Code NOTE: for each distance case, the sample command to submit the application to Spark is included in 1st two lines in .py file.

Solutions for a(2) and b(2):

a(2):
The cost estimates the quality of the clustering. It can be observed from Euclidean.png that c2 is better than c1 in Euclidean case because it distributes the initial clusters far apart.

b(2):
The cost estimates the quality of the clustering. It can be observed from Manhattan.png that c1 is better than c2 in Manhattan case. NOTE: It is because c2 initializations are chosen according to the Euclidean distance metric.