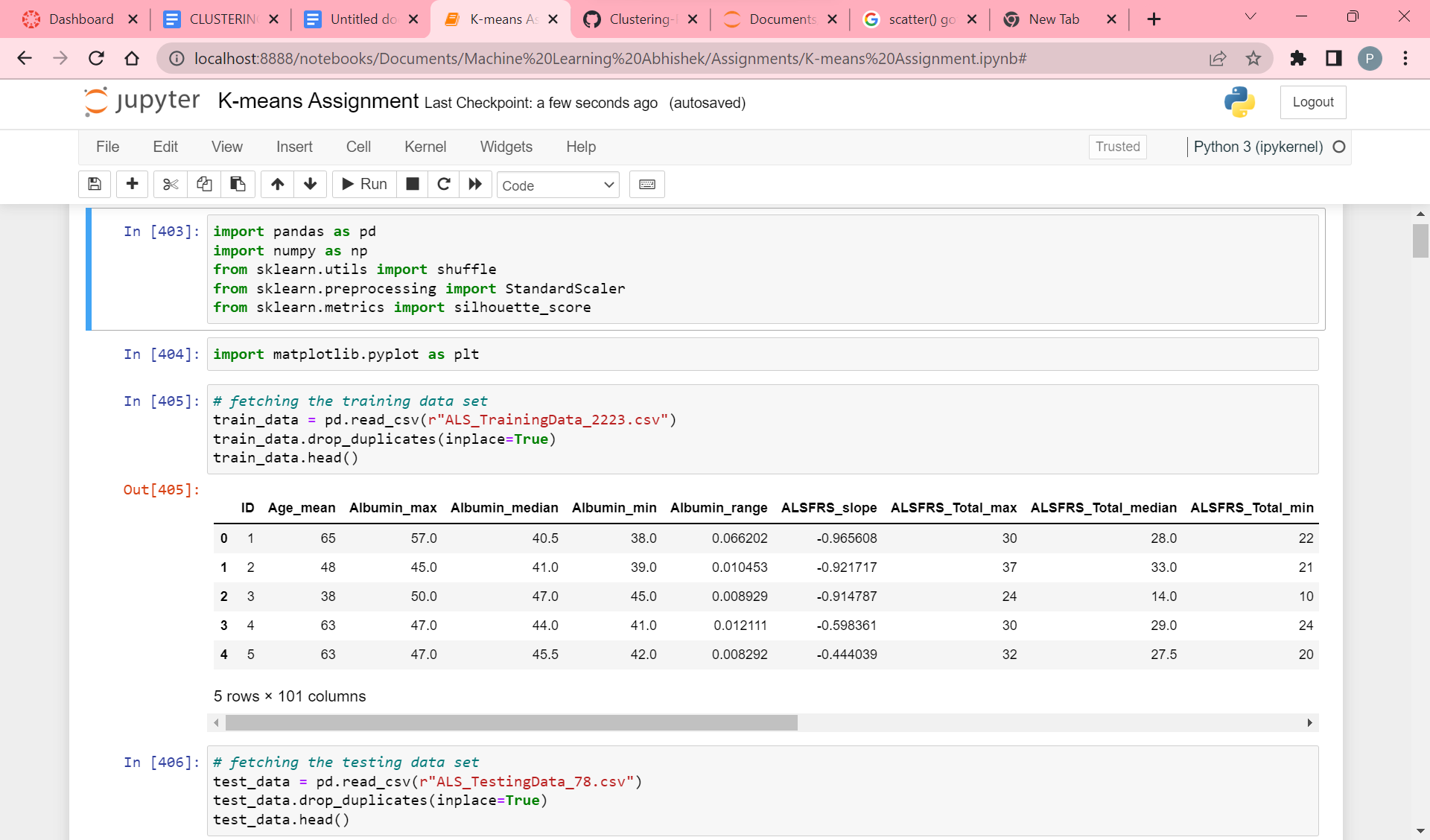
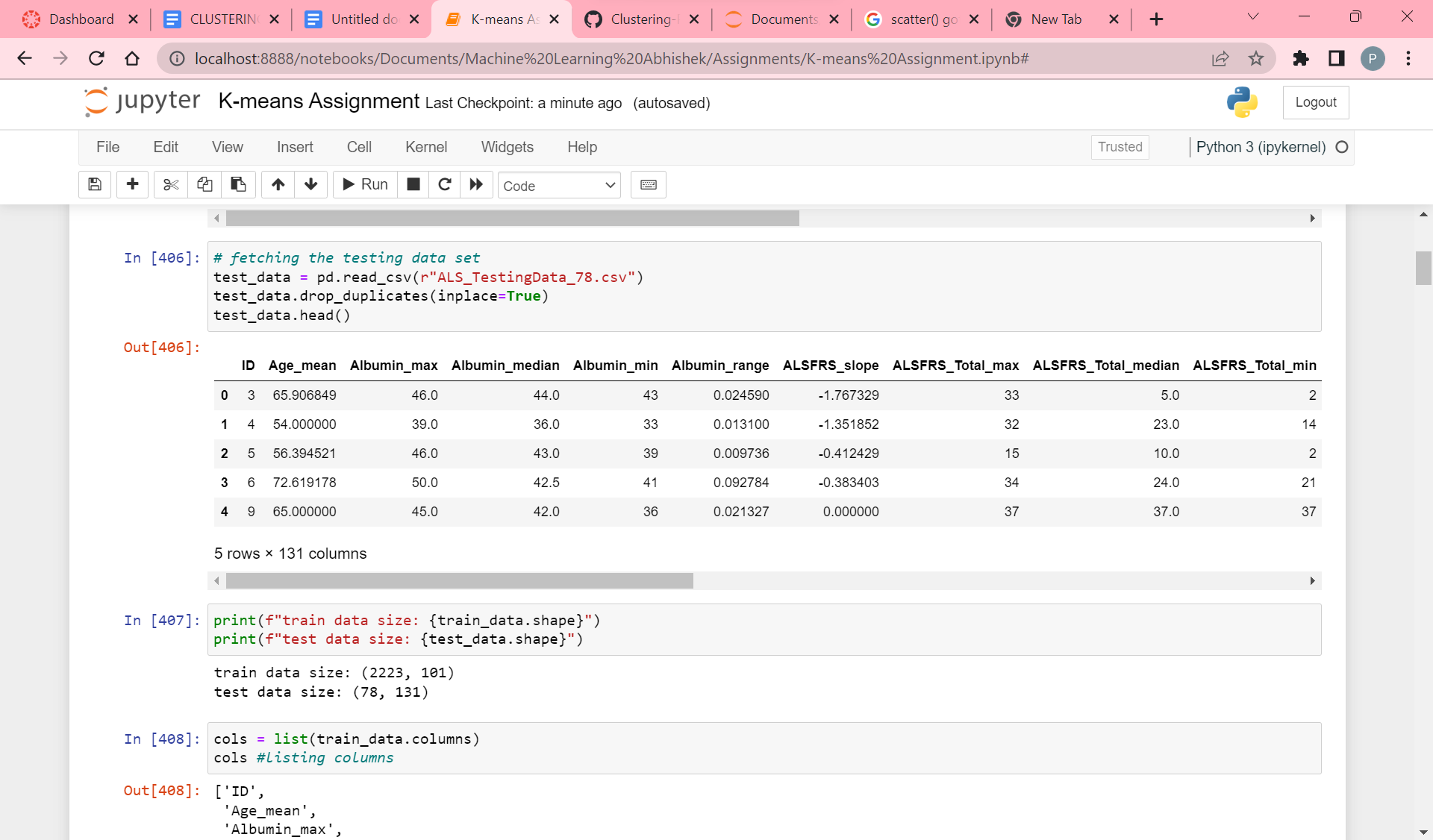
**CLUSTERING**

