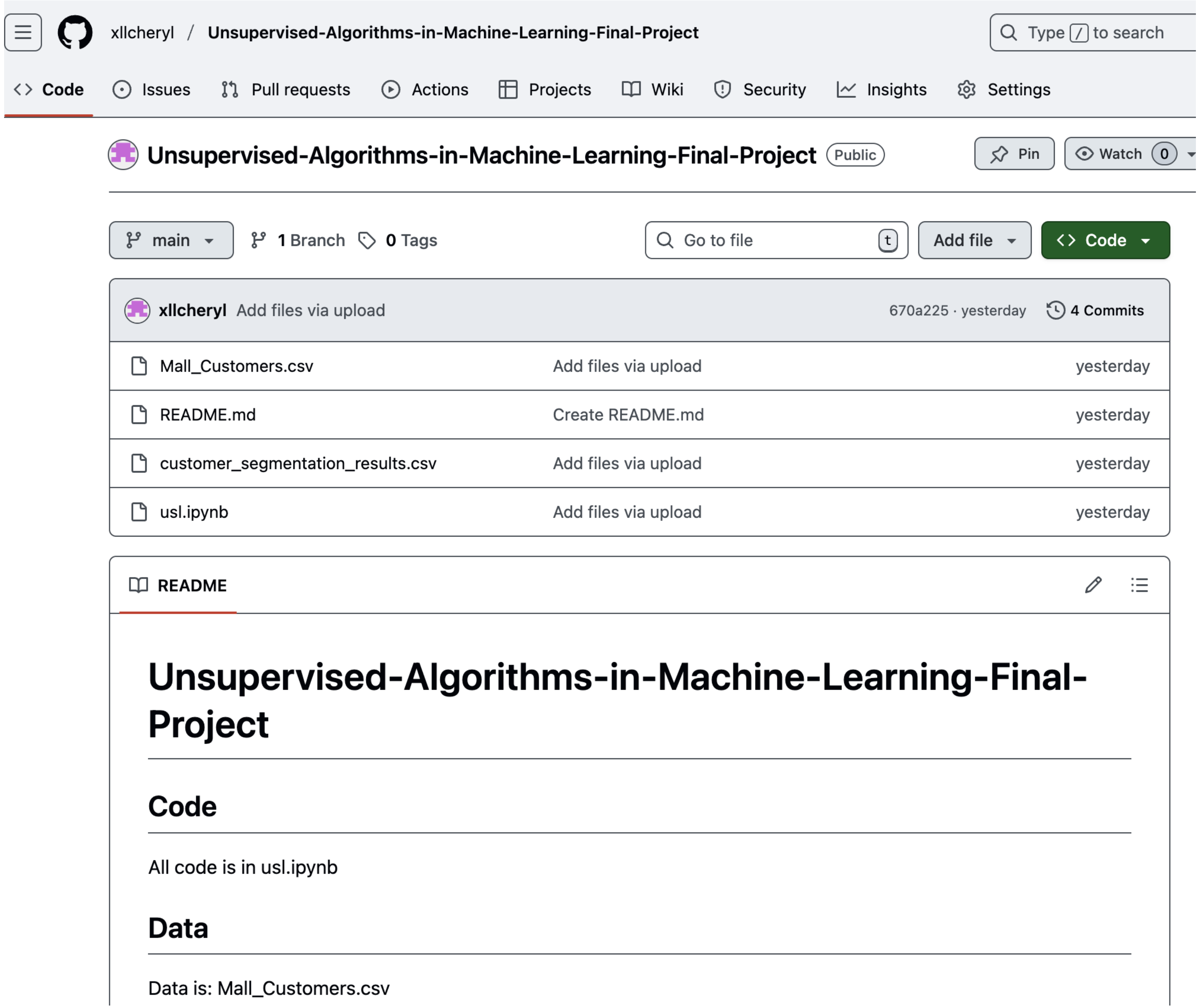
CSCA 5632 Unsupervised Algorithms in Machine Learning Final Project

LINLI XIANG

GitHub Link

https://github.com/xllcheryl/Unsupervised-Algorithms-in-Machine-Learning-Final-Project.git



xllcheryl Add files via upload 670a225 · yesterday 4 Commits

Mall_Customers.csv	Add files via upload	yesterday
README.md	Create README.md	yesterday
customer_segmentation_results.csv	Add files via upload	yesterday
usl.ipynb	Add files via upload	yesterday

README

Unsupervised-Algorithms-in-Machine-Learning-Final-Project

Code

All code is in usl.ipynb

Data

Data is: Mall_Customers.csv

1. Project Overview

Focuses on customer segmentation using unsupervised learning techniques.

- K-Means
- Gaussian Mixture Model
- Hierarchical Clustering
- DBSCAN

The goal is to identify distinct groups of customers based on their purchasing behavior and demographic characteristics.

This segmentation can help businesses develop targeted marketing strategies and personalized customer experiences.

2. Data Collection

Dataset Size: 200 records with 5 customer attributes

Collection Method: The data was likely collected through mall membership cards and customer surveys

The dataset used is the "Mall Customer Segmentation Data" from Kaggle, which contains basic information about mall customers.

- CustomerID
- Gender
- Age
- Annual Income (k$)
- Spending Score (1-100)

