1) Explanation of The Souce Code

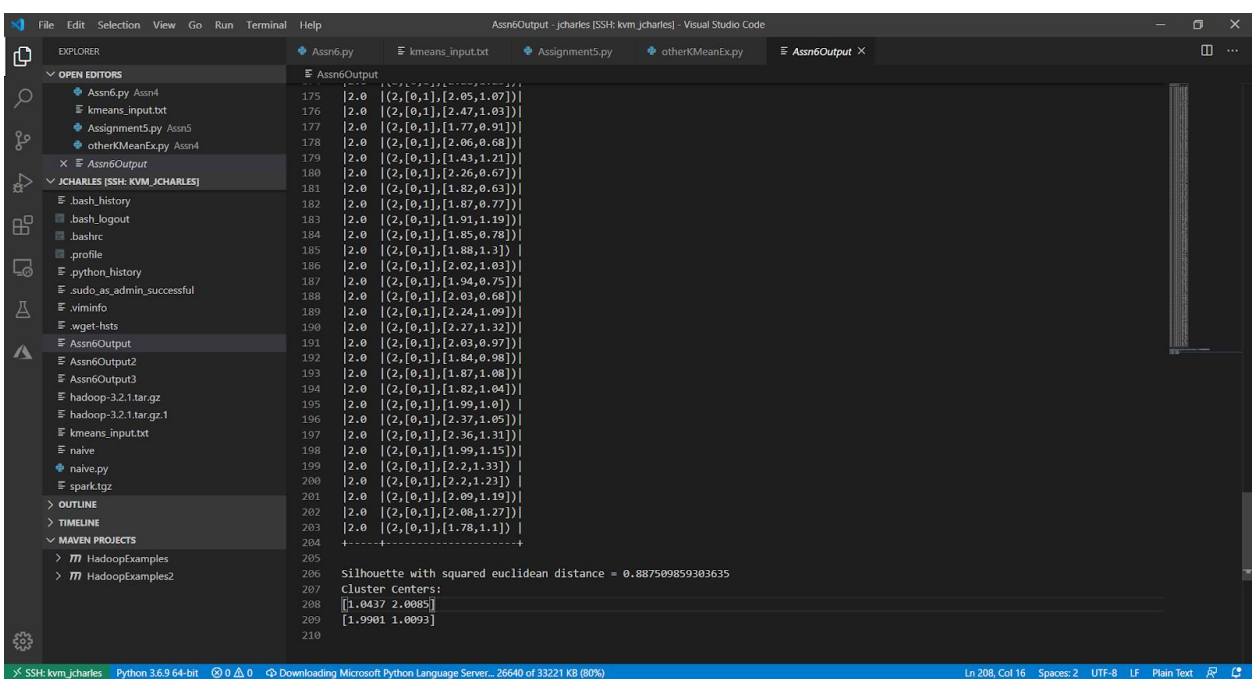
k-Means is often used to pre-cluster massive data sets before other processing. Although it is very simple and fast it is very sensitive to 2 items which are the initial choice of centroid and the Outliers. It is designed to be used on continuous data.

This gif basically explains what the k-means algorithm does, and it is how I perceived it.
https://thumbs.gfycat.com/NecessarySociableCentipede-size_restricted.gif

To Start, we need some additional import like our Kmean import and the ClusteringEvaluator import. These help use their functions later. We start by reading in our dataset with our pregven file by the professor and show our 200 point data set. Afterwards we set the number of clusters, our means, to 2. Following that we then fit our dataset into the model and transform it so our CulteringEvaluator can evaluate it. This gives us the silhouette with sq'd Euclidean distance. Finally we can then get the cluster centers by calling clusterCenters() function on our model. Then we can print.

2) Experimental Results:

Issues that I had to deal with was getting my numpy installed and getting the starter code from Spark (which was pretty helpful) , as well as looking for the proper commands.



I’m also just going to include my source code in the document as well to avoid any confusion with submission.

My source code has comments explaining what each portion of my code does. Although it was short, the simplicity was very appreciated.