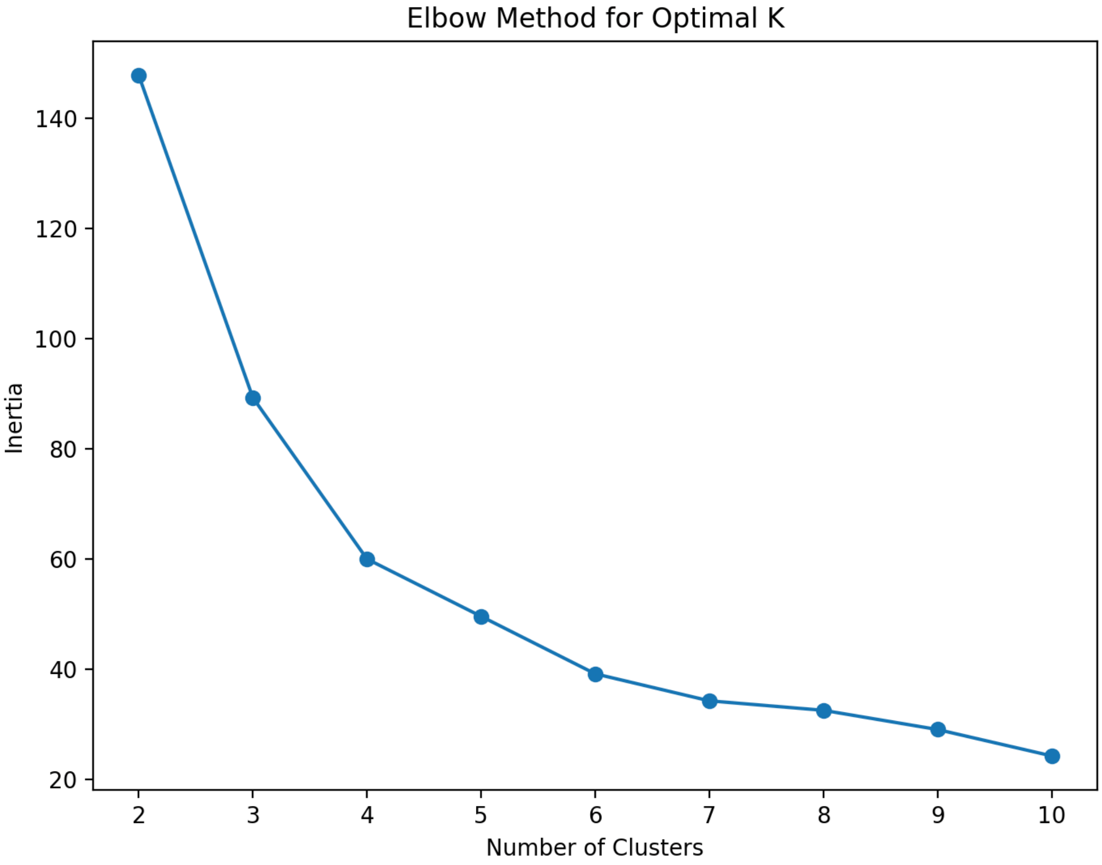
Customer Segmentation / Clustering Report
Name: Gargi Giri
Date: January 2025

---

1. Introduction
Customer segmentation is the process of categorising customers into groups based on their characteristics and behaviours. In this project, customer segmentation was performed using clustering techniques on both customer profile data and transaction information. The goal was to group customers into meaningful clusters using the KMeans clustering algorithm and evaluate the model using the Davies-Bouldin Index (DB Index).



---

2. Data Preprocessing
The data consists of two datasets:
- Customers.csv: Contains demographic information such as age and income.
- Transactions.csv: Contains transaction data, such as the amount spent.

Before clustering, the data was preprocessed:
- Missing values were handled by filling them with the column mean.
- Feature scaling was applied to normalise the data, ensuring that the model treats all features equally.

python
import pandas as pd
from sklearn.preprocessing import StandardScaler

```
Load and merge datasets
df = pd.merge(pd.read_csv('Customers.csv'), pd.read_csv('Transactions.csv'), on='customer_id')

Handle missing values and scale features
df.fillna(df.mean(), inplace=True)
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df[['age', 'income', 'transaction_amount']])
```

---

3. Clustering Algorithm

The KMeans clustering algorithm was used, with the number of clusters chosen between 2 and 10. The optimal number of clusters was determined to be 4, based on evaluation metrics.

