Mall Customer Segmentation using Unsupervised Learning Clustering Algorithms

Student Mubarak Babslawal
Course Statistical Learning
Programme Data Science in Economics
Github Repo https://github.com/Mubarakbabs/SL_final_project/

Abstract

This report explores the application of two clustering techniques, k-means and hierarchical clustering, to a dataset of 200 customers of a mall. The dataset contains information on age, gender, annual income, and spending scores. The objective is to identify meaningful customer clusters that can inform key elements of the marketing mix, including product pricing, placement, development, and promotion strategies.

Six distinct clusters were identified using k-means, as well as using three different linkage methods with hierarchical clustering. The clusters were however unidentical across the methods, with Ward D2 linkages providing similar clusters to Kmeans, while Complete and Average Linkages produced identical clusters. In all cases, the clusters highlighted demographic factors and spending habits that could be used as a means to provide a more personalized shopping experience for customers

A comparative analysis of the two methods reveals that while hierarchical clustering offers flexibility in selecting the number of clusters, k-means directly optimizes within-cluster variance. The analysis of various hierarchical clustering methods also shows that while Complete Linkages optimize the within-cluster variance, the final decision on the best clustering method is context-dependent.

The report concludes that all methods provide valuable insights, and any analysis should iteratively test the various methods when selecting a clustering method to use. The context, the results, the goals and the subject matter will all influence the final selection of clustering methods.

1 Statement of the Problem

The goal of the analysis is to produce clusters that can influence the key elements of the marketing mix which in this case includes:

1. Product pricing: Introducing pricing systems that target specific clusters
2. Product placement: Ensuring products targeted at the same clusters are placed close to each other to optimize customer experience, and ensuring each customer comes in contact with all products targeted at them
3. Product development: Allow the mall to develop products that will match the taste of their customers and get maximum value
4. Promotion: Run appropriate promotions that considers the customer demographics that are most common in each cluster (so we should look at age distribution in each cluster)

We perform this analysis using two different clustering methods and compare the results of both methods.

2 Description of Dataset

The dataset consists of 200 observations across four major variables: age, gender, annual income in USD, and a scaled measure of spending propensity ranging from 1-100 called the Spending Score. A higher spending score indicates that the customer tends to spend more on each order.

The data has no missing values

The split betweeen male and female in the data is almost equal

Female	Male
112 (56%)	88 (44%)

Table 1: Gender distribution

Distribution of other variables

We also examine the distribituon of customers across age groups, annual income and spending score.

We observe that more than half of the customers are below the age of forty. The annual income is right-skewed, with income peaking between $70-$80k. Finally, the spending score is a bit more bell-shaped, with the most users having a medium spending score and an almost even distribution of users with conservative and lavish spending habits.