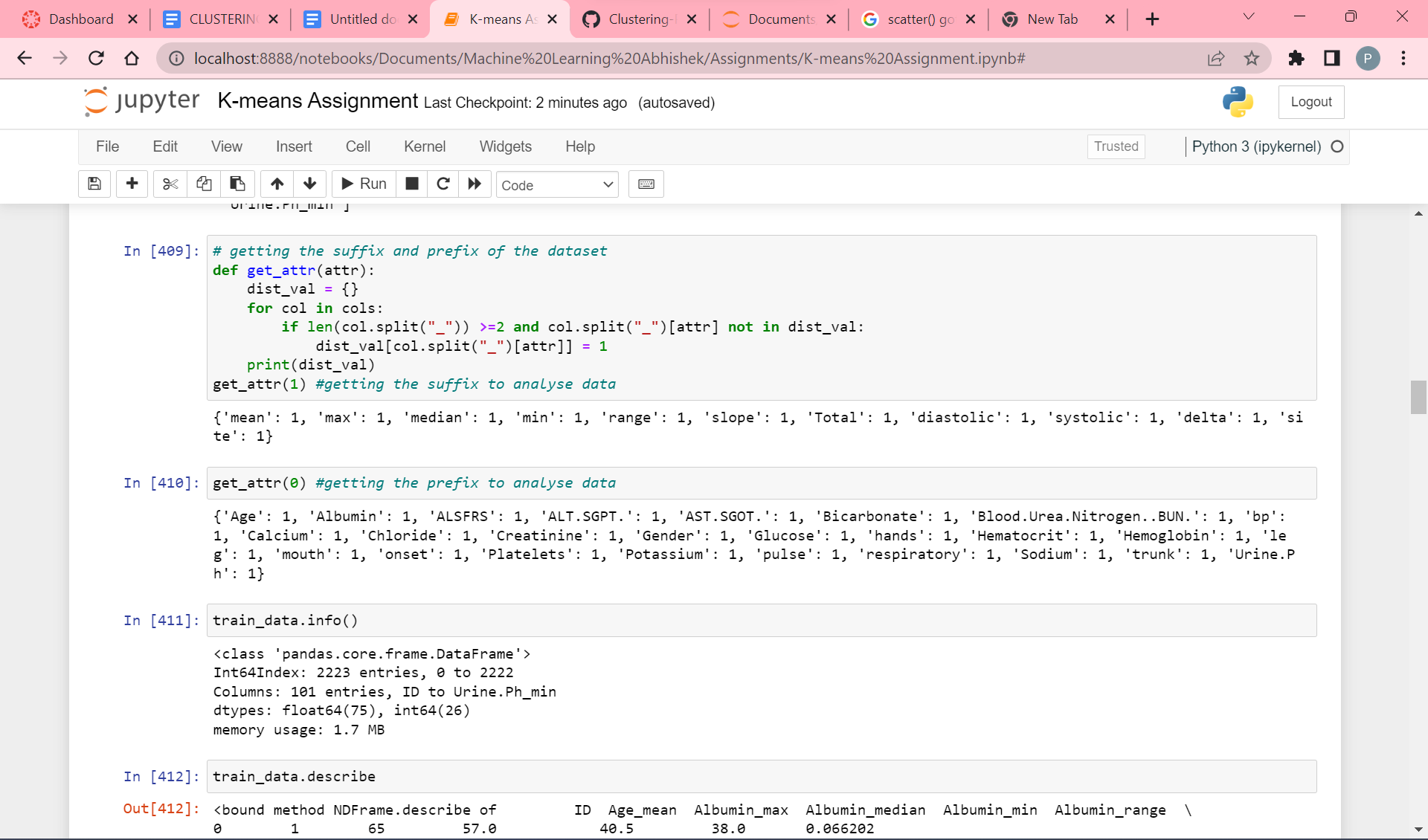
Clustering is an unsupervised type of learning that has the task of dividing the items into groups, such that data points in the same group are more similar to other data points in the same group and dissimilar to the data points in other groups.

Given the datasets with training and testing, as ALS\_TestingData\_78.csv, ALS\_TrainingData\_2223.csv

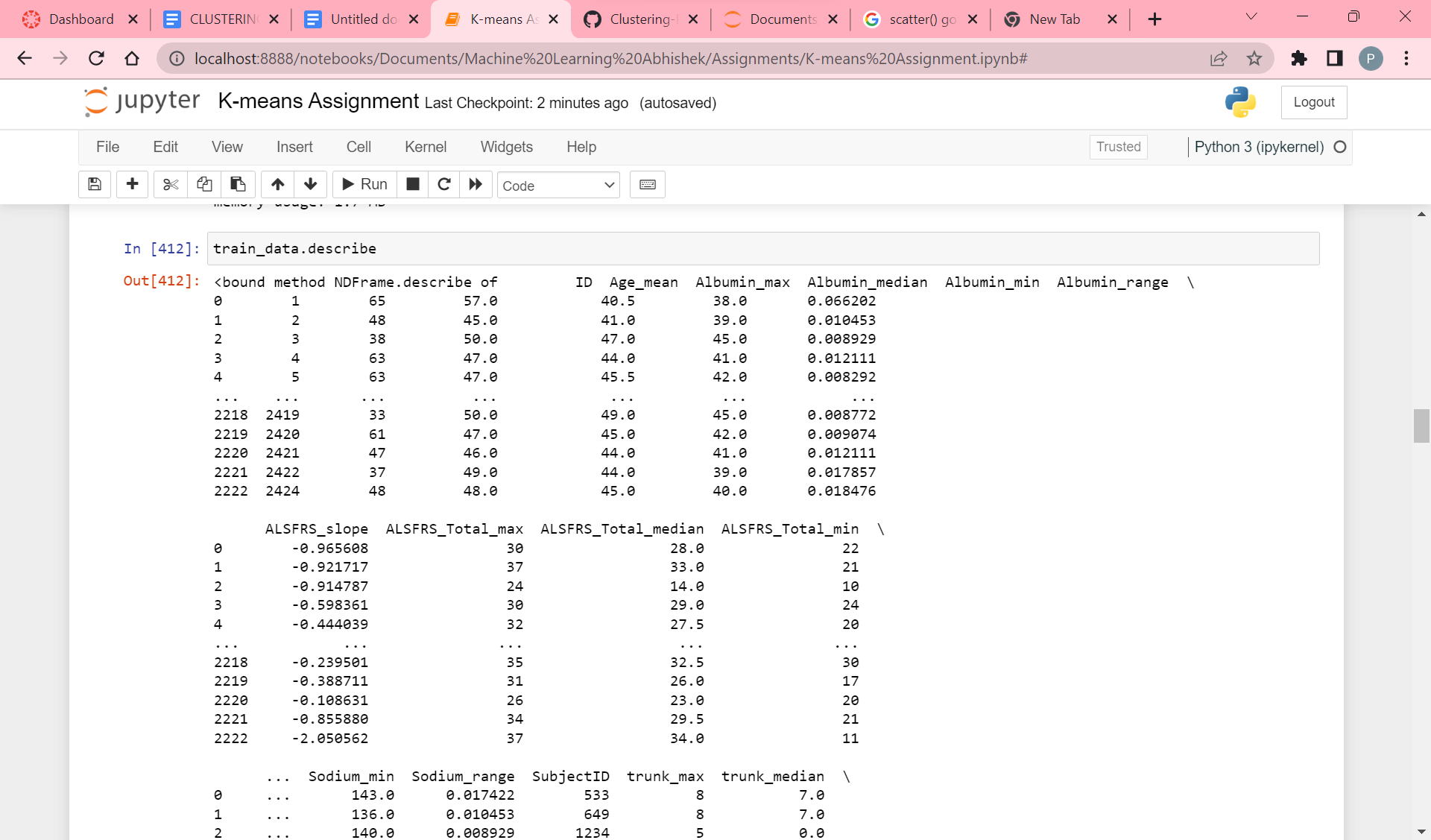




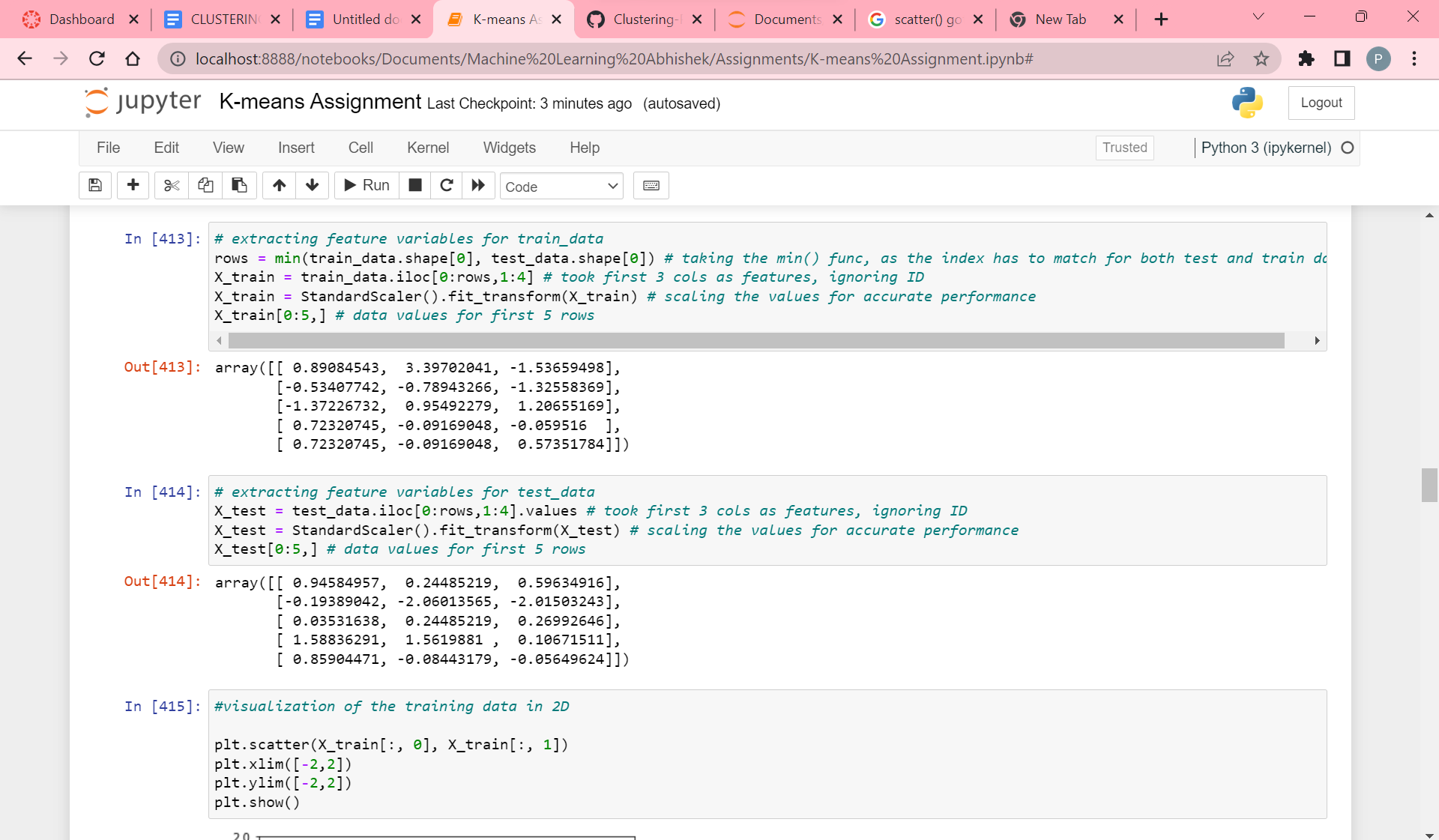
Fetching the training and testing data by dropping off duplicate values. Then printing the dimensions of datasets. Where training data has 2223 instances and 101 features. Testing data has 78 instances and 131 features. And then listing the columns.



