Customer Segmentation Using K-means and Hierarchical Clustering

Unsupervised Learning for Targeted Marketing Strategies

Mehrdad Naderi

2024 - Fall 2

University of Colorado Boulder

What is Customer Segmentation?

- Grouping customers based on shared attributes.

Why Customer Segmentation Matters:

- Personalized marketing.
- Improved resource allocation.
- Enhanced customer satisfaction.

Objective:

- Identify meaningful customer groups using clustering techniques.

Dataset Description

Dataset Source:

Kaggle: Mall Customers Dataset.

Features:

Age, Gender, Annual Income, Spending Score.

Data Characteristics:

200 rows, no missing values.

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Genre                   200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB

Summary Statistics:

Missing Values:
CustomerID                0
Genre                     0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

Problem:

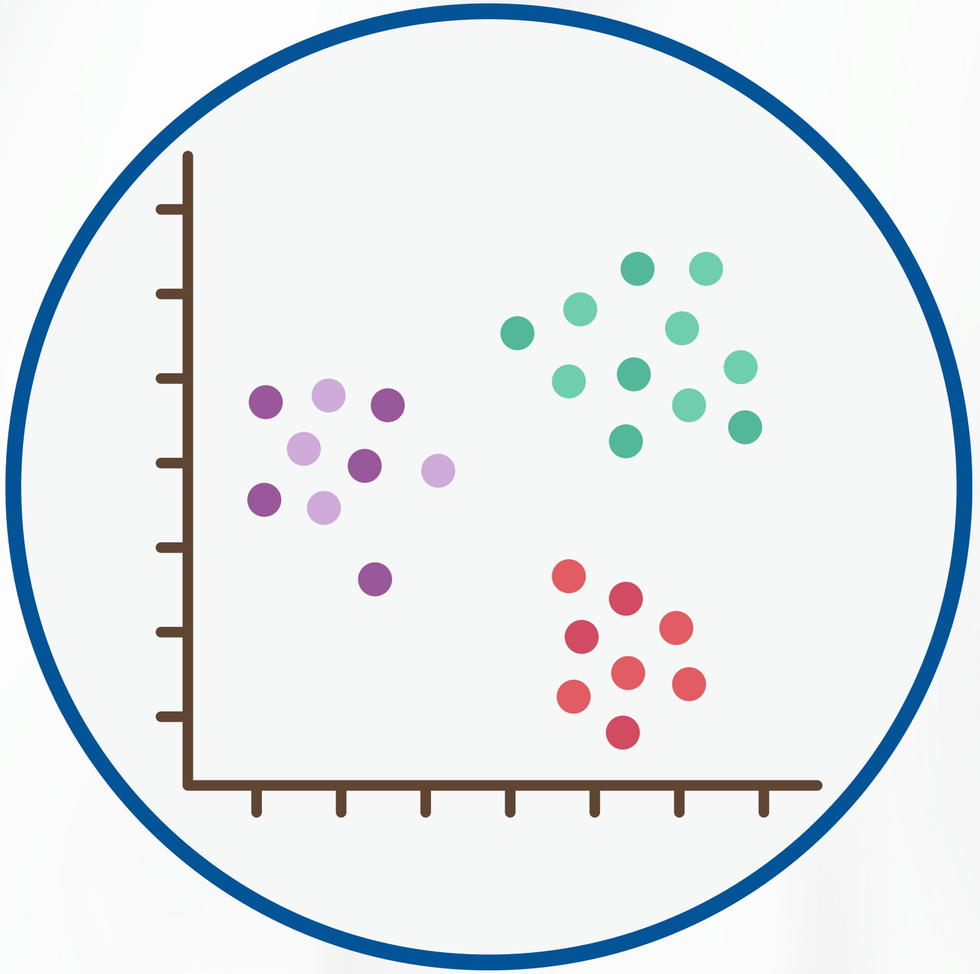
- "Group customers into clusters for targeted marketing strategies."