Source URL: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

Initial Inspection

	CustomerID	Gender	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

3. Exploratory Data Analysis (EDA)

01

Distribution analysis of all features

02

Correlation analysis between variables

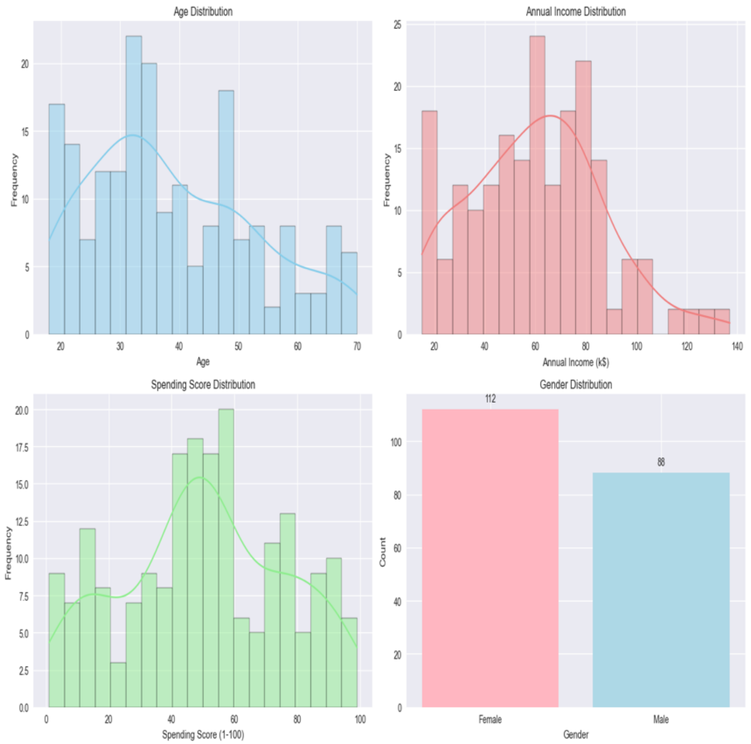
03

