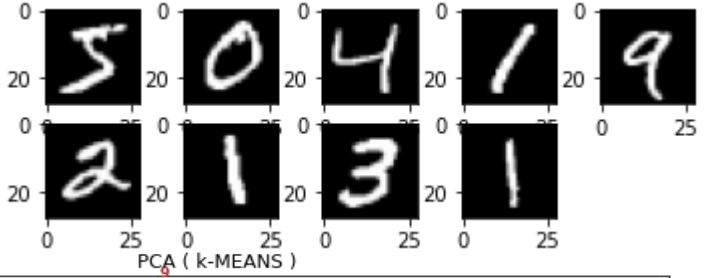
ASSIGNMENT - 4

ELL 784 (INTRODUCTION TO MACHINE LEARNING)

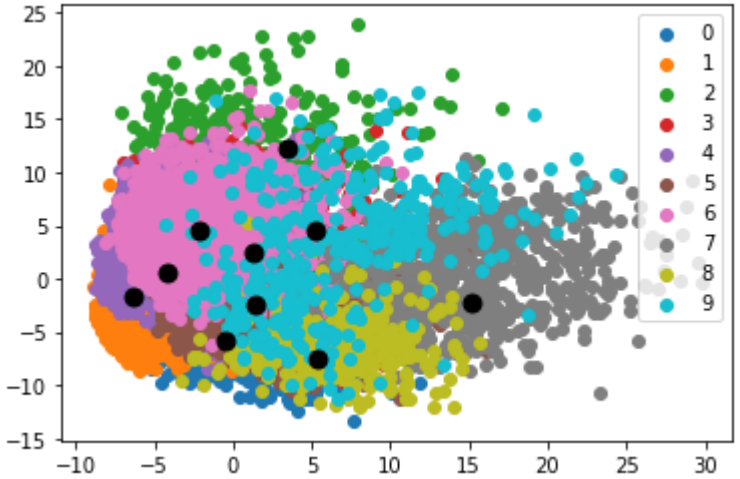
ROHAN KUMAR BOHARA (2021AMA2095)

