List and explain the different iterations you went through while approaching the dataset

- **Explain the output you got in each**

First let's explain the dataset. It’s composed of data about flowers. These flowers have four different features, and are classified in three different classes.There are 50 sample of each classes, for a total of 150. All of the costs are from the TestSet.

First Iteration:

- 3 clusters
- TrainSet = 100
- TestSet = 50
- Initial Centroids = TrainSet[0],TrainSet[1],TrainSet[2]
- Iterations = 2
- Initial Cost = 23.86
- Final Cost = 3.49

Increasing the iterations does not reduce the final cost. This is probably because it found the best position.

Second iteration:

- 3 clusters
- TrainSet = 100
- TestSet = 50
- Initial Centroids = TrainSet[RANDOM],TrainSet[RANDOM],TrainSet[RANDOM]
- Iterations = 1
- Initial Cost = 4.72
- Final Cost = 3.49

I use srand(1) for the randomness. It seems it has a better initial cost, and only needs 1 iteration to converge to the best one.

Third iteration:

- 3 clusters
- TrainSet = 75
- TestSet = 75
- Initial Centroids = TrainSet[RANDOM],TrainSet[RANDOM],TrainSet[RANDOM]
- Iterations = 3
- Initial Cost = 19.12
- Final Cost = 2.36

I have increased the TestSet, to try to improve it, and I did. Now the cost is less, which means that it’s better than before, but needs more iterations to get to it.

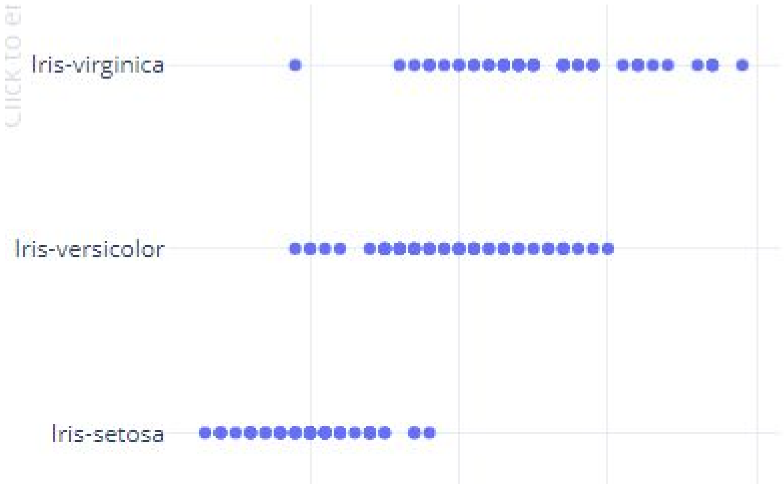
Final iteration:

- 3 clusters
- TrainSet = 75
- TestSet = 75
- Initial Centroids = TrainSet[0],TrainSet[25],TrainSet[50]
- Iterations = 3
- Initial Cost = 1.18
- Final Cost = 0.58

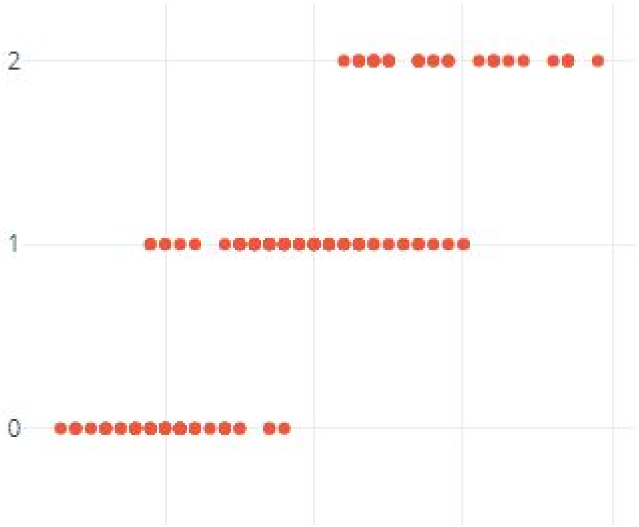
This is probably the best result, and the one that is actually correct. I thought that the dataset randomized the data, but no. Which means that the dataset was ordered, which made the sets be biased. Now, the TrainSet is formed by 25 samples of each classes, and the TestSet is composed by the missing ones.

- **Select the best and explain why that is the best you got**

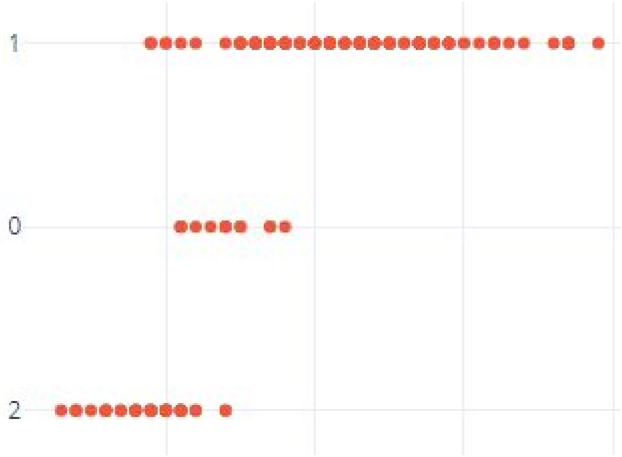
I think the last iteration is the best one. Here are some pictures to prove it. The first picture is a visual representation of the original dataset.



The next one is using the final iteration, clustering by K-means.



For comparison, this is the first iteration, which I think it's the worst one:



So it’s clear that the last iteration is the best one.

How many clusters did you find on your best guess? What does each cluster contain (interpret those results looking at the meaning of each value)?

In the dataset of the Iris, there are clearly three different classes, so I used three clusters for all of the iterations. Each cluster should correctly contain each type of flower, depending on their features. Of course, the model is not perfect, so there may be some Iris in a class where they don’t really belong.