K-MEANS IMAGE CLUSTERING

Introduction

I first tried running the algorithm on the cifar-10/100 datasets, but, the results were rather poor as K-Means is too simple for such data, the runs for them are included below, but, I resorted to working on the mist dataset as it is much simpler and gave much, much better results.

Accuracy

I calculated the accuracy per the following equation:

First, the label of the cluster is the label that's repeated the most in this cluster.

Second, the accuracy of each cluster is calculated as $\frac{Number\ of\ occurrence\ of\ cluster\ label}{Total\ number\ of\ labels\ in\ cluster}$.

Finally, the total accuracy is calculated as the average of each cluster accuracy.

$$Accuracy = {\sum_{i=0}^{n} purity(cluster)_i}/_{n}$$

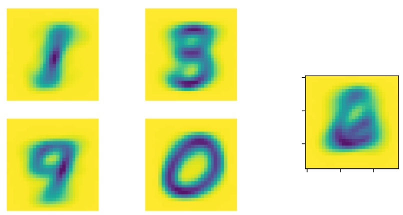
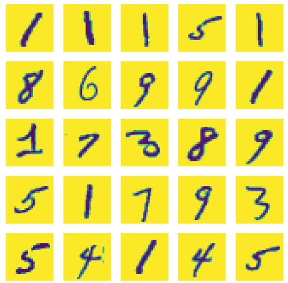
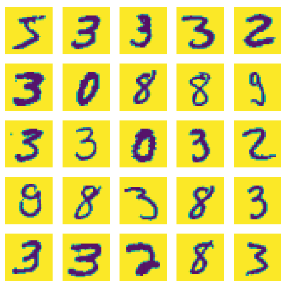
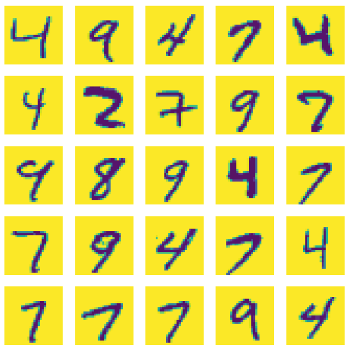
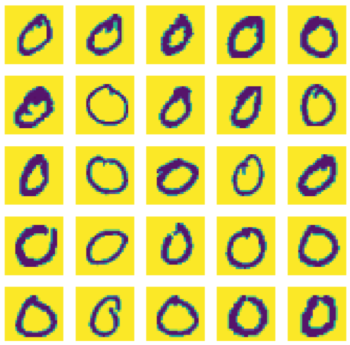
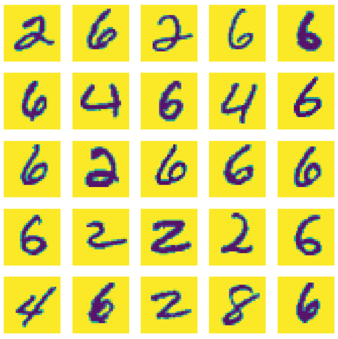
MNIST

Tests were run for k values equal to [5, 7, 10, 20] on all 60k samples results are as follows:

5 Clusters [49 iterations to converge]:

Accuracy: 0.516

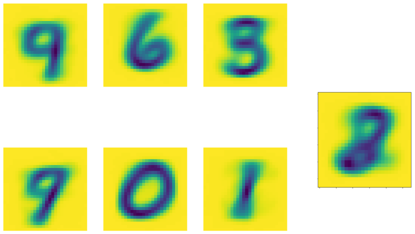
F1-Score: 0.321

Mean Images	**Cluster 1 – Label (1)**	**Cluster 2 – Label (3)**
		
Cluster 3 – Label (7)	**Cluster 4 – Label (0)**	**Cluster 5 – Label (6)**
		

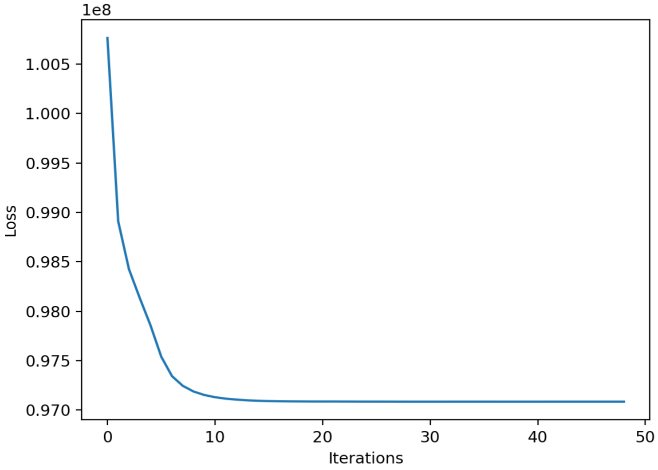
7 Clusters [48 iterations to converge]:

Accuracy: 0.566
F1-Score: 0.45

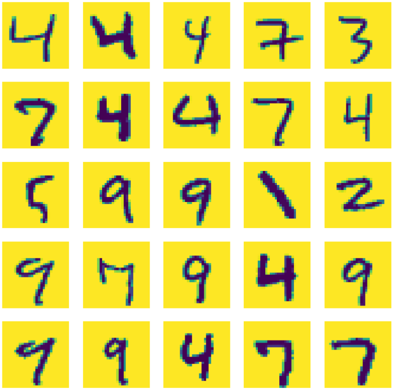
Mean Images



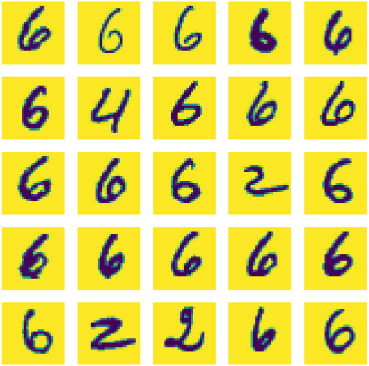
Objective vs Iterations



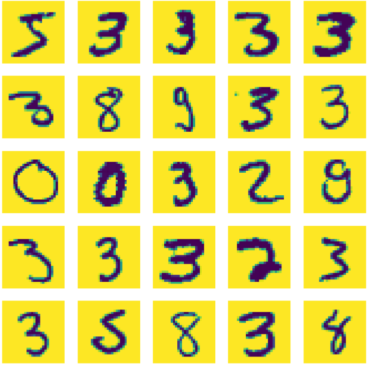
Cluster 1 – Label (4)



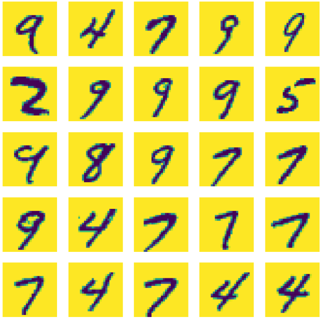
Cluster 2 – Label (6)



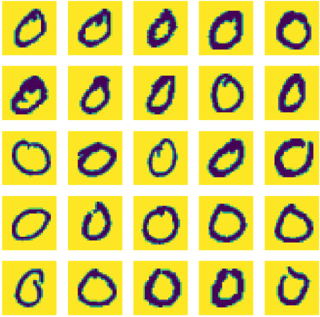
Cluster 3 – Label (3)



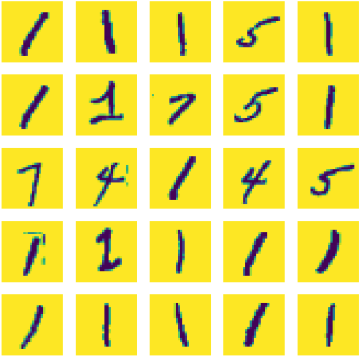
Cluster 4 – Label (7)



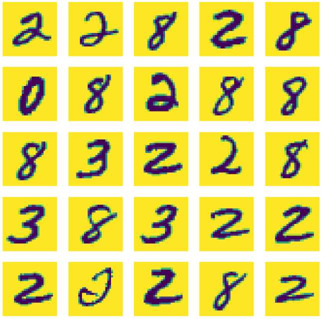
Cluster 5 – Label (0)



Cluster 6 – Label (1)



Cluster 7 – Label (2)