Clustering Techniques:

- K-means and Hierarchical Clustering.

Goal:

- Reveal patterns in customer spending and demographics.



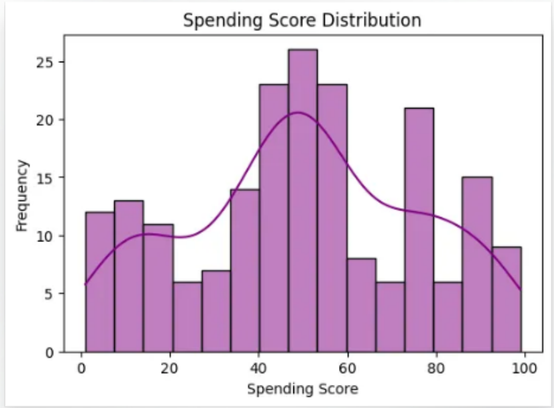
Insights from EDA:

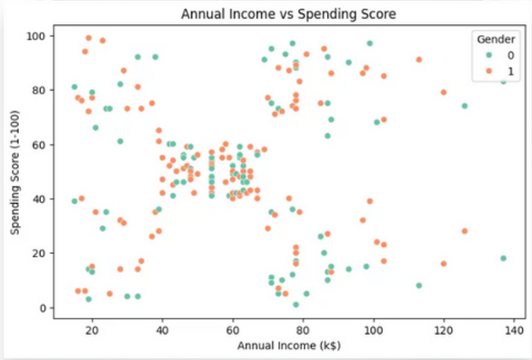
Feature distributions:

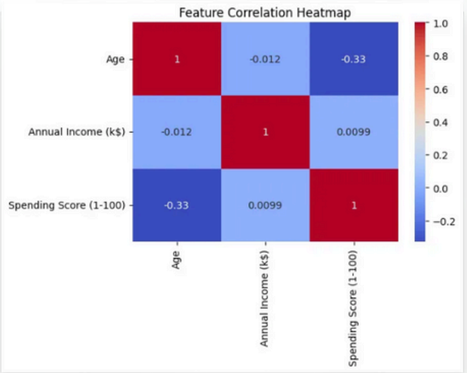
- Spending Score: "Most customers have moderate spending scores."
- Income: "Annual Income ranges widely."

Relationships:

- Income vs. Spending Score shows clear patterns for clustering.





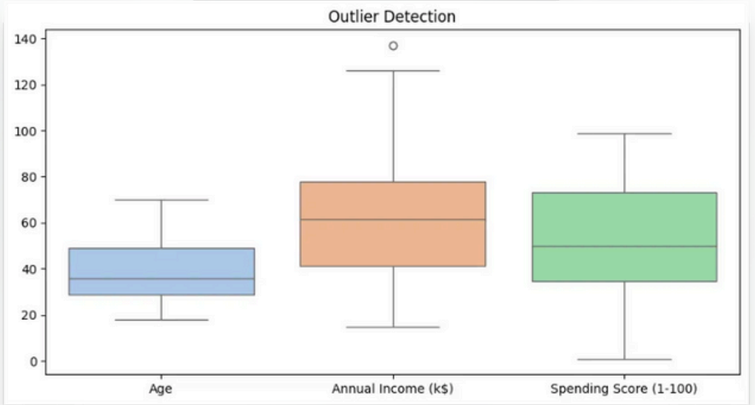


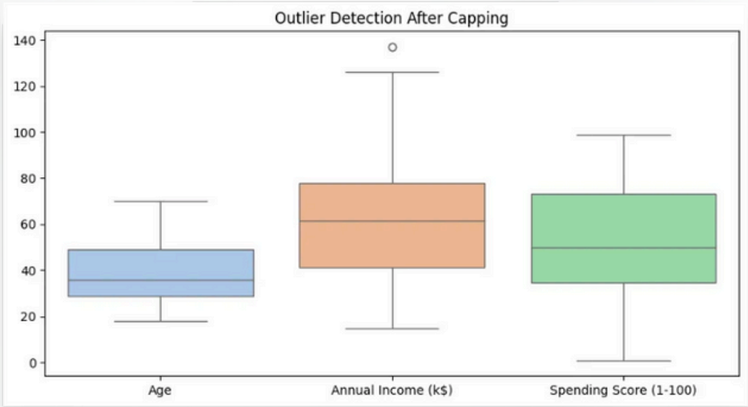
Outlier Handling:

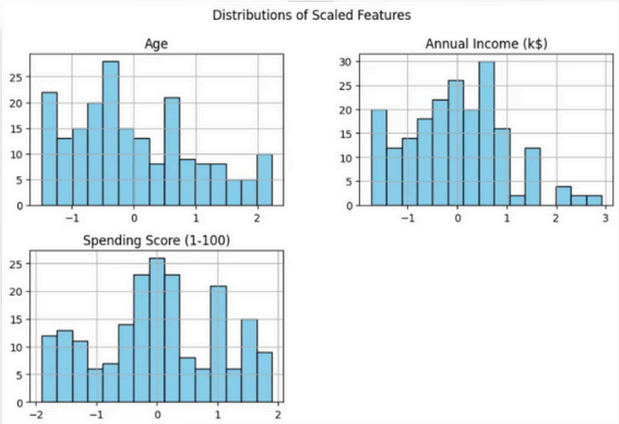
Detected and capped outliers using IQR.

Feature Scaling:

Standardized numeric features to ensure equal weighting.

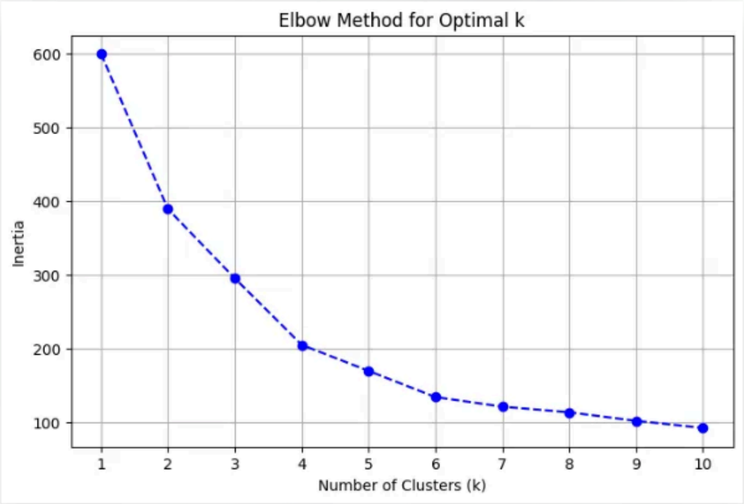






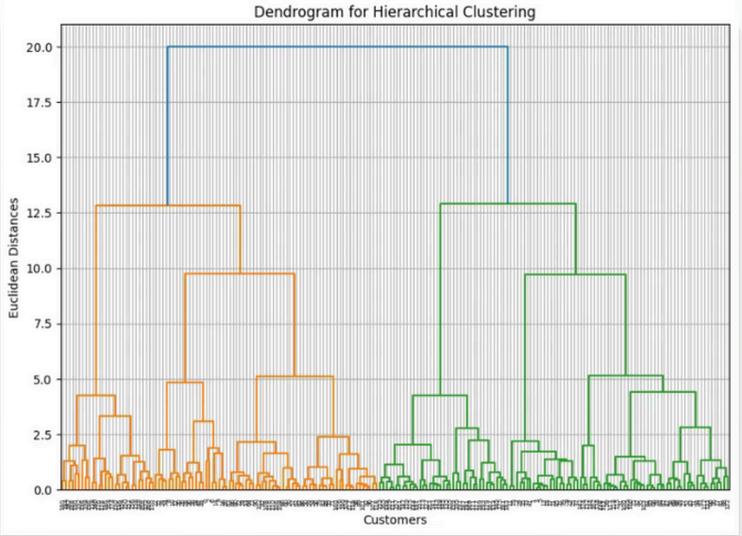
K-means Clustering:

- "Finds clusters by minimizing within-cluster variance."
- Optimization:
  - Elbow Method to determine optimal k.
  - Silhouette Score to validate clusters.