PART 1(A): K-means on mnist handwritten digits dataset

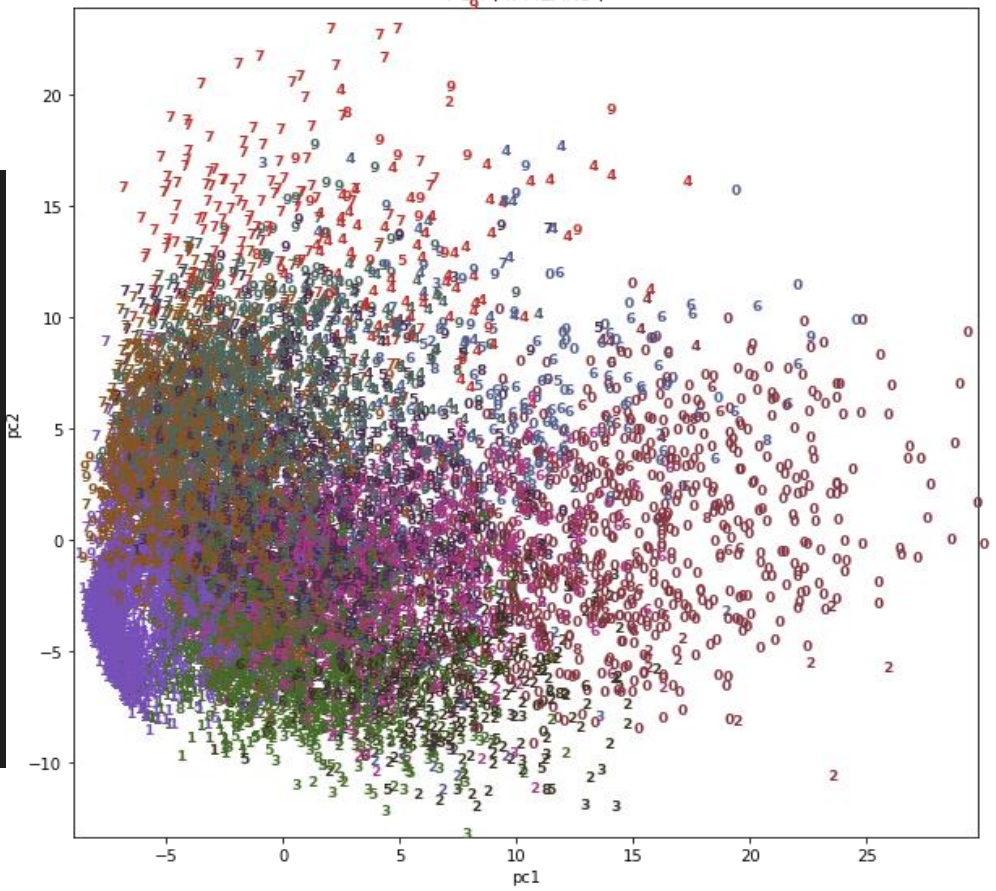
- Used sklearn library for k-means



K-means with k =10 on handwritten digits

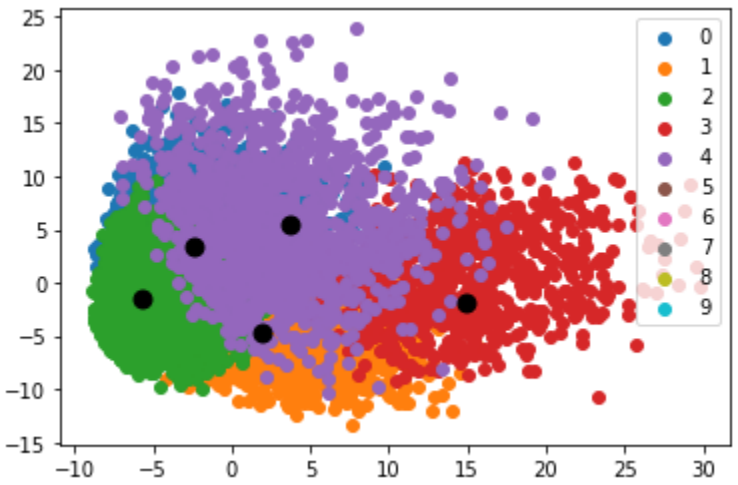


	0	1	2	3	4	5	6	7	8	9
0	19	13	0	140	11	50	33	622	63	29
1	72	1052	0	6	0	3	0	0	1	1
2	205	170	3	44	1	208	28	20	190	163
3	257	193	3	489	12	20	3	1	18	14
4	3	84	74	34	81	19	650	9	0	28
5	155	125	1	339	167	12	53	4	25	11
6	10	62	0	7	0	683	7	32	60	97
7	29	147	149	37	550	1	112	0	0	3
8	349	127	2	302	100	6	51	9	11	17
9	11	57	19	58	366	1	473	4	0	20