Identification of patterns and relationships in the data

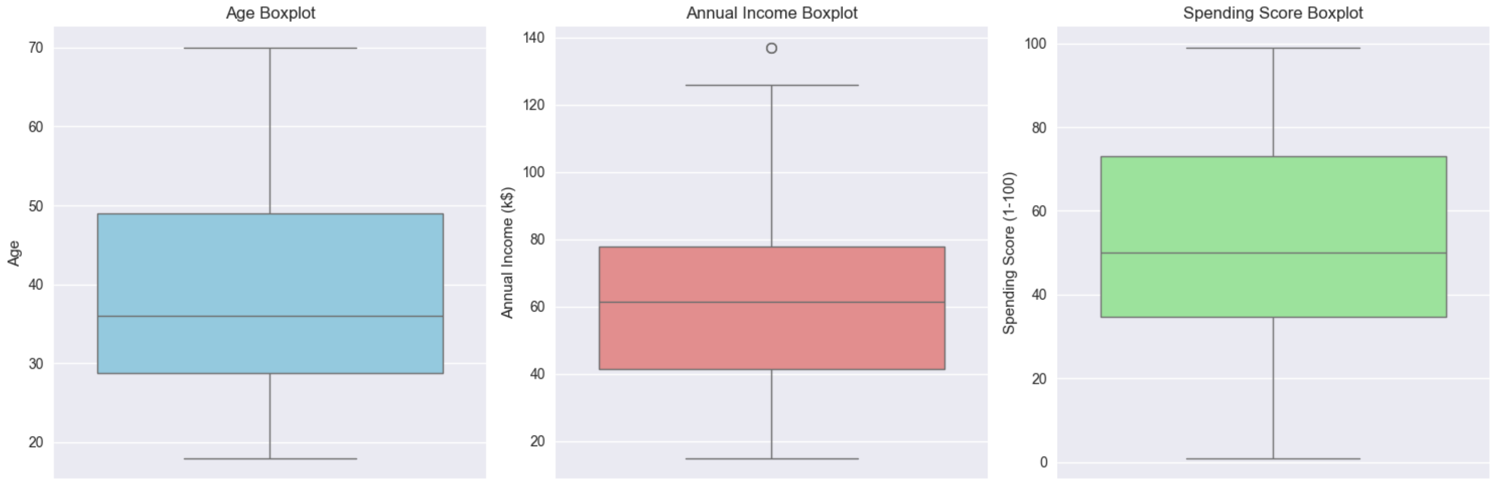
04

Outlier detection and treatment

Univariate Analysis

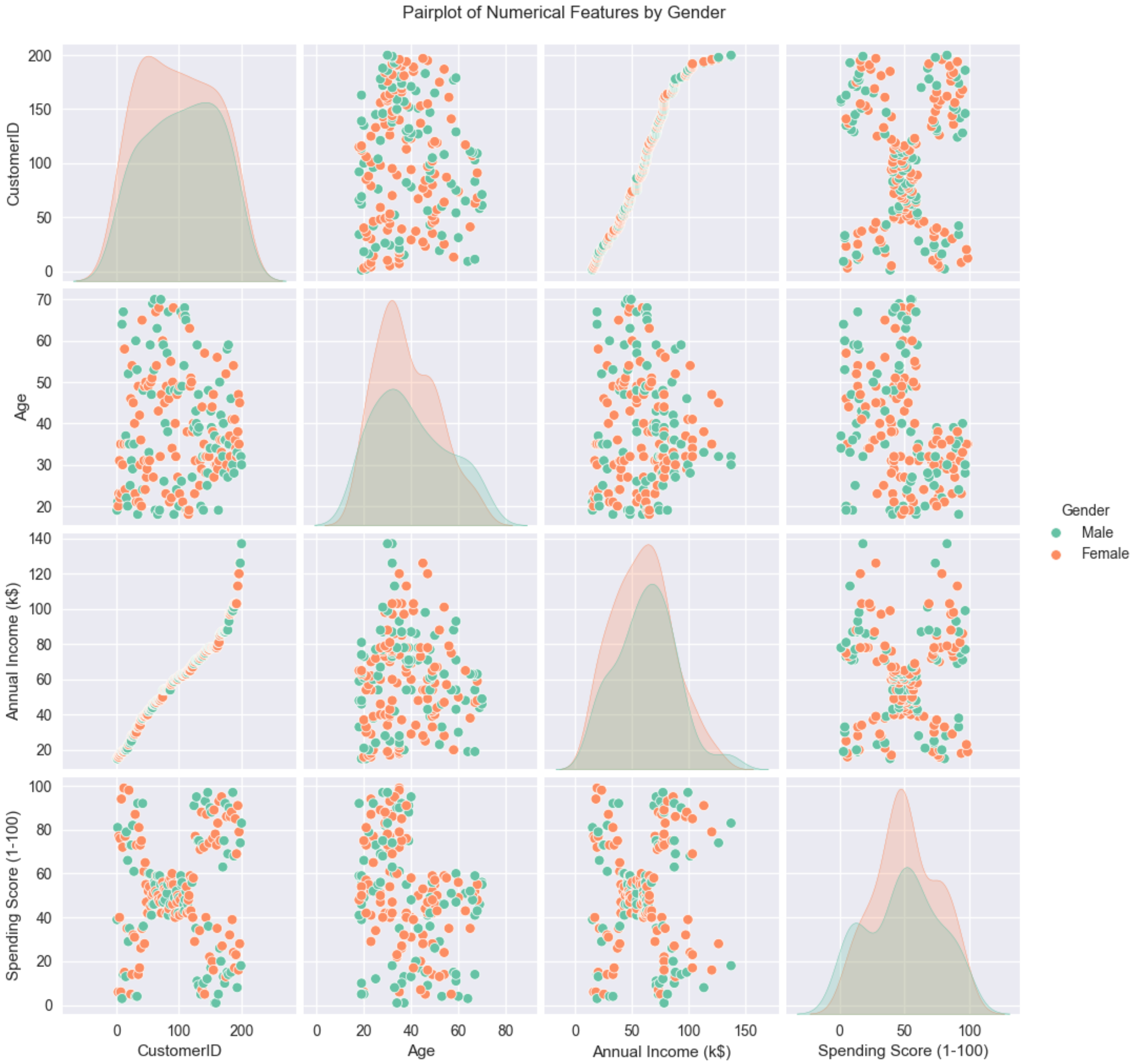


Distribution of numerical features

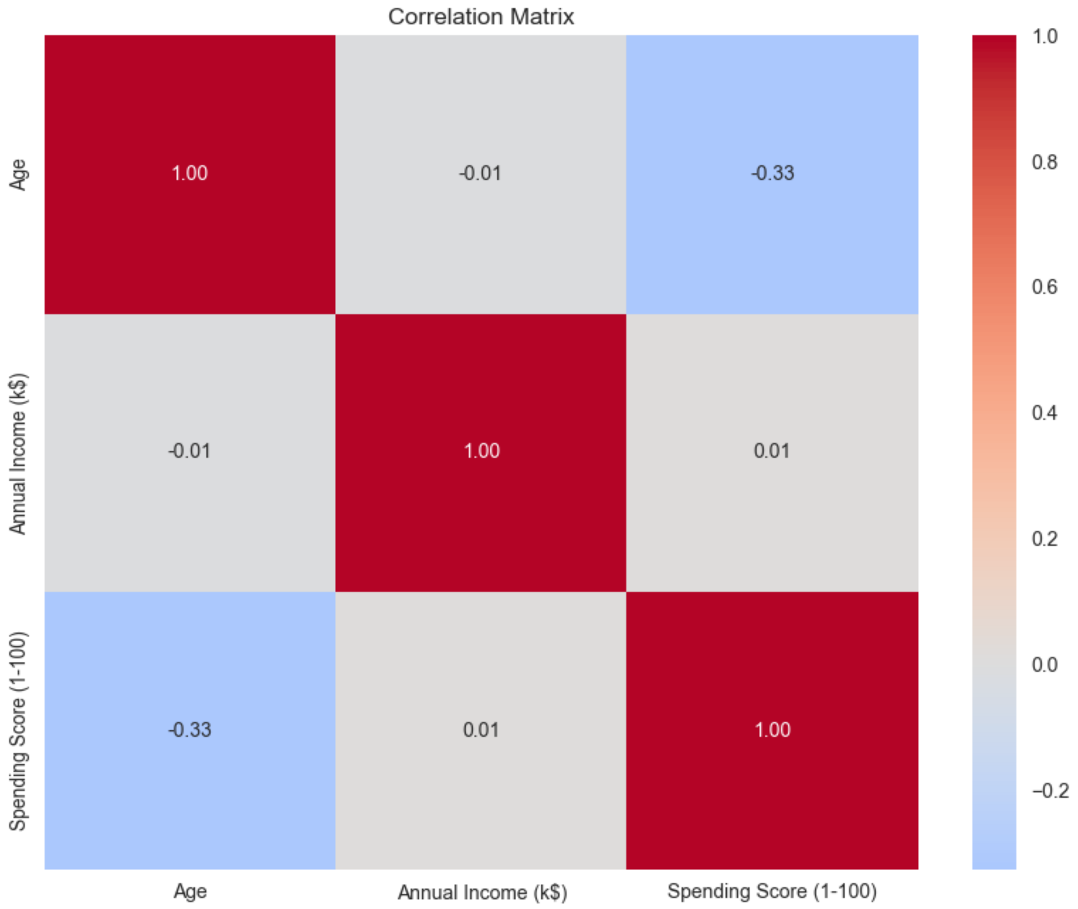


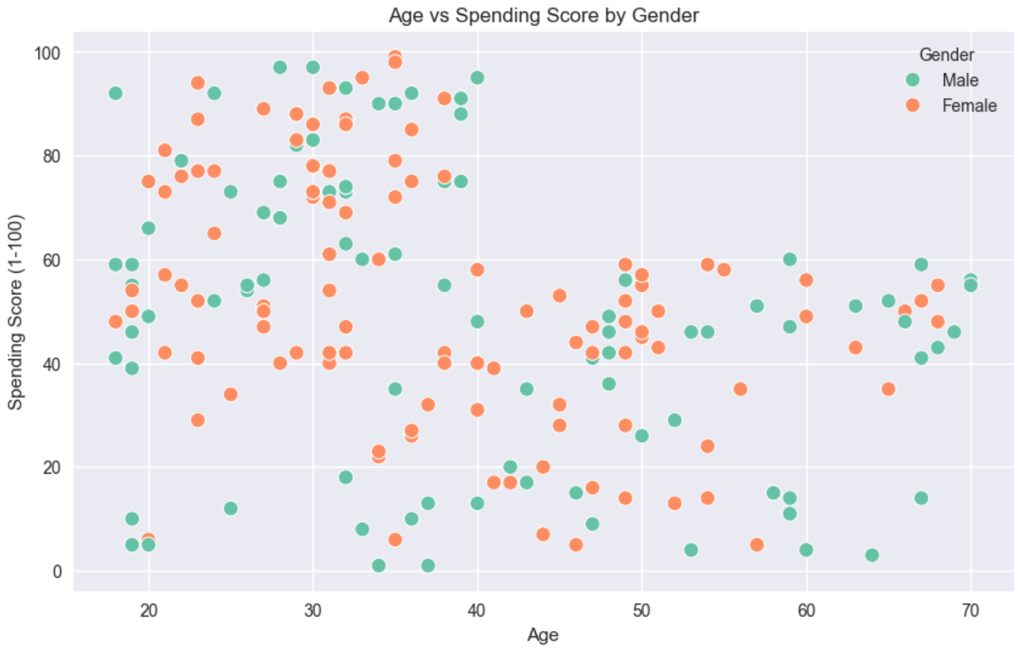
Boxplots for numerical features to identify outliers

