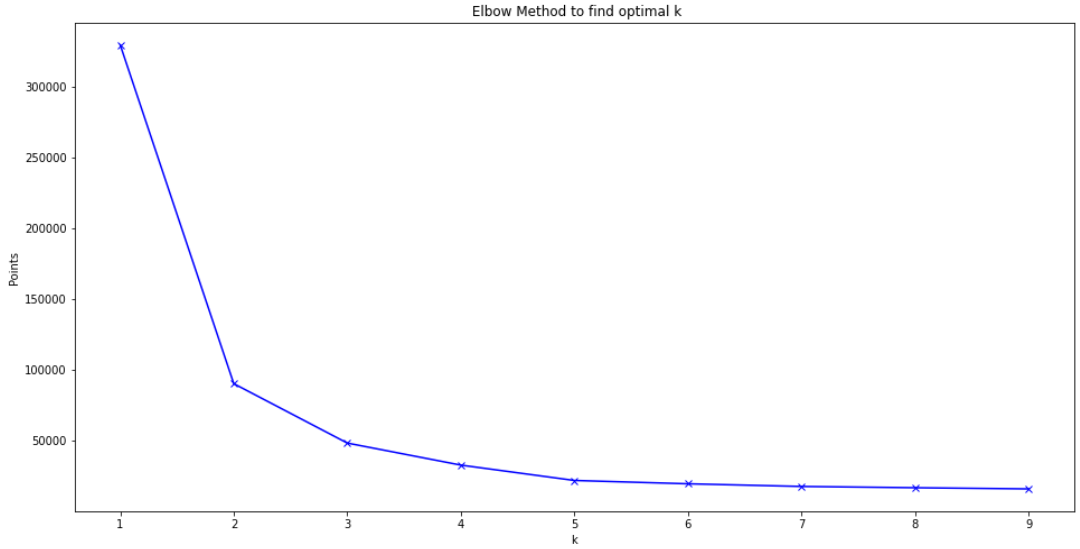
Report(Assignment 3)

Q1:

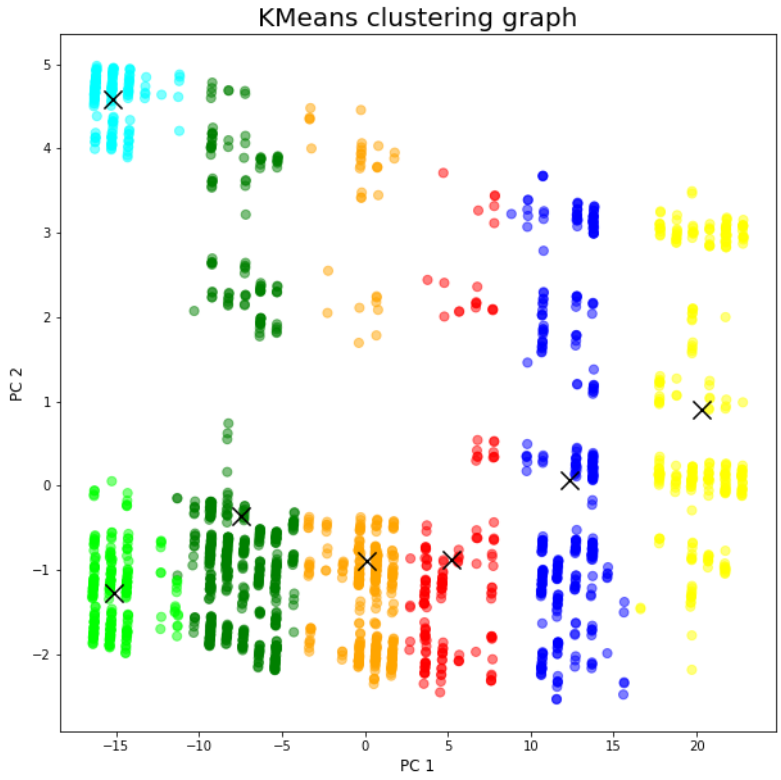
Hypothesis: Categorical data is being converted into numerical data
Duplicate data is being dropped assuming it is not useful
Dimensionality reduction is also done on data to plot in 2d