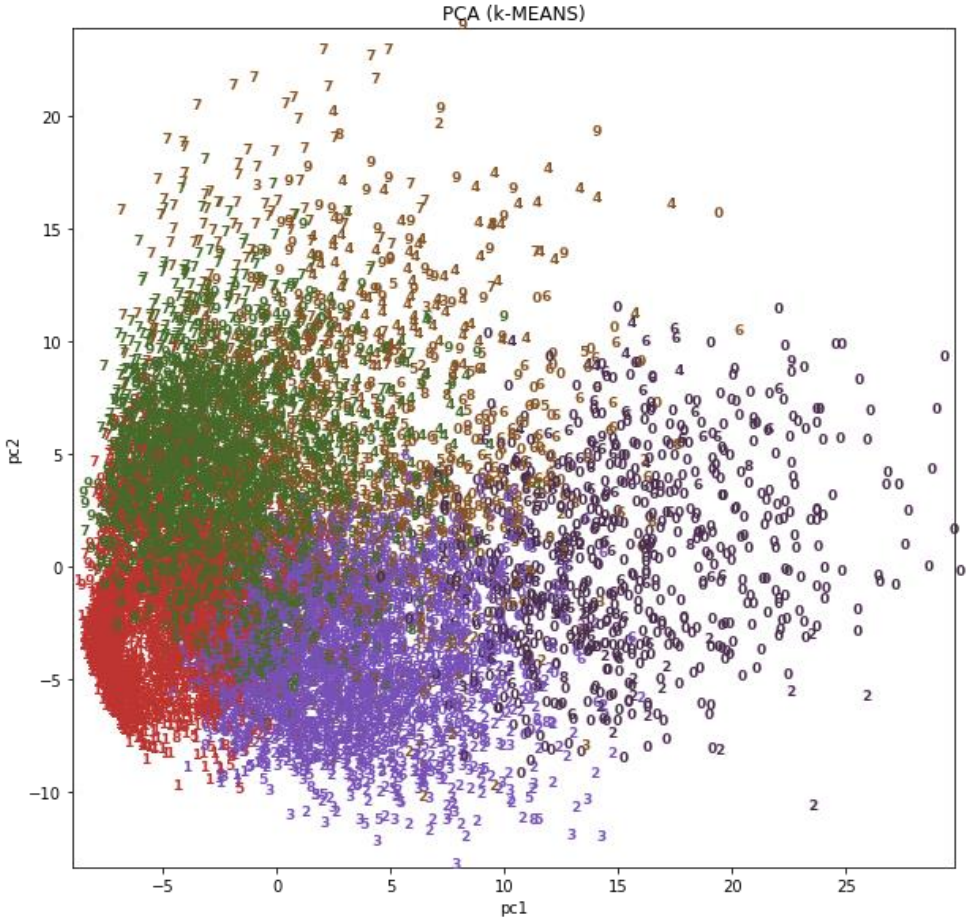


Accuracy is 20 %

K-means with k =5 on handwritten digits

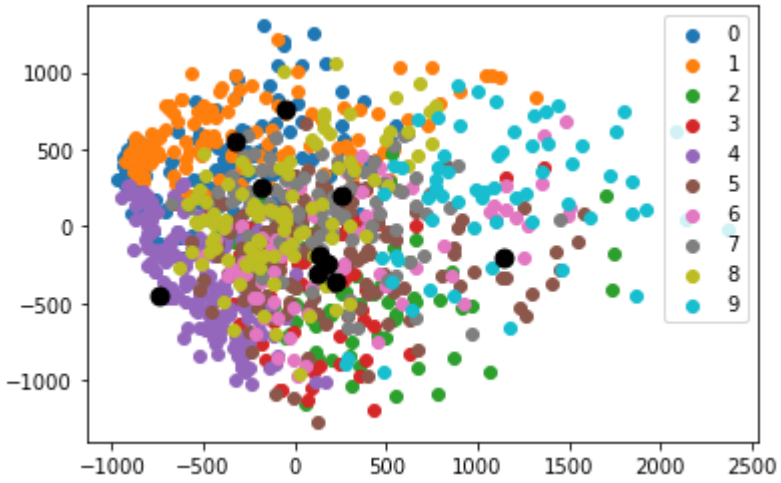


	0	1	2	3	4	5	6	7	8	9
0	14	268	17	614	67	0	0	0	0	0
1	0	17	1115	0	3	0	0	0	0	0
2	24	590	206	26	186	0	0	0	0	0
3	29	403	359	1	218	0	0	0	0	0
4	657	19	123	11	172	0	0	0	0	0
5	257	282	214	6	133	0	0	0	0	0
6	6	678	70	86	118	0	0	0	0	0
7	695	11	196	1	125	0	0	0	0	0
8	226	279	295	13	161	0	0	0	0	0
9	775	9	95	5	125	0	0	0	0	0

