Final Course Project: Mall Customer