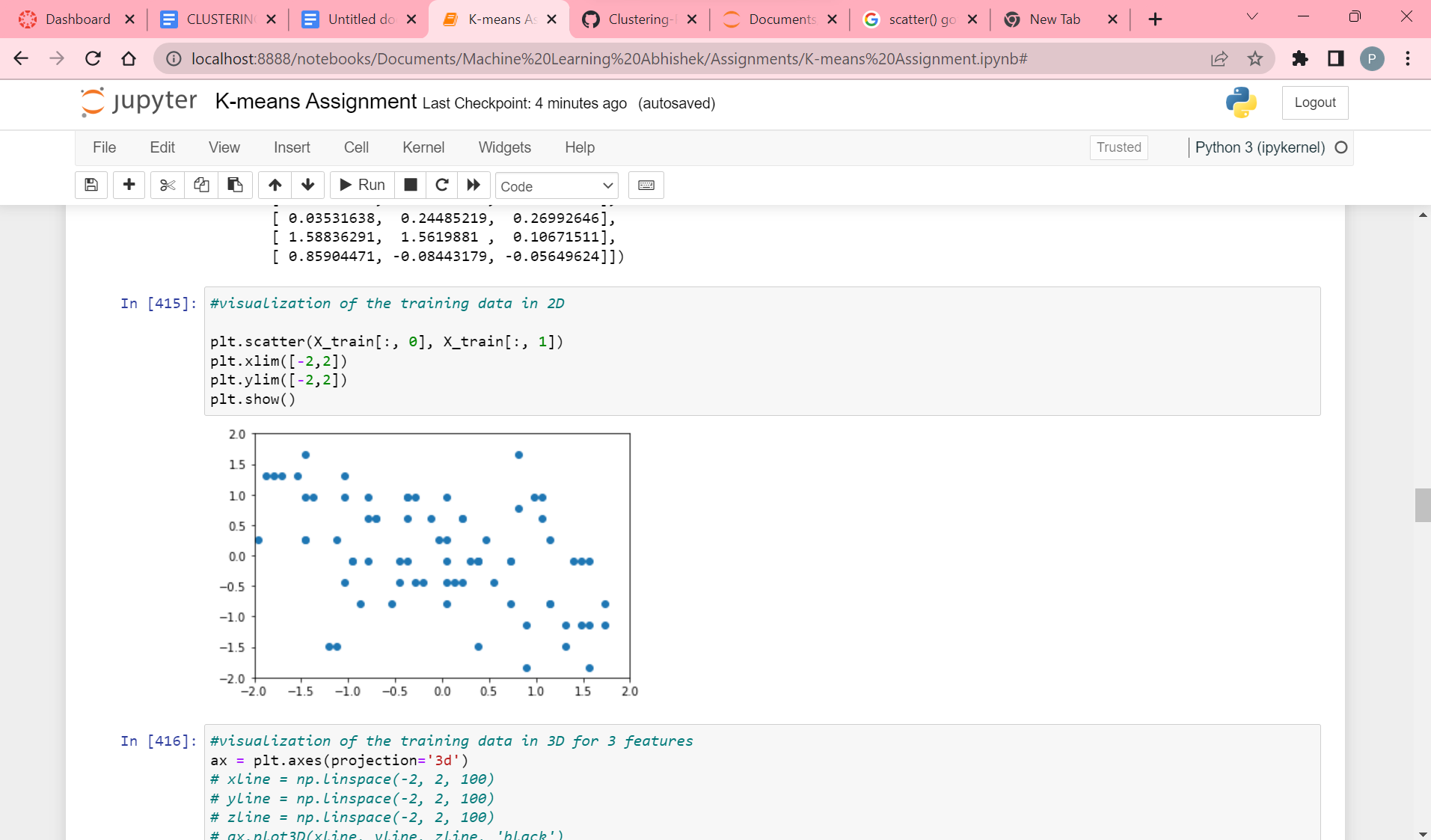
Getting the prefix and suffix of column names of dataset, to analyze the data.



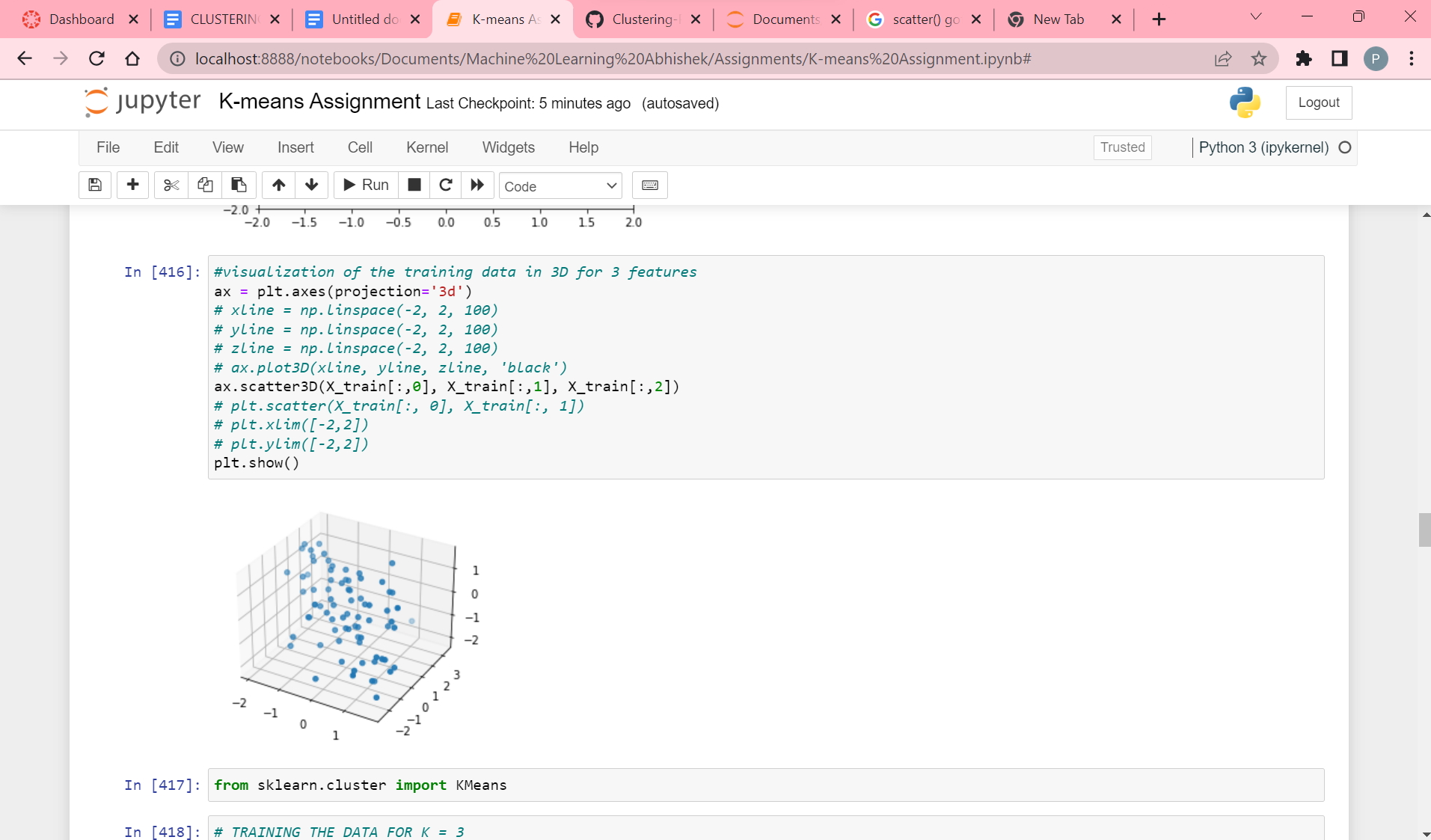
Getting the info and description of the training dataset to analyze its data for better performance.



Preparing the data. In this, we took min(number of rows) as there has to be the same index while data is processed. And took the first 3 columns as feature variables. And then scaling the data for better accurate results using StandardScaling.

