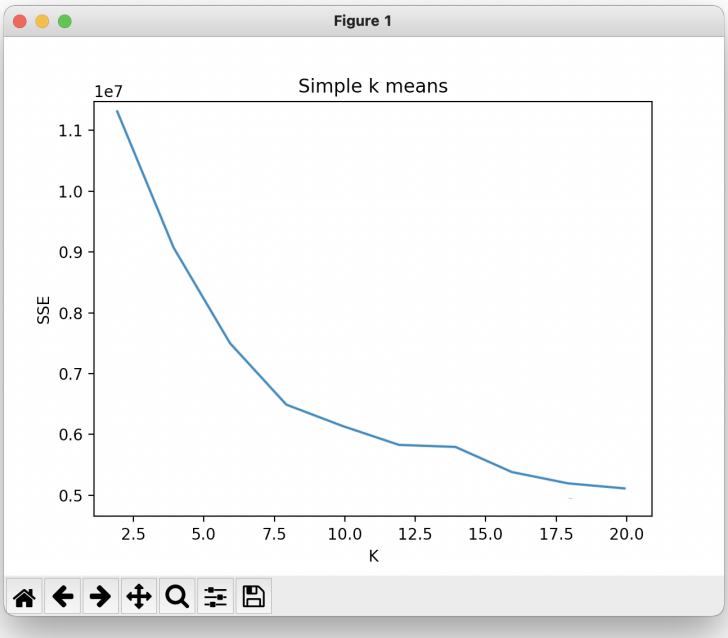
HW2 report

How to use my code:

After you run the “litong.py” file, it will promt you to enter the file name. And I export the .xlsx file to .csv file, which is in the zipped folder. You can download it and enter the file path now. Then you will get the plot for question 1. After you close the window of plot, the silhouette coefficient for question 2 will show.

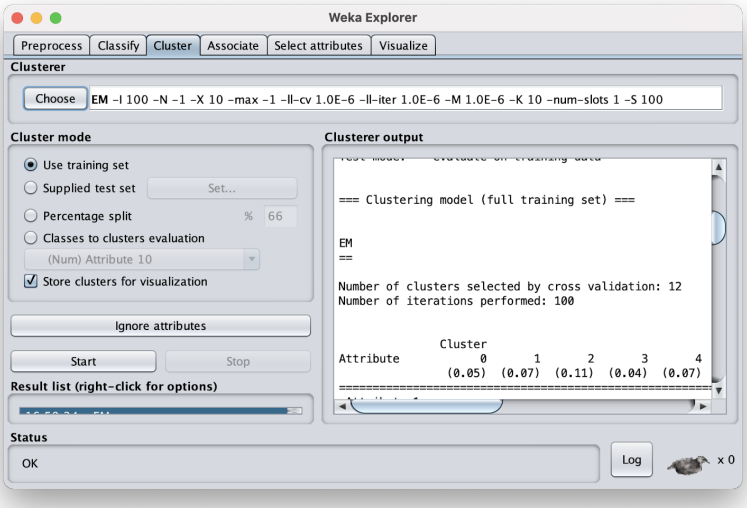
1.

A. Here is the plot of the sum of squared error for the iterations.



And I find that the “knee” of the graph is when k = 8.

B. Weka's EM algorithm generated k = 12.



Even though the EM k value does not exactly match the one I chose for k means in Part-1, 8 and 12 are still closed to each other. I think it is because k means is EM'ish, only makes 'hard' assignments to cluster. And if the covariaEM ans keans are similar in the sense that they allow model refining of an iterative process to find the best congestion. The difference between them is the K-means algorithm differs in the method used for calculating the Euclidean distance while calculating the distance between each of two data items; and EM uses statistical methods.

2.

For k means method, the silhouette coefficient is 0.9557787873520192, and for EM method, the silhouette coefficient is 0.5339370126182955. K means has better performance.