# COMPENG 4SL4

# Assignment 5

Instructor: Dr. Dumitrescu

Hritheekka Chinnakonda – chinnakh – 400292782 – C01 – L03

As a future member of the engineering profession, the student is responsible for performing the required work in an honest manner, without plagiarism and cheating. Submitting this work with my name and student number is a statement and understanding that this work is my own and adheres to the Academic Integrity Policy of McMaster University and the Code of Conduct of the Professional Engineers of Ontario. Submitted by [Hritheekka Chinnakonda, chinnakh, 400292782]

## Two initialization strategies (method explanation):

**1) Randomly pick the centers from the data points (use this two times):**

The 'random_centers' strategy in k-means clustering starts the algorithm by randomly (no prioritized data points, randomly chosen without replacement) selecting 'k' data points from the dataset as initial centroids. This method could result in centroids being close together initially, which could affect the convergence speed and the quality of the final clusters. This method is simple and efficient in handling larger datasets.

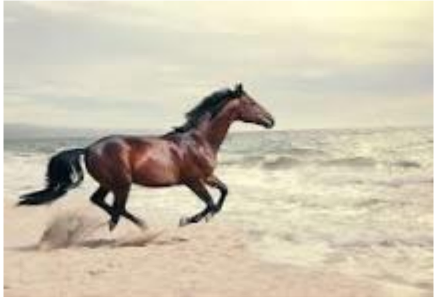
**2) Pick the centers such that they have a sufficiently large distance between them:**

The 'max_distance' strategy sets the initial centroids in a way that maximizes their spatial separation. It starts by randomly selecting the first centroid from the dataset, and identifies other centroids based on their maximum distance from the chosen centroid. It iterates and picks the data points that are the furthest from the nearest centroid, consequently ensuring that the initial centroids are well-distributed and far apart. This strategy allows the k-mean algorithm to converge to a more optimal cluster representation.

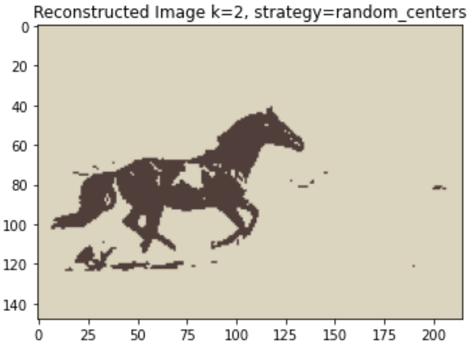
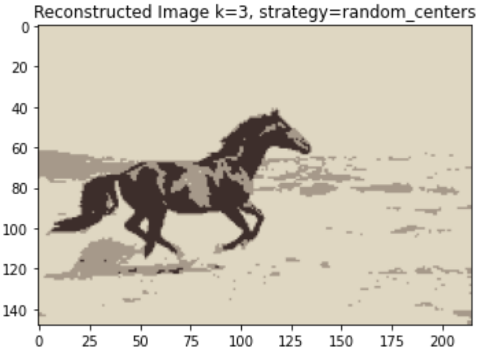
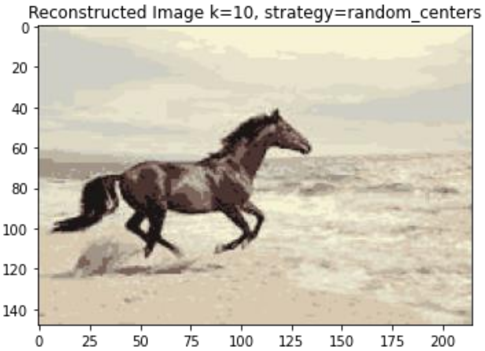
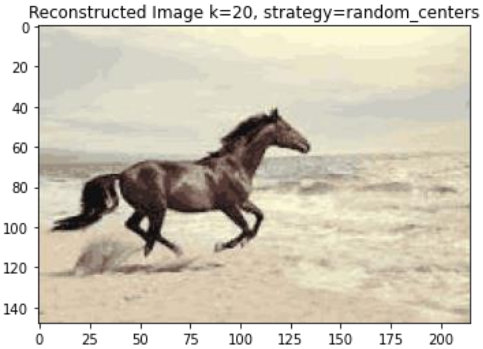
For each image below, the clustering algorithm was run for k = 2, 3, 10, 20, 40 until convergence, with a different initialization strategy (the first strategy run twice, and the second run once). The results for two colour images are seen below.

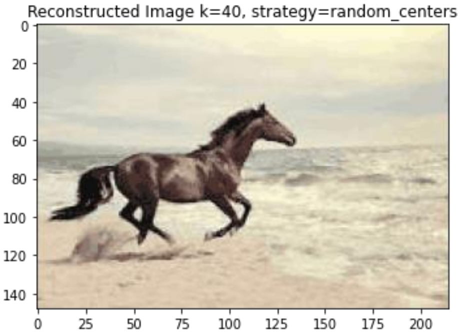
## Image 1: ‘horse.jpg’

Original Image:

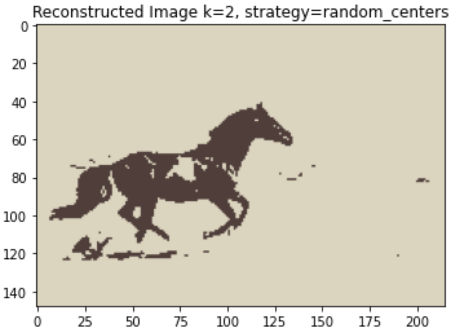
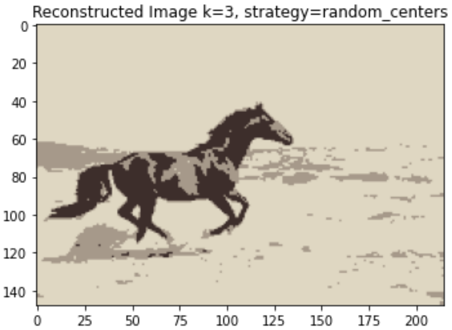
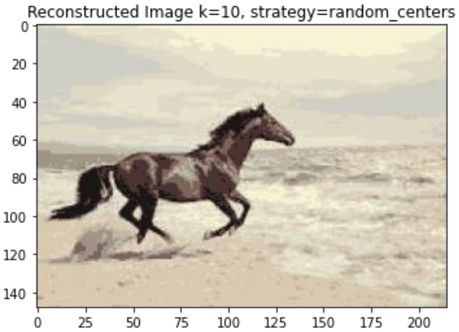


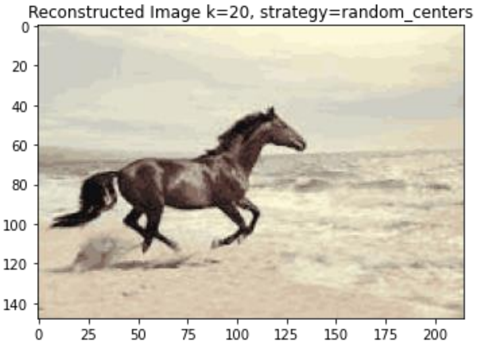
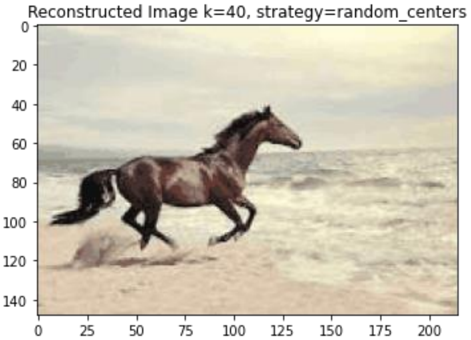
1. random_centers
    a. Trial 1:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 471.18340539890113 | 8 |  |
| 3 | 271.13973813195423 | 34 |  |
| 10 | 55.53956673065252 | 69 |  |
| 20 | 29.99692800875767 | 166 |  |

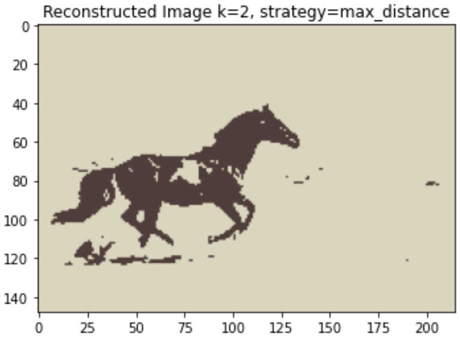
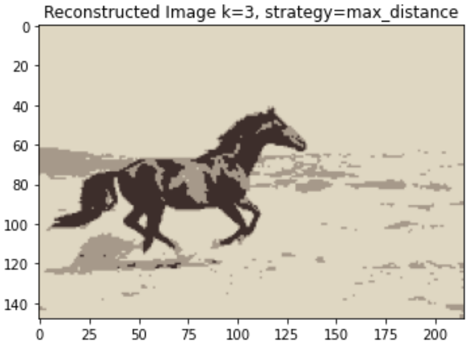
|  |  |  |  |
|---|---|---|---|
| 40 | 20.174049573253292 | 171 |  Reconstructed Image k=40, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |

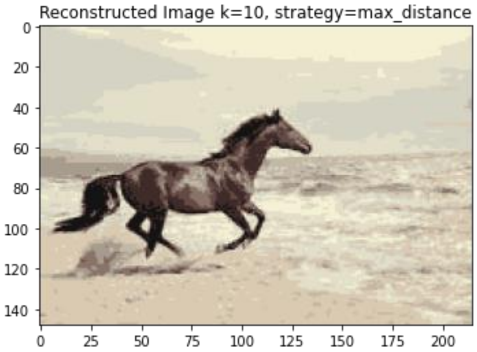
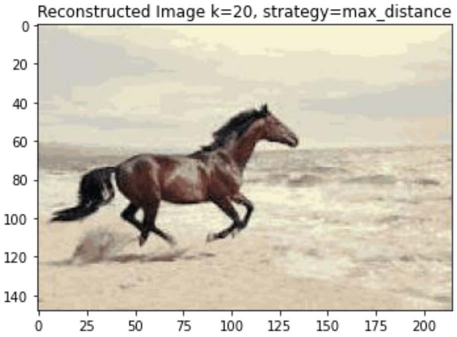
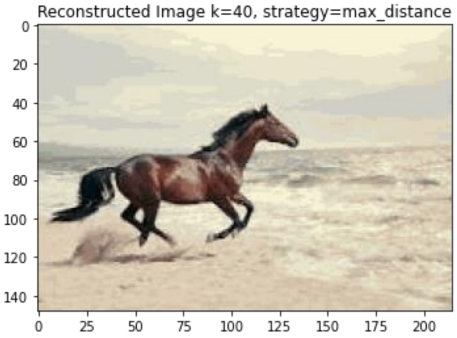
b. Trial 2:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 471.18340539890113 | 9 |  Reconstructed Image k=2, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |
| 3 | 271.13956526956486 | 9 |  Reconstructed Image k=3, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |
| 10 | 55.53956673065252 | 65 |  Reconstructed Image k=10, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |

| | | | |
|---|---|---|---|
| 20 | 29.762638159422508 | 199 |  Reconstructed Image k=20, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |
| 40 | 19.004637816442372 | 199 |  Reconstructed Image k=40, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |

c. max_distance:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 471.18340539890113 | 9 |  Reconstructed Image k=2, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |
| 3 | 271.13973813195423 | 36 |  Reconstructed Image k=3, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |

| | | | |
|---|---|---|---|
| 10 | 54.02838205501795 | 147 |  Reconstructed Image k=10, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |
| 20 | 31.130809013680178 | 199 |  Reconstructed Image k=20, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |
| 40 | 23.38612065497985 | 199 |  Reconstructed Image k=40, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175<br>200 |

For each k the following strategy was better based on the MSE:

*k = 2: Strategy 1 (Trial 1)*

*k = 3: Strategy 1 (Trial 2)*

*k = 10: Strategy 2*

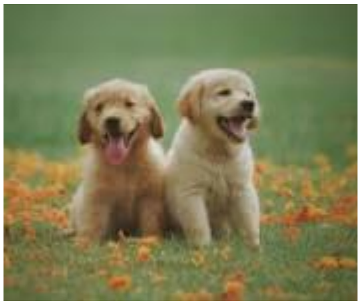
*k = 20: Strategy 1 (Trial 2)*

*k = 40: Strategy 1 (Trial 2)*

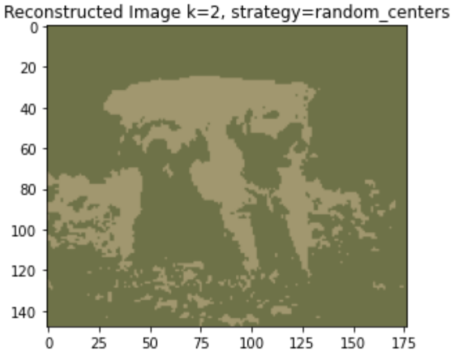
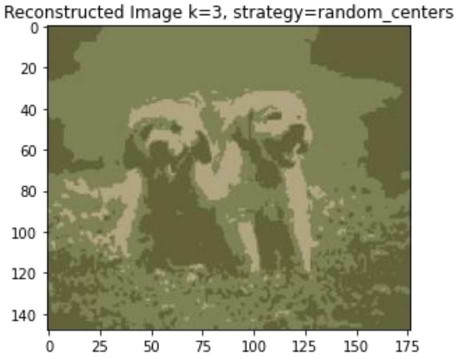
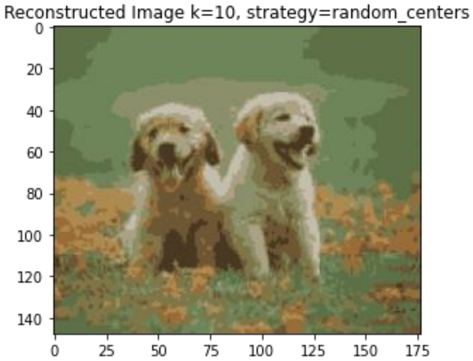
The second trial of the first strategy produces better results. The visual reconstruction of each k is extremely similar and difficult to judge in terms of choosing a better initialization strategy.

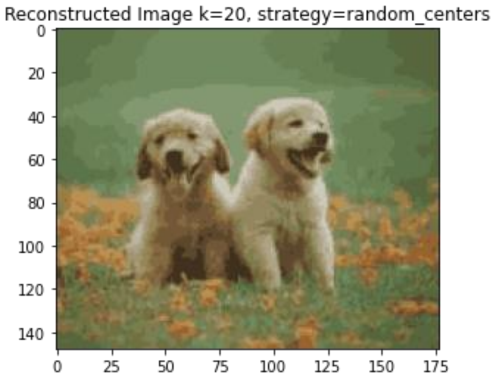
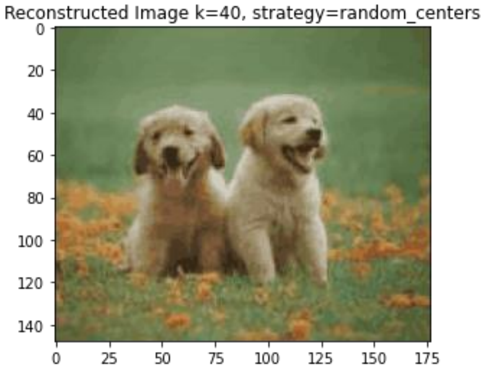
## Image 2: ‘dogs.jpg’

Original Image:

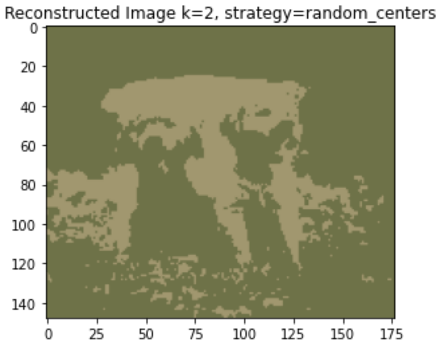
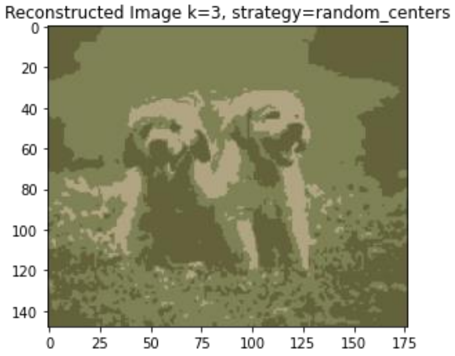