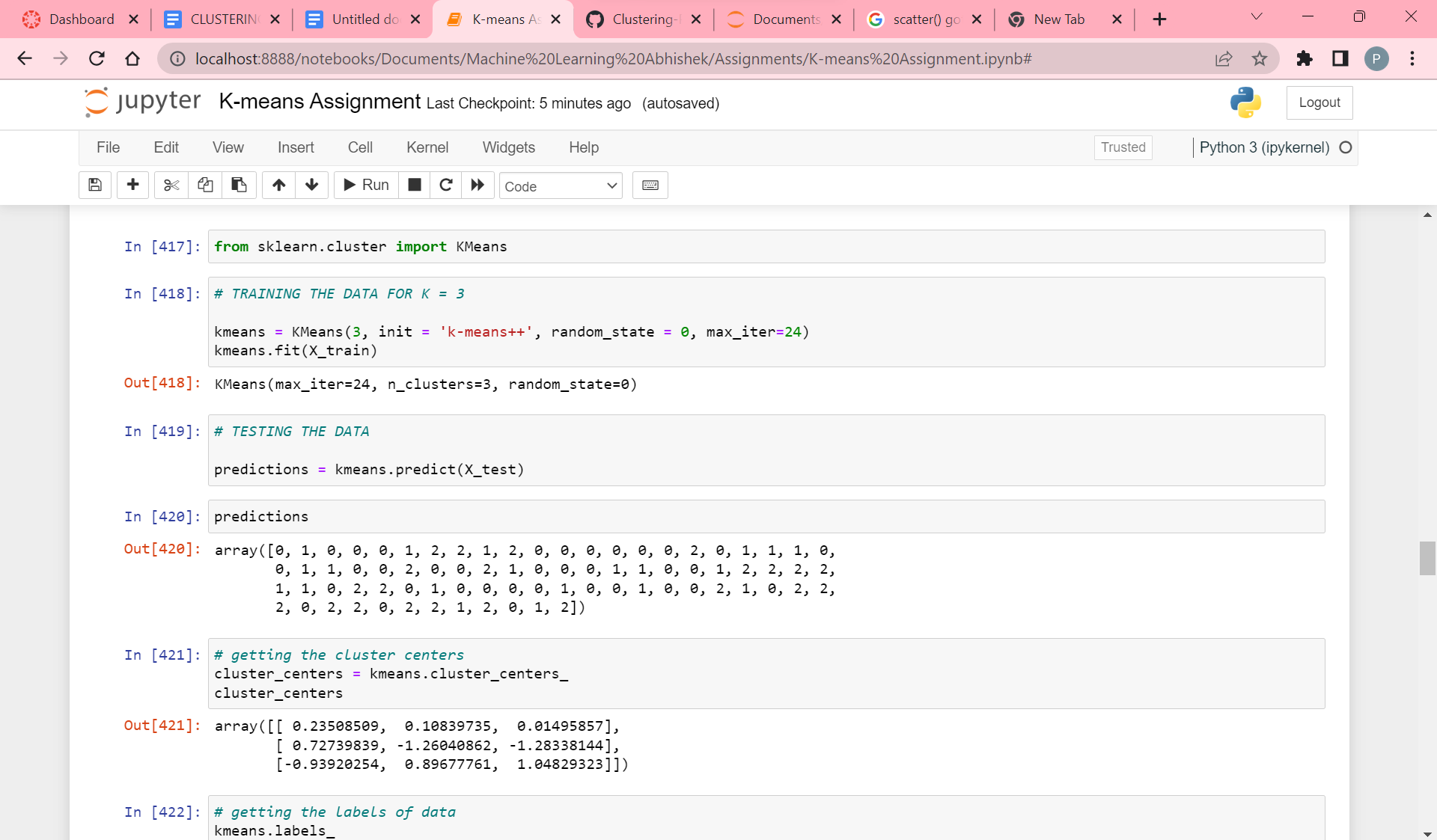


Visualizing the data for the first two features in 2D.



Visualizing the data for the 3 features taken in 3D

**K-Means clustering with K value as 3**



Processing the K-means algorithm with K value as 3. I.e It has 3 clusters.

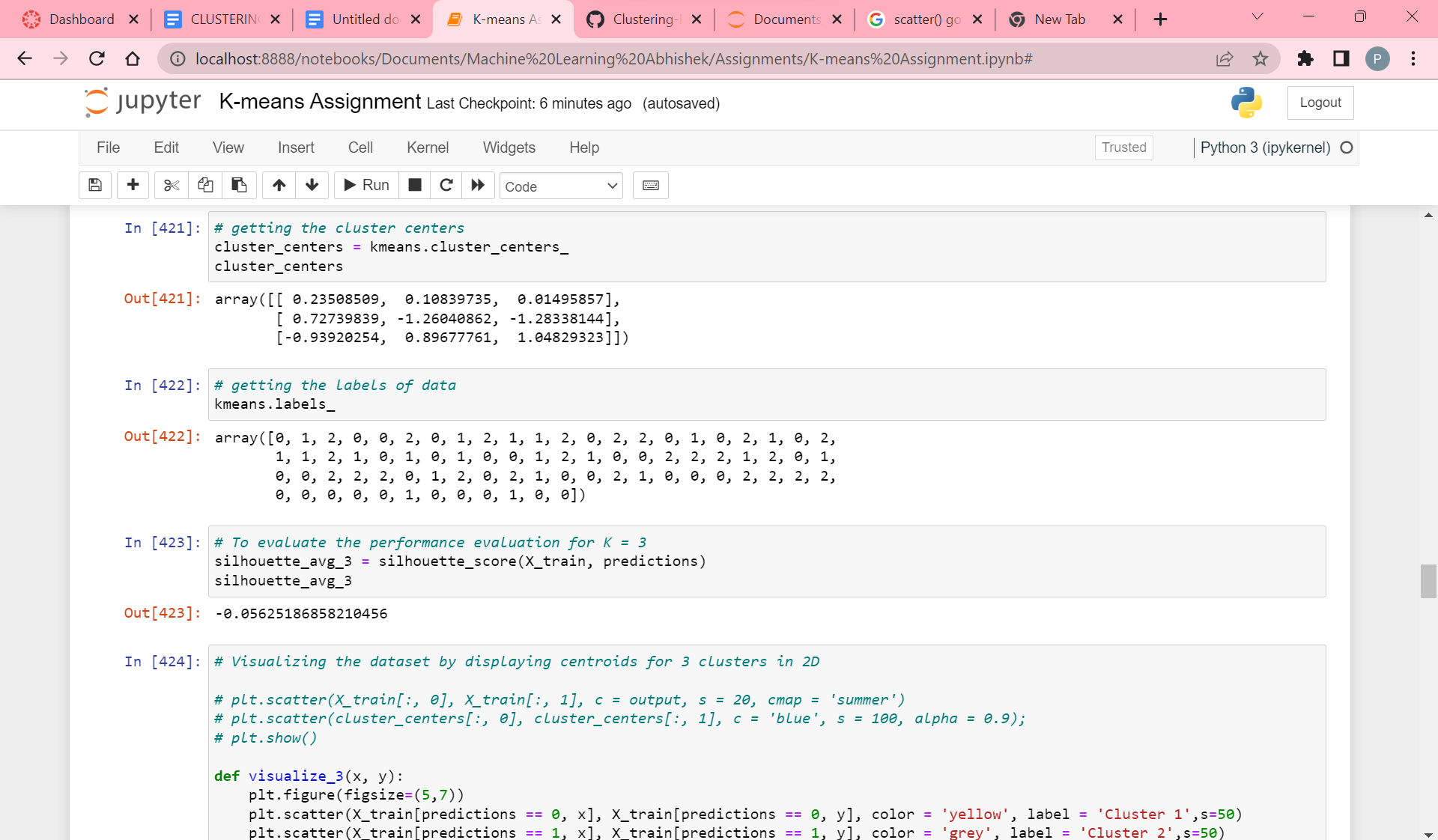
Used sklearn for KMeans.

Then trained the model using X\_train data with the help of the fit function.

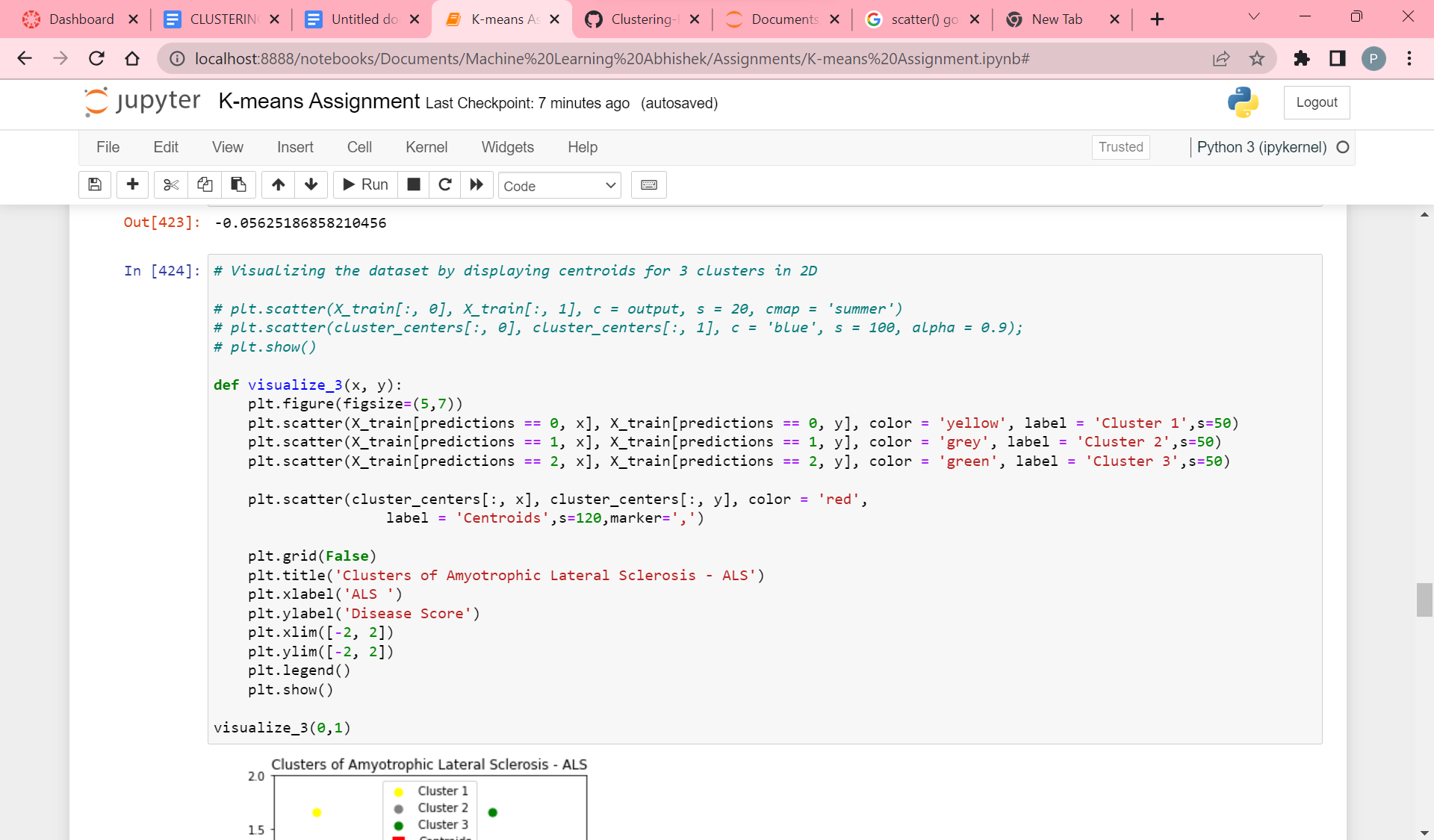
Then tested the model using X\_test data with the help of the predict function.

