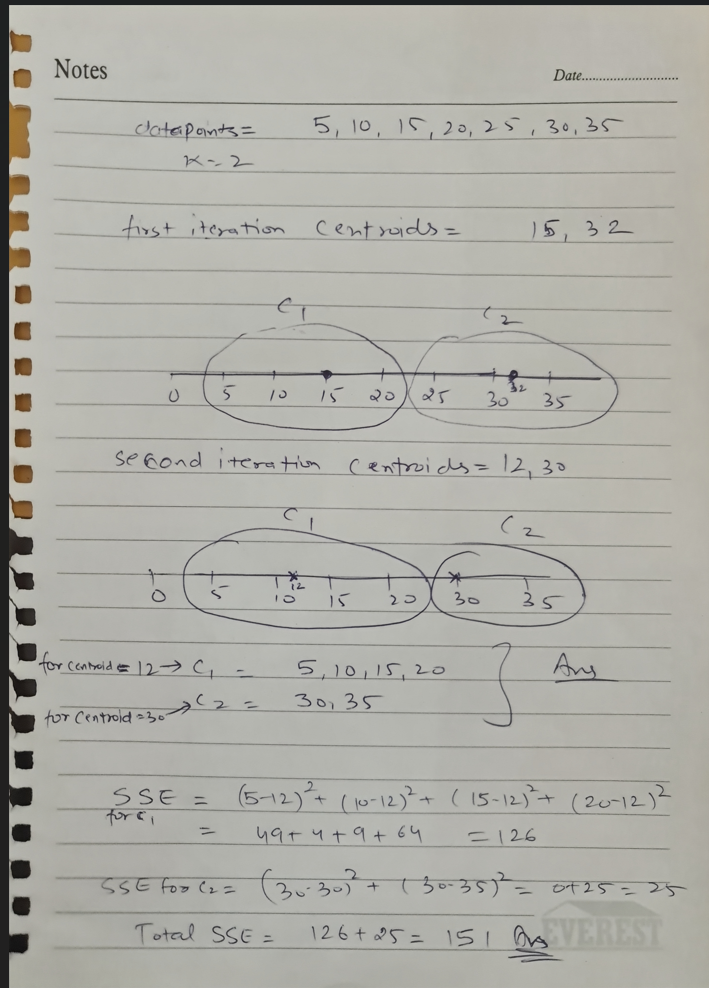
1. A set of one-dimensional data points is given to you: 5, 10, 15, 20, 25, 30, 35. Assume that k = 2 and that the first set of random centroid is 15, 32, and that the second set is 12, 30.

a) Using the k-means method, create two clusters for each set of centroid described above.

b) For each set of centroid values, calculate the SSE.

Ans:



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

Ans: **Association Analysis concept in Market Basked Research allows you to find two or more associated products that are bought together or have higher probability of getting sold together. Unsupervised algorithms can be applied to create cluster for those items which have higher chances of getting sold.**

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

Ans: **Apriori algorithm is used to determine the association between items which are sold together. For example an association can be made between milk and bread as these two items are bought most frequently. Another example is the set of pages which are visited most frequently one after the other.**

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

Ans: **different distance measurements can be used to calculate the distance between clusters such as Euclidean distance, Manhattan Distance or Hamming distance. The iteration stops after finding the best K value. K value is found be forming a dendogram on which we try to find the longest vertical line where no horizontal passes through it. After choosing the vertical line, we create a horizontal line an find the points on which vertical lines intersect the horizontal lines. The number of points represent the best value of K.**

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

Ans: **Centroids are recomputed based on the average value of the cluster points within a cluster.**

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

Ans: **in K means, we use Elbow method to determine the right number of clusters. We find the WCSS values for each K value. WCSS stands for within cluster sum of squares. We take that value of K where the WCSS changes abruptly.**

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

Ans: **K means advantage includes faster computing and easy to implement and disadvantage is no exclusion of outliers/noise like DBScan which creates separate clusters for noise or outliers so a decision can be made easily where they need to be excluded or not.**

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

Ans: **clustering concept revolves around grouping or clustering the data points based on similarities. Similar points tend to be in a same cluster. K-means and Hierarchical clustering initialize centroids to be able to find similar data points and create clusters.**

Chart, scatter chart

Description automatically generated

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?

Ans: **Considering the points after the first iteration is below:**

**C1: {(2,2), (4,4), (6,6)}**

**C2: {(0,4), (4,0)}**

**C3: {(5,5), (9,9)}**

**Then the centroids position in the 2nd iteration would be the average value of all the points within the clusters. So the value would be:**

**C1: (4,4), C2: (2,2), C3: (7,7)**

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.

Ans:

A picture containing venn diagram

Description automatically generated

5 Clusters have been formed to put 20 defect points into different clusters. The diagram would look like the above.