k-Means Clustering
Lab Assignment

By: Kenneth Marenco

(10) 1. Provide three example situations (maybe ones relevant to your major) that seem analogous to our "pizza-stores" model. For each situation, what are the "pizza customers" and what are the "pizza stores?"

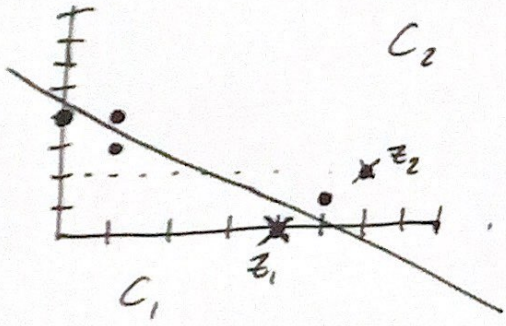
K-means Clustering

1) 1) Engineering Students and their proximity to Engineering classrooms

2) Workers and their job locations

3) features inside of an image

2)



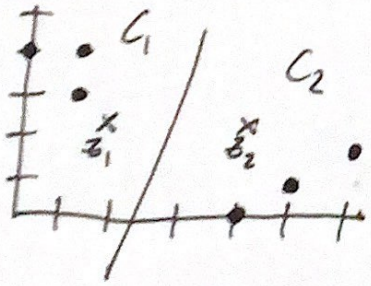
z_1)

$x: \frac{1+0+4}{3} = 5/3$

$y: \frac{4+3+0}{3} = 7/3$

z_2) $x: \frac{1+5+6}{3} = 4$

$y: \frac{4+1+2}{3} = 7/3$



z_1)

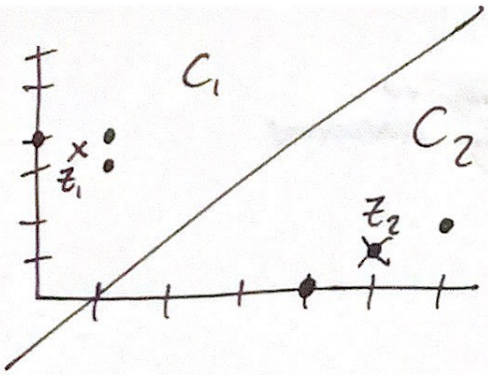
$x: \frac{0+1+1}{3} = 2/3$

$y: \frac{4+4+3}{3} = 11/3$

z_2)

$x: \frac{4+5+6}{3} = 5$

$y: \frac{0+1+2}{3} = 1$

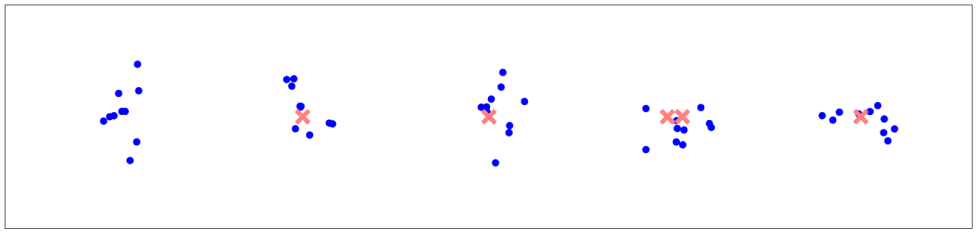


The k-means clustering has reached convergence

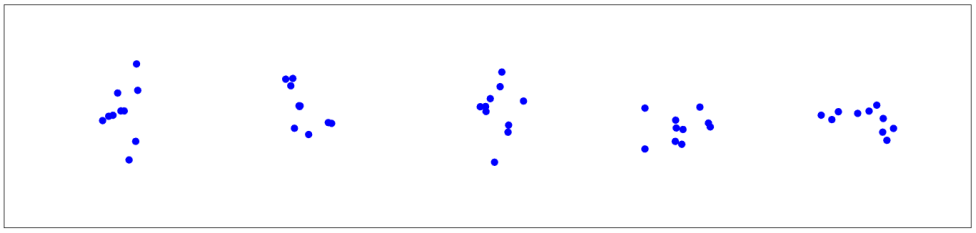
(20) 2. Perform k-Means clustering by hand on the following data with $k = 2$ and assuming the initial cluster centers are located at (4,0) and (6,2). Show your work step-by-step... what are the clusters at each step, and the new cluster centers at each step? Be sure to identify the final cluster assignments after the k-Means algorithm has converged.

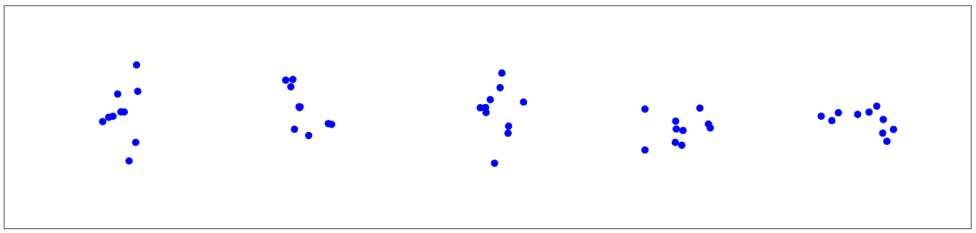
Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

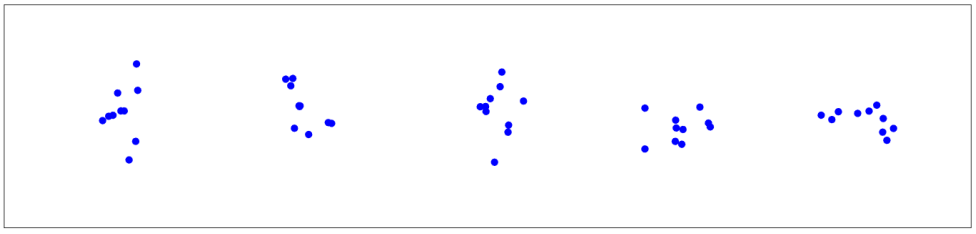
(20) 3. Consider the data set shown below, with initial cluster locations denoted by the $\times$'s.



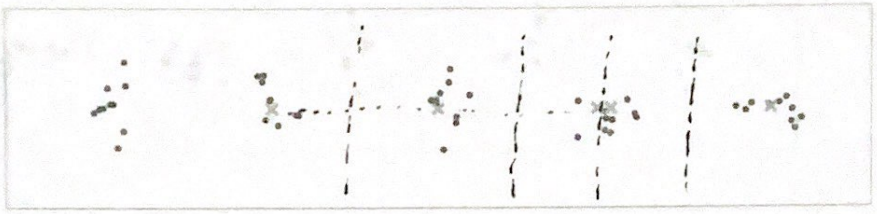
(a) What is the result of *k*-Means clustering of this data with these initial cluster locations? (Data are replicated below to assist you in iterating through *k*-Means.)



