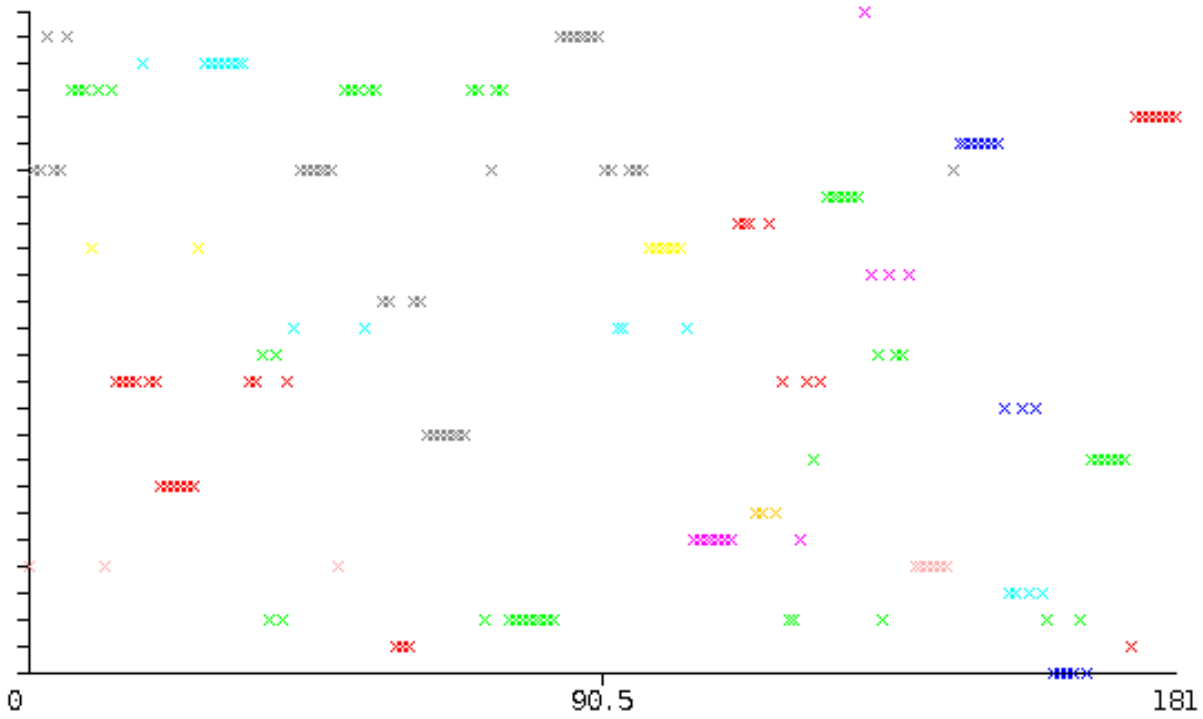
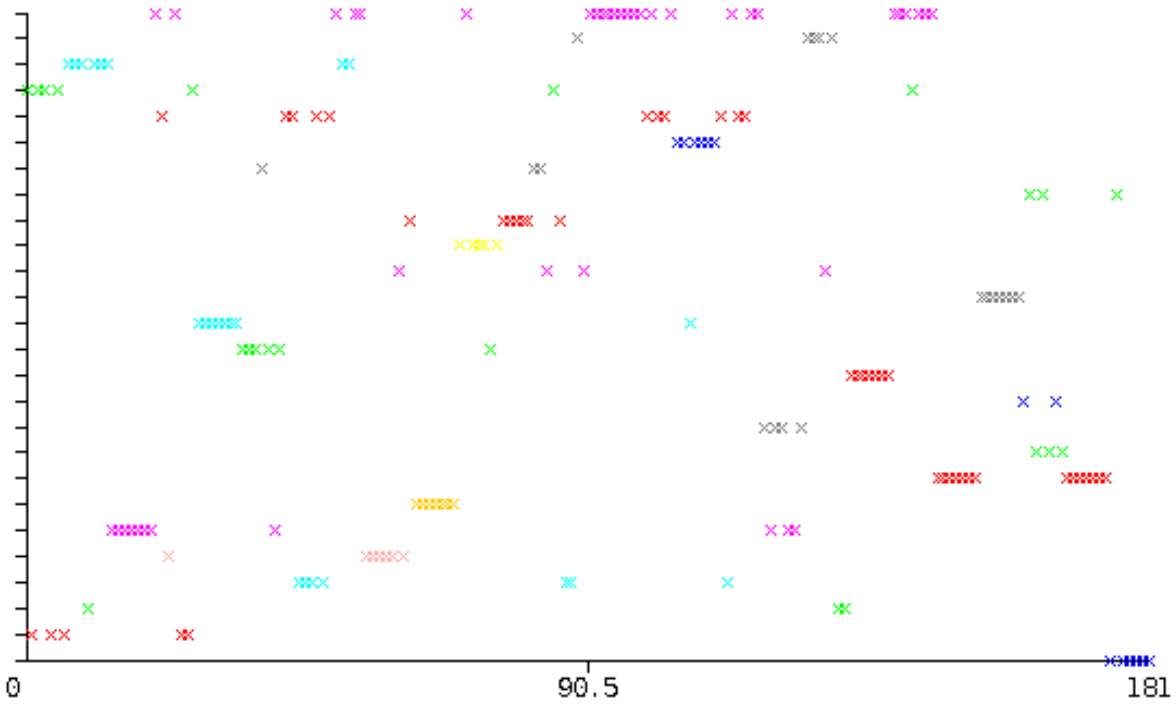
Assignment3
Youssef Ayman Taher Kandil, 34-1920