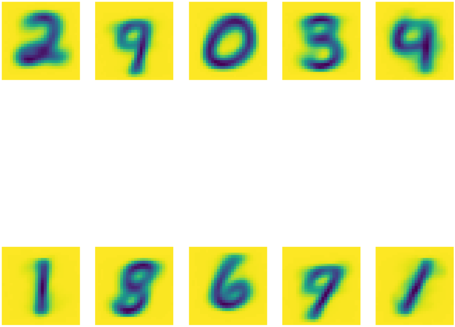


Mnist k = 10
Iterations=50
Acc = 0.604
F1 = 0.55

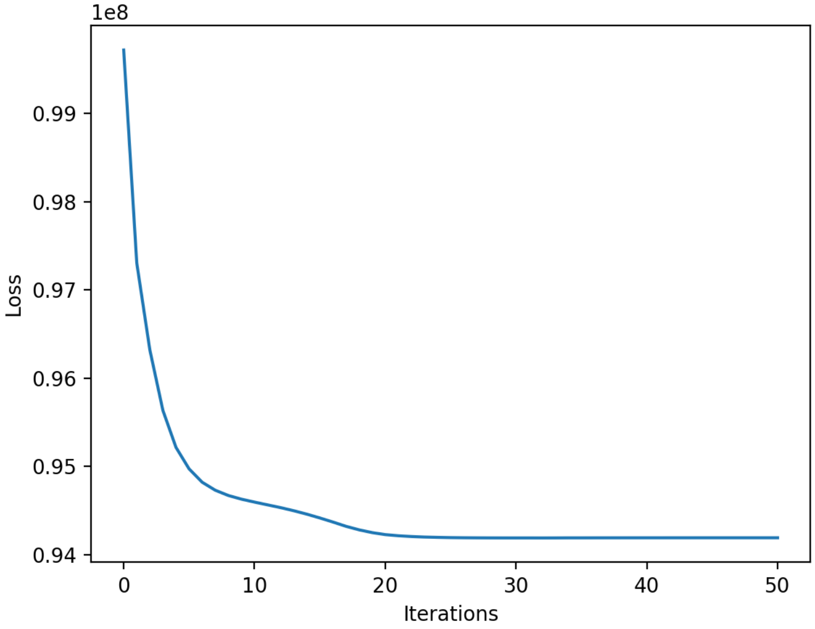
10 Clusters [50 iterations to converge]:

Accuracy: 0.604
F1-Score: 0.55

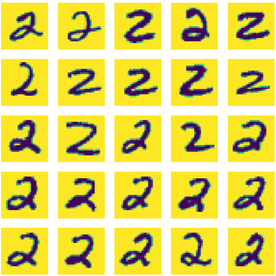
Mean Images



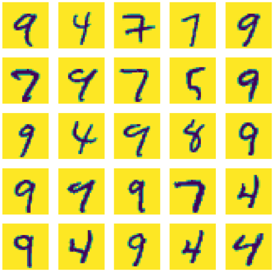
Iterations vs Objective



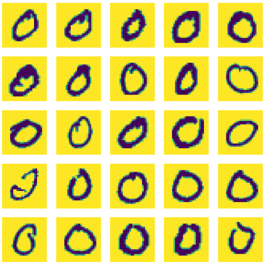
Cluster 1 – Label (2)



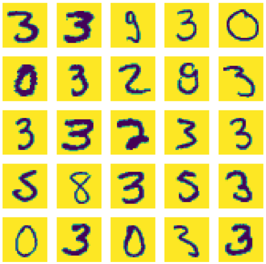
Cluster 2 – Label (9)



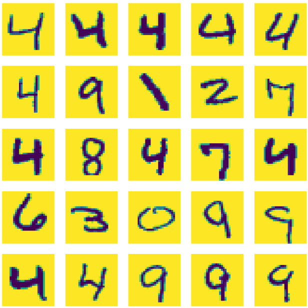
Cluster 3 – Label (0)



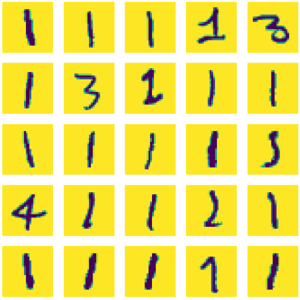
Cluster 4 – Label (3)



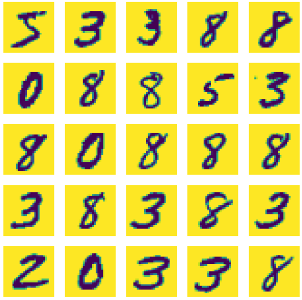
Cluster 5 – Label (4)



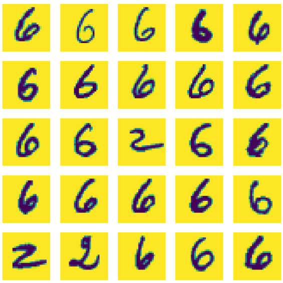
Cluster 6 – Label (1)

