Report - Assignment 4 - Clustering

(Using "`clusteringdata.csv`" for this analysis)

Task 1: K-Means Clustering

Loading Dataset:

```
1 df_cd = pd.read_csv("clusteringdata.csv")
2 df_cd.head()
```

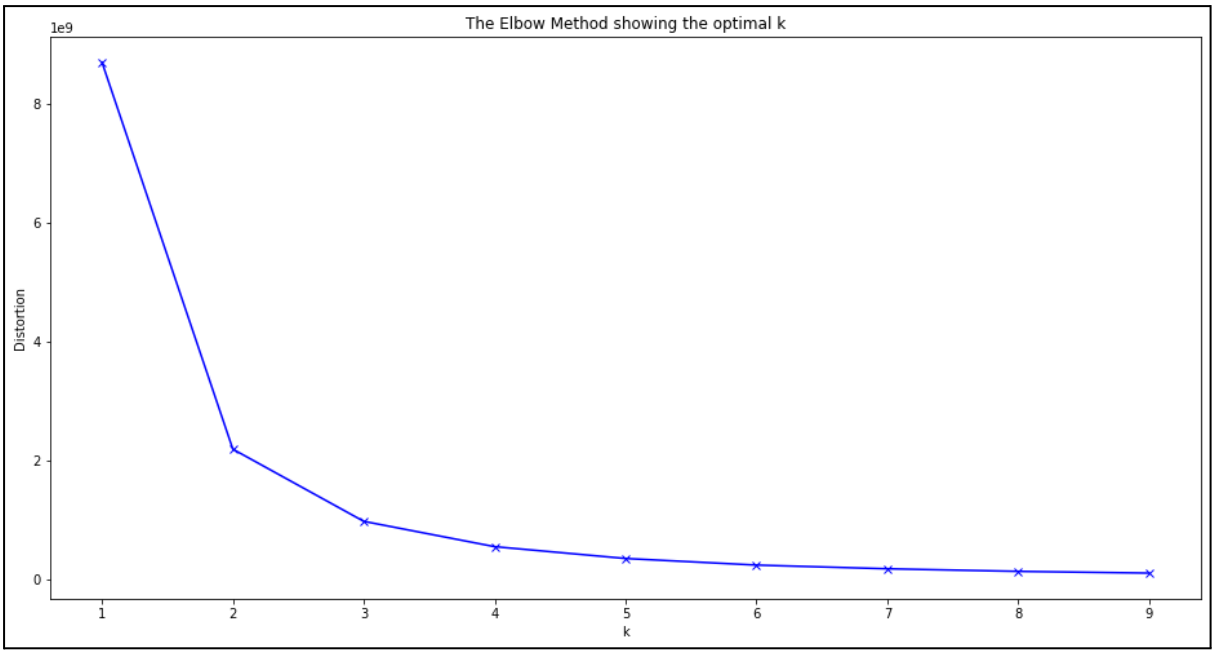
	Age	WorkClass	Fnlwght	Education	EducationNumber	MaritalStatus	Occupation	Relationship	Race	Sex	CapitalGain
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0

Label Encoding the string columns to apply k-mean clustering:

	Age	WorkClass	Fnlwght	Education	EducationNumber	MaritalStatus	Occupation	Relationship	Race	Sex	CapitalGain
0	22	6	576	9	12	4	1	1	4	1	16
1	33	5	633	9	12	2	4	0	4	1	0
2	21	3	3093	11	8	0	6	1	4	1	0
3	36	3	3332	1	6	2	6	0	2	1	0
4	11	3	4145	9	12	2	10	5	2	0	0
...	...	...	...	...	...	...	...	...	...	...	...
4995	26	3	3193	4	2	4	7	4	4	0	0
4996	14	3	3604	11	8	2	3	0	4	1	0
4997	30	4	3044	11	8	2	3	0	4	1	0
4998	9	3	3482	11	8	4	1	1	4	1	0
4999	41	2	393	15	9	4	10	1	4	0	0

