ASSIGNMENT 1

Task 1:

Clustering on the given dataset using k-means algorithm was performed with the following specs:

Initialization: Random

Distance measure: Euclidean Distance

Evaluation metrics: Rand Index, SSE (same as WCSS)

Stopping Criteria:200 iterations

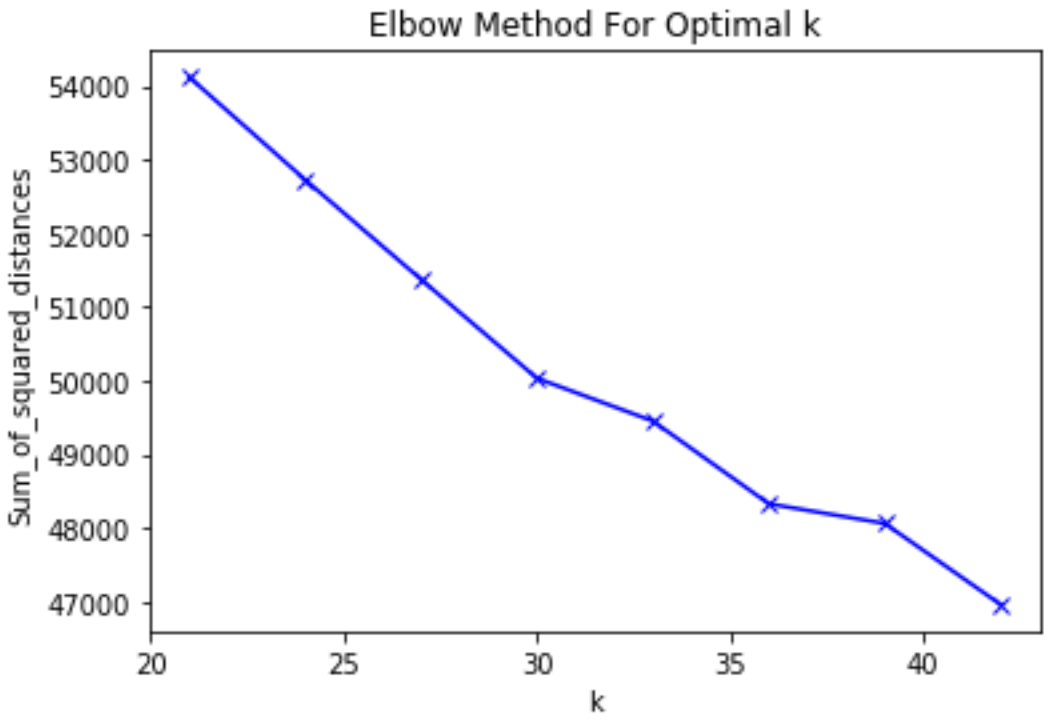
The Table with the Number of Clusters and Rand Index and SSE is stated below.

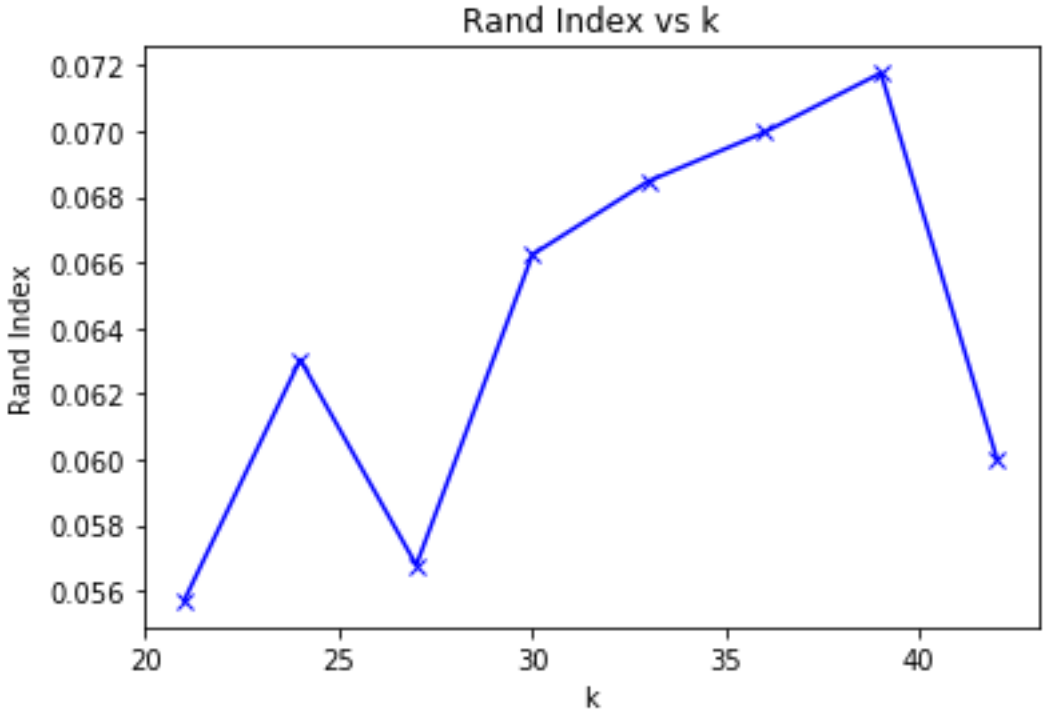
The Plot of the Rand Index and SSE with the 8 different k values in the range of 21,42 is given below.

Label Encoder used to convert categorical target data to numeric data for getting true labels for sklearn adjusted_rand_score.

Table1: K-means clustering.

Number of Clusters	Rand Index	SSE
K=21	0.05569730810355433	54122.235650559785
K=24	0.06304249340247113	52728.939604744955
K=27	0.056782340400405806	51375.37420744925
K=30	0.06624218959318448	50033.744372261346
K=33	0.06848011201584682	49452.89059049289
K=36	0.06997145983180315	48329.29386225876
K=39	0.07177266227576506	48063.27923723095
K=42	0.06001619608776324	46961.146664791835


Elbow Method For Optimal k
Sum_of_squared_distances
54000
53000
52000
51000
50000
49000
48000
47000
20
25
30
35
40
k


Rand Index vs k
Rand Index
0.072
0.070
0.068
0.066
0.064
0.062
0.060
0.058
0.056
20
25
30
35
40
k

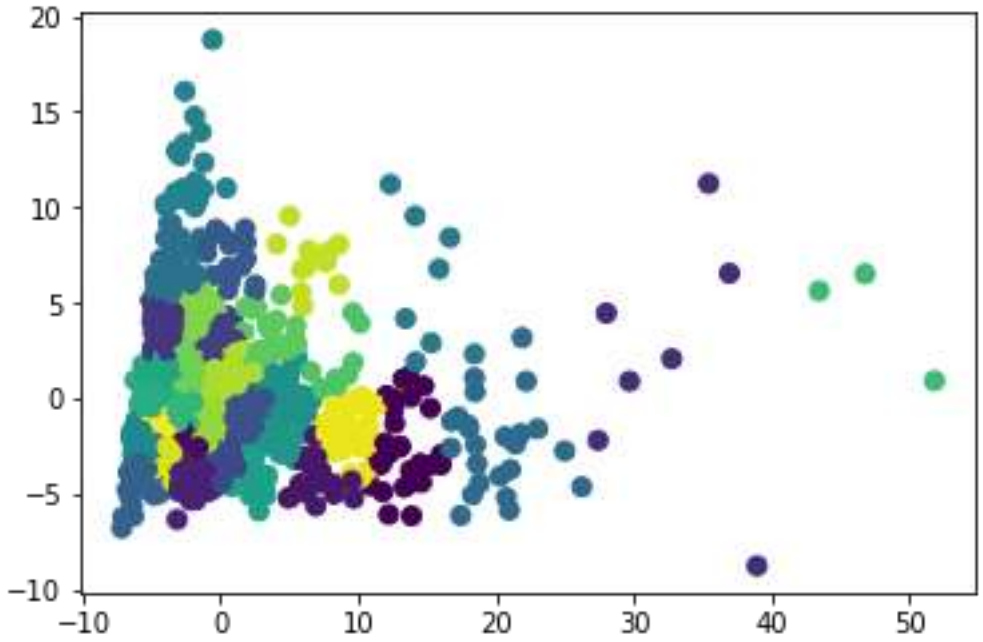
From the given elbow curve, it can be inferenced that the optimal number of clusters i.e. choice of k should be 33 as there is a sharp bend after the point representing k as 33 and straight line is observed for points post it.

Task 2:

PCA is used for dimensionality reduction and visualization. The optimal k value is used for the below:

a) Project your data into 2D plane using PCA and color code the clusters. Use

scatter plot.



b) Choose the following number of components for PCA: 2, 4, 8, 16, 32. Do the clustering on the reduced dataset and compute the Rand index.

Table 2: K-means clustering after dimensionality reduction

Number of components in PCA	Rand Index (using chosen k)
N=2	0.023896393688517988
N=4	0.04191059635677282
N=8	0.055996705690040595
N=16	0.06252004475915843
N=32	0.06434305344523152