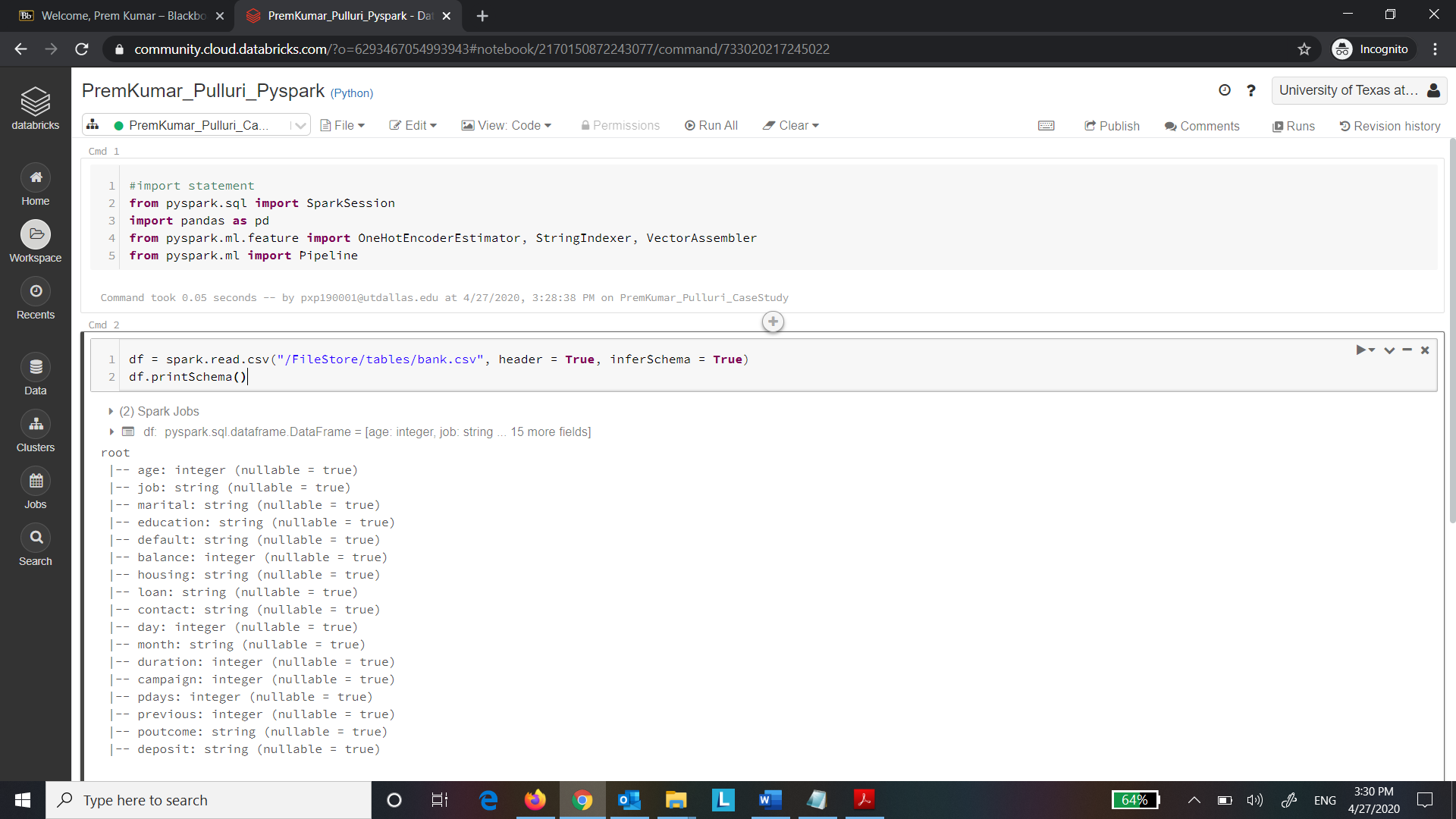


Step 12:

Q. Take a screenshot of the output of the code and paste it below.

Here our output variable id is “deposit”. List all the input variables.

age: integer (nullable = true) job: string (nullable = true) marital: string (nullable = true) education: string (nullable = true) default: string (nullable = true) balance: integer (nullable = true) housing: string (nullable = true) loan: string (nullable = true) contact: string (nullable = true) day: integer (nullable = true) month: string (nullable = true) duration: integer (nullable = true) campaign: integer (nullable = true) pdays: integer (nullable = true) previous: integer (nullable = true) poutcome: string (nullable = true)



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

Q. Take a screenshot of the output of the code and paste it below.

i. Describe “Select”, “Group by” operations in one line.

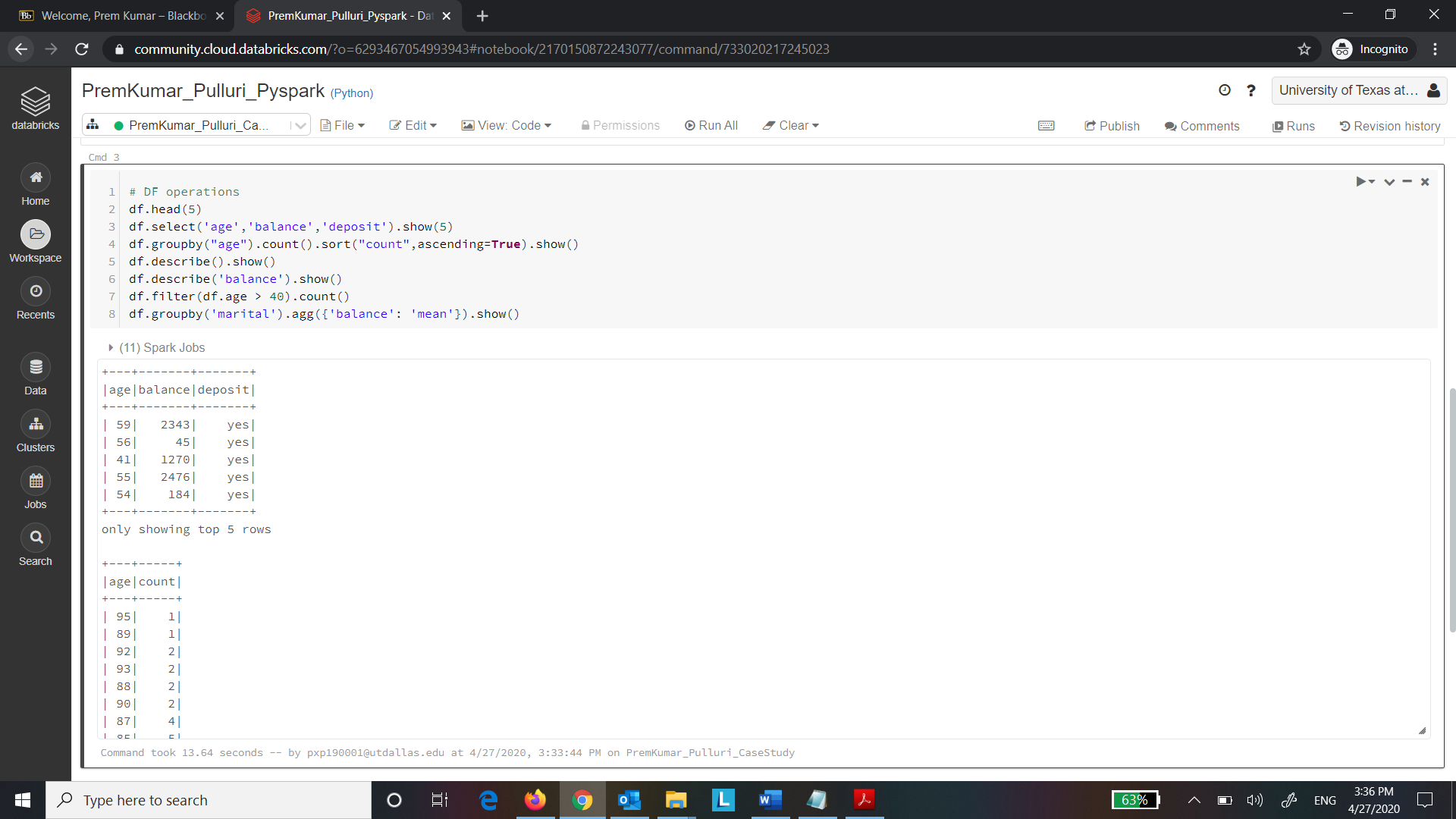
“Select” will select the mentioned columns and “Group by” will aggregate all the rows by the mentioned columns

ii. How many customers are there with an age greater than 40?

There are 4967 customers whose age is greater than 40

iii. Why are there only 5 records viewable?

