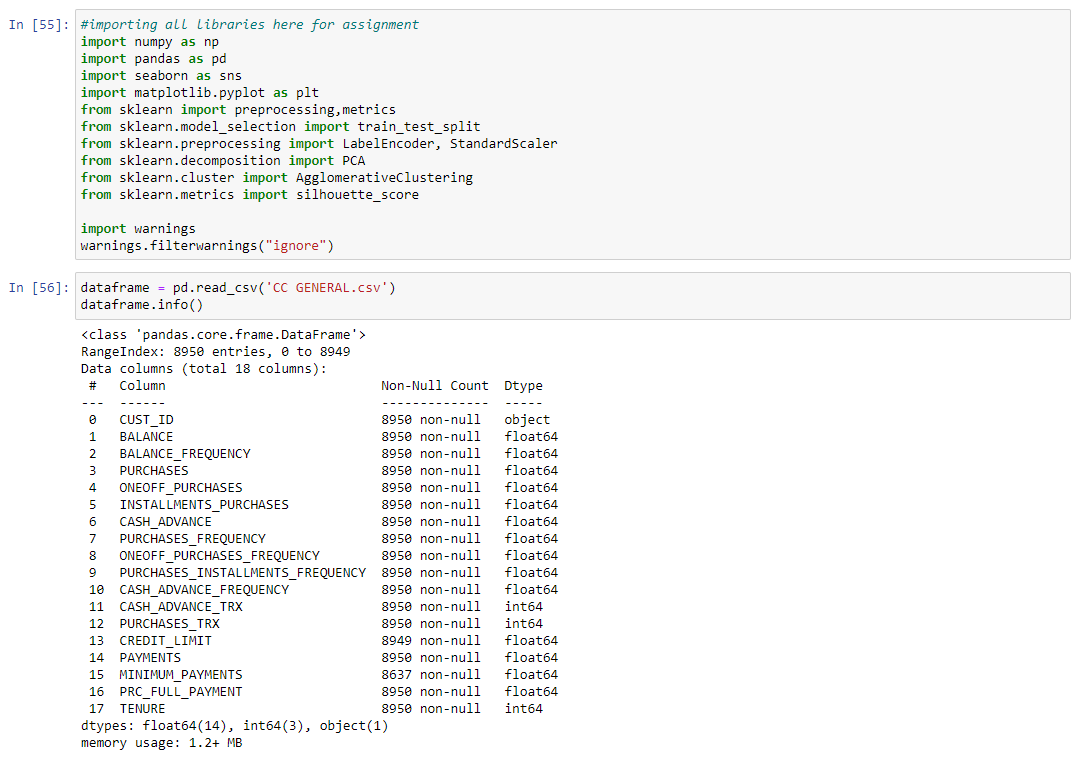
NAME: Sandeep Kumar Rai

GITHUB\_LINK: https://github.com/raisandeepkmr/assignment\_6.git

VIDEO\_LINK: <https://drive.google.com/drive/folders/1Vdx4KE5dXfa7xQApfBVou3D9oAi_ofjQ?usp=sharing>

