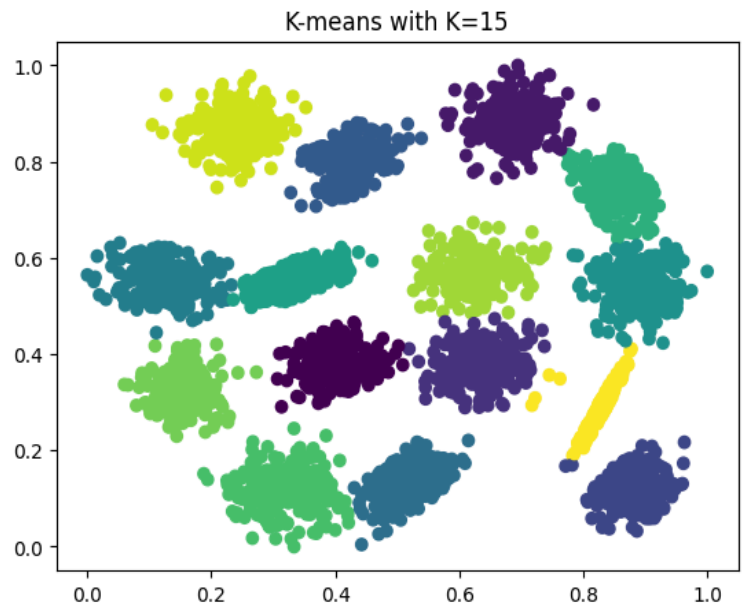# Question # 1

(A)
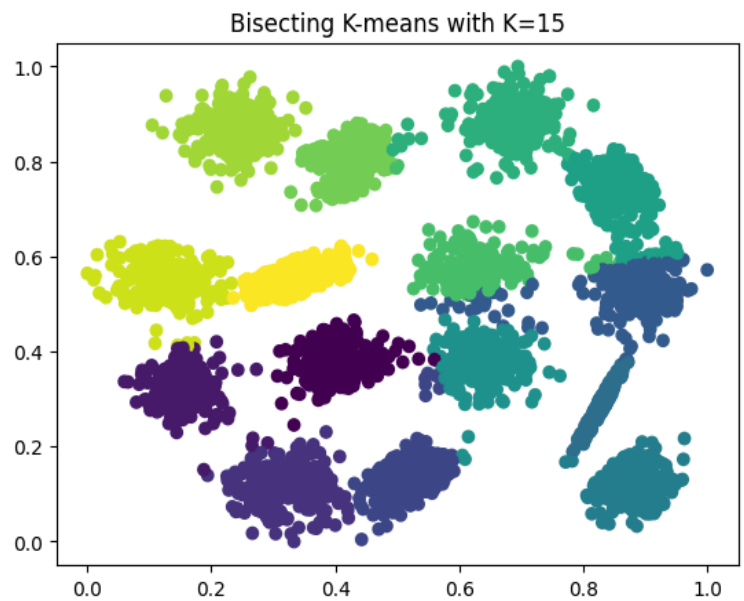
K-means

- K = 15
- Execution Time: 0.0151 seconds
- WCSS = 10.2871
- BCSS = 655.0790
- SSE: 665.3661
- Silhouette Score: 0.7116
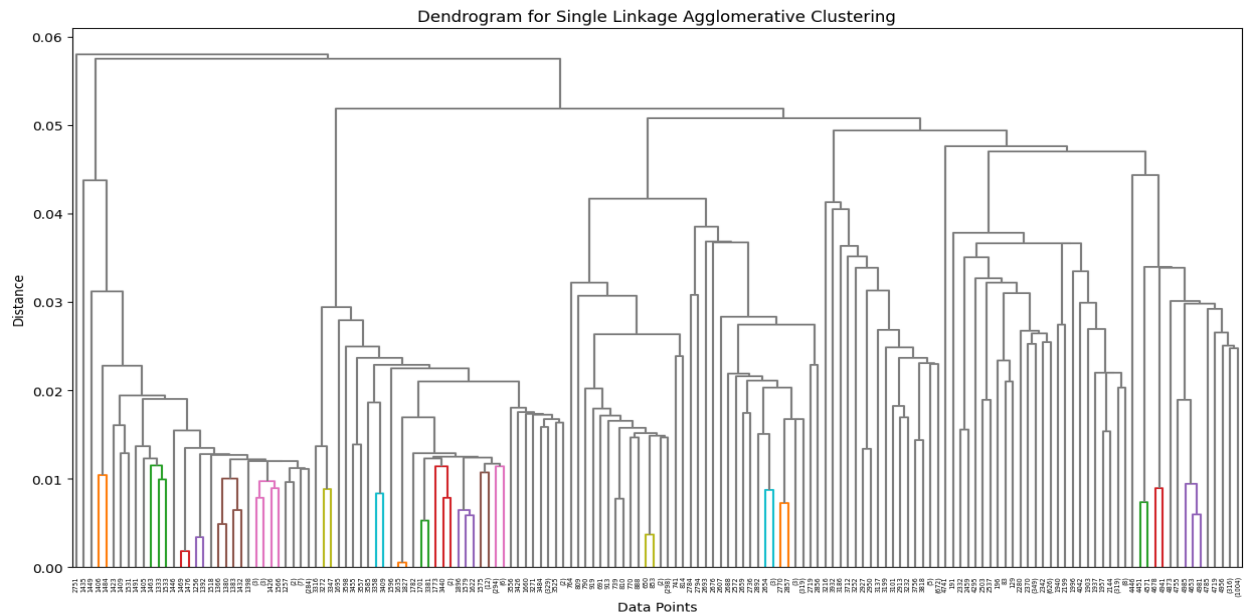

K-means with K=15

Bisecting K-means

- K = 15
- Execution Time: 0.0137 seconds
- WCSS = 12.9334
- BCSS = 652.9139
- SSE: 665.8473
- Silhouette Score: 0.6905