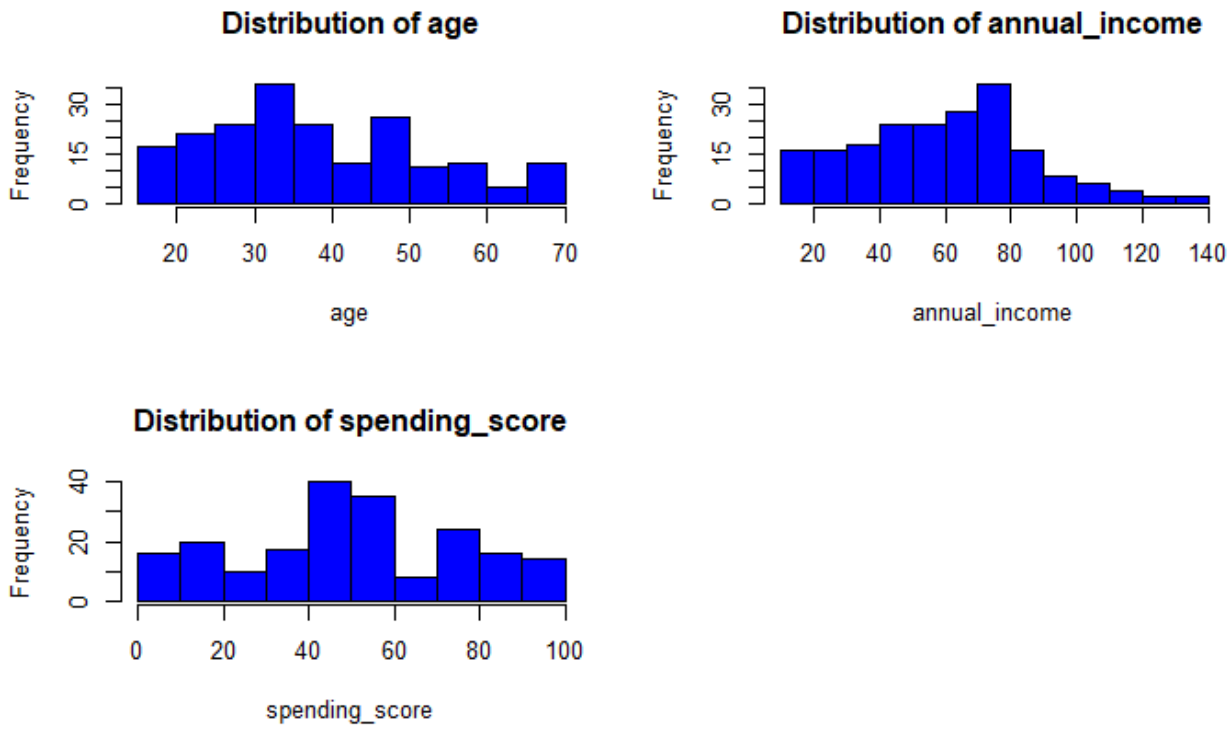

Figure 1: Numeric variables distribution

3 Key Findings

The comparison of the two clustering methods reveals the following key insights:

1. Using different methods such as kmeans and hierarchical clustering can produce different clusters, and produce more insights as to how the data can be divided.
2. Hierarchical clustering can be an easier way to select the number of clusters.
3. Modifying model parameters (such as number of clusters for kmeans, cutoff height and linkage types for hierarchical) is also a way of producing various clusters.
4. There is no single best method for evaluating the performance of clustering models.

4 Analysis of Key Findings

In this section, we will analyze the evidence that led to the above key findings. These will be discussed under three sections:

1. Analysis of Clusters
2. Selecting Number of Clusters Using Hierarchical Clustering
3. Comparison of Model Performance

4.1 Analysis of Clusters

Using kmeans algorithm to divide the data into clusters produced six clusters. The data was divided into six clusters as shown below. Since the aim is to use the clusters for marketing, I took the liberty of giving them memorable names that captured their characteristics across all the features that would be fit for building user personas.

1. Mama Savers: All female, older, low to average income, medium spending score.s
2. Gentlemen Bargainers: All male, older, low to average income, medium spending score.
3. Elite Spenders: All male, across all income levels, high spending score, younger males.
4. Chic and Stylish: 92% female, high income, high spending score, younger (average age 32)
5. Vibrant Shoppers: All female, young (average age: 25), low to average income, across spending scores.
6. Frugal Affluents: High income, low spending score, with a near-even mix of male and female, age indiscriminate.

These clusters give the mall an opportunity to design products not just along gender lines, but also to design their pricing and promotions to target each cluster. For example, promotions targeted at Chic and Stylish shoppers will focus more on high quality than lower prices. On the other hand, products targeted at Elite Spenders may be high quality, but the mall may want to offer credit options that allow the lower income members of the cluster access it. These clusters can also be a basis for building user personas of the ideal customers.

HierarchicalCompleteCluster	mean_age	percent_male	mean_spending_score	mean_annual_income
⟨*fctr*⟩	⟨*dbl*⟩	⟨*dbl*⟩	⟨*dbl*⟩	⟨*dbl*⟩
1	25.12500	1.0000000	59.45833	39.83333
2	25.94595	0.0000000	57.45946	42.21622
3	53.37143	0.4142857	40.01429	50.05714
4	33.27778	1.0000000	82.66667	87.11111
5	32.19048	0.0000000	81.66667	86.04762
6	39.86667	0.5666667	16.10000	90.50000

6 rows

Table 2: Hierachical Clusters Using Complete Linkage

Using hierarchical clustering, we also got six clusters from the dataset, both using average and complete linkage. Some of these clusters are exactly the same as those from kmeans. Particularly, Frugal Affluents and Vibrant Shoppers have the same exact splits as Clusters 2 and 6 from Complete Hierarchical Clustering. In addition, Cluster 5 is almost the same as Chic and Stylish, except that it's an exclusively female cluster in the Complete Hierarchical Splitting. The other three clusters are completely different.

Using Ward D2 linkages however, we also obtain six clusters. These clusters are more similar to clusters obtained from Kmeans.

