**Assignment 19 Solutions**

**1. A set of one-dimensional data points is given to you: 5, 10, 15, 20, 25, 30, 35. Assume that k = 2 and that the first set of random centroid is 15, 32, and that the second set is 12, 30. ?**

1. Using the k-means method, create two clusters for each set of centroid described above.
2. For each set of centroid values, calculate the SSE.

In [ ]:

**2. Describe how the Market Basket Research makes use of association analysis concepts ?**

In [ ]:

**3. Give an example of the Apriori algorithm for learning association rules ?**

In [ ]:

**4. In hierarchical clustering, how is the distance between clusters measured? Explain how this metric is used to decide when to end the iteration ?**

In [ ]:

**5. In the k-means algorithm, how do you recompute the cluster centroids ?**

In [ ]:

**6. At the start of the clustering exercise, discuss one method for determining the required number of clusters ?**

In [ ]:

**7. Discuss the k-means algorithm's advantages and disadvantages ?**

In [ ]:

**8. Draw a diagram to demonstrate the principle of clustering ?**

In [ ]:

**9. During your study, you discovered seven findings, which are listed in the data points below. Using the K-means algorithm, you want to build three clusters from these observations. The clusters C1, C2, and C3 have the following findings after the first iteration ?**

* C1: (2,2), (4,4), (6,6); C2: (2,2), (4,4), (6,6); C3: (2,2), (4,4),
* C2: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4), (0,
* C3: (5,5) and (9,9)

What would the cluster centroids be if you were to run a second iteration? What would this clustering's SSE be?

In [ ]:

**10. In a software project, the team is attempting to determine if software flaws discovered during testing are identical. Based on the text analytics of the defect details, they decided to build 5 clusters of related defects. Any new defect formed after the 5 clusters of defects have been identified must be listed as one of the forms identified by clustering. A simple diagram can be used to explain this process. Assume you have 20 defect data points that are clustered into 5 clusters and you used the k-means algorithm ?**

In [ ]: