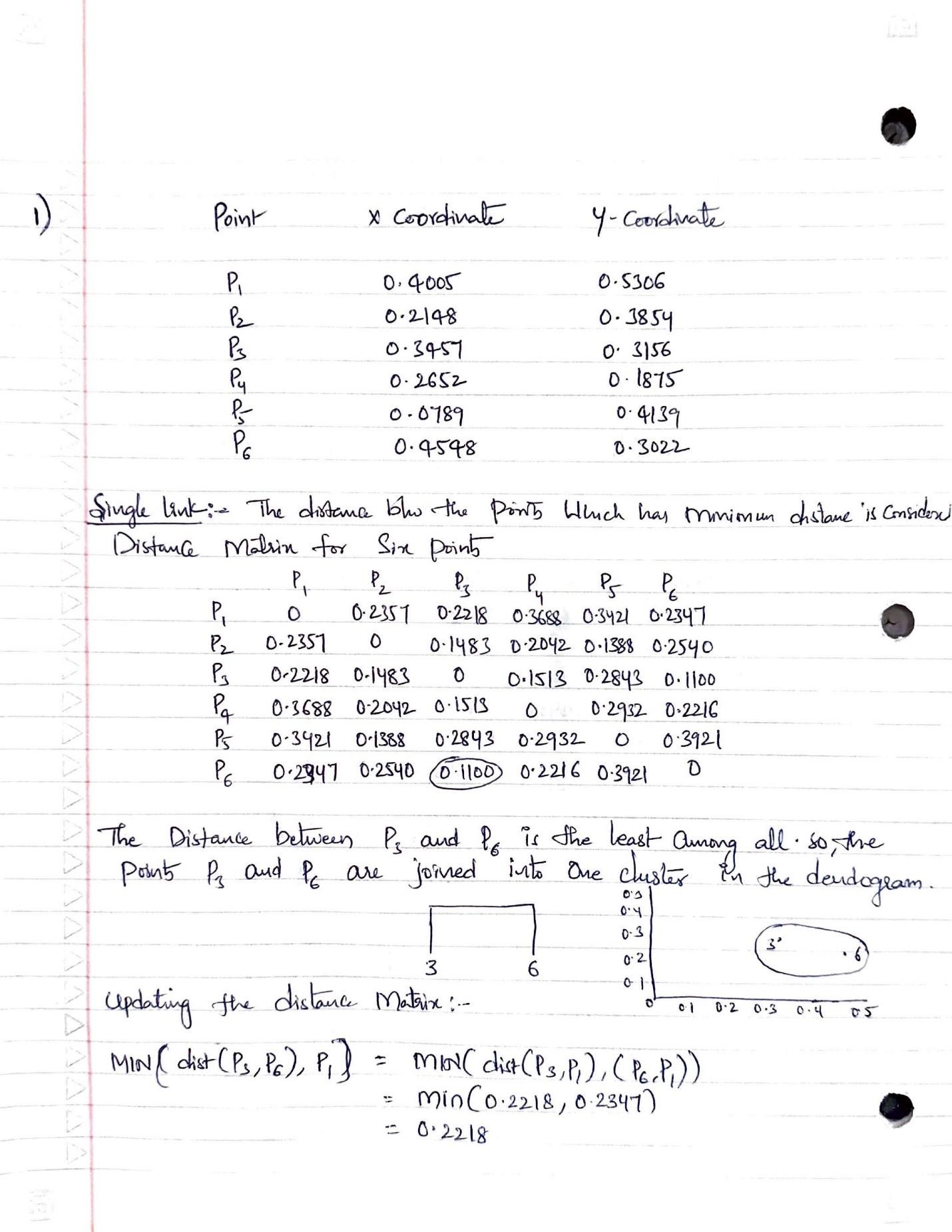
Assignment – 6(ML)

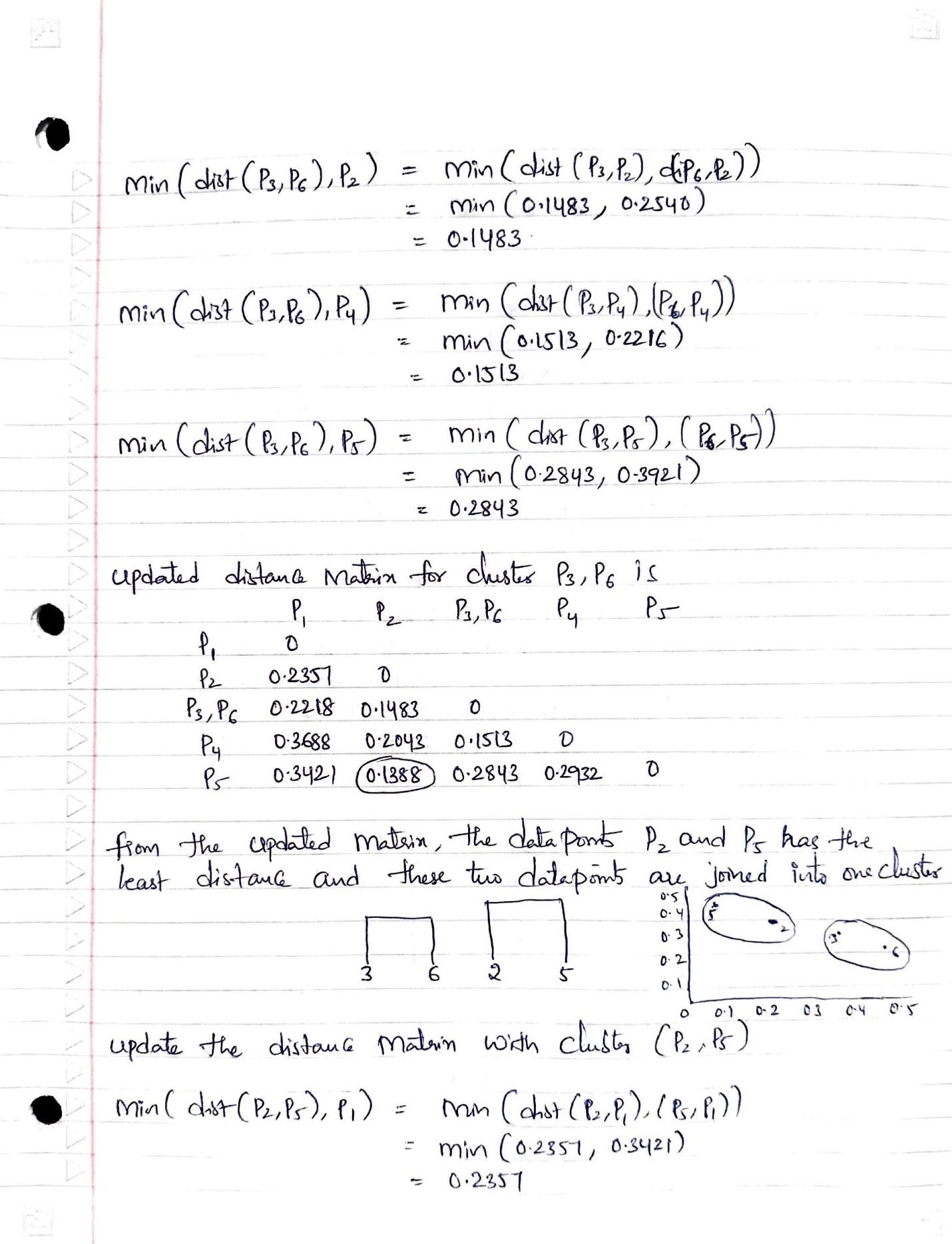
GitHub link to access the source code and video which demonstrates my work. <https://github.com/AdityaSaiVarma/Machine_Learning_Assignment-6>

Source code can be found under the source code file.

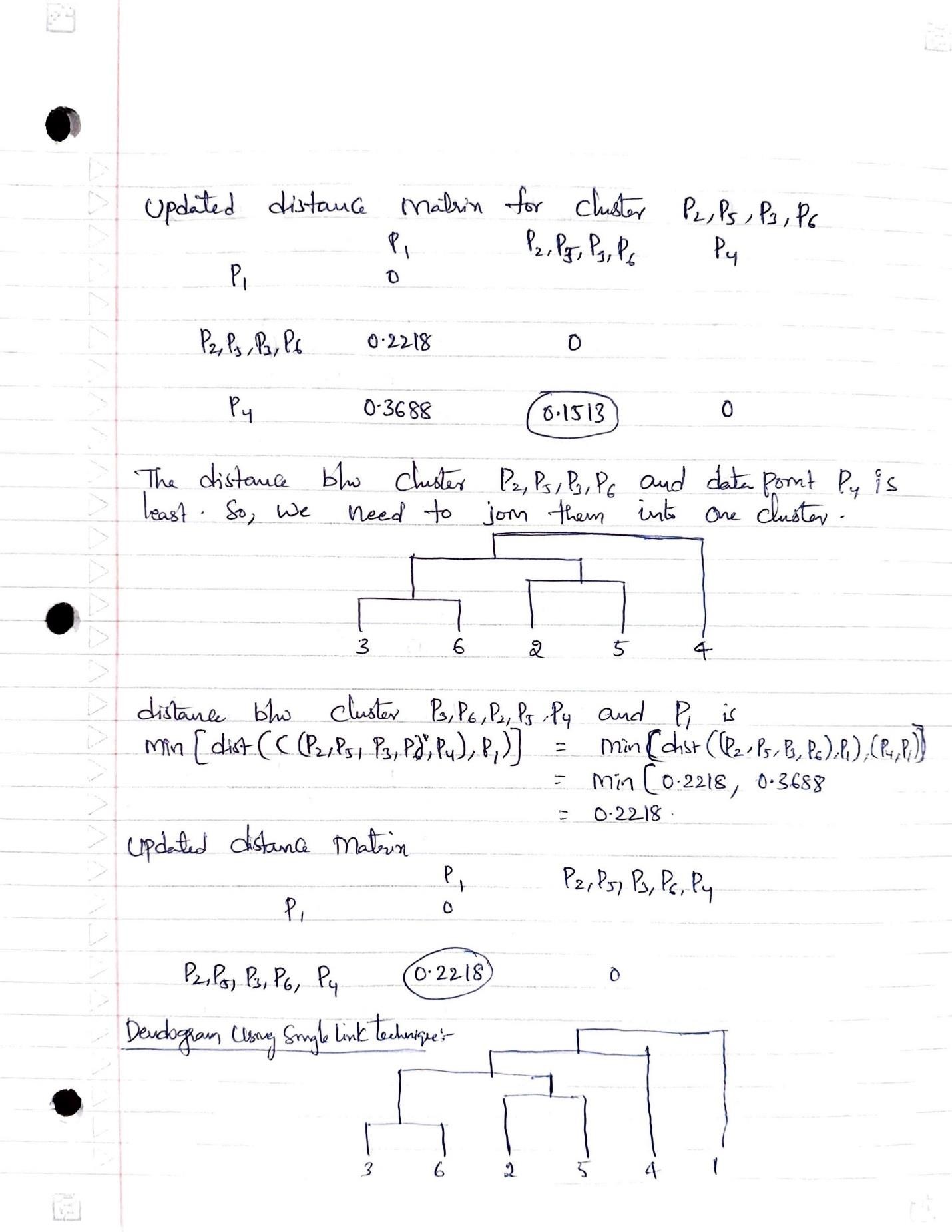
There is a README file which contains student Info, short description about assignment. A Video is available in README file under Video section.

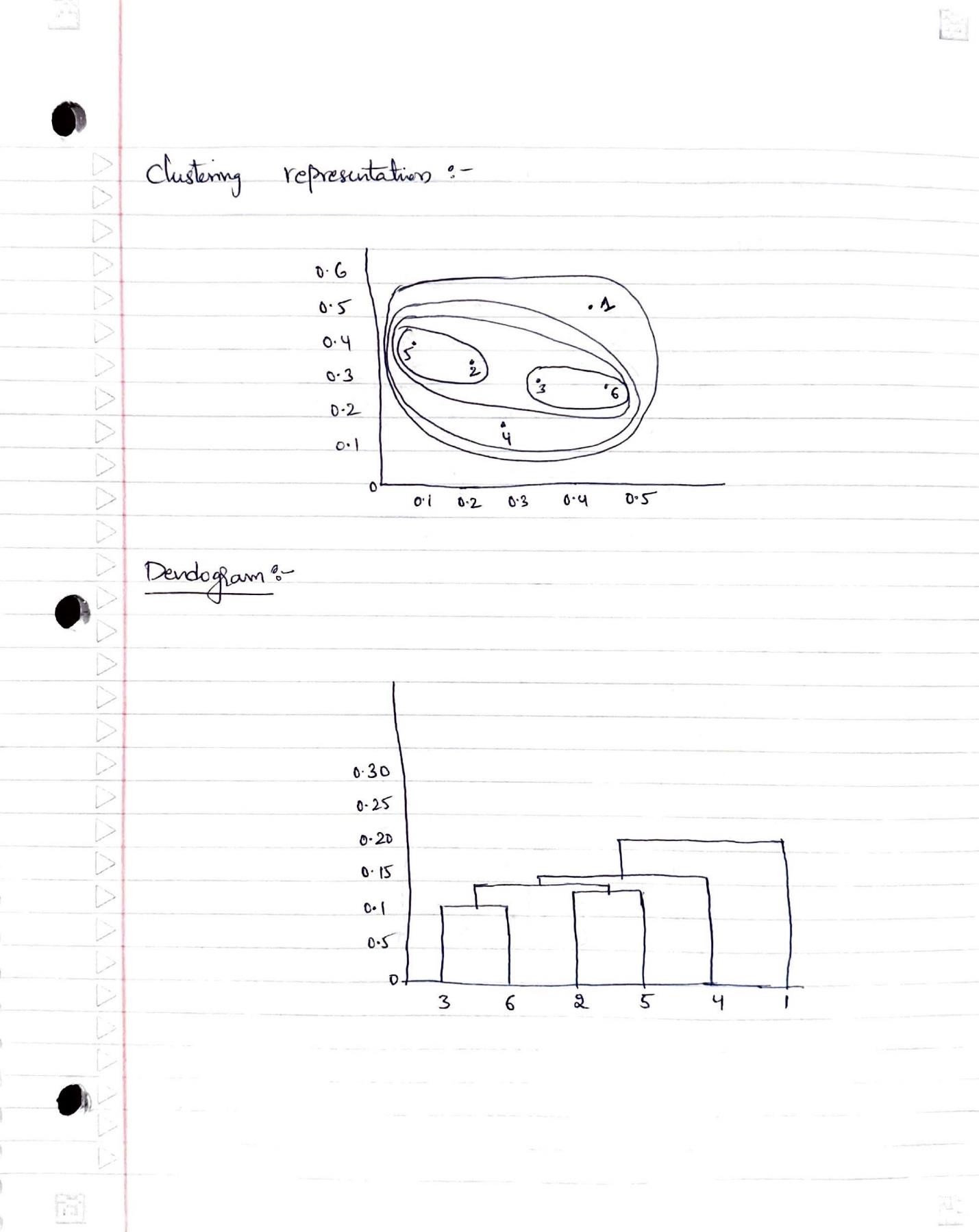
# 1. (Provide only mathematical solutions for this question) Six points with the following attributes are given, calculate and find out clustering representations and dendrogram using Single, complete, and average link proximity function in hierarchical clustering technique.

