Student Name: Yogesh Haresh Bojja.
Student Number: s3789918.

Step 1 - Folder **s3789918_BDP_A2/target** has "**s3789918_BDP_A2-0.0.1-SNAPSHOT.jar**"

Step 2 - Upload "**s3789918_BDP_A2-0.0.1-SNAPSHOT.jar**" to **Hue.**

Step 3 - Copy "s3789918_BDP_A2-0.0.1-SNAPSHOT" to EMR node by executing "**hadoop fs -copyToLocal <HDFS jar path> ~/**"

Step 4 - Create a folder with files **RMIT**, **MELBOURNE, 3littlepigs** and **NYTaxiLC1000** in it, this will be your Input folder.

> **<input_path>**: path of Input folder
> **<output_path>**: path of desired output folder.
> **<K>:** number of medoid.

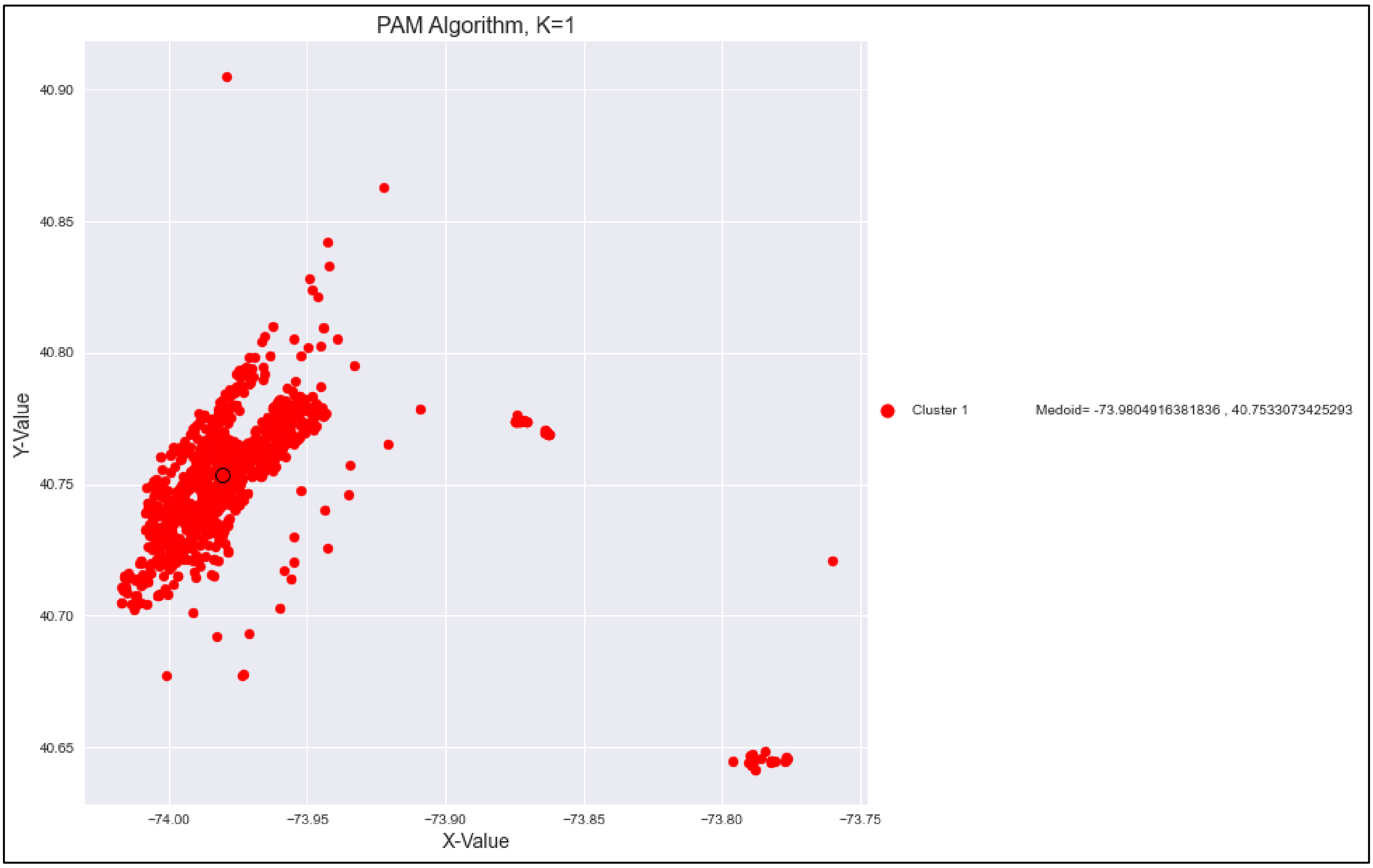
Task 1 Execution – hadoop jar s3789918_BDP_A2-0.0.1-SNAPSHOT.jar edu.rmit.cosc2367.s3789918_BDP_A2.Yogesh_Task1 <input_path> <output_path>

