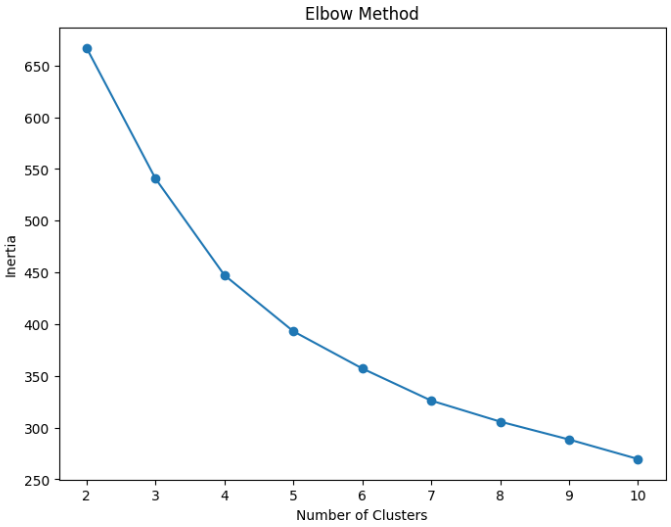
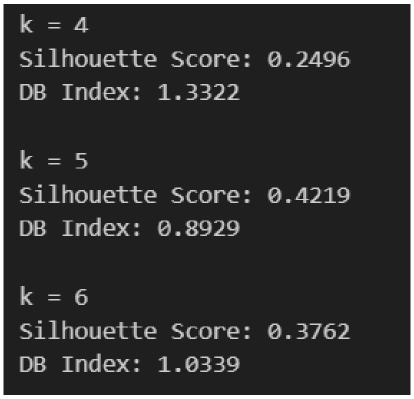
Report on Clustering Results

1. **Goal:** Customer segmentation based on profile information (from Customers.csv) and transaction information (from Transactions.csv)
2. **Data Preprocessing:** Merged the Customers.csv and Transactions.csv datasets to create a customer profile (customer_profile.csv). Applied necessary preprocessing steps (encoding categorical data, handling missing values, scaling features).
3. **Feature Set:** Listed features used for clustering, such as TotalSpent, AvgOrderValue, TotalOrders, UniqueProducts, TotalQuantity, and region variables.
4. **Clustering Algorithm:** I have chosen KMeans clustering due to its efficiency and interpretability.
5. **Number of Clusters (k):** Number of clusters (k) was selected using the Elbow method, which identified a reasonable range for k between 2 and 10.
   I have tested for 3 different k values: k = [4, 5, 6]
   Below is the plot using the Elbow method:



6. **Evaluation Metrics:** I have measured the DB index and also the Silhouette score for the list of k values.



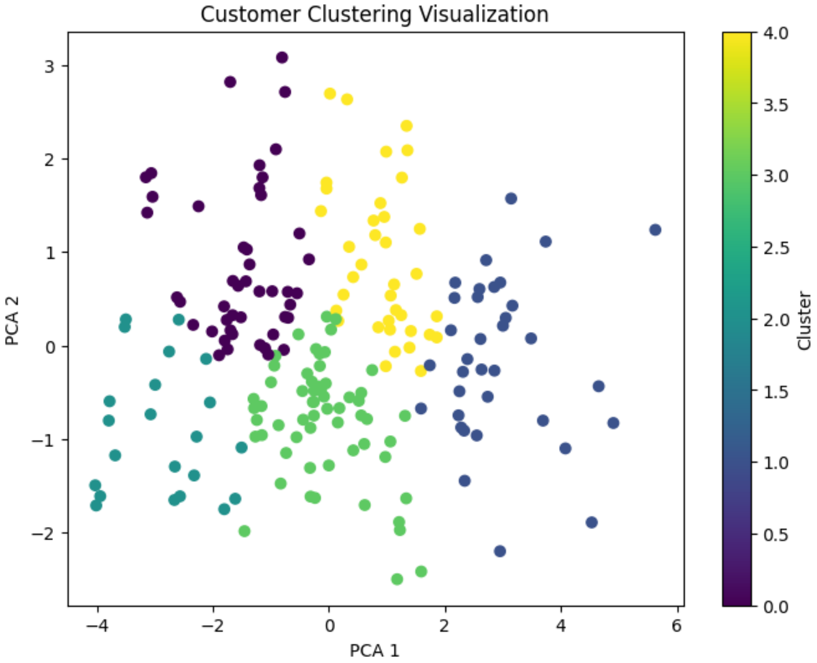
```
k = 4
Silhouette Score: 0.2496
DB Index: 1.3322

k = 5
Silhouette Score: 0.4219
DB Index: 0.8929

k = 6
Silhouette Score: 0.3762
DB Index: 1.0339
```

From the above result, k = 5 has the lowest DB index and highest Silhouette score. So I choose 5 as my k value.

