Unsupervised Machine Learning

Course Project

Evgeny Zorin

17. October 2021

Mall Customers Segmentation

Abstract

The mall owner, through membership cards, has some basic customer data such as customer ID, age, gender, annual income and spending score. The spending score is what is assigned to the customer based on certain parameters such as customer behavior and purchasing data.
The mall owner wants to understand which customers are easy to approach (target customers) so the marketing team can develop appropriate strategy for each customer segment.

Main Objective

The main objective of the analysis is to create an unsupervised learning model using clustering approach that can identify various customer segments and help the marketing team develop an appropriate marketing strategy for each customer segment.

Data Summary

The dataset was taken from Kaggle and it contains 200 observations of mall customers and 5 features as shown in Table 1:

Table 1: Columns and Descriptions

Column	Description
CustomerID	Unique ID assigned to the customer
Gender	Gender of the customer
Age	Age of the customer
Annual Income (k$)	Annual Income of the customer
Spending Score (1-100)	Score assigned by the mall based on customer behavior and spending nature

The dataset consists of 4 numerical features (**CustomerID**, **Age**, **Annual Income (k$)**, **Spending Score (1-100)**) and 1 categorical feature (**Gender**). There are no missing or duplicated values in the dataset.

Exploratory Data Analysis

Data Cleaning

CustomerID feature was dropped from the dataset, since it doesn't provide any useful information for the customer segmentation.

Summary Statistics of Numerical Features

Age feature is represented by a distribution from 18 to 70 years old. The average age of customers is 39 years old, but 50% of customers are between 18 and 36 years old. This distribution will help better understand the behavior of younger and older customers and develop an appropriate marketing strategy for each customer segment. Annual Income feature is represented by a distribution from $15K to $137K, where the average income is about $60K. Summary statistics of the numerical features is shown in Table 2.

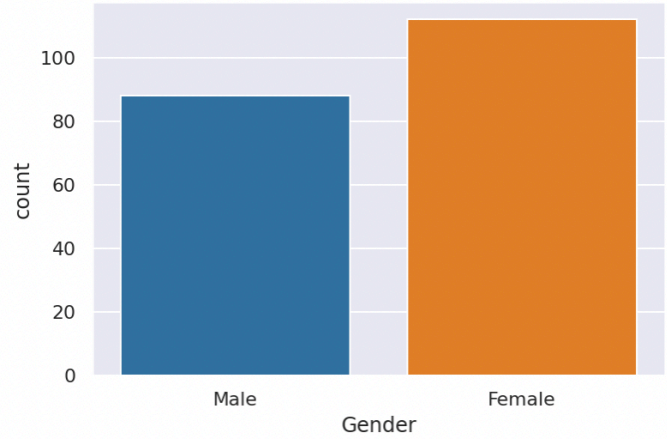
Table 2: Summary Statistics of Numerical Features

	count	mean	std	min	25%	50%	75%	max
Age	200.0	38.85	13.969007	18.0	28.75	36.0	49.0	70.0
Annual Income (k$)	200.0	60.56	26.264721	15.0	41.50	61.5	78.0	137.0
Spending Score (1-100)	200.0	50.20	25.823522	1.0	34.75	50.0	73.0	99.0

Feature Distribution

Distribution of Gender feature (Figure 1) is slightly unbalanced, with 112 female customers (56%) and 88 male customers (44%).

Figure 1: Gender Distribution



Distribution of the average feature values in relation to Gender (Table 3) shows that female customers, on average, are younger than male customers, tend to have lower incomes, but spend more.