HierarchicalWardCluster	mean_age	percent_male	mean_spending_score	mean_annual_income
⟨*fctr*⟩	⟨*dbl*⟩	⟨*dbl*⟩	⟨*dbl*⟩	⟨*dbl*⟩
1	28.61905	1.0000000	69.40476	60.09524
2	25.94595	0.0000000	57.45946	42.21622
3	50.60976	0.0000000	40.14634	49.65854
4	57.27586	1.0000000	39.82759	50.62069
5	32.19048	0.0000000	81.66667	86.04762
6	39.86667	0.5666667	16.10000	90.50000
6 rows				

Table 3: Hierachical Clusters Using Ward D2 Linkage

The same clusters present in the Complete hierarchical clustering: (Frugal Affluents and Vibrant Shoppers) are present here. In addition, Mama savers is present in cluster 3. The other three clusters, while not exactly the same figures, have similar characteristics. So Cluster 1: Elite Spenders; Cluster 3: Gentlemen Bargainers; and Cluster 5: Chic and Stylish.

While there is no easy way to decide on which clustering is more ideal, domain knowledge can be used to deduce which clusters more closely resemble reality, and is more actionable for achieving the goal. For example, a marketer could decide that Elite Spenders as a cluster would be difficult to target because it includes individuals from all income levels.

This establishes the importance of trying different methods when carrying out unsupervised learning. Even within the same model, varying parameters can be used to obtain different results.

4.2 Selecting Number of Clusters Using Hierarchical Clustering

Hierarchical clustering may provide a more versatile way of choosing the best cluster. While the elbow method may be a good heuristic for selecting the number of clusters for a kmeans algorithm, it may not always offer clear cut results. For example, while building the model, using the elbow method resulted in the below chart

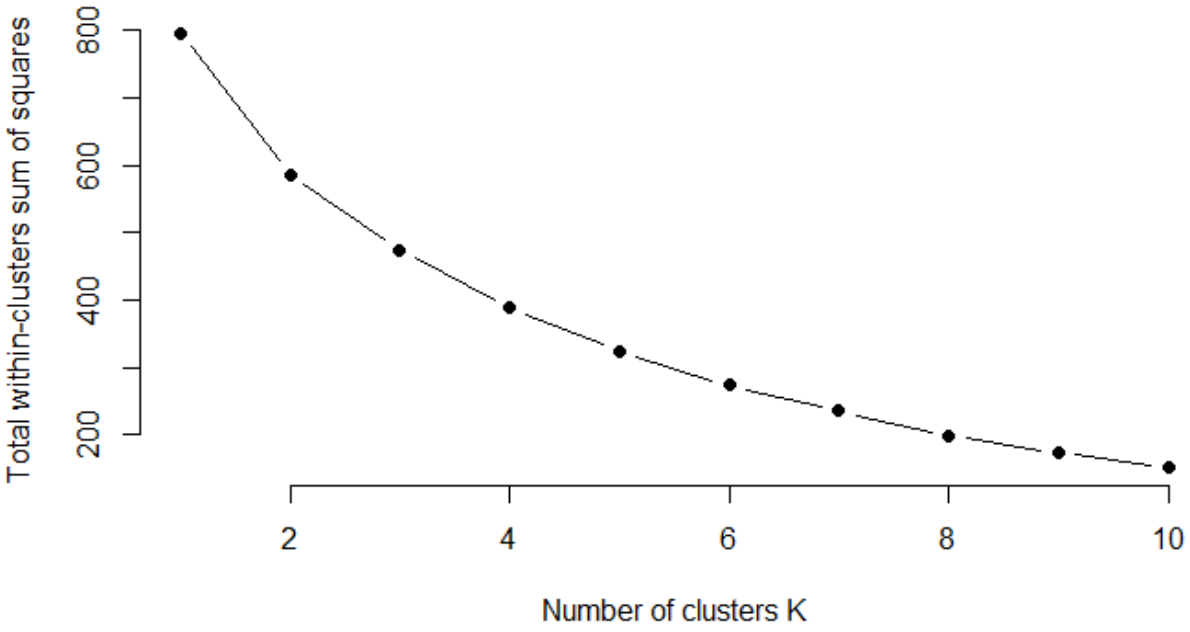

Figure 2: Selecting Number of Clusters using the Elbow Method

Looking at the chart, there is no clear point where the within sum of squares starts to decelerate. Hence, a choice of four clusters would be as reasonable as 6, 8 or 10. This would thus require iteration, and rebuilding the model several times using several number of clusters. This can be time consuming and resource-intensive.