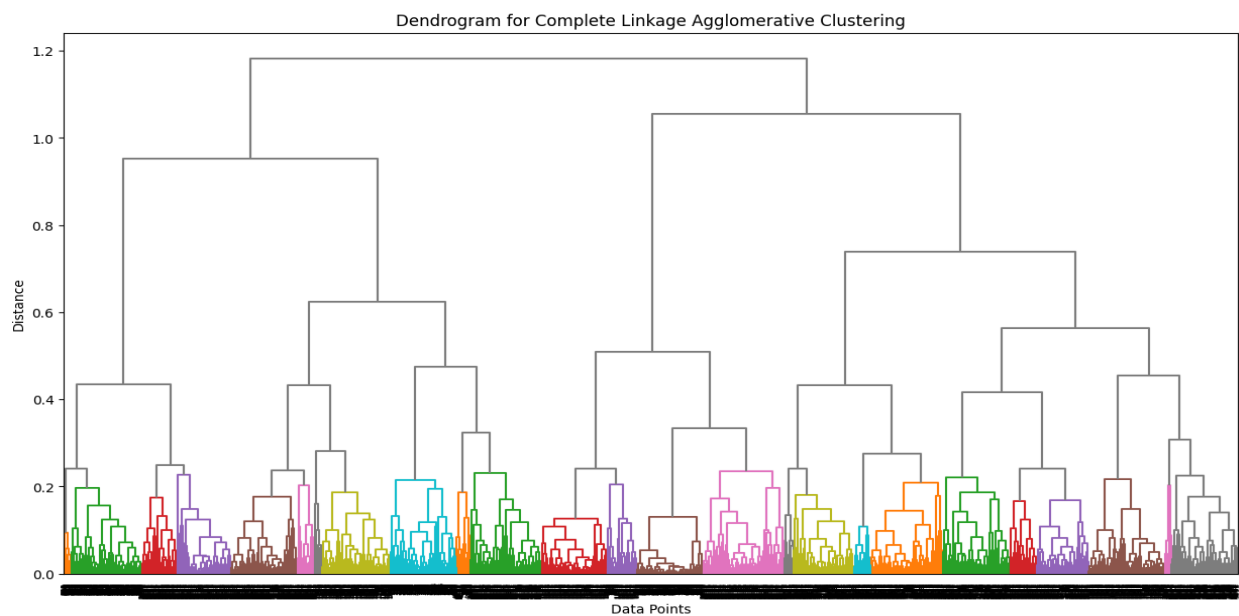Bisecting K-means with K=15

Hierarchical (single-link)

- K = 15
- Execution Time: 0.3742 seconds
- Silhouette Score: -0.0469

Dendrogram for Single Linkage Agglomerative Clustering

Hierarchical (complete-link)

- K = 15
- Execution Time: 1.3388 seconds
- Silhouette Score: 0.6832

Dendrogram for Complete Linkage Agglomerative Clustering

Hierarchical (average-link)

- K = 15
- Execution Time: 1.9115 seconds
- Silhouette Score: 0.7085

Dendrogram for Average Linkage Agglomerative Clustering
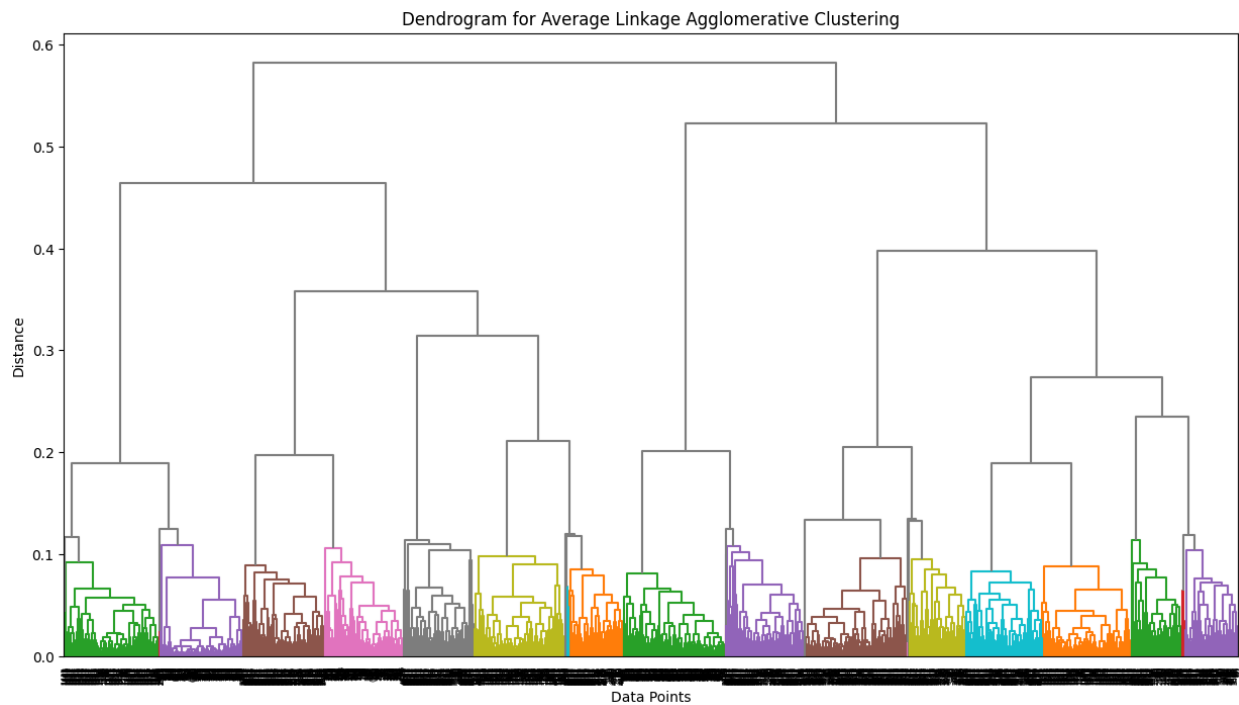
DBSCAN

- eps = 0.016
- min_points = 5
- Execution Time: 0.0392 seconds
- Silhouette Score: 0.7116

DBSCAN with eps=0.016
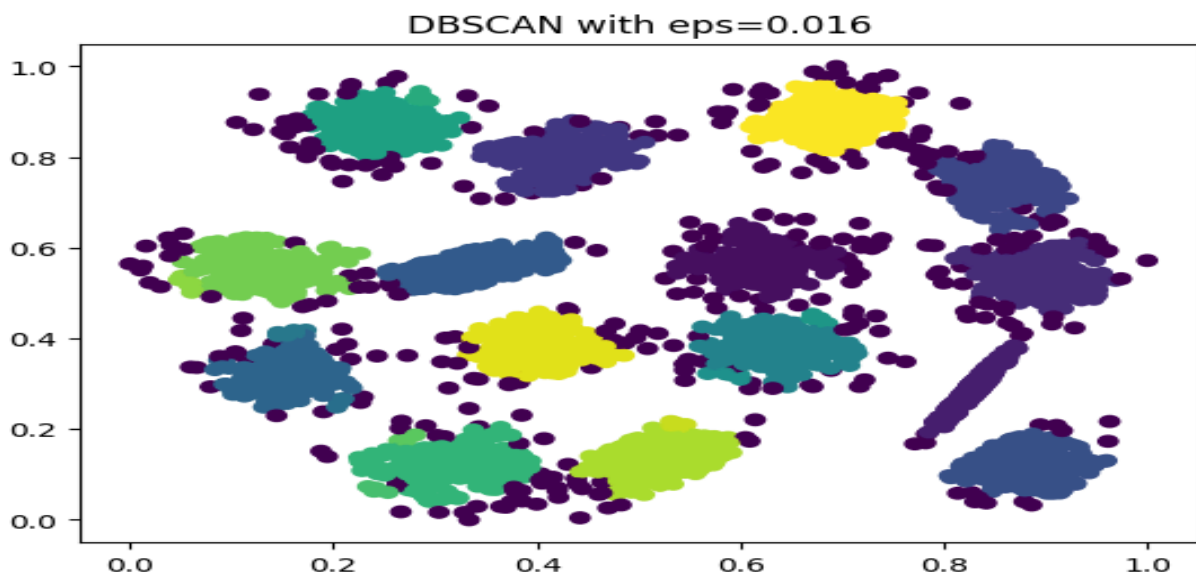
(B)