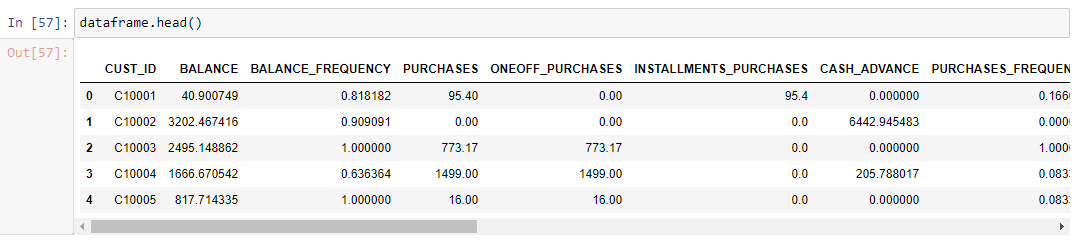
**In class programming:**

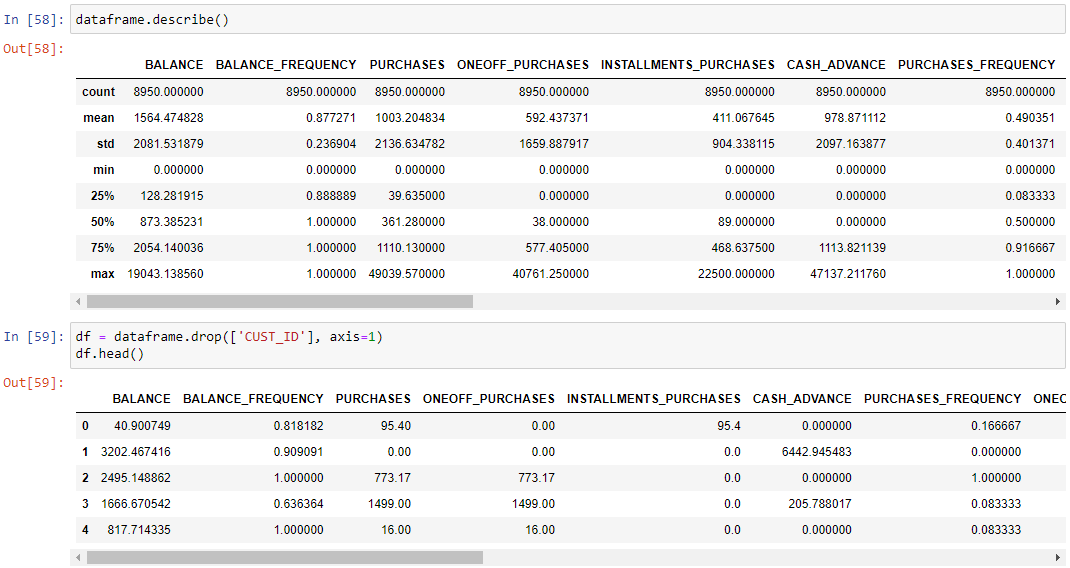
**Question 2:**

****

First, I have imported the required libraries. Then imported the dataset ‘CC GENERAL.csv’.

Dataset is also displayed using head () function and there is description of the dataset.