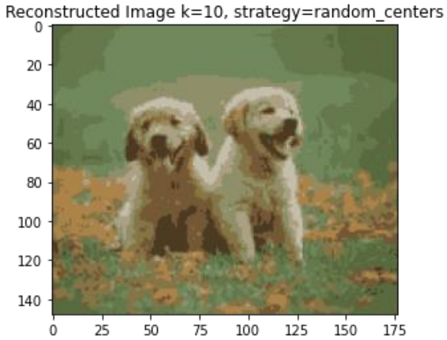
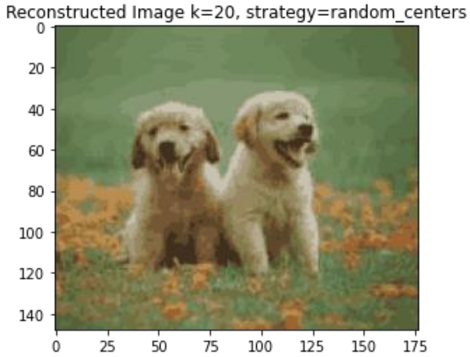
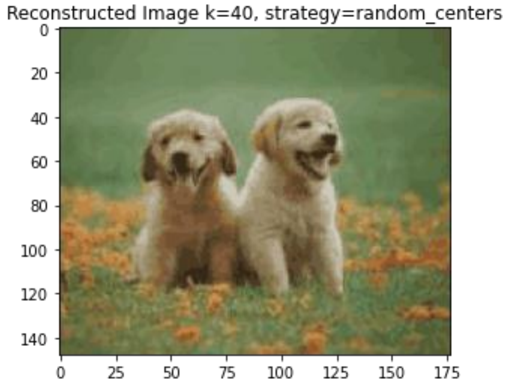
1. random_centers
   a. Trial 1:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 434.0003361222744 | 17 |  |
| 3 | 301.93501864905653 | 37 |  |
| 10 | 85.97552339460118 | 51 |  |

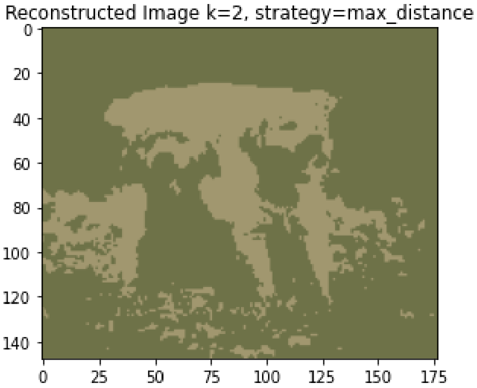
|  |  |  |  |
|---|---|---|---|
| 20 | 44.51547000039864 | 71 | <br> Reconstructed Image k=20, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 40 | 27.075272777791074 | 94 | <br> Reconstructed Image k=40, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |

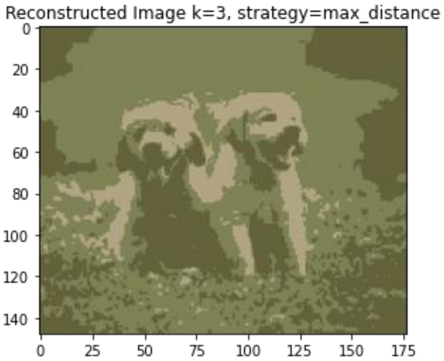
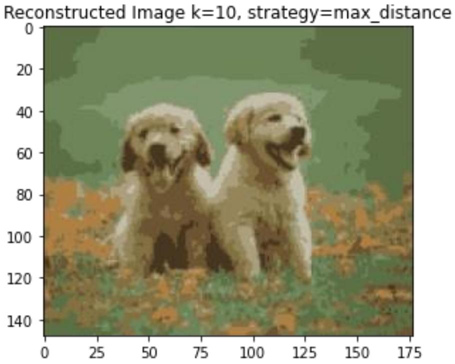
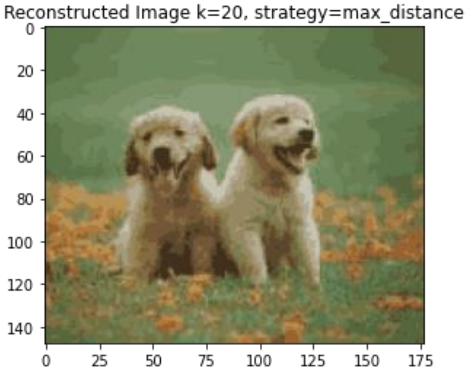
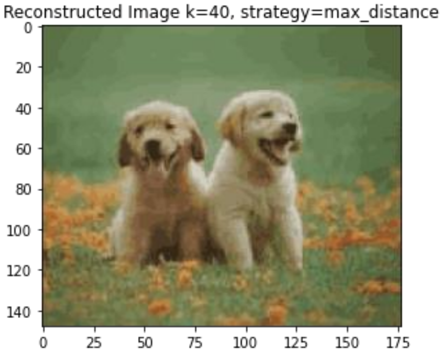
b. Trial 2:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 434.0003361222744 | 23 | <br> Reconstructed Image k=2, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 3 | 301.93501864905653 | 36 | <br> Reconstructed Image k=3, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |

|  |  |  |  |
|---|---|---|---|
| 10 | 86.44827574084721 | 55 |  Reconstructed Image k=10, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 20 | 45.824075925849165 | 42 |  Reconstructed Image k=20, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 40 | 27.4888626557069 | 118 |  Reconstructed Image k=40, strategy=random_centers<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |

c. max_distance:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 434.0003361222744 | 19 |  Reconstructed Image k=2, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |

|  |  |  |  |
|---|---|---|---|
| 3 | 301.93501864905653 | 42 |  Reconstructed Image k=3, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 10 | 85.8788868661963 | 67 |  Reconstructed Image k=10, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 20 | 44.95401569635182 | 84 |  Reconstructed Image k=20, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |
| 40 | 27.3197537007335 | 133 |  Reconstructed Image k=40, strategy=max_distance<br>0<br>20<br>40<br>60<br>80<br>100<br>120<br>140<br>0<br>25<br>50<br>75<br>100<br>125<br>150<br>175 |

For each k the following strategy was better based on the MSE:

*k = 2: Strategy 1 (Trial 1)*

*k = 3: Strategy 1 (Trial 1)*

*k = 10: Strategy 2*

*k = 20: Strategy 1 (Trial 1)*

*k = 40: Strategy 1 (Trial 1)*

The first trial of the first strategy produces better results based on the MSE. The visual reconstruction of each k is extremely similar and difficult to judge in terms of choosing a better initialization strategy.

**The report should contain a discussion of the results: for each image and each k, which initialization strategy led to better clustering judging based on a) the MSE; b) the visual reconstruction? Does always a smaller MSE correspond to a more pleasing visual reconstruction? Is one initialization strategy better than the other all the time or almost all the time? Include any other observations you might find useful.**

For each image and each k, I found that both initialization strategies are effective and similar in the regards to the reconstructed image. The MSE is very similar for each k for both strategies as well. When analysing the MSE values and the reconstructed images, the conclusion can be drawn that a smaller MSE corresponds to a more pleasing reconstruction. The peak value for MSE was around k = 3 for each strategy and trial (MSE around 300) and decreased significantly for k = 10, 20 and 40, which resulted in significantly better reconstructions. In terms of a better initialization strategy, I found that the randomly chosen centers were faster to produce an output versus the strategy of choosing centers in which the distances were sufficiently large. The run time for the second strategy was extended and does not produce a significantly clearer/better reconstructed image to choose this strategy over the first. The second strategy overall had more iterations until convergence for each k value.

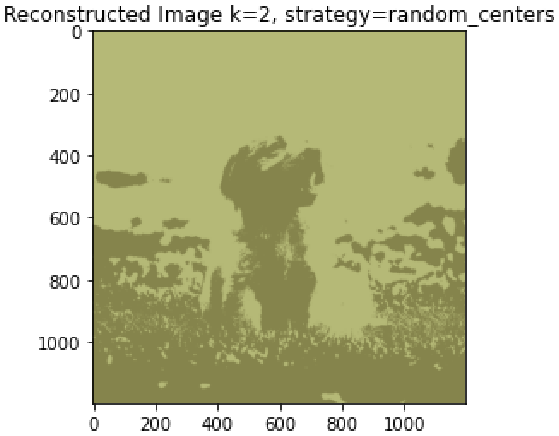
## Additional Testing:

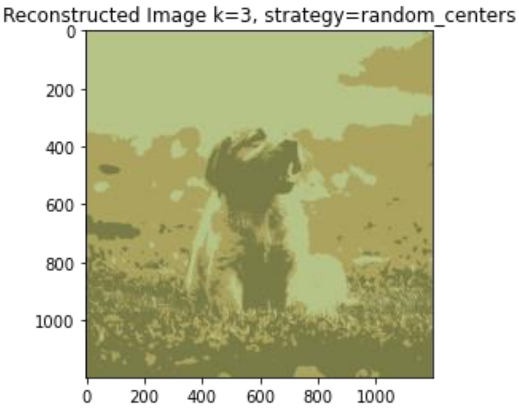
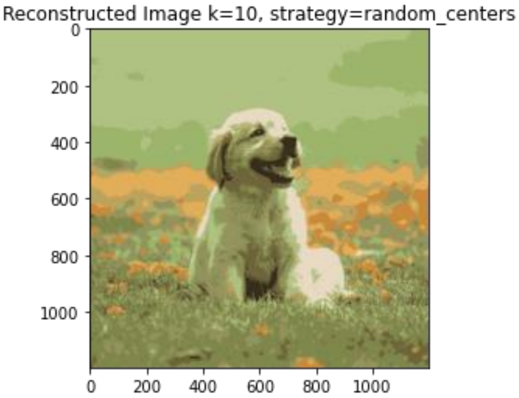
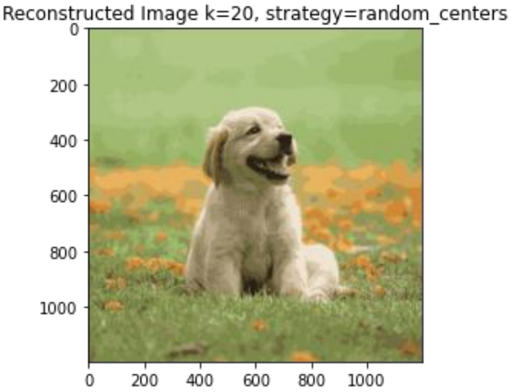
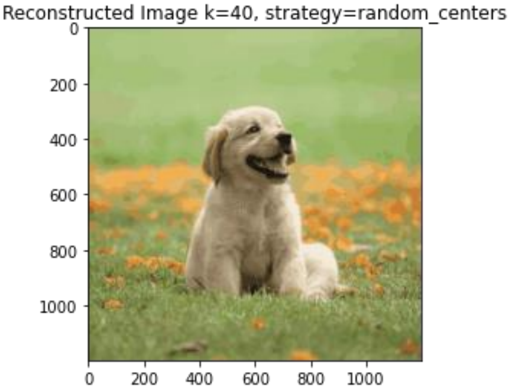
Initially, the two chosen images were of high resolution, the size of the picture was much larger. This resulted an extremely long run time for the code because there are many pixels/data points. This showed me that this code, especially the second initialization strategy is not very efficient for higher quality images. The results of these two images have been added below. Overall, the MSE values were larger with a lower k value, and the number of iterations were much higher especially for a larger image like 'lego.jpg'. The visual reconstruction of each k is extremely similar for both images and it is difficult to judge in terms of choosing a better initialization strategy.