We get the output as [0,1,2] with length=3, as we took K value 3 for 3 clusters.

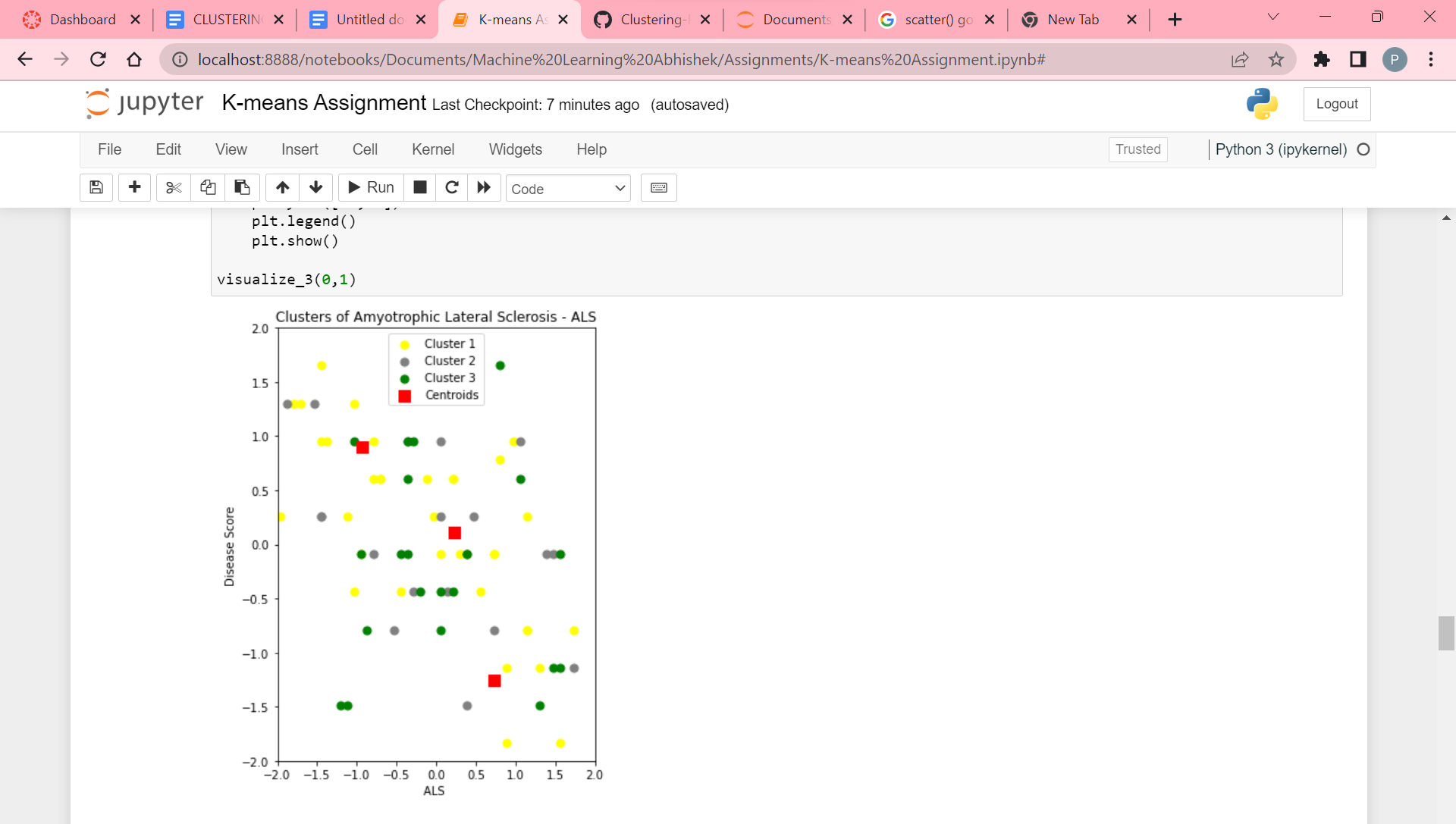
Now, we try to fetch the cluster centers using cluster\_centers\_. We get 3 rows for k value 3 and 3 columns for 3 feature variables.



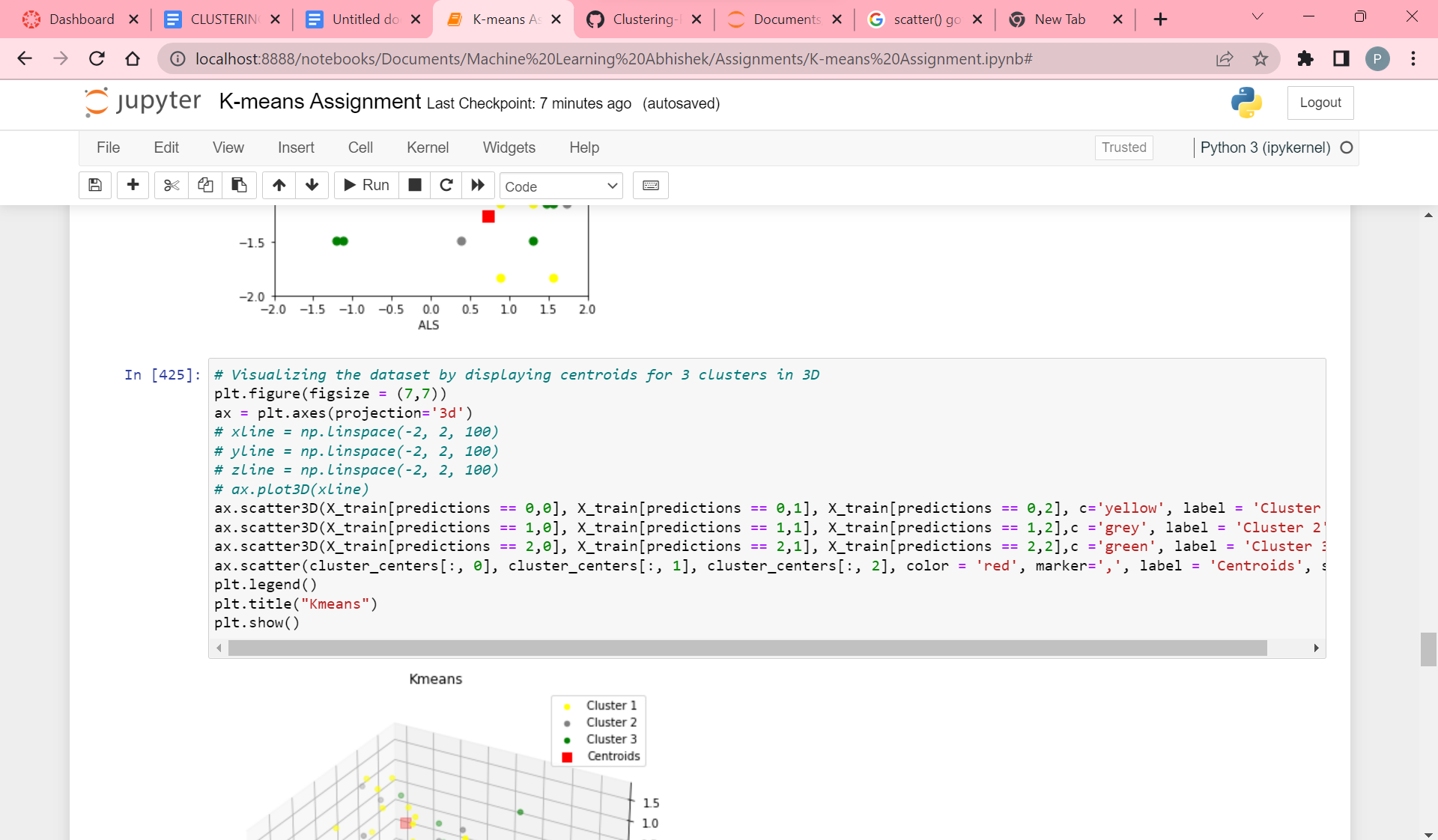
Getting labels of data and evaluating the performance of clusters using silhouette\_score. Here the score is -0.05625186858210456.