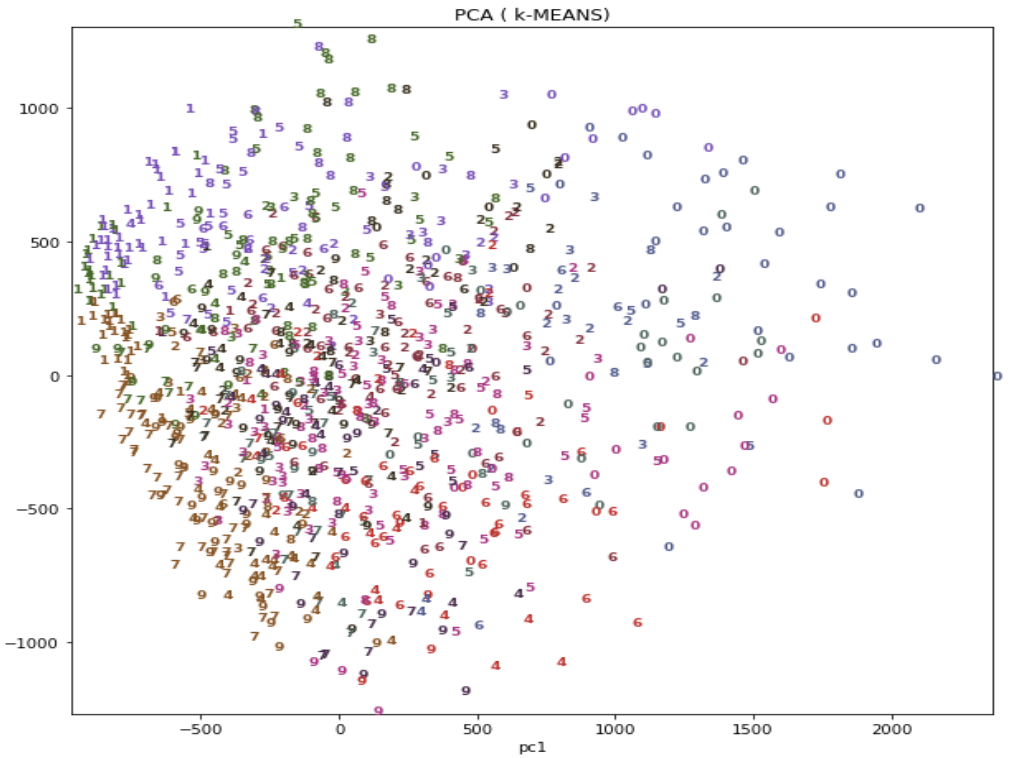


Accuracy is 5 %

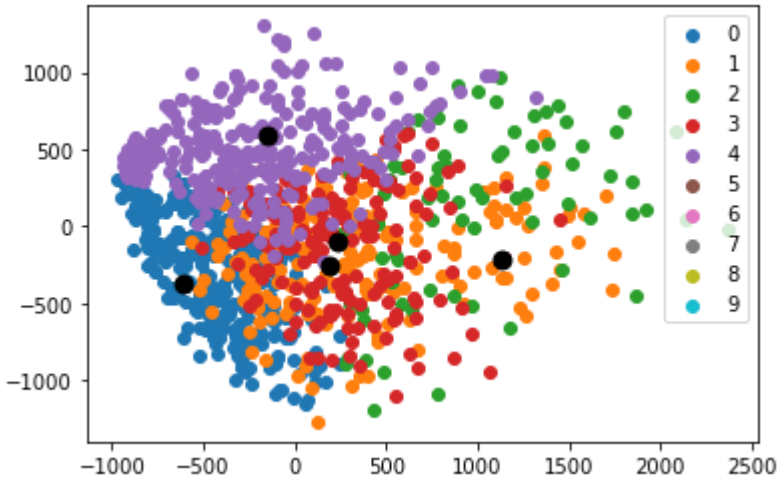
K-means with k =10 on handwritten digits 3000



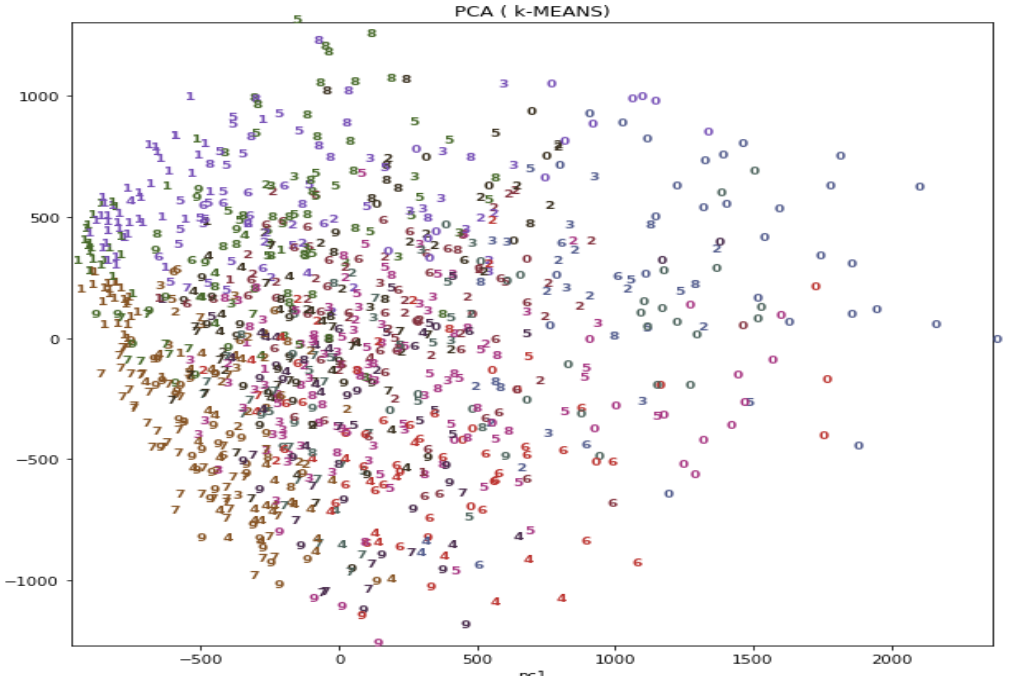
	0	1	2	3	4	5	6	7	8	9
0	0	13	10	3	0	15	25	4	6	29
1	24	42	0	1	27	2	0	3	1	0
2	1	12	8	1	11	6	2	27	19	10
3	11	9	0	0	8	50	4	1	1	7
4	6	2	15	14	30	0	2	1	15	2
5	18	23	1	10	1	28	11	2	1	5
6	1	9	32	3	6	0	0	51	1	4
7	8	1	1	11	44	0	20	0	21	0
8	35	11	2	2	4	27	4	1	13	6
9	13	0	4	17	35	4	6	0	23	0



K-means with k =5 on handwritten digits 3000

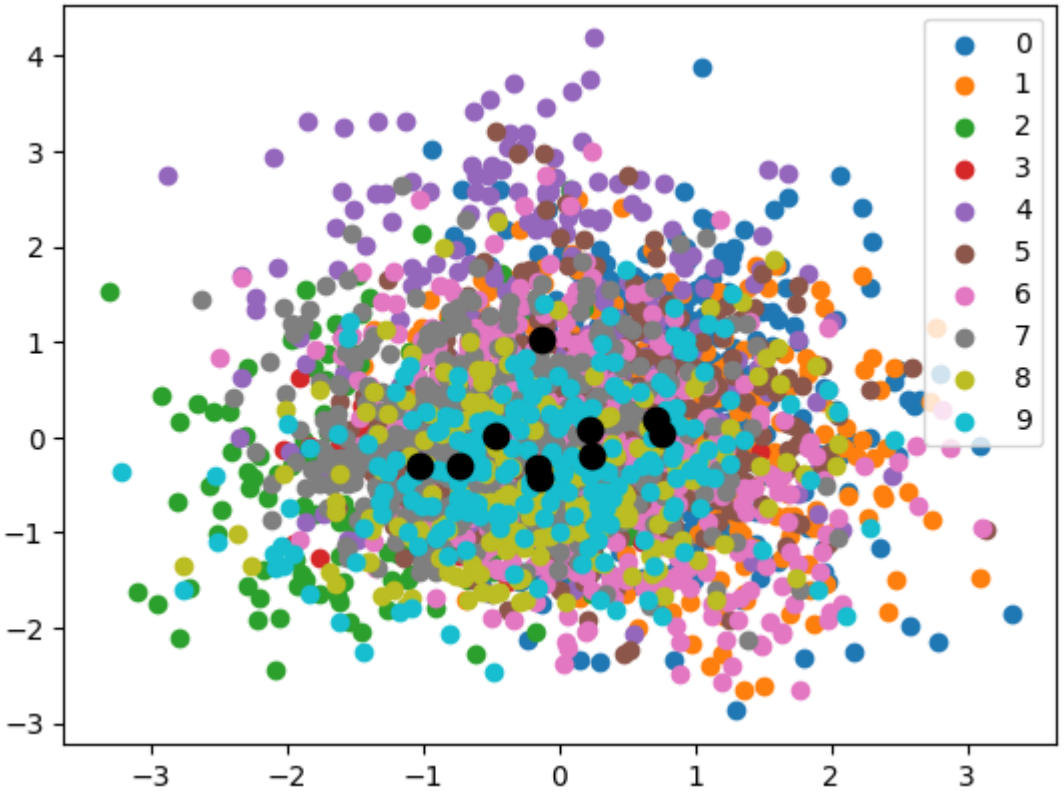


	0	1	2	3	4	5	6	7	8	9
0	0	46	29	13	17	0	0	0	0	0
1	35	1	0	2	62	0	0	0	0	0
2	15	6	13	44	19	0	0	0	0	0
3	8	56	7	1	19	0	0	0	0	0
4	50	0	4	21	12	0	0	0	0	0
5	1	50	3	2	44	0	0	0	0	0
6	11	1	8	71	16	0	0	0	0	0
7	84	3	0	3	16	0	0	0	0	0
8	6	28	6	3	62	0	0	0	0	0
9	78	6	5	4	9	0	0	0	0	0



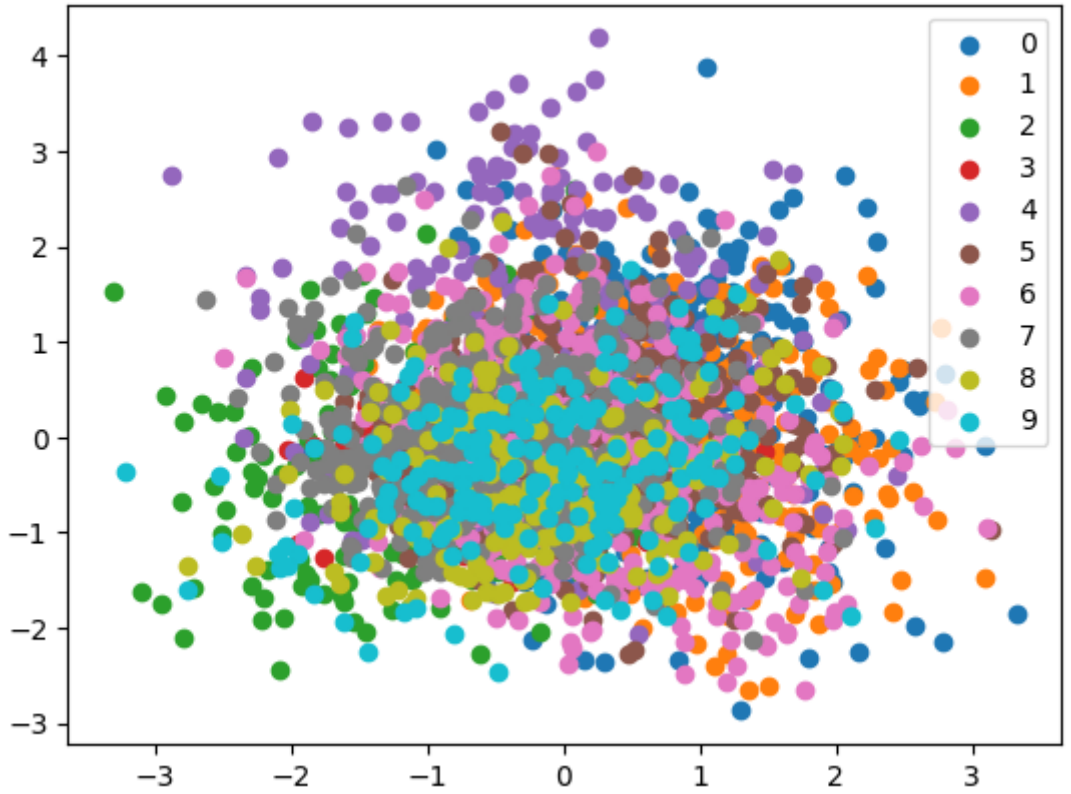
PART 1(B): K-means using PCA (25 features)

K-means with k =10 on handwritten digits 3000



Accuracy is 13%

Visualization of PCA



Conclusion:

- PCA dimension data gives good accuracy as compared to raw pixel data.
- With k=10 k-means performs well as compared to k=5 due to misclassification happening with all the digits.

