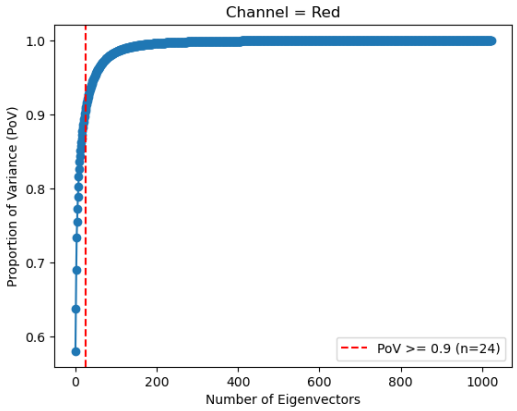
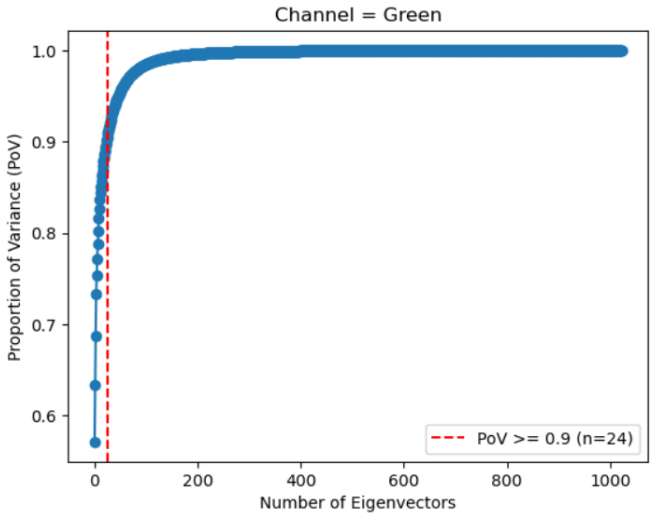
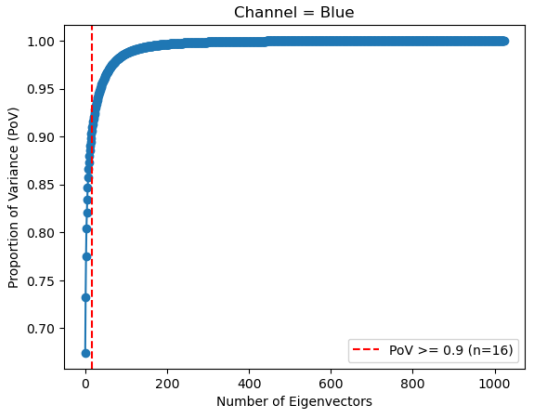
Q.1 Principal component analysis

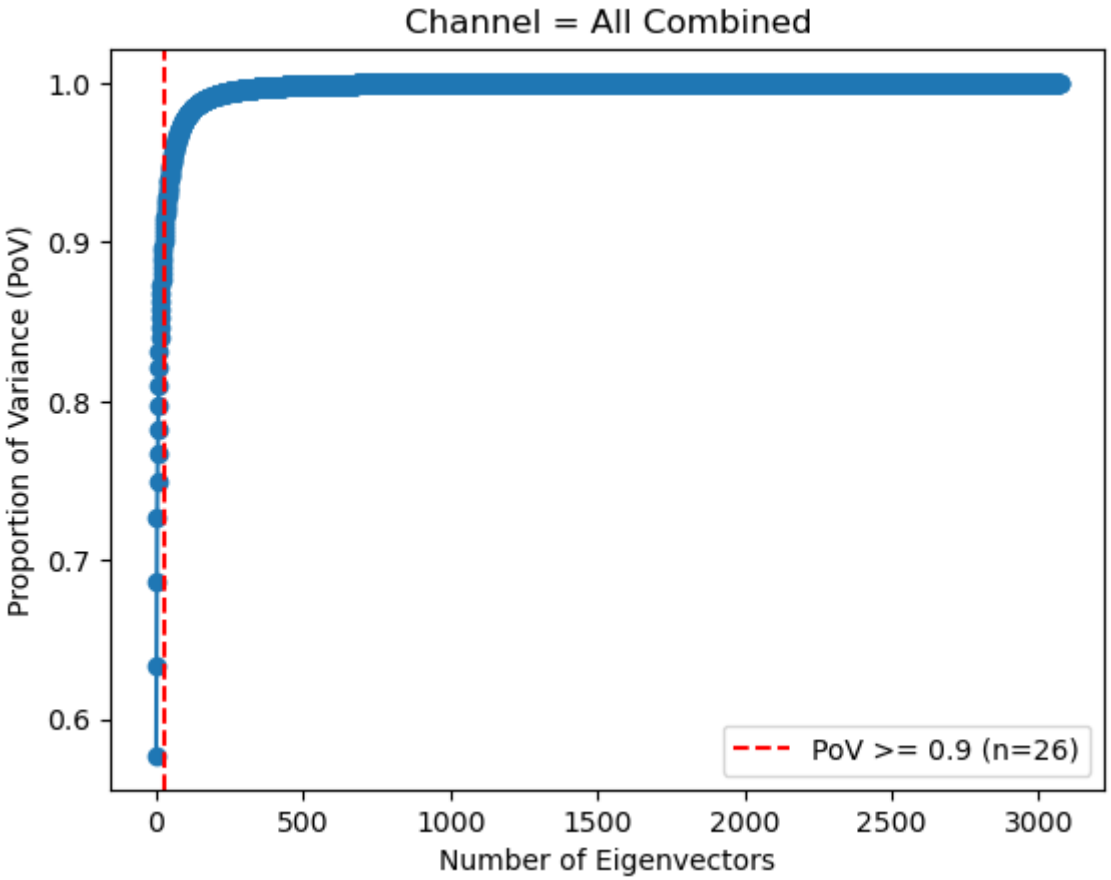
This illustrates the number of top eigenvectors needed to maintain a variance proportion above 0.9. Initially, we segregated the three channels of the image and plotted the proportion of variance (POV) for each channel against the number of eigenvectors and then plotting with all combined channels