Hierarchical Clustering:

- "Builds a hierarchy of clusters based on linkage.
- " Ward's method used for merging clusters.



Optimal Clusters:

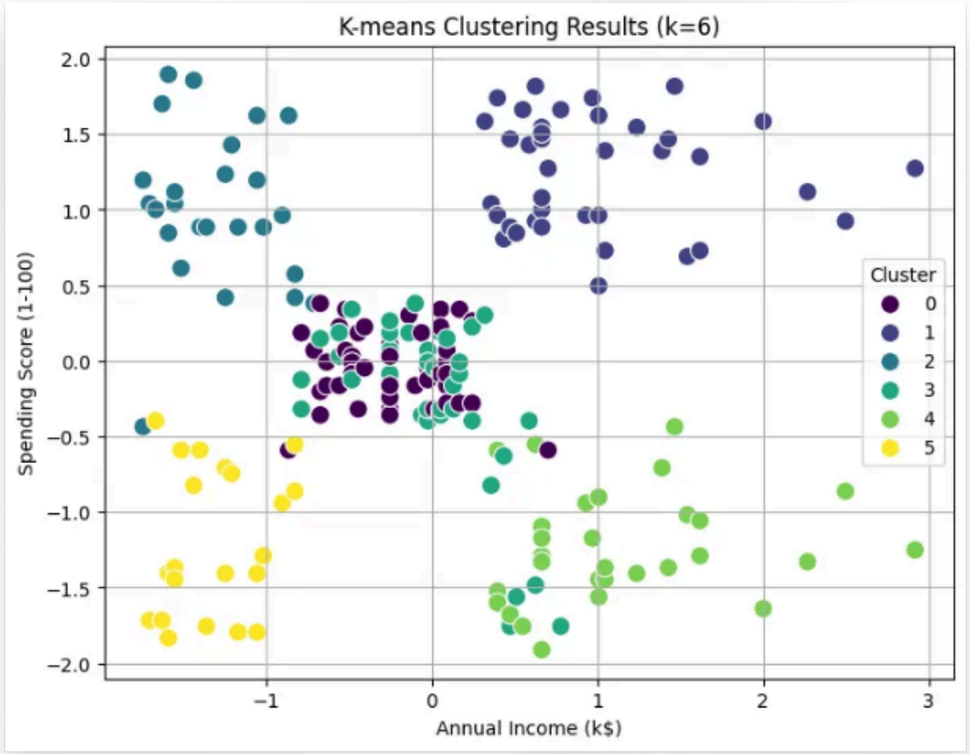
k=6.

Cluster Characteristics:

- Cluster 1: High-income, high spenders
- Cluster 2: Young, low-income, high spenders
- Cluster 3: Middle-aged, low spenders, etc

Evaluation Metrics:

- Silhouette Score: 0.431.