Contrast this with hierarchical clustering

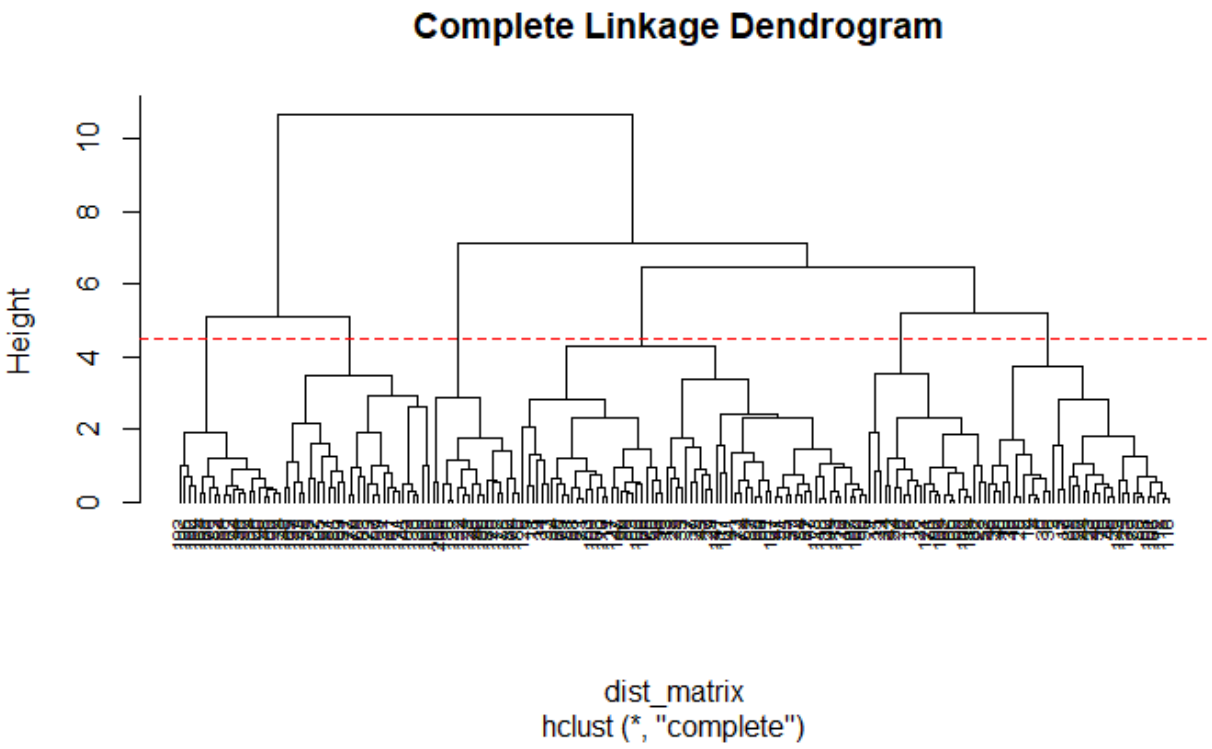

Figure 3: Dendrogram with Complete Linkage

While in this case, we chose to cut the tree at 4.5, we could also cut the tree at 4 or 3.8 resulting in a different number of clusters. However, in this case, we do not have to retrain the model in order to obtain a different number of clusters. We simply have to train the model once, and then cut the tree at different levels to examine different possible splits.

4.3 Comparison of Model Performance

Just like it is when selecting clustering parameters, evaluating the efficacy of the clusters selected is not a straightforward task.

One way of examining the efficiency of clusters is by using the within sum of squares and between sum of squares. Within sum of squares evaluates how similar the values within a single cluster are, while between sum of squares tells us how different each cluster is from the other. From all our models, we obtain the following figures for the within sum of squares and between sum of squares:

Model	Tot.Withinss	Tot.Betweenss
Kmeans	273.31	522.69
Complete Hierarchical	1076.34	489.03
Average Hierarchical	1080.45	489.03
Ward.D2 Hierarchical	1095.50	516.37

Table 4: Model Performance

It is interesting to note that while the clusters produced from complete and average clusters are exactly the same (see 4), complete clusters still have a slightly lower within sum of squares. This could be attributed to the different methods of identifying the linkages, thereby producing a different final figure.