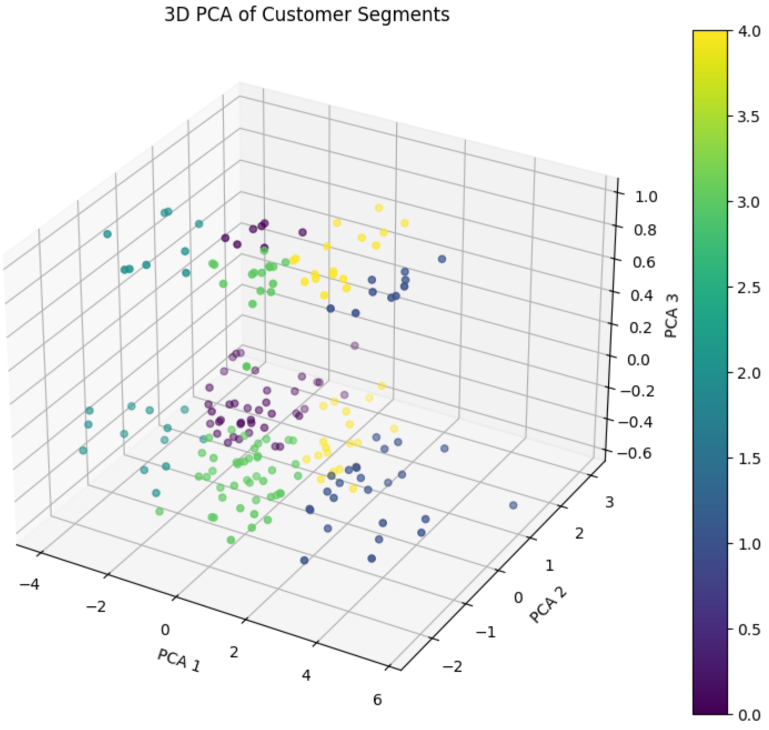
7. **Visualization:** I have created a 2D plot of Clusters after performing PCA to reduce the dimensions.

   Below is my clustering visualizations:

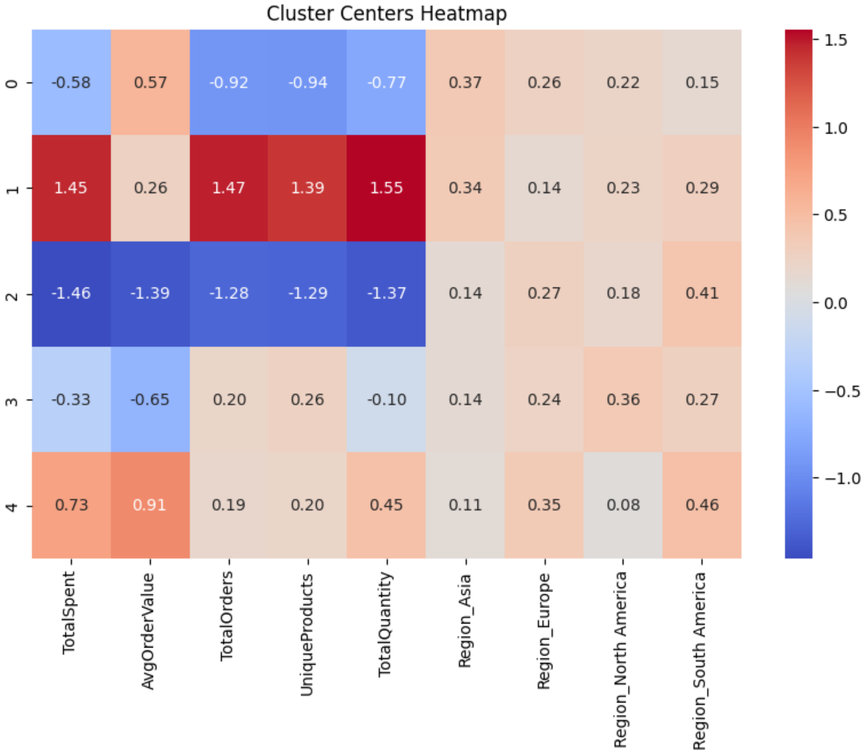
   A) 2D plot:



   B) 3D plot:



C) Cluster centers heatmap:


Cluster Centers Heatmap

	TotalSpent	AvgOrderValue	TotalOrders	UniqueProducts	TotalQuantity	Region_Asia	Region_Europe	Region_North America	Region_South America
0	-0.58	0.57	-0.92	-0.94	-0.77	0.37	0.26	0.22	0.15
1	1.45	0.26	1.47	1.39	1.55	0.34	0.14	0.23	0.29
2	-1.46	-1.39	-1.28	-1.29	-1.37	0.14	0.27	0.18	0.41
3	-0.33	-0.65	0.20	0.26	-0.10	0.14	0.24	0.36	0.27
4	0.73	0.91	0.19	0.20	0.45	0.11	0.35	0.08	0.46