Red	Green	Blue





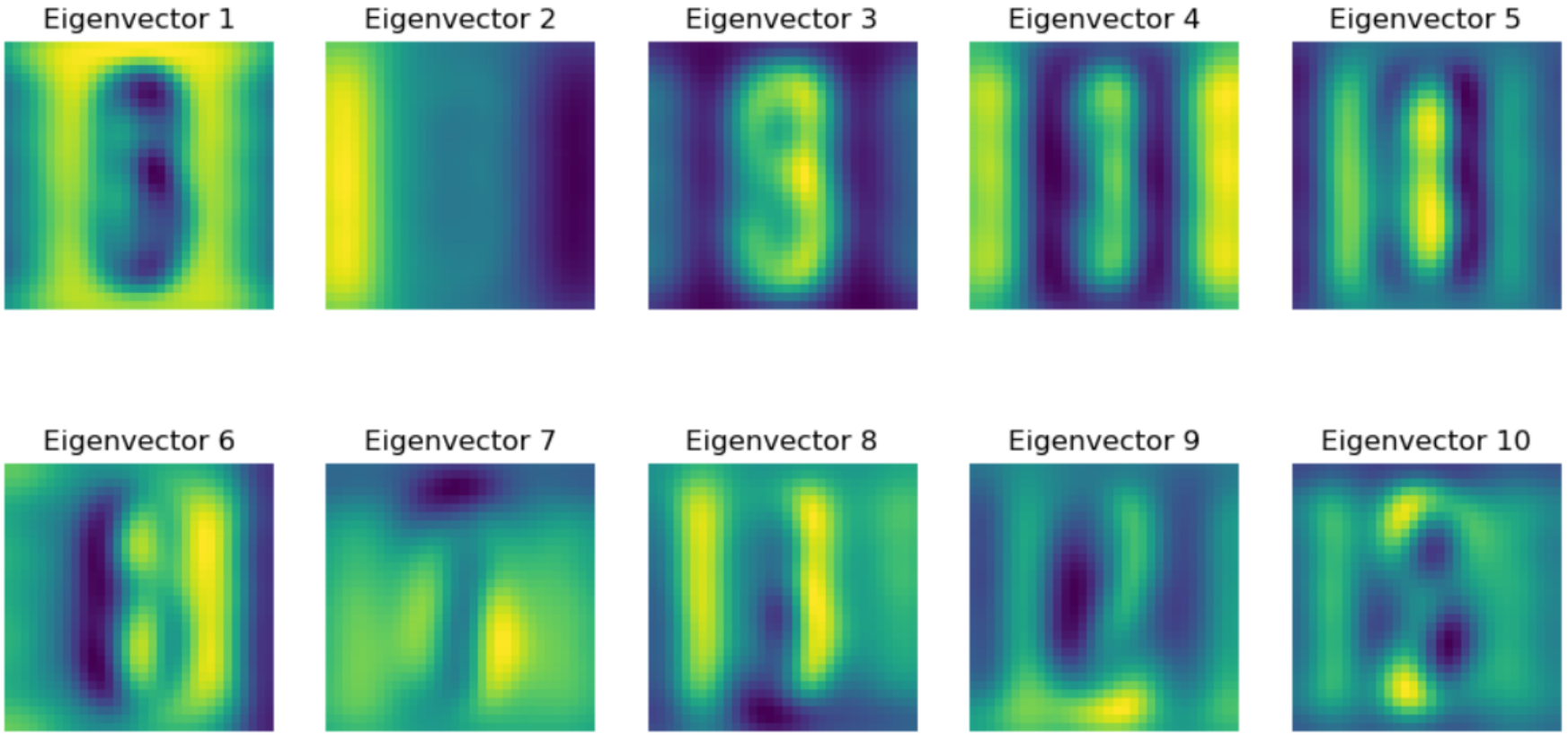




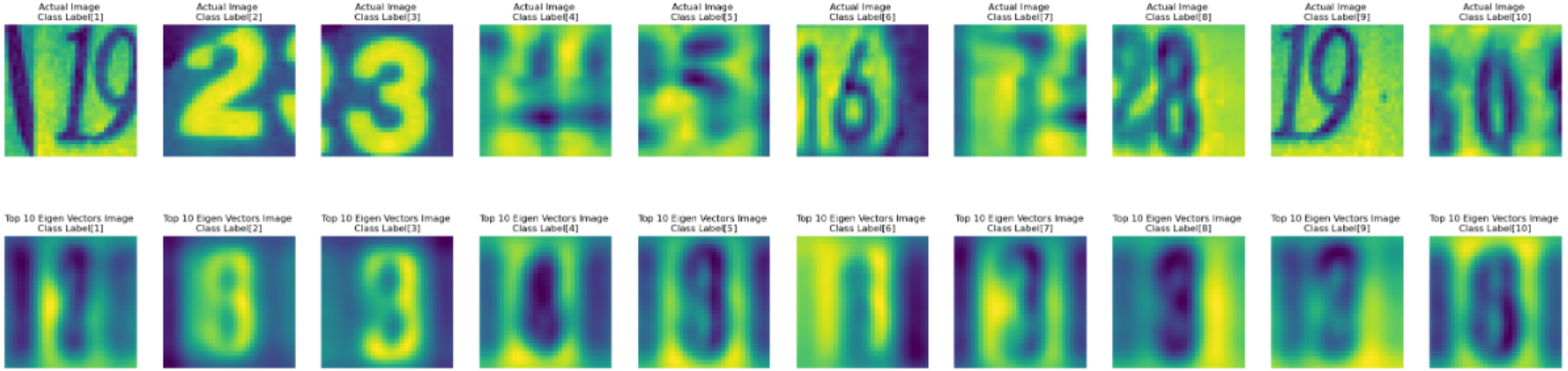
The charts above reveal that, for Points of View (POV) greater than or equal to 0.9, the red and green channels necessitate 24 components each, while the blue channel requires 16 components. However, when amalgamating all channels and applying Principal Component Analysis (PCA) to attain a POV greater than or equal to 0.9, 26 components are needed. Consequently, in our subsequent calculations, we will employ **26 components for POV greater than or equal to 0.9**.

Visualize top 10 eigenvectors and provide reconstruction of 10 SVHN samples (one from each class) using top 10 eigenvectors.