K-means (3 clusters)
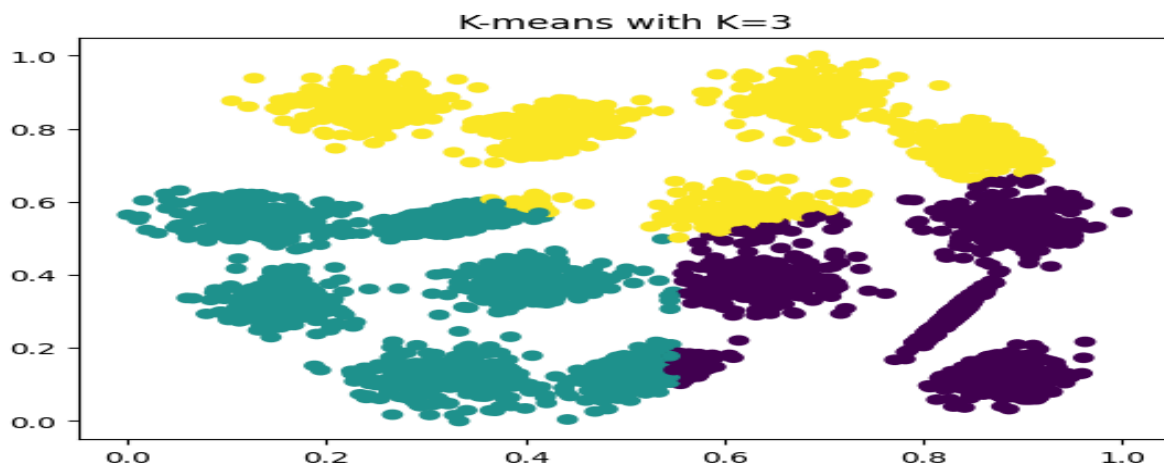
- **Run 1:**
    - Execution Time: 0.0096 seconds
    - Number of iterations: 7
    - WCSS = 249.8162
    - BCSS = 415.5964
    - SSE: 665.4127
    - Silhouette Score: 0.4072

- **Run 2:**
    - Execution Time: 0.0106 seconds
    - Number of iterations: 6
    - WCSS = 250.9309
    - BCSS = 414.3736
    - SSE: 665.3046
    - Silhouette Score: 0.4178

- **Run 3:**
    - Execution Time: 0.0107 seconds
    - Number of iterations: 6
    - WCSS = 246.5607
    - BCSS = 418.7799
    - SSE: 665.3407
    - Silhouette Score: 0.4151

K-means with K=3
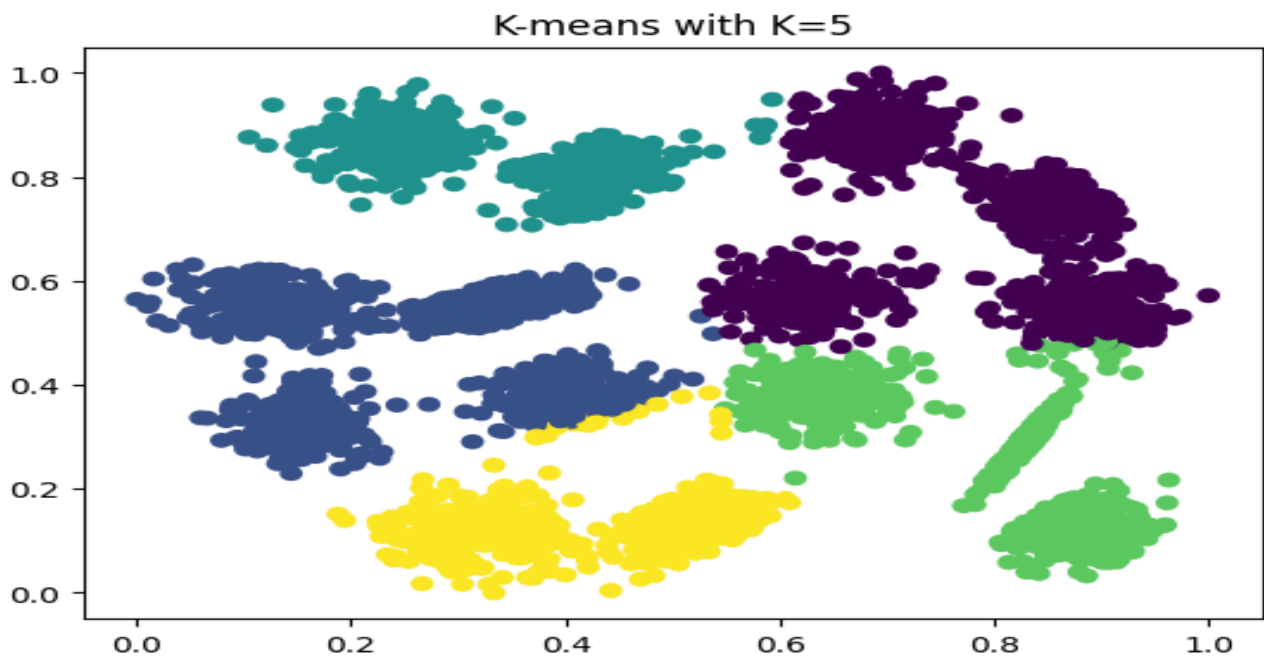
K-means (5 clusters)

- **Run 1:**
  - Execution Time: 0.0138 seconds
  - Number of iterations: 8
  - WCSS = 134.4806
  - BCSS = 530.8383
  - SSE: 665.3190
  - Silhouette Score: 0.4492

- **Run 2:**
  - Execution Time: 0.0148 seconds
  - Number of iterations: 7
  - WCSS = 135.0286
  - BCSS = 530.2922
  - SSE: 665.3208
  - Silhouette Score: 0.4399

- **Run 3:**
  - Execution Time: 0.0136 seconds
  - Number of iterations: 9
  - WCSS = 120.9673
  - BCSS = 544.3614
  - SSE: 665.3287
  - Silhouette Score: 0.4658