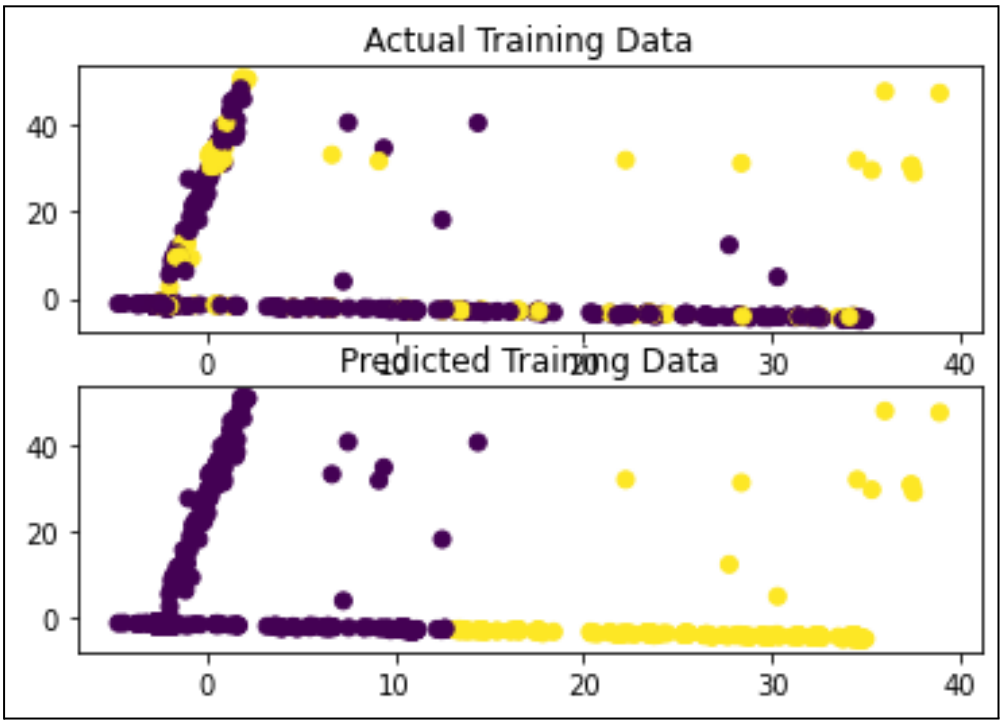
5000 rows × 15 columns

Elbow Method to get optimal value of K:

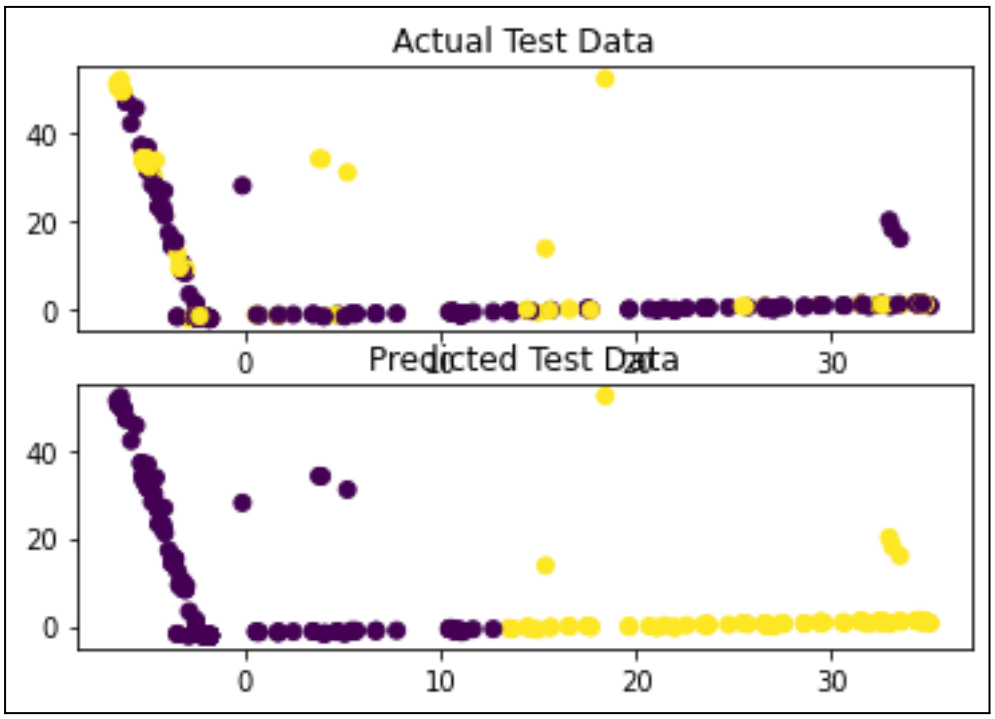


From the figure, we get the elbow at k=2, So the optimal value of k is 2 in k-means clustering.

Visualizing the k-means clustering for training data:



Visualizing the k-means clustering for Test data:



Confusion Matrix for Test data prediction:

```
[[1078   69]
 [ 322   31]]
```

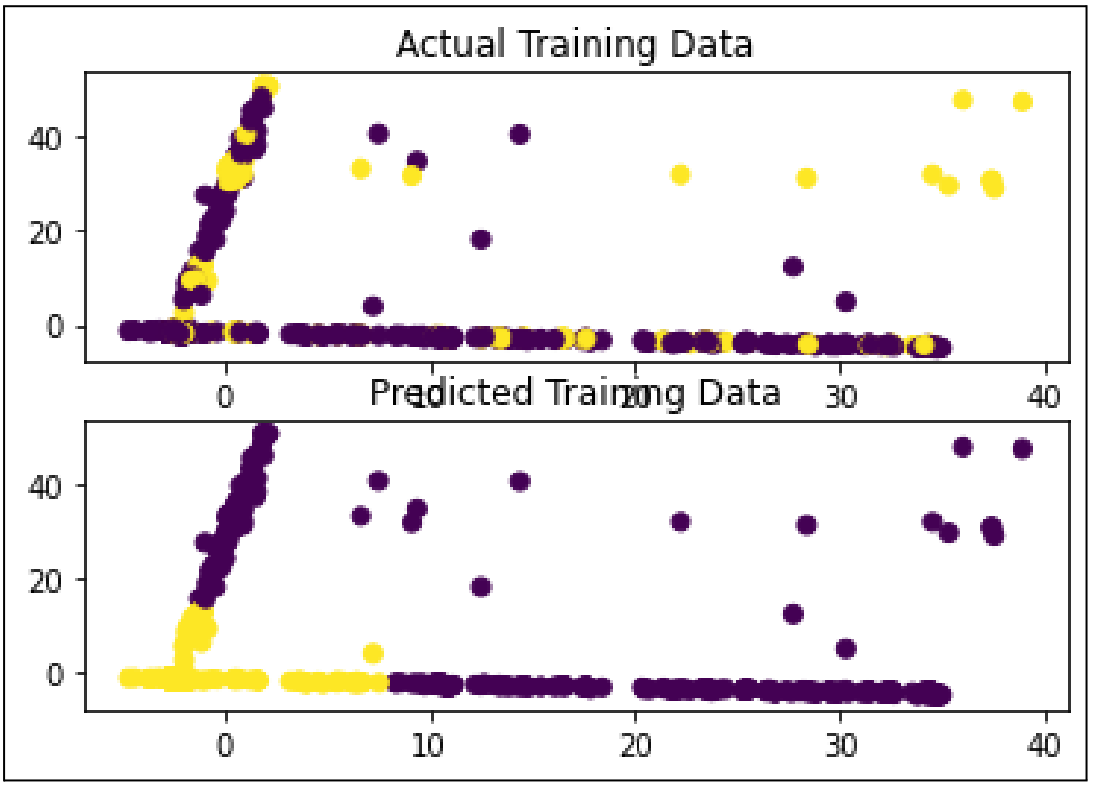
Task 2: Hierarchical Agglomerative Clustering

Finding the best Hierarchical Agglomerative Clustering Model

```
F1-score for complete linkage + cosine 0.36749297214413496
F1-score for complete linkage + euclidean 0.03551609322974472
F1-score for complete linkage + manhattan 0.03551609322974472
F1-score for average linkage + cosine 0.370502679254912
F1-score for average linkage + euclidean 0.004597701149425287
F1-score for average linkage + manhattan 0.004597701149425287
```