Kmeans Clustering:

Elbow graph gives us the optimal value of k



We choose k=7 which is same as no of labels



Below are the 7 coordinates of centroid

5.19602985, -0.87961759
-7.49513969, -0.36778561
12.3083726 , 0.06186192
-15.20524395, 4.57871685
20.29885041 , 0.89872538
-15.16118422 , -1.27644496
0.11770502 , -0.89097185

Accuracy of the model:

0.5314872711031711

Q3:

True label count (label no,count value)

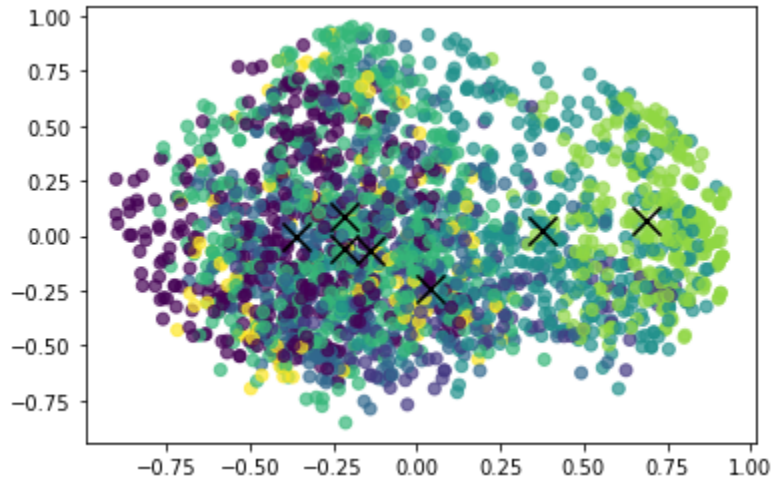
(3, 38), (1, 781), (6, 233), (4, 135), (5, 149), (2, 197), (0, 706)

Predicted label count (label no,count value)

2, 280), (1, 1553), (6, 163), (0, 243)

Birch Clustering:

Graph when no of cluster =7:



Coordinates of centroid are :
0.68494156, 0.06476856
-0.13950582, -0.0706801
-0.2147147, 0.08327481
0.37597863, 0.01924579
-0.35994927, -0.00490497
0.04114547, -0.2432552
-0.21809382, -0.06236326

Accuracy of Birch clustering:
0.47967842786958464

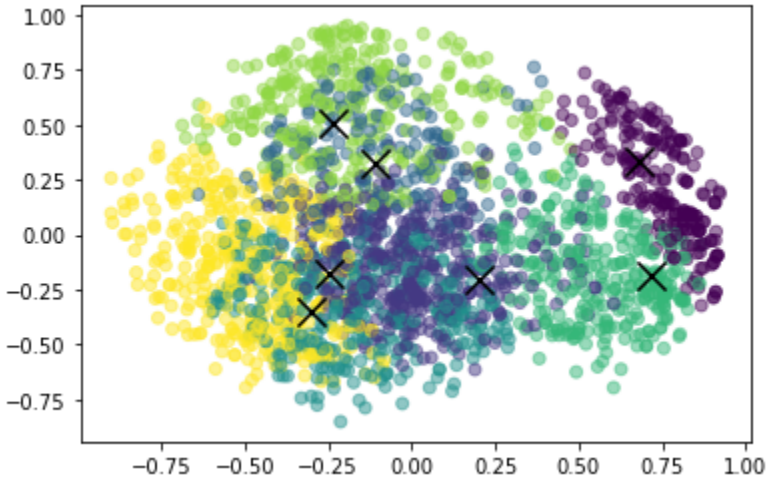
Q3:
True label count (label no,count value)
(3, 38), (1, 781), (6, 233), (4, 135), (5, 149), (2, 197), (0, 706)

predicted label count (label no,count value)

(2, 280), (1, 1433), (0, 401), (4, 125)