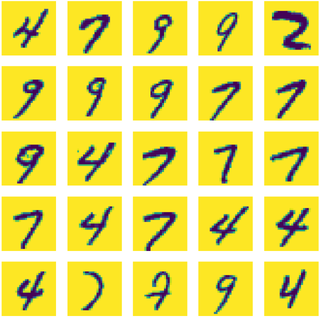
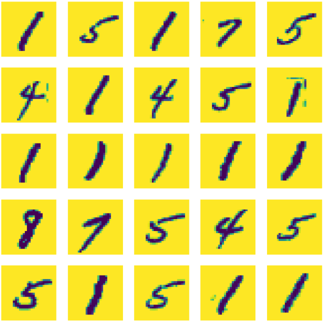


Cluster 7 – Label (8)



Cluster 8 – Label (6)

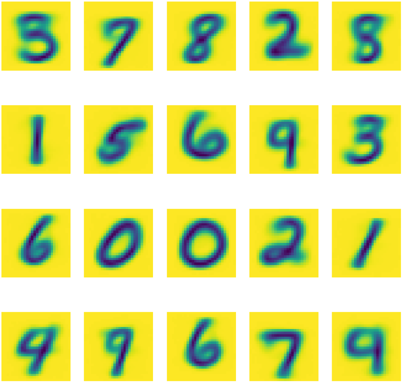
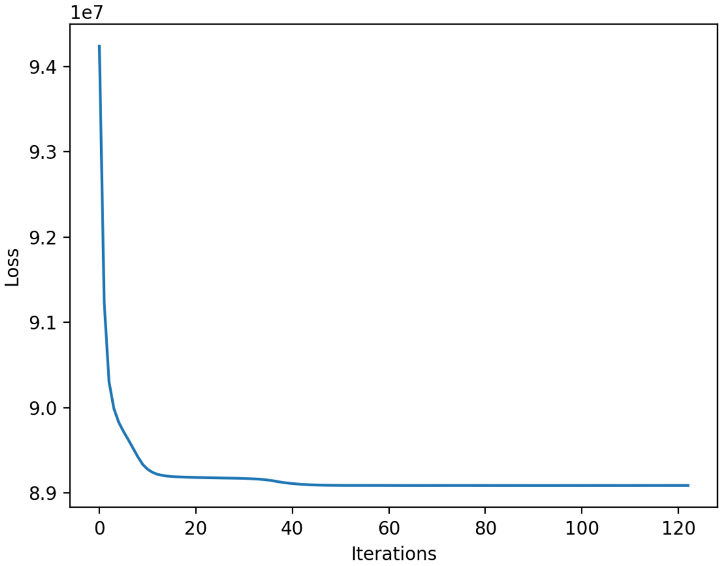


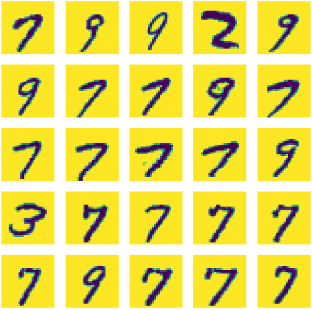
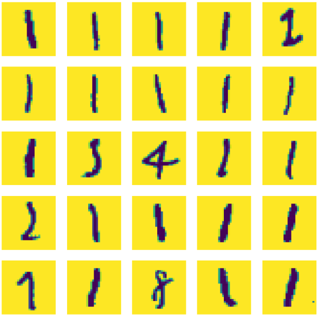
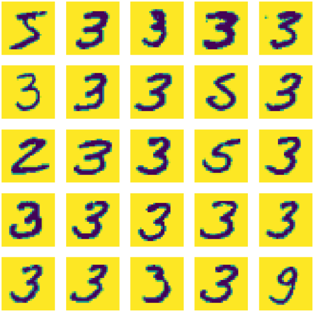
Cluster 9 – Label (7)	Cluster 10 – Label (1)
	