Only 5 records are viewable because we have passed 5 as parameter in show function

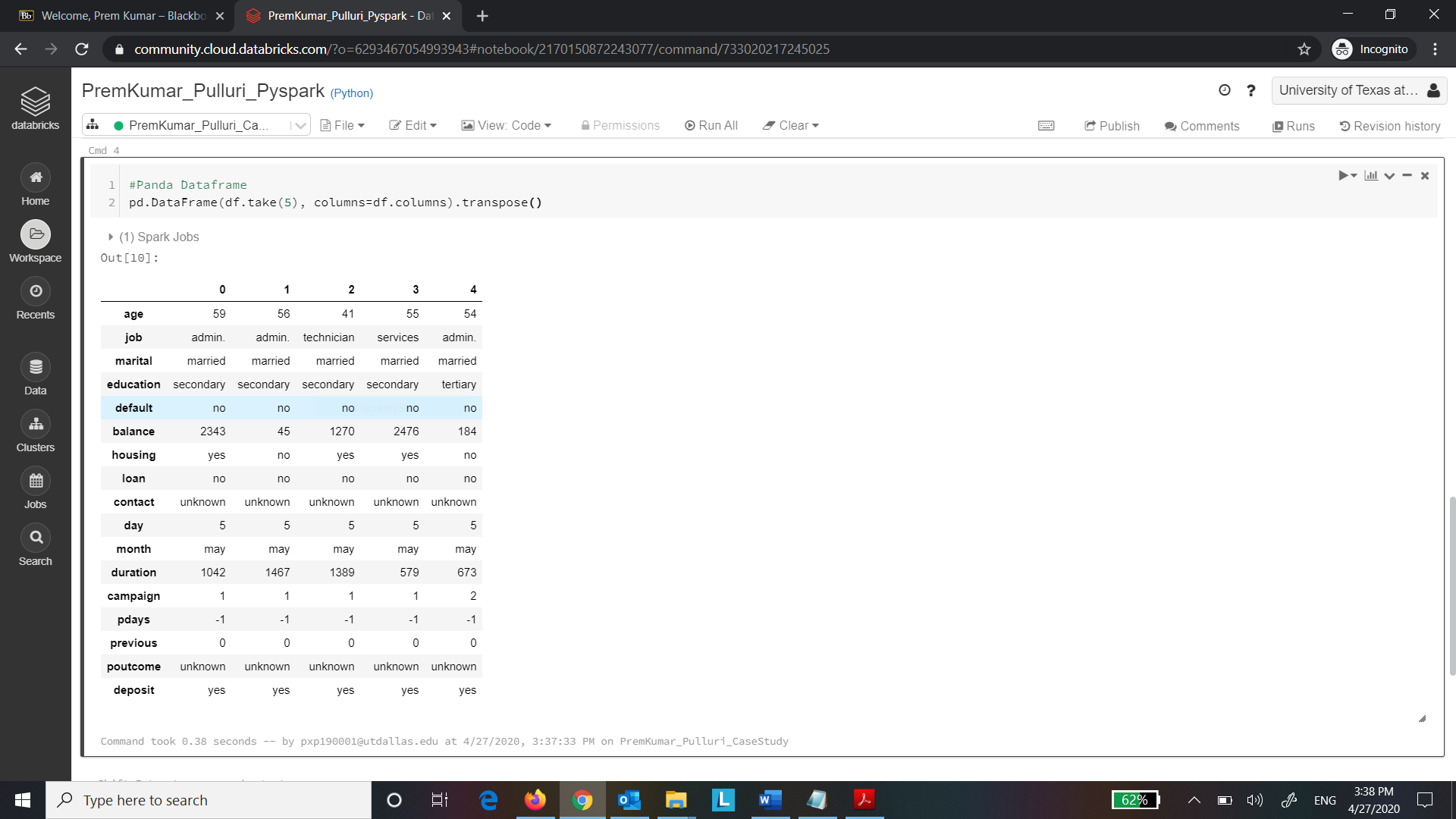


Step 14: Execute the following code:

Q. Take a screenshot of the output of the code and paste it below.

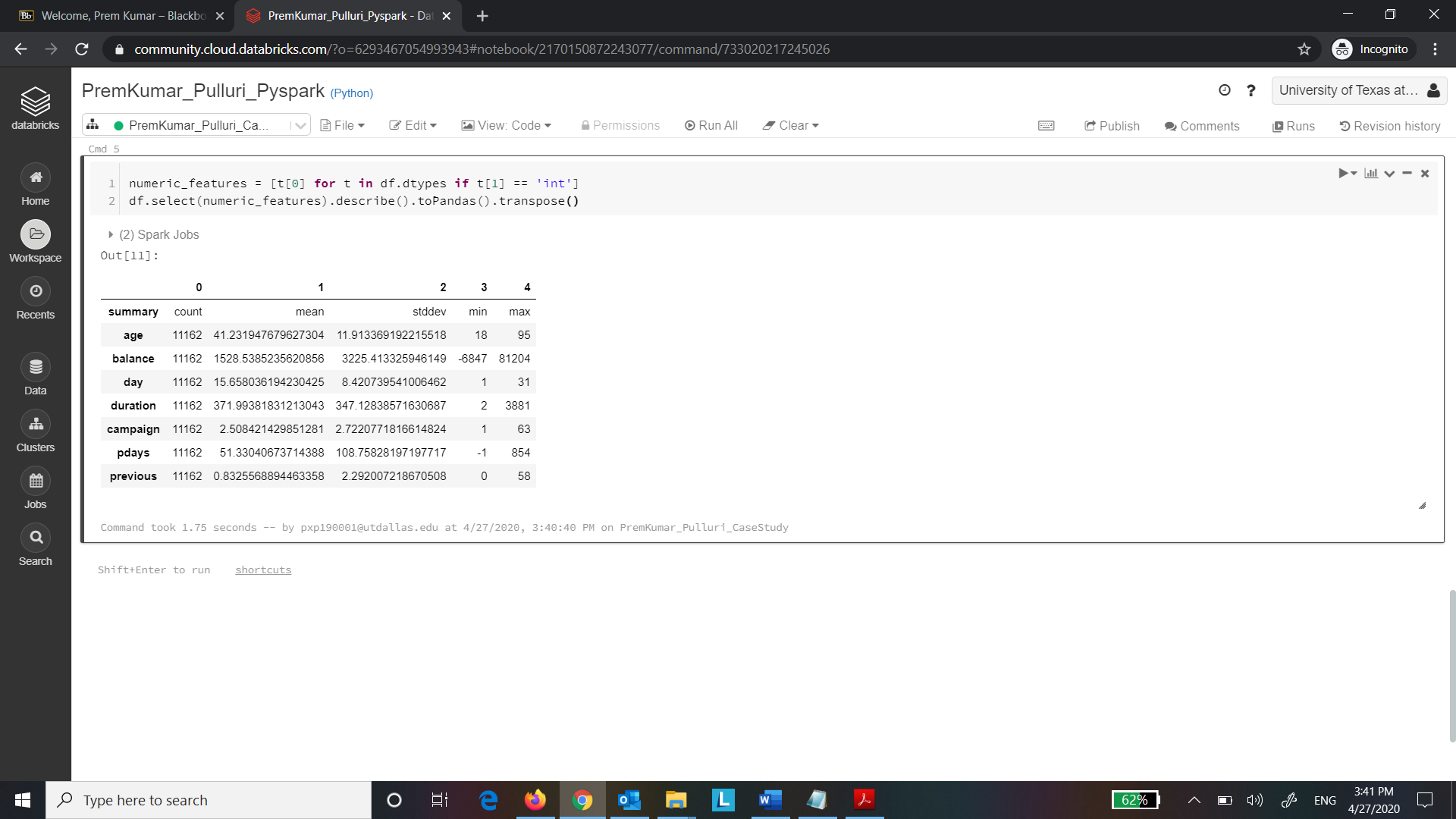
What is seen in the output of the transpose?

The rows and columns are interchanged.