K-means (8 clusters)
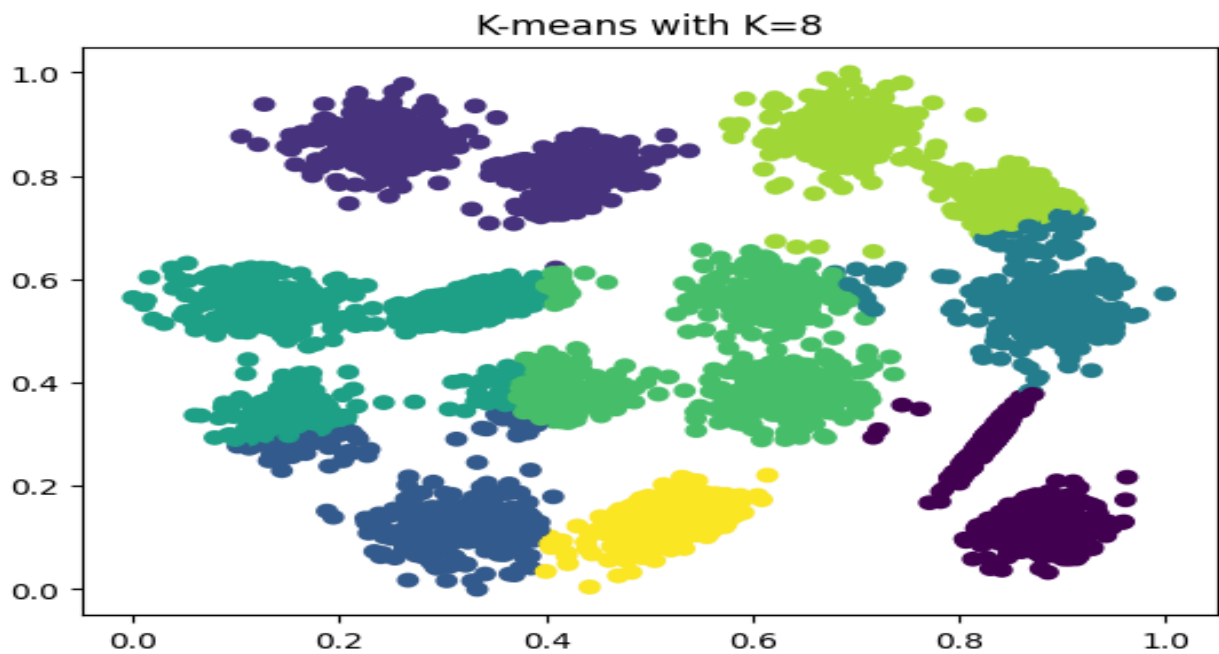
- **Run 1:**
    - Execution Time: 0.0157 seconds
    - Number of iterations: 8
    - WCSS = 60.8940
    - BCSS = 604.5608
    - SSE: 665.4548
    - Silhouette Score: 0.5347

- **Run 2:**
    - Execution Time: 0.0130 seconds
    - Number of iterations: 7
    - WCSS = 60.8943
    - BCSS = 604.3547
    - SSE: 665.2490
    - Silhouette Score: 0.5345

- **Run 3:**
    - Execution Time: 0.0113 seconds
    - Number of iterations: 9
    - WCSS = 60.8941
    - BCSS = 604.5156
    - SSE: 665.4097
    - Silhouette Score: 0.5347

K-means (12 clusters)
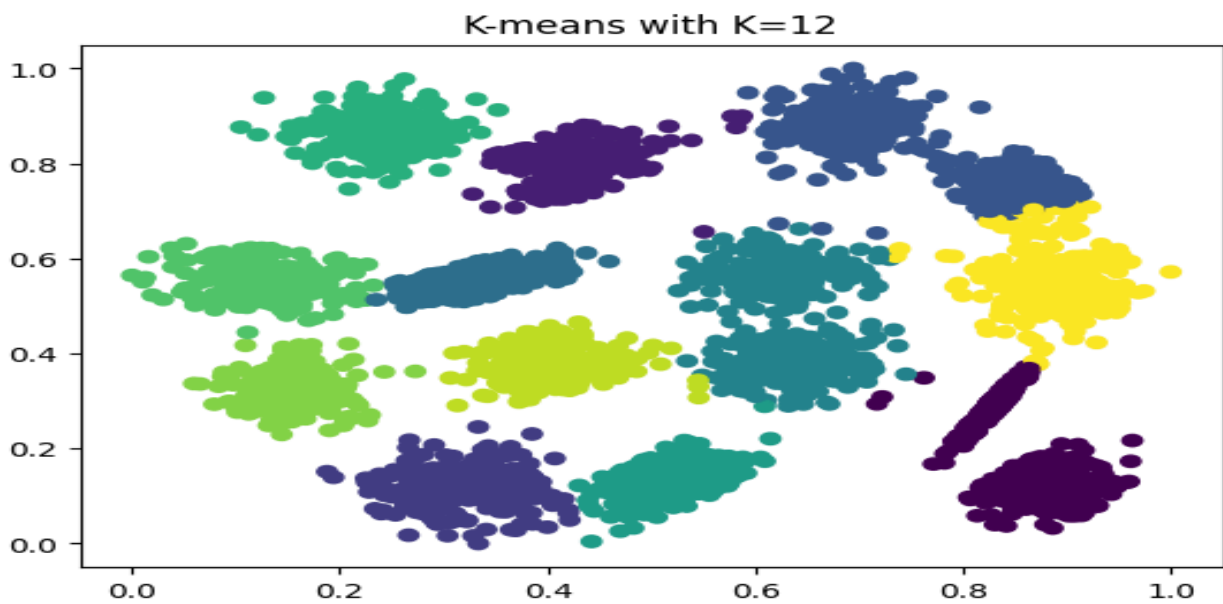
- **Run 1:**
    - Execution Time: 0.0192 seconds
    - Number of iterations: 3
    - WCSS = 28.6926
    - BCSS = 636.7573
    - SSE: 665.4499
    - Silhouette Score: 0.6384

- **Run 2:**
    - Execution Time: 0.0181 seconds
    - Number of iterations: 5
    - WCSS = 28.2016
    - BCSS = 637.1644
    - SSE: 665.3661
    - Silhouette Score: 0.6401

- **Run 3:**
    - Execution Time: 0.0160 seconds
    - Number of iterations: 4
    - WCSS = 27.4861
    - BCSS = 637.8800
    - SSE: 665.3661
    - Silhouette Score: 0.6366

K-means with K=12

K-means (20 clusters)
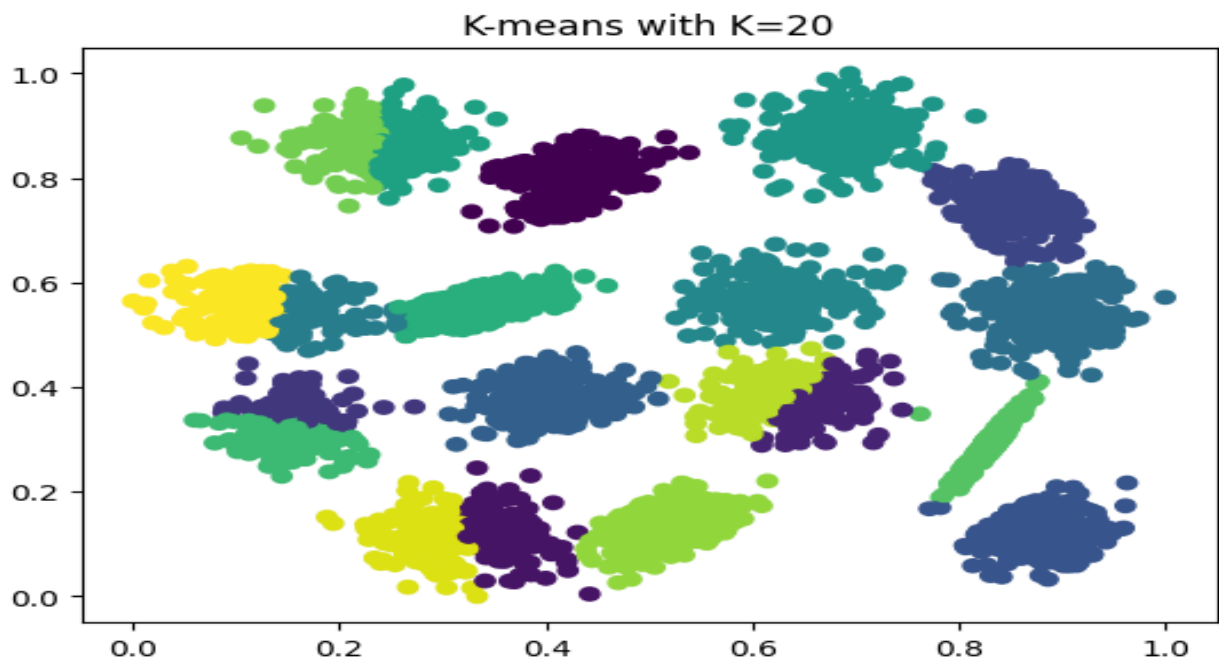
- **Run 1:**
    - Execution Time: 0.0280 seconds
    - Number of iterations: 14
    - WCSS = 9.1251
    - BCSS = 656.2219
    - SSE: 665.3470
    - Silhouette Score: 0.5785