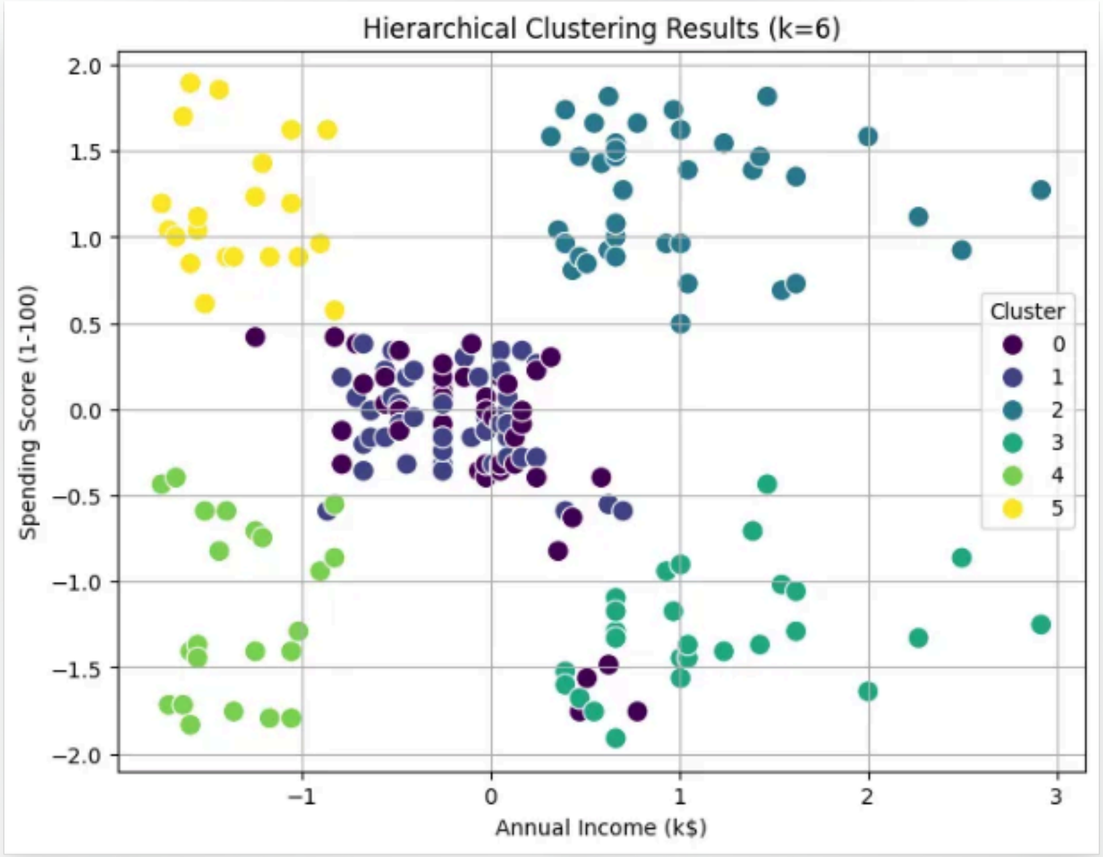


Cluster Characteristics:

Similar patterns to K-means clusters.

Evaluation Metrics:

Silhouette Score: 0.420.