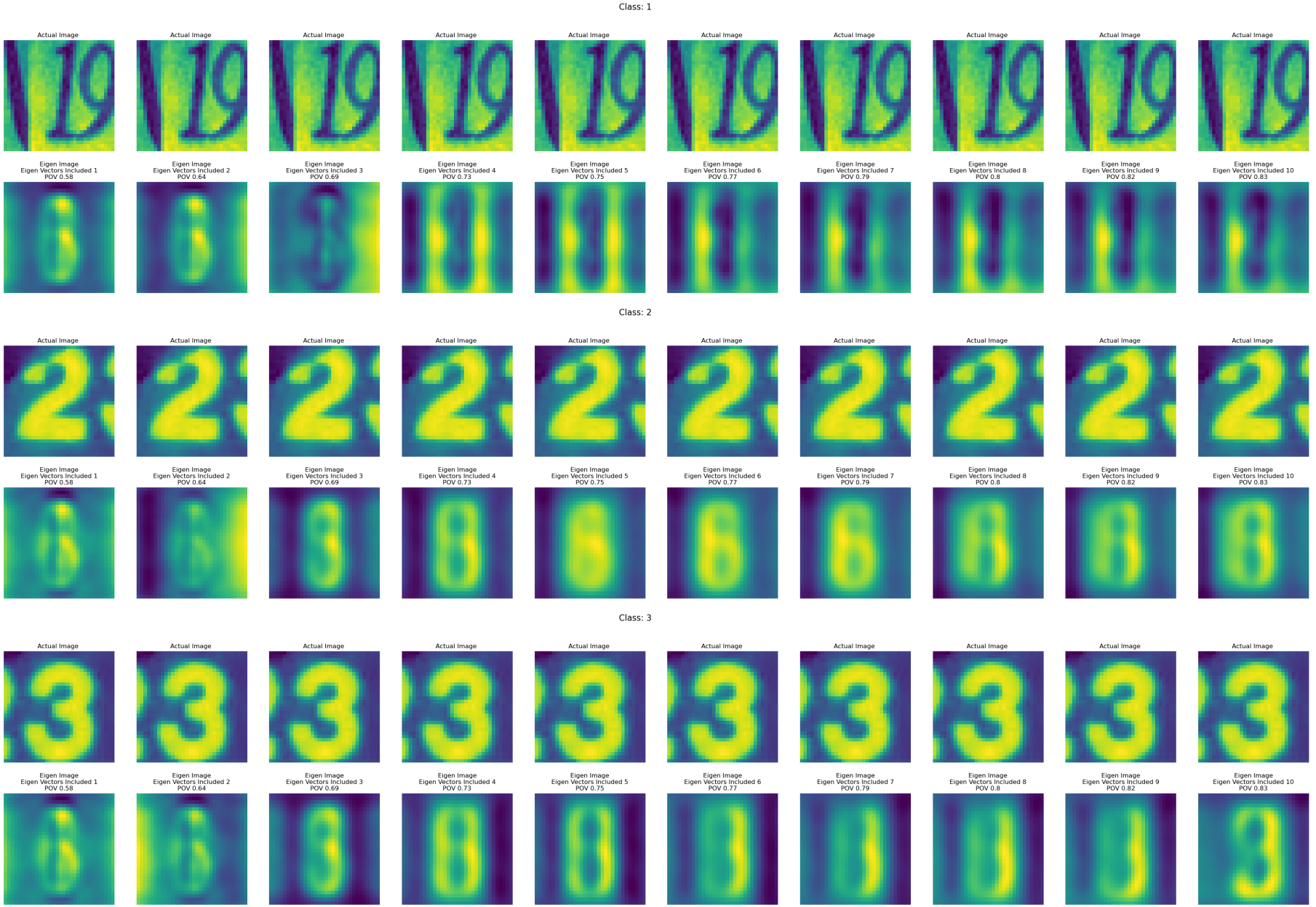
Visualizing top 10 Eigenvector



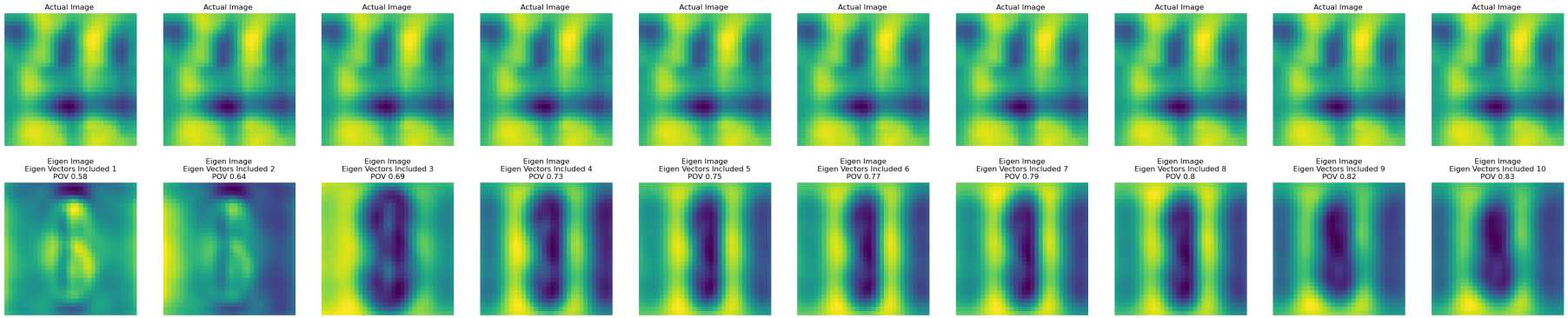
Reconstruction of 10 sample form each class using top 10 eigenvectors



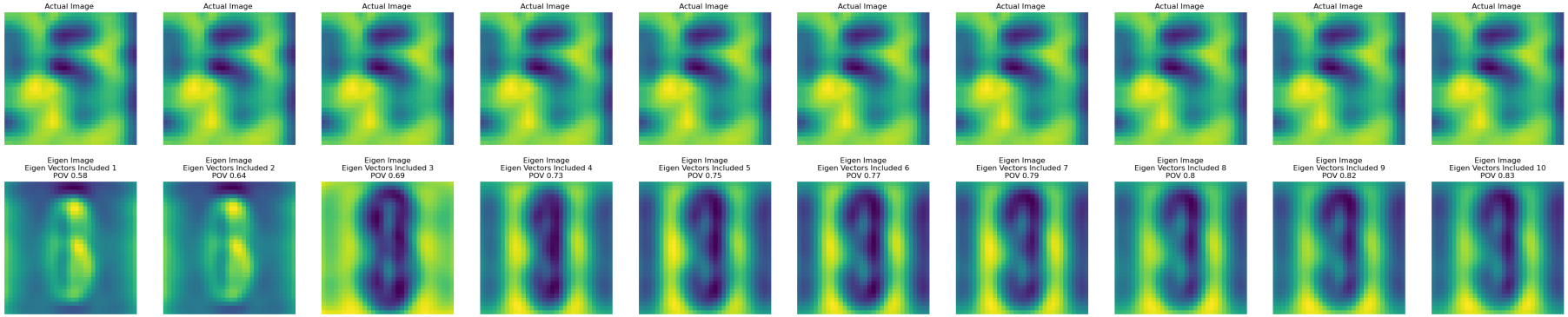
Reconstruction of 10 images from each class by cummulative increase of POV with 10 eignevectors