- **Run 2:**
    - Execution Time: 0.0238 seconds
    - Number of iterations: 11
    - WCSS = 9.1976
    - BCSS = 656.1666
    - SSE: 665.3642
   - Silhouette Score: 0.5751

- **Run 3:**
    - Execution Time: 0.0241 seconds
    - Number of iterations: 8
    - WCSS = 9.3327
    - BCSS = 656.2640
    - SSE: 665.5968
    - Silhouette Score: 0.5979

K-means with K=20

Elbow Method Results

Using the Plot obtained by plotting the WCSS against the No. of Clusters, it can be observed that the Elbow Point lies around **k = 15** clusters which is supported by the experimental results recorded earlier in the report.

Elbow Method: SSE vs Number of Clusters
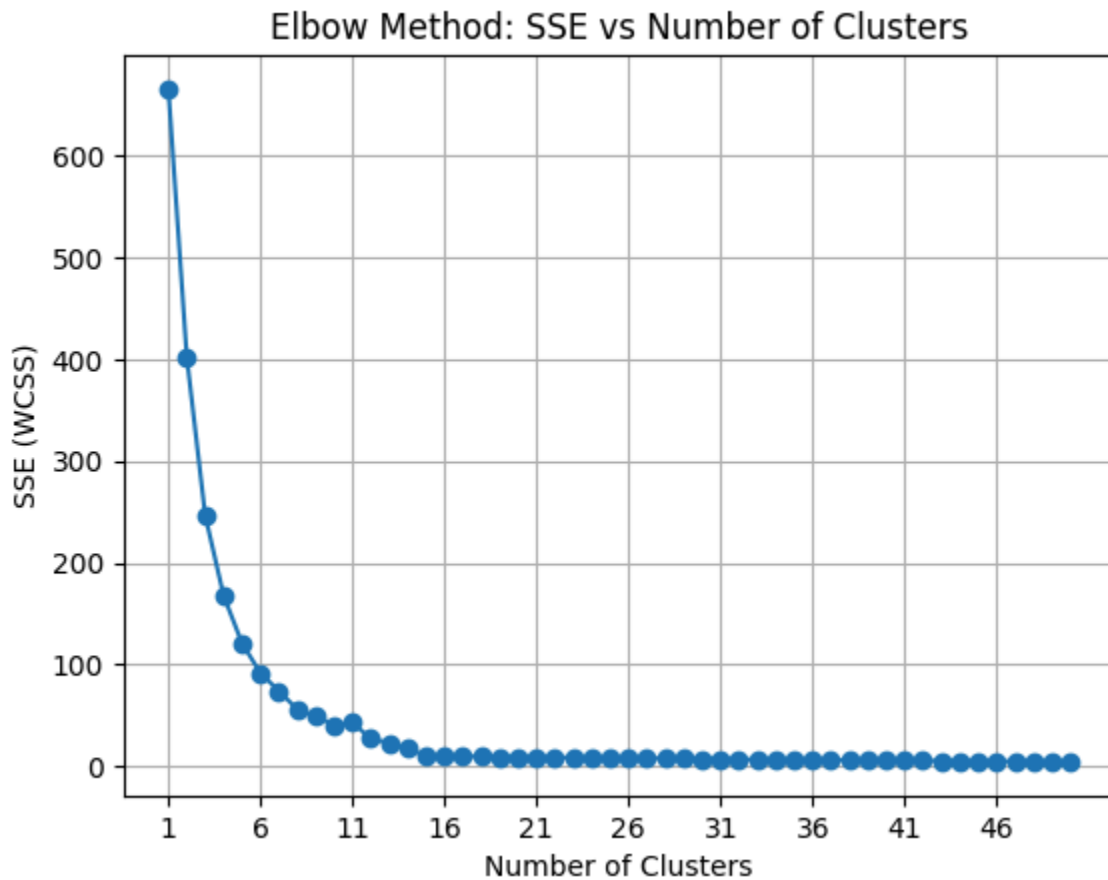
(C)

Best Values of K

Using Elbow Method & Performing Kmeans using different values, it is concluded that K = 15 would result in optimal clustering for the given dataset.

Concluding Remarks

Kmeans and Bisecting Kmeans algorithms performed good clustering when given k = 15. DBSCAN was successful in clustering and removing the noise adequately. Hierarchical Clustering with complete & average link also clustered the data nicely for k = 15. However, Hierarchical clustering with single linkage performed very poor clustering, which is probably due to the fact that it merged bigger, denser clusters that were nearby based on the minimum distance between them.

Question # 2

Scenario

Can we identify distinct student profiles based on reading scientific books, taking notes in class, and attending conferences?

Selected Attributes:

- Taking notes in classes
- Reading frequency (non-scientific books/journals)
- Attendance to the seminars/conferences related to the department
- Cumulative grade point average in the last semester

Preprocessing

1) Z-Score Normalization: This step is being done to ensure that all the attributes weigh in the data equally even if their magnitudes vary significantly
2) Encoding: The selected attributes are already encoded which makes this step redundant

Elbow Method Results

Using the Plot obtained by plotting the WCSS against the No. of Clusters, it can be observed that the Elbow Point lies around **k = 26**. We will be running K-means for k = 21, 26, 30.