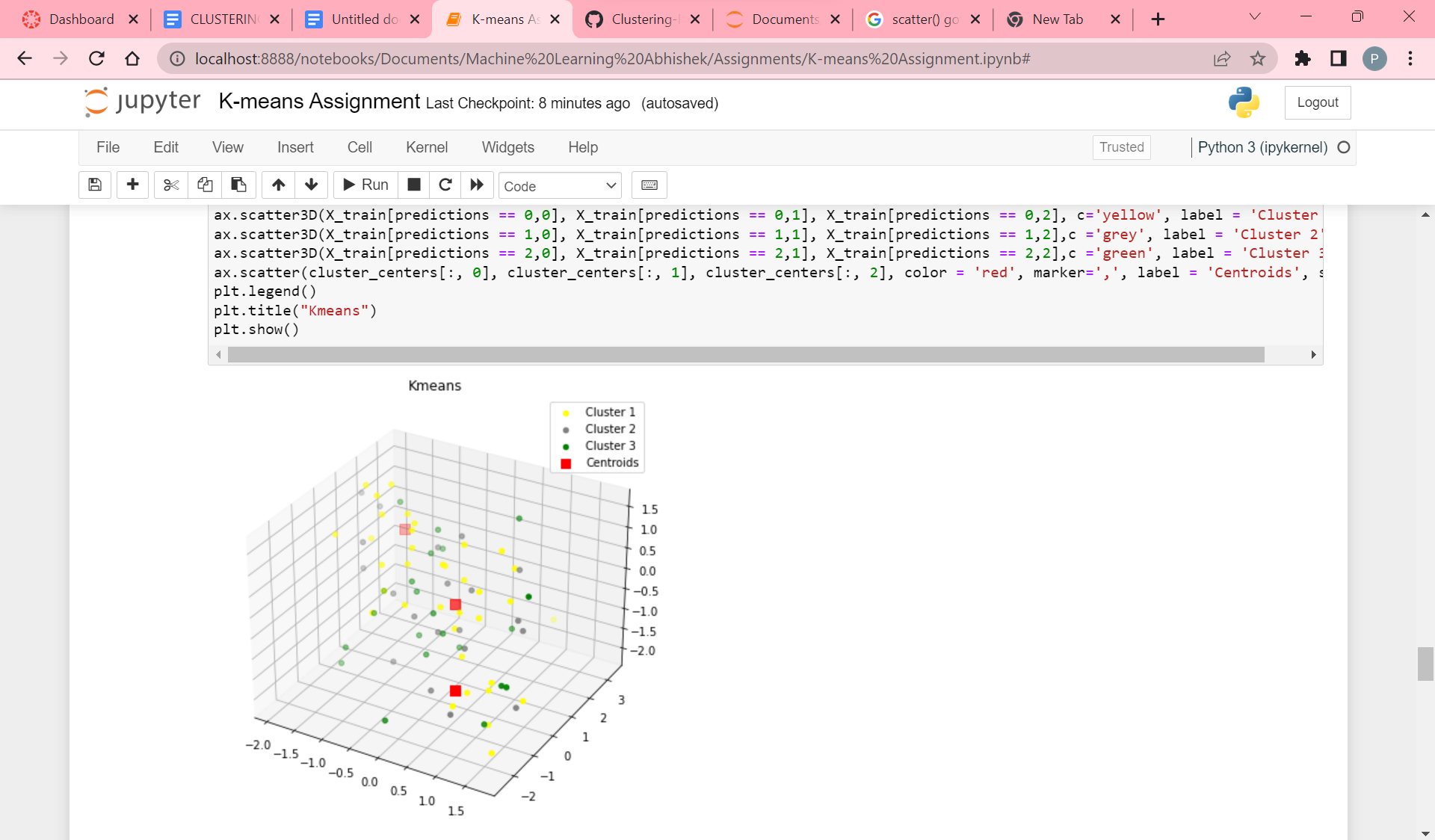
Code for visualizing the clusters through centroids for K value as 3 in 2D.



Visualization of clustering with K as value 3 in 2D.

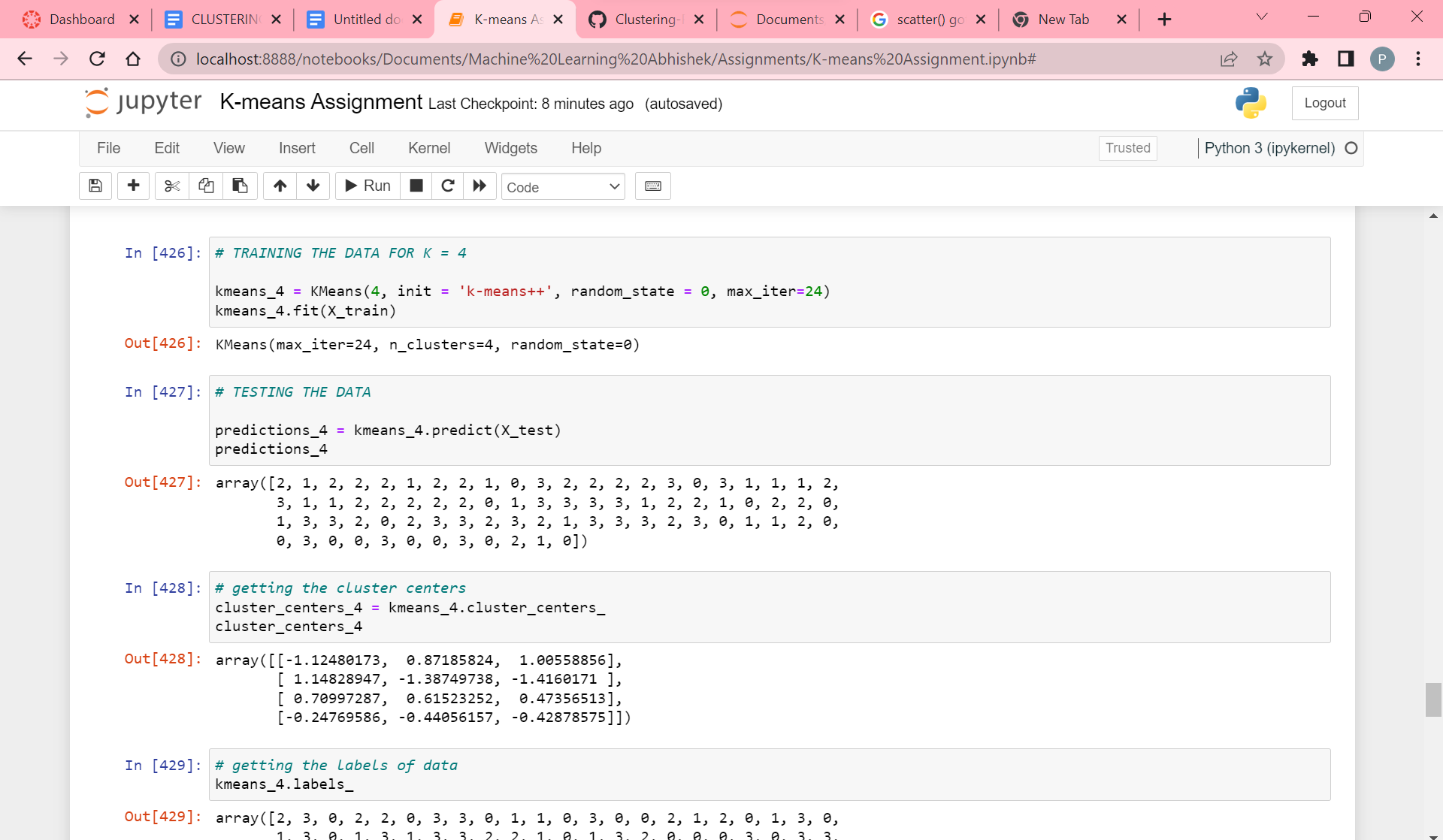


Code for visualizing the clusters through centroids for K value as 3 in 3D.



Visualization of clustering with K as value 3 in 3D and here yellow color data points indicate cluster\_1, gray color data points indicate cluster\_2, green color data points indicate cluster\_3 and red color indicate the centroids.

**K-Means Clustering with K value as 4**



Processing the K-means algorithm with K value as 4. I.e It has 4 clusters.