Guassian Mixture:

Graph for number of cluster =7



Here we have reduced dataset into 3 components using pca thus have each coordinate have 3 axes

Coordinates of centroid are :
0.66569938 , 0.01396668, 0.01230368
-0.13721734, -0.29237947, -0.21589035
-0.09231701, -0.31254835, 0.41700747
-0.16259177, 0.70616846, -0.14158195
-0.52682399, 0.03730288, -0.22055236
0.01403345, 0.07407912, -0.39089778
-0.11479241, 0.37549658, 0.57124209

Accuracy of Gaussian Mixture clustering:
0.4242965609647164

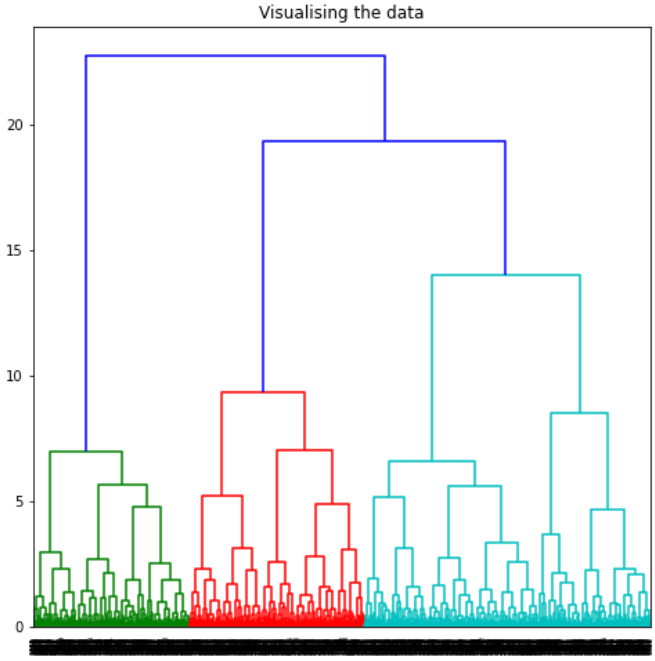
Q3:

True label count (label no,count value)
(3, 38), (1, 781), (6, 233), (4, 135), (5, 149), (2, 197), (0, 706)

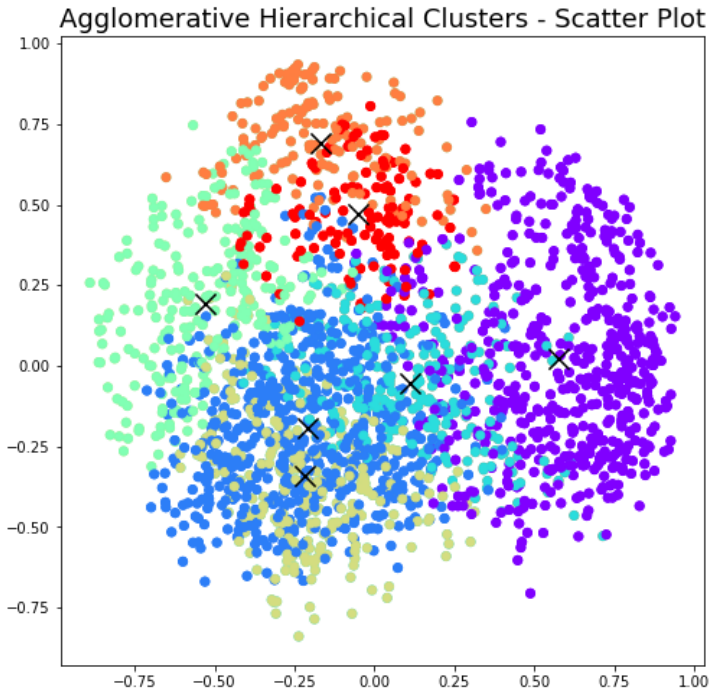
predicted label count (label no,count value)
(2, 280), (1, 1433), (0, 401), (4, 125)

Hierarchial clustering : Agglomerative clustering

Dendogram Graph:



Graph plotting :



Coordinates of centroid are :

0.5764235 , 0.02099307, -0.08420748
-0.21031729, -0.19309133, -0.28323077
0.11152992, -0.05240793, 0.41716702
-0.05172056, 0.47001617, 0.47032419
-0.53160811, 0.19467304, -0.09955337
-0.16773611, 0.69237907, -0.12287428
-0.21932041, -0.33972807, 0.41939733

Accuracy of Hierarchial clustering : Agglomerative clustering:
0.44707458686913804

Q3

True label count (label no,count value)

(3, 38), (1, 781), (6, 233), (4, 135), (5, 149), (2, 197), (0, 706)

predicted label count (label no,count value)

(2, 568), (1, 787), (6, 373), (0, 511)

Comparing accuracy of 4 models:

CLUSTERING	ACCURACY
Kmeans:	0.5314872711031711
Birch Clustering:	0.47967842786958464
Hierarchical clustering : Agglomerative clustering	0.44707458686913804
Gaussian Mixture:	0.4242965609647164

Observation: Kmeans is performing better than others as its accuracy is higher and giving better results after that Birch is performing better than hierarchical and gaussian.
Gaussian performing lower than other 3 clustering algorithms on the given data.

Q2:
Hypothesis: One hot encoding is used on data to convert categorical data
Duplicate data is being dropped

Approach:

For testing on the train data, it is split into 80:20 ratio

Kmeans is used for the data the results of the **grid search** on number of clusters and algorithms is as follows:

auto 7 0.45982142857142855
auto 9 0.45982142857142855
auto 11 0.45982142857142855
auto 13 0.45982142857142855
auto 15 0.45089285714285715
auto 17 0.46651785714285715
auto 19 0.43973214285714285
auto 21 0.453125
auto 23 0.484375
auto 25 0.44866071428571436
auto 27 0.46875
auto 29 0.45982142857142855
auto 31 0.48214285714285715
auto 33 0.4732142857142857
auto 35 0.48660714285714285
auto 37 0.45535714285714285

auto 39 0.44642857142857145
auto 41 0.46651785714285715
auto 43 0.43080357142857145
auto 45 0.4419642857142857
auto 47 0.47098214285714285
auto 49 0.45089285714285715
full 7 0.45982142857142855
full 9 0.45982142857142855
full 11 0.45982142857142855
full 13 0.45982142857142855
full 15 0.453125
full 17 0.46651785714285715
full 19 0.44642857142857145
full 21 0.453125
full 23 0.45535714285714285
full 25 0.43080357142857145
full 27 0.43080357142857145
full 29 0.44866071428571436
full 31 0.4642857142857143
full 33 0.45982142857142855
full 35 0.42410714285714285
full 37 0.46875
full 39 0.46651785714285715
full 41 0.4799107142857143
full 43 0.43080357142857145
full 45 0.45089285714285715
full 47 0.47767857142857145
full 49 0.45089285714285715
elkan 7 0.45982142857142855
elkan 9 0.45982142857142855
elkan 11 0.45982142857142855
elkan 13 0.45982142857142855
elkan 15 0.45089285714285715
elkan 17 0.46651785714285715
elkan 19 0.43973214285714285
elkan 21 0.453125
elkan 23 0.484375
elkan 25 0.44866071428571436
elkan 27 0.46875
elkan 29 0.45982142857142855
elkan 31 0.48214285714285715
elkan 33 0.4732142857142857
elkan 35 0.48660714285714285
elkan 37 0.45535714285714285

elkan 39 0.44642857142857145
elkan 41 0.46651785714285715
elkan 43 0.43080357142857145
elkan 45 0.4419642857142857
elkan 47 0.47098214285714285
elkan 49 0.45089285714285715

Gaussian is also applied with 7 clusters but was giving F1 score : 0.3794642857142857

So, After analyzing all things it came out that the best model is Kmeans with 35 clusters , algorithm: auto or elkan gives results on the training data as:
F1 Score as :0.48660714285714285