Step 15: Execute the following code:

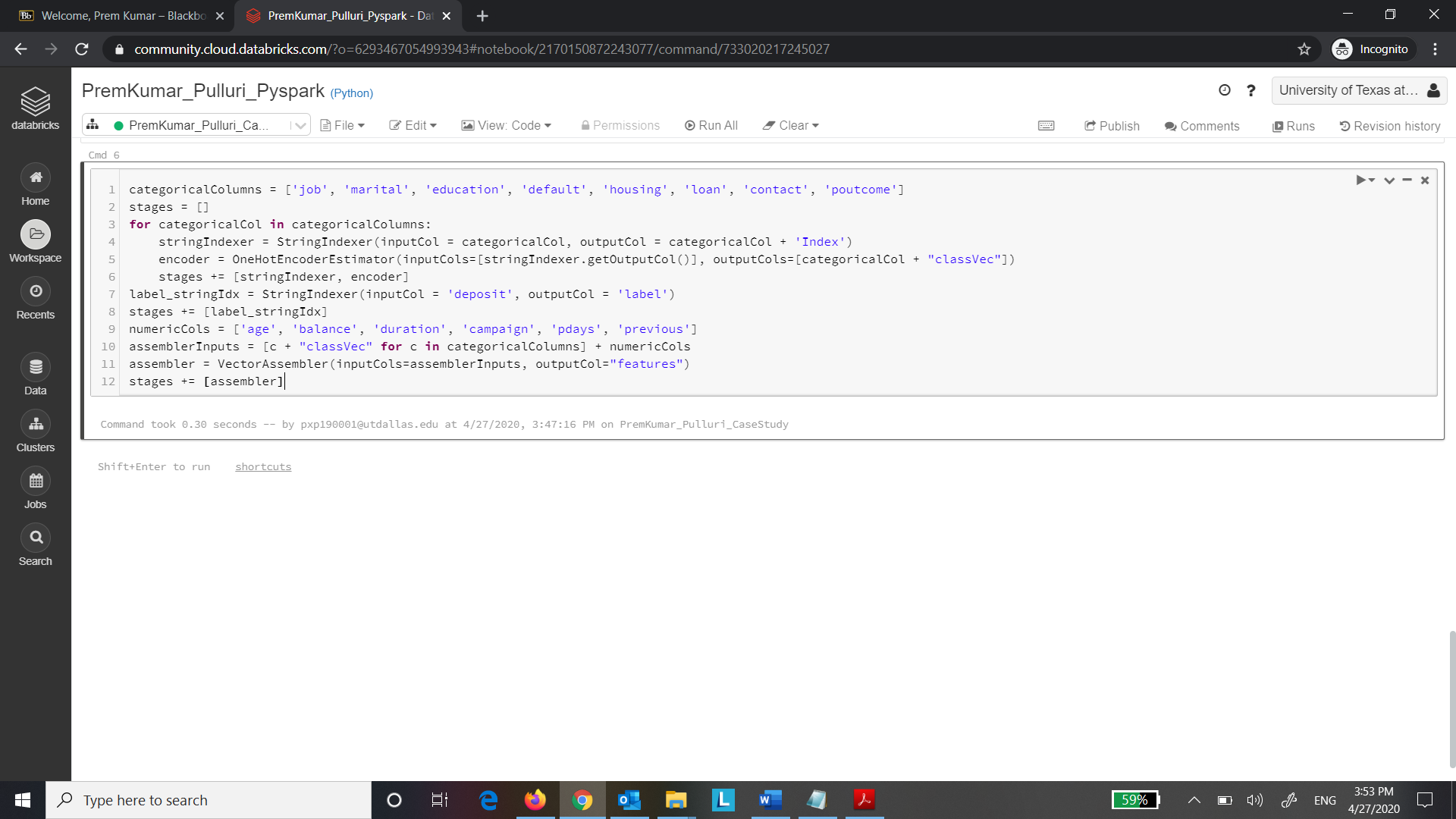
Q. Take a screenshot of the output of the code and paste it below.



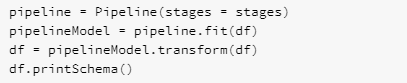
Part 5: Data Preprocessing:

Step 16: Execute the following code:

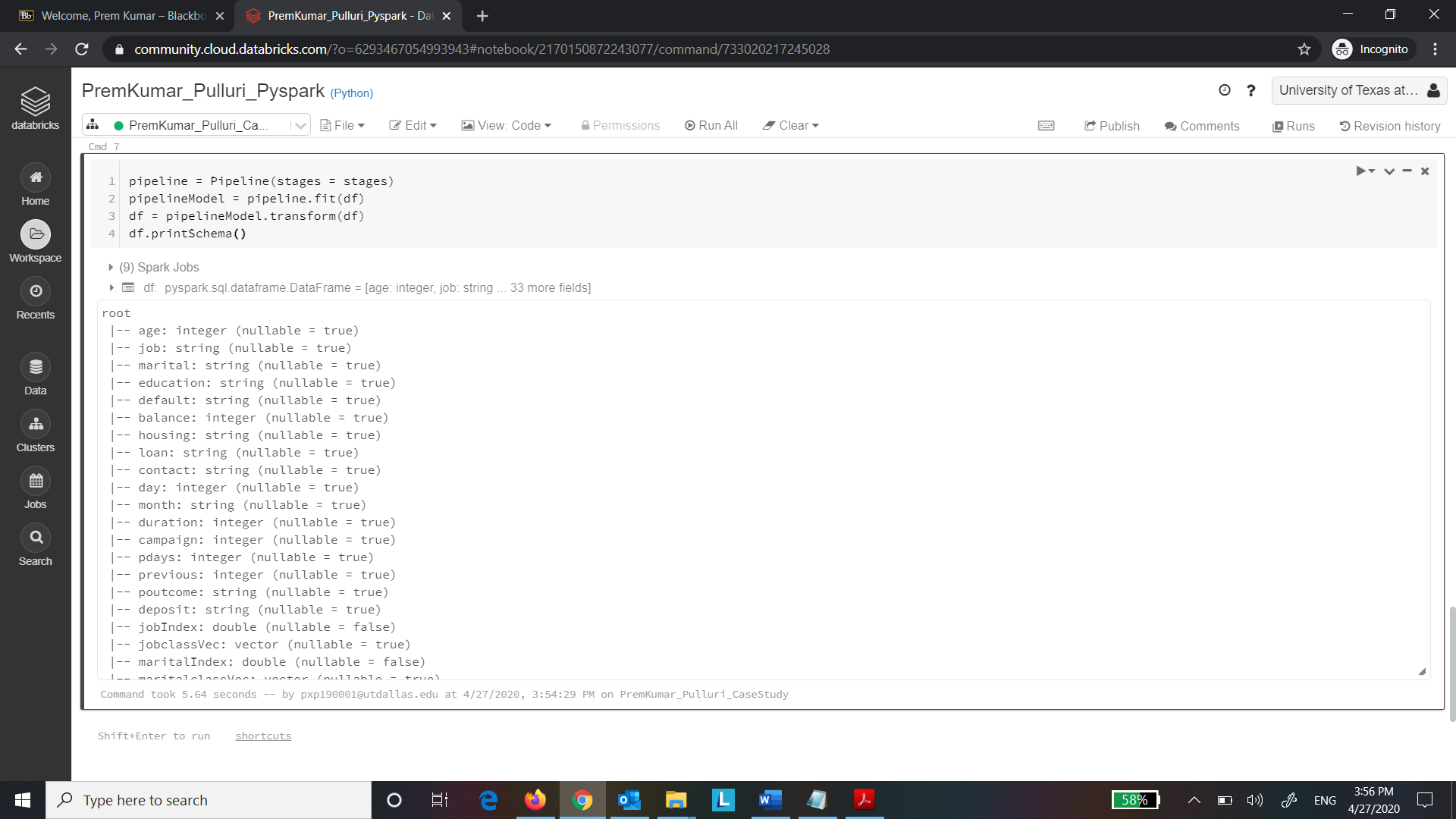
Q. Take a screenshot of the output of the code and paste it below.



Step 17: Execute the following code:



Q. Take a screenshot of the output of the code and paste it below.



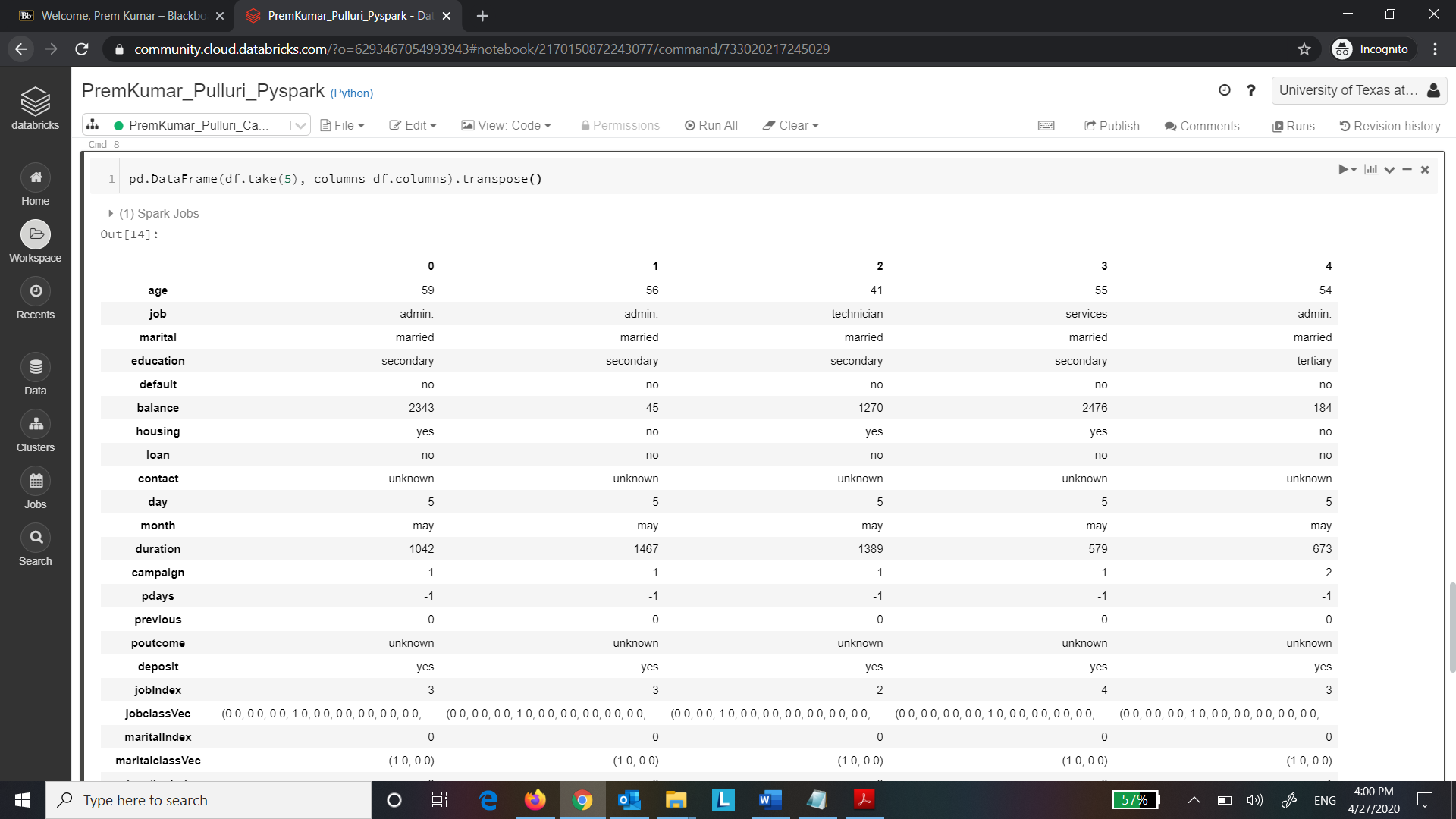
Step 18: Execute the following command:



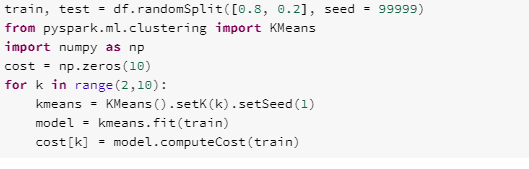
Q. Take a screenshot of the output of the code and paste it below.

Define the purpose of the transpose function?

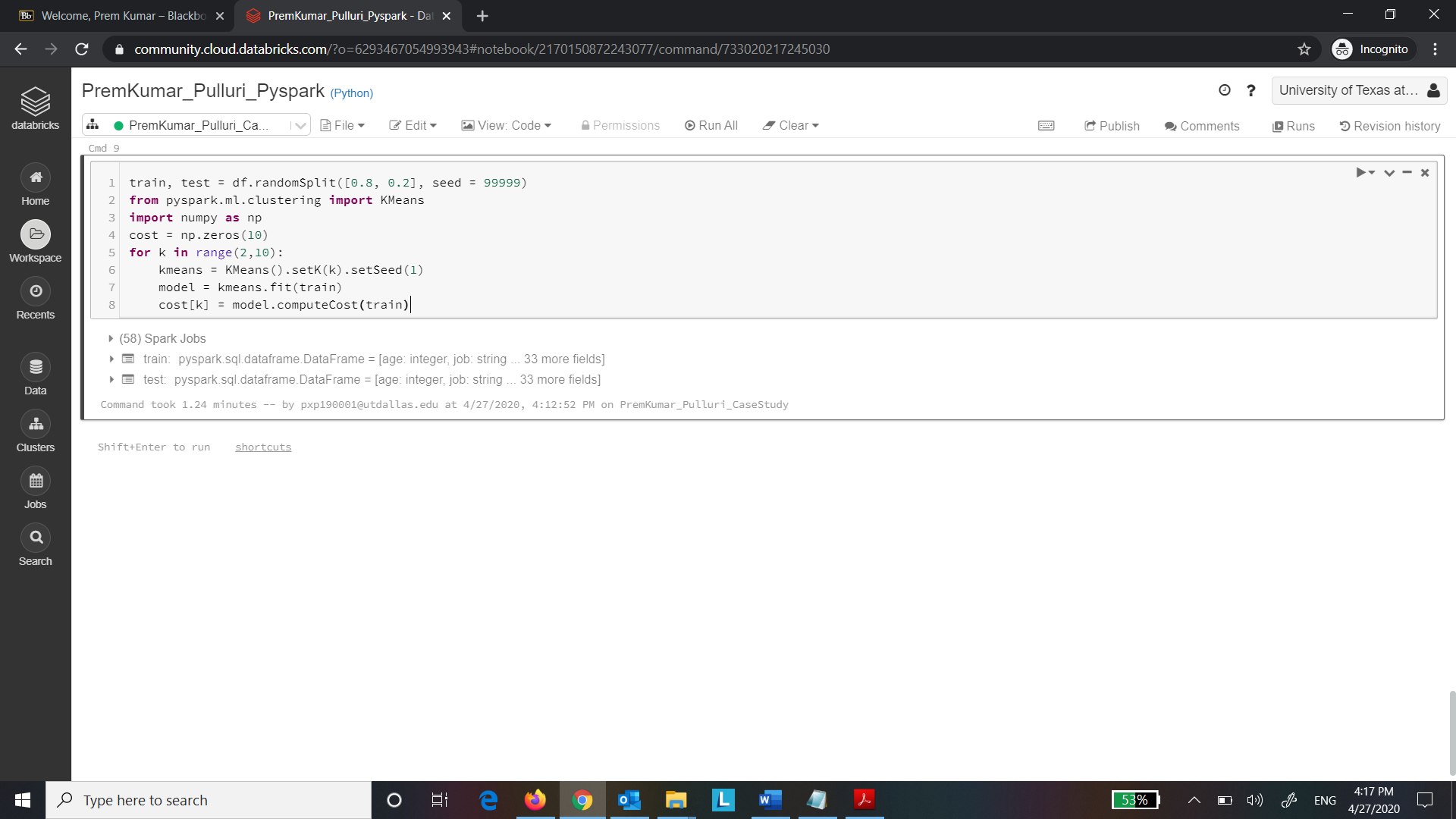
The purpose of the transpose function is to swap the rows and columns



Step 19: Execute the following code:



Q. Take a screenshot of the output of the code and paste it below.

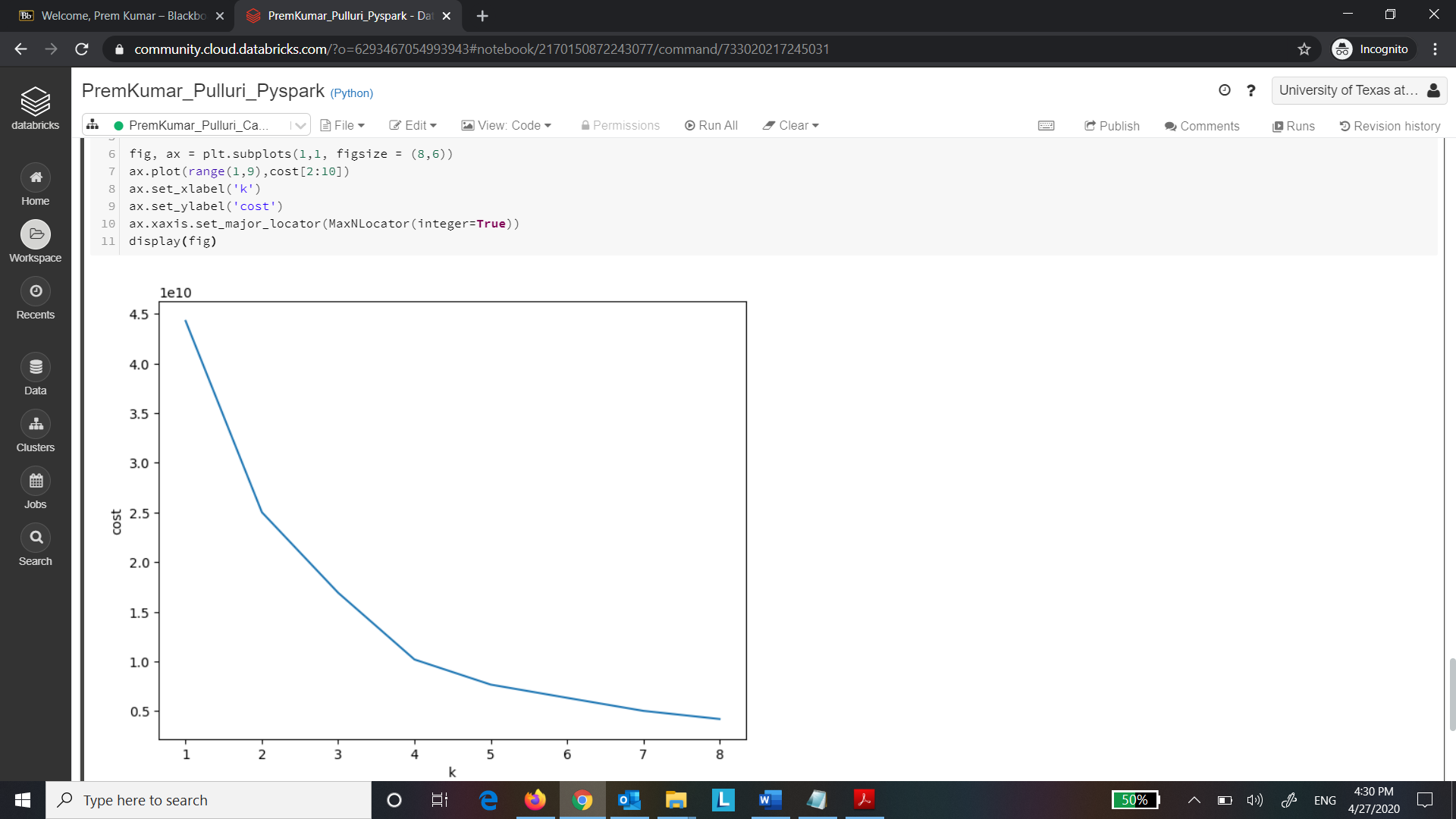


Step 20: Execute the following code:

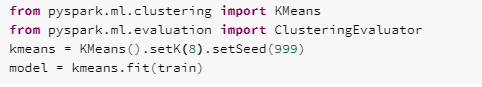
Q. Take a screenshot of the output of the code and paste it below.

Comment about the graph that is plotted here.

From the elbow plot, we can say that the cost decreases as number of clusters increases



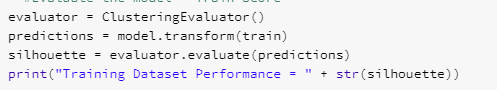
Step 21: Execute the following code:



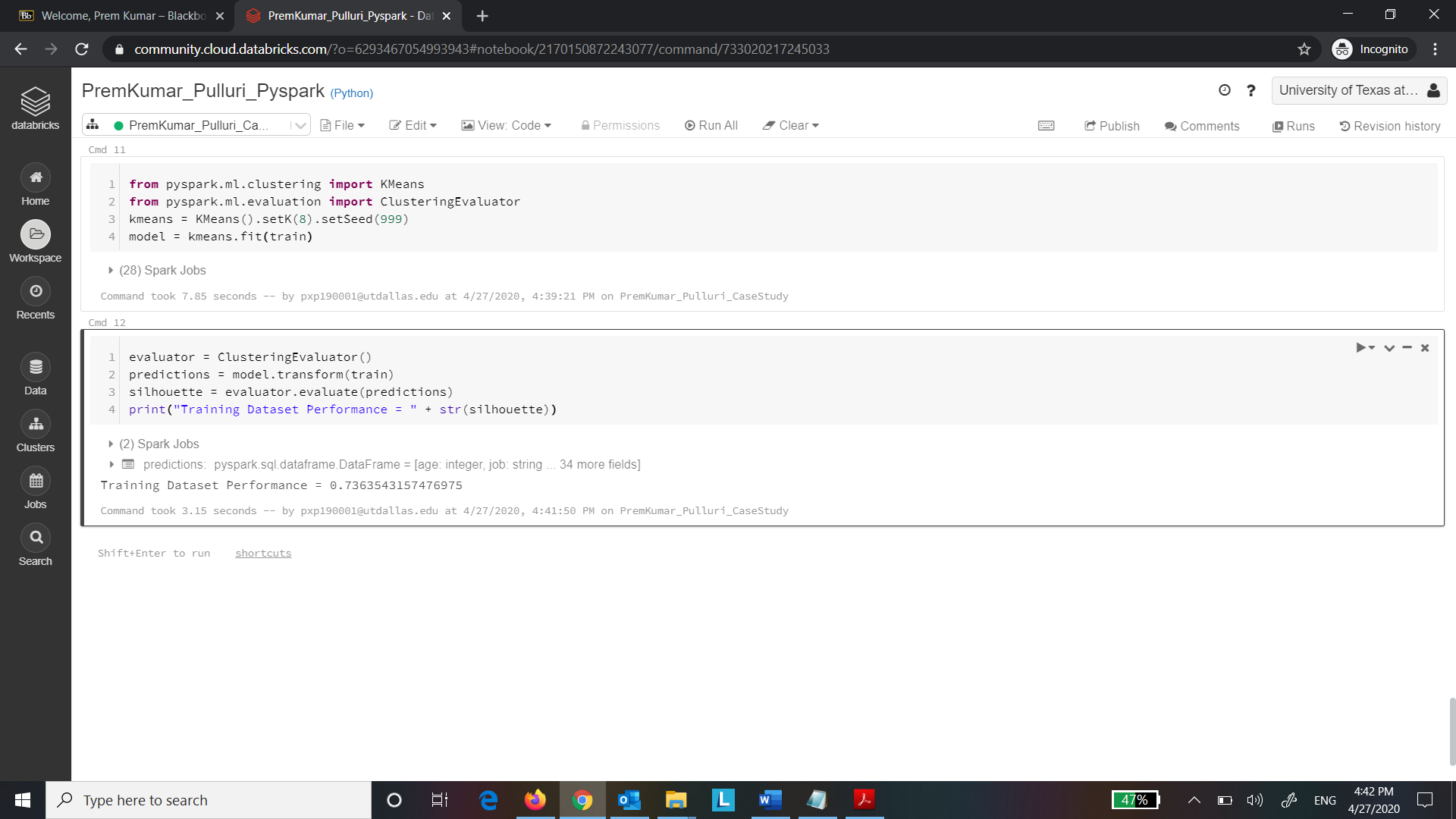
Q. Take a screenshot of the output of the code and paste it below.



Step 22: Execute the following code:



Q. Take a screenshot of the output of the code and paste it below.



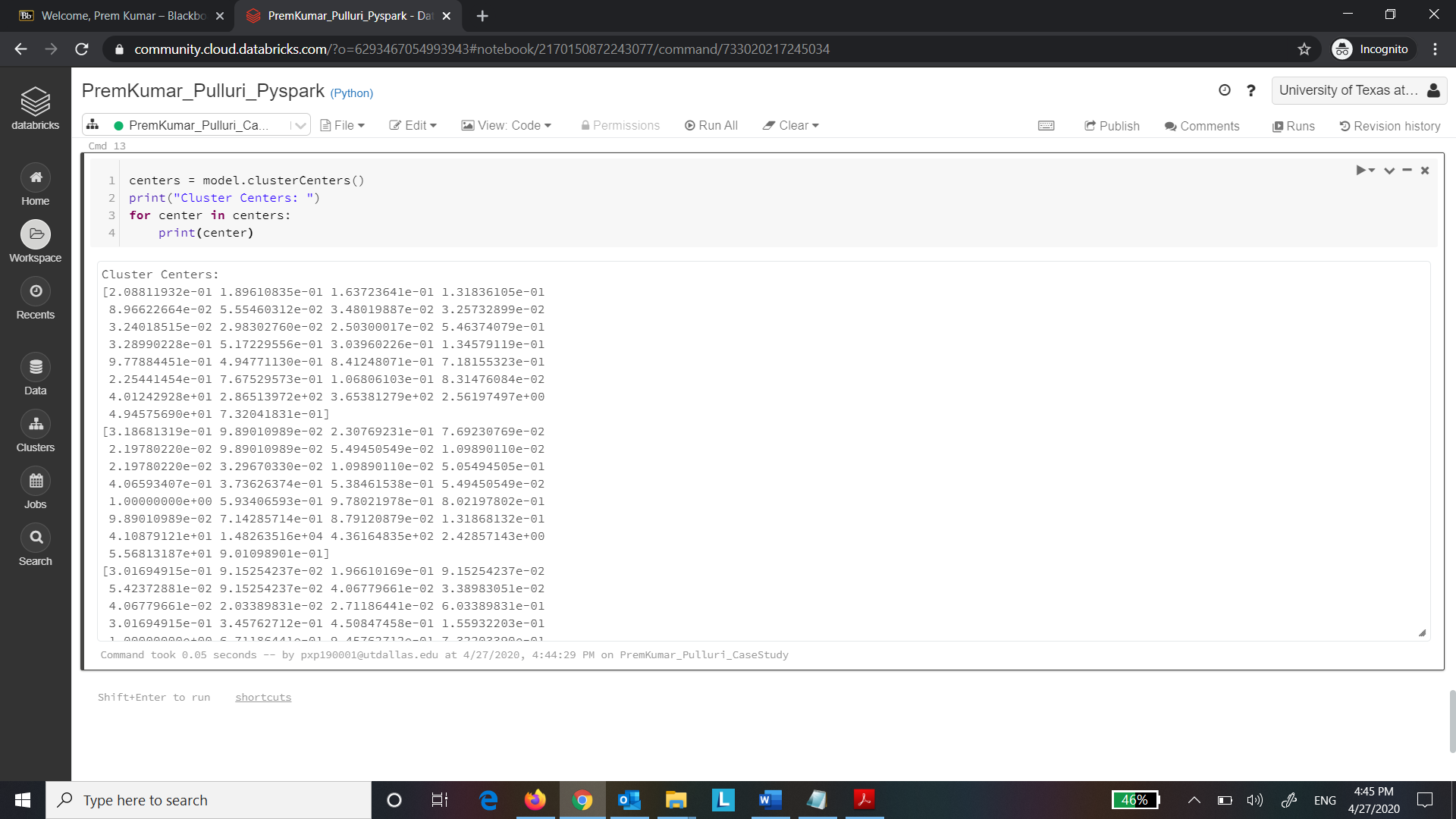
Step 23: Execute the following code:



Q. Take a screenshot of the output of the code and paste it below.

What is Centroid and how is it calculated in K-means?

A centroid is a data point at the center of a cluster. Centroid is calculated by taking the mean of all the data points in a cluster

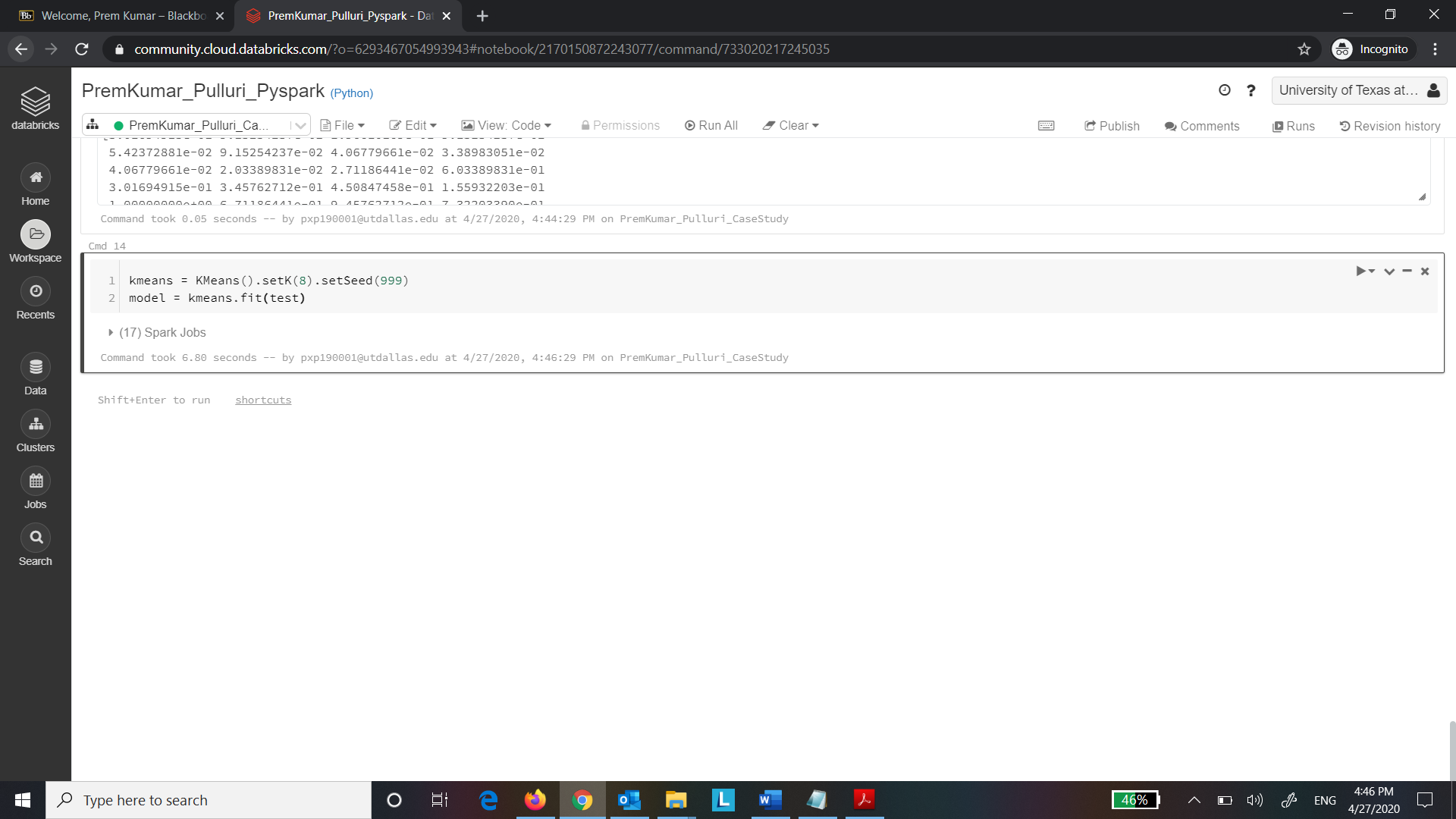


Step 24:

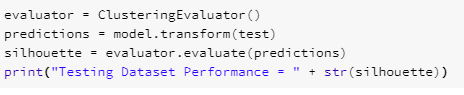
Execute the following code:



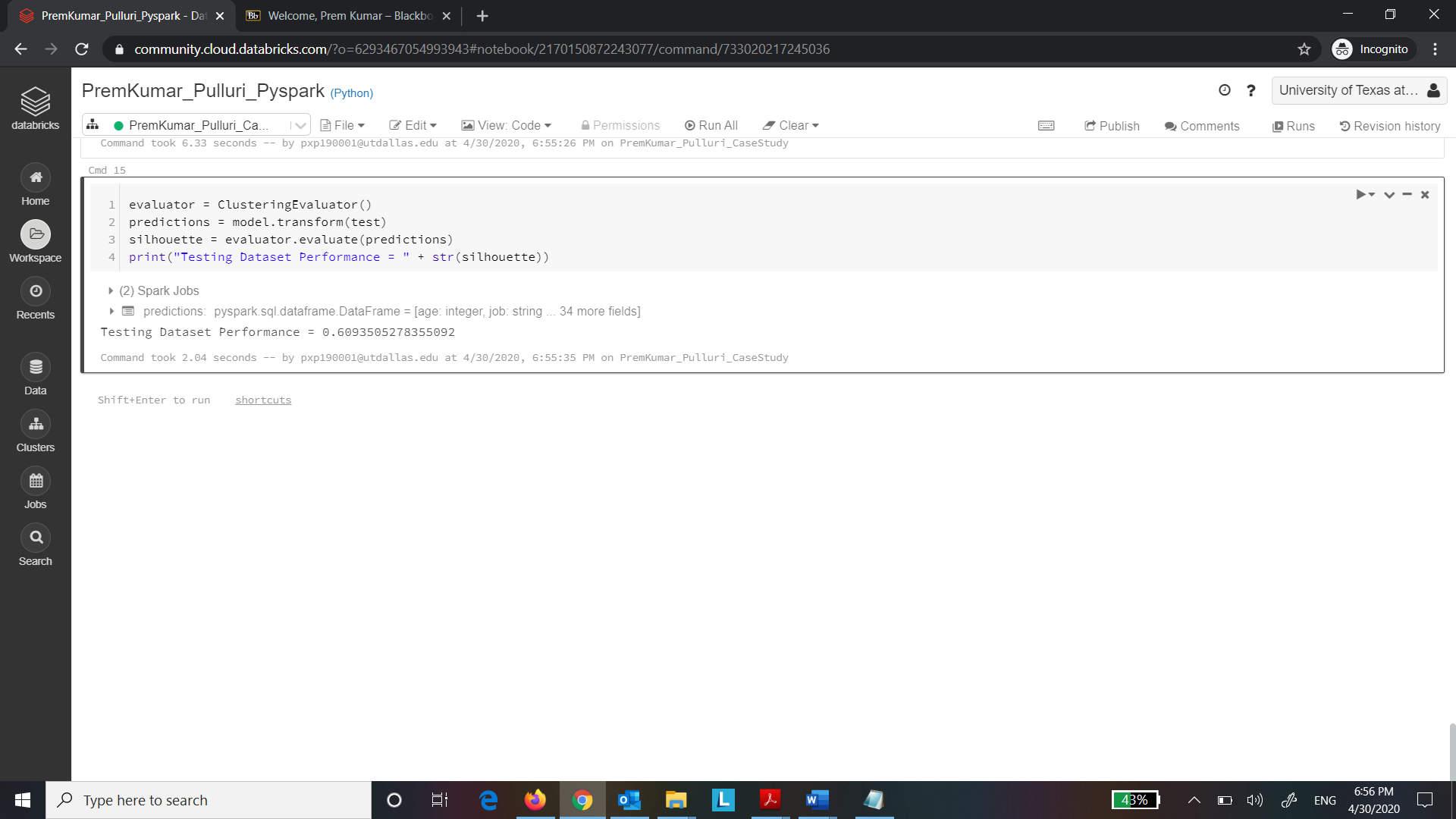
Q. Take a screenshot of the output of the code and paste it below.