## Image 3: 'dog.jpg'

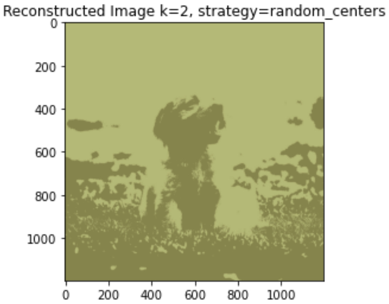
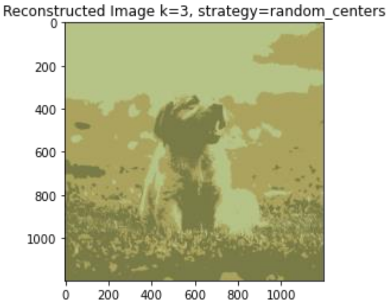
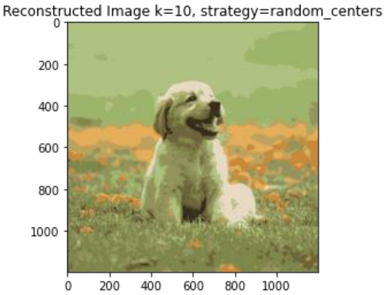
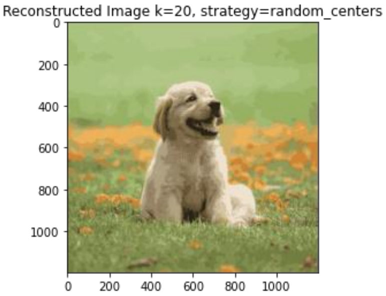
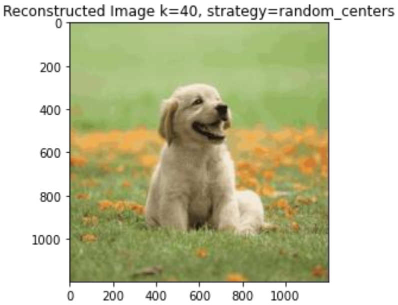
Original Image:

1. random_centers:
    a. Trial 1:

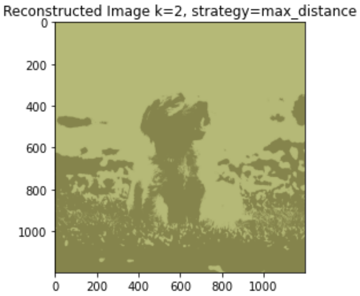
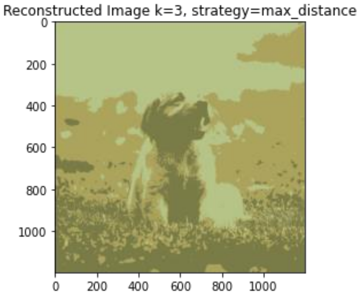
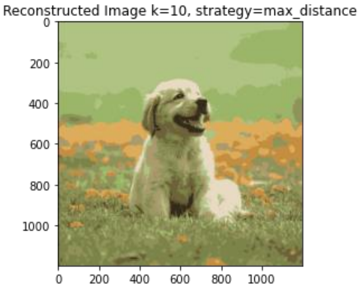
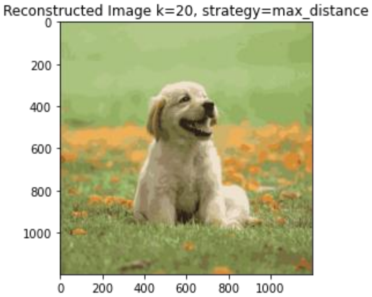
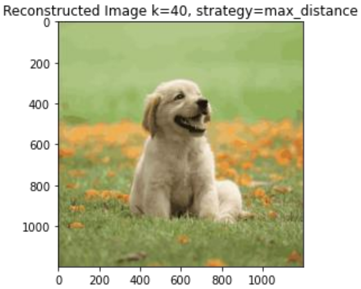
| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 583.2324494968353 | 13 |  |

| | | | |
|---|---|---|---|
| 3 | 452.9683637864333 | 16 |  Reconstructed Image k=3, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0 200 400 600 800 1000 |
| 10 | 116.67764349367474 | 58 |  Reconstructed Image k=10, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0 200 400 600 800 1000 |
| 20 | 64.70564672018222 | 145 |  Reconstructed Image k=20, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0 200 400 600 800 1000 |
| 40 | 38.04048870181048 | 199 |  Reconstructed Image k=40, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0 200 400 600 800 1000 |

b. Trial 2:

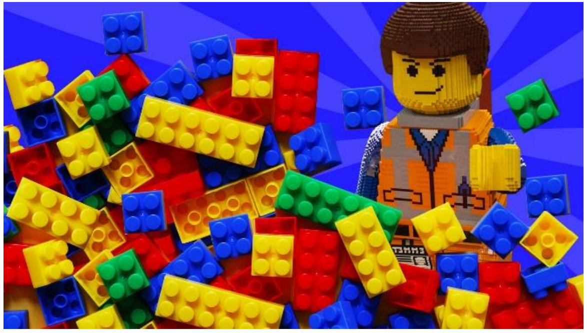
| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 583.2324505187506 | 11 |  Reconstructed Image k=2, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0<br>200<br>400<br>600<br>800<br>1000 |
| 3 | 452.96890570018917 | 40 |  Reconstructed Image k=3, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0<br>200<br>400<br>600<br>800<br>1000 |
| 10 | 116.67764349367474 | 106 |  Reconstructed Image k=10, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0<br>200<br>400<br>600<br>800<br>1000 |
| 20 | 62.95798035504031 | 84 |  Reconstructed Image k=20, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0<br>200<br>400<br>600<br>800<br>1000 |
| 40 | 37.42988937588162 | 199 |  Reconstructed Image k=40, strategy=random_centers<br>0<br>200<br>400<br>600<br>800<br>1000<br>0<br>200<br>400<br>600<br>800<br>1000 |

c. max_distance:

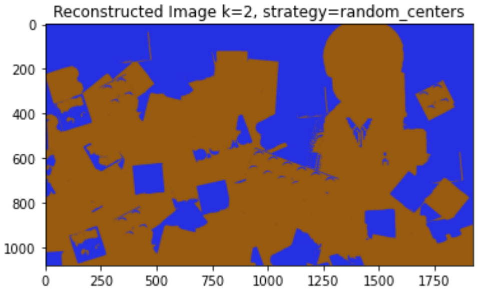
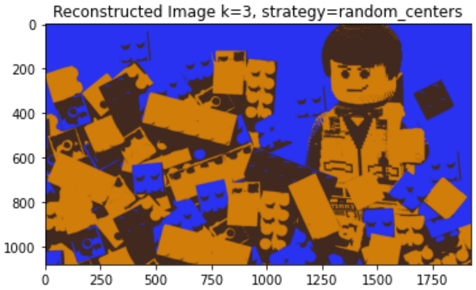
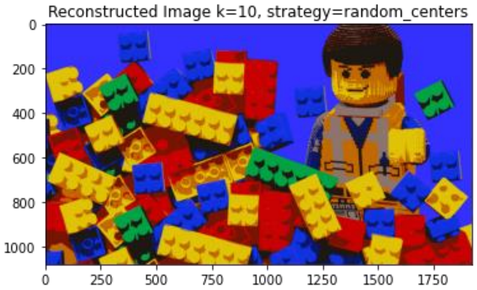
| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 583.2324494968353 | 11 |  |
| 3 | 452.96890570018917 | 69 |  |
| 10 | 117.92706244977819 | 94 |  |
| 20 | 64.47874878254204 | 199 |  |
| 40 | 37.574107040516786 | 199 |  |

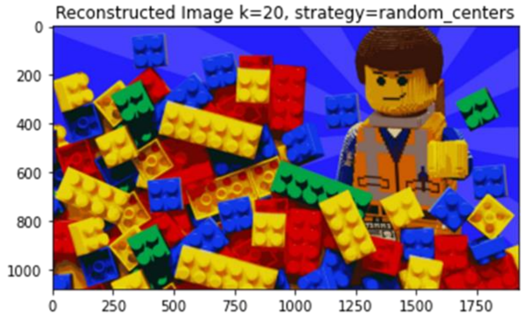
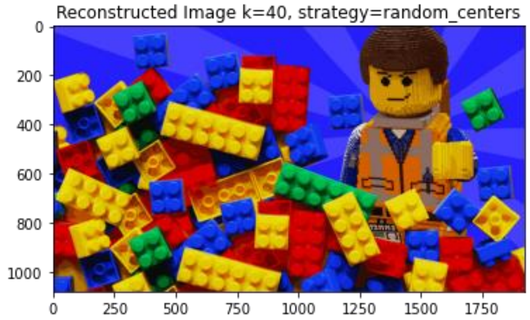
## Image 4: ‘lego.jpg’

Original Image:

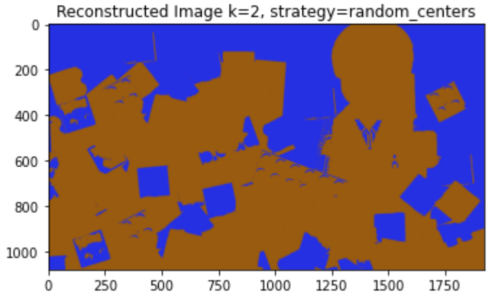
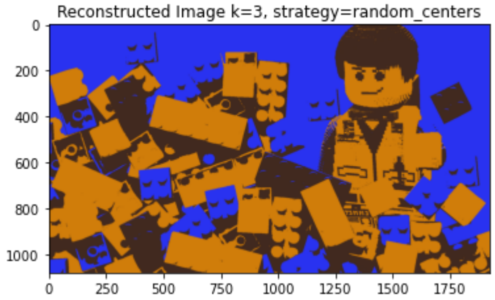