Task 2 Execution – hadoop jar s3789918_BDP_A2-0.0.1-SNAPSHOT.jar edu.rmit.cosc2367.s3789918_BDP_A2.Yogesh_Task2 <input_path> <output_path>

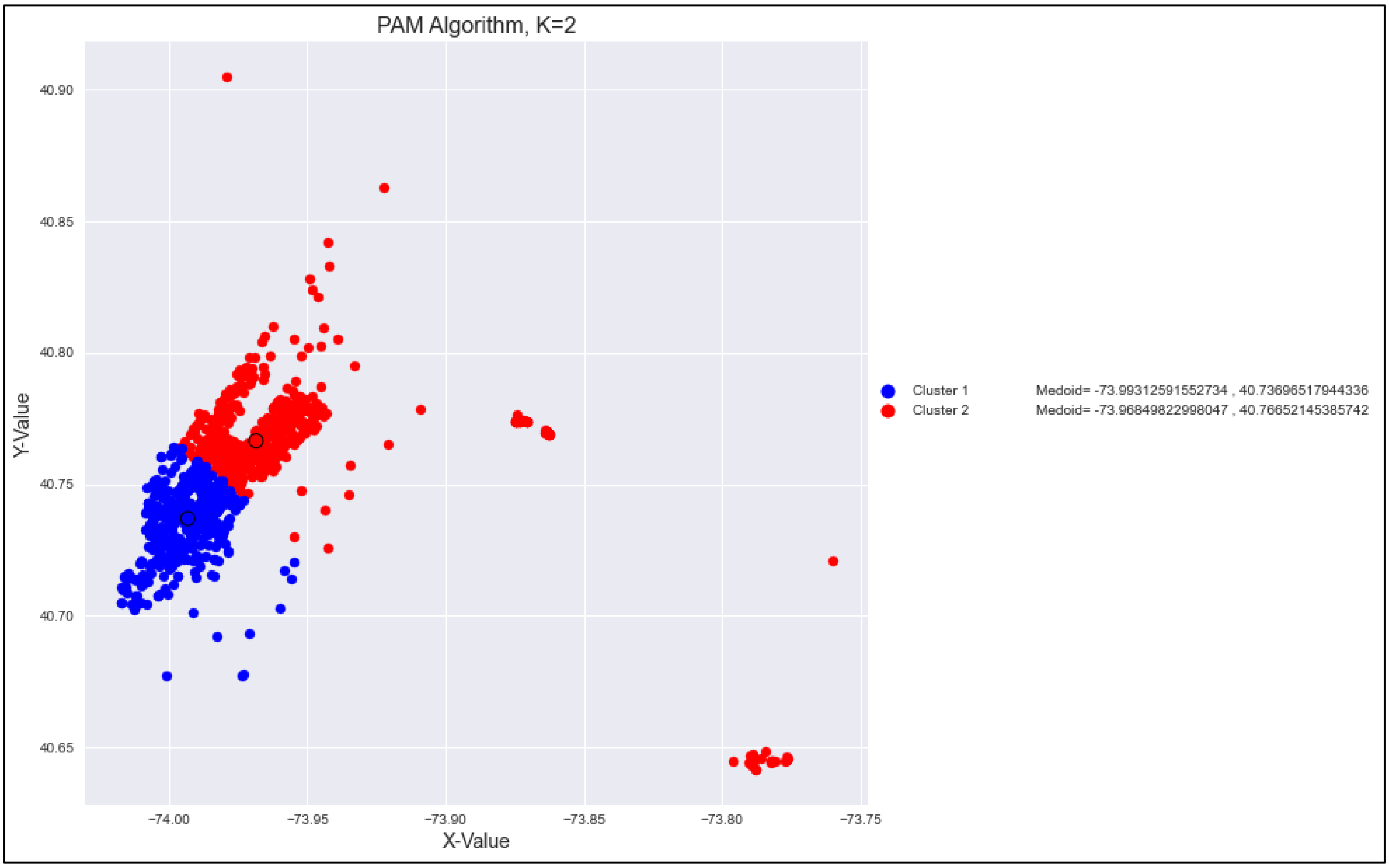
Task 3 Execution – hadoop jar s3789918_BDP_A2-0.0.1-SNAPSHOT.jar edu.rmit.cosc2367.s3789918_BDP_A2.Yogesh_Task3 <input_path> <K> <output_path>