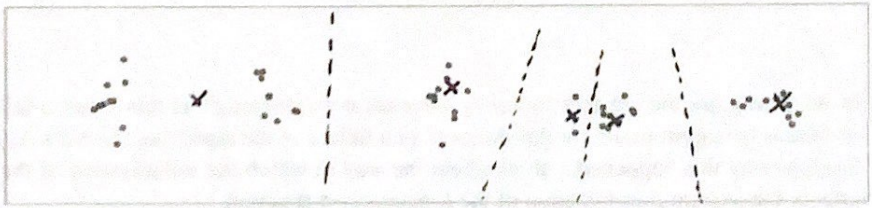


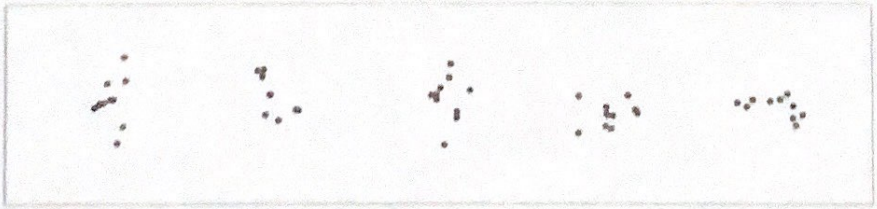


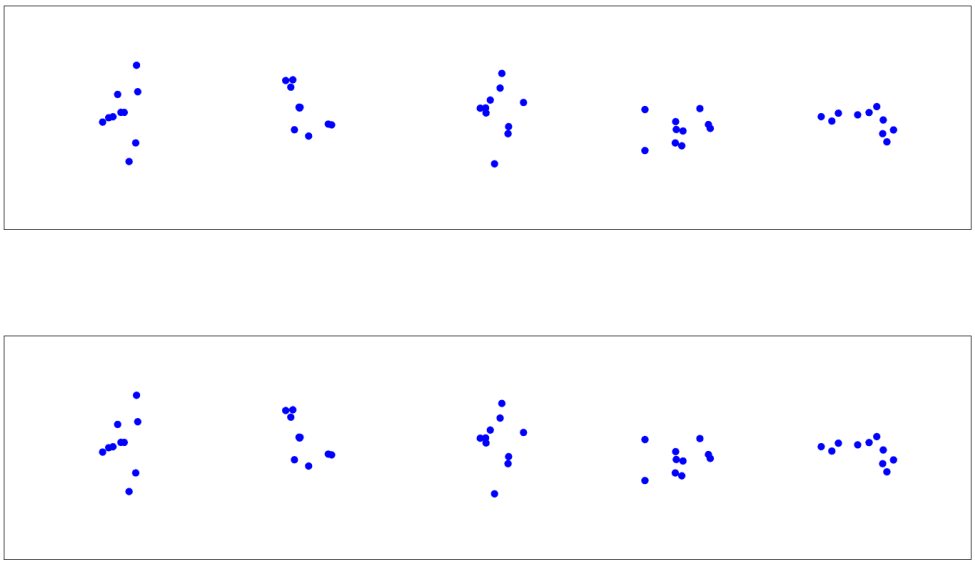
(20) 3. Consider the data set shown below, with initial cluster locations denoted by the $\times$'s.



(a) What is the result of *k*-Means clustering of this data with these initial cluster locations? (Data are replicated below to assist you in iterating through *k*-Means.)