Elbow Method: SSE vs Number of Clusters
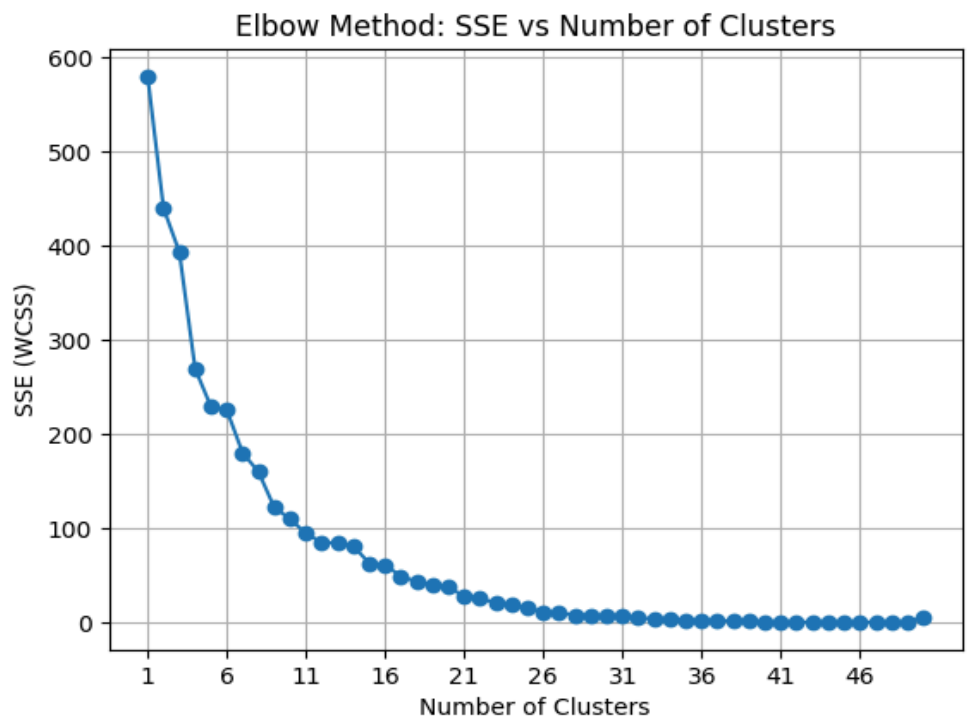
K-means (21 clusters)

- **Run 1:**
  - Execution Time: 0.0077 seconds
  - Number of iterations: 2
  - WCSS = 32.9591
  - BCSS = 3702.9719
  - SSE: 3735.9310
  - Silhouette Score: 0.7306

- **Run 2:**
  - Execution Time: 0.0090 seconds
  - Number of iterations: 3
  - WCSS = 37.7039
  - BCSS = 3698.2271
  - SSE: 3735.9310
  - Silhouette Score: 0.7594

- **Run 3:**
  - Execution Time: 0.0084 seconds
  - Number of iterations: 2
  - WCSS = 28.1427
  - BCSS = 3707.7883
  - SSE: 3735.9310
  - Silhouette Score: 0.6905

K-means (26 clusters)

- **Run 1:**
  - Execution Time: 0.0077 seconds
  - Number of iterations: 2
  - WCSS = 16.1257
  - BCSS = 3719.8054
  - SSE: 3735.9310
  - Silhouette Score: 0.7819

- **Run 2:**
  - Execution Time: 0.0125 seconds
  - Number of iterations: 2
  - WCSS = 12.7034
  - BCSS = 3723.2276
 - SSE: 3735.9310
  - Silhouette Score: 0.8158

- **Run 3:**
  - Execution Time: 0.0092 seconds
  - Number of iterations: 2
  - WCSS = 13.1783
  - BCSS = 3722.7527
  - SSE: 3735.9310
  - Silhouette Score: 0.8327

K-means (30 clusters)

- **Run 1:**
  - Execution Time: 0.0100 seconds
  - Number of iterations: 2
  - WCSS = 5.5518
  - BCSS = 3730.3793
  - SSE: 3735.9310
  - Silhouette Score: 0.9014

- **Run 2:**
  - Execution Time: 0.0094 seconds
  - Number of iterations: 2
  - WCSS = 5.6708
  - BCSS = 3730.2603
  - SSE: 3735.9310
  - Silhouette Score: 0.8992

- **Run 3:**
  - Execution Time: 0.0114 seconds
  - Number of iterations: 2
  - WCSS = 6.8034
  - BCSS = 3729.1276
  - SSE: 3735.9310
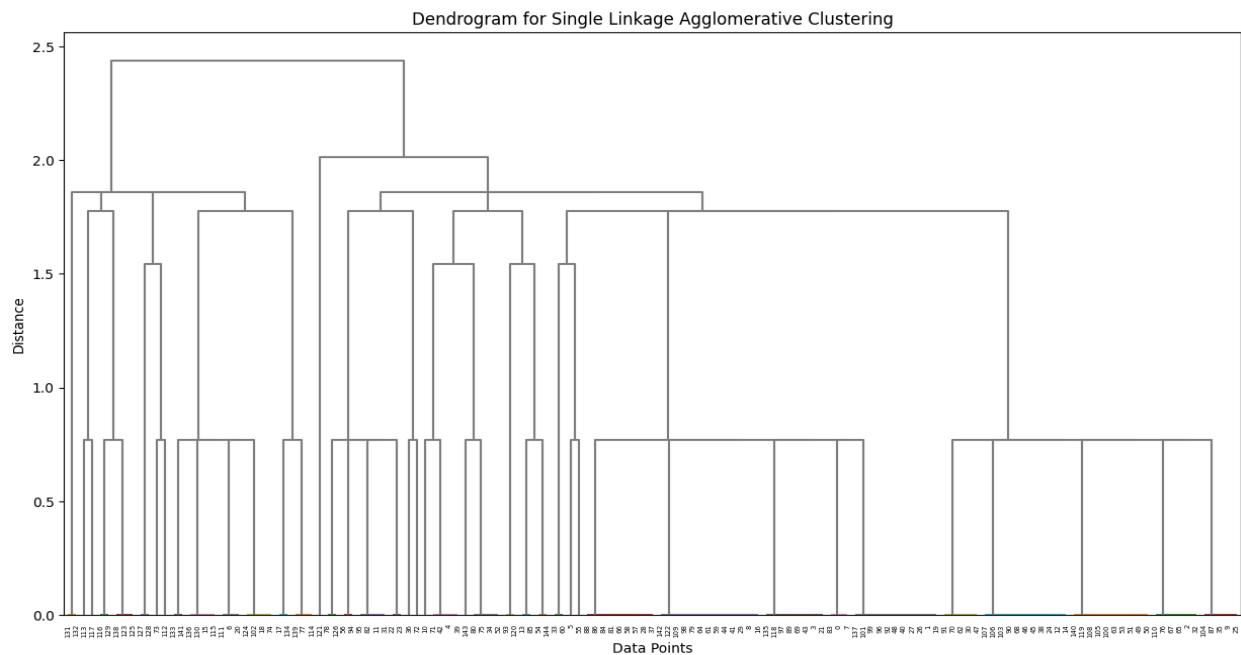 - Silhouette Score: 0.8796

Hierarchical (single-link, k = 22)