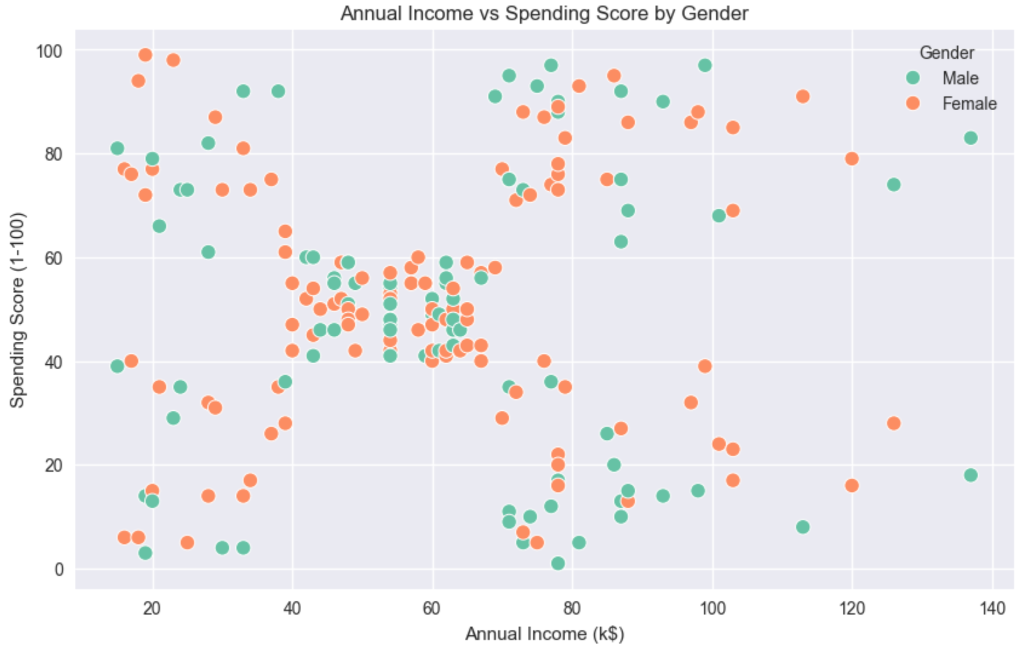
Bivariate Analysis



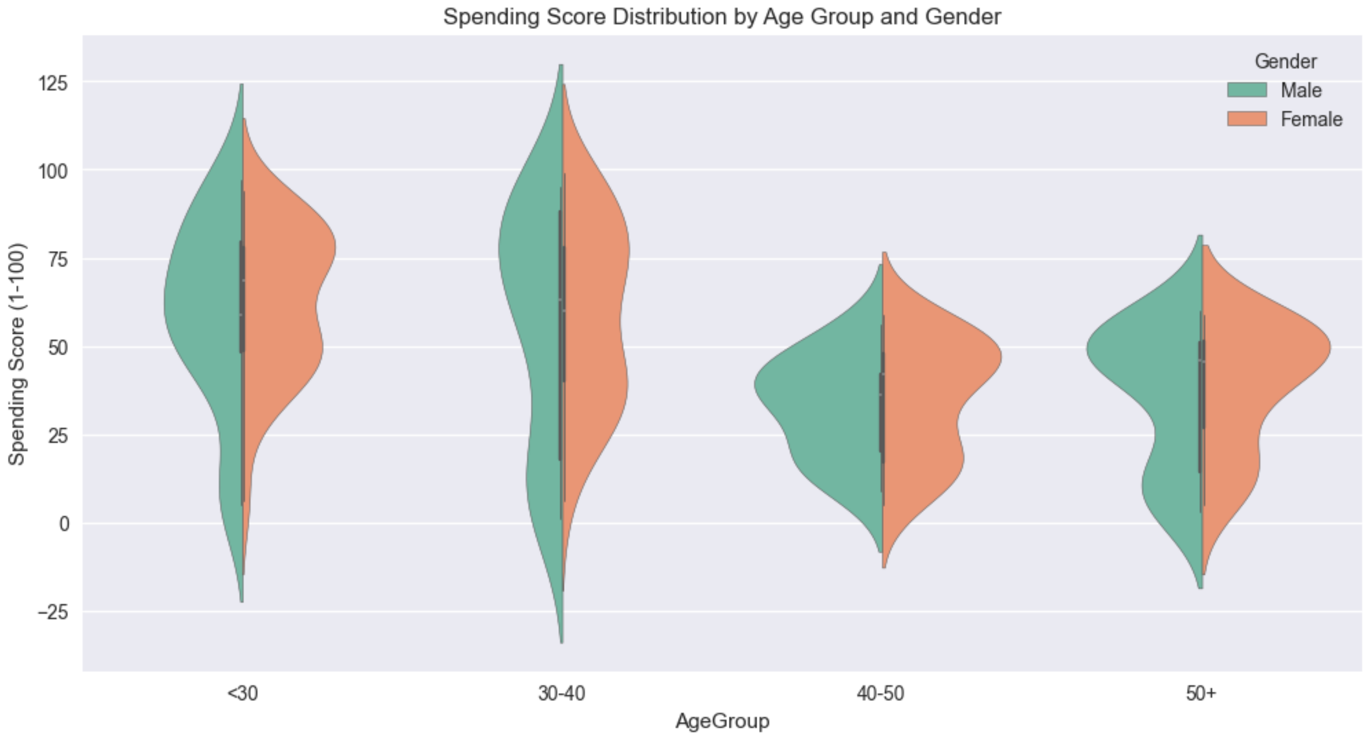
Correlation Analysis



