# 43 Overview

## Algorithm

//k means overview

We utilized the k means clustering method to analyze our datasets. The method involved randomly determining *centroids* that fall within a particular interval particular to the dataset. Then we find the Euclidean distance between each point and each centroid and calculate the minimum distance between each point and a centroids. We then assign that point to the "cluster" corresponding to that centroid. Next, we update the centroid to be the average of points corresponding to that centroid. We repeat the process until no update is performed (we have reached convergence) or until

## Ideas

We decided to create an algorithm to calculate the value of "k". For each "k", we calculated the silhouette coefficient, then chose the "k" with the highest "k" value.

# Discussion

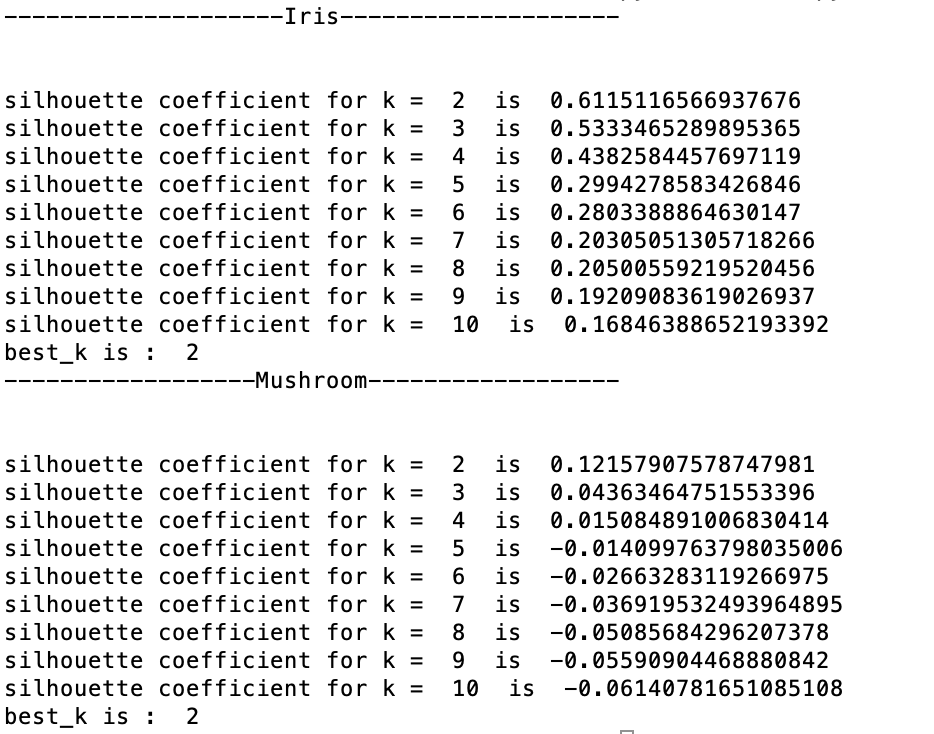
## **Determine the best k in both datasets**

iris = load\_iris("Iris.csv") #load iris data from local file

number\_of\_clusters(iris) # determine best k using silhouette coefficient method

mushroom = load\_mushroom("mushrooms.csv") #load mushroom data from local file

number\_of\_clusters(mushroom) # determine best k using silhouette coefficient method



As we can see, in both cases, the best k are 2. Observed that best k for iris data is 2 while there are actually three species in total. This is understandable because from distribution diagram of data, we are inclined to divide them to two clusters without knowing there are three species. Also, chances are silhouette coefficient method may not be the optimum method for best k determination within iris dataset.

