PROJECT 5

Michael Encinas
November 9 of 2024
Project 5

Objectives

1. *Apply K-Means clustering algorithm to real data*
2. *Analyze and optimize the parameters of the K-Means*
3. *Analyze and compare the clustering results.*

Problem 1 (100 points)

a) Use the *K-Means* algorithm to cluster the provided data. Vary the number of clusters from 5 to 15 and select the optimal number. Justify your choice based on the **SSE vs. No. clusters plot**.

b) Using the number of clusters selected in (a), generate the silhouette plot.

c) Using the silhouette coefficients, identify 5 samples that are at the core of each cluster and 2 samples that are at the boundary of any two clusters (if they exist). Display the original images associated with these samples and comment on the results.

d) Compute the adjusted rand index by comparing the generated clusters to the provided ground truth (**this should be the only time you use the ground truth**).

Intro

I will be using the K-Means Algorithm to cluster my data. This is my first attempt at unsupervised learning. I will vary the number of clusters from 5 to 15 and based on my results from SSE vs No. Clusters plot I will then select the optimal number of clusters.

Once I have my number of optimal solutions selected as my parameter, I will generate a silhouette plot. Then I will use the silhouette coefficients and identify 5 samples in each cluster that are at the core as well as 2 samples that are at the boundary of any two clusters. I will display these images and comment on my observation and results.

I will then compute the adjusted rand index by comparing my generated clusters to the provided ground truth.

Dataset

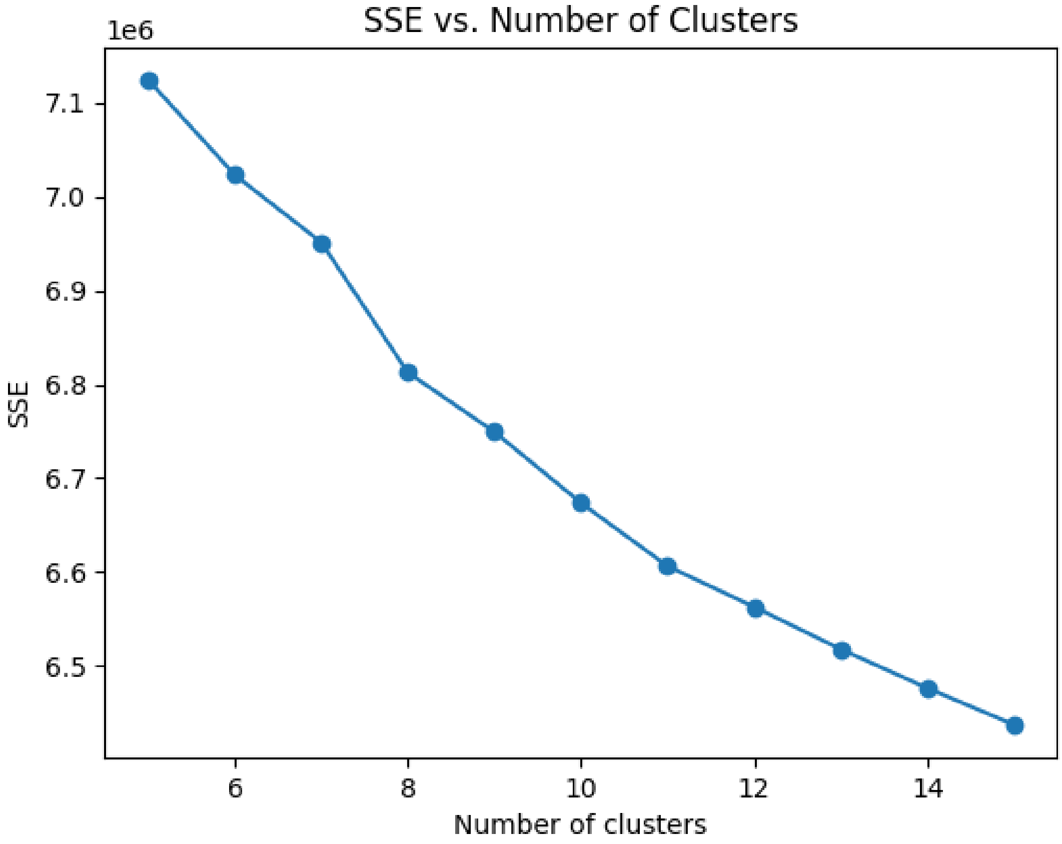
Data was given to me by professor and its provided data that contains 10k images in which has 5 classes and 128 features. Classes being 'plane', 'car', ' bird', 'horse' and 'ship'. Features are extracted from the 10k images using a CNN.