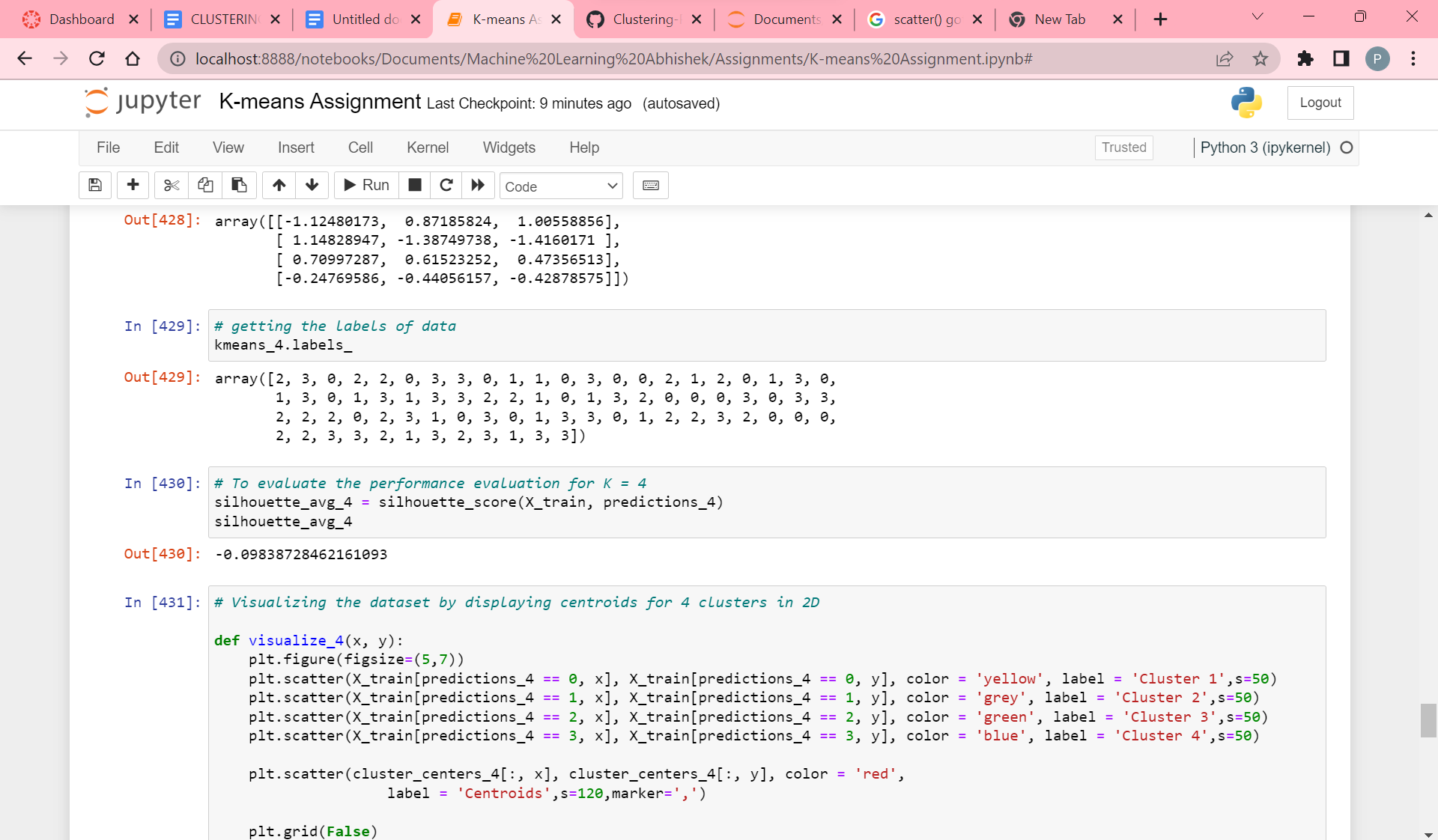
Used sklearn for KMeans.

Then trained the model using X\_train data with the help of the fit function.

Then tested the model using X\_test data with the help of the predict function.

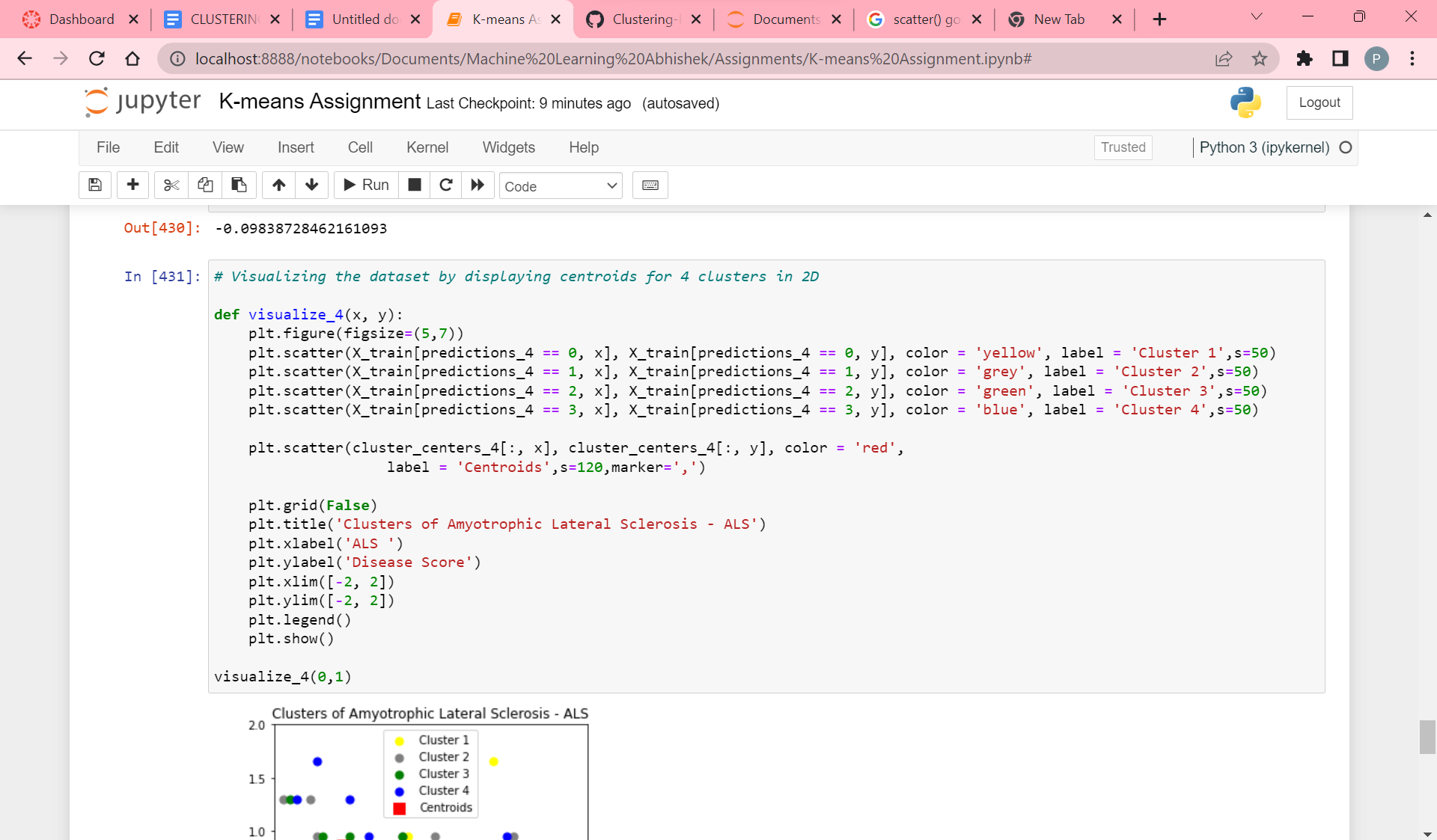
We get the output as [0,1,2,3] with length=4, as we took K value 4 for 4 clusters.

Now, we try to fetch the cluster centers using cluster\_centers\_. We get 4 rows for k value 4 and 3 columns for 3 feature variables

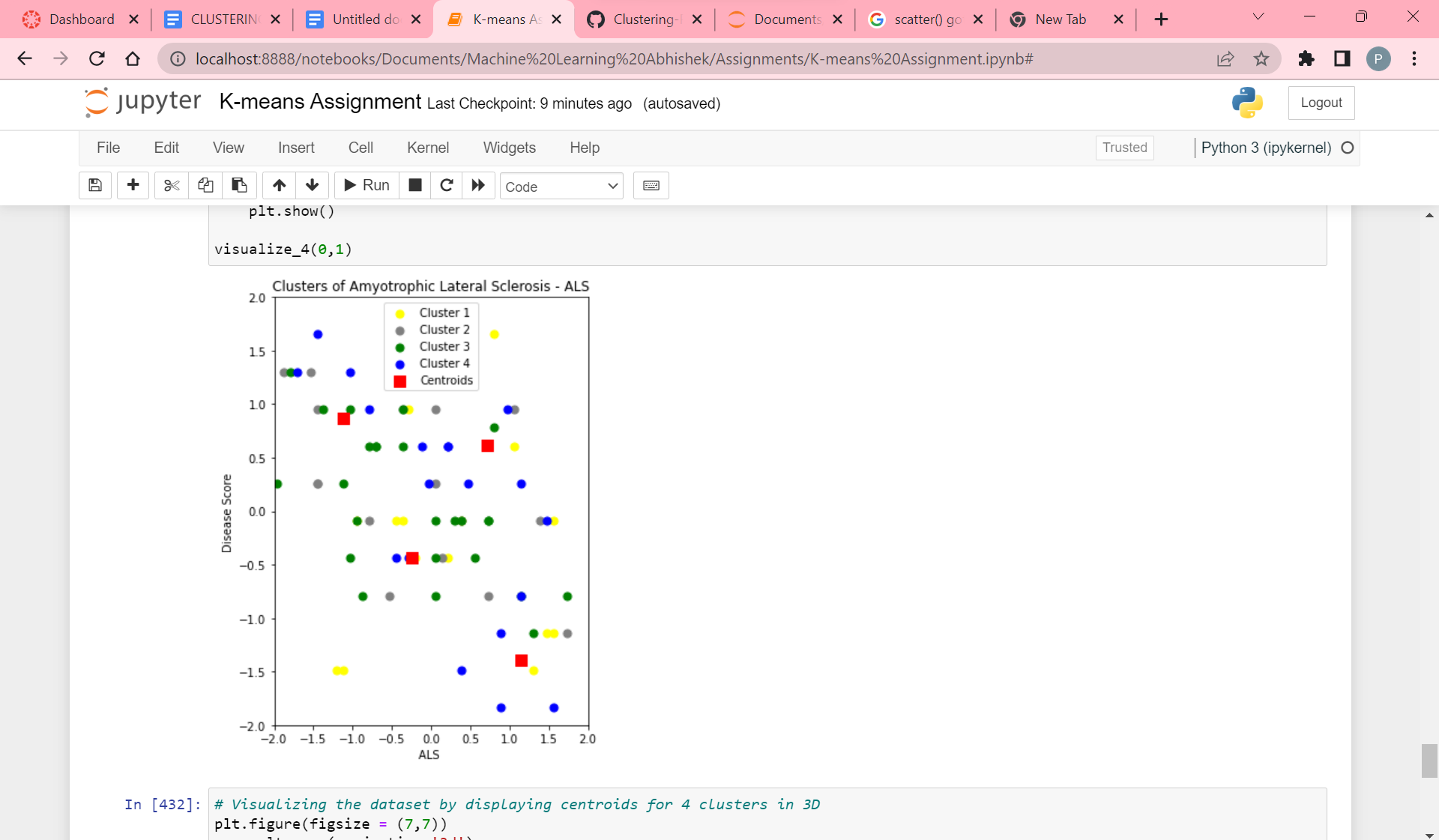


Getting labels of data and evaluating the performance of clusters using silhouette\_score. Here the score is -0.09838728462161093.