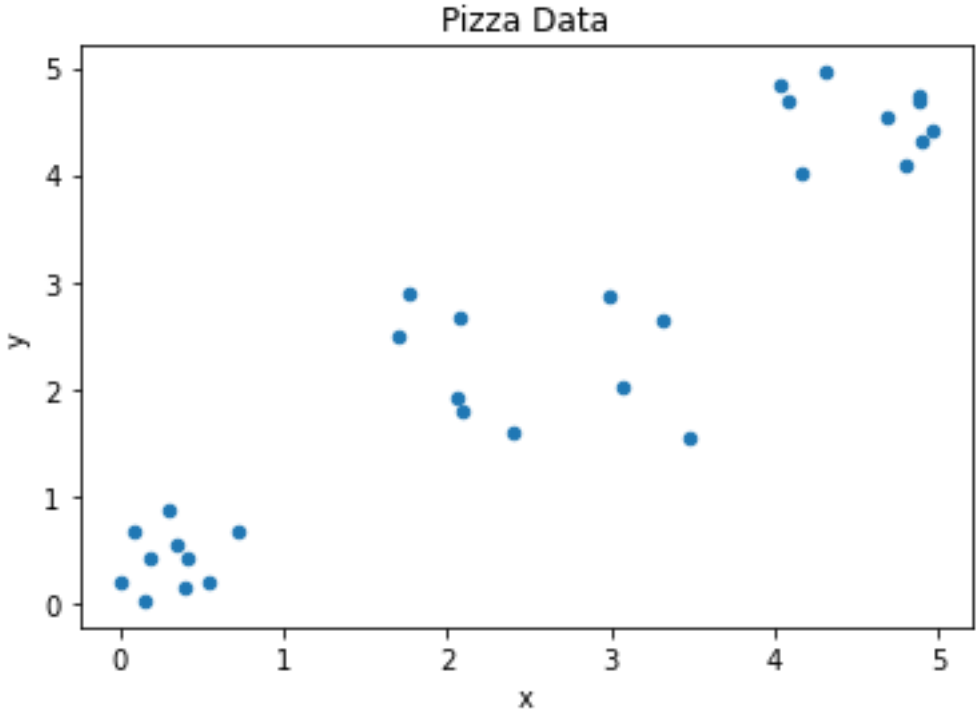


(b) In what way are the clusters found by *k*-means not satisfying? Is this result a failure of the goal (*k*-Means is inappropriate for this dataset), or a failure of the algorithm (Lloyd's Algorithm failed)? Explain why this happened... think about the way in which the initialization of the cluster centers affects subsequent minimization of the *k*-means cost function.

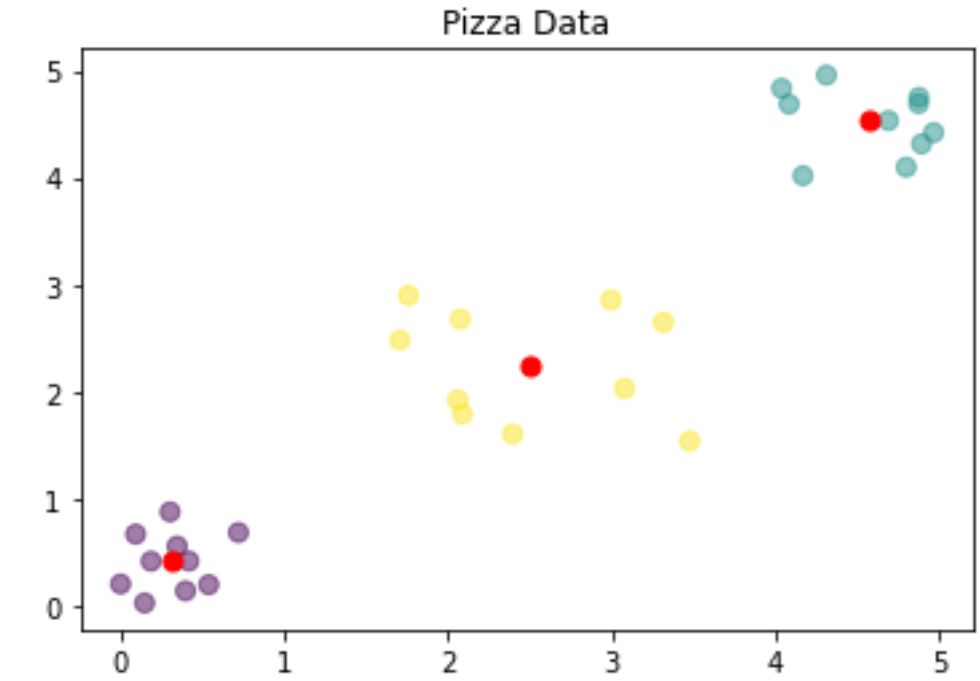
The clusters are not satisfying because they are all inline with anchors either too far from what should be a cluster or has too many anchors in what should. This appears to be a failure of the algorithm itself due to the initialization of the anchors. If the anchors were to be placed one near every group seen above, the algorithm would have created the accurate 5 clusters one would assume there to be.

(30) 4. This example concerns the sample dataset `PizzaData.csv`.