Data is from CIFAR-10 dataset and I was given three files, X.csv, y.csv and images.csv

SSE vs. Number of Clusters

I vary my numbers of clusters(k) from 5 to 15. I will be using K means while varying the K and observing the SSE vs. Number of Clusters. SSE is the sum of squared errors which measures the total variance within the clusters from where each data point is away from the centroid of its assigned cluster. The lower the SSE the more aligned your cluster should be. I will be checking the 'elbow' of my number of clusters vs SSE to decide on what K I will be choosing. The 'elbow' is just an observation where my SSE vs number of clusters starts to gradually change instead of just drop off over time.

Figure 1

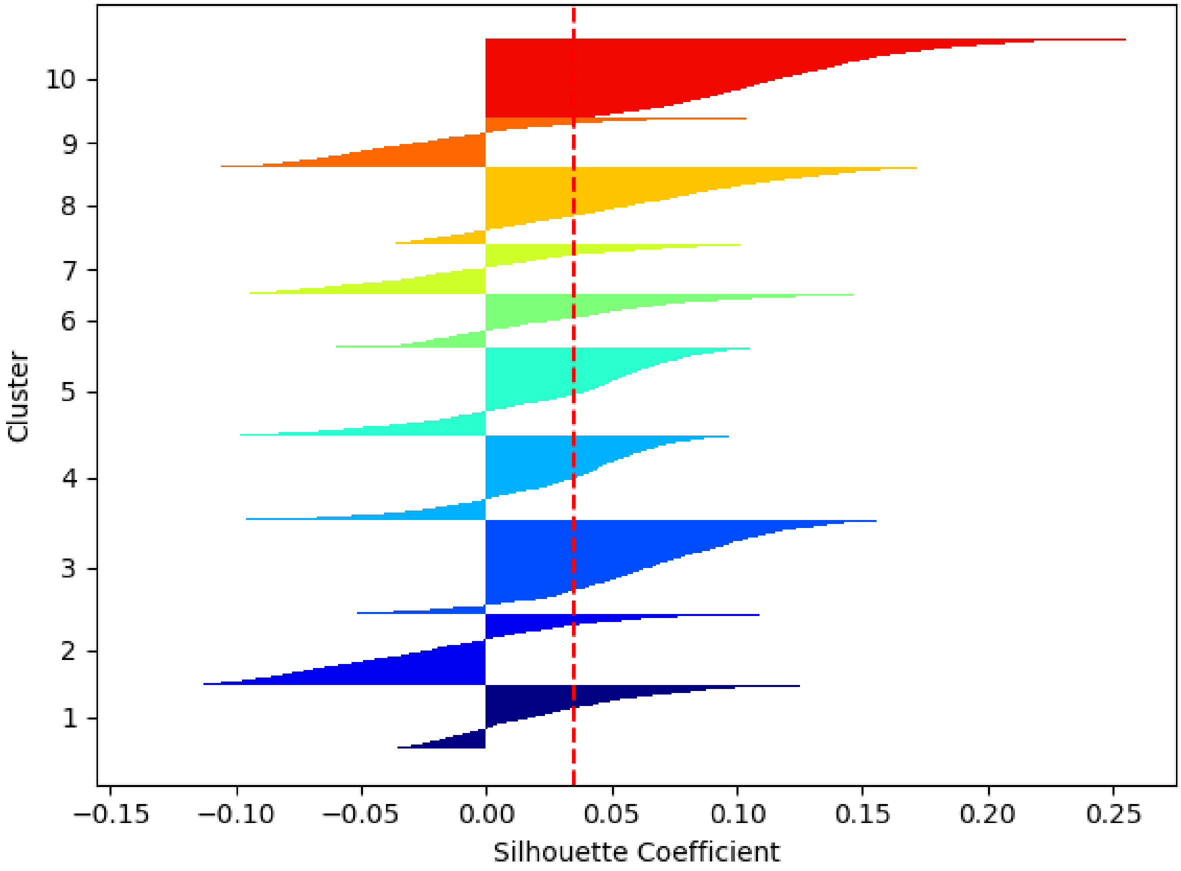


Based on Figure 1 when number of clusters is lower from 5 to 10, I see larger drops on SSE when clusters are iterated. There is drops off at least .1 per iteration. From cluster 11 to 15 to drops are now gradually and not as must lost in SSE. SSE does lower as clusters increase it is more applicable to lower your clusters. I selected cluster 10 since it appears that's the last big drop but looking back, I could have selected maybe 11 since it could elbow there.

Silhouette

Silhouette scores should range between -1 to 1 while 1 is a stronger pull towards being in the right cluster while -1 is the opposite.

Figure 2



Based on this score. My clusters don't seem to be well separated from each other due to average silhouette score being .04. Also, my red cluster is the most well separated since my scores are higher than the others but at its lowest coefficient score which is the boundary samples might be more like other clusters than the one it's in. Clusters 1 and 2 have low correlations which tells me that those clusters might not be the right fit for the data.

I could improve these scores most likely if I lower my K to 7 or 8 so I can have less clusters which might improve my Silhouette coefficient scores.

Core Samples - Boundary Samples

I displayed my 9 clusters with randomly selected 5 core samples and 2 boundary samples. Samples are displayed with its image. I printed out a pdf file that contains my images.