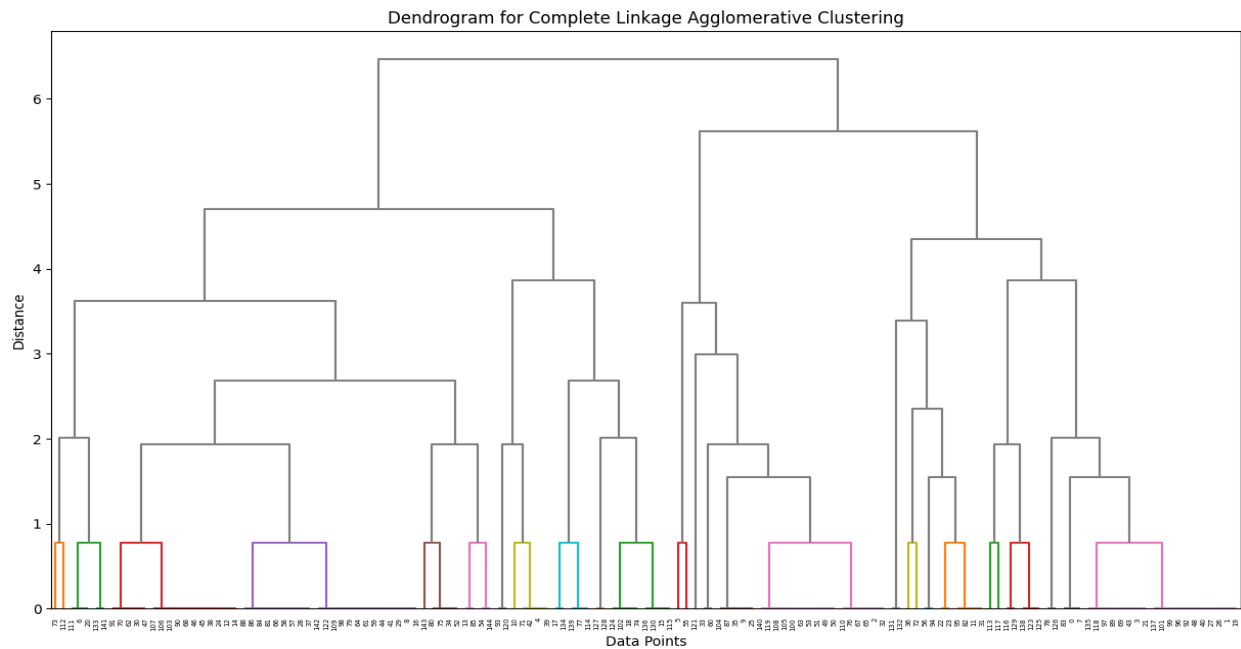- Execution Time: 0.0031 seconds
- Silhouette Score: 0.5337

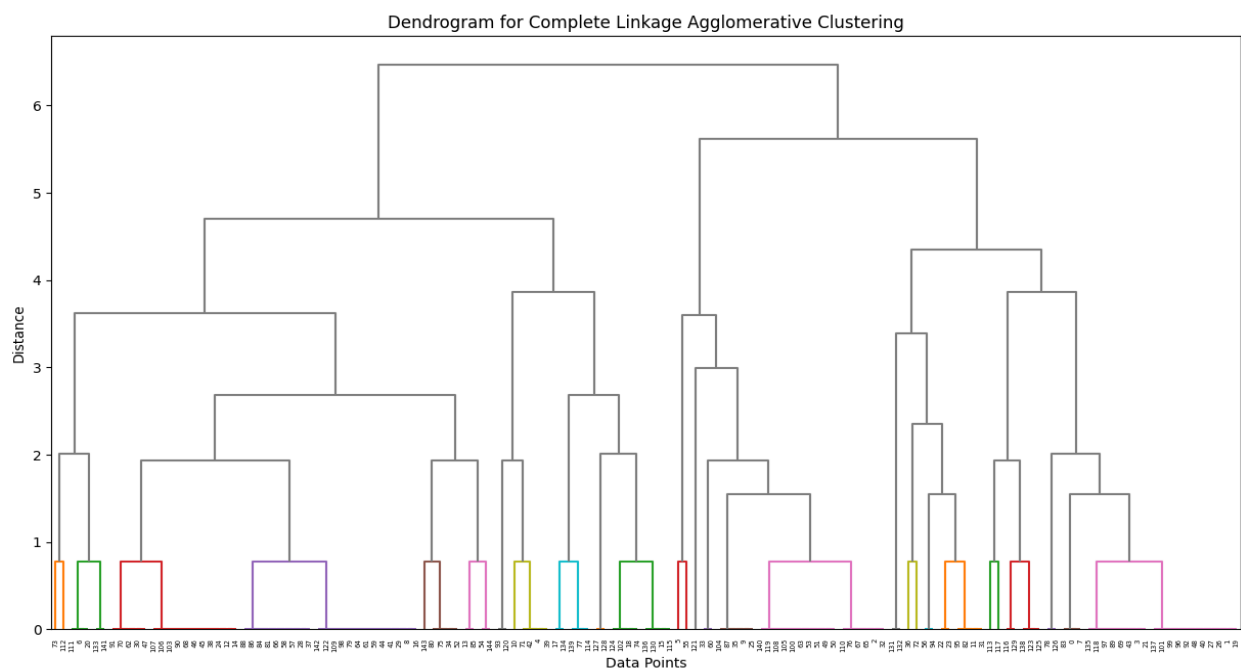Dendrogram for Single Linkage Agglomerative Clustering
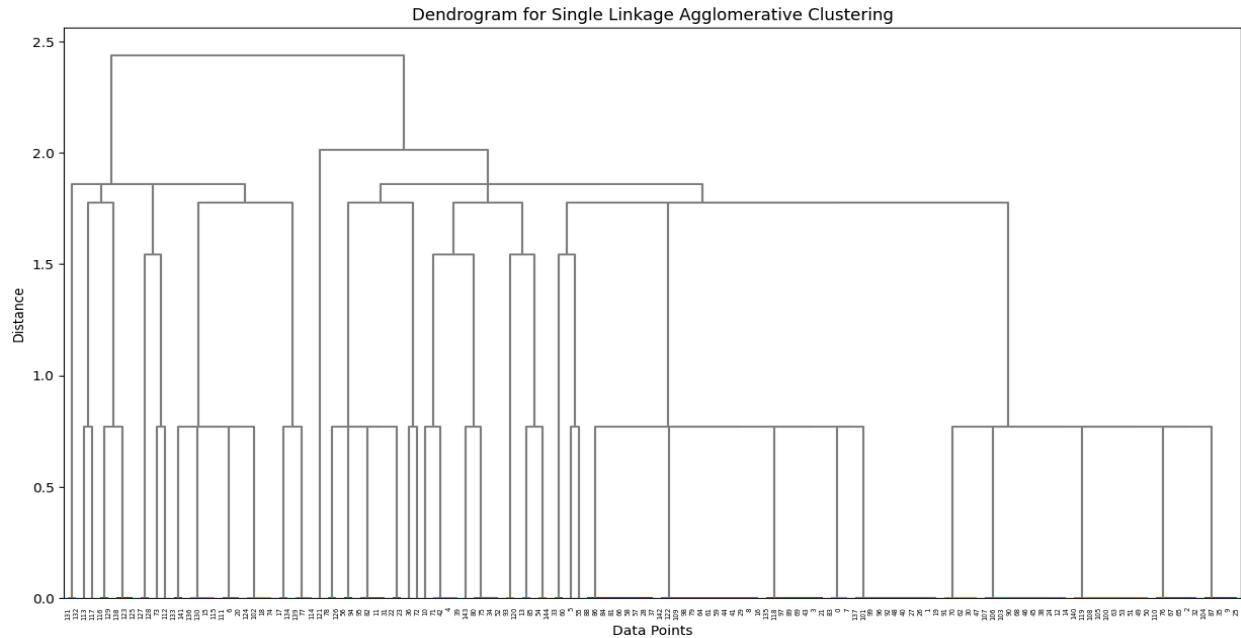
Hierarchical (single-link, k = 26)