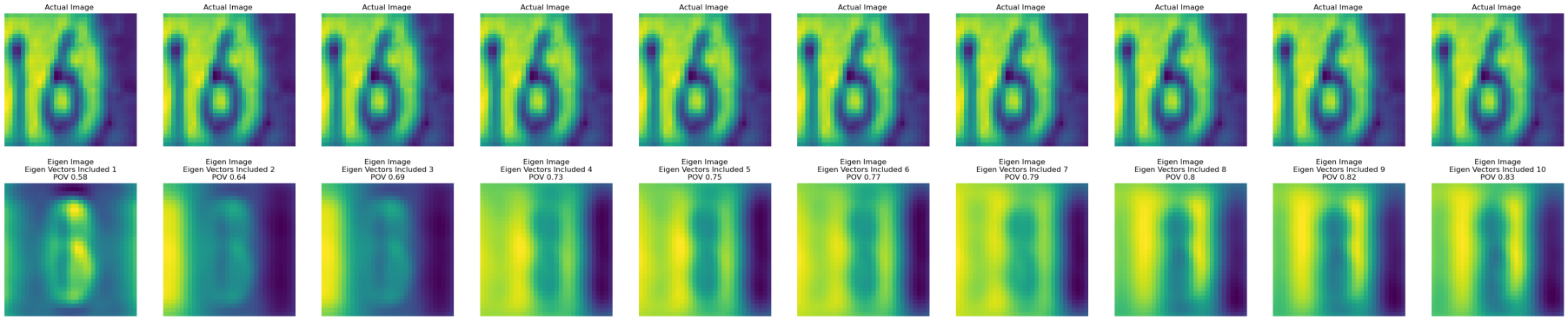
Class: 4



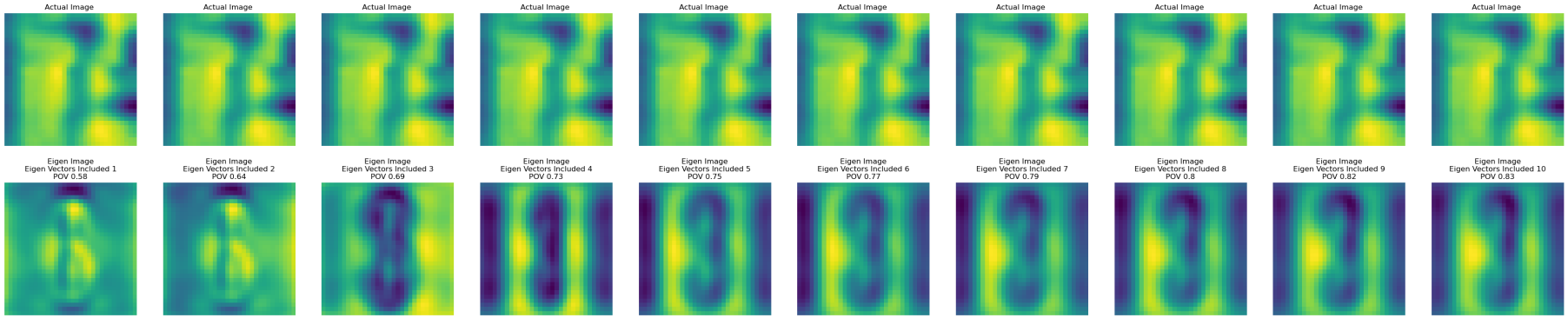
Class: 5



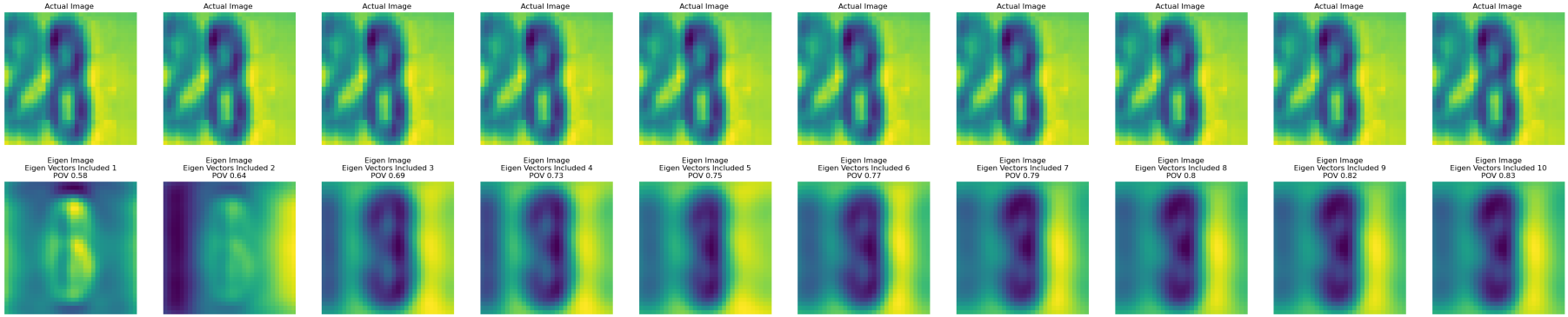
Class: 6

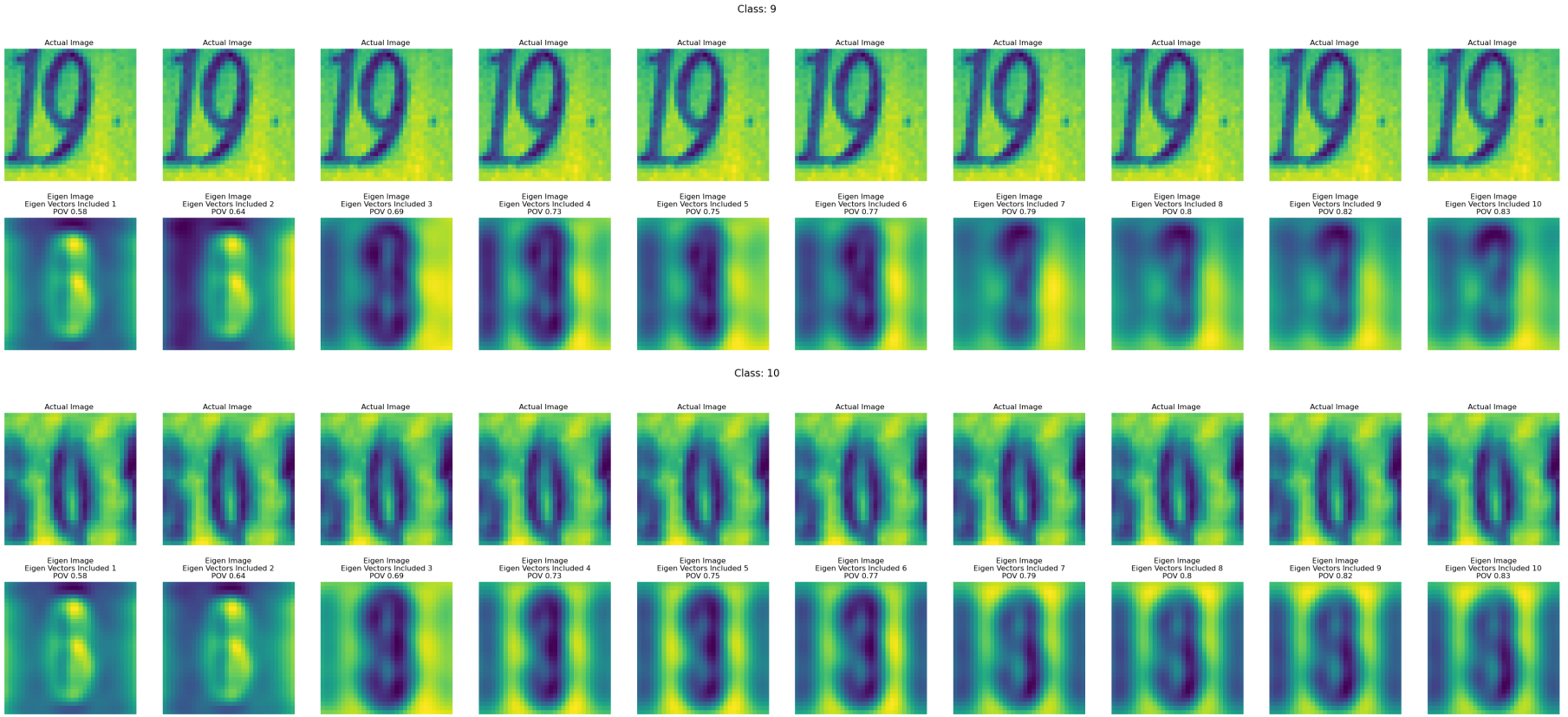


Class: 7



Class: 8





1. Image constructed using Top 10 eigen vectors gave indicative features of the actual image and was blurred as POV=0.83 is only explained by top 10 eigen vectors
2. Image quality improves with the increase in number of top eigen vectors gets included as variance explained by them increases

Runing k-NN (for k=5 and k=7) on raw data and data obtained after PCA dimensionality reduction for dimension as found in part (a) and for dimension 10 as in part (b).