Clustering Algorithm Graphs:

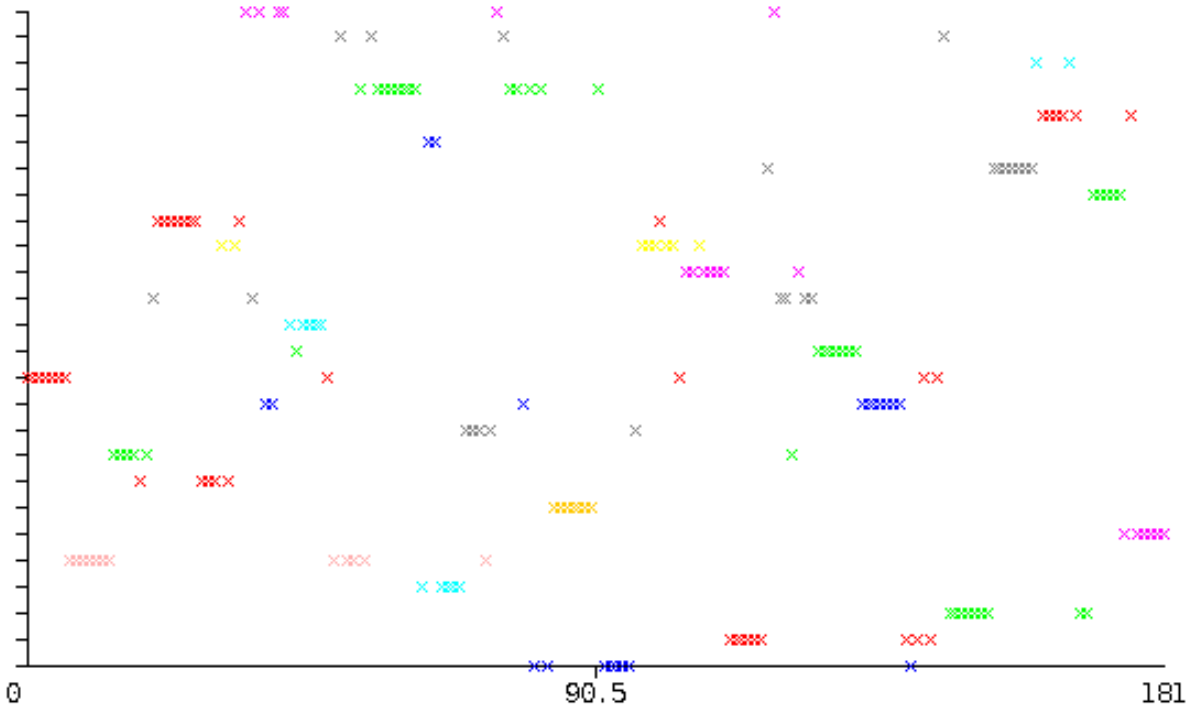
1. Canopy:



2. EM:

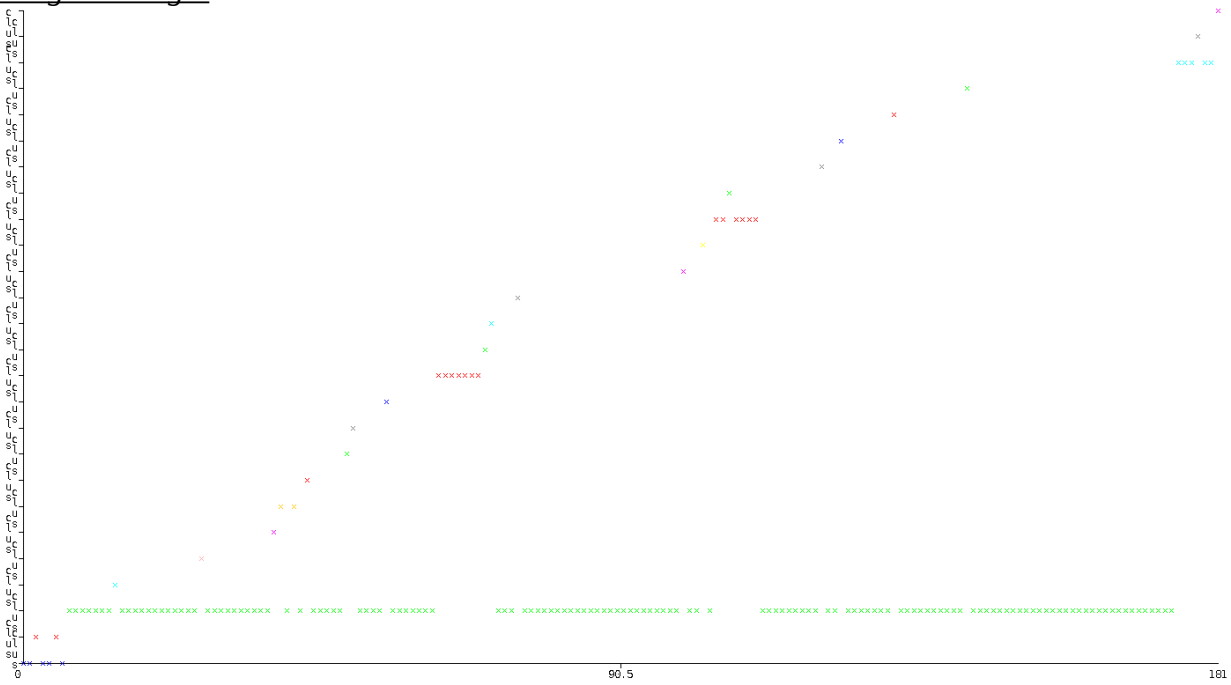


3. FarthestFirst:

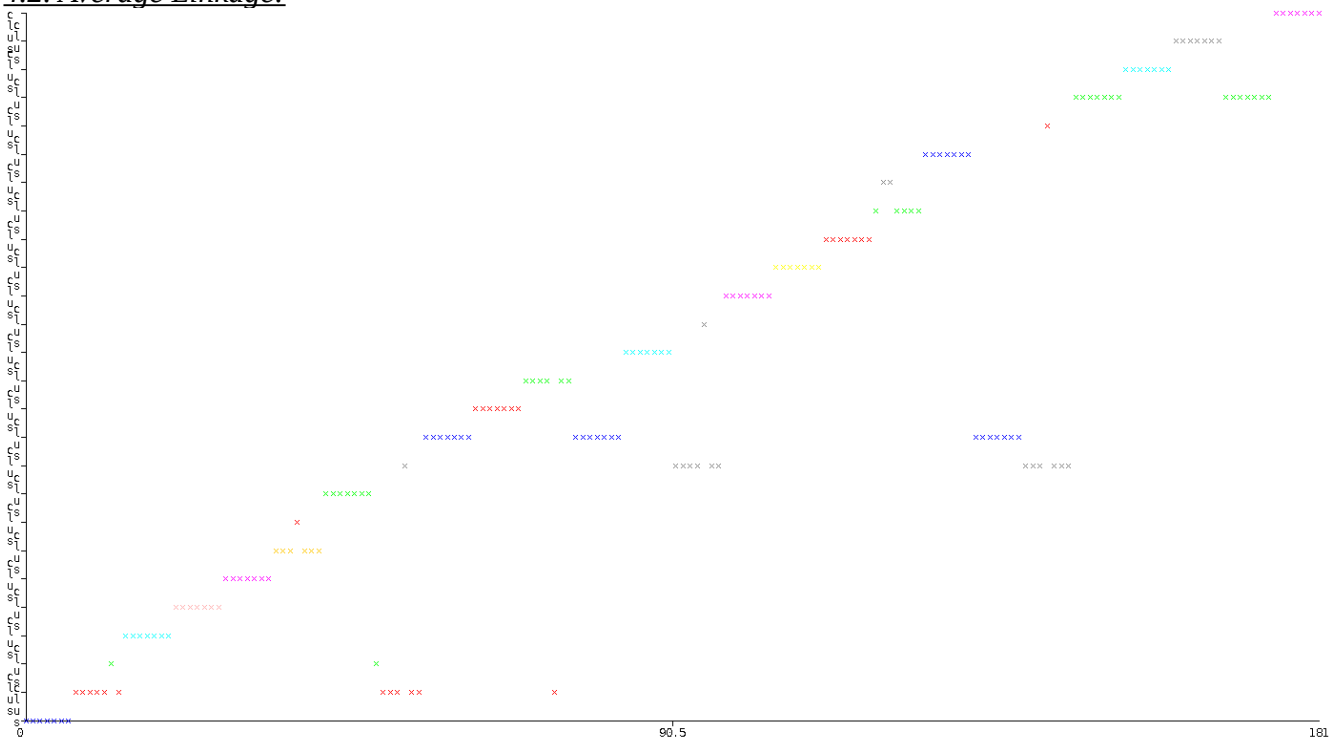