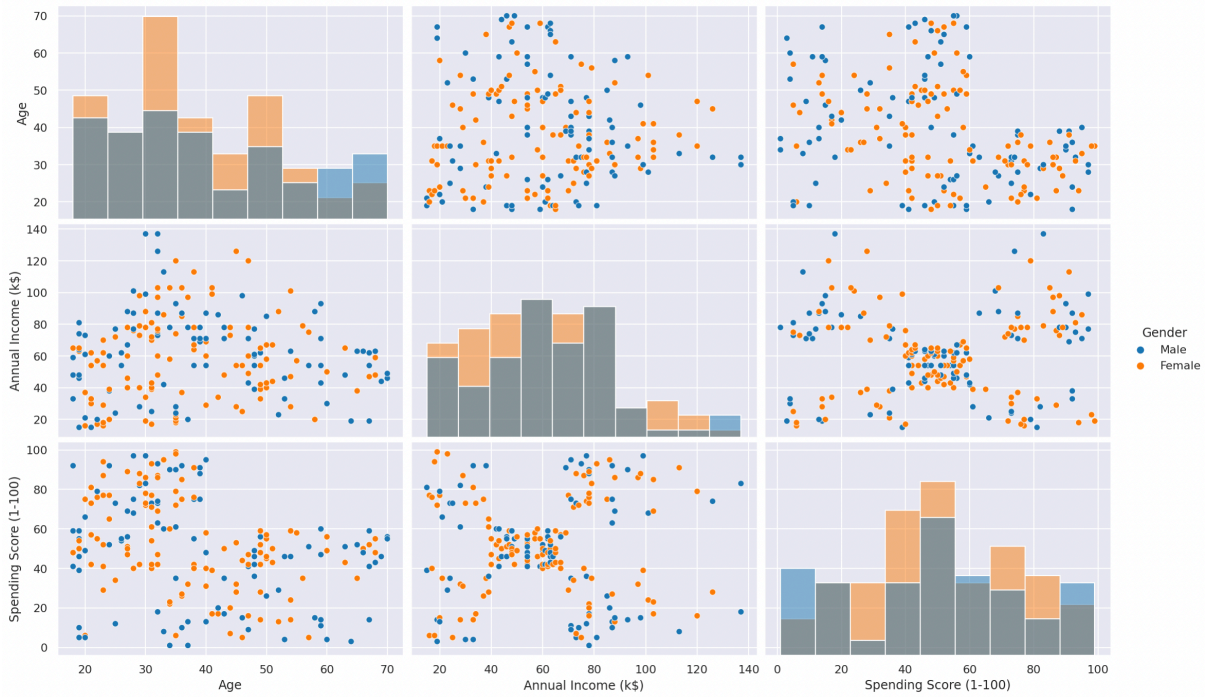
Table 3: Average Distribution by Gender

Gender	Age	Annual Income (k$)	Spending Score (1-100)
Female	38.098214	59.250000	51.526786
Male	39.806818	62.227273	48.511364

Analyzing the pairwise relationships between the features (Figure 2), the following insights can be made:

- Most customers (>50%) are between the ages of 18 and 40. Male customers have a more uniform distribution, while female customers stand out in the group of 30-35 year old.
- The main distribution of annual income for all customers is between $15K and $85K. There are a small number of customers between the ages of 27 and 60 who have incomes above $85K, and only a few have incomes above $100K. Female customers initially have a lower annual income, which increases with age. Male customers initially have a higher annual income, which decreases as they get older.
- The main distribution of spending scores for all customers is between 40 and 60. Spending scores higher than 60 are common for customers younger than 40 years old, the majority of them are females. Spending scores lower than 20 are common for male customers.
- The relationship between annual income and spending score indicates that there is a clear segmentation between these features, pointing to 5 different clusters. This assumption will be verified using the 4 most popular clustering algorithms.
- The Gender and Age features will not be used in building a clustering models, as they do not provide any clear patterns allowing the segmentation of customers.

Figure 2: Pairwise Relationship



Cluster Modeling

Feature and Clustering Algorithms Definition

Based on the Exploratory Data Analysis, it was decided to fit the dataset with only two features Annual Income (k$) and Spending Score (1-100) to the most common clustering algorithms, such as:

- K-Means
- Hierarchical Agglomerative Clustering (HAC)
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
- Mean Shift

K-Means

Selecting a number of clusters is the most challenging part of setting this algorithm. One of the simplest and the most popular approaches is the elbow method. In this method, inertia will be calculated for the number of clusters from 2 to 10. The rule is to choose the number of clusters at which an inflection or "elbow" is visible on the plot. Additionally, a silhouette score method will be used to confirm the choice based on the elbow score method. The higher the