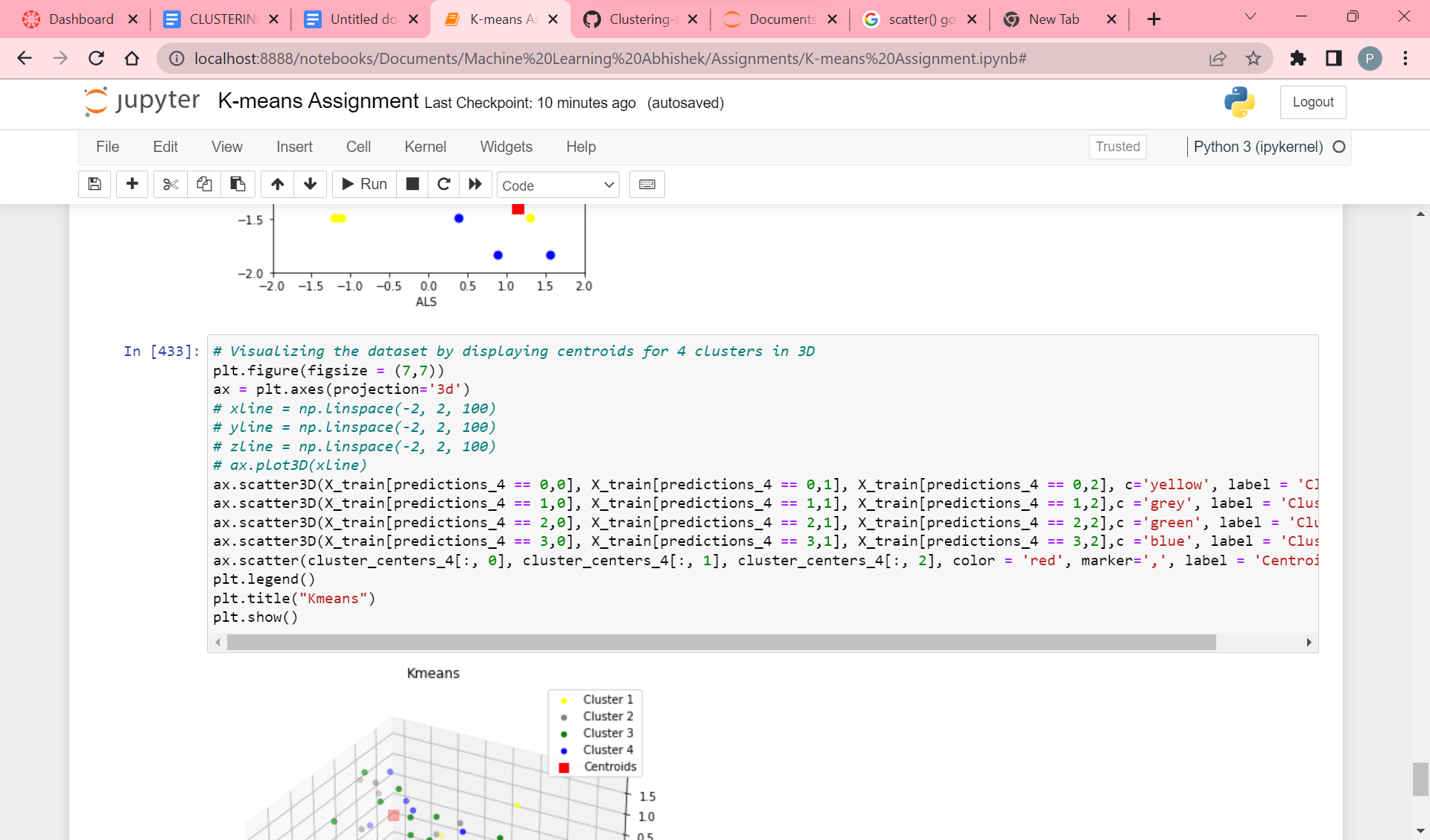
.



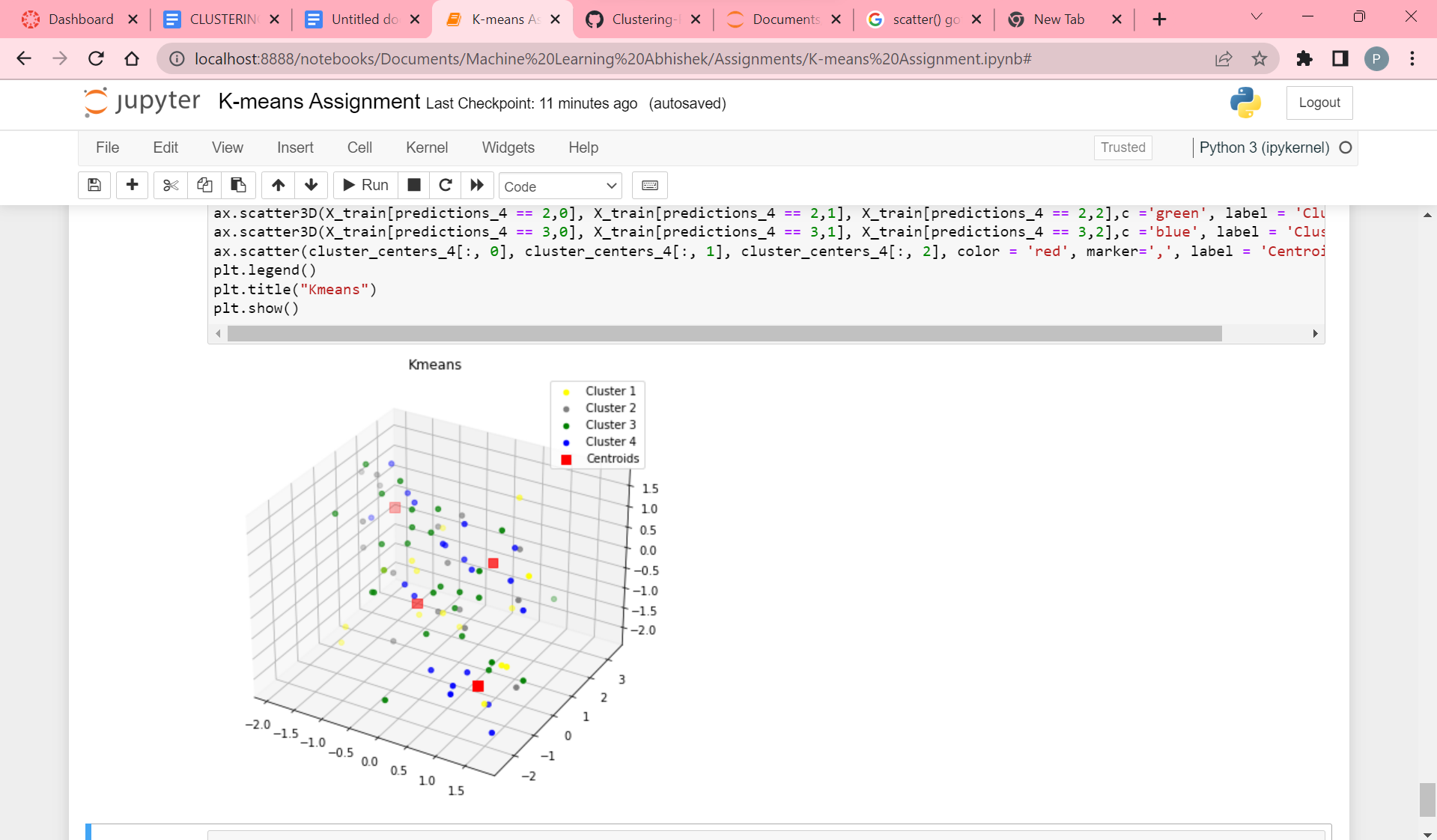
Code for visualizing the clusters through centroids for K value as 4 in 2D.



Visualization of clustering with K as value 4 in 2D.



Code for visualizing the clusters through centroids for K value as 4 in 3D.



Visualization of clustering with K as value 4 in 3D and here yellow color data points indicate cluster\_1, gray color data points indicate cluster\_2, green color data points indicate cluster\_3, blue color indicate cluster\_4 and red color data points indicate the centroids.

**Which K Value is better choice**

As I took the 2 K values, with one as K=3 and other as K=4. Hence, from the above processing of the data, I got the silhouette\_score. Where the score for K=3 is -0.05 and K=4 is -0.09. Hence from the above scores, we can say that K=3 is a better choice. Because the higher the silhouette\_score, the better the algorithm.