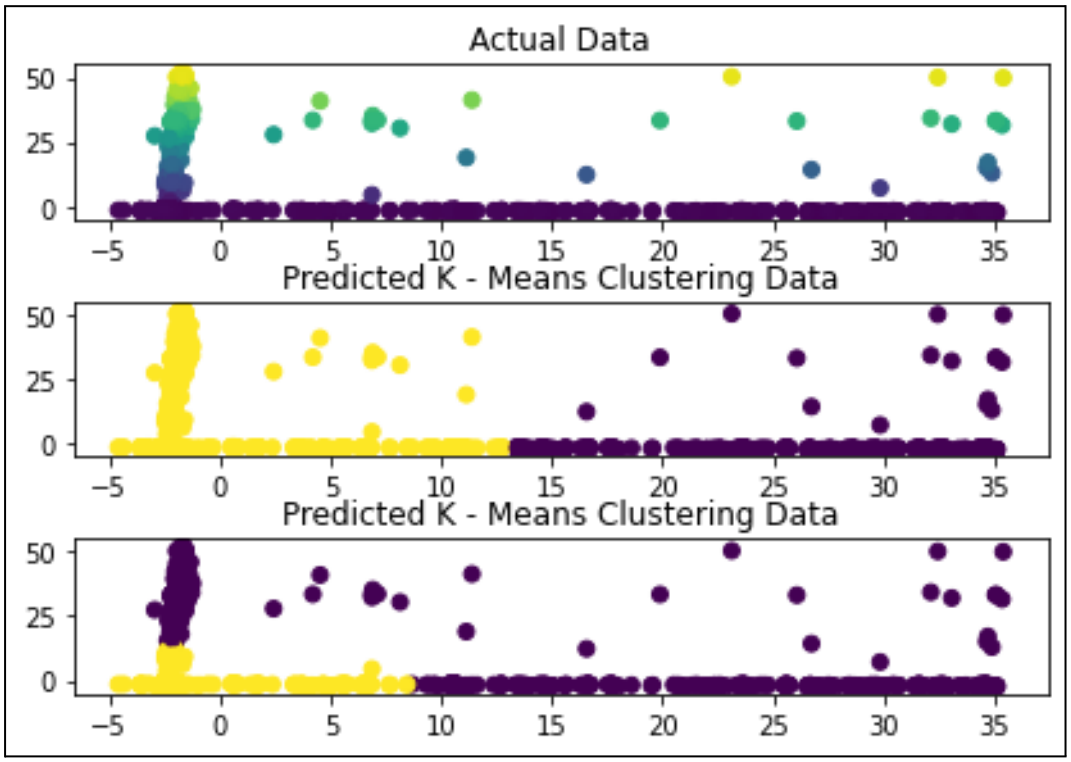
> "average linkage + cosine" is best among all.

Visualization for Hierarchical Agglomerative Clustering



Task 3: Compare K-Means Clustering and Hierarchical Agglomerative Clustering

Visualize Clusters



Comparing precision, recall, and F1-score for both model

```
K-means Clustering Scores
              precision    recall  f1-score   support

           0       0.72      0.06      0.11      3779
           1       0.24      0.93      0.38      1221

    accuracy                           0.27      5000
   macro avg       0.48      0.49      0.25      5000
weighted avg       0.61      0.27      0.18      5000

Confusion matrix:

[[ 229 3550]
 [  88 1133]]


Hierarchical Agglomerative Clustering Scores
              precision    recall  f1-score   support

           0       0.68      0.12      0.20      3779
           1       0.23      0.83      0.36      1221

    accuracy                           0.29      5000
   macro avg       0.46      0.47      0.28      5000
weighted avg       0.57      0.29      0.24      5000

Confusion matrix:

[[ 438 3341]
 [ 204 1017]]
```

Reasoning:

K-mean Clustering performance is much better than heirarchical Agglomerative Clustering.