20 Clusters [122 iterations to converge]:

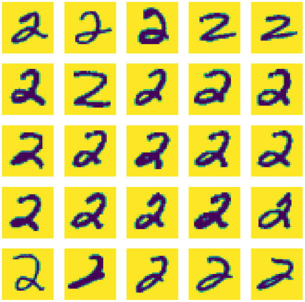
Accuracy: 0.726

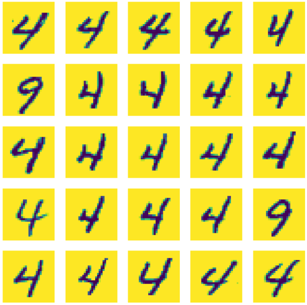
F1-Score: 0.70

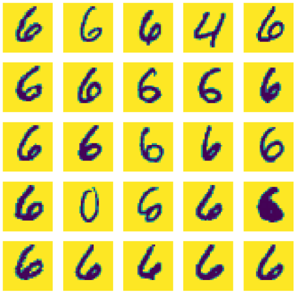
Mean Images	Iterations vs Objective
	

Cluster 2 – Label (7)	Cluster 6 – Label (1)	Cluster 10 – Label (3)
		

Cluster 14 – Label (2) | **Cluster 16 – Label (4)** | **Cluster 18 – Label (6)**







CIFAR-10

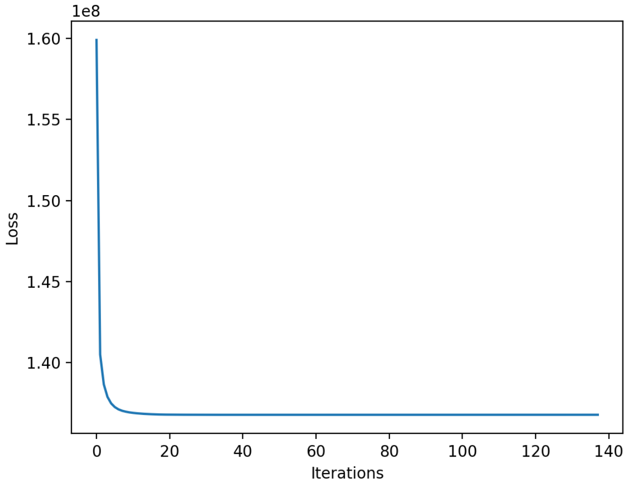
10 Clusters [137 iterations to converge]:

Accuracy: 0.236

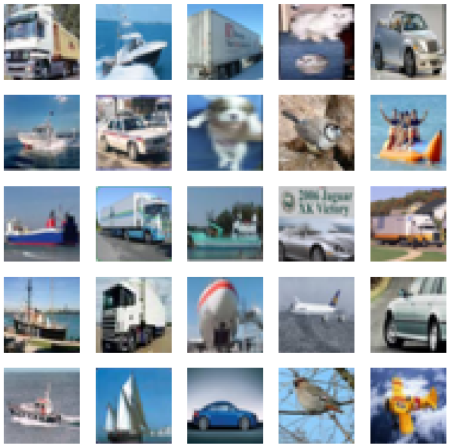
F1-Score: 0.182

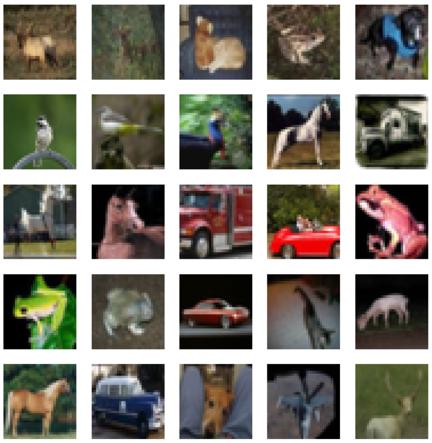
Mean Images

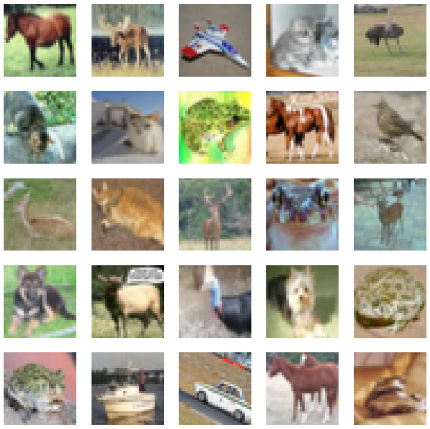
Iterations vs Objective



Cluster 2 – Label (Ship) | **Cluster 4 – Label (Frog)** | **Cluster 1 – Label (Dog)**