Figure 3

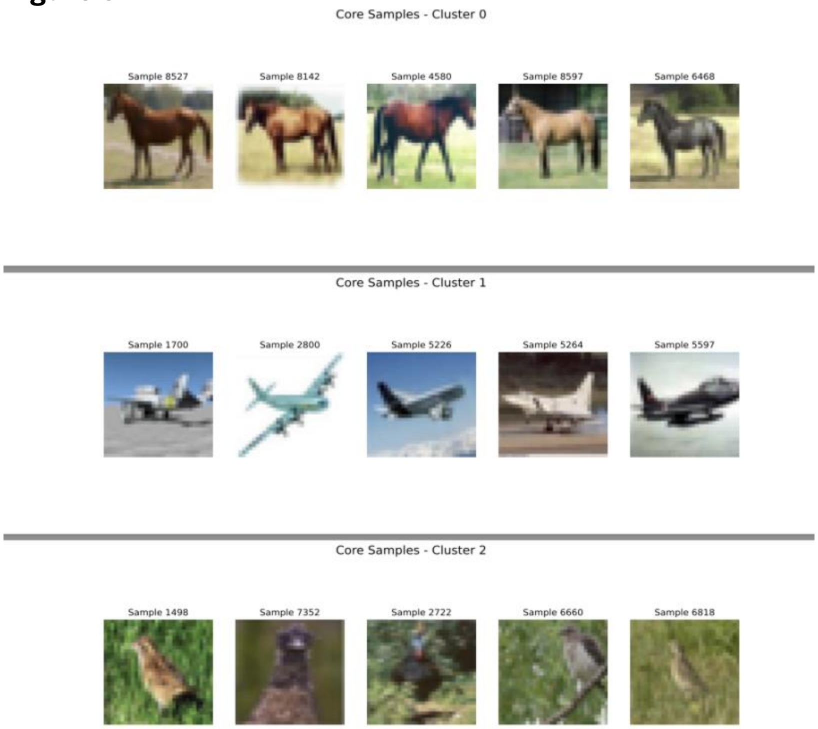


Figure 4

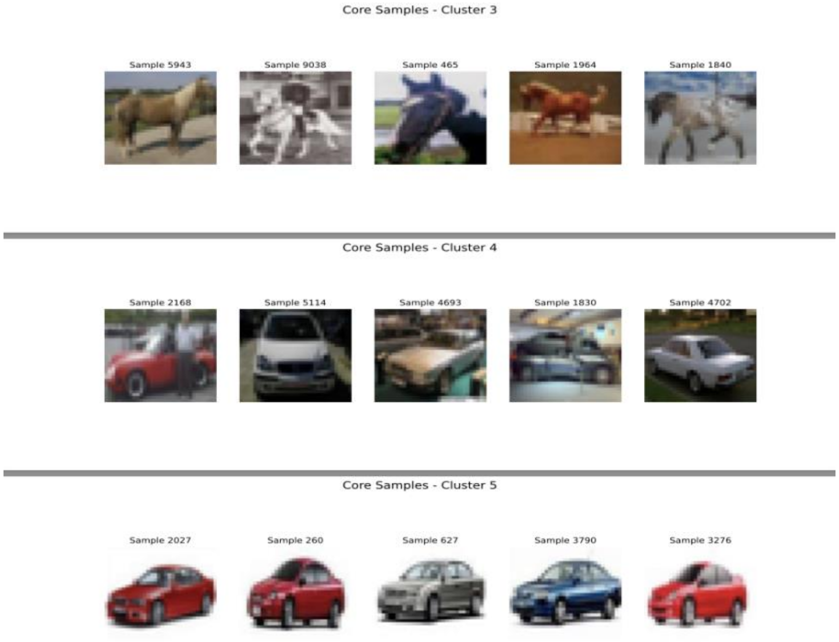


Figure 5