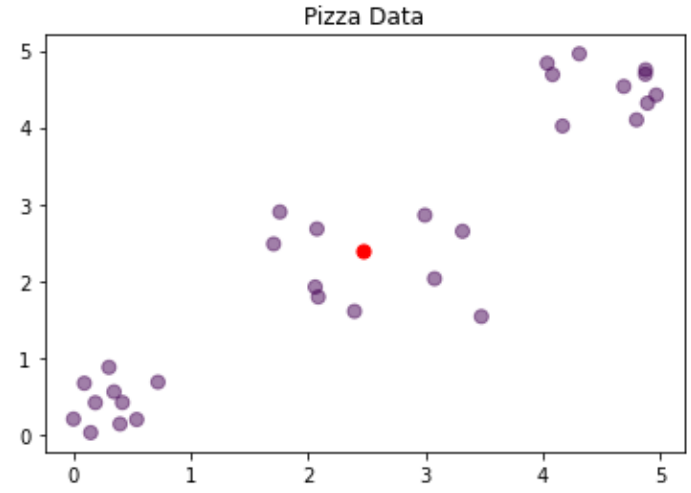
(a) Load the dataset and plot it (paste your picture below).

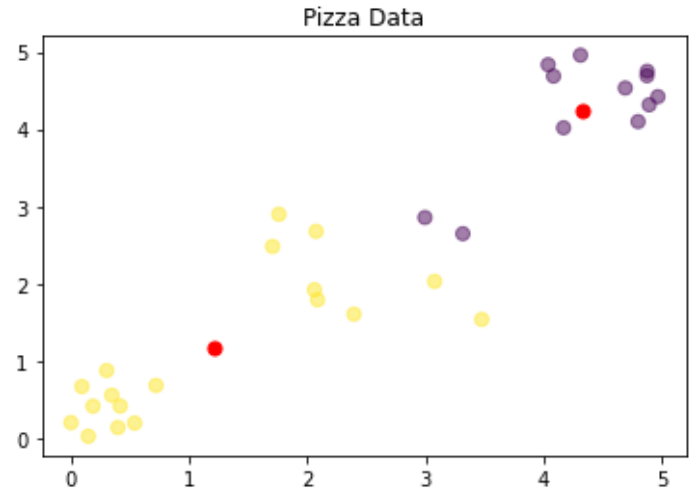


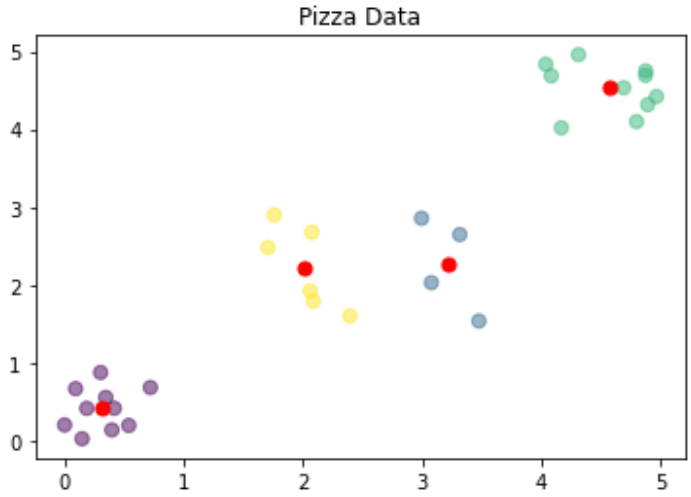
(b) Run k-means on the pizza dataset with $k = 3$, and plot the output below. (Figure out a useful way to visualize the output clusters and cluster-centers, perhaps by using three colors and/or marker shapes.) Use a convergence criterion of = 0.02, but feel free to experiment with what happens when you tweak this parameter.



(c) Repeat k-Means clustering of the pizza dataset for $k = 1$ then $k = 2$ and then $k = 4$. What is a principled way to think about $k = 3$ being "the best choice" for this particular dataset?