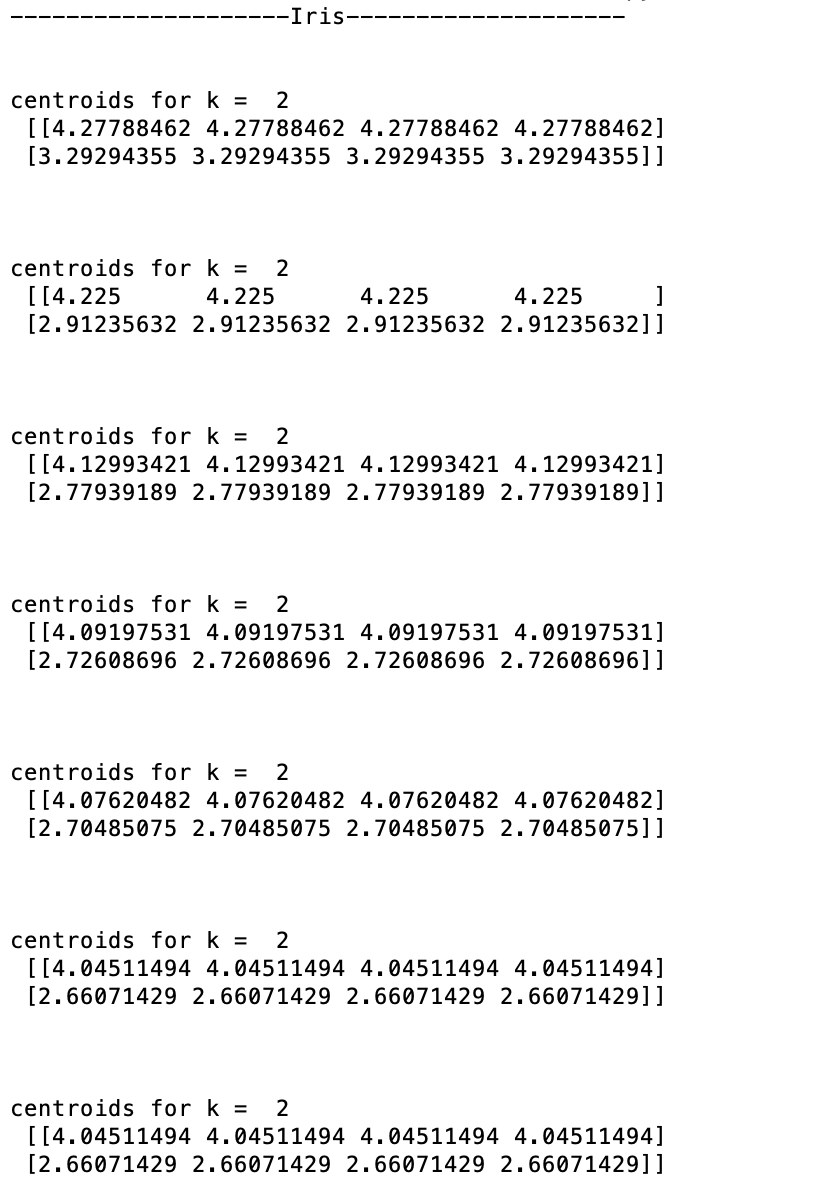
## **Track changes of centroids**

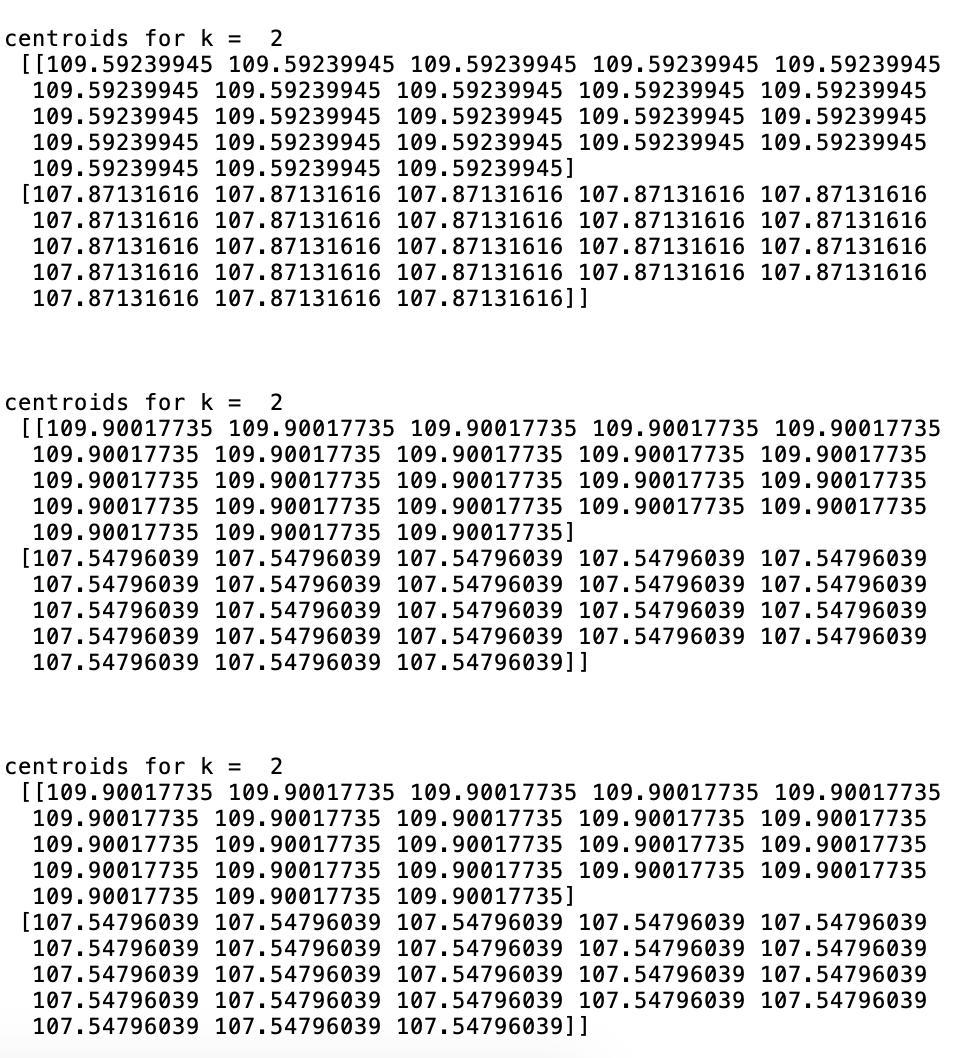
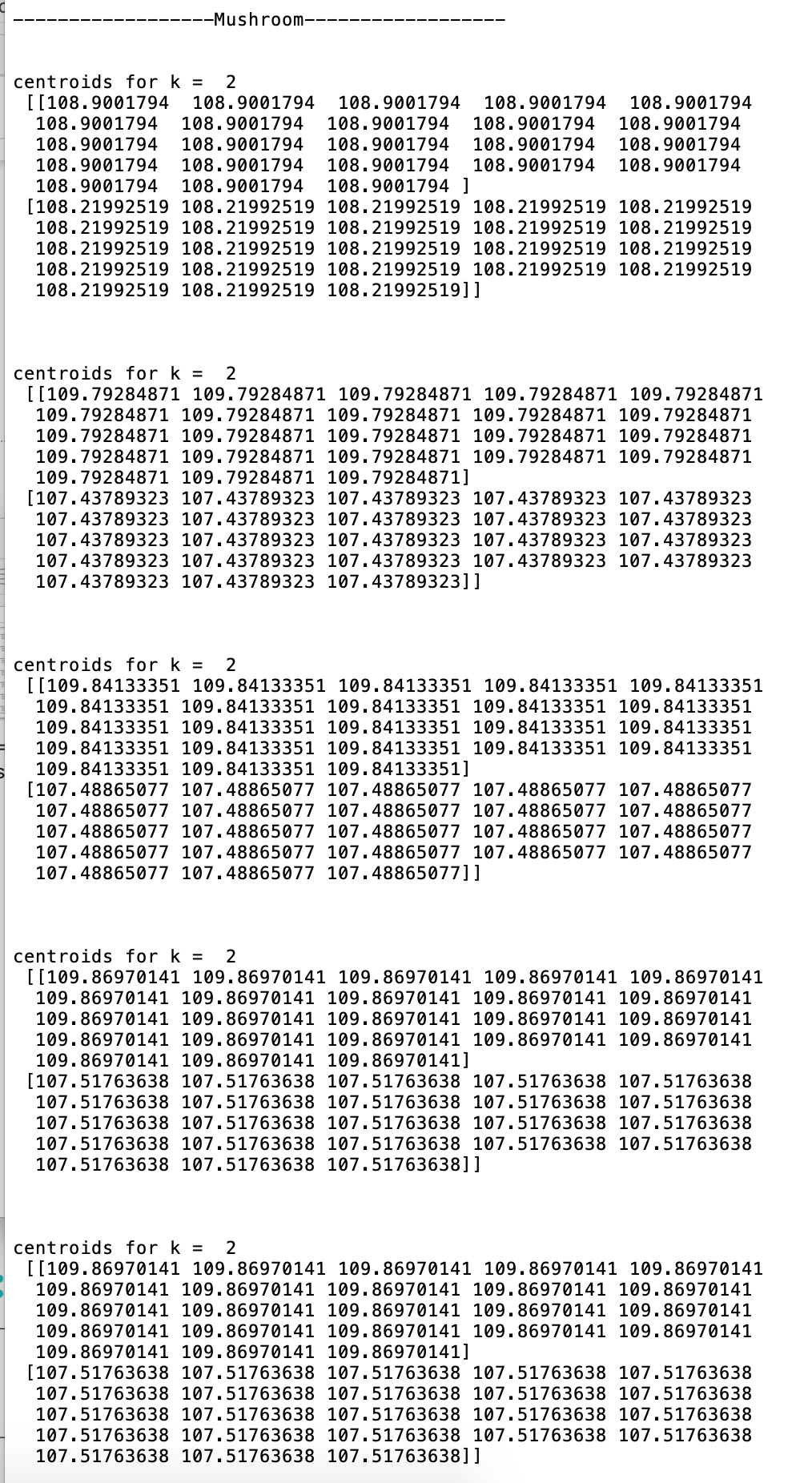
iris = load\_iris("Iris.csv") #load iris data from local file

kmeans(iris, 2)

mushroom = load\_mushroom("mushrooms.csv") #load mushroom data from local file

kmeans(mushroom, 2)