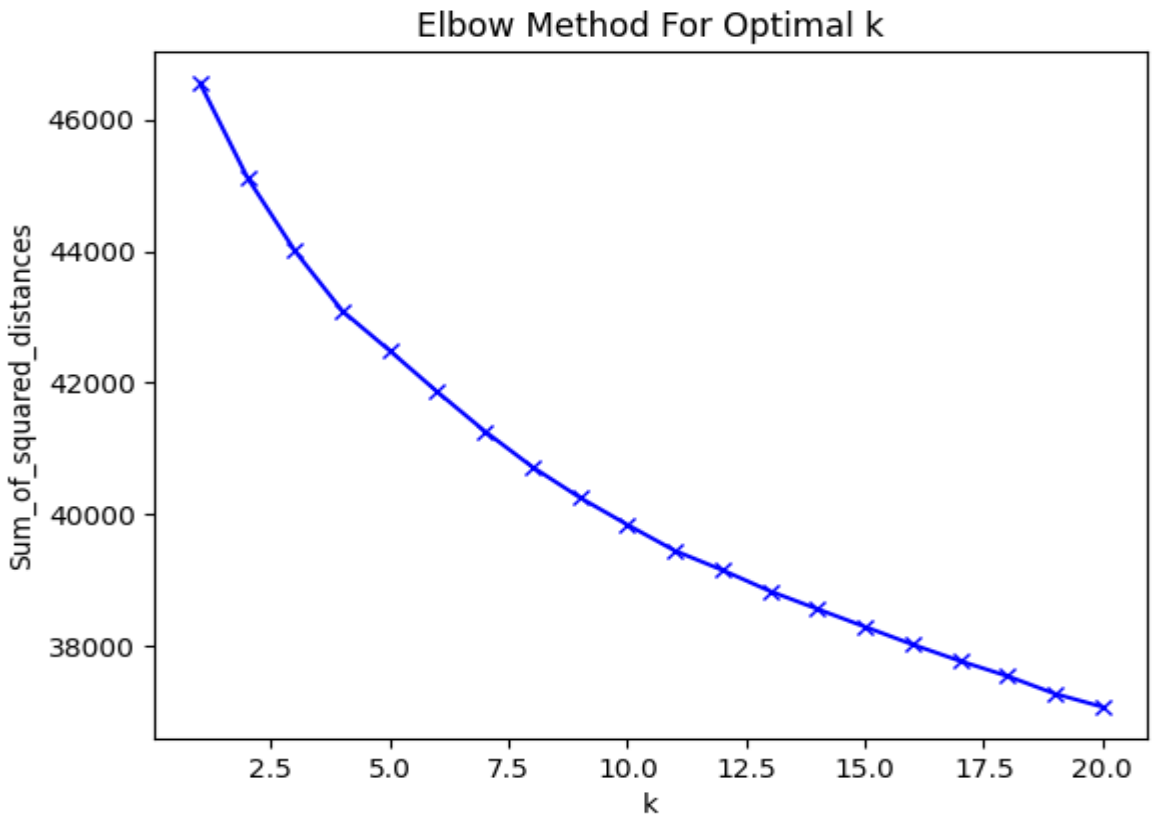
PART 2: K-means on mystery dataset with 15,000 datapoints and 127 features.

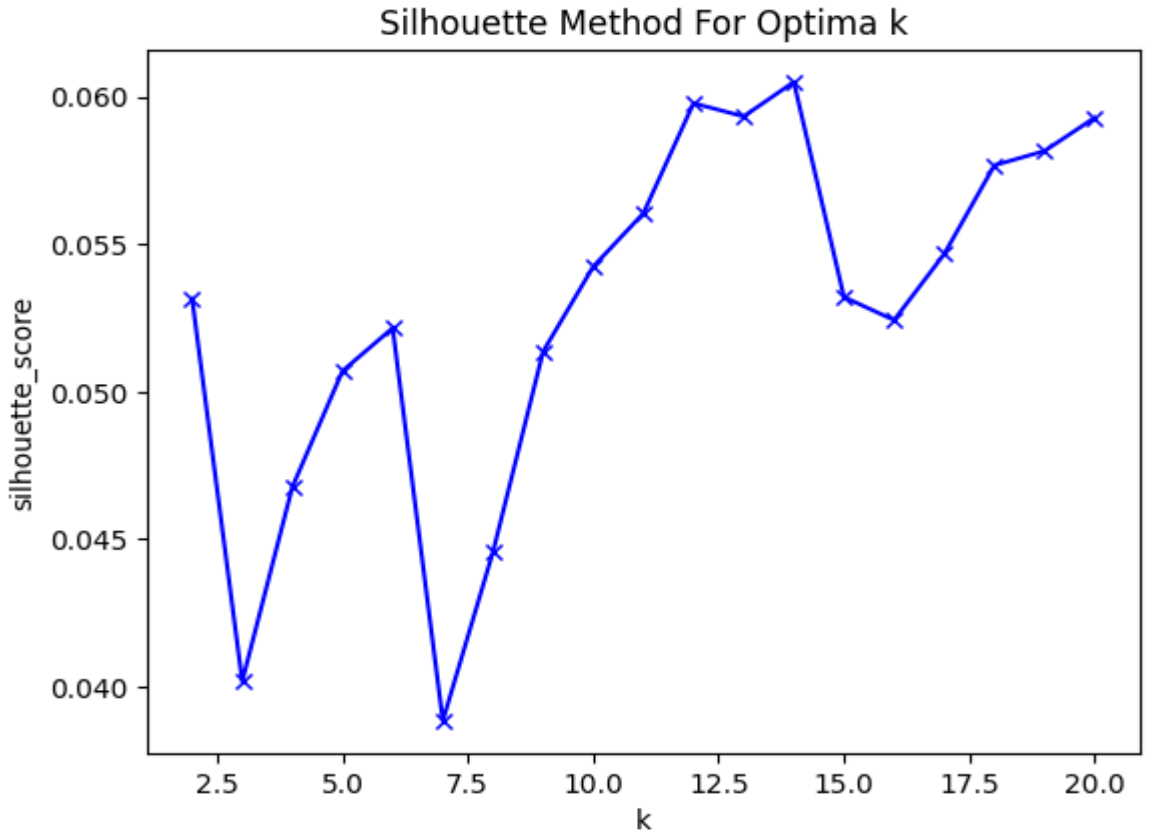
```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=20, random_state=0)

kmeans.fit(df)
```