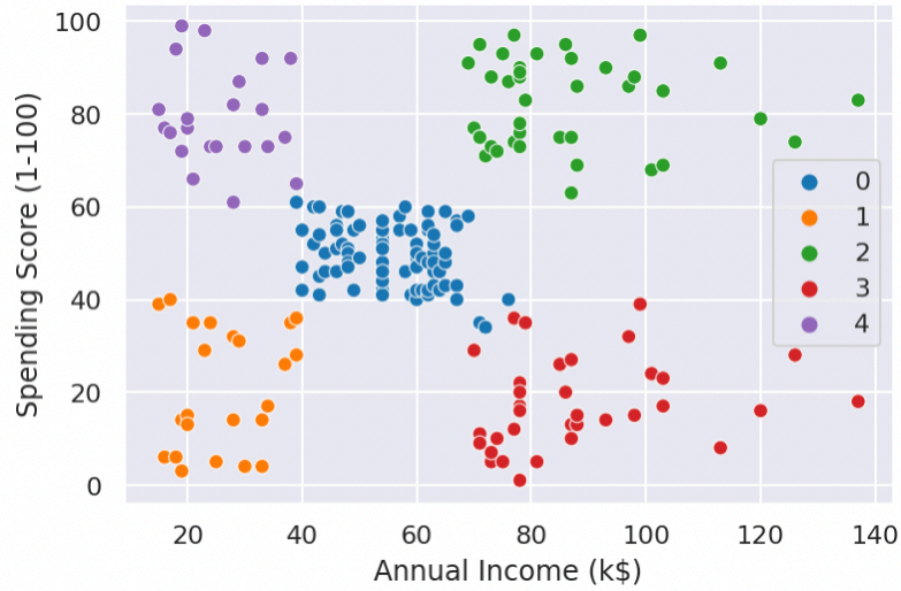
silhouette score, the better the selected number of clusters can segment the data. The combined plot of the elbow method and the silhouette score method (Figure 3) shows that the appropriate number of clusters is 5, which also confirms the assumption made in the exploratory data analysis.

Figure 3: The Elbow Method and The Silhouette Score Method

K-Means algorithm generated the following 5 clusters (Figure 4):

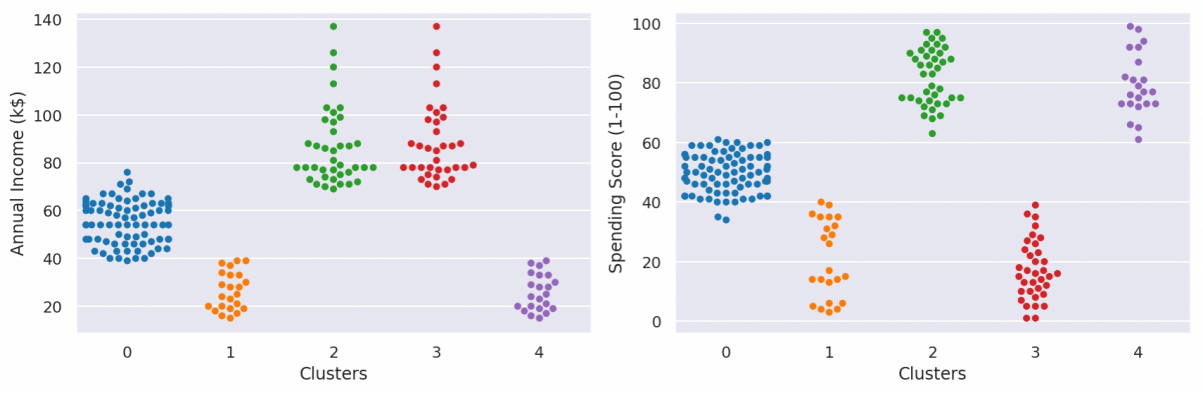
- Customers with low annual income and high spending score
- Customers with low annual income and low spending score
- Customers with medium annual income and medium spending score
- Customers with high annual income and low spending score
- Customers with high annual income and high spending score

Figure 4: K-Means Results



The distribution of features by clusters can be seen more clearly in Figure 5.

