Q3 Report

**Analysis of Cluster Merging and Mean Squared Error (MSE) Change in Hierarchical Clustering**

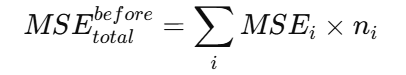
**1. Total MSE Before Merging**

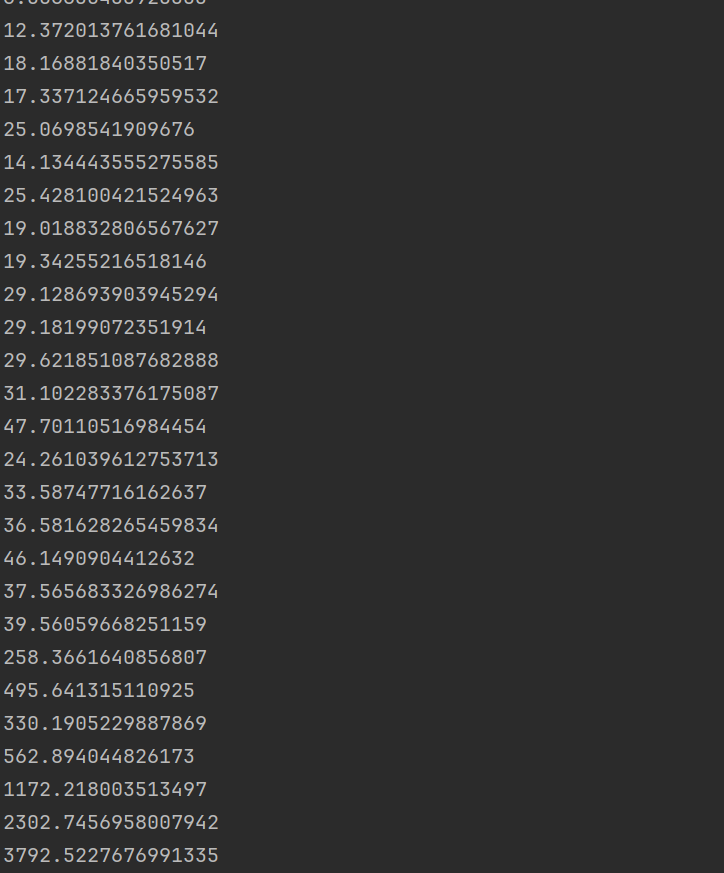
The total mean squared error (MSE) in hierarchical clustering is a measure of how spread out the data points are within each cluster. Initially, when each point is considered its own cluster, the total MSE is zero. As clustering progresses, the total MSE changes depending on how clusters are merged.

In our implementation, we track the total MSE at each step of the merging process using:

* The sum of the MSEs of all existing clusters before a merge.
* The new MSE calculated after merging two clusters.

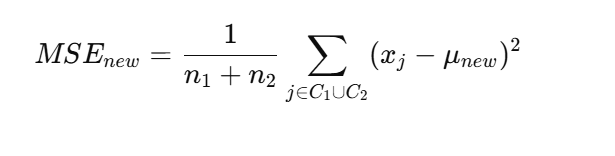
At each step, the MSE before merging is computed as:





**2. MSE of the New Cluster After Merging**

After merging two clusters C1C\_1 and C2C\_2, the new cluster's MSE is determined using:

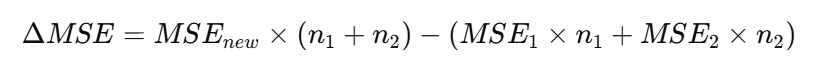


where μ(new) is the centroid of the newly merged cluster.

We calculate this value in the code by merging the points from both clusters and computing their new centroid and MSE.

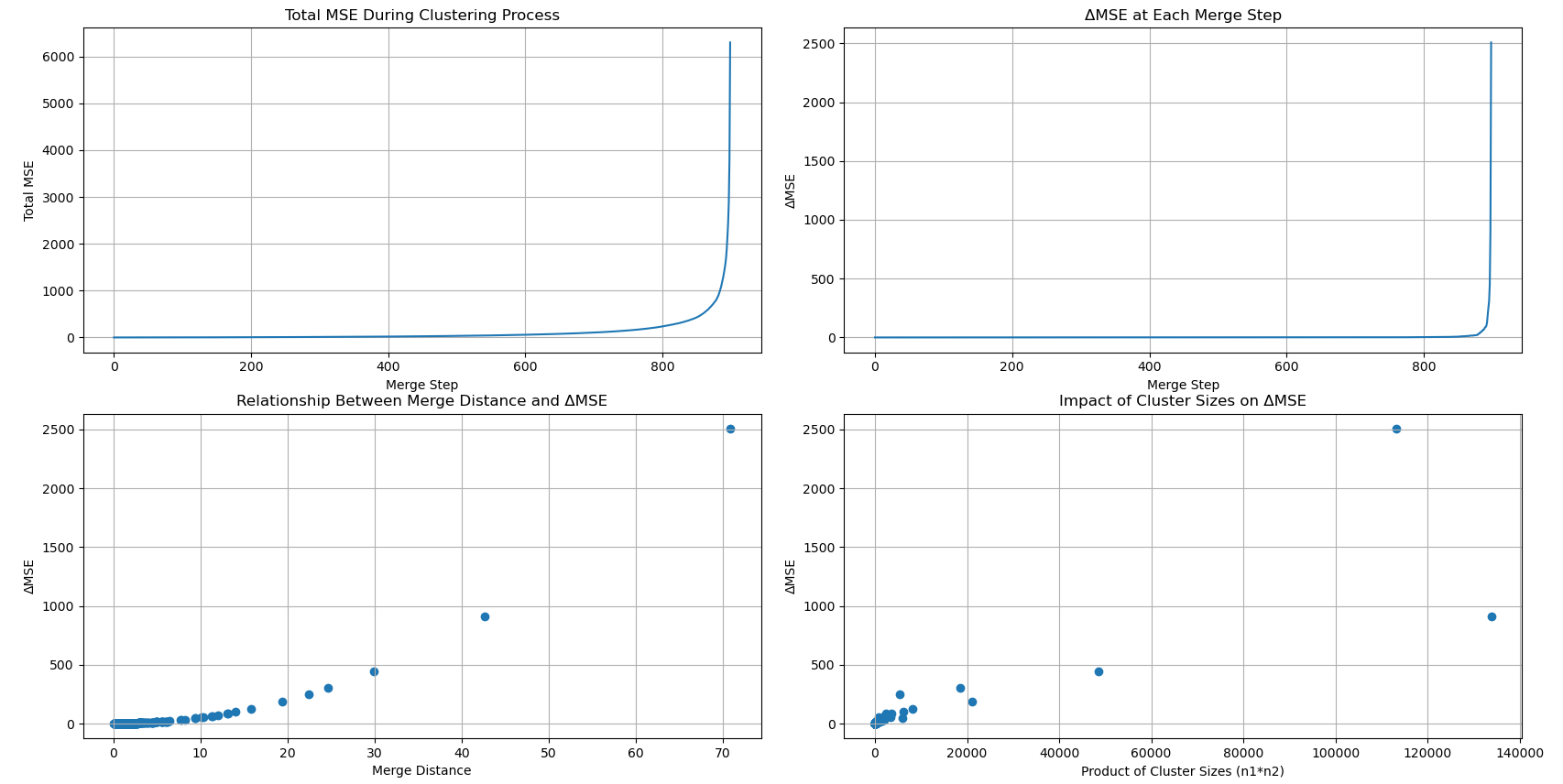
**3. Relationship Between MSE Change and Distance Between Centers**