Task 3 Visualization –

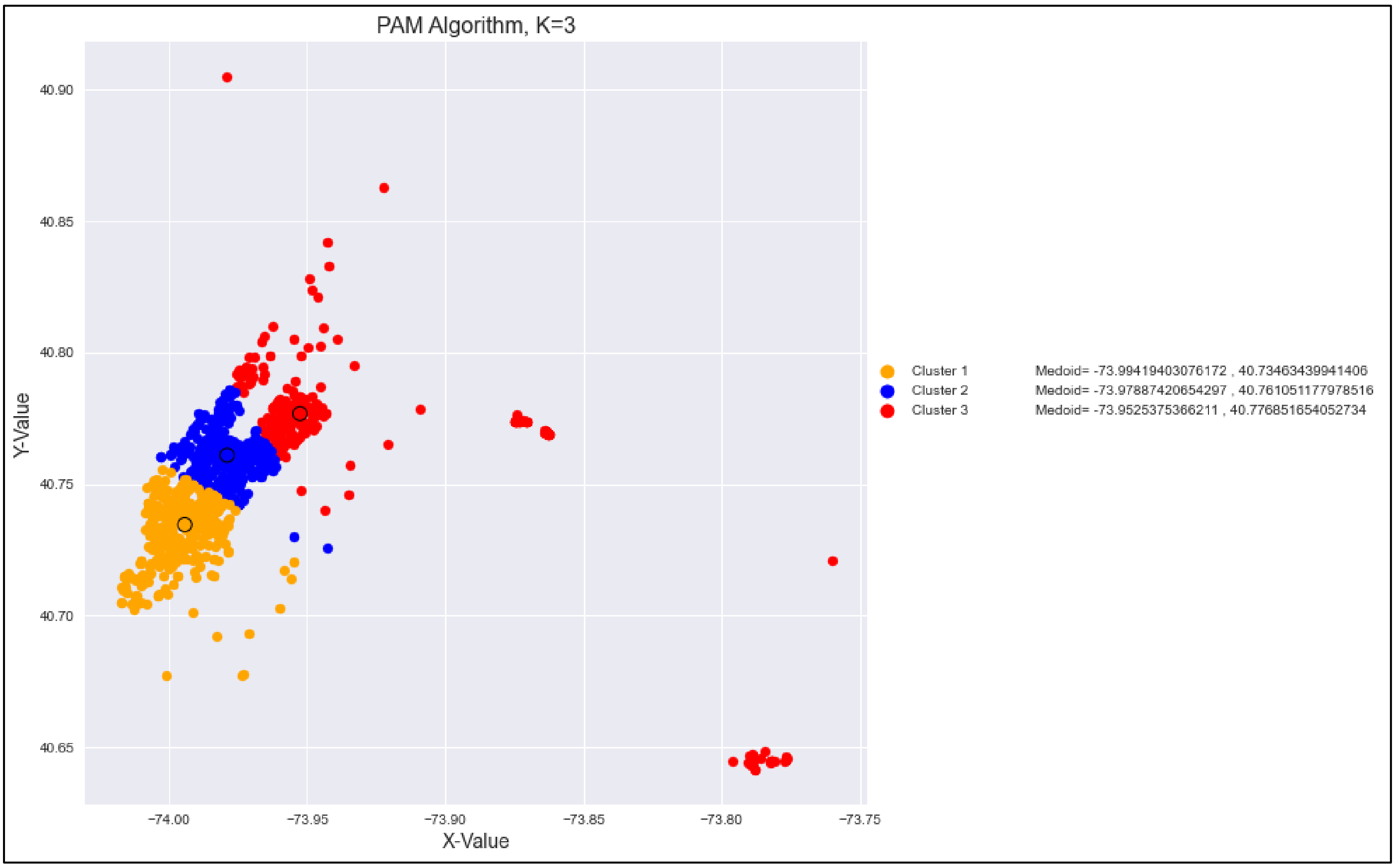
Visualization for k = 1:



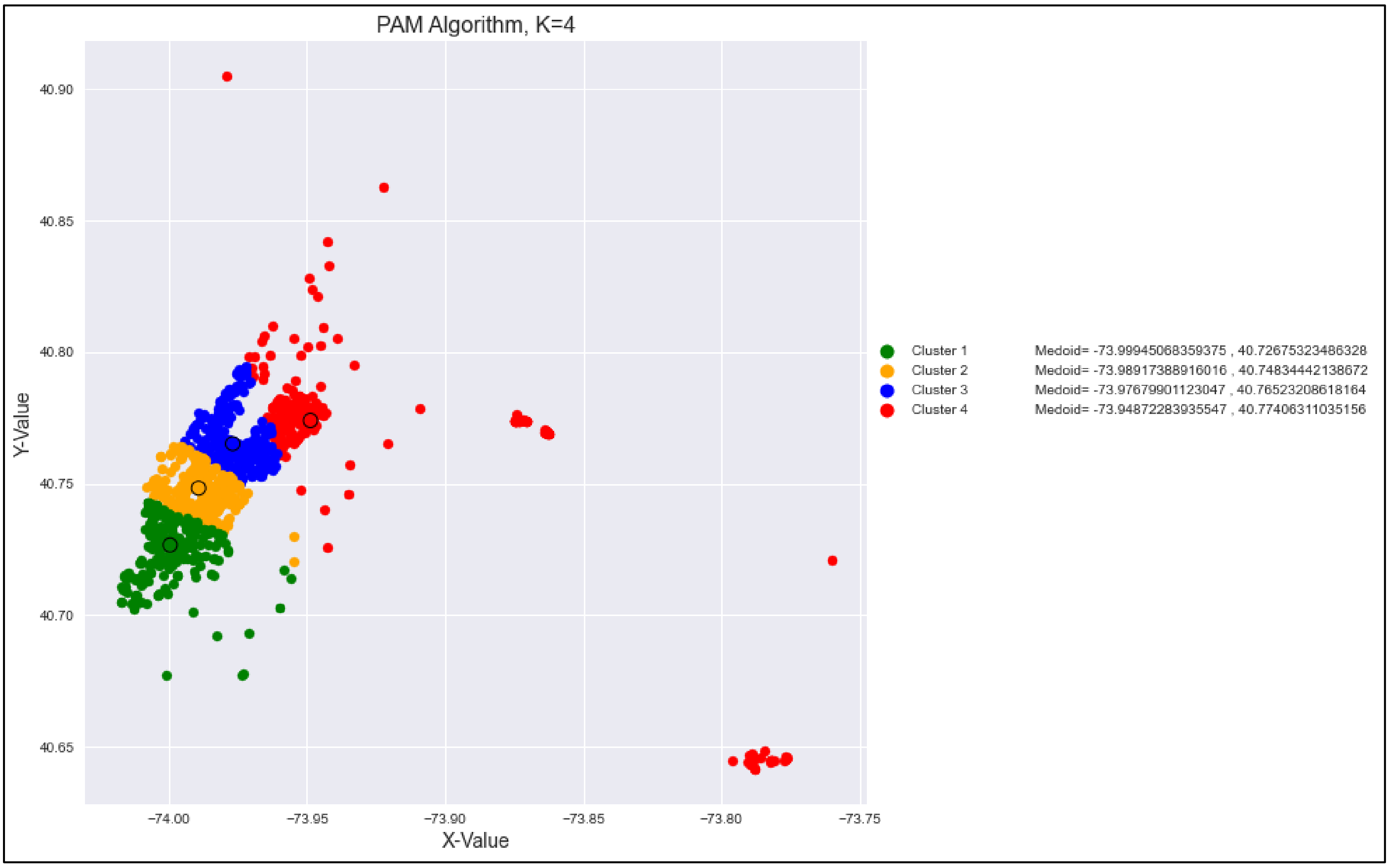
Visualization for k = 2:



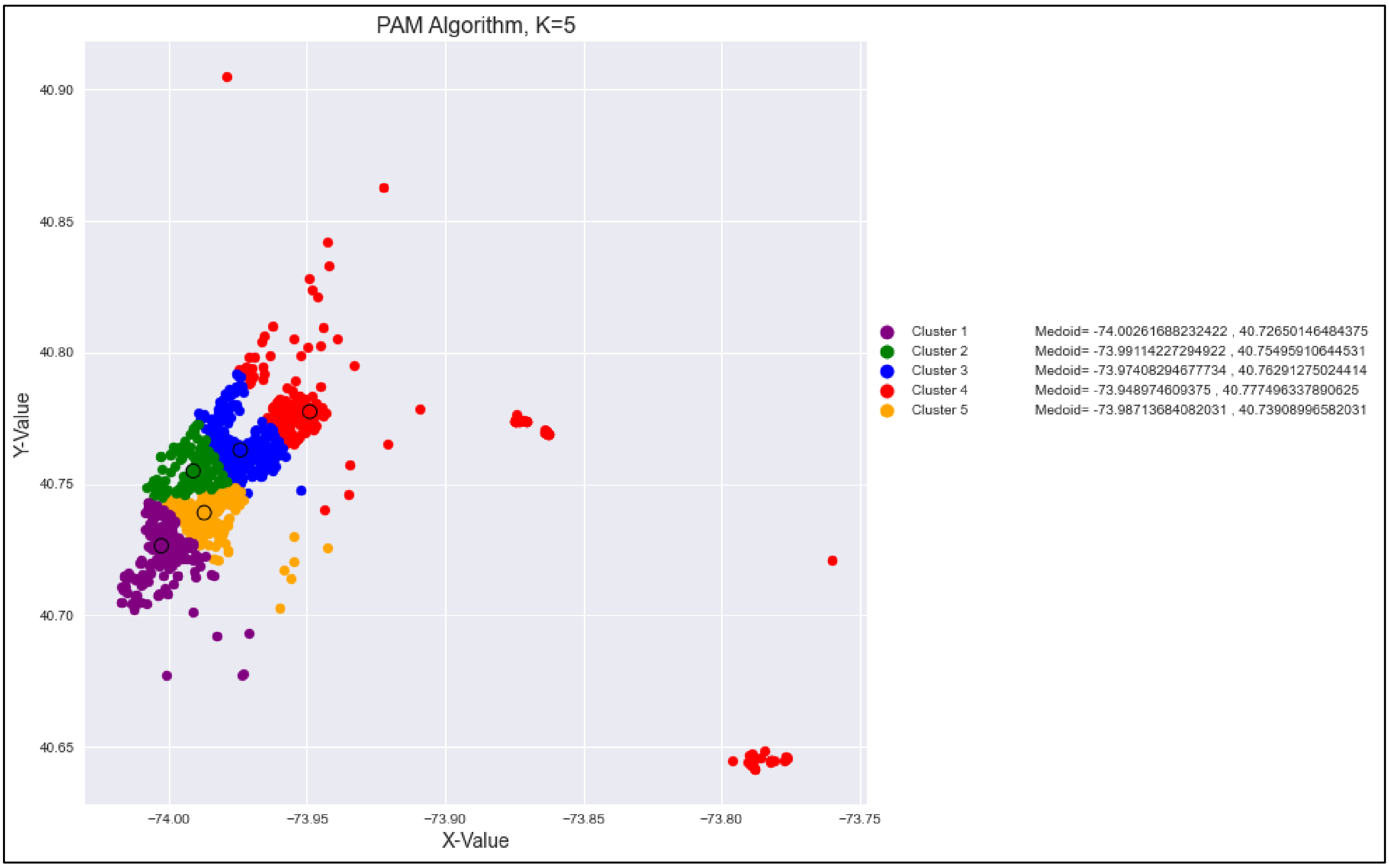
Visualization for k = 3:



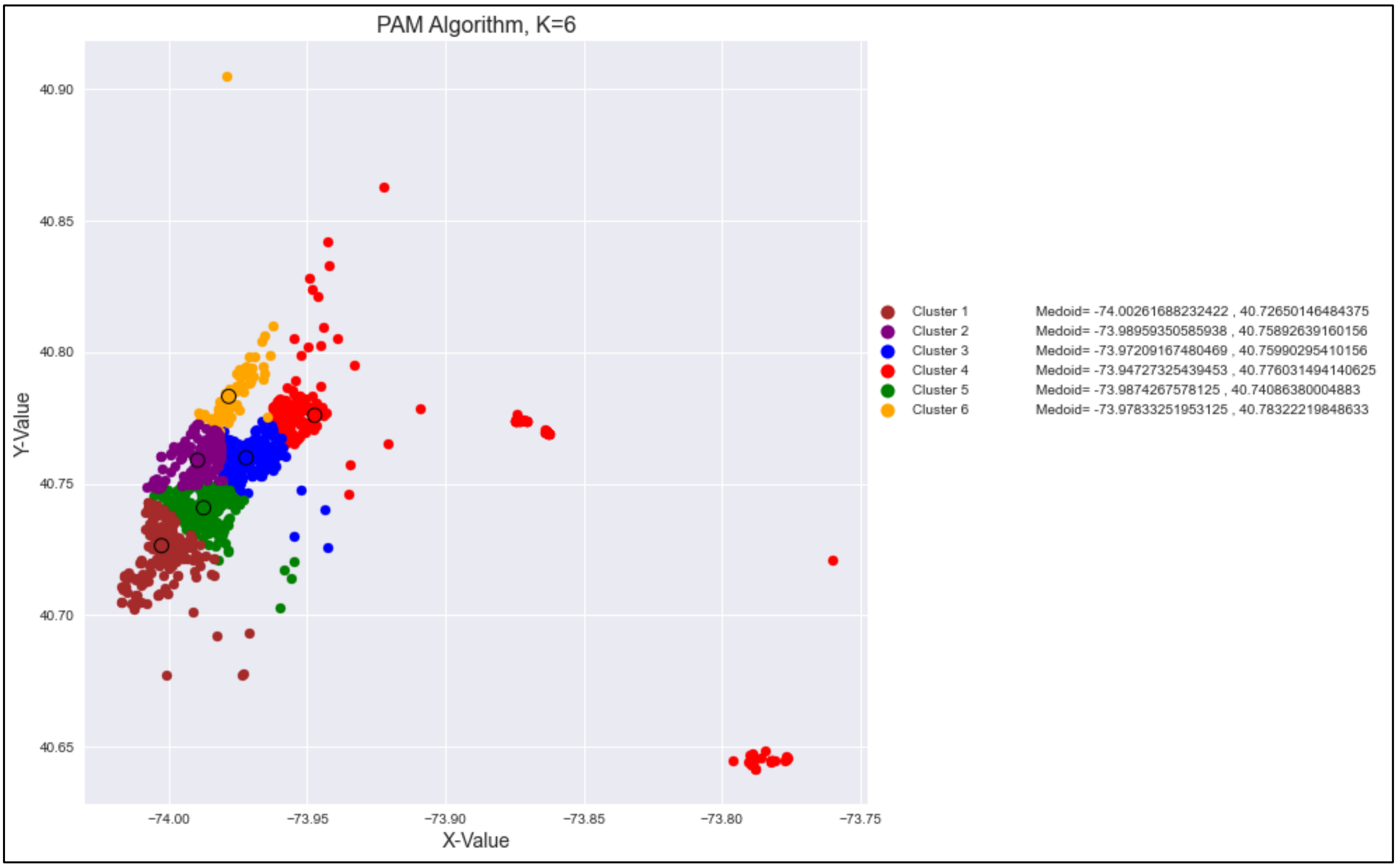
Visualization for k = 4:



Visualization for k = 5:



Visualization for k = 6:



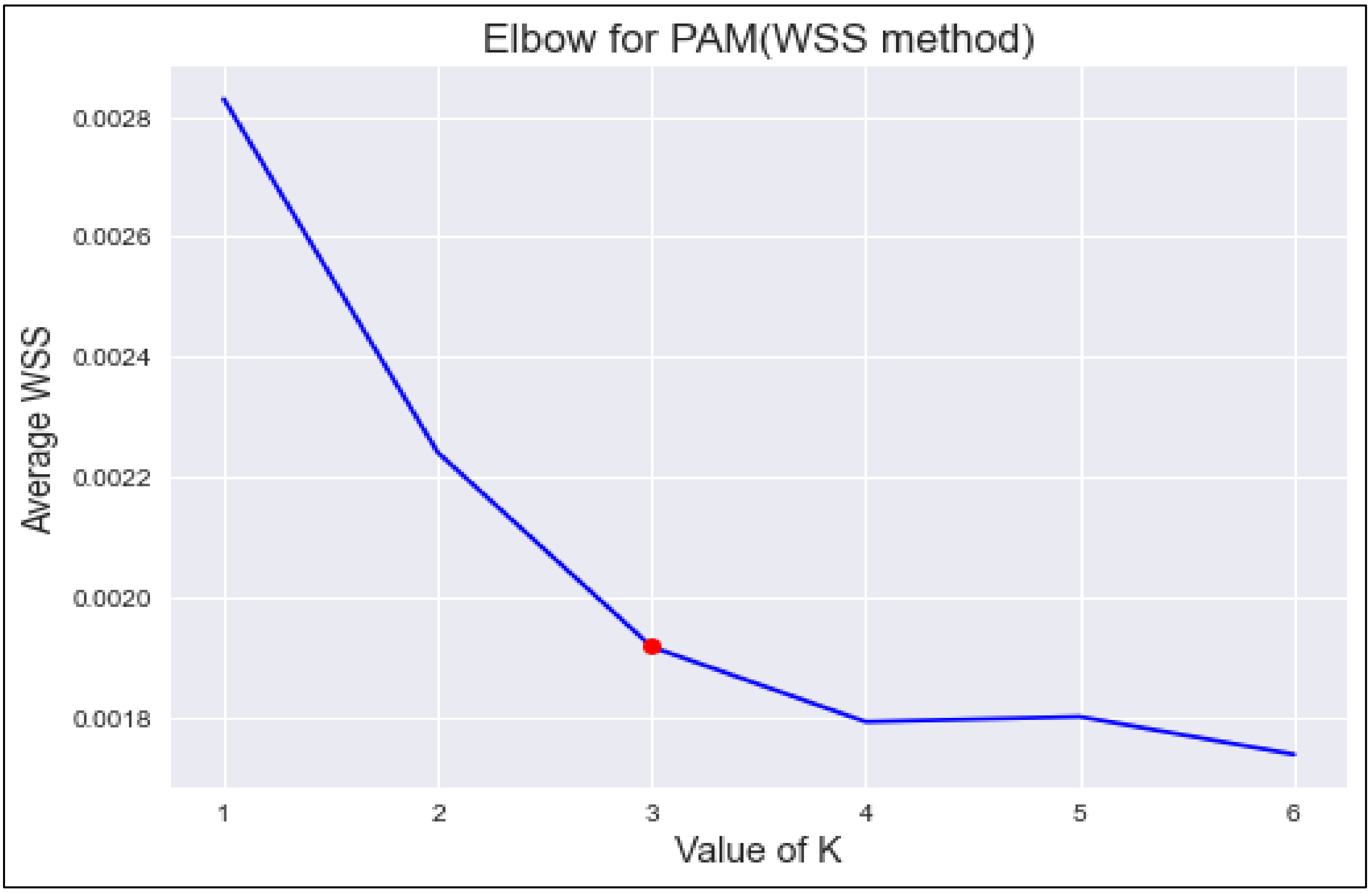
<u>Task 3 Analysis –</u>

Iteration analysis –

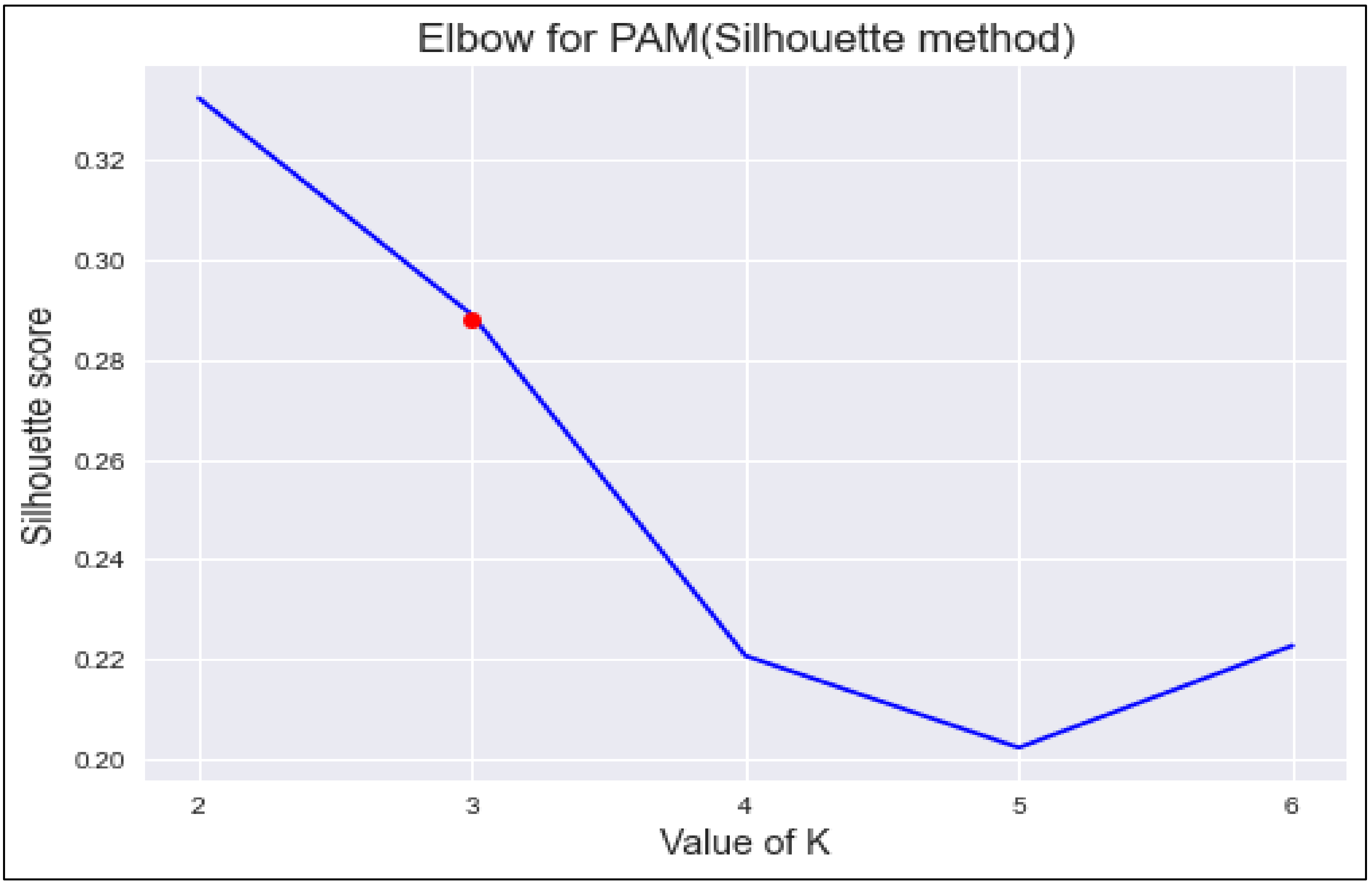
Value of K	Number of Iterations
3	10
4	6
5	8
6	6

The above table shows number of iterations taken for MR jobs to converge. Number of iterations seems to decrease as value of K increases but this is not necessarily true.

To choose right number of clusters we will perform Elbow method and Silhouette method. Below Image shows the graph of Elbow method. We use average WSS(Within sum of squares) as error metric. Average WSS is mean of distance of all the datapoints to its respective medoid. Analysing the graph below, value of k equal to 3 or 4 should remain ideal for PAM algorithm. After k=3 or 4, Average WSS is not decreasing rapidly or at huge rate, thus K=3/4 should be chosen.



In the below graph we have visualized results got from Silhouette method performed on each value of K. Silhouette tells us how similar the point is to its own cluster as compared to the other/neighbouring clusters. Higher the value of Silhouette's coefficient better is the clustering. K=2 and K=3 has high Silhouette coefficient.



Conclusion: Elbow method suggests the K value to be 3 or 4 and Silhouettes method suggests the K value to be 2 or 3, Therefore, considering both the results we choose value of K = 3 which falls common in both the analysis.