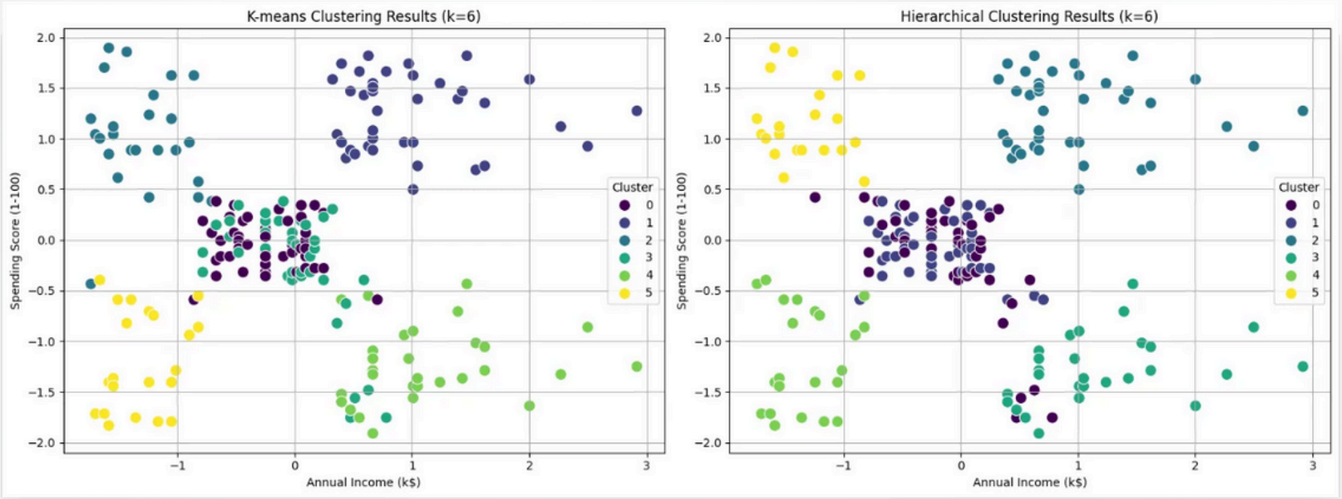


Metrics:

Metric	K-means (k=6)	Hierarchical (k=6)
Silhouette Score	0.431	0.420
Davies-Bouldin Index	0.835	0.852

Preferred Model:

"K-means is preferred due to better performance and scalability."



Cluster Summaries:

- Cluster 1: "Target high spenders with premium offers.”
- Cluster 5: "Provide cost-effective options to low-income groups.”

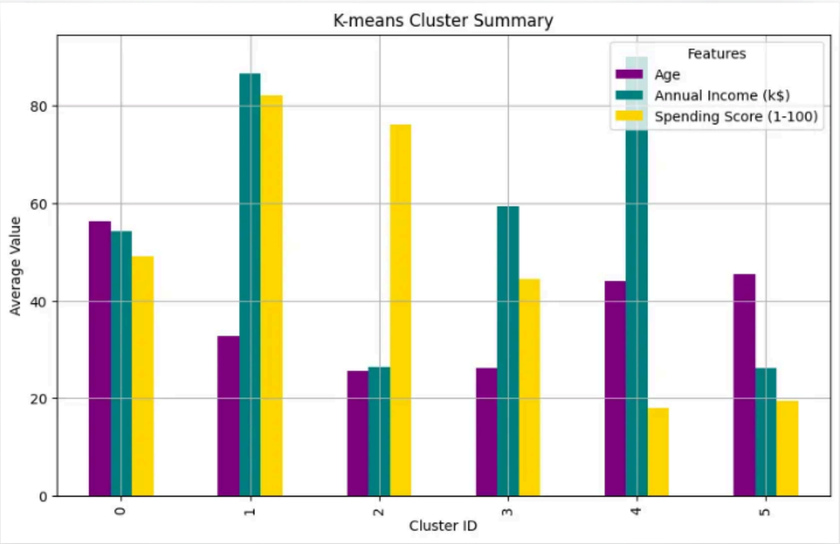
Actionable Insights:

- "High-value clusters should receive personalized promotions."

K-means Cluster Summary Table:

	Age		Annual Income (k$)		\
	mean	std	mean	std	
K-means Cluster					
0	56.333333	8.453079	54.266667	8.975725	
1	32.692308	3.728650	86.538462	16.312485	
2	25.560000	5.439669	26.480000	8.525061	
3	26.125000	7.031750	59.425000	10.587577	
4	44.000000	8.081482	90.133333	16.919145	
5	45.523810	11.766984	26.285714	7.437357	

	Spending Score (1-100)	
	mean	std
K-means Cluster		
0	49.066667	6.300794
1	82.128205	9.364489
2	76.240000	13.562448
3	44.450000	14.279176
4	17.933333	9.888807
5	19.380952	12.555780



Key Takeaways:

- Clustering revealed actionable customer segments.
- K-means performed slightly better than Hierarchical Clustering

Limitations:

- Small dataset size.
- Subjective Spending Score introduces bias.

Future Work:

- Expand dataset with additional features.
- Test other clustering methods (e.g., DBSCAN)