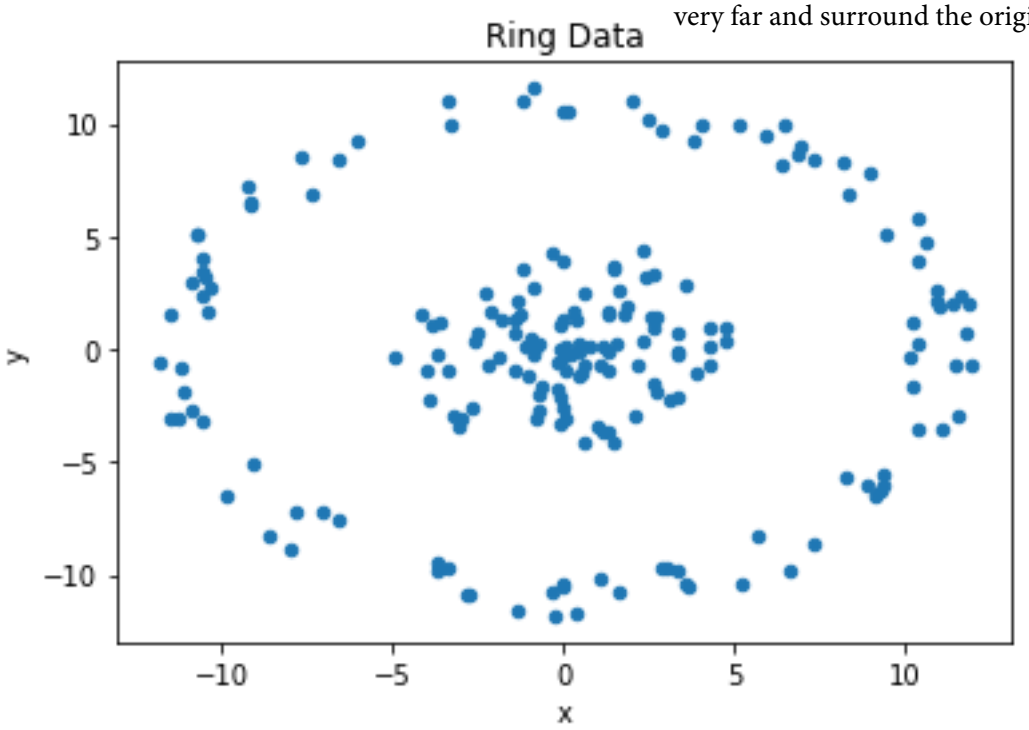


The reason for k=3 to be considered the best choice is from the observer's perspective there are visually 3 distinct groups with points that are far with high x and y values, points that are not as far with x's and y's around 2.5, and close points that have x's and y's that are about .5.

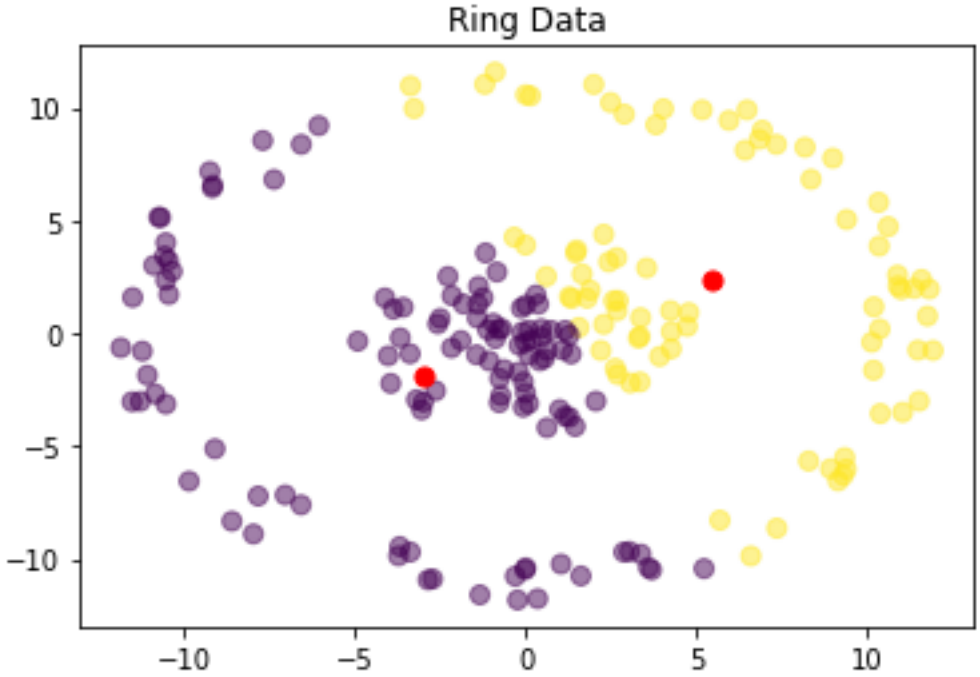
(30) 5. This example concerns the sample dataset `RingData.csv`.