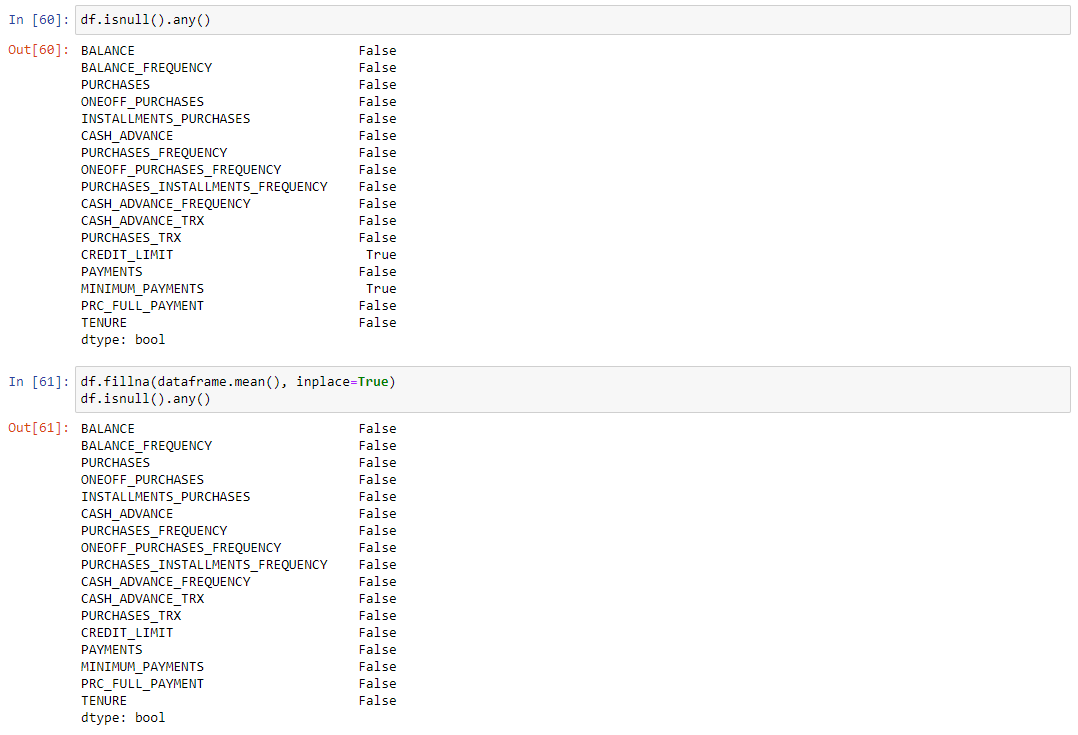




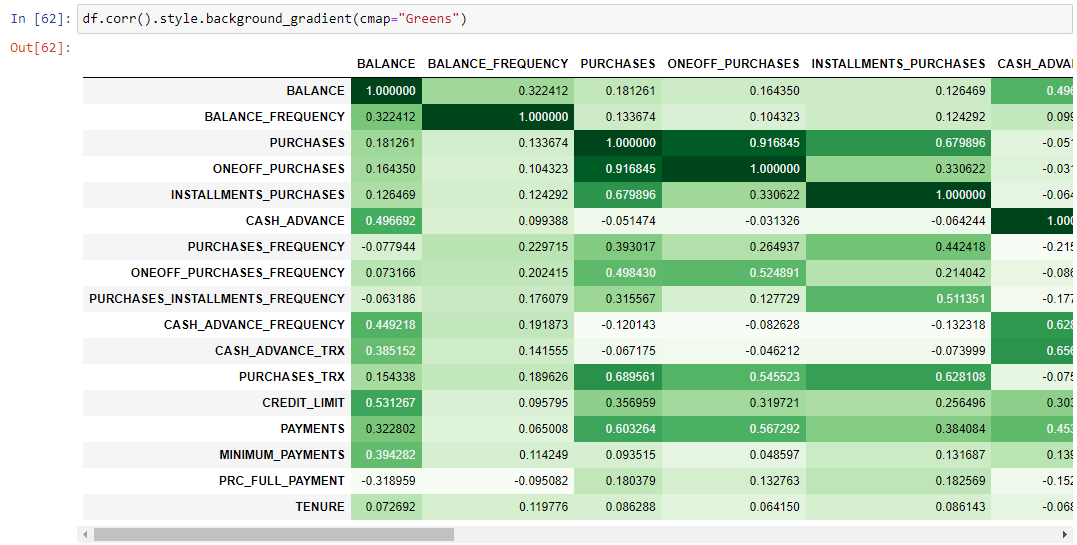
For Question 2(a), I have deleted the first column which is ‘CUST\_ID’. I have checked for the null

values in the dataset there are 2 attributes with the null values. I have used the mean values to

fill the null values of those two attributes.



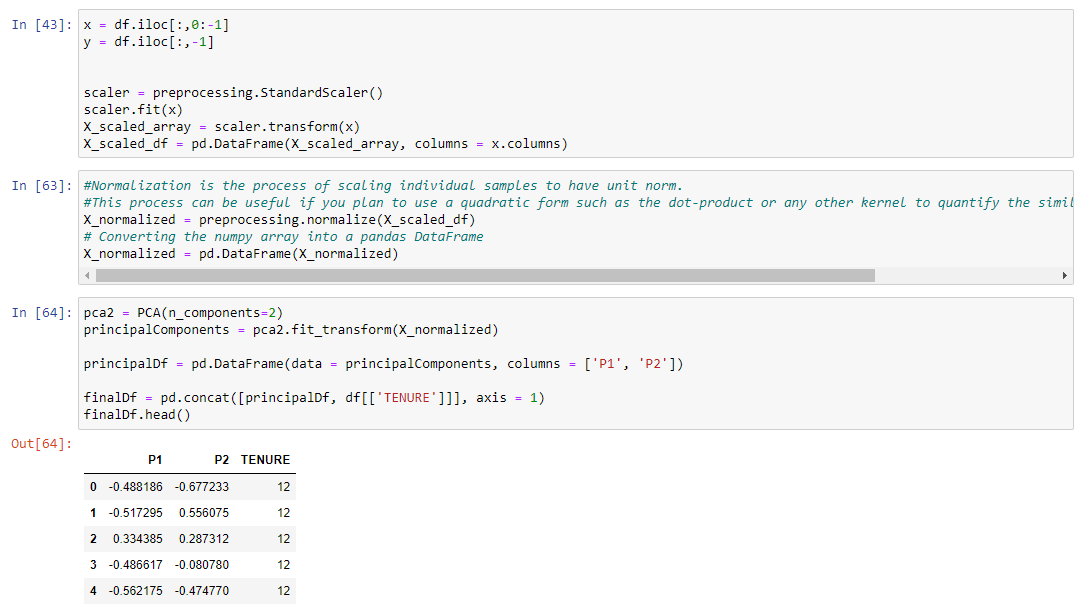
Use corr() function to find the correlation among the columns in the Dataframe using the ‘Pearson’ method. With green gradient.



For question 2(b), first I have applied the standard scaler. And then I have normalized the data

using normalize () function. In above screenshot I have displayed the dataset after the standard

scaler and after normalizing to see how dataset changes.