4. Hierarchical:

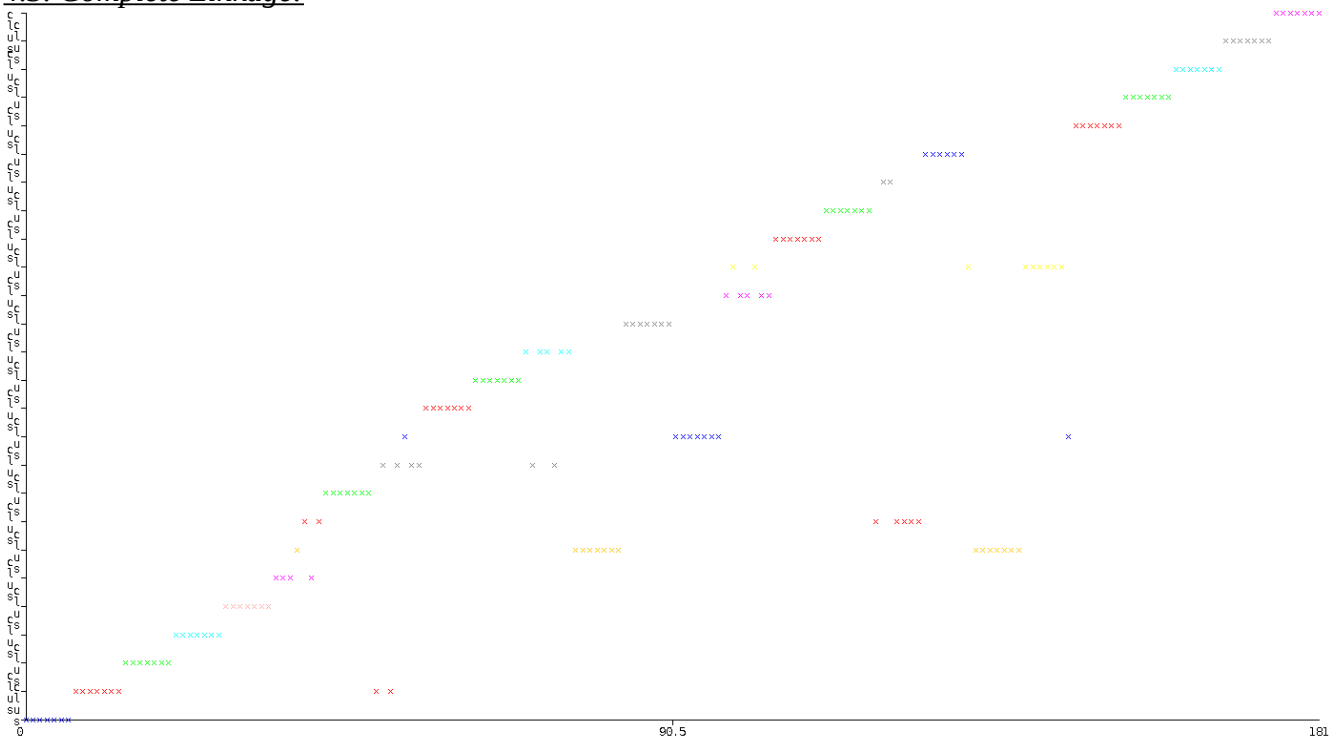
4.1. Single Linkage:



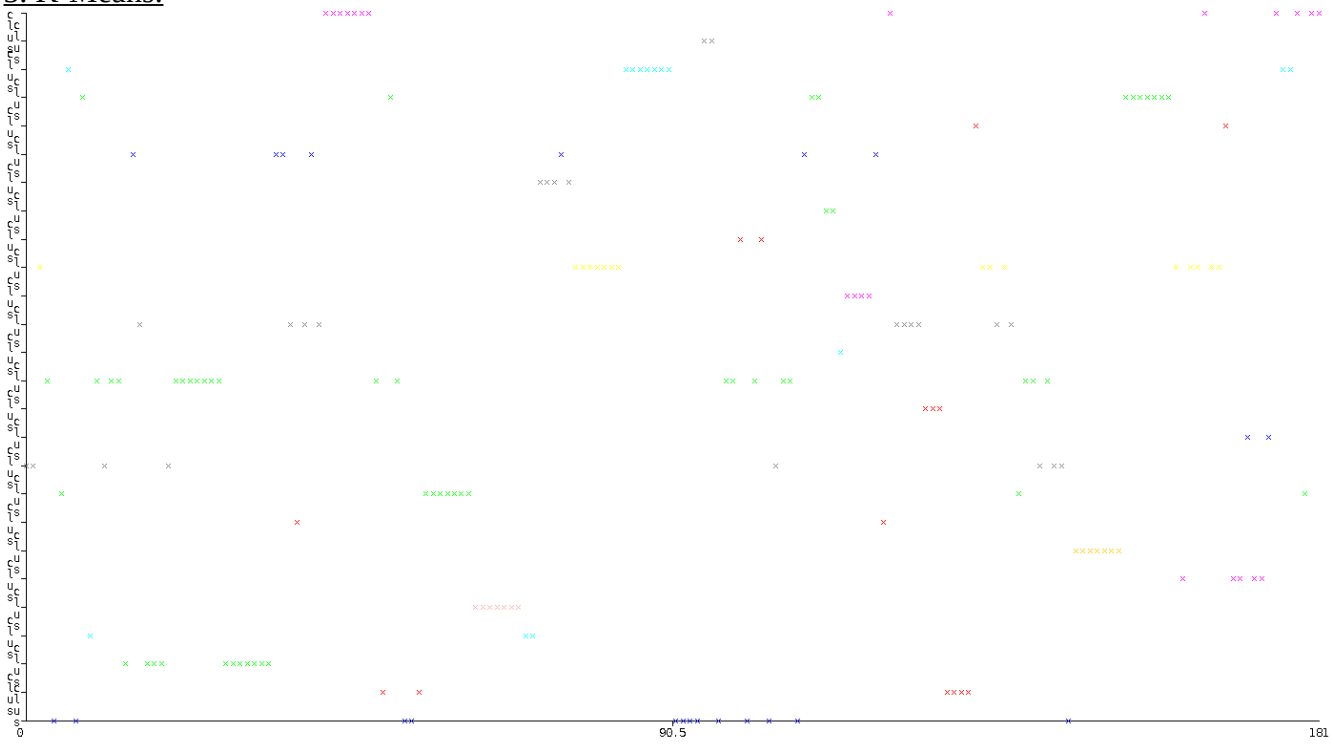
4.2. Average Linkage:



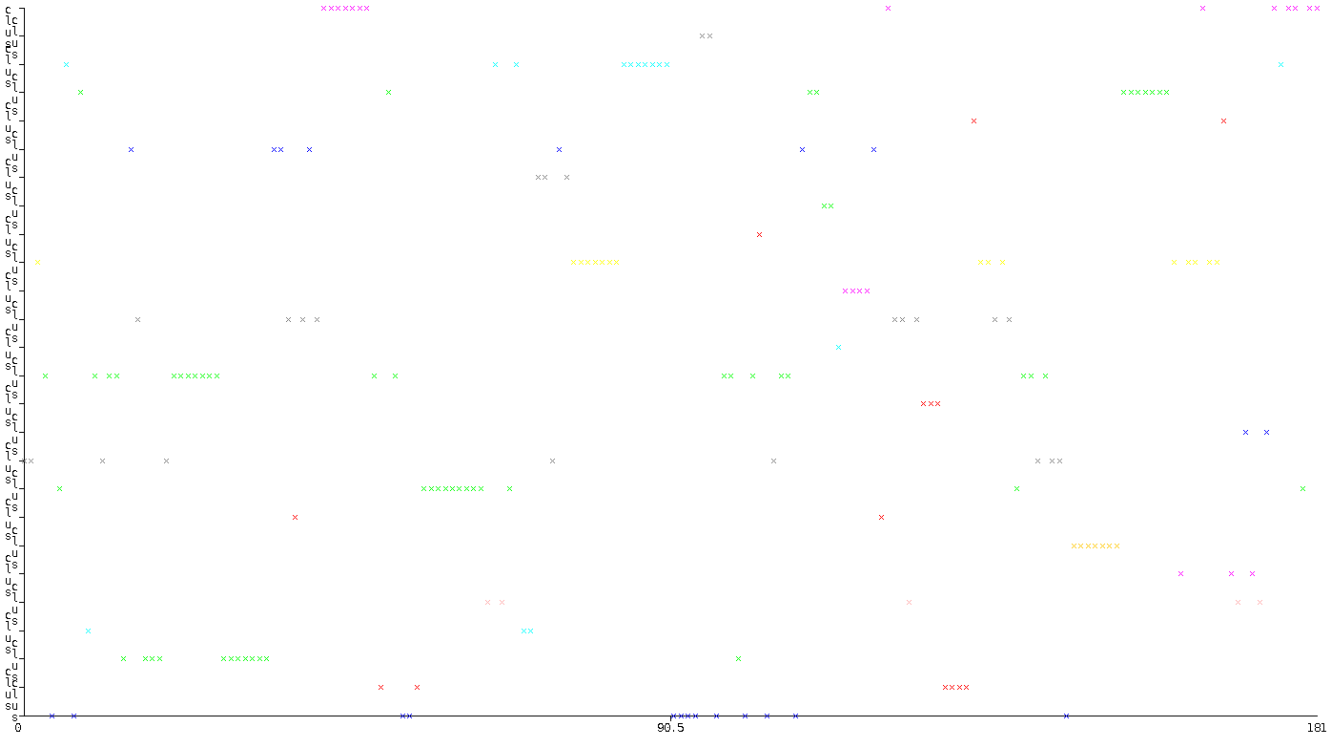
4.3. Complete Linkage:



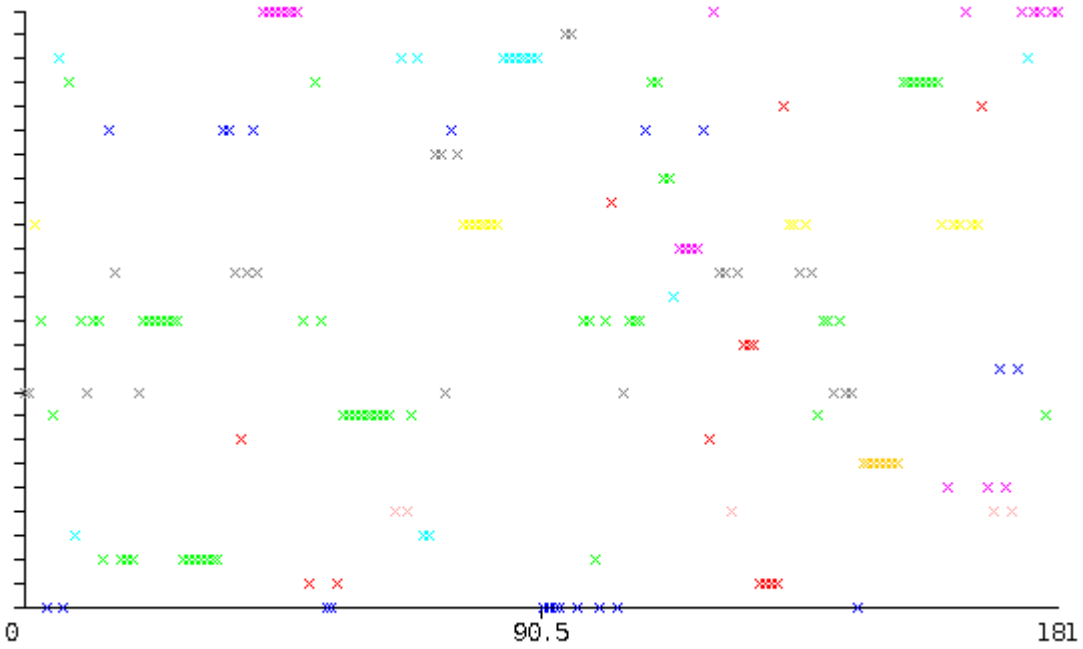
5. K-Means:



6. FilteredClusterer:



7. MakeDensityBasedClusterer



Best Clustering Technique

Before determining the best clustering technique we need to define what would be the visual metric for identifying the best visual structure. In the provided problem with 7 images for each of the alphabetical letters, the perfect clustering technique would cluster all 7 images of each image into 1 cluster, resulting in a graph of 7 consecutive instances belonging in the same cluster(Same horizontal line).

The 2 graphs that reached a visual goal very similar to the ideal were the Hierarchical clustering with average and complete linkages, because they, in contrary to all other techniques do have several instances belonging on the same horizontal line without any obvious fluctuations in the y-axis(Cluster group) such as Canopy, EM, FarthestFirst, and K-Means, nor do they cluster most of the instances in one cluster like the Hierarchical single linkage algorithm. It is also important to note that the FilteredClusterer and MakeDensityBasedClusterer do not add much since they are basically K-Means with just the possibility of adding filters.

It is quite difficult to determine the better algorithm between the hierarchical average and complete linkage algorithms just be visual inspection, so we resort to investigating the cluster elements for both algorithms:

Average:	Complete:
Clustered Instances 0 7 (4%) 1 9 (5%) 2 7 (4%) 3 7 (4%) 4 9 (5%) 5 9 (5%) 6 1 (1%) 7 2 (1%) 8 7 (4%) 9 19 (10%) 10 9 (5%) 11 1 (1%) 12 21 (12%) 13 7 (4%) 14 6 (3%) 15 8 (4%) 16 1 (1%) 17 7 (4%) 18 7 (4%) 19 2 (1%) 20 6 (3%) 21 1 (1%) 22 1 (1%) 23 14 (8%) 24 7 (4%) 25 7 (4%)	Clustered Instances 0 7 (4%) 1 7 (4%) 2 5 (3%) 3 7 (4%) 4 7 (4%) 5 9 (5%) 6 5 (3%) 7 6 (3%) 8 7 (4%) 9 5 (3%) 10 4 (2%) 11 5 (3%) 12 9 (5%) 13 7 (4%) 14 7 (4%) 15 12 (7%) 16 6 (3%) 17 8 (4%) 18 13 (7%) 19 7 (4%) 20 8 (4%) 21 2 (1%) 22 6 (3%) 23 2 (1%) 24 14 (8%) 25 7 (4%)

I think it is quite fair to say that the complete linkage algorithm performed better than the average linkage one since the complete linkage does not have several clusters with extreme anomalies such as clusters containing only 1% or 12% of the instances, where it is usually hovers around the desired 4%.