(a) Load the dataset and plot it (paste your picture below). How many clusters do you see? What are they?

I see 2 clusters (those near the origin and those in very far and surround the origin).



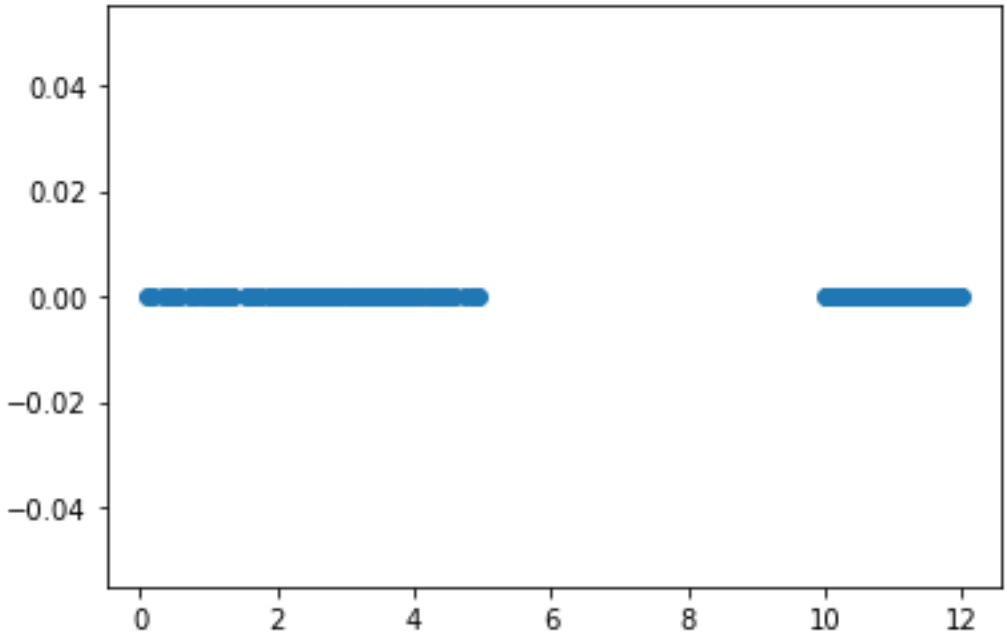
(b) Run k-means on the ring dataset with $k = 2$, and plot the output below, visualizing it as you did in the previous exercise. (WARNING: It's going to look odd!)



