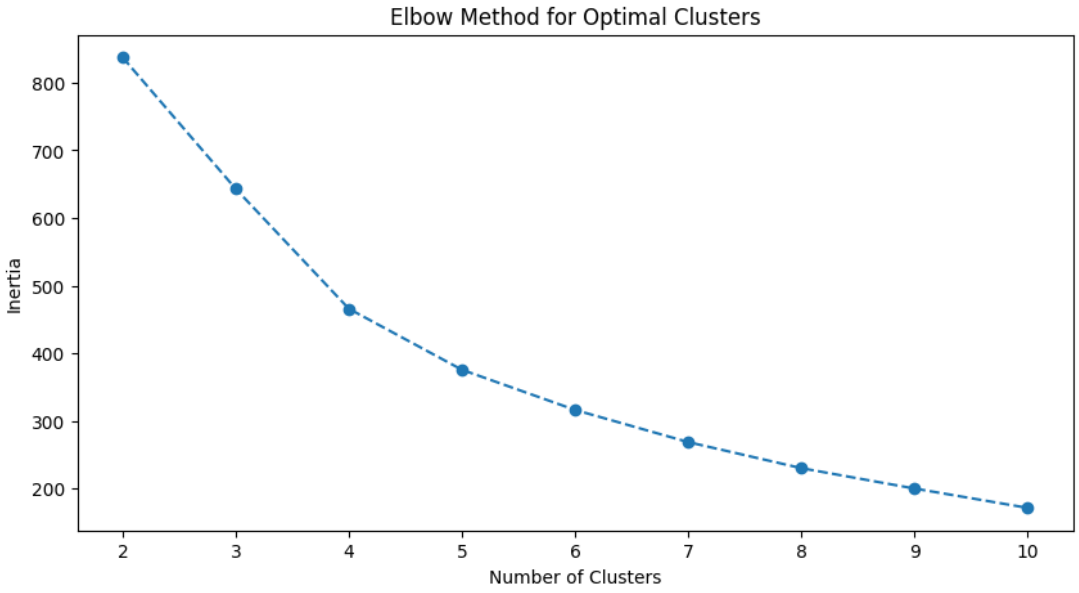
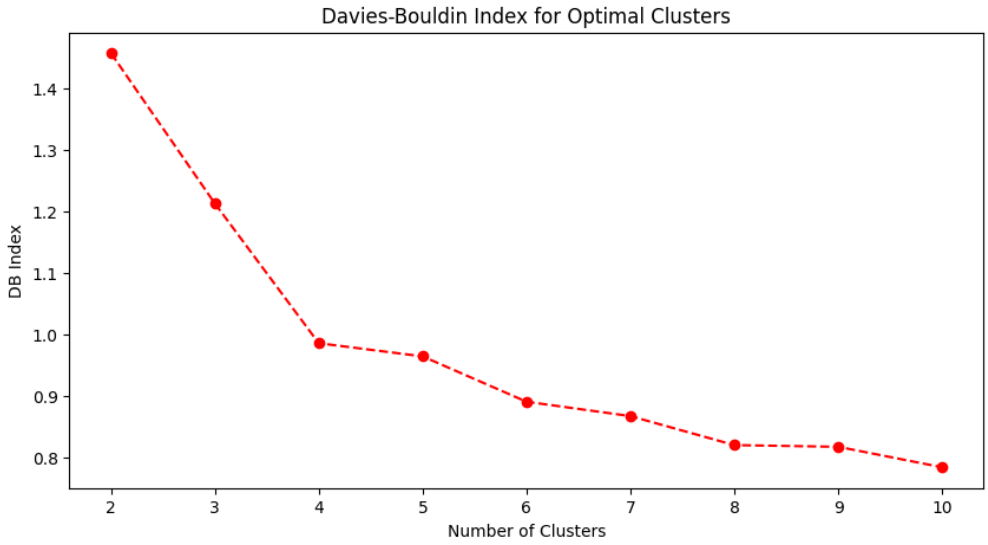
Clustering

Clustering is an unsupervised machine learning technique used to group similar data points into clusters based on similar characteristics. It is particularly relevant for customer segmentation in transactional data, as it allows businesses to identify distinct groups of customers with similar purchasing behaviors, preferences, or spending patterns. By analyzing transactional data, clustering algorithms such as K-Means can discover meaningful patterns, enabling businesses to personalize marketing strategies, customer experiences, and optimize resource allocation. This segmentation helps organizations better understand their customer base, predict future behaviors, and drive data-driven decision making to enhance customer satisfaction.

Here, we have segmented the customers using aggregate metrics such as Quantity, Total Value and Number of Transactions.

To find the number of optimal clusters, we use the Elbow method to analyze the clustering metrics for a range of clusters to find the optimal number of clusters. A few important evaluation metrics are Inertia (Within Cluster Sum of Squares Error) and Davies-Bouldin Index. The plots corresponding to these metrics are as follows.





Here, we can observe that the DB Index is lowest for 10 clusters. Thus, this is the optimal number of clusters in the range of 2 to 10 for our scenario.

By implementing K-Means clustering for 10 clusters, we get the following result.

```
Davies-Bouldin Index: 0.7848748526738408
Silhouette Score: 0.4421262539240562
```

Since the number of features is large, it is not feasible to show the plot of the dataset to distinctively visualize the clusters. Instead, we may use PCA to reduce the number of features to 2, this representing the clusters in a 2 dimensional plot. The plot obtained is as follows.