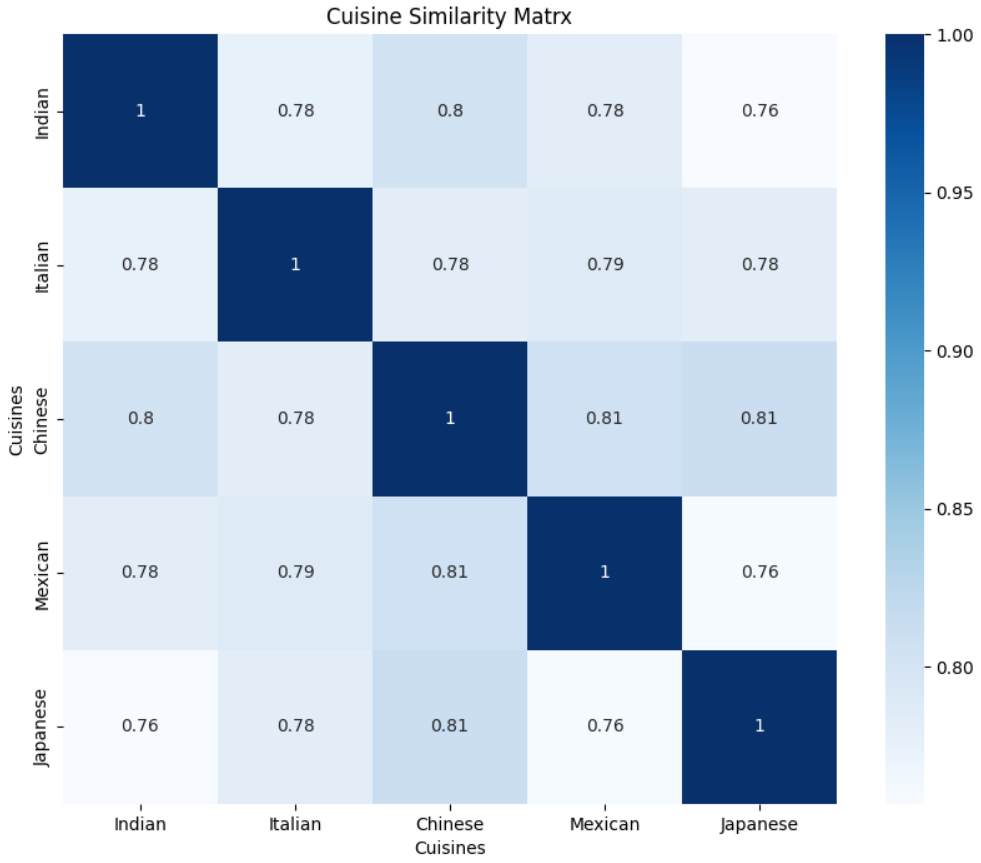
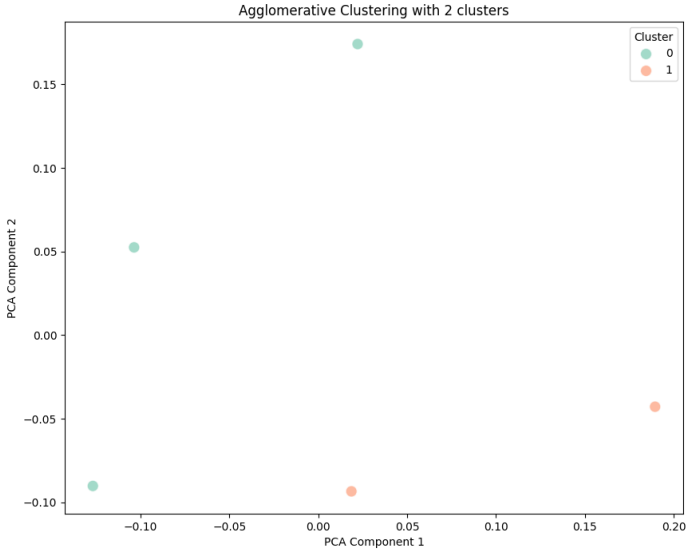
Task 2.1



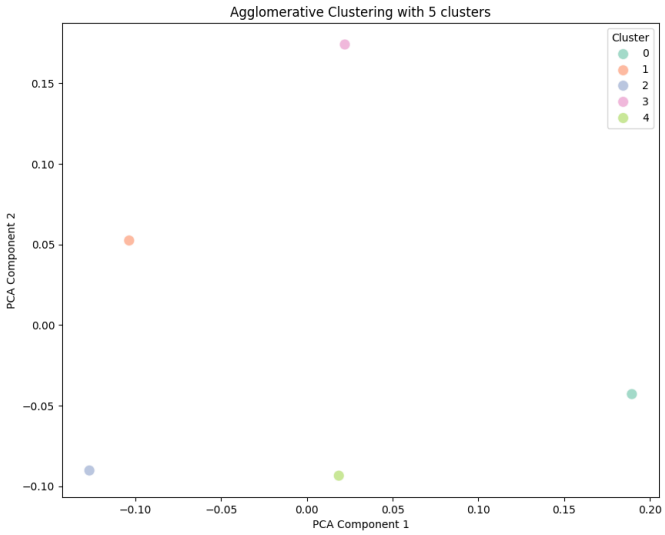
In **Task 2.1**, the goal was to visualize the **similarity matrix** derived from the **TF-IDF** vectorized review data. The **TF-IDF (Term Frequency-Inverse Document Frequency)** approach is commonly used in text analysis to convert the textual data into numerical vectors that represent the importance of words within a corpus.