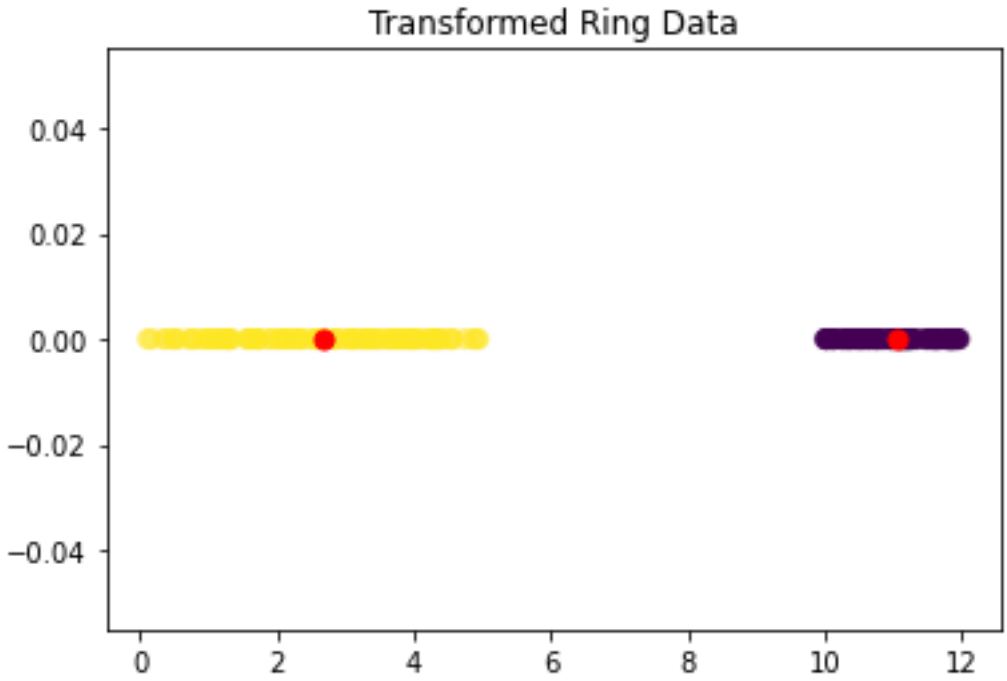
(c) In what way were the two "clusters" found by running 2-means on the ring dataset not satisfying? Is this result a failure of the goal (*k*-Means is inappropriate for this dataset), or a failure of the algorithm (Lloyd's Algorithm failed)? Explain why this happened... think about the way in which the *k*-means cost function defines a "good clustering."

The found two clusters were not satisfying because they were not the clusters intended to be found. Unlike the Pizza Data where the issue was the initial anchors with the algorithm failing, this Ring Data fails due to a failure of the goal. The method of k-Means is not appropriate for this type of dataset. The reason for this is how the cost function defines a cluster. Typically from what has been seen from the algorithm, good clusters are when the centroids are almost unique from other clusters and when its points are nearby each other. However in this case, the group of nearby points are now very far from each other.

(30) 6. (a) For each data point in the ring dataset, compute its distance from the origin $(0,0)$, and store these distances as a many-by-one array (`distFromOrigin`). Plot the transformed ring dataset (`distFromOrigin`). Since this dataset is now 1-dimensional, think about using a visualization technique to help you see multiple data points that may be at the same radius from the origin.



(b) Run k-means on the transformed ring dataset (`distFromOrigin`) with $k = 2$ and visualize the resulting clusters for both the transformed dataset (`distFromOrigin`) and the original (2-dimensional) ring dataset.



(c) In what way were the two "clusters" found by running 2-means on the transformed ring dataset more satisfying? Is this result due to greater consistency with the goal (*k*-Means is more appropriate for the transformed dataset), or greater consistency with the algorithm (Lloyd's Algorithm is more appropriate for the transformed dataset)? Explain why this happened... think about the way in which the *k*-means cost function defines a "good clustering."

The two clusters are more satisfying because the transformed dataset is now 1D. The method used was attempting to project the 2D dataset into something that would not fail in terms of goal for the k-means clustering. The cost being the distance from all the points in the dataset are now in line with the goal because by first taking the distance from the origin before projecting the data allowed there to be two distinct groups of points rather than what would have likely been three if the projection occurred beforehand.