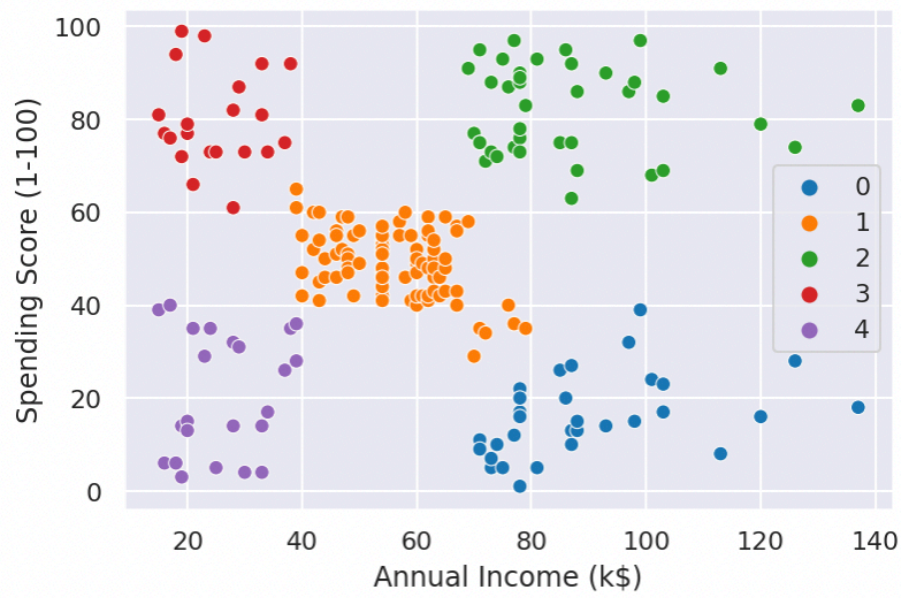
Figure 5: Feature Distribution by Clusters



Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering (HAC) is a bottom-up approach that requires two inputs: the number of clusters and the linkage criteria. As in the K-Means algorithm, 5 clusters will be used in this approach as well, and Ward's method will be used as the linkage criteria. HAC seems to provide slightly worse results than the K-Means algorithm (Figure 6). In particular, the central cluster includes some customers explicitly belonging to other clusters.

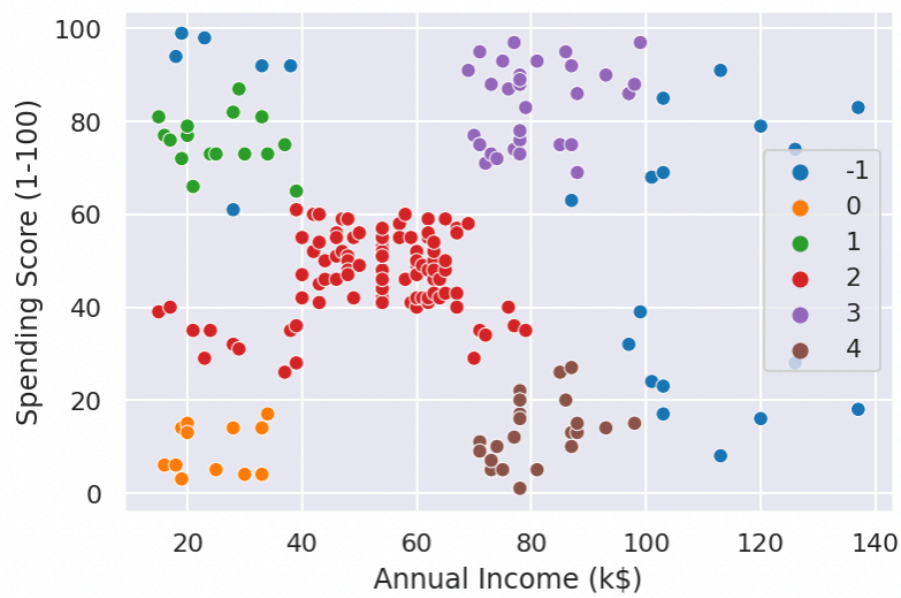
Figure 6: Hierarchical Agglomerative Clustering Results



Density-Based Spatial Clustering of Applications with Noise

Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). For this approach, 7 points will be used as the minimum number of points, and epsilon 11 will be used as the radius of the density area. DBSCAN provides much worse results than the previous two algorithms (Figure 7). Cluster -1 indicates outliers, which are quite a lot seen in the plot. This could be because the data density is not high enough. The results would probably be much better if the dataset included more data.

