Task 3: Customer Segmentation

Objective:

To segment customers into distinct groups based on their transaction history and profile information using clustering techniques.

Steps Performed:

1. Feature Engineering

- Combined data from `Customers.csv`, `Products.csv`, and `Transactions.csv` to form a unified dataset.
- Derived a new `Price` column from `TotalValue` and `Quantity`.
- Aggregated transaction-level data into customer-level features:
    - **TotalValue**: Total spending by the customer.
    - **Quantity**: Total quantity of products purchased.
    - **Price**: Average price of items purchased.
    - **Category**: Most frequent product category.
    - **Region**: Customer's region.

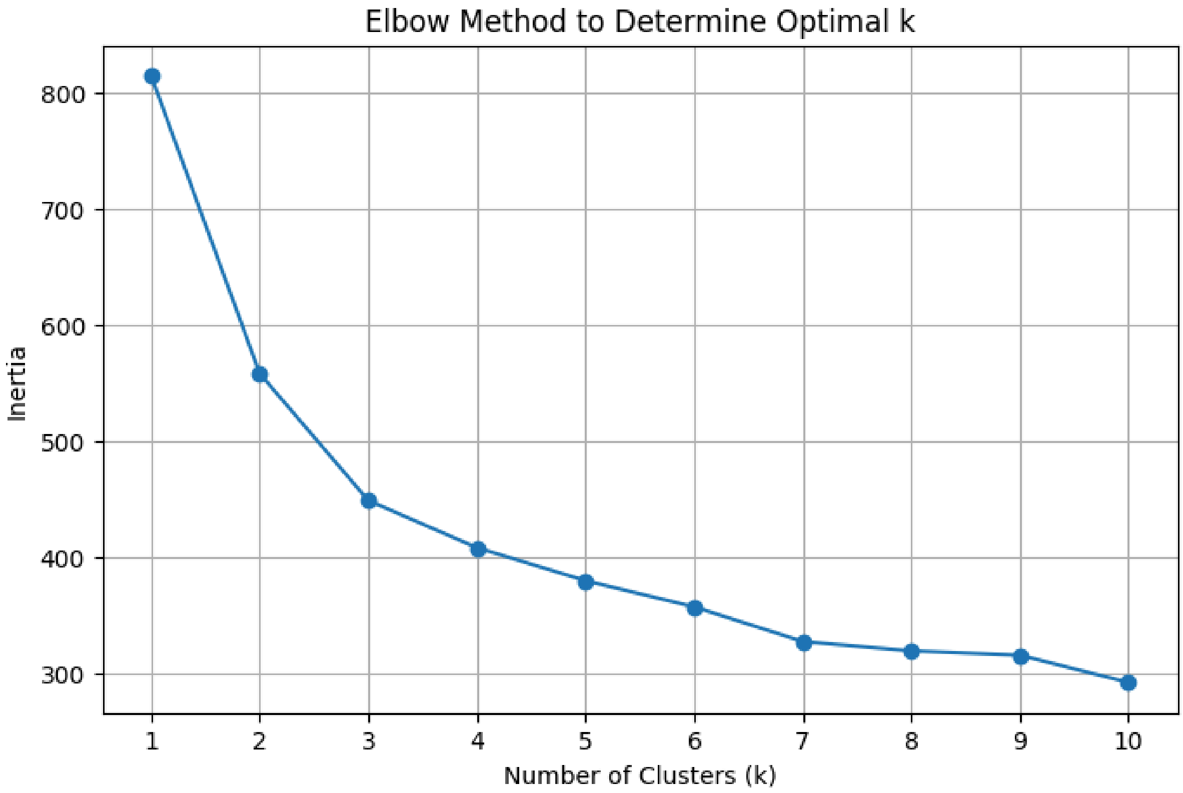
2. Data Preprocessing

- One-hot encoded categorical features (`Category` and `Region`).
- Standardized numerical features (`TotalValue`, `Quantity`, `Price`) using `StandardScaler`.

---

3. Elbow Method

- The **elbow method** was used to determine the optimal number of clusters (k).



- **Graph Description**:
  - The graph plots the number of clusters (kkk) against inertia (within-cluster sum of squares).
  - The optimal kkk is identified at the "elbow point," where the decrease in inertia slows significantly.
  - Based on the graph, k=6k = 6 was chosen as the optimal number of clusters.

---

4. Clustering

- Applied the **KMeans clustering algorithm** with k=6.
- Assigned each customer to one of the six clusters..

---

5. Visualization

- Used **Principal Component Analysis (PCA)** to reduce the dimensionality of the feature set for visualization purposes.
- Plotted the clusters in a 2D space using the first two principal components (`PCA1` and `PCA2`).

Results and Insights:

1. **Number of Clusters Formed**:
   - Six clusters were identified, each representing a distinct customer group.
2. **Cluster Characteristics**:
   - Each cluster has unique spending behavior, quantity preferences, and product choices.
   - Regional preferences and dominant product categories were also evident in certain clusters.
3. **Elbow Method Results**:
   - Optimal k (number of clusters): 6.
   - The elbow graph illustrates this point clearly.
4. **Visualization**:
   - The cluster visualization in PCA space shows distinct groupings of customers, confirming the segmentation.
5. **Clustering Metrics**:
   - **DB Index**: The Davies-Bouldin Index was calculated to evaluate the clustering performance.
   - A lower DB Index indicates well-separated and compact clusters.

Conclusion:

The clustering analysis successfully segmented customers into six distinct groups. These insights can be used for targeted marketing, product recommendations, and personalized customer engagement strategies.