Name:

1. **(2 Pts.)** State ***two*** assumptions that the *k*-means clustering algorithm makes about the shape or distribution of clusters in a data set.

Any 2 of:

Points in a cluster are relatively close to one-another (non-sparse/dense data)

Disjoint/separate/non-hierarchical clusters

Spherical/Gaussian distributed clusters

Roughly similarly sized clusters

Must know number of clusters

(Any phrasing is fine)

2. **(2 Pt.)** State ***two*** data preparation steps that should be considered before conducting a clustering analysis.

Any two of:

Standardization, removing/replacing outliers, binning of rare categorical input levels, appropriate numeric encoding of categorical inputs, combining or removing highly correlated features

(Any phrasing is fine)

3. **(1 Pt.)** ***True or False***: Squared error from cluster centroids nearly always decreases when adding more clusters into an analysis.

True

4. **(2 Pt.)** State ***two*** mathematical or statistical techniques for determining the number of clusters in a data set.

Any two of: ABC, CCC, Gap Statistic, Silhouette, 'Elbow Method'

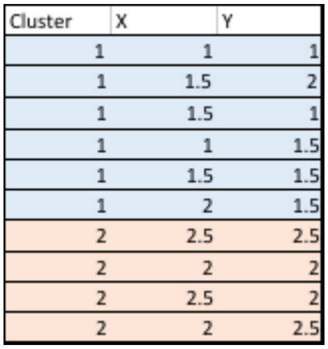
5. **(1 Pt.)** ***True or False***: k-Means clustering tends to be slower but more accurate than hierarchical clustering techniques.

False

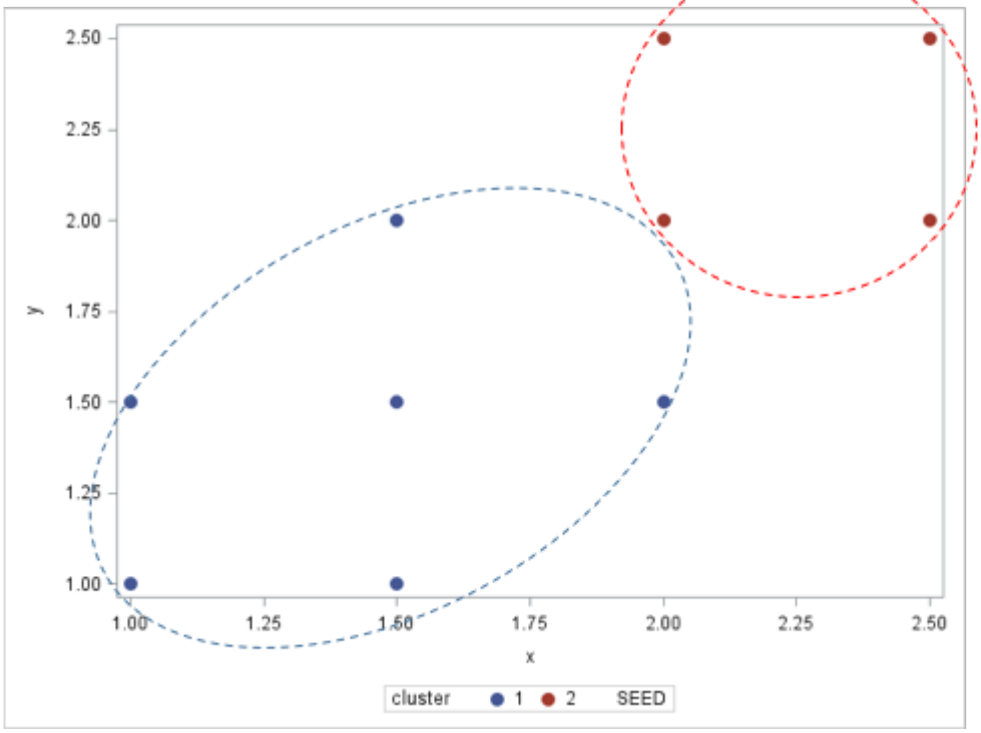
Name:

(Continue to next page)

6. **(2 Pts.)** Calculate the centroids (or archetypal members) of these two clusters. Show your work.



| Cluster | X | Y |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1.5 | 2 |
| 1 | 1.5 | 1 |
| 1 | 1 | 1.5 |
| 1 | 1.5 | 1.5 |
| 1 | 2 | 1.5 |
| 2 | 2.5 | 2.5 |
| 2 | 2 | 2 |
| 2 | 2.5 | 2 |
| 2 | 2 | 2.5 |



1 pt each

(x1, y1) = (AVERAGE(1, 1.5, 1.5, 1, 1.5, 2), AVERAGE(1, 2, 1, 1.5, 1.5, 1.5)) = **(1.41, 1.41)**

(x2, y2) = (AVERAGE(2.5, 2, 2.5, 2), AVERAGE(2, 2, 2.5, 2.5)) = **(2.25, 2.25)**