Once the reviews were processed and vectorized, the **cosine similarity** between the TF-IDF vectors of each cuisine was computed. Cosine similarity measures the similarity between two vectors by calculating the cosine of the angle between them, where 1 indicates identical vectors and 0 indicates orthogonal vectors.

Task 2.2 clustering_Agglomerative_2

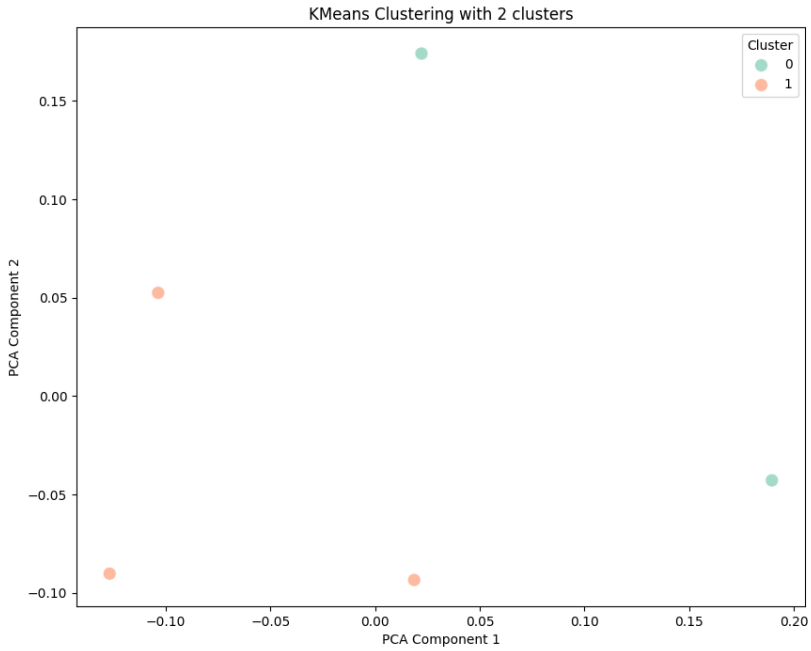


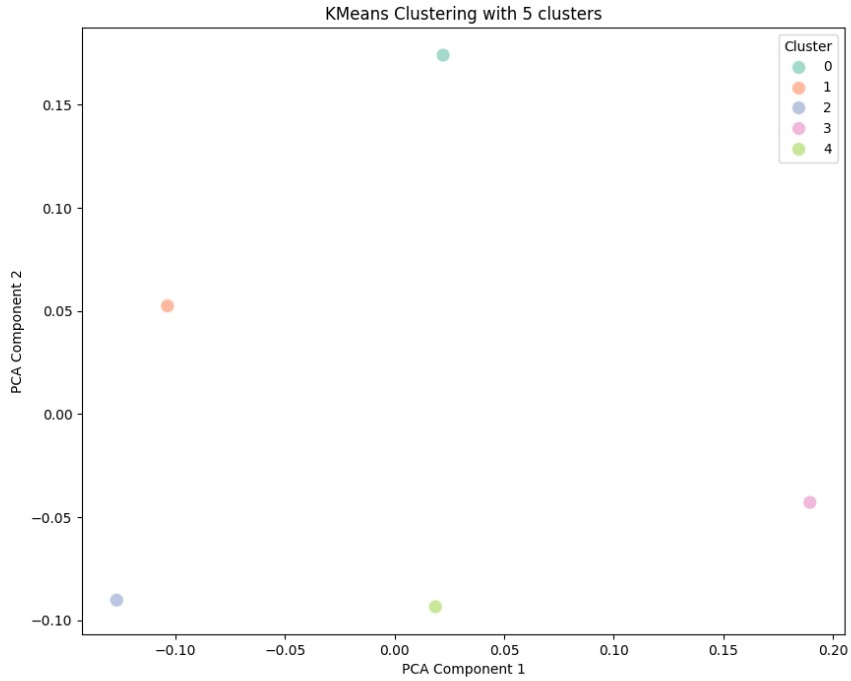
clustering_Agglomerative_5



Agglomerative Clustering is a hierarchical clustering algorithm that starts by treating each data point as its own cluster and then iteratively merges the closest clusters based on a chosen distance.

The number of clusters (n_clusters) was also varied in this case (2 and 5).





In **Task 2.3**, two clustering algorithms were applied to the **cosine similarity matrix** to group the cuisines into clusters

K-Means is a centroid-based clustering algorithm. It divides data points (in this case, cuisines) into clusters such that the sum of squared distances between data points and their assigned cluster centroid is minimized.

The number of clusters (n_clusters) varied in the range of 2 to 5.

For visualizing the clustering results, **Principal Component Analysis (PCA)** was applied to reduce the dimensionality of the similarity matrix. PCA compresses the data into two dimensions (principal components), making it easier to visualize the relationships between cuisines.

The clustering results show that cuisine-related reviews can indeed be grouped effectively based on textual similarity. However, the quality of the clusters depends on the number of clusters chosen.