Optimum k=14



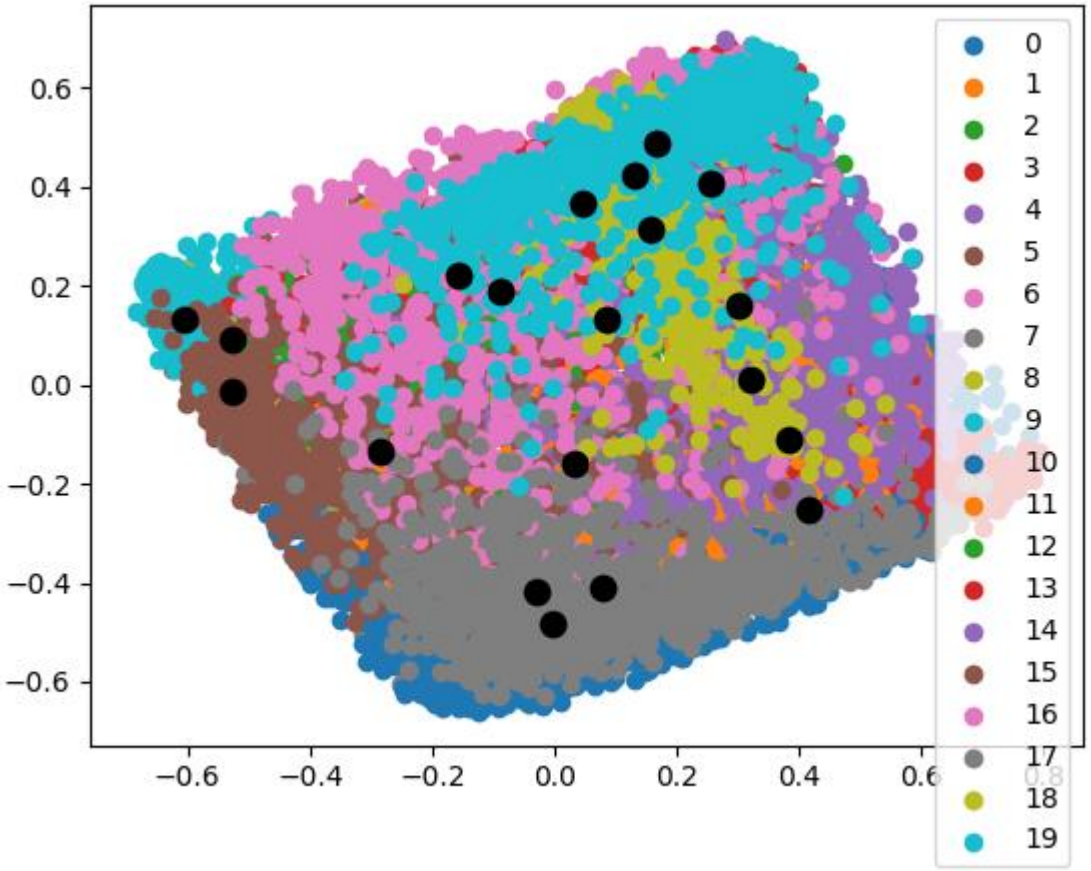


- Predicted the cluster labels for given data at k = 14.

PCA visualization

```
pca = PCA(n_components=25, random_state=0)
```

K-means for 25 pca features



With 127 features