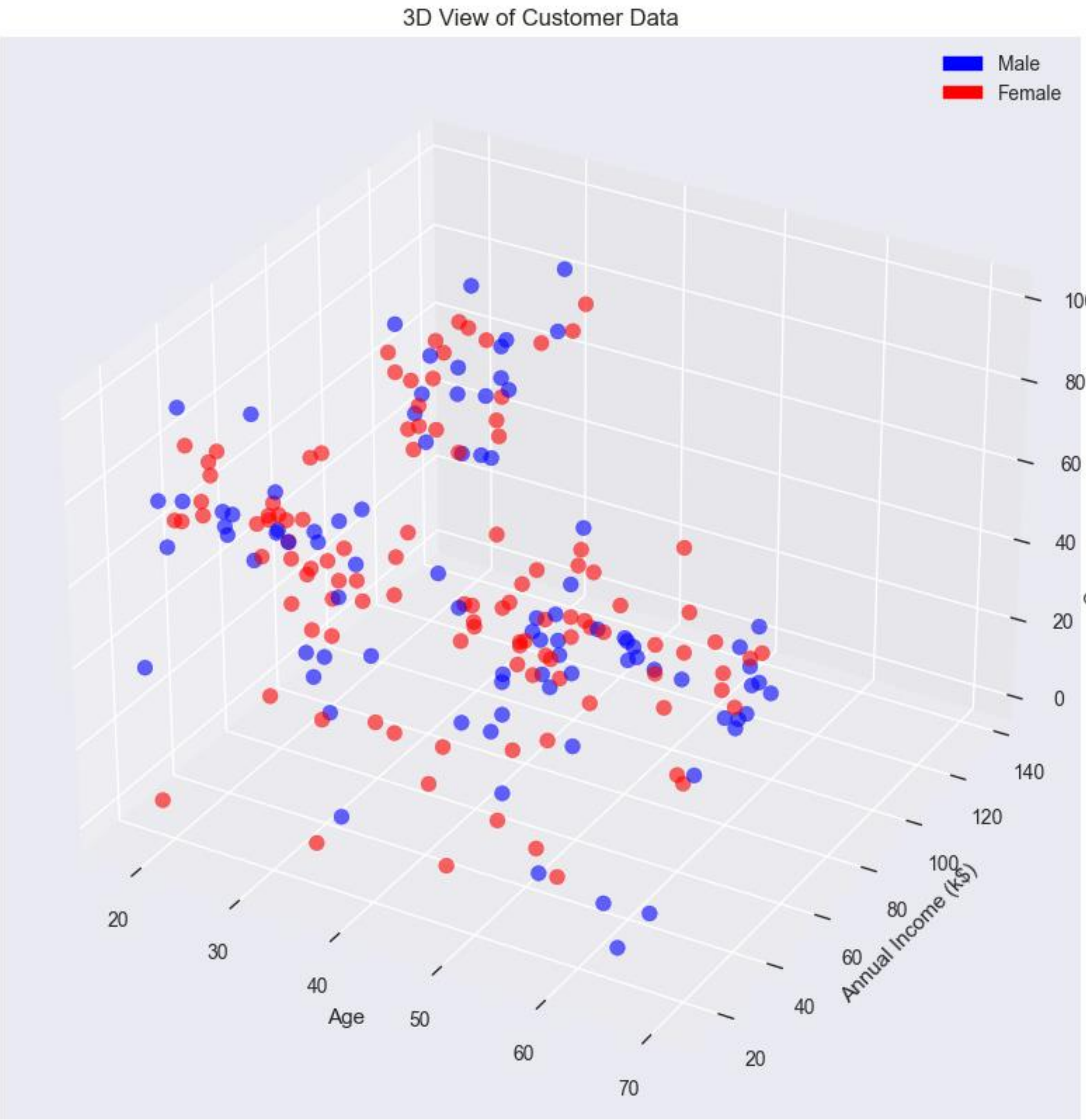




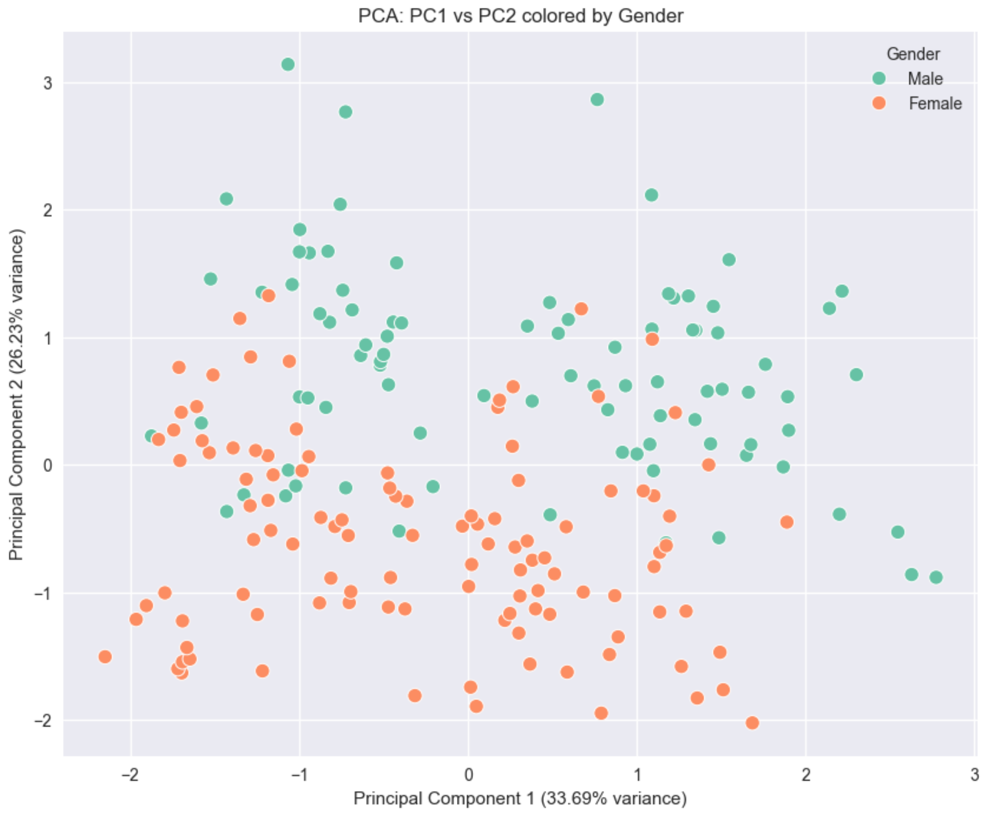
Distribution of spending by gender and age groups