Figure 7: DBSCAN Results

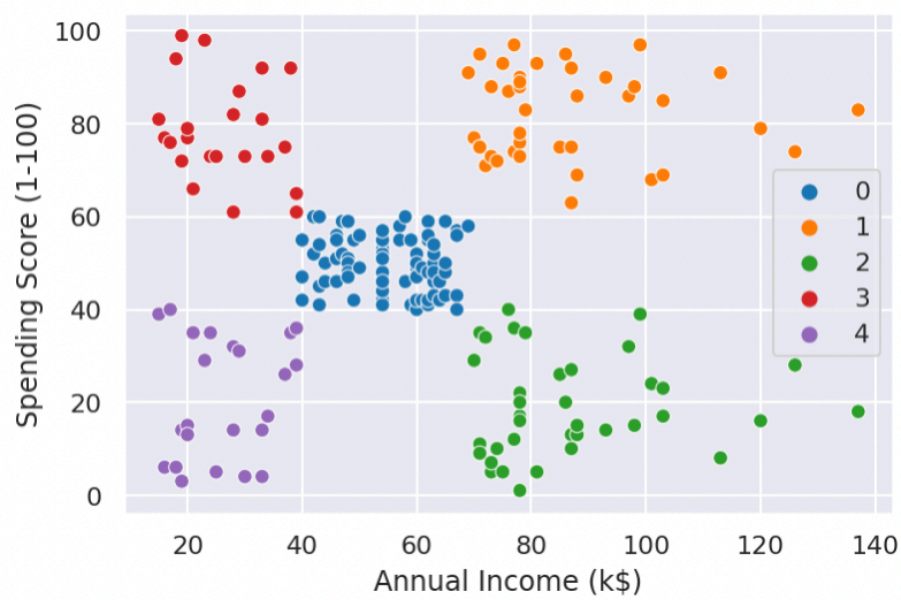


Mean Shift

MeanShift clustering aims to discover “blobs” in a smooth density of samples. It is a centroid based algorithm, which works by updating candidates for centroids to be the mean of the points within a given area. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids.

The algorithm automatically sets the number of clusters, instead of relying on a parameter bandwidth, which dictates the size of the area to search through. This parameter can be set manually, but can be estimated using the provided estimate_bandwidth function, which is called if the bandwidth is not set.
Mean Shift seems to provide the best results among the algorithms evaluated so far (Figure 8). Cluster boundaries are clearly defined, and each customer is in the cluster in which he should be.

Figure 8: Mean Shift Results



Clustering Model Evaluation

Evaluating all four clustering models based on the visual accuracy (Figure 9) it can be seen that both the K-Means and Mean Shift algorithms have the highest segmentation accuracy. It was decided to use the model based on the Mean Shift algorithm as the preferred model for customer segmentation, because of the more accurate results. The clustering results of all models were added to the dataset for a better evaluation of customer segmentation (Table 4). Based on a small sample of customers, it can be seen how they were segmented by each of the algorithms.

