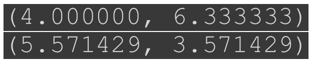
Task 1:

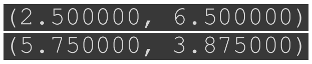
all graphs of the clusters are in the code

(1):



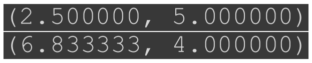
```
(4.000000, 6.333333)
(5.571429, 3.571429)
```

(2):



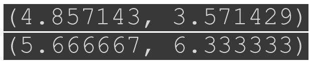
```
(2.500000, 6.500000)
(5.750000, 3.875000)
```

(3):



```
(2.500000, 5.000000)
(6.833333, 4.000000)
```

(4):



```
(4.857143, 3.571429)
(5.666667, 6.333333)
```

Task 2:

Q1:

I would say that Jaccard is probably the best out of the three based on the results I got. Jaccard had the lowest values overall. While Euclidian produced the lowest values for Centroid 2 and had similar values for Centroid 1, it had much higher values for Centroid 3. While Jacccard did have larger values for Centroid 2, the values for Centroid 2 were overall large, making the percent difference much smaller compared to the changes in Centroid 3. Cosine did the worst out of all of them.

Q2:

Jaccard and Cosine seem to be equally accurate with 100% accuracy on centroids 1 and 2 while Euclidian took two from Centroid 2 and added it to Centroid 3 bringing the overall accuracy down

Q3:

Jaccard and Euclidian seems to take only a few iterations for the data I tested. Cosine took several more iterations to stop changing.

Q4:

Based upon my results, there isn't a time where SSE increases and since the centroids settle for Euclidian and Jaccard very quickly, the one that takes the most time would be Cosine.

Task 3:

I am not quite sure whether we are doing these measurements for each cluster or one point between each cluster, so I just did both.

A.

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

for red cluster:

$$\sqrt{(4.6 - 4.7)^2 + (2.9 - 3.2)^2}\ or\ \sqrt{(5.0 - 4.7)^2 + (3.0 - 3.2)^2}$$

$$\sqrt{(-0.1)^2 + (-0.3)^2}\ or\ \sqrt{(0.3)^2 + (-0.2)^2}$$

$$\sqrt{0.1}\ or\ \sqrt{0.13}$$

$$0.3162\ or\ 0.3606$$

The farthest distance seems to be 0.3606 according to euclidians for the red cluster

for blue cluster:

$$\sqrt{(6.2 - 5.9)^2 + (2.8 - 3.2)^2}\ or\ \sqrt{(6.7 - 5.9)^2 + (3.1 - 3.2)^2}$$

$$\sqrt{(0.3)^2 + (-0.4)^2}\ or\ \sqrt{(0.8)^2 + (-0.1)^2}$$

$$\sqrt{0.25}\ or\ \sqrt{0.65}$$

$$0.5\ or\ 0.8062$$

The farthest distance seems to be 0.8062 according to euclidians for the blue cluster

for both clusters:

$$\sqrt{(6.7 - 4.6)^2 + (3.1 - 2.9)^2}$$

$$\sqrt{(2.1)^2 + (-0.2)^2}$$

$$\sqrt{4.45}$$

$$2.1095$$

The largest distance between two points of different clusters seems to be 2.1095

B.

for red cluster:

$$\sqrt{(5.0-4.9)^2+(3.0-3.1)^2}$$

$$\sqrt{(0.1)^2+(-0.1)^2}$$

$$\sqrt{0.02}$$

$$0.1414$$

The shortest distance seems to be 0.1414 according to euclidians for the red cluster

for blue cluster:

$$\sqrt{(6.0-5.9)^2+(3.0-3.2)^2}$$

$$\sqrt{(0.1)^2+(-0.2)^2}$$

$$\sqrt{0.05}$$

$$0.2236$$

The shortest distance seems to be 0.2236 according to euclidians for the blue cluster

for both clusters:

$$\sqrt{(5.9-5.0)^2+(3.2-3.0)^2}$$

$$\sqrt{(0.9)^2+(0.2)^2}$$

$$\sqrt{0.85}$$

$$0.9220$$

The shortest distance between two points of different clusters seems to be 0.9220

C.

As I understand the question, it says that I should find the distance between each point for all points giving six distances per cluster

For red cluster:

$$\frac{\sqrt{0.1}+\sqrt{0.13}+\sqrt{0.02}+\sqrt{0.05}+\sqrt{0.17}+\sqrt{0.13}}{6}=0.3024$$

For blue cluster:

$$\frac{\sqrt{0.25} + \sqrt{0.65} + \sqrt{0.05} + \sqrt{0.08} + \sqrt{0.5} + \sqrt{0.34}}{6} = 0.5171$$

For both clusters:

$$\frac{\sqrt{1.44} + \sqrt{4.01} + \sqrt{1.73} + \sqrt{2.41} + \sqrt{1.01} + \sqrt{3.24} + \sqrt{1.22} + \sqrt{1.78} + \sqrt{0.85} + \sqrt{2.9} + \sqrt{1} + \sqrt{1.48} + \sqrt{1.78} + \sqrt{4.45} + \sqrt{1.97} + \sqrt{2.57}}{16}$$

$$= 1.4129$$

D.

Noise is data that misleads algorithms or errors. The first form this could take that comes to mind is outlier points. However, data points that are added in error also come to mind. However, this is less likely to happen. The dataset that would be most susceptible to change would most likely be A. Outliers usually are usually data points that are far away from a cluster, but still considered part of it. Since this one is far from the rest of the data, it would be the farthest distance and skew this measurement. C would also be affected by outliers, but lesser so because it's an average of everything and therefore, that one point would be a fraction of the total value. B would be completely unaffected by outliers except in the case that there are two that are closer than two data points in the cluster. Faulty information is hard to gauge the affect on these measurements because it's not exactly consistent in how it would manifest in a dataset. Also, adding data that has no effect on prediction would have equal effect on all stats just making it harder in general to come to conclusions considering you have to weed through more data. Therefore, I would rank them as B>C>A in terms of robustness