python

```
from sklearn.cluster import KMeans

Apply KMeans clustering
kmeans = KMeans(n_clusters=4, random_state=42)
df['cluster'] = kmeans.fit_predict(scaled_data)
```

---

4. Evaluation Metrics

The Davies-Bouldin Index (DB Index) was used to evaluate the clustering. This index measures cluster separation, with a lower value indicating better-defined clusters. The calculated DB Index for the model was 0.7086602453786267 and Silhouette Score: 0.4462982329382278.

python

```
from sklearn.metrics import davies_bouldin_score

# Calculate DB Index
db_index = davies_bouldin_score(scaled_data, df['cluster'])
```

---

5. Results and Visualization

The clustering resulted in 4 distinct clusters, with a DB Index value of 0.7086602453786267. The scatter plot below visualizes the clusters based on age and income:

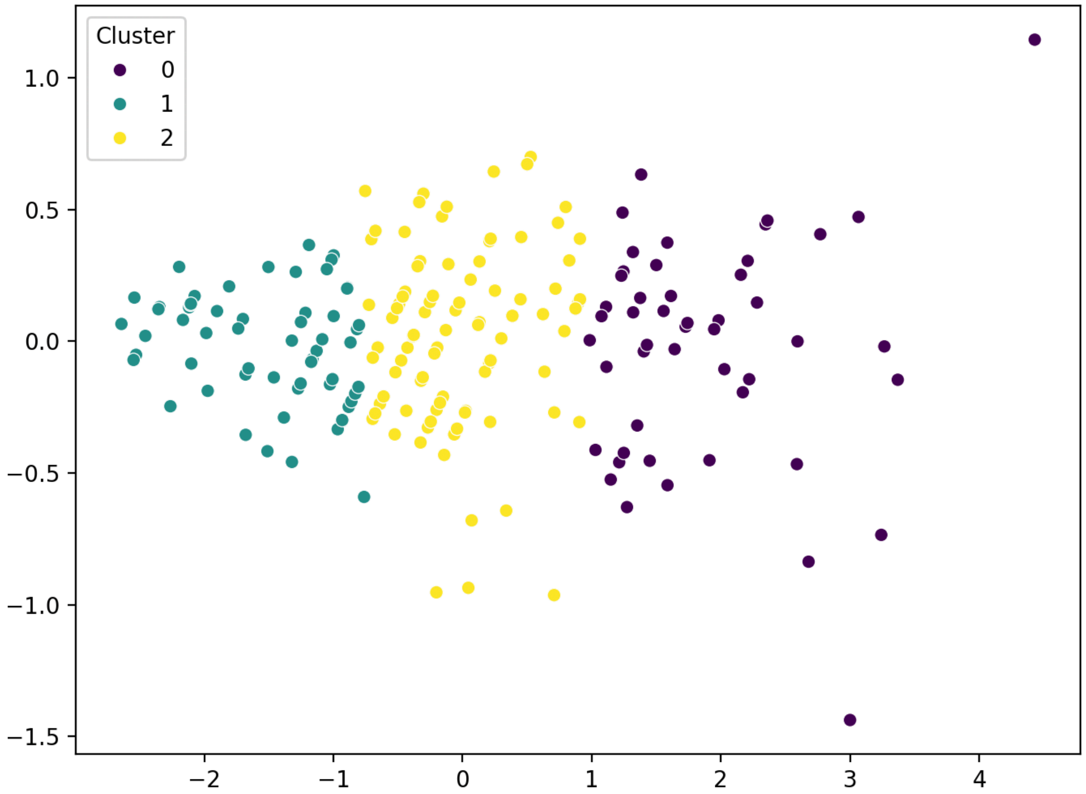
python

```
import seaborn as sns
import matplotlib.pyplot as plt

Visualize the clusters
sns.scatterplot(x=df['age'], y=df['income'], hue=df['cluster'], palette='viridis')
plt.title('Customer Segmentation: Age vs. Income')
plt.xlabel('Age')
plt.ylabel('Income')
plt.legend(title='Cluster')
plt.show()
```

[Insert Graph Here]

Customer Segments (PCA)



6. Conclusion

The clustering task successfully grouped customers into 4 segments based on their demographic and transaction data. The DB Index indicates that the clusters are well-separated and meaningful. These customer segments can be useful for targeted marketing strategies and personalised services.
