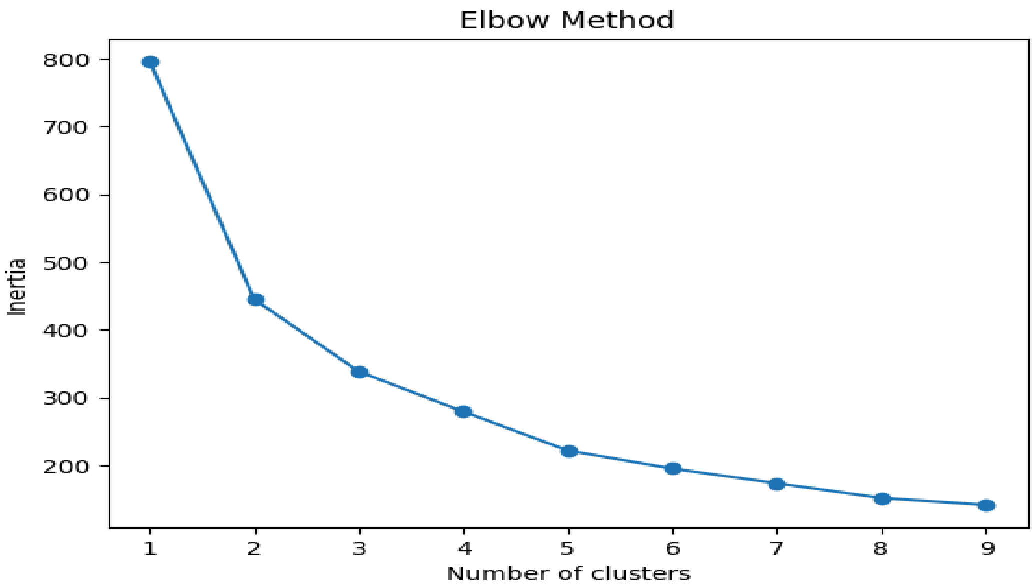
Customer Segmentation & Clustering

The task was to perform customer segmentation using clustering techniques, utilizing both customer profile data and transaction information. The primary objective was to determine the optimal number of clusters, evaluate the clustering performance using the Davies-Bouldin Index (DBI), and visualize the resulting clusters.

For the clustering analysis, the **Elbow method** was first applied to determine the optimal number of clusters for K-Means. This method involves plotting the **inertia**, which is the sum of squared distances from each point to its assigned cluster center, against the number of clusters. The optimal number of clusters was identified where the inertia began to decrease at a slower rate. Based on the elbow plot, the optimal number of clusters was found to be **3**.

Next, **K-Means clustering** was performed with both **k=3** and **k=5(**visualisation**)**. The Davies-Bouldin Index (DBI) was calculated for each clustering configuration. The DBI is a metric used to evaluate the quality of clusters, where a lower value indicates better cluster separation. For **K-Means with k=3**, the DBI value was **0.964**, indicating relatively well-separated clusters. When the number of clusters was increased to **k=5**, the DBI value rose to **2.615**, suggesting that the clusters were less distinct and the clustering performance worsened as the number of clusters increased.

To further confirm the optimal number of clusters, the **KneeLocator** method was used. The KneeLocator helps identify the "elbow" point more accurately in the inertia curve. The optimal k value identified by the KneeLocator was also **3**, aligning with the results from the Elbow method.



Additionally, **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** was applied as an alternative clustering technique. DBSCAN is a density-based algorithm that can identify clusters of arbitrary shape and handle noise (outliers). The algorithm was configured with parameters `eps=1.5` and `min_samples=5`. The resulting DBI for DBSCAN was **0.541**, which was lower than that of K-Means, indicating that DBSCAN provided better separation of clusters.

The clustering visualizations confirmed the findings from the metrics. Scatter plots were generated for both K-Means and DBSCAN results, showcasing the clustering behavior. For K-Means with **k=3**, the clusters were well-separated, while increasing the number of clusters to **k=5** resulted in less distinct groupings. DBSCAN also identified well-separated clusters, with some points marked as noise.

The final clustering evaluation showed that **K-Means with k=3** provided the best clustering results, based on the Davies-Bouldin Index and visual analysis. **DBSCAN** performed better than K-Means in terms of DBI, although it identified fewer clusters and labeled some points as noise.