Step 25: Execute the following code:



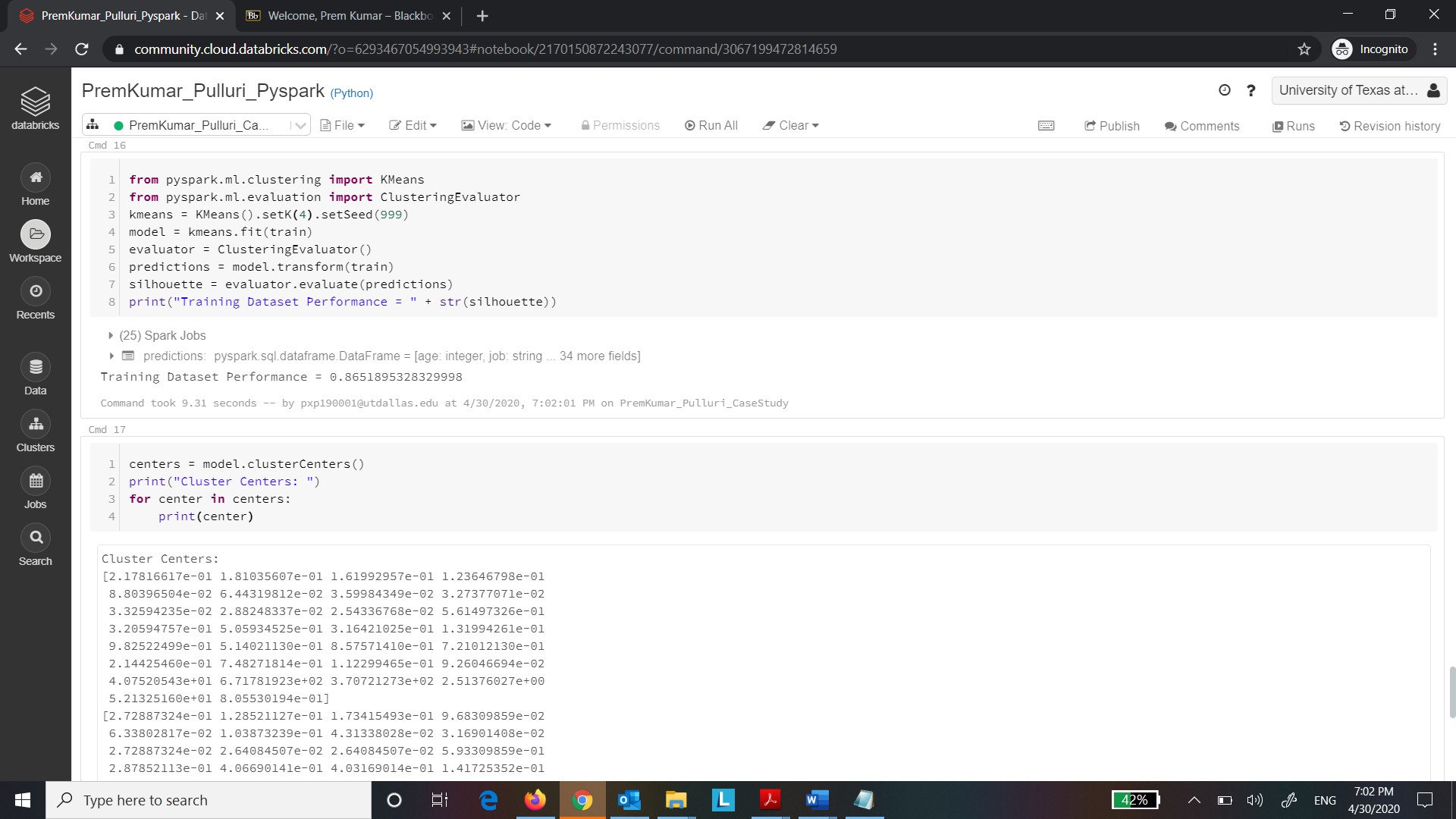
Q. Take a screenshot of the output of the code and paste it below.

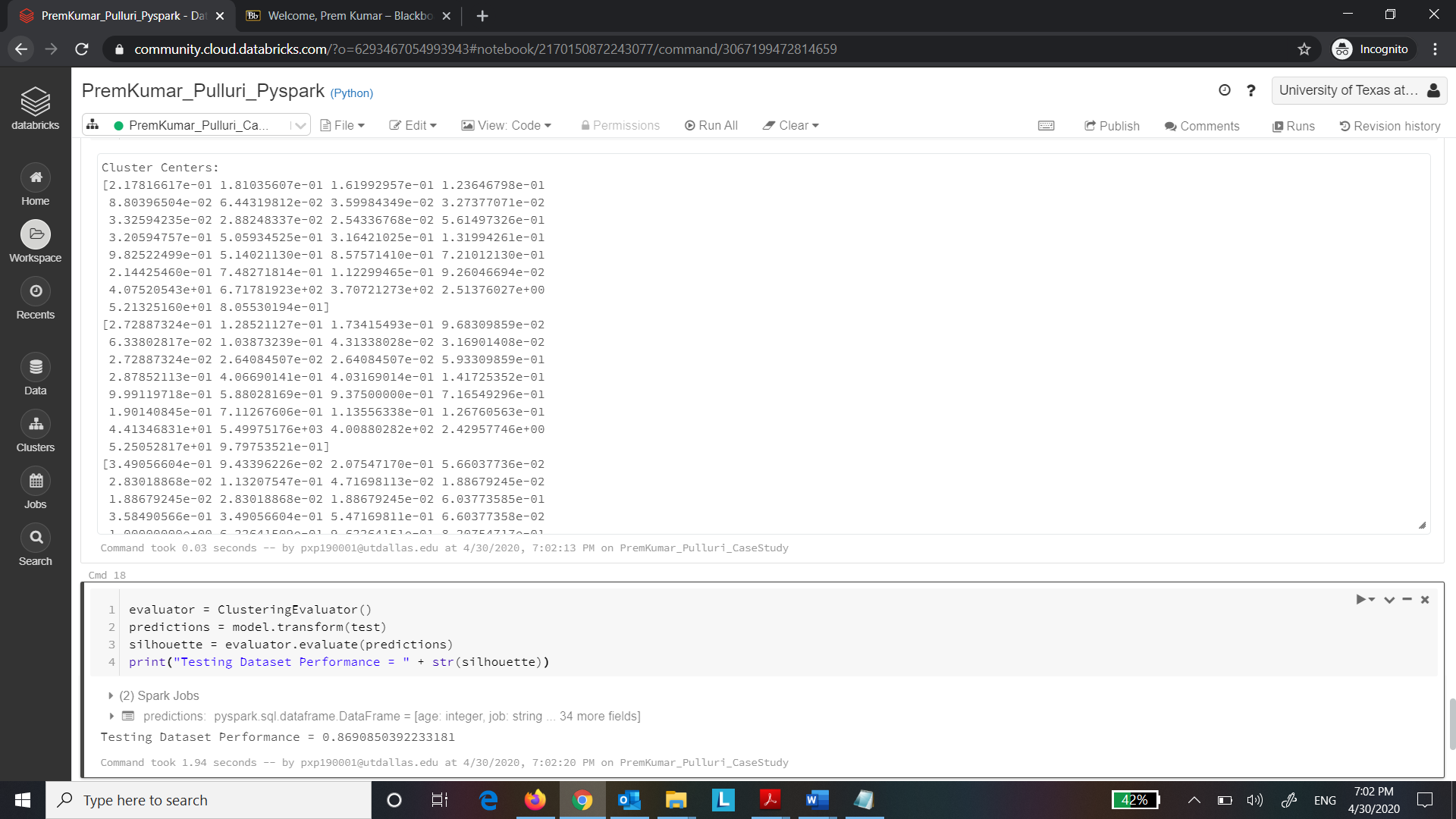


Copy the Train and Test performance results of all the clusters here. Compare the results and comment about the performance of each model. Select your best model and explain why you chose that number of clusters as the optimal level.