Analysis of k-NN with Raw data, with dimension 26, with dimension 10 for K = 5 and K = 7

Data	Test Accuracy(K=5)	Test Accuracy(K=7)
Raw Data	46.83	47.51

Data	Test Accuracy(K=5)	Test Accuracy(K=7)
Data with 26 dim	46.07	46.85
Data with 10 dim	29.93	30.85

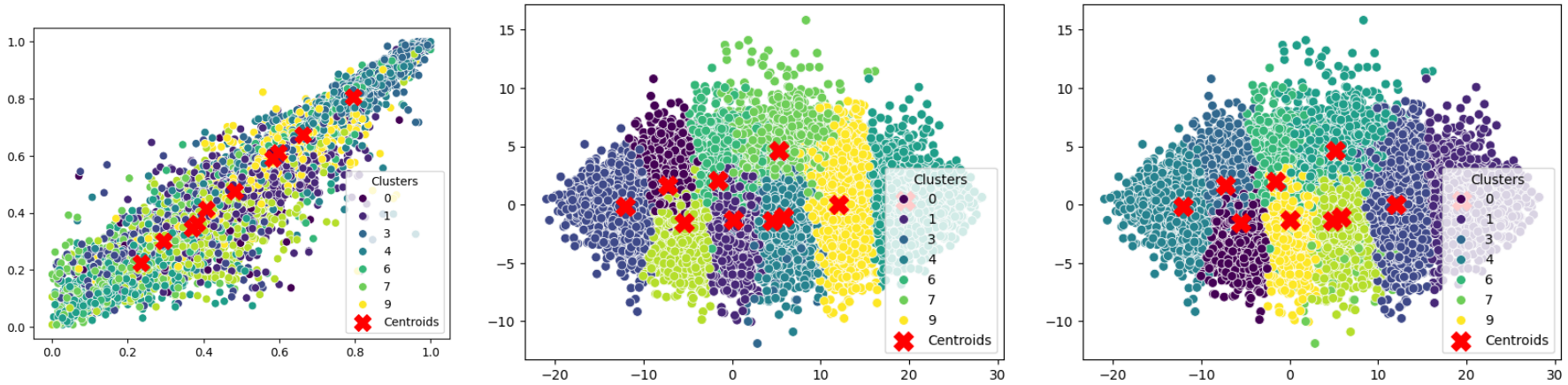
From the above results we observe that

1. Test Accuracy drops only minutely when dimension is reduced to 26 as POV explained by it exceeds 0.90
2. Test Accuracy drops significantly as POV explained by it is nearly 0.80

Q2. K-Means clustering

K-Means(K = 10) clustering with raw data and data obtained by PCA with POV >= 0.9(where we got POV >= 0.9 with top 26 components) and data obtained by PCA with 10 component

Raw Data	PCA(components = 26)	PCA(components = 10)