After applying normalizing we get array as an output so I have converted the array into panda

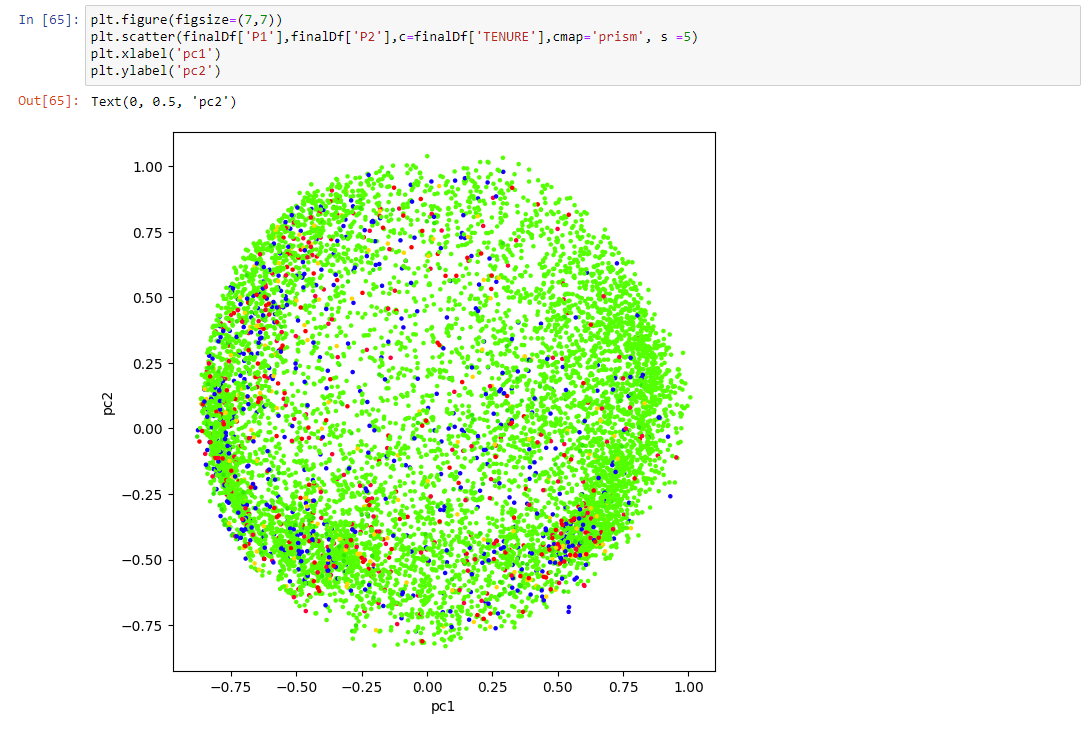
dataframe and displayed the dataset named ‘principalDf’.

For Question 2(c), I have implemented PCA where I taken k = 2. So, the dataset x\_norm has

been transformed into array. I have again transform the array into panda dataframe which has

2 column named ‘P1’,’P2’ and the name of the dataset is x\_pca. It is displayed in the

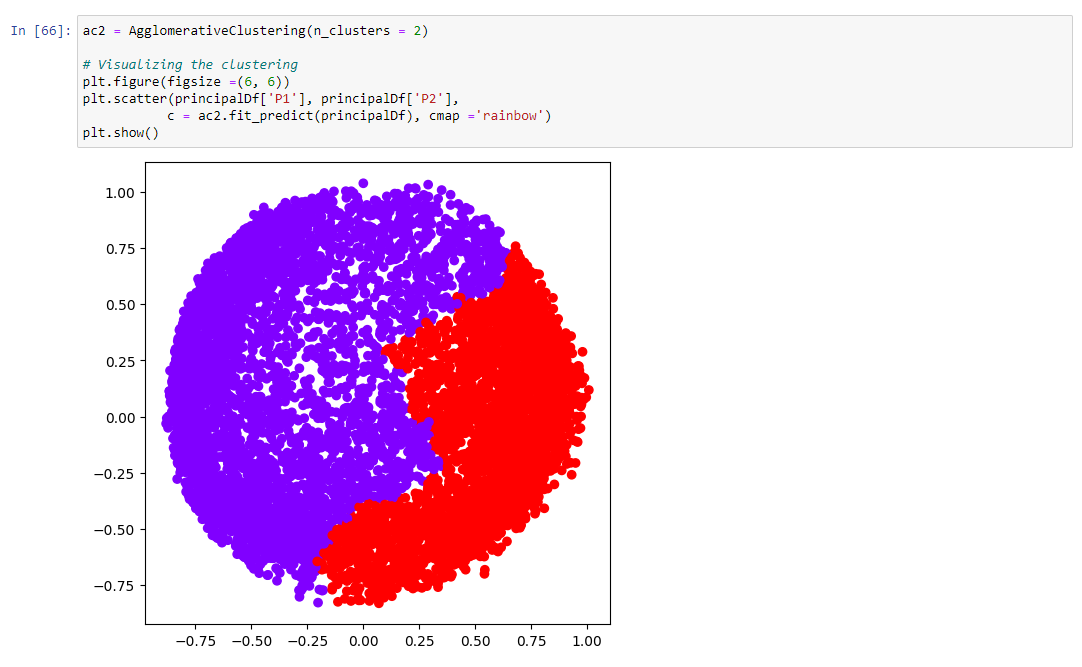
screenshot.

