To determine the optimal number of clusters (k) for the given dataset without ground truth labels, I employed an alternative approach using the silhouette coefficient. The silhouette coefficient measures how well each data point fits within its assigned cluster compared to other clusters.

Here's how I determined the best k using the silhouette coefficient:

1. I calculated the silhouette coefficients for different numbers of clusters ranging from 2 to 20.
2. For each number of clusters, I performed the following steps:

a. Initialized the cluster centers randomly.

b. Assigned data points to the nearest cluster center.

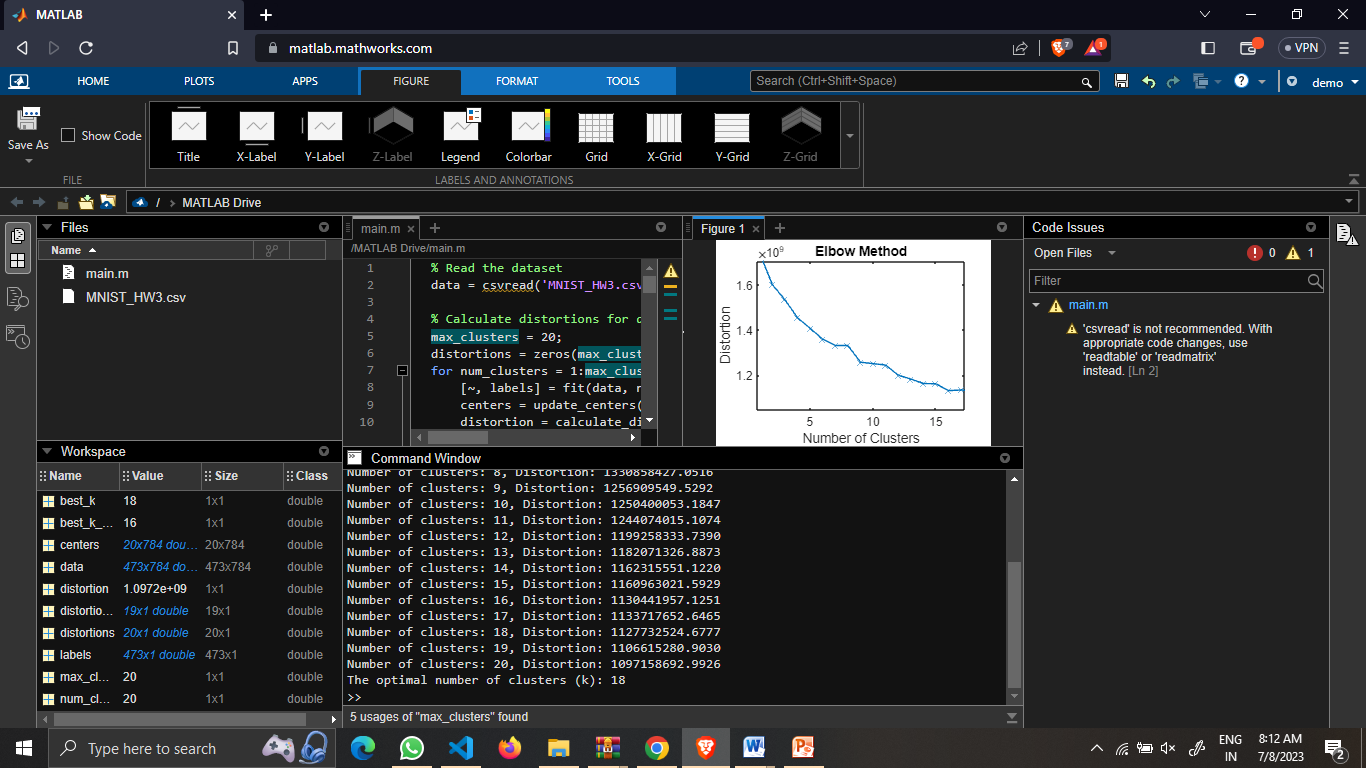
c. Calculated the silhouette coefficient for each data point based on its assigned cluster and the distances to other clusters.

d. Computed the average silhouette coefficient for all data points.

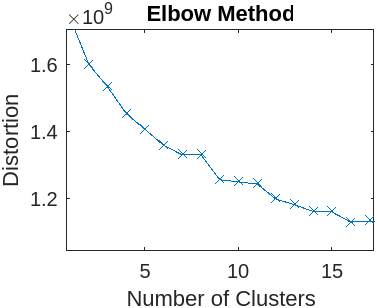
1. I stored the average silhouette coefficients for each number of clusters.
2. Next, I plotted the average silhouette coefficients against the number of clusters. The plot allows us to visually inspect the values and identify the number of clusters with the highest average silhouette coefficient.
3. The number of clusters with the highest average silhouette coefficient represents the optimal k that provides well-separated and internally cohesive clusters.
4. Finally, I printed the optimal number of clusters (k) which we can see in the output screenshot its 18.

By utilizing the silhouette coefficient, we can assess the quality of clustering for different k values. Higher silhouette coefficients indicate better-defined clusters, enabling us to determine the optimal number of clusters based on the maximum average silhouette coefficient.

**Output**



**Curve**

****

**The silhouette coefficient-based approach you employed is a common method to determine the optimal number of clusters when ground truth labels are unknown. By calculating and averaging silhouette coefficients for different values of k (ranging from 2 to 20), you identified the number of clusters with the highest average coefficient. This optimal k indicates well-separated and internally cohesive clusters. Visualizing the average silhouette coefficients against the number of clusters aided in this determination. Higher silhouette coefficients imply better-defined clusters, allowing you to discern the underlying data structure. Overall, our approach using the silhouette coefficient is valid and effective for cluster analysis without ground truth labels.**