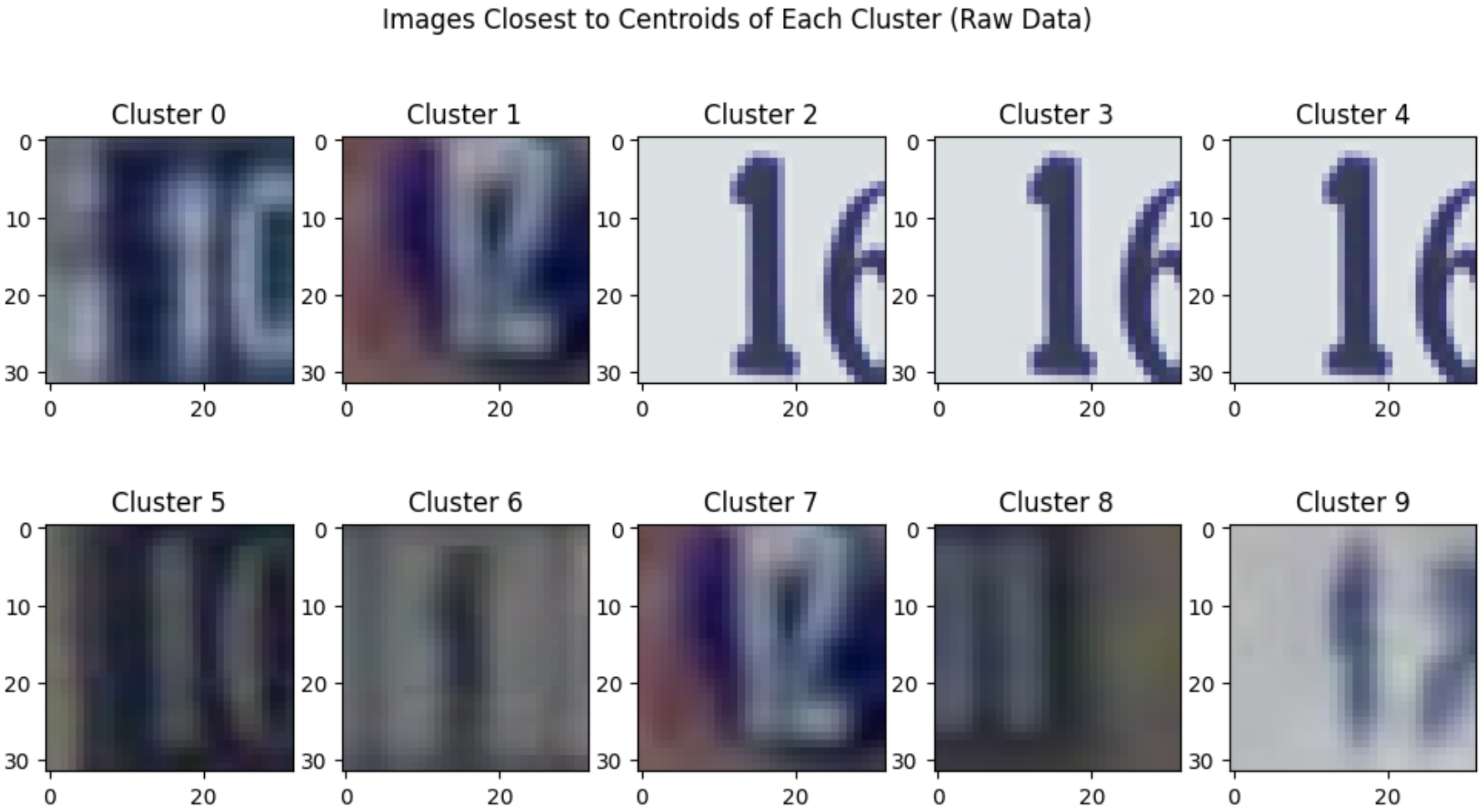


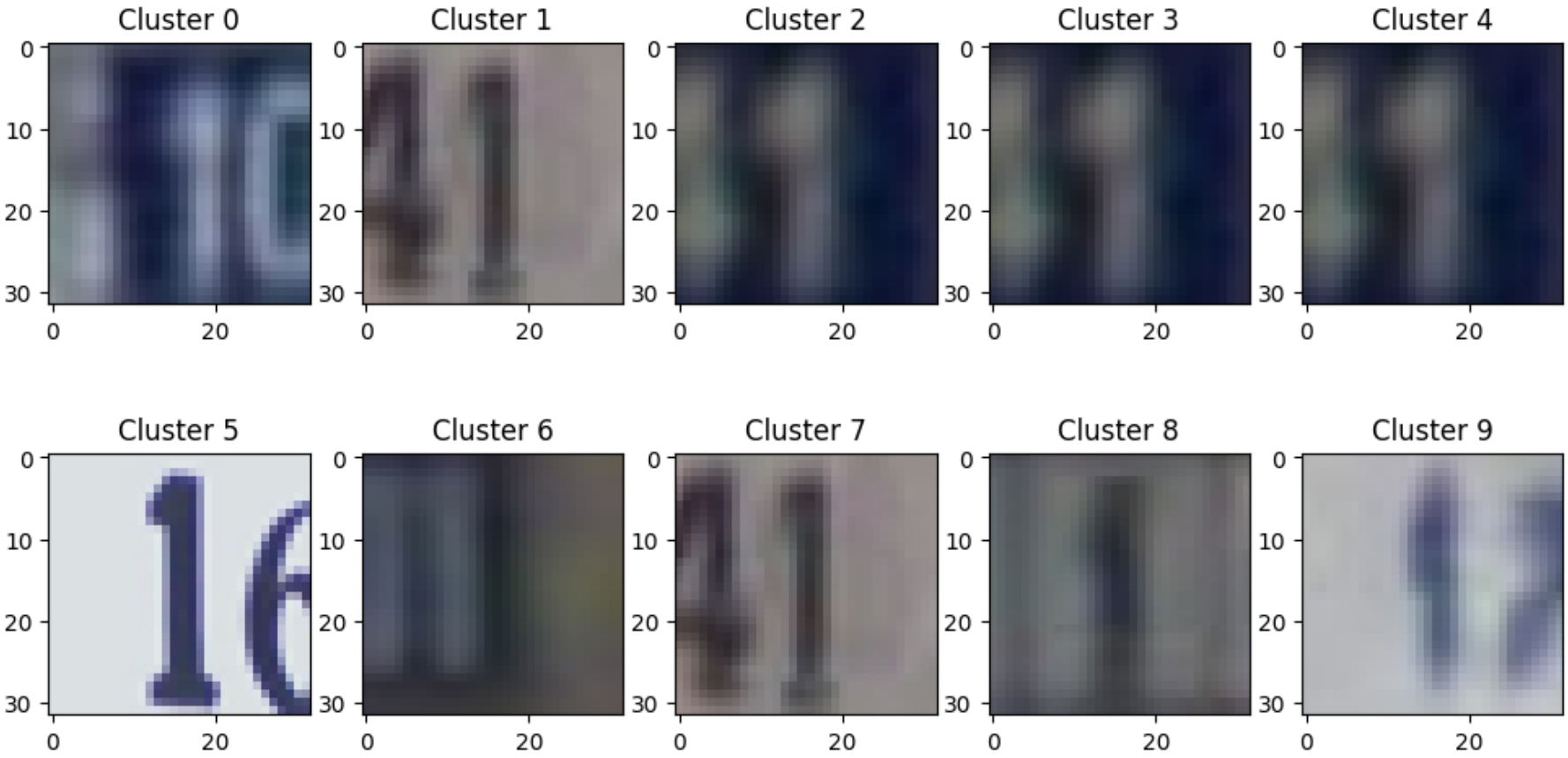
As we can see above clustering images the cluster of raw data is very overlapping with each other but with PCA components = 26 and 10 we can see the most of the cluters are well separated except two cluster centroid are very closer