- Execution Time: 0.0034 seconds
- Silhouette Score: 0.5686

Dendrogram for Single Linkage Agglomerative Clustering

Hierarchical (complete-link, k = 22)

- Execution Time: 0.0032 seconds
- Silhouette Score: 0.6795

Dendrogram for Complete Linkage Agglomerative Clustering

Hierarchical (complete-link, k = 26)

- Execution Time: 0.0037 seconds
- Silhouette Score: 0.7185

Dendrogram for Complete Linkage Agglomerative Clustering

Hierarchical (average-link, k = 22)

- Execution Time: 0.0036 seconds
- Silhouette Score: 0.6633

Dendrogram for Average Linkage Agglomerative Clustering
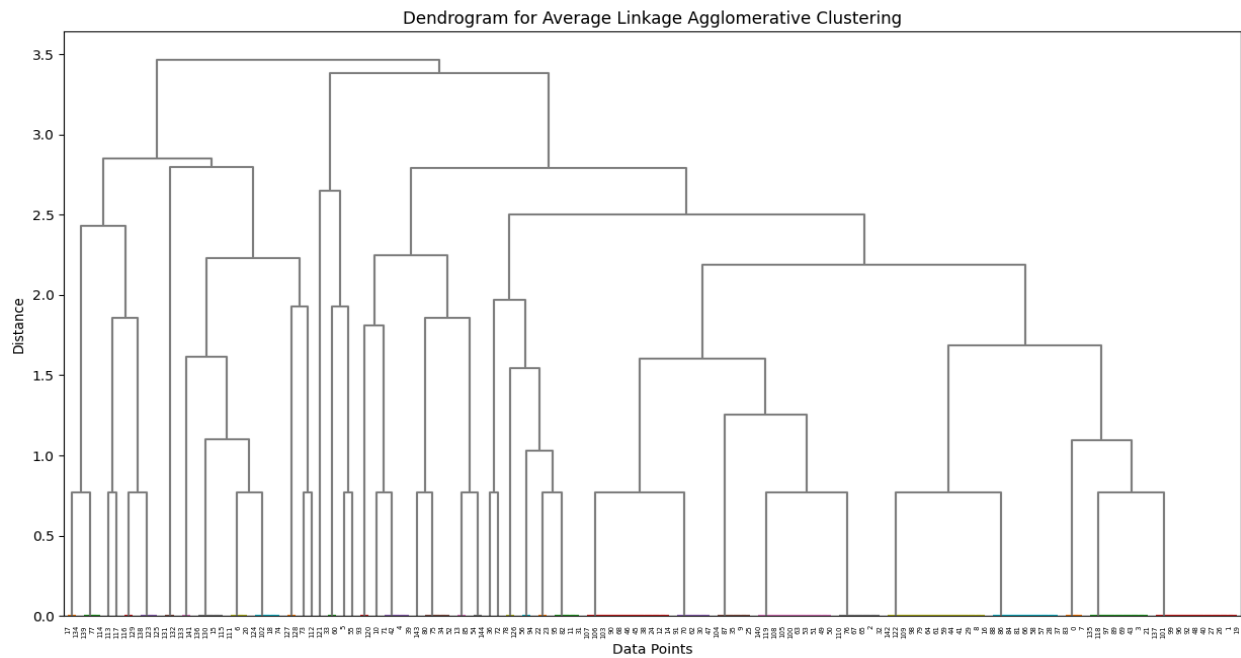
Hierarchical (average-link, k = 26)

- Execution Time: 0.0039 seconds
- Silhouette Score: 0.7063

Dendrogram for Average Linkage Agglomerative Clustering
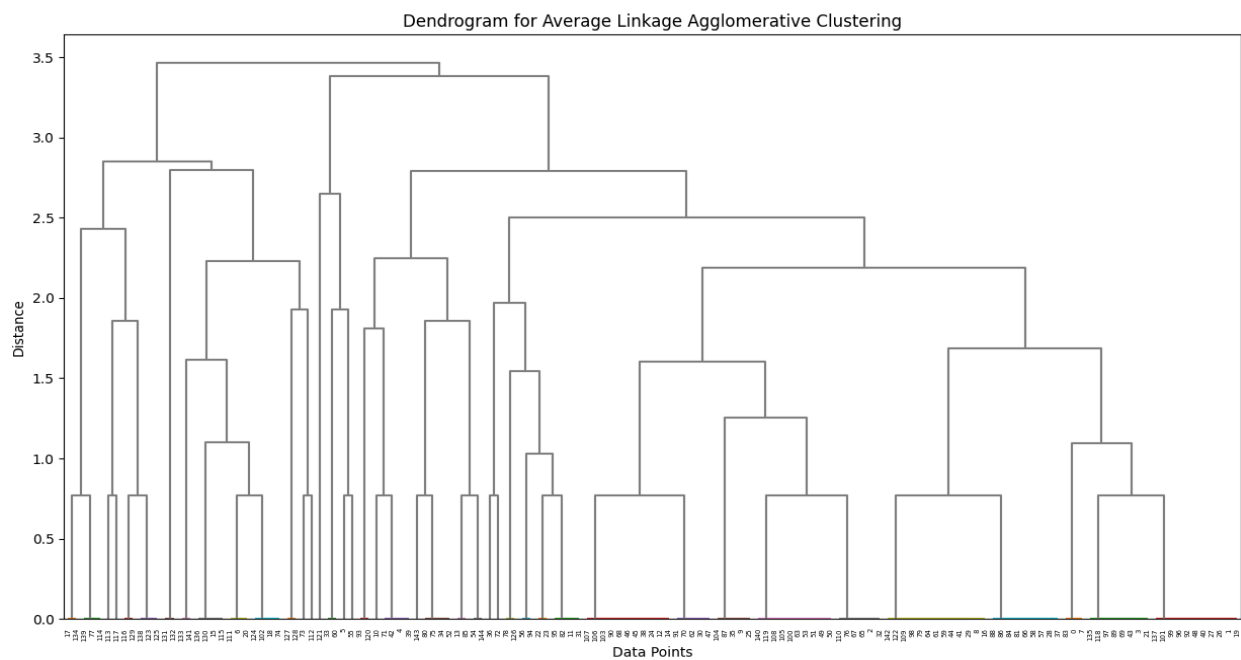
Results & Comparison

The comparison between K-Means and Hierarchical Clustering highlights a trade-off between efficiency and clustering quality. K-Means achieved a **higher silhouette score (0.81012)**, indicating better-defined clusters, but required **0.01718 seconds** on average, making it slower. In contrast, Hierarchical Clustering was significantly faster, averaging **0.003483 seconds**, but resulted in **a lower silhouette score (0.64498)**, suggesting weaker cluster separation. This suggests that K-Means is preferable when cluster quality is the priority, whereas Hierarchical Clustering is more suitable when speed is a key concern, especially for smaller datasets or when interpretability through dendrograms is needed.