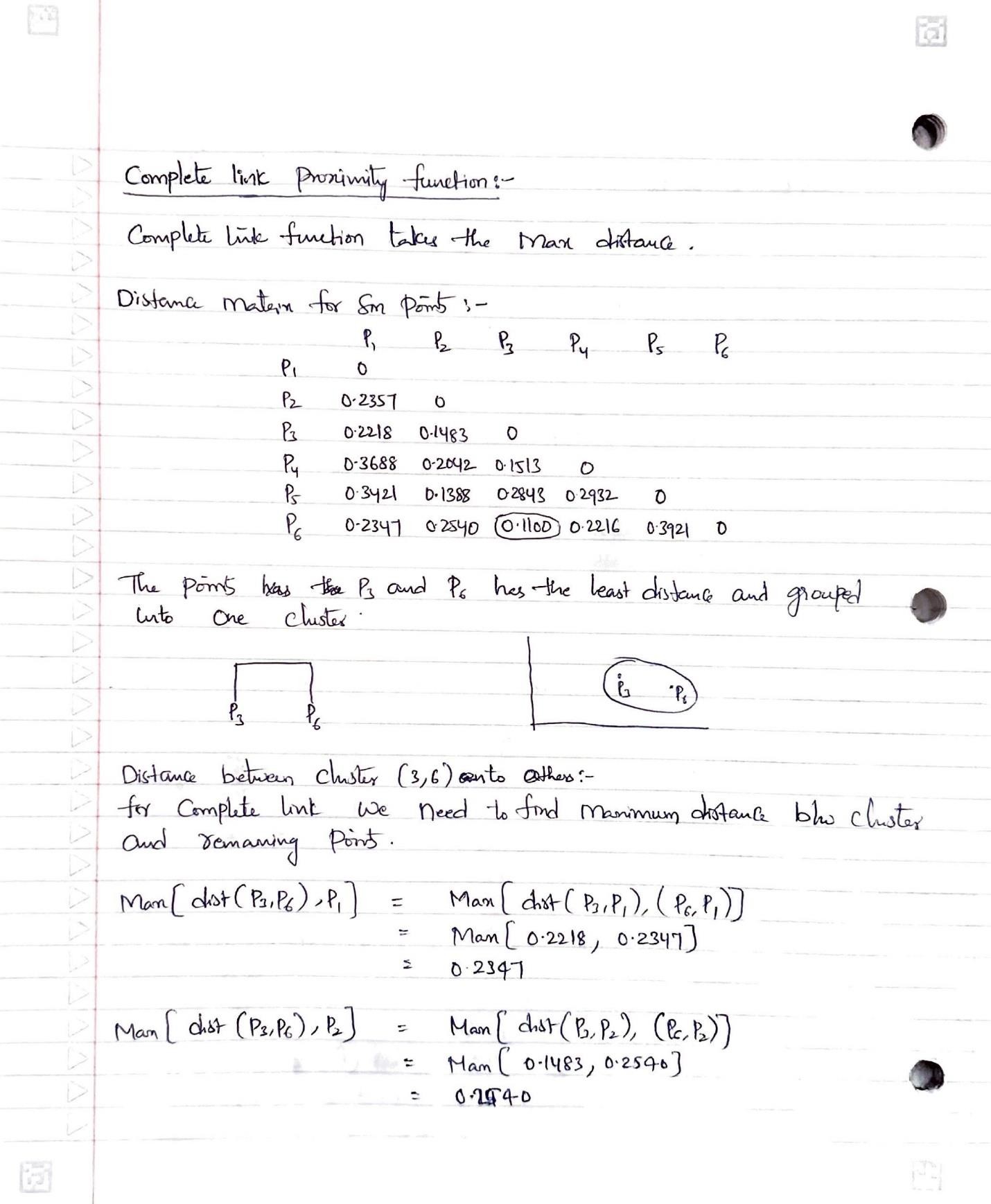


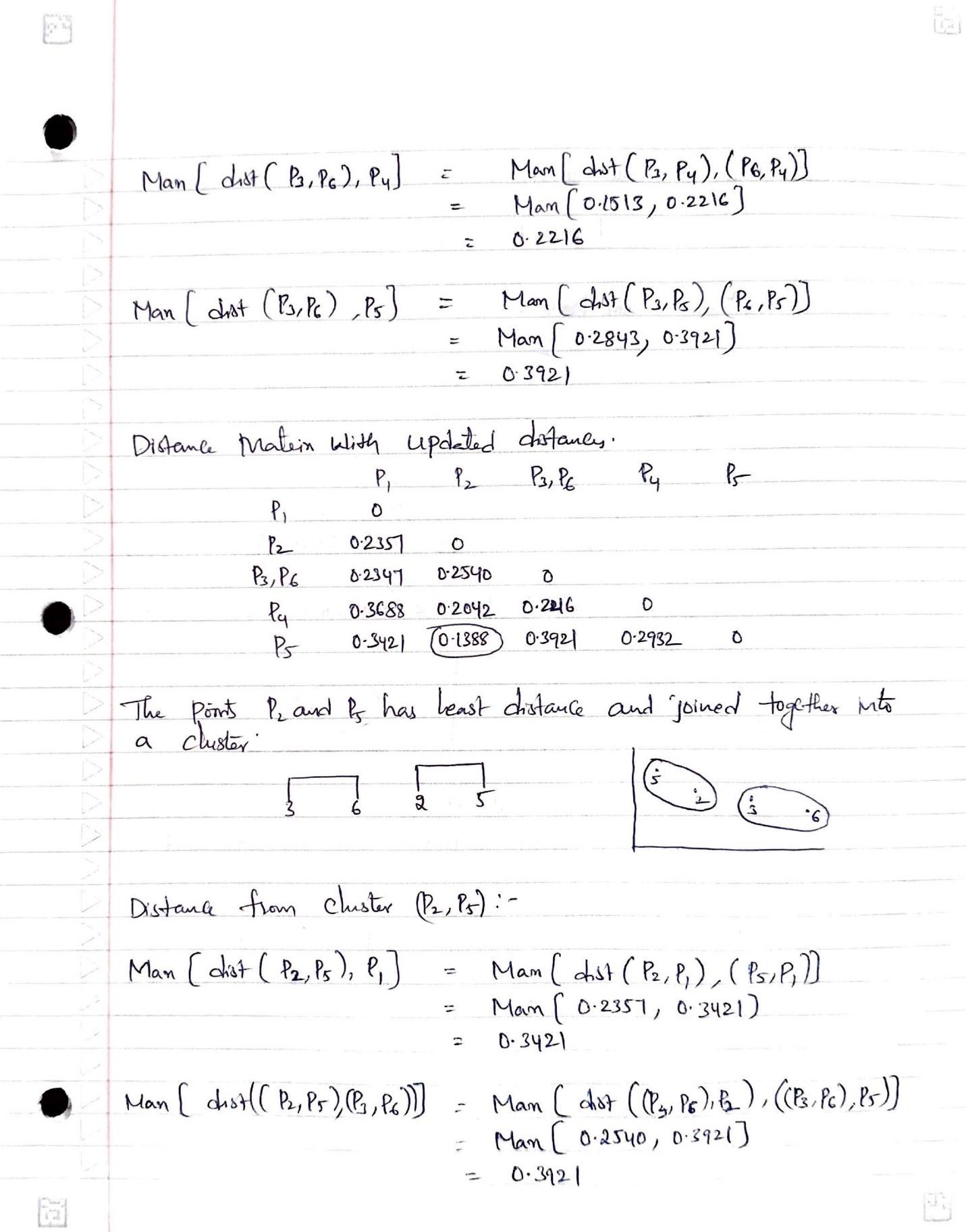


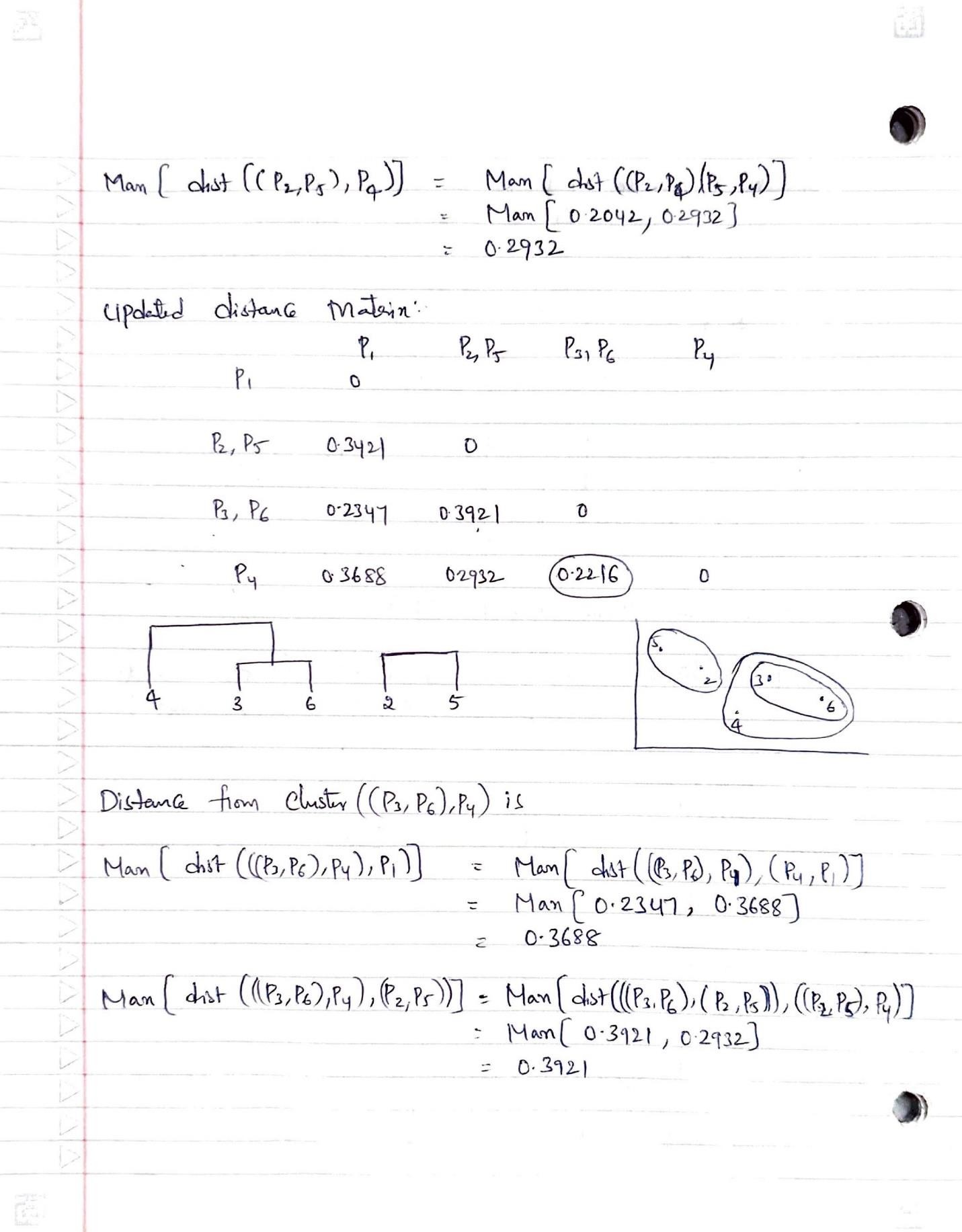


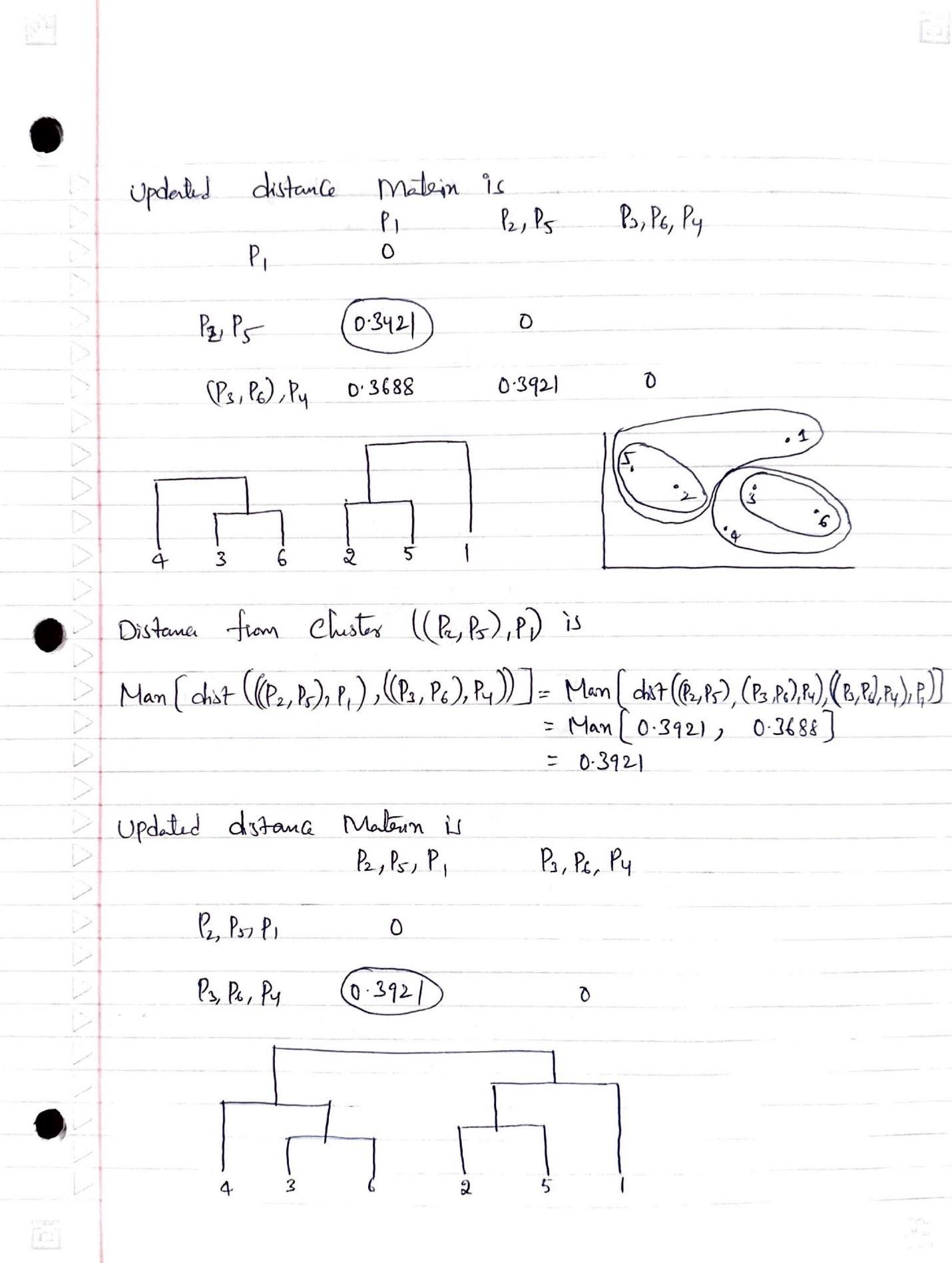


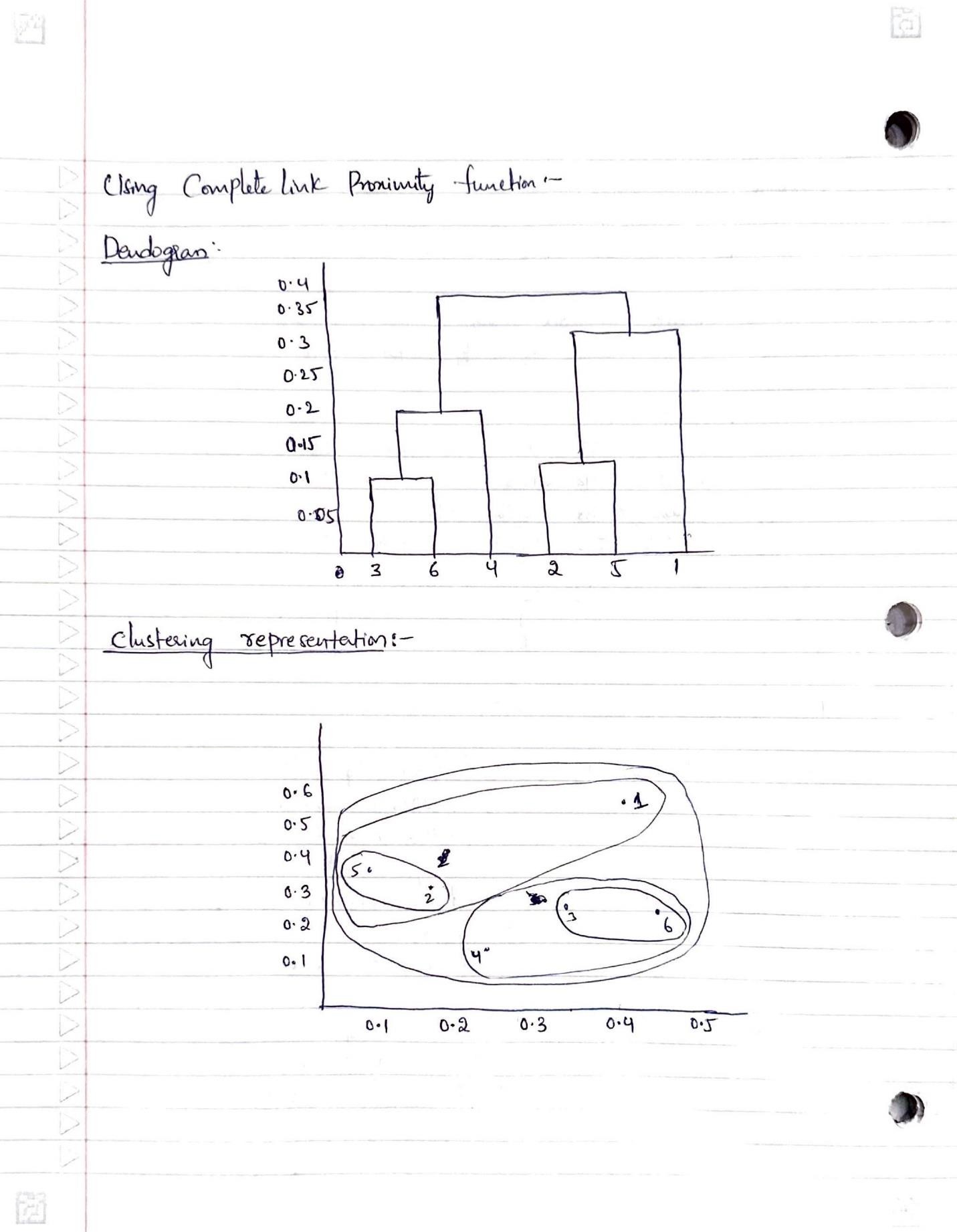


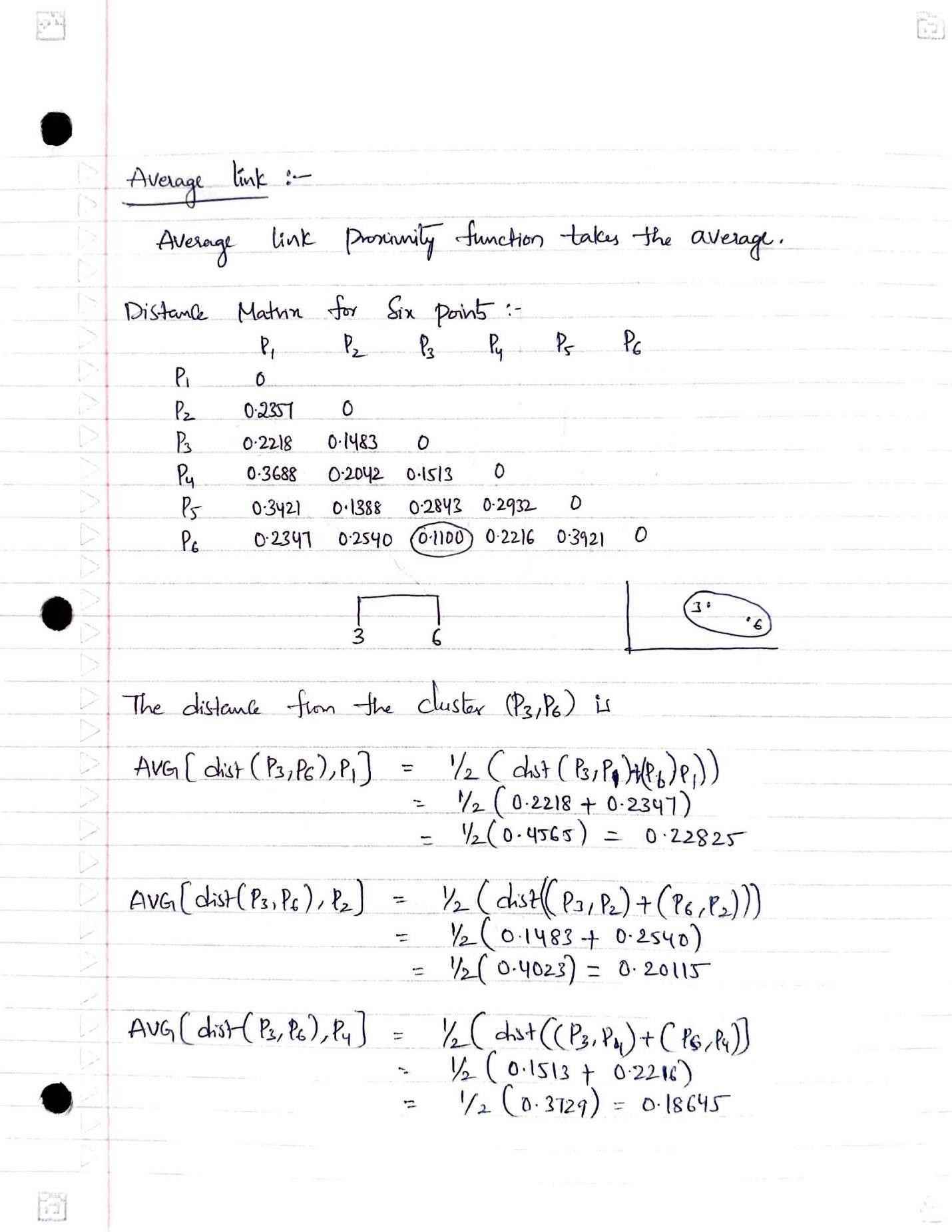


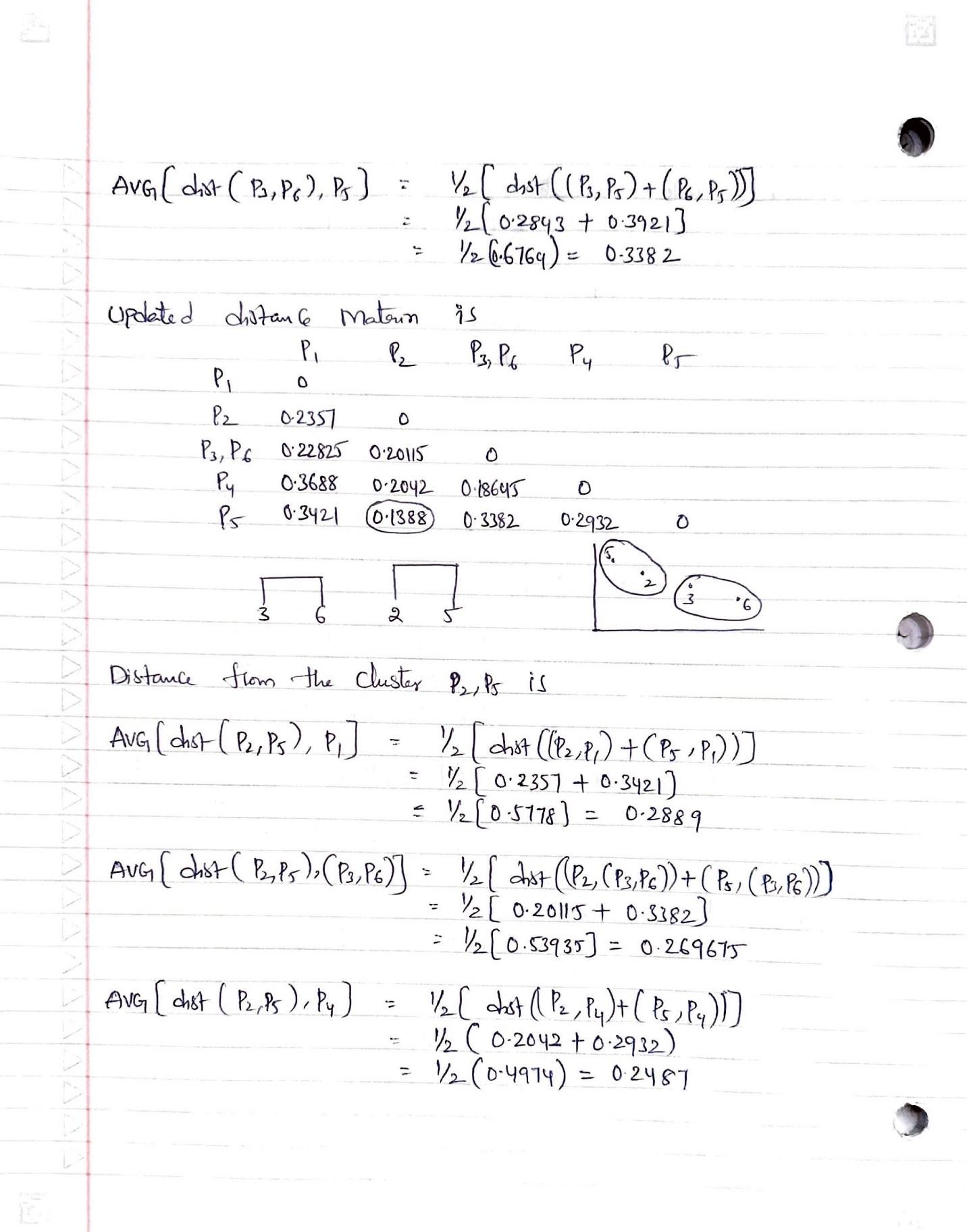


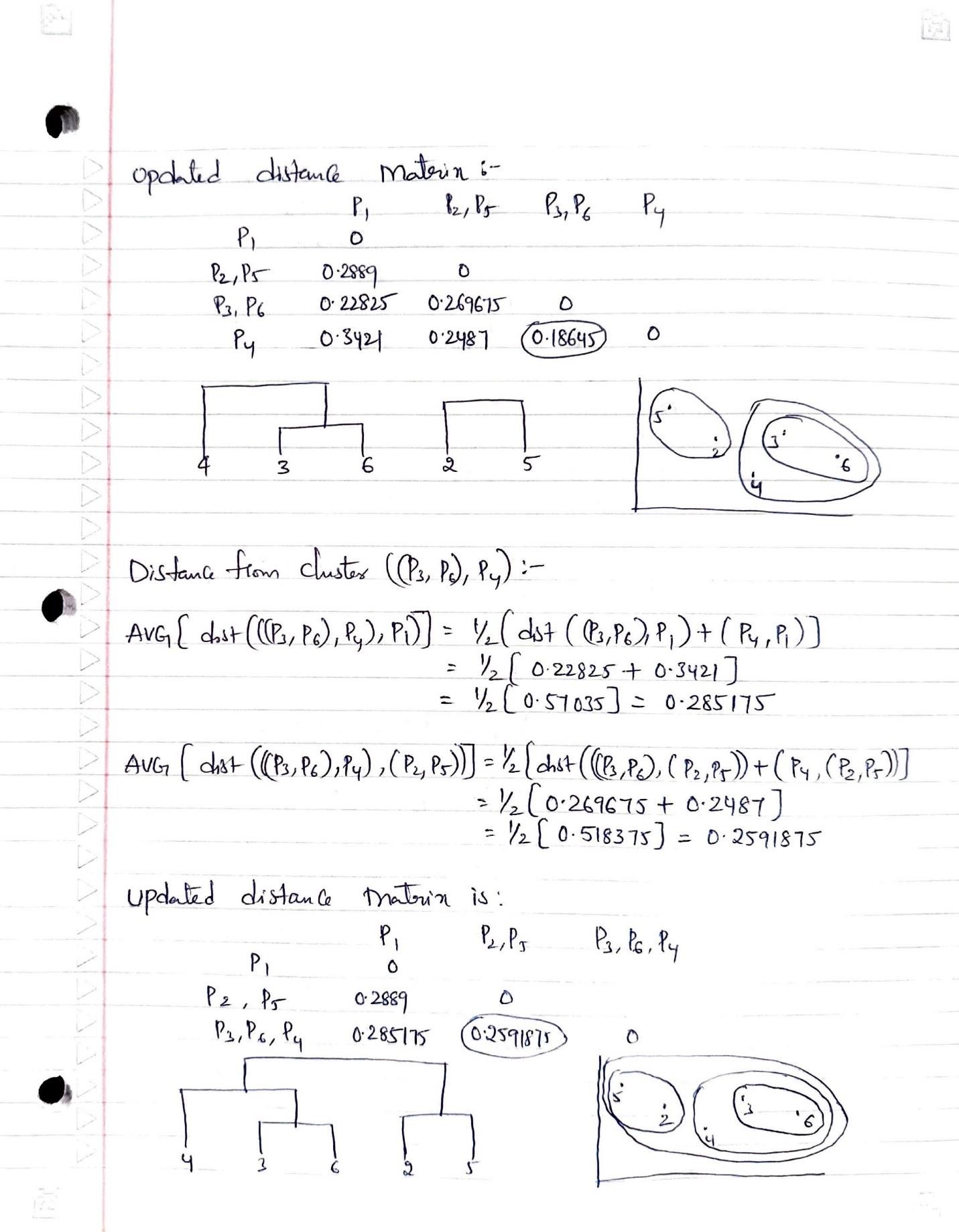


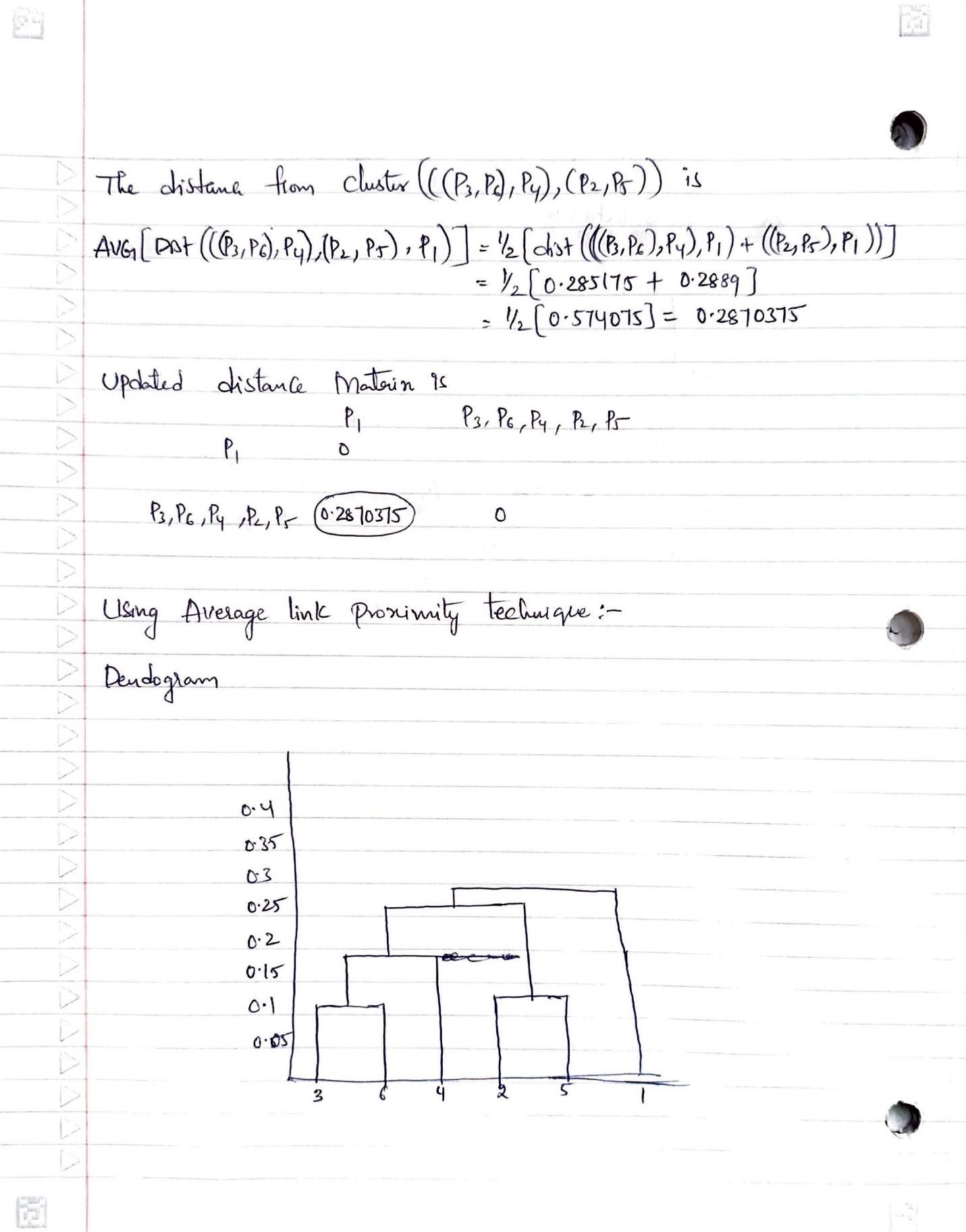








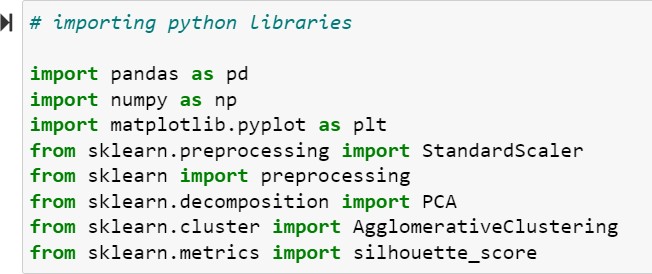




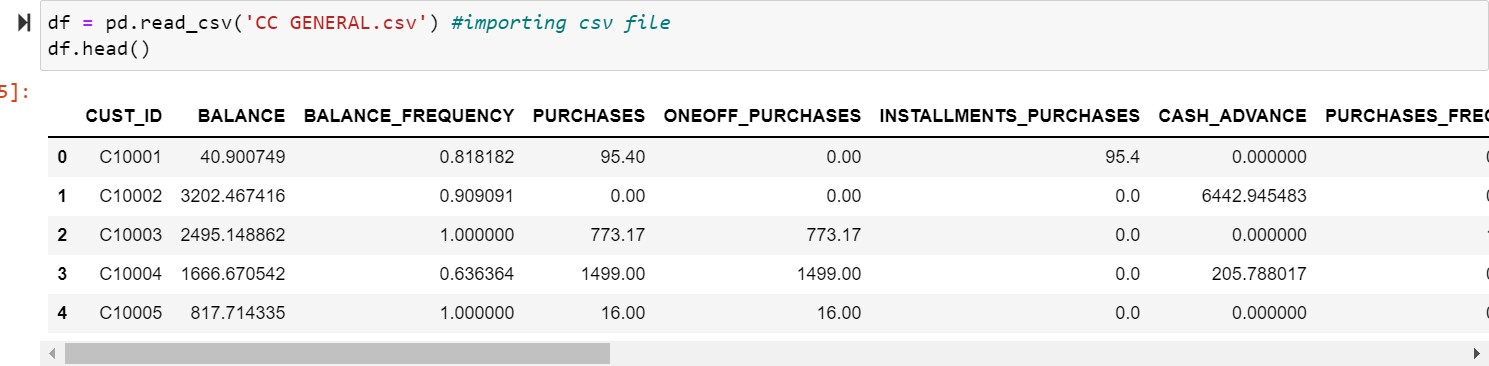


1. **Use CC\_GENERAL.csv given in the folder and apply:**

A few python libraries are imported

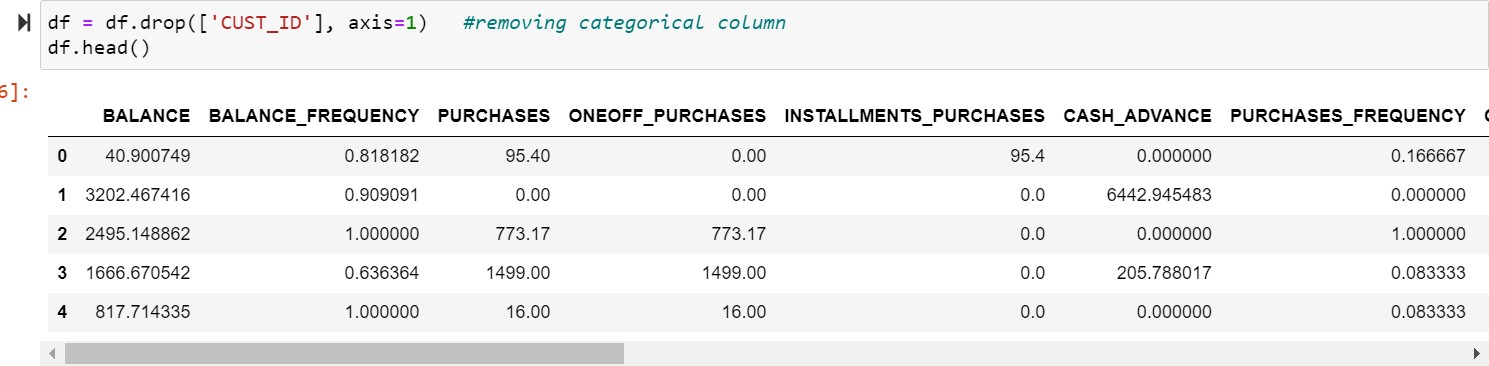


Using read\_csv method of pandas’ library imported cc general csv file



# Preprocess the data by removing the categorical column and filling in the missing values.

Using drop method removed a categorical column



To check null values, I used isnull method

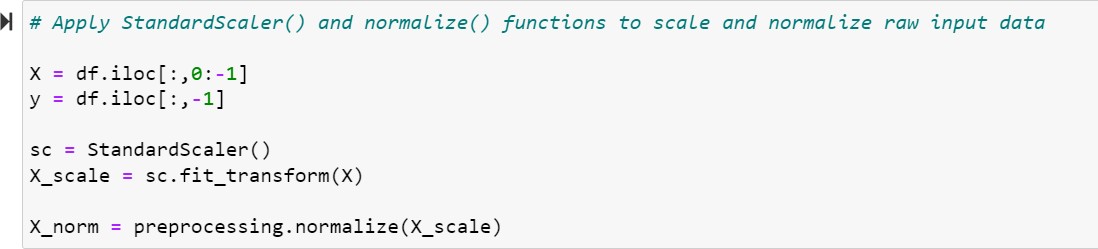


With mean of individual columns missing values are being replaced



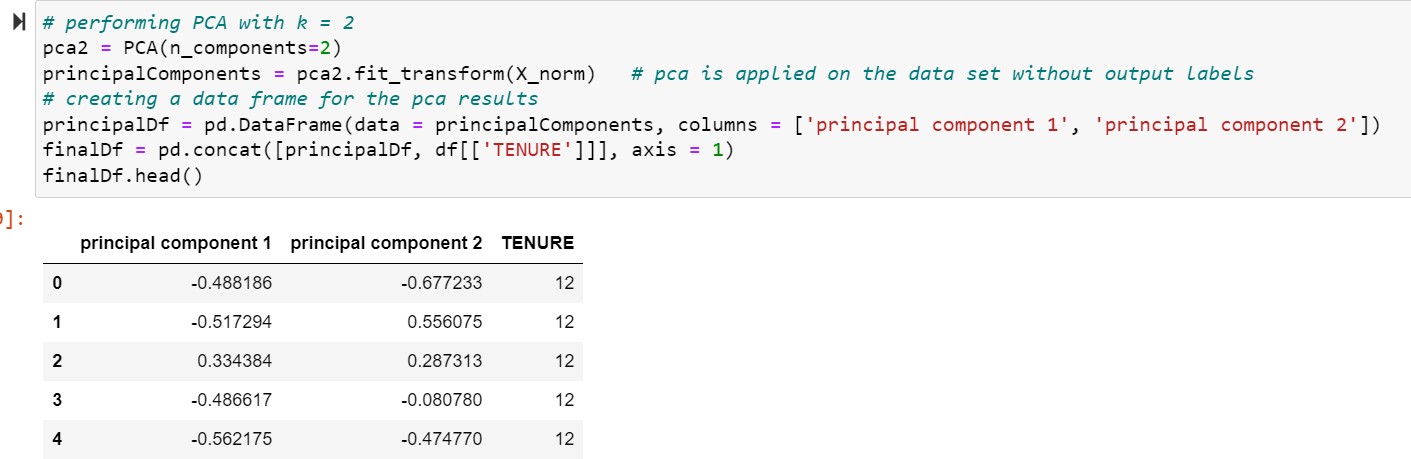
# Apply StandardScaler() and normalize() functions to scale and normalize raw input data.

Using StandardScalar and normalize methods we performed feature scaling on the data set. Feature scaling is used to normalize the range of all features.



# Use PCA with K=2 to reduce the input dimensions to two features.

From sklearn python library we imported PCA method to perform PCA on the data set. PCA results in a data frame with features having maximum variance with other features by ignoring the duplicate features. Here we reduced the dimensionality of data into two components by keeping k value is equal to 2.