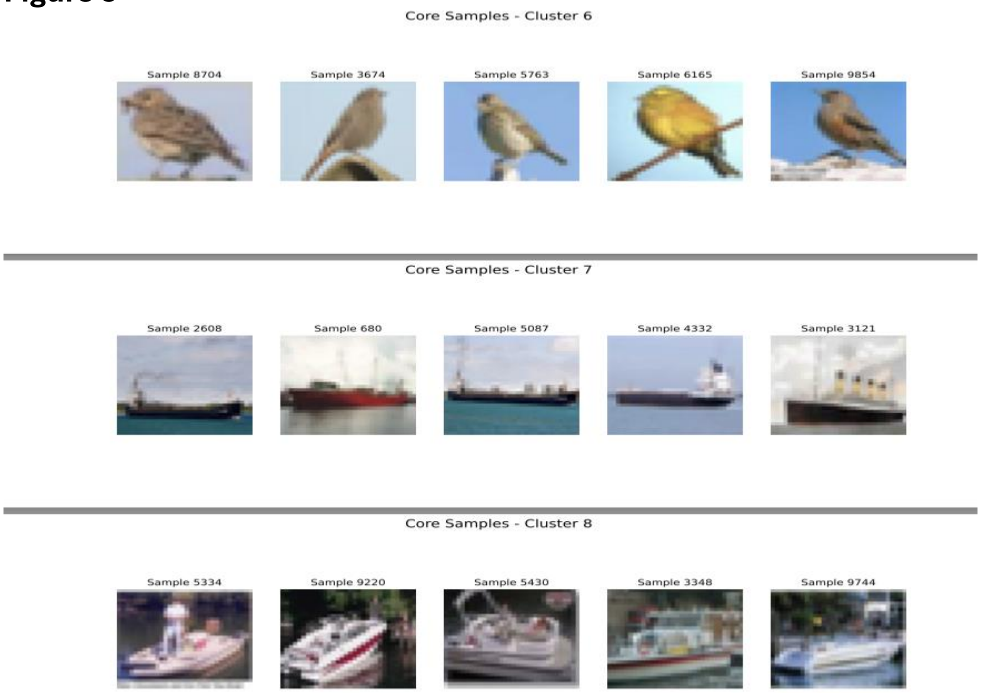


Figure 6

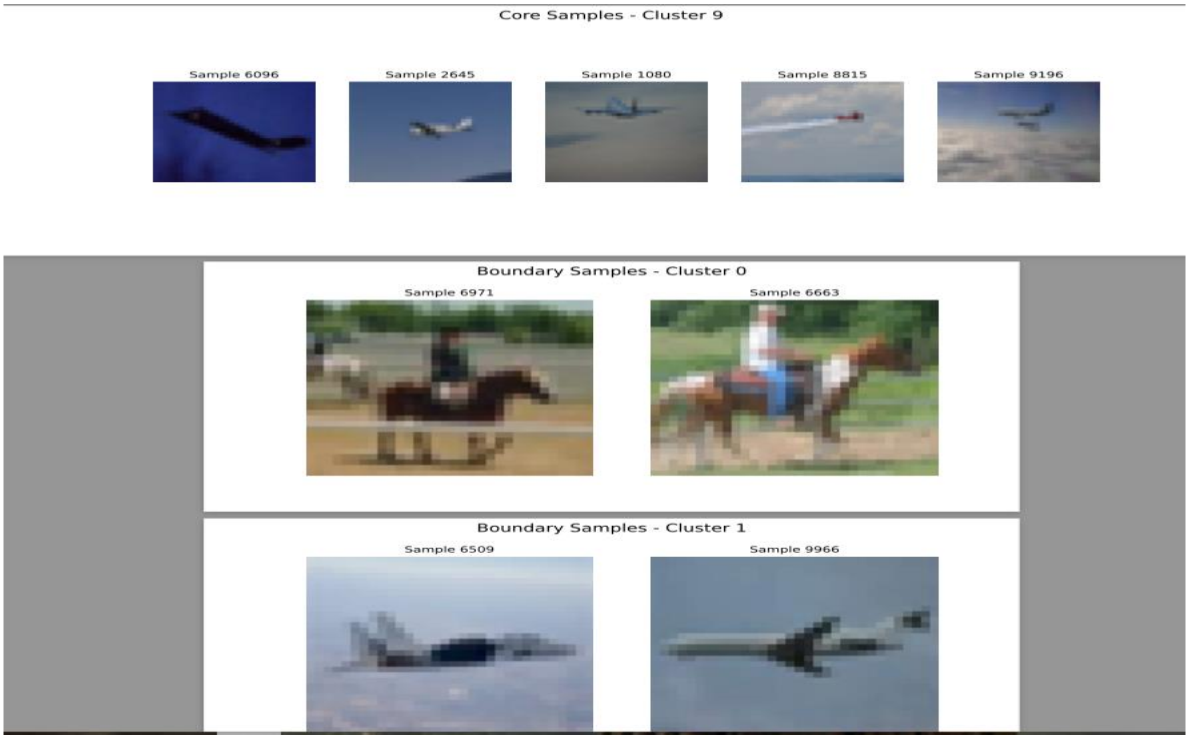


Figure 7

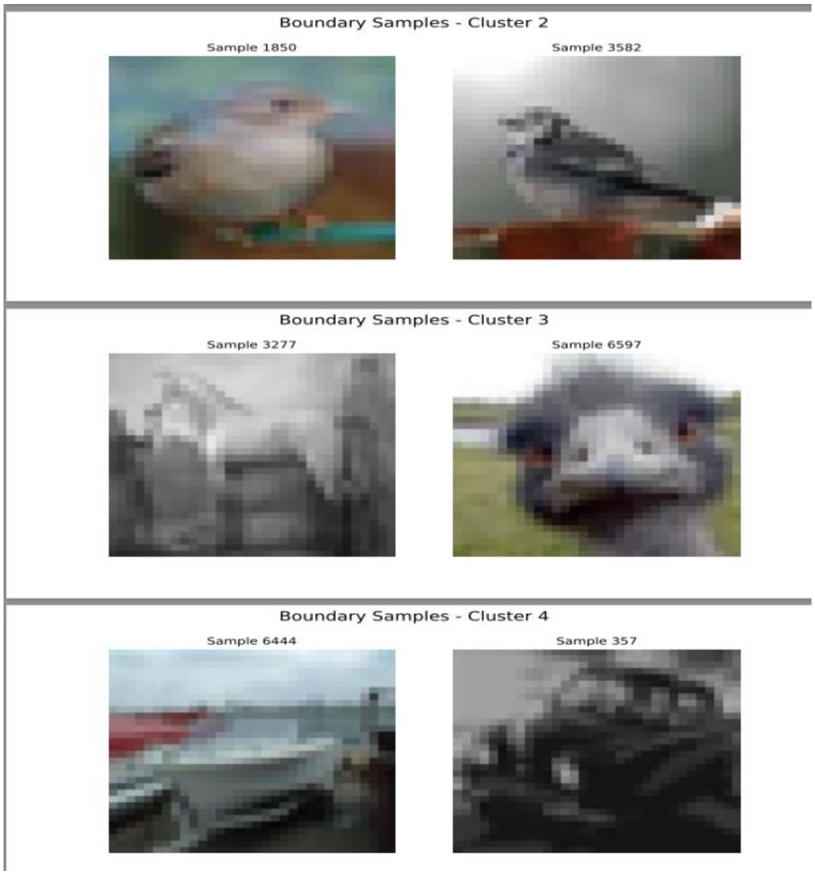