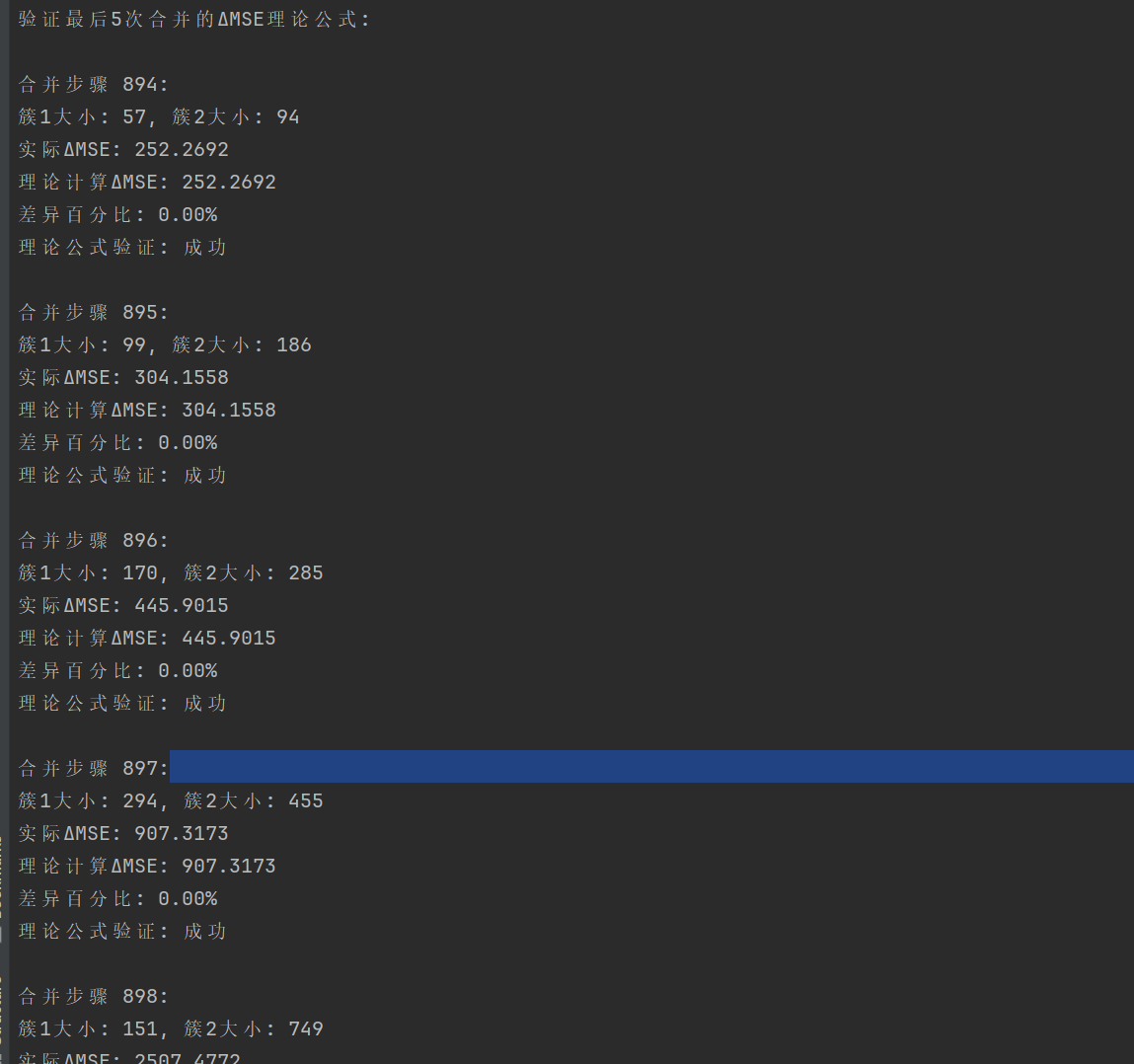
The change in total MSE after a merge is given by:



A key observation is that the change in MSE is related to the distance between the centroids of the merging clusters. As the centroid distance increases, the new cluster's variance increases, leading to a higher MSE.

In our visualization, we plot ΔMSE\Delta MSE against the merge distance (from the linkage matrix) and observe a positive correlation. This indicates that merging distant clusters leads to a larger increase in MSE.





**4. Physical Meaning Explanation**

The physical interpretation of these observations is as follows:

* **Merging clusters that are too far apart** introduces significant new variance, increasing the overall MSE.
* **Merging clusters that are close together** results in a smaller increase (or even a decrease) in MSE, as the data points are already relatively compact.
* **The impact of cluster sizes** is also significant: merging a small and large cluster tends to increase MSE more than merging two similarly sized clusters.

**Key Takeaways**

1. **Merging distant clusters is suboptimal** if the goal is to minimize MSE, as it introduces high variance.
2. **The Ward method** (used in our implementation) seeks to minimize the increase in total variance at each step, making it a preferred choice for hierarchical clustering.
3. **Tracking MSE changes provides insights** into the structure of the data and helps refine clustering strategies.

By analyzing these MSE variations, we can make informed decisions about choosing appropriate cluster merging strategies in hierarchical clustering.