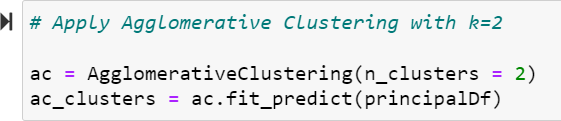


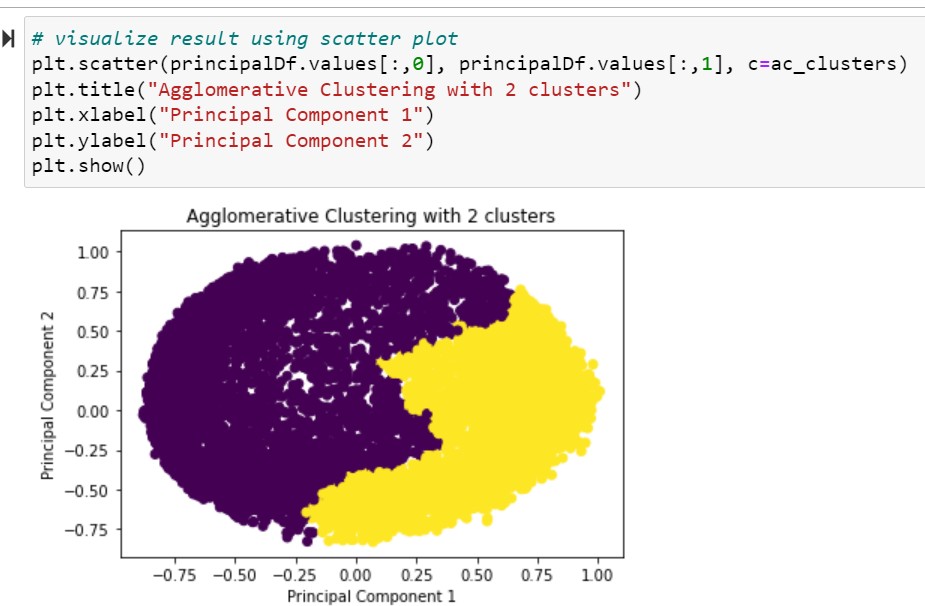
# Apply Agglomerative Clustering with k=2,3,4 and 5 on reduced features and visualize result for each k value using scatter plot.

Agglomerative clustering recursively combines two sample data clusters using distance.

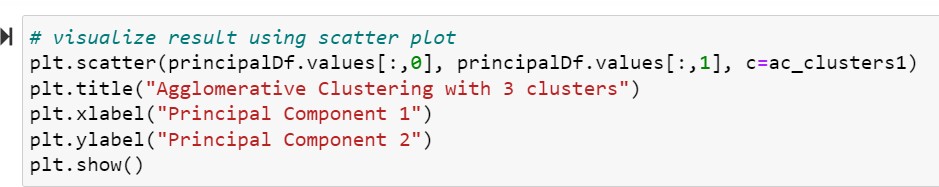
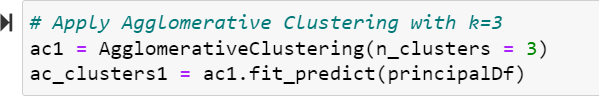
From sklearn python library we imported Agglomerative method to perform clustering on the preprocessed and PCA applied data set.

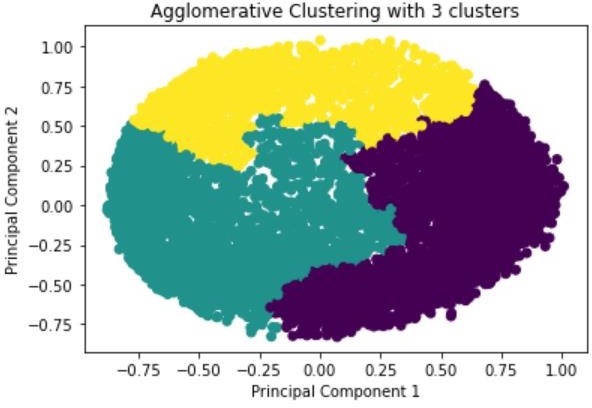
Using no of clusters is equal to 2 we mapped our data into 2 clusters.





Similarly, using no of clusters is equal to 3 we mapped our data into 3 clusters.

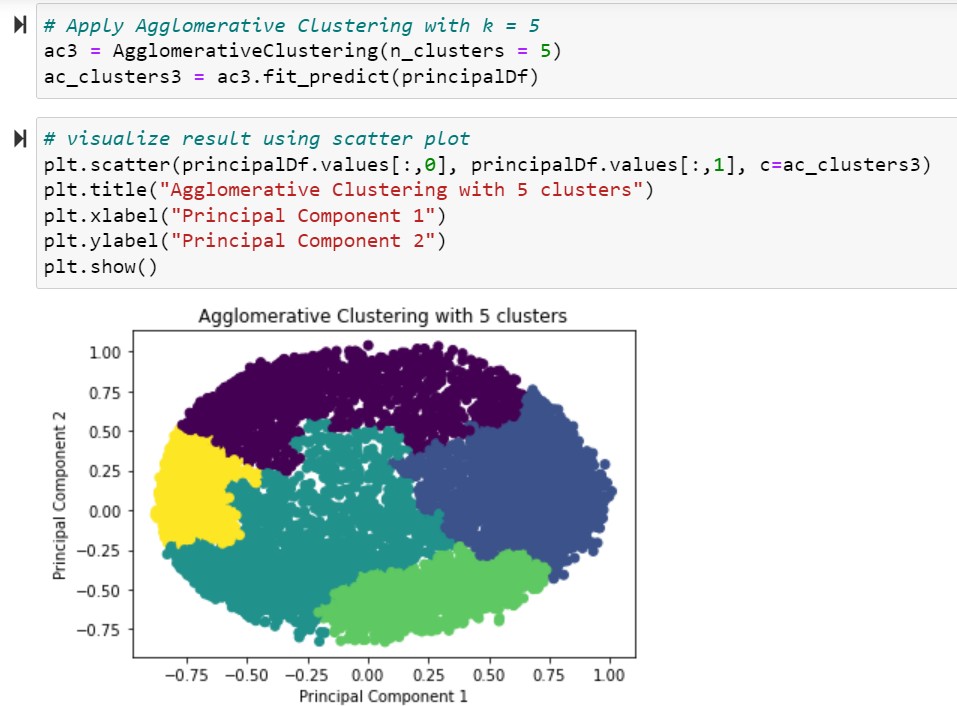




Similarly, using no of clusters is equal to 4 we mapped our data into 4 clusters.

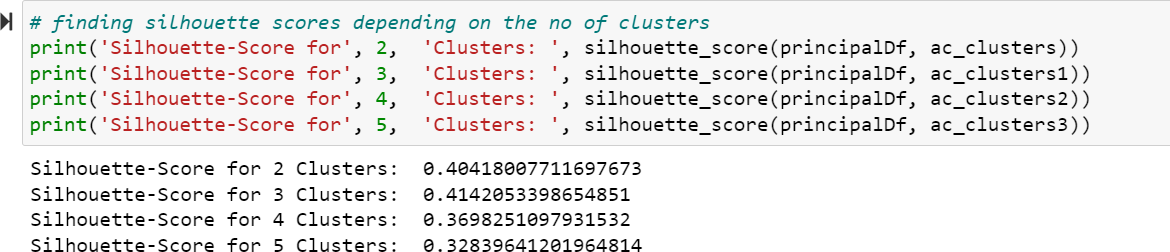


Similarly, using no of clusters is equal to 5 we mapped our data into 5 clusters.



# Evaluate different variations using Silhouette Scores and Visualize results with a bar chart.

With k = 2, 3, 4, and 5 we can train Agglomerative clustering algorithm on dataset and mapped the data into clusters depending on the no of clusters using Agglomerative clustering method of sklearn library. Depending on the number of clusters, we can find the silhouette score for each category. Silhouette score is used to calculate how many clusters required to be a good clustering technique. Silhouette score is the difference between the point and the nearest cluster that the point is not part of the cluster.



Taking no of clusters on the x- axis and silhouette scores on the y-axis we plot the bar graph using matplotlib library.