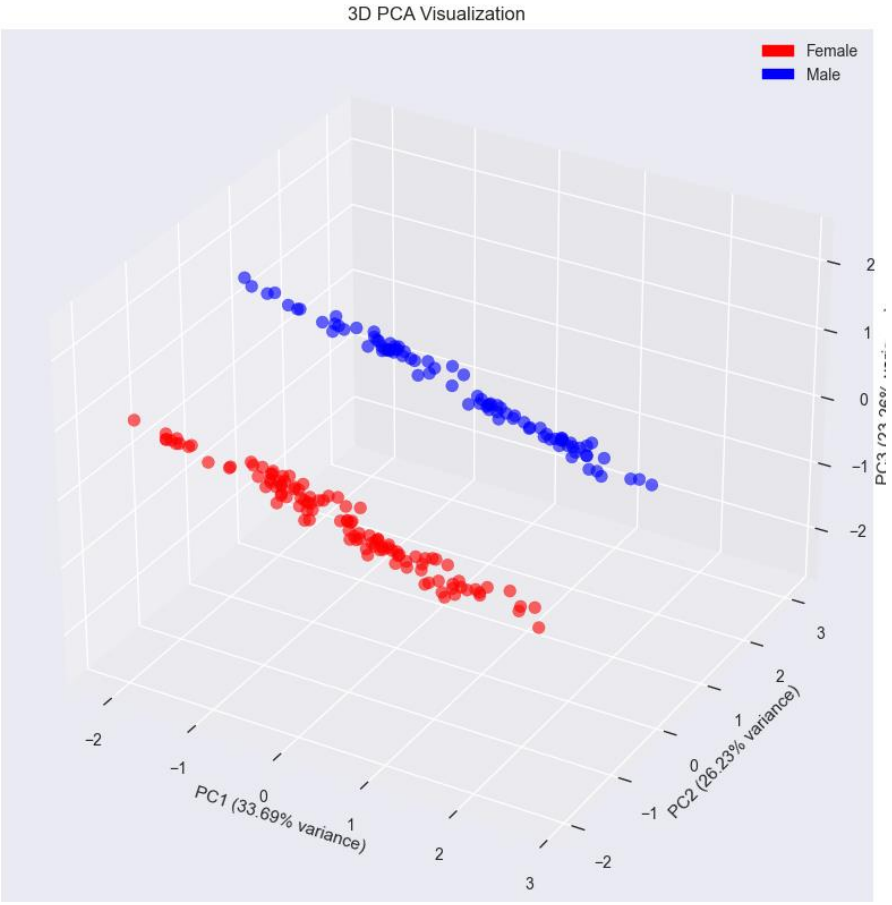


Multivariate Analysis



Dimensionality Reduction with PCA





4. Model Building and Training

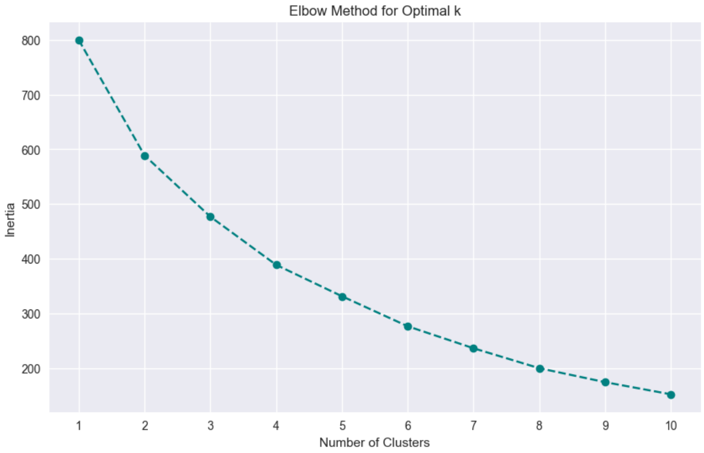
Determining Hyperparameter

Evaluation Metrics

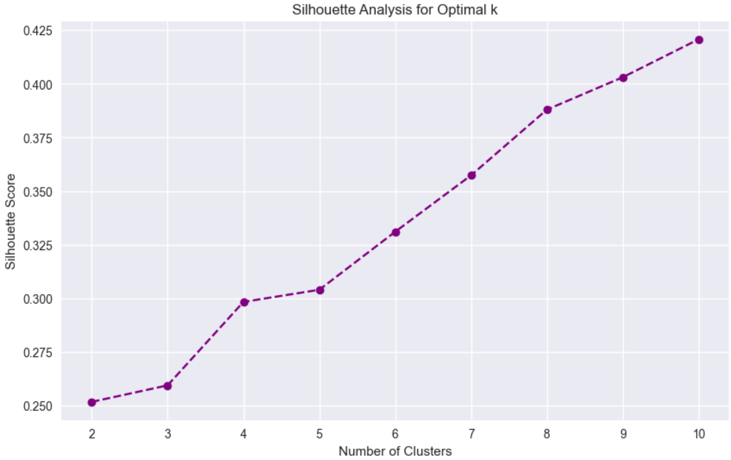
Model Training and Evaluation

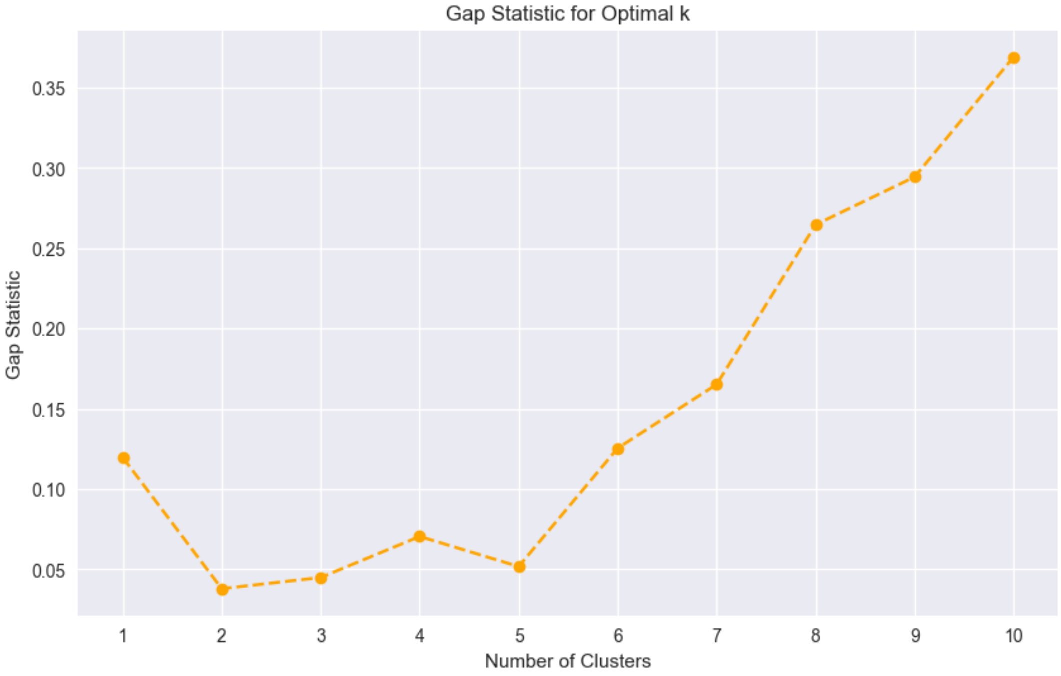
Determining Optimal Number of Clusters

Elbow Method

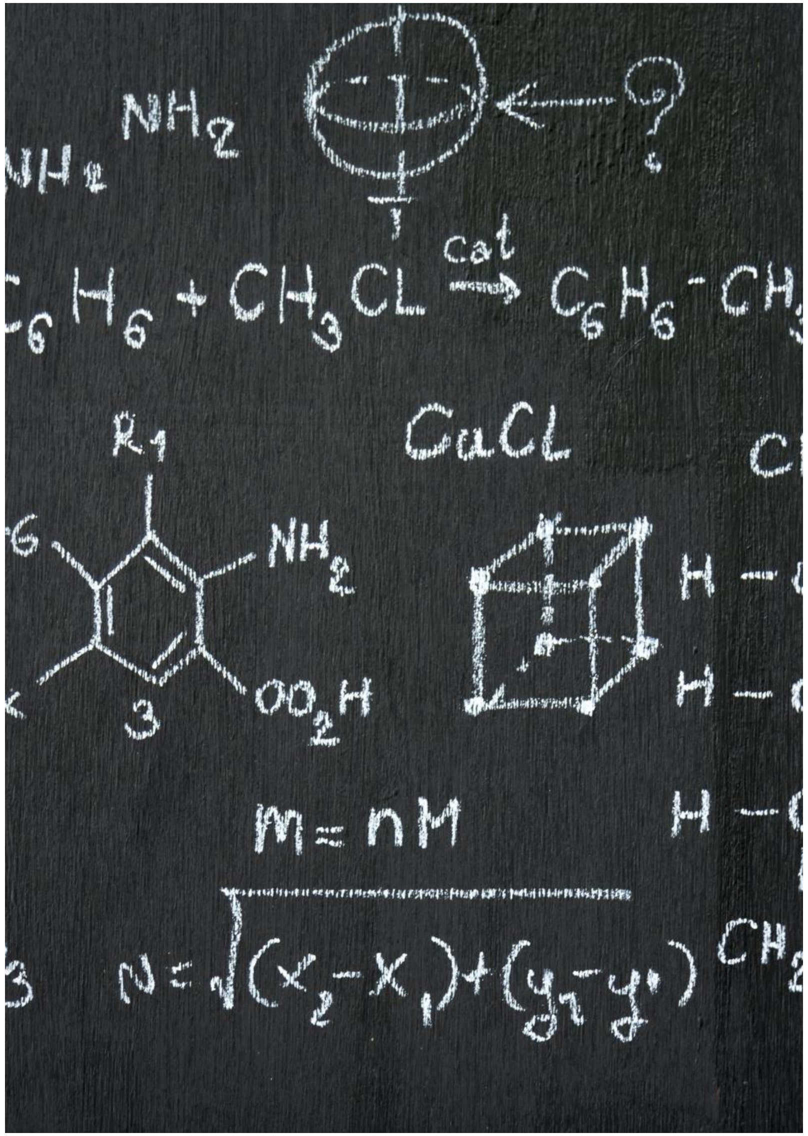

Elbow Method for Optimal k

Silhouette Analysis


Silhouette Analysis for Optimal k


Gap Statistic for Optimal k

Gap Statistic



Evaluation Metrics

- Silhouette Score
  - Measures how similar an object is to its own cluster compared to other clusters
- Calinski-Harabasz Index
  - Ratio of between-clusters dispersion to within-cluster dispersion
- Davies-Bouldin Index
  - Average similarity measure of each cluster with its most similar cluster

Model Training and Evaluation

Algorithm	Silhouette (↑)	Calinski-Harabasz (↑)	Davies-Bouldin (↓)	Notes
K-Means	**0.304**	**68.965**	**1.167**	Best overall metrics
GMM	0.222	45.817	1.211	Lower scores than K-Means
Hierarchical	0.287	64.469	1.220	Close 2nd place
DBSCAN	0.012	12.099	1.389	9 clusters + 105 noise points