Figure 9: Clustering Model Evaluation

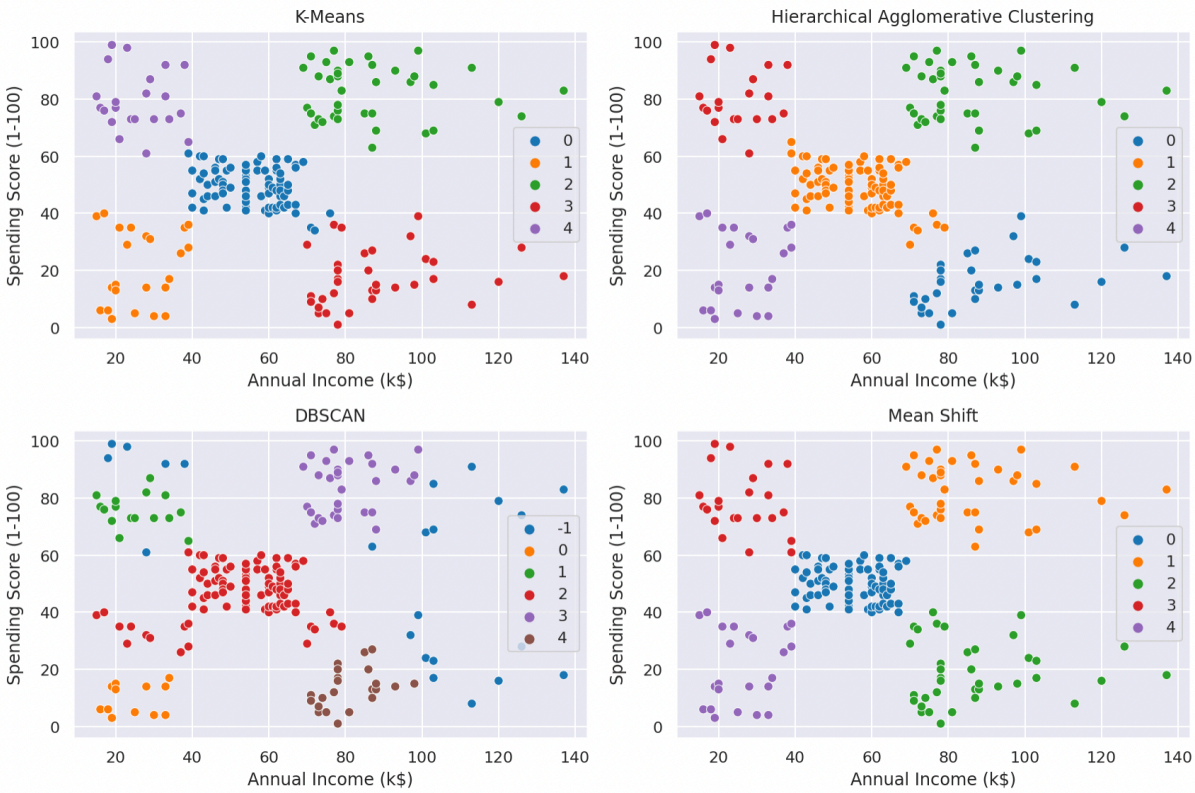


Table 4: Dataset Sample

	Gender	Age	Annual Income (k$)	Spending Score (1-100)	K-Means	Agglomerative	DBSCAN	Mean Shift
26	Female	45	28	32	0	4	2	4
130	Male	47	71	9	2	0	4	2
135	Female	29	73	88	3	2	3	1
19	Female	35	23	98	4	3	-1	3
66	Female	43	48	50	1	1	2	0
60	Male	70	46	56	1	1	2	0
115	Female	19	65	50	1	1	2	0
36	Female	42	34	17	0	4	0	4
53	Male	59	43	60	1	1	2	0
95	Male	24	60	52	1	1	2	0

Key Findings and Insights

For a more accurate customer segmentation, it is not enough just to select an appropriate clustering algorithm. It is also important to understand which

features to use, which patterns they form, and how those patterns can be interpreted. It is recommended to use only numerical features in cluster models, and it is strongly not recommended to use categorical features, encoding them into numerical ones. From the above evaluation, it is clear that DBSCAN failed to generate reasonable clusters. It is due to its problems in recognising clusters of various densities (which are present in this case). In turn, K-Means, Hierarchical Agglomerative Clustering and Mean Shift algorithms created reasonable clusters. The best clustering algorithm in this case is Mean Shift, which provides a clear customer segmentation.

Next Steps

The next steps to achieve more accurate and appropriate segmentation would be to obtain and analyze a larger dataset with more observations and features. Perhaps new features will be able to identify new patterns, and more customer observations will be able to identify more clusters that better explain customer behavior.

Appendix

GitHub URL:
https://github.com/evgenyzorin/IBM-Machine-Learning/tree/main/Clustering