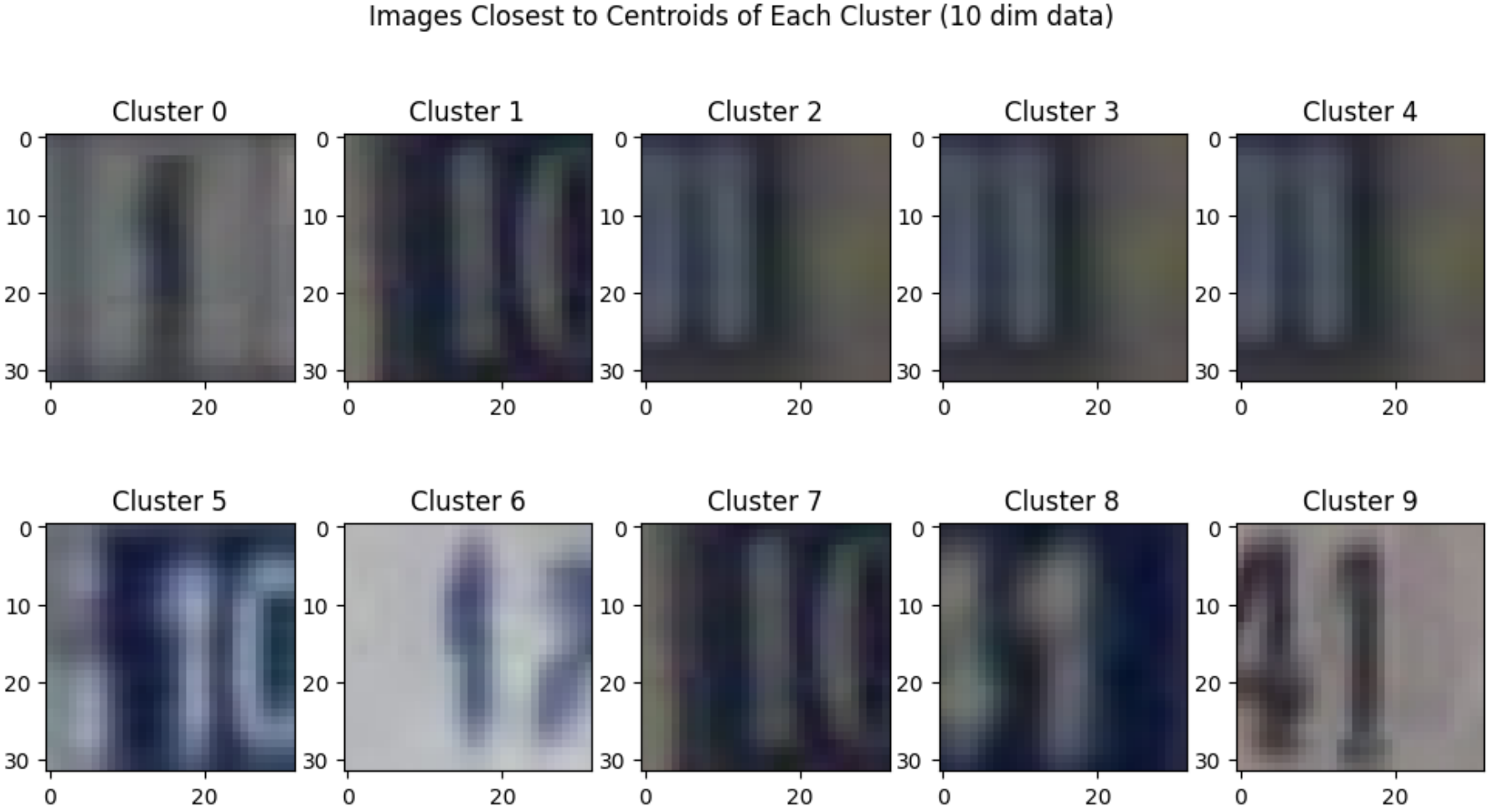
When clustering is performed on raw image data, especially if the data has a high dimensionality, the distance between data points (images) may not reflect the true underlying structure of the data. High-dimensional spaces often suffer from the curse of dimensionality,

where the distance between points becomes less meaningful, and the data points are more likely to be equidistant or nearly equidistant from each other. This can lead to overlapping clusters, as the raw features might not capture the true patterns or variations in the data.

On the other hand, PCA is a dimensionality reduction technique that transforms the original features into a new set of orthogonal features called principal components. These principal components are ordered in terms of the amount of variance they explain in the data. By selecting a subset of these principal components that capture most of the variance, you effectively reduce the dimensionality of the data.




Images Closest to Centroids of Each Cluster (26 dim data)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
0
10
20
30
0
20



Evaluting different cluster with metrices

Data	Accuracy	SSE	Purity	Rand Index
Raw Data	0.1158	1242934.50	0.1530	0.0005
Data with 26 dim	0.11576	921024.78	0.15372	0.0006
Data with 10 dim	0.11588	703045.50	0.15364	0.0005

Possible Reasons Behind Results:

- The SVHN dataset might have inherent complexities or noise that make it challenging for clustering algorithms to accurately group data points.
- The dataset might have outliers or anomalies that affect the clustering performance.
- The choice of the number of clusters may not align well with the inherent structure of the data, affecting the accuracy and SSE values.

Each cluster with the digit that occurs most frequently within it.

PCA Dimension	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
26 Class Labels	10	8	2	1	7	9	4	1	10	10
10 Class Labels	8	4	7	1	10	10	9	10	7	10
Raw Data Class Labels	4	10	7	10	2	8	10	7	1	9

Explanation:

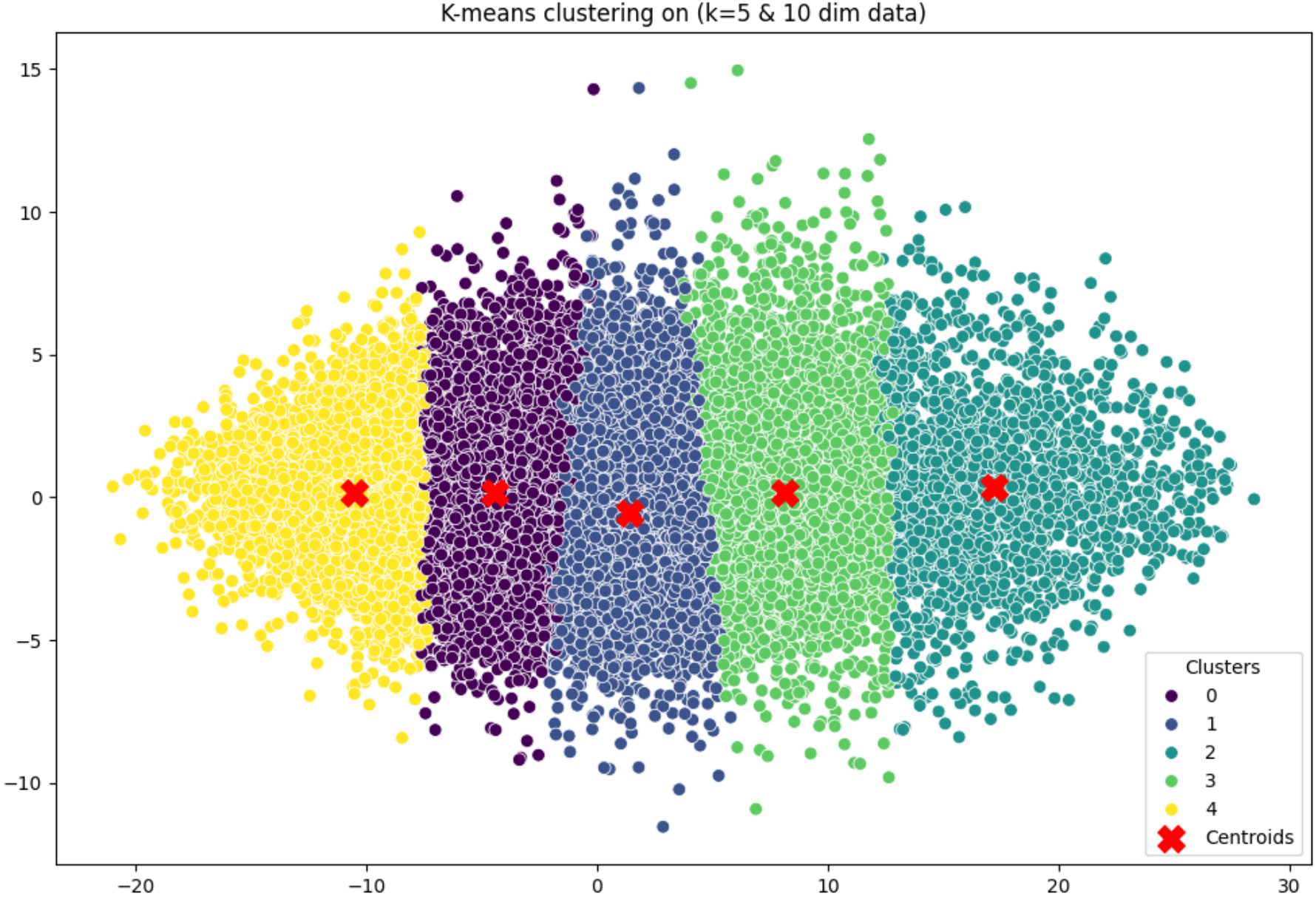
1. **PCA with dim = 26:**
    - For each cluster, the majority class label and its frequency are provided, along with the second majority class label and its frequency. This information gives insight into the dominant classes within each cluster after PCA with 26 dimensions.
2. **PCA with dim = 10:**
    - Similarly, for PCA with 10 dimensions, the majority class labels for each cluster are presented. The clusters are now based on the reduced-dimensional representation obtained from PCA with 10 dimensions.
3. **Raw data:**
    - The clusters formed on the raw data without dimensionality reduction are presented with the majority class labels for each cluster.