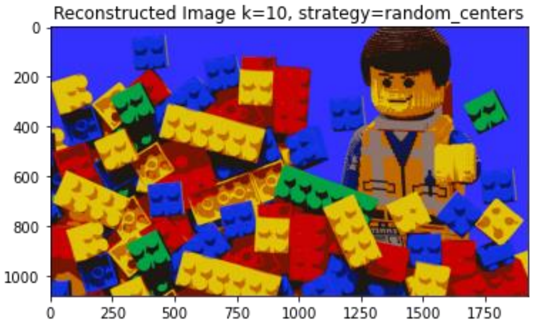
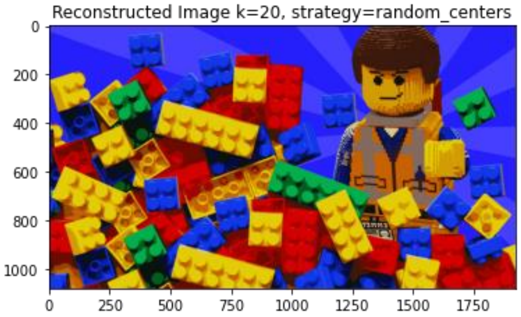
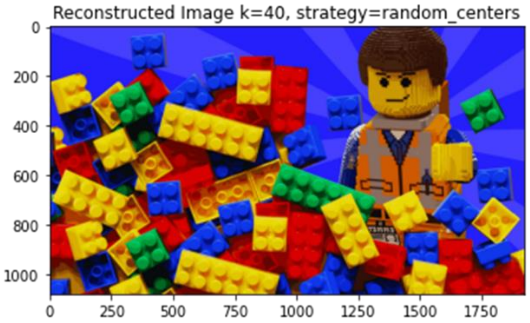
1) random_centers:
   a. Trial 1:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 3241.0451666200215 | 13 |  |
| 3 | 1841.9171034183448 | 32 |  |
| 10 | 346.1819378238659 | 22 |  |

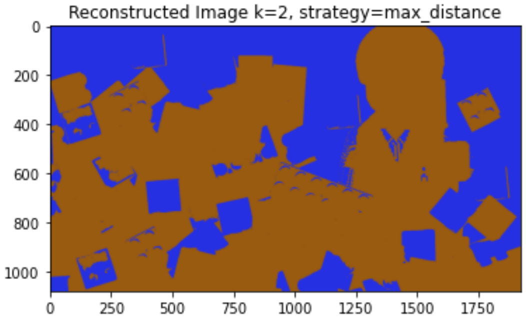
|  |  |  |  |
|---|---|---|---|
| 20 | 154.69499713776395 | 152 |  |
| 40 | 71.29969158658749 | 150 |  |

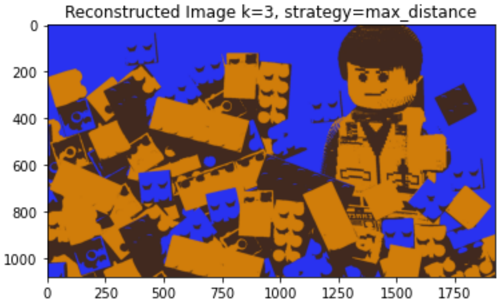
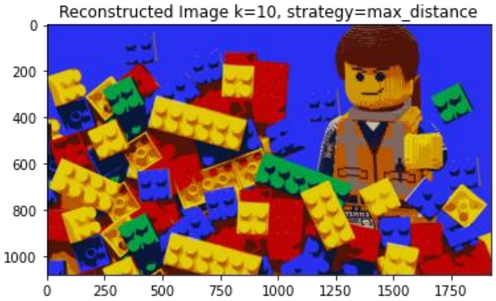
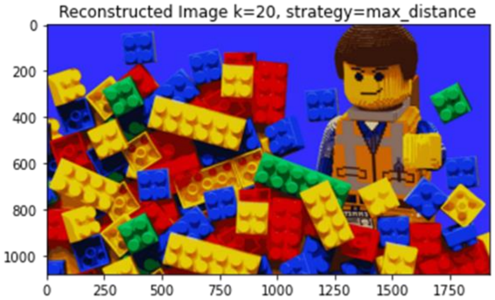
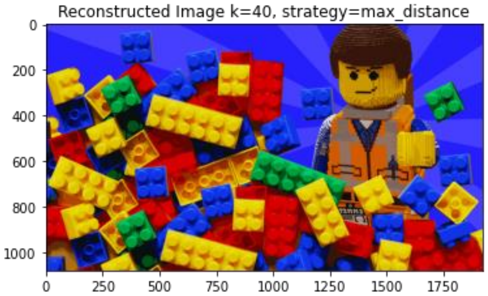
b. Trial 2:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 3241.0451666200215 | 14 |  |
| 3 | 1841.9171034183448 | 35 |  |

| | | | |
|---|---|---|---|
| 10 | 346.1819445328958 | 19 |  Reconstructed Image k=10, strategy=random_centers |
| 20 | 147.56469992890658 | 42 |  Reconstructed Image k=20, strategy=random_centers |
| 40 | 71.29043393171408 | 179 |  Reconstructed Image k=40, strategy=random_centers |

c. max_distance:

| k | MSE | # of iterations | Reconstructed Image |
|---|---|---|---|
| 2 | 3241.045166587815 | 15 |  Reconstructed Image k=2, strategy=max_distance |

| | | | |
|---|---|---|---|
| 3 | 1841.9171034183448 | 36 |  |
| 10 | 406.9019134315619 | 36 |  |
| 20 | 189.30081095494268 | 55 |  |
| 40 | 77.15398118540381 | 109 |  |