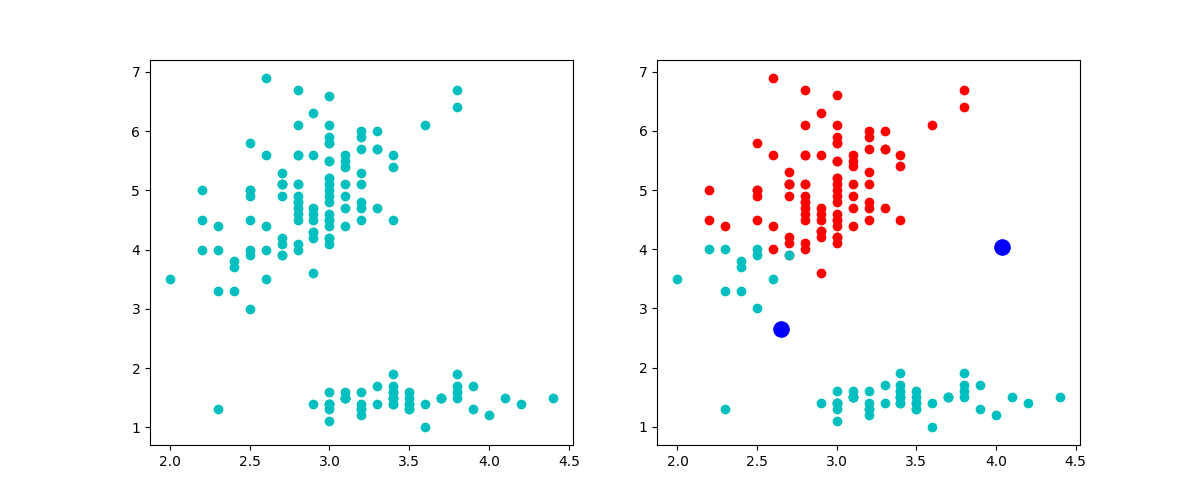


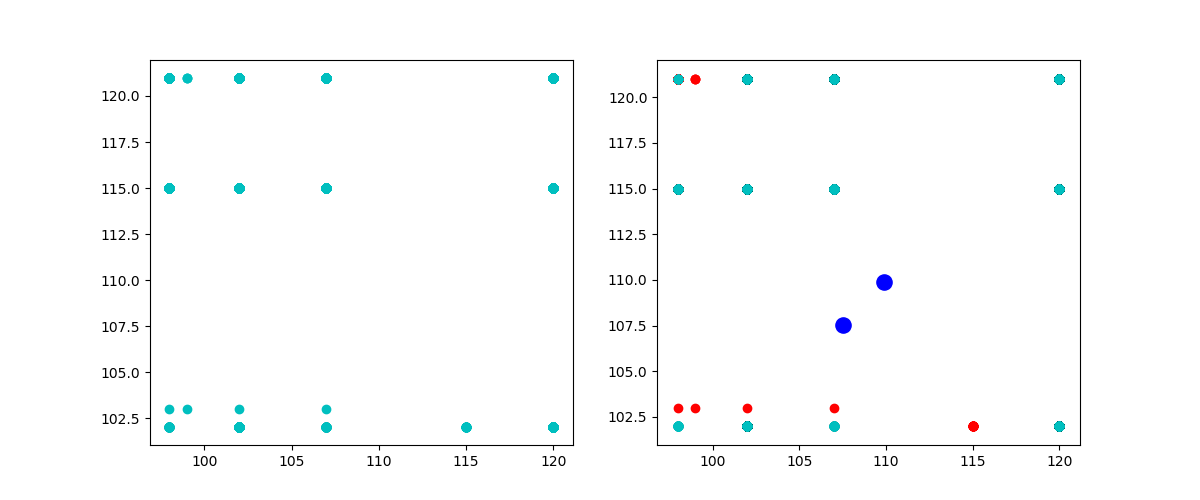
As we can see, centroids stop updating and converge to some points eventually.

## **Model the distribution of data**

PlotClusters(iris, label, centroids); #plot iris dataset

PlotClusters(mushroom, label, centroids); #plot mushroom dataset



In distribution of iris dataset, two clusters are depicted in red and cyan. Two blue dots represent the centroids. As we are using 2d diagram to show the distribution of data, relations between points are not very obvious since there are four features for each item.

In distribution of mushroom dataset, items are divided to two clusters, red and cyan. Centroids are represented by two blue dots. As we can see this is not a good approach to model data since there are 8125 data in the file, but we only see few in the diagram. First is because for each item, there are 22 features. 22 dimensional value won’t be plotted well in 2D diagram. Also, the values in the dataset are characters, which means values are discrete and some data are highly repetitive as the size of dataset goes larger.