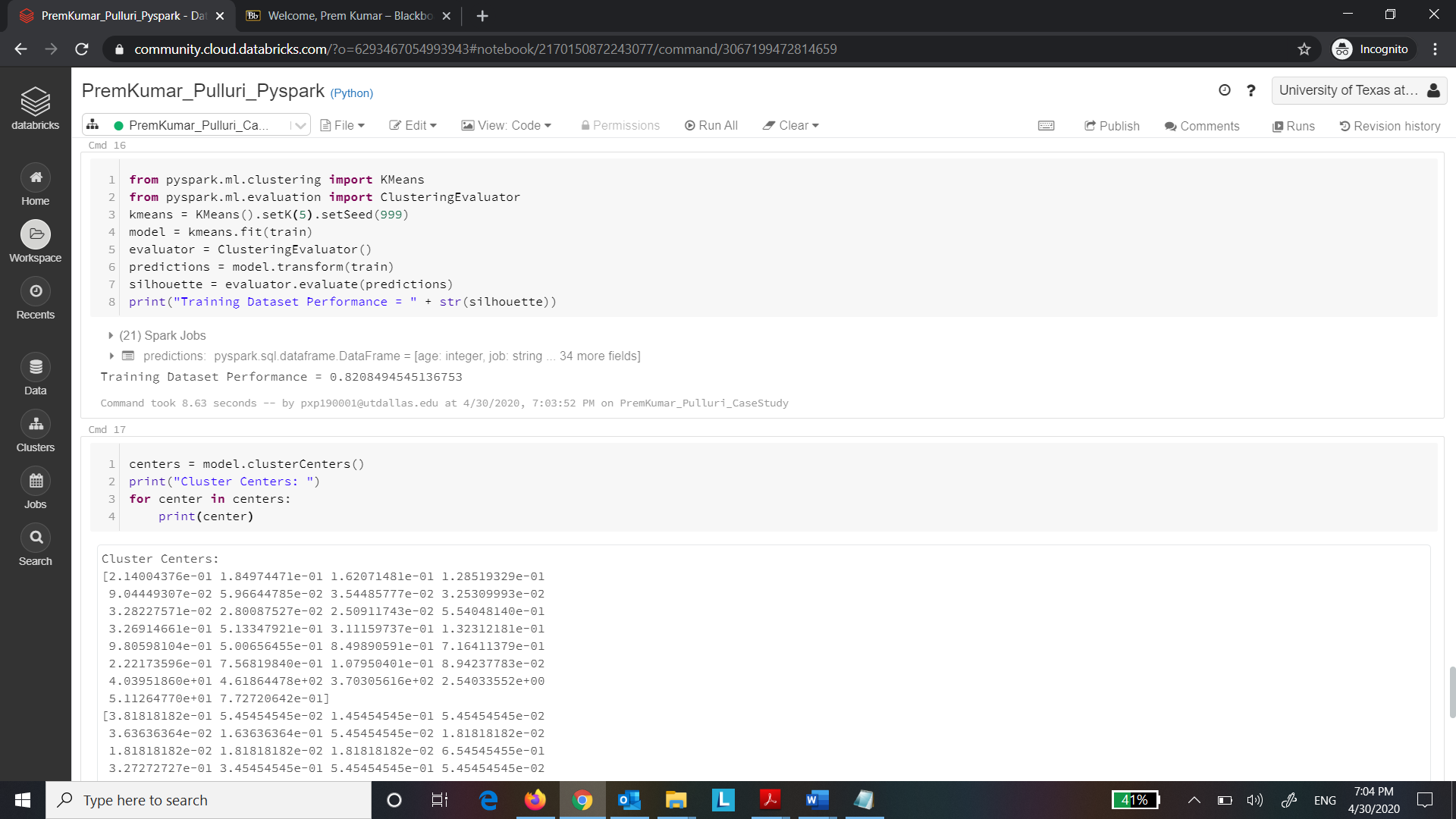
From the elbow plot, we can see that the cost function decreases significantly until K=4 but after that there isn’t significant decrease. If K is less, it leads to overfitting and higher value of K leads to underfitting. So, 4 or 5 might be the optimal value for number of clusters K

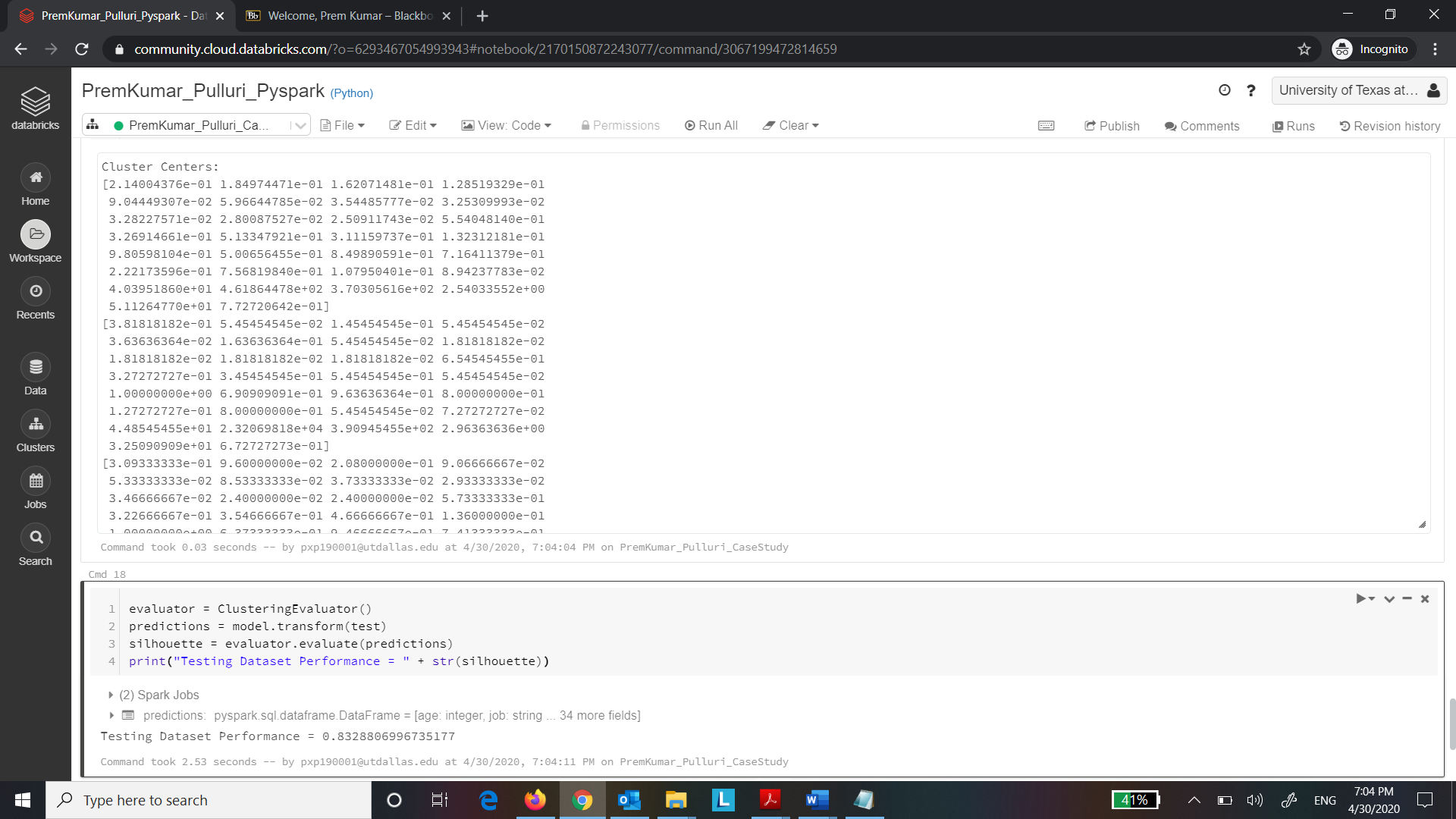
For K=4





For K=5





The train and test dataset performance is better for K = 4 compared to K = 5. Hence K = 4 is the best choice for the value of K