Boundary Samples - Cluster 2
Sample 1850
Sample 3582
Boundary Samples - Cluster 3
Sample 3277
Sample 6597
Boundary Samples - Cluster 4
Sample 6444
Sample 357

Figure 8

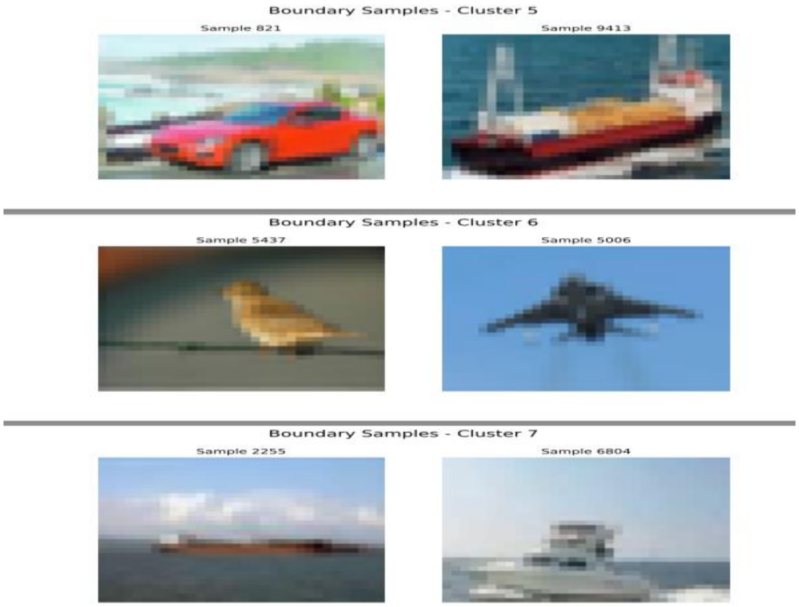
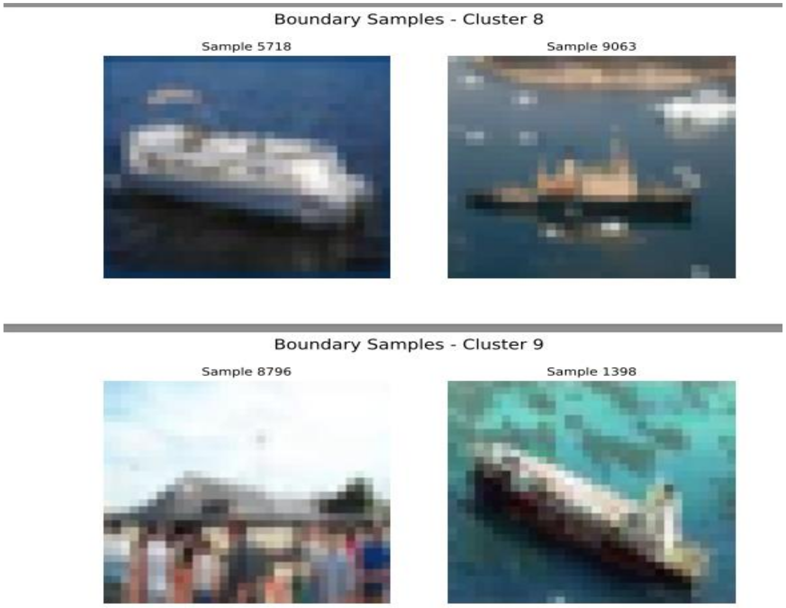

Boundary Samples - Cluster 5
Sample 821
Sample 9413
Boundary Samples - Cluster 6
Sample 5437
Sample 5006
Boundary Samples - Cluster 7
Sample 2255
Sample 6804

Figure 9



Cluster core samples
Cluster 0 - Horses
Cluster 1 - Airplanes
Cluster 2 - Birds
Cluster 3 - Horses
Cluster 4 - Cars
Cluster 5 - Cars
Cluster 6 - Birds
Cluster 7 - Boats
Cluster 8 - Boats
Cluster 9 -Airplanes

Cluster 0 appears to be the same as cluster 3 since horses are in both pictures in core samples. Cluster 4 and Cluster 5 appear to be the same clusters since its both cars. Cluster 2 and Cluster 6 appear to be the same as well on core samples since images show birds. Cluster 7 and cluster 8 appear to always have core samples from boat images. Cluster 1 and Cluster 9 have its core images as displayed as a plane.

Just off the results above based on the core samples above we can make our clusters above maybe 5 since I have 10 clusters here but with 5 I can get them potentially all together since each cluster shares with another cluster. However, just based on looking at core samples, Cluster 7 and Cluster 8 are boats but one is big boats and the other is small boats. Cluster 1 and Cluster 9 might be different because airplanes appear to be far away and the other cluster group the planes are closer in image. In cluster 5 and Cluster 4, Cars in one image are just the car by itself with no background and in Cluster 4 the image is taking in real live with background to it. The birds in Cluster 2 and Cluster 6 could be because of the light color of the birds in one cluster compared to the other. For the horse clusters It could be that green grass in the one image might make it distinct compare to the other horse images where its dirt, snow or some other image shoot.

Boundary Samples

Now for Cluster 0 the images appear to be horses with a human on top which based on Cluster 3 for the core sample there is a human on top of a horse so that could be the link where if the clusters were shortened and SSE was larger it might grab those boundary images into another group. Cluster 7 and Cluster 8 boundary samples are not small or large but in the middle in terms of boat size which might confuse the cluster. It clusters 4 and 5 for the boundary images it appears some of the car and boats are in the same angle which might confused the images to mistake a boat for car and car for boat. Some of the birds in Cluster 6 are shown to be airplanes because shape does appear similar.

Based on everything above I could make my cluster numbers shortened and get those boundary samples maybe in better cluster groups by limiting my cluster number since some of them can be confused.

Adjusted Rand Index (ARI)

Adjusted Rand Index (ARI): 0.48704916264662157
This was my ARI score which ranges from -1 to 1.
-1 is worse than random clustering, 0 is no better than random assignments and 1 is perfect agreement between predicted clusters and ground truth labels.

Since my score is .487 its between random and great prediction. This is not a bad score but I think if I lower my clusters than I can get this score up.

Video Link: https://youtu.be/gsSdS3YA3PQ