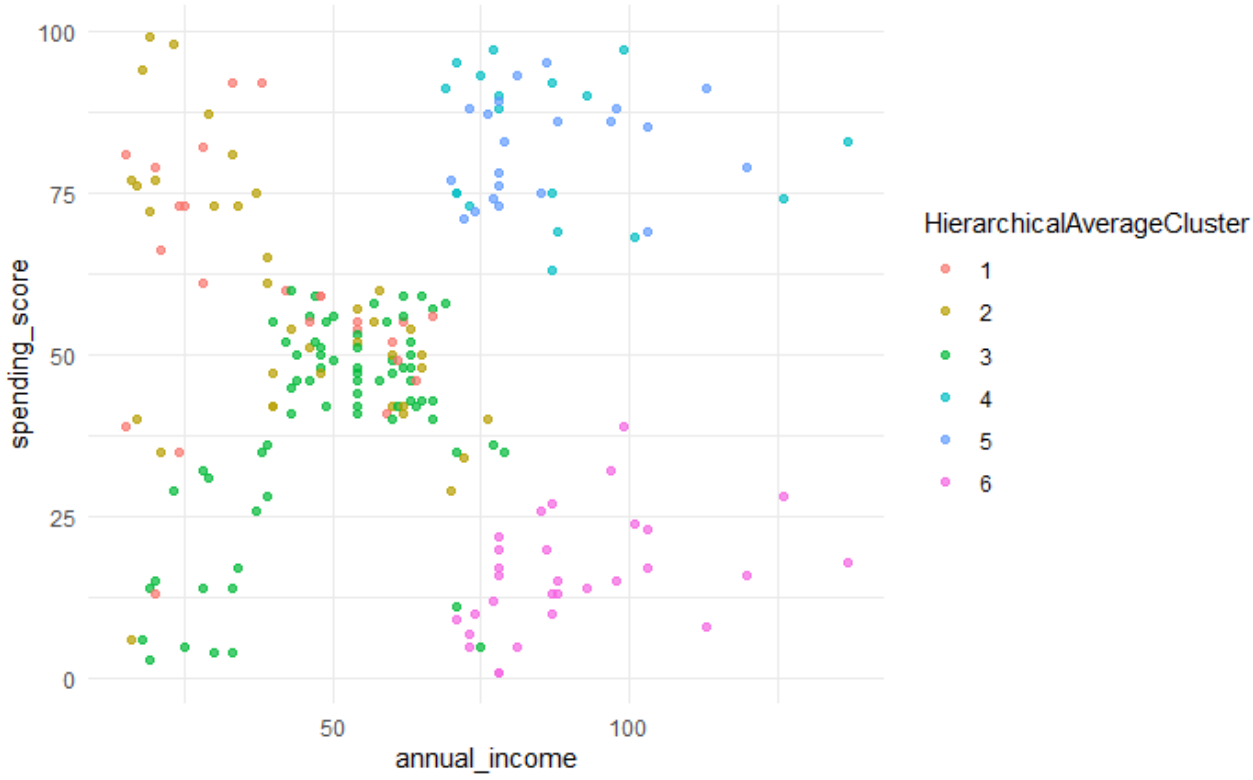

Figure 4: Average Clusters

Figure 5: Complete Clusters

Another interesting observation from Table 4 is the gulf between the within sum of squares from Kmeans and those from the hierarchical models. The cause of this difference is due to the fact that Kmeans directly optimizes for withinss when training the model. Hierarchical clustering on the other hand optimizes for proximity of branches

in the dendrogram. Thus the calculation of withinss is post-hoc and the centers aren't predefined. We thus find that it's not ideal to use this method for comparing performance across models. However, we can use the withinss to compare hierarchical models with different methods. This shows us that complete linkages provide the best clustering on this dataset compared to other linkage methods explored in terms of efficiency of clusters.

5 Conclusion

Unsupervised learning methods are generally less used than supervised learning methods. Some of the reasons for this have been highlighted in this analysis such as semi-scientific methods of choosing hyperparameters and the lack of a single standard for evaluating model performance. However, where ambiguity is high and labelled data isn't available, clustering can be used as a method of exploring and understanding the dataset. Clustering can also be useful when there is a ground truth to compare the results with.

Appendix

The step-by-step analysis, including R code is available in this Github repo

The dataset used can also be accessed here