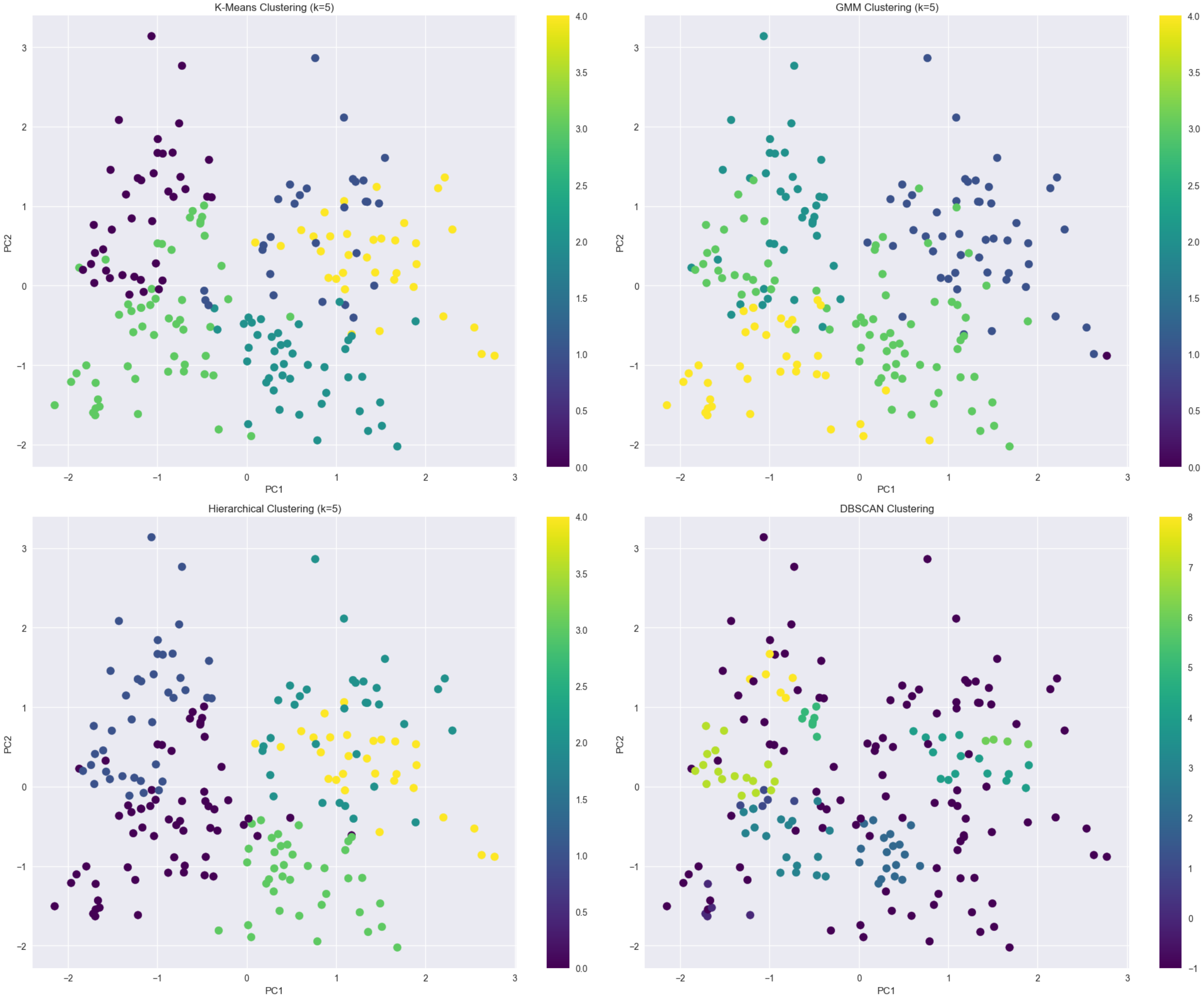
5. Results and Analysis

Cluster Visualization

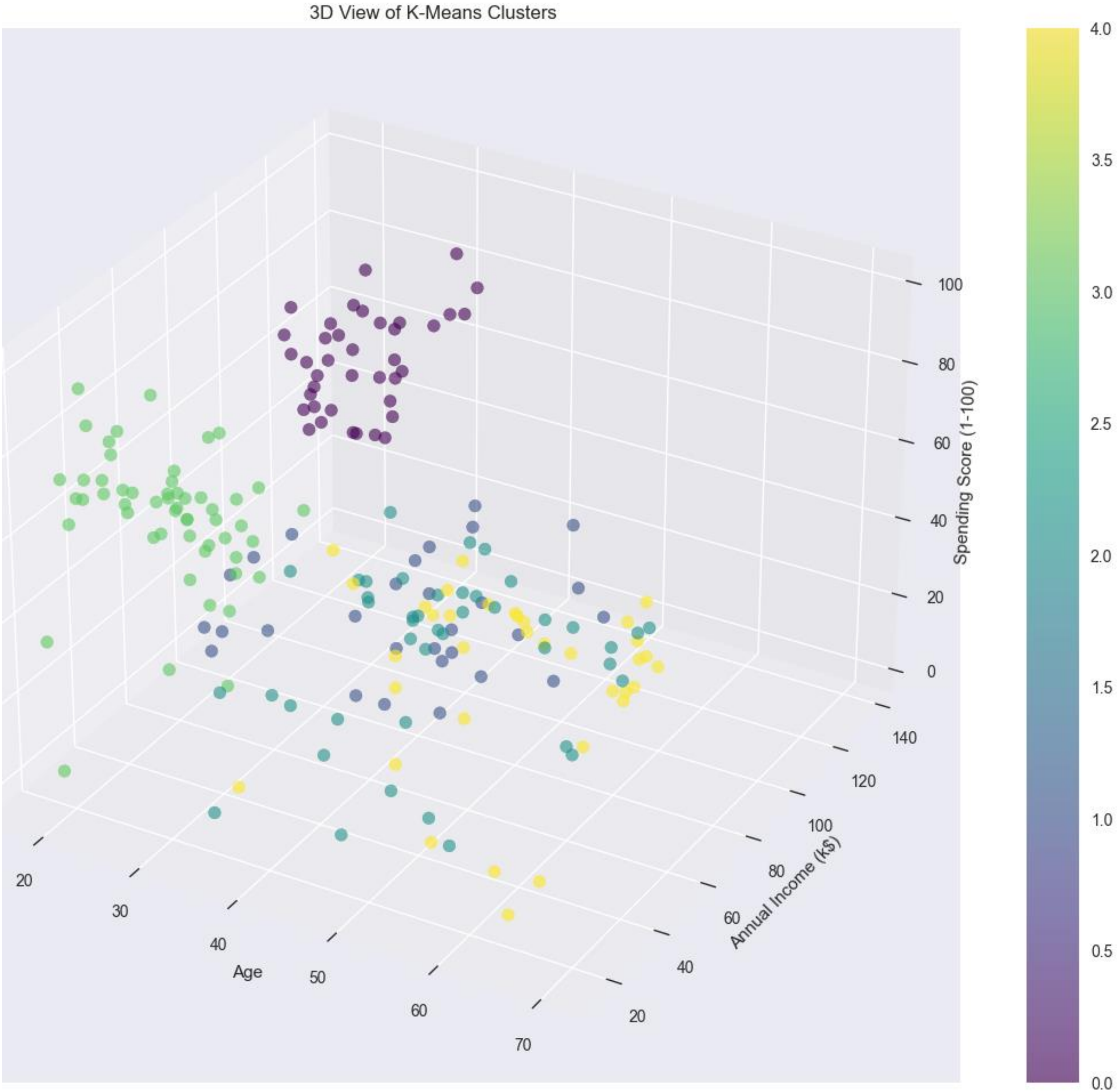
Cluster Profiling

Comparison of Algorithms

Cluster Visualization



3D visualization of K-Means clusters



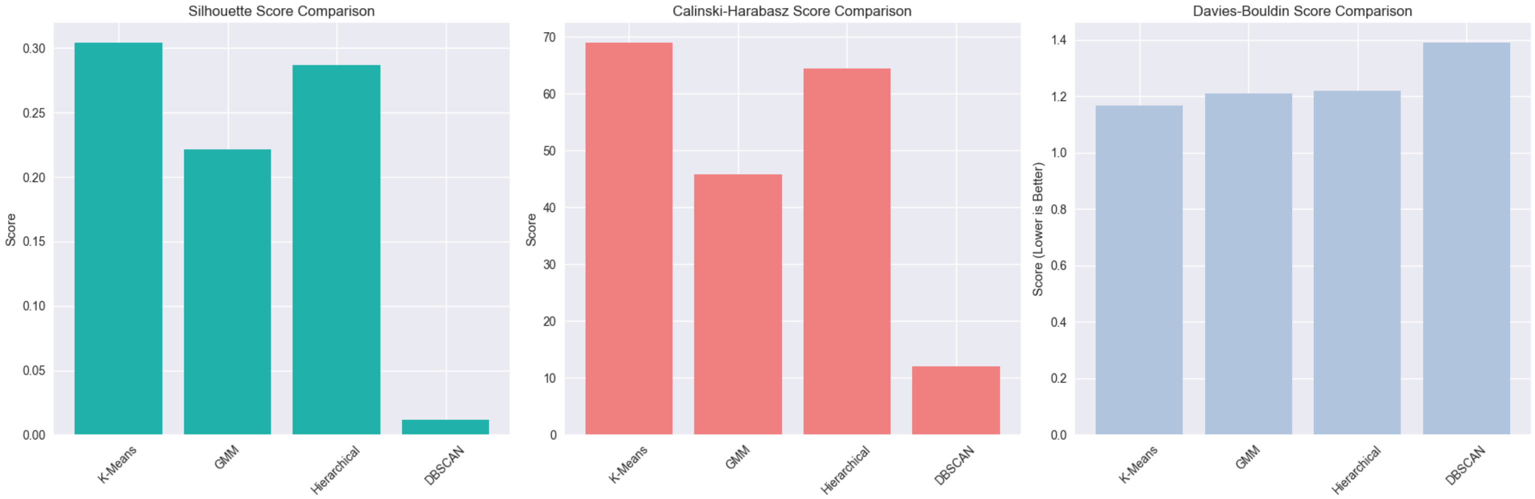
Cluster Profiling

KMEANS_CLUSTER	AVG AGE	AGE STD	AVG INCOME	INCOME STD	AVG SPENDING	SPENDING STD	FEMALE %
0	32.69	3.73	86.54	16.31	82.13	9.36	53.85
1	36.48	9.68	89.52	17.42	18.00	10.58	55.17
2	49.81	9.47	49.23	15.60	40.07	15.56	100.00
3	24.91	5.35	39.72	16.98	61.20	18.42	59.26
4	55.71	9.60	53.69	18.71	36.77	17.99	0.00

K-Means Customer Snapshots (5 Segments)

ID	Label	Profile	Go-To Strategy
0	Premium Spenders	Mid-age, high income & spend	Luxury drops, VIP perks
1	Saver-Investors	High income, low spend	Value-led, investment stories
2	Pragmatic Seniors	Older female, balanced I/S	Quality & utility focus
3	Trend Impulsives	Young, low income, high spend	Flash sales, influencer codes
4	Conservative Gents	Older male, low spend	Durability & function messaging

Comparison of Algorithms



Algorithm	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
K-Means	0.304	68.965	1.167
GMM	0.222	45.817	1.211
Hierarchical	0.287	64.469	1.220
DBSCAN	0.012	12.099	1.389

Algorithm Performance Snapshot

1

K-Means ➔ Winner
Best Silhouette, Calinski-Harabasz & Davies-Bouldin → data = well-separated, spherical clusters.

2

Hierarchical ➔ 2nd
Respectable scores; structure exists but spheres > trees here.

3

GMM ➔ 3th
Low scores refute Gaussian-distribution assumption.

4

DBSCAN ➔ Weakest
No meaningful density peaks; 105 noise points, 9 tiny clusters.