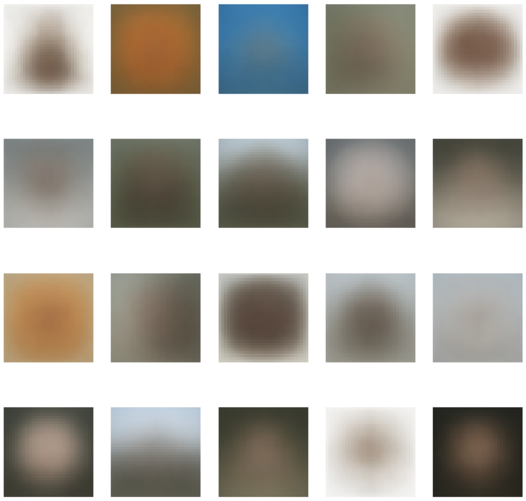
CIFAR-100

20 Clusters [140 iterations to converge]:

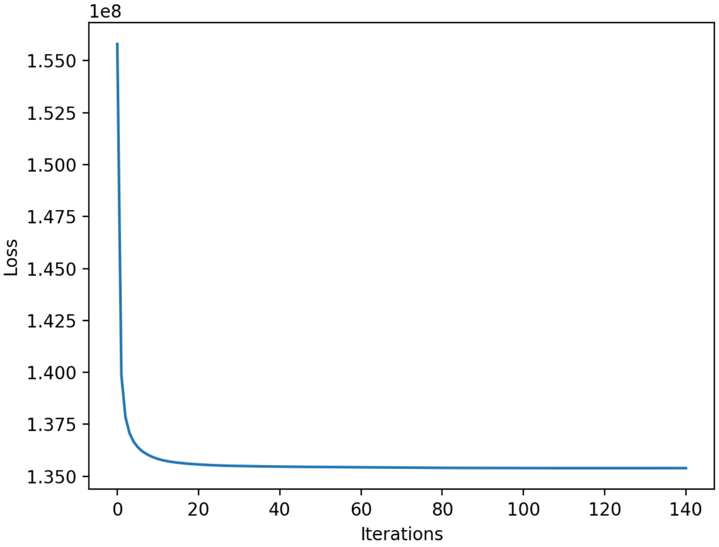
Accuracy: 0.15

F1-Score: 0.12

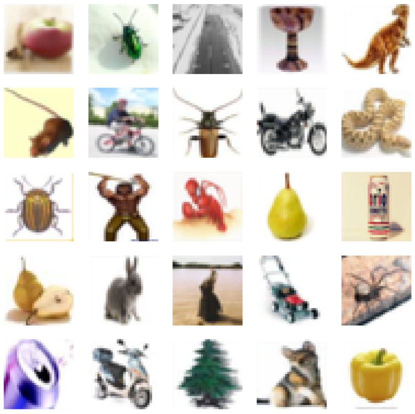
Mean Images



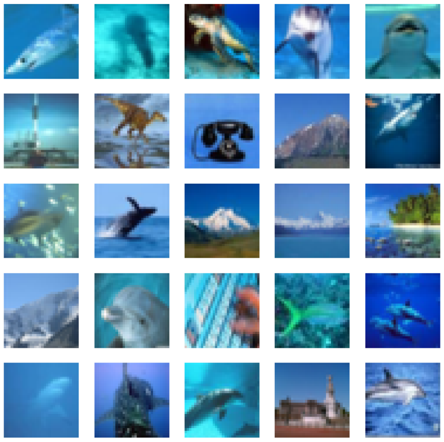
Iterations vs Objective



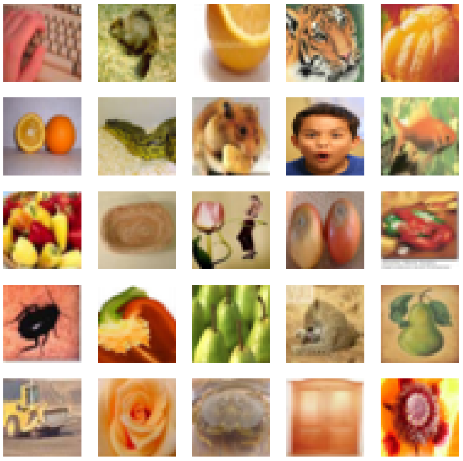
Cluster 1 – Label (Vehicles-2)



Cluster 3 – Label (Aquatic Mammals)



Cluster 11 – Label (Fruits and Vegetables)



Conclusion

As seen in all iterations vs objective plots, the objective never increases and always decreases till it convergres.

K-Means isn't good for image clustering, as the dimensionality is just too high, also with complex images it just doesn't give results. Maybe some dimensionality reduction should be done would give a better result.