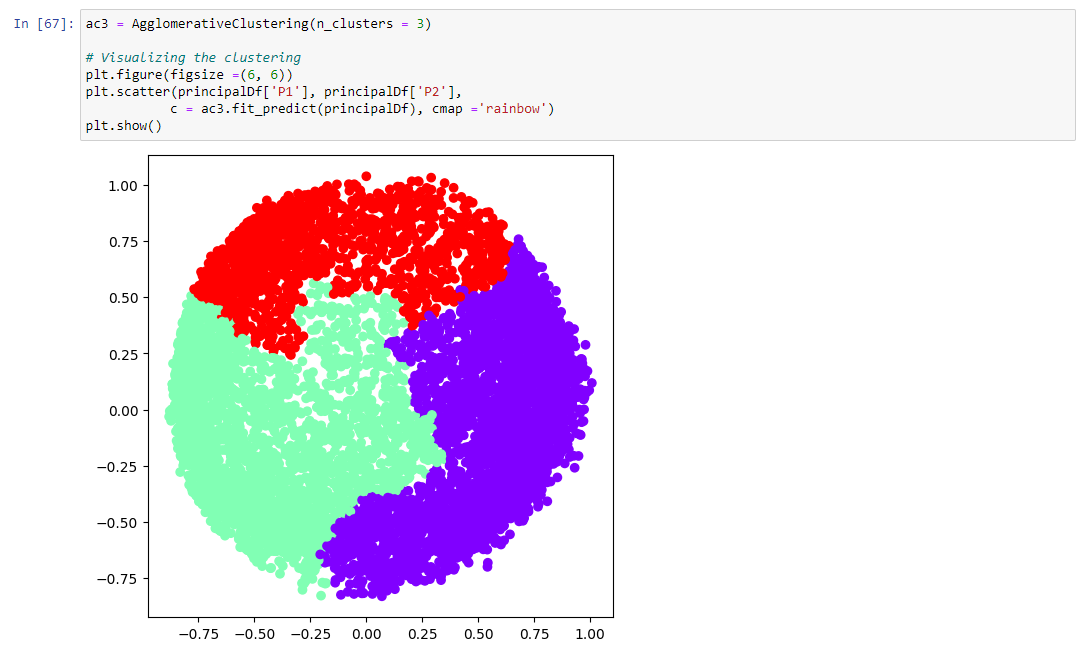


For question 2(d), I have implemented agglomerative clustering using sklearn library. Where the

number of clusters is 2. Also, the output has been displayed using the scatterplot. 2 different

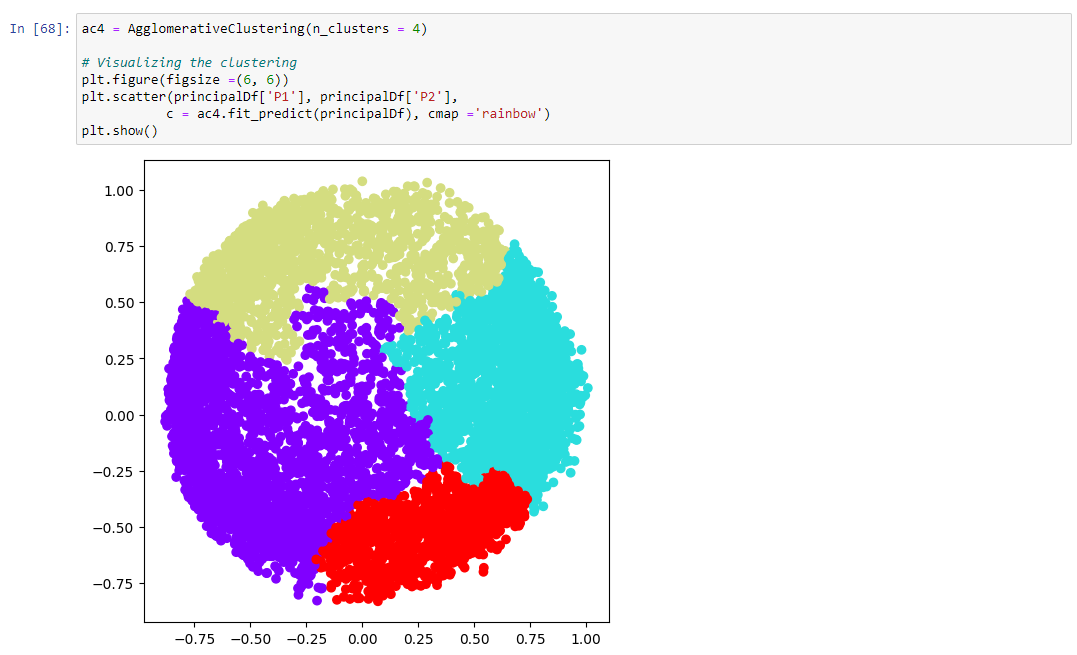
cluster has been displayed using two different colors.



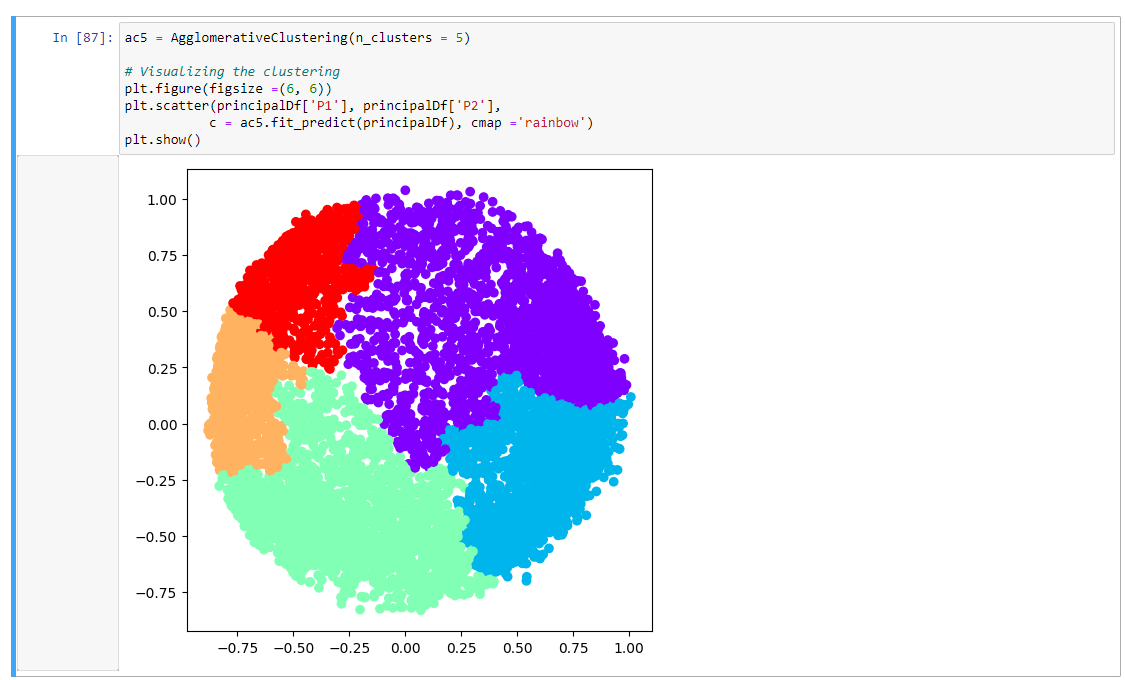


Above I have implemented the agglomerative cluster where number of clusters is 3. Three

different colors represent three clusters.



Above is the implementation of the agglomerative cluster with number of the cluster 4.



Above is the implementation of agglomerative cluster where number of cluster is 5.

For question 2(e), first I have calculated the silhouette score for all clusters model named

“S2,S3,S4,S5” and added to the list named “ss”.

I have used the bar graph to represent the silhouette score of each model. In bar graph y-axis

represent the silhouette score and x-axis represent cluster models.