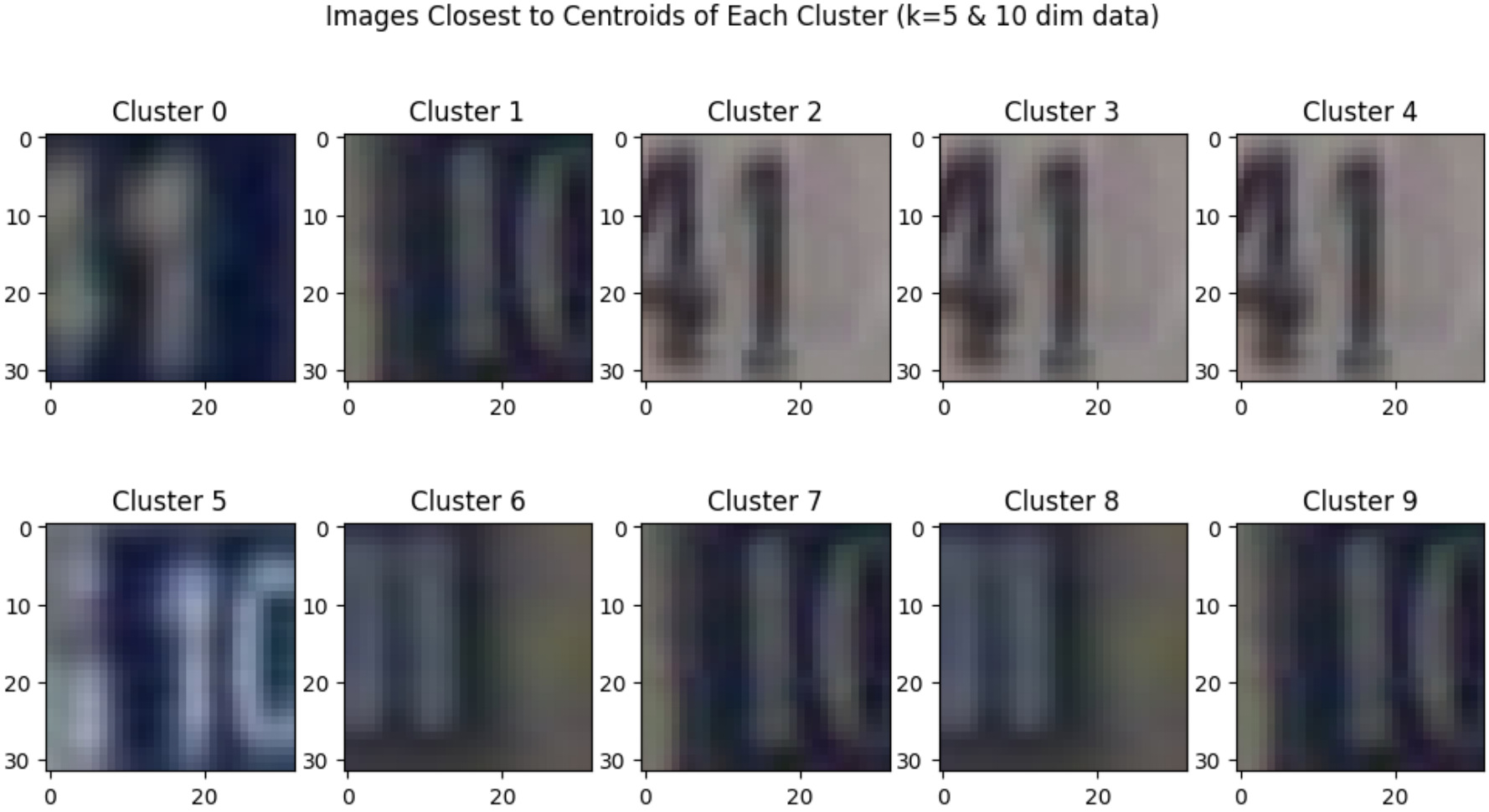
Comparison of accuracy by Q-1(c) and here

Data	Accuracy(K-Means - 2(b))	Accuracy KNN(Q1(c))
Raw Data	0.1158	48.51
Data with 26 dim	0.11576	46.85
Data with 10 dim	0.11588	30.85

In summary, the results suggest that the raw data may be challenging for K-Means to cluster effectively. Dimensionality reduction to 26 dimensions has a minor impact, and further reduction to 10 dimensions leads to a decrease in KNN accuracy, indicating that some essential information for KNN classification might have been lost in the dimensionality reduction process. The choice of dimensionality reduction technique, the number of dimensions retained, and the characteristics of the dataset can all influence the performance of clustering and classification algorithms.

Performing k-means clustering on the dataset from question 1 after applying PCA with a POV threshold of 0.9 and using k=5.


K-means clustering on (k=5 & 10 dim data)
15
10
5
0
−5
−10
−20
−10
0
10
20
30
Clusters
0
1
2
3
4
Centroids



Cluster assignment to different class labels

Cluster	Majority Class (Label/Frequency)	Second Majority Class (Label/Frequency)
0	9 (800)	3 (743)
1	4 (662)	7 (649)
2	6 (235)	10 (235)
3	2 (521)	7 (476)

Cluster	Majority Class (Label/Frequency)	Second Majority Class (Label/Frequency)
4	7 (520)	3 (513)

1. **Sum of Squared Error = 1082550.52**
2. **We observe that images from different classes are clustered together as number of clusters defined is half of the actual number of classes.**
3. **The grouping of classes within each cluster reflects the algorithm's perception of similarities among digits based on the features used for clustering. The organization suggests that digits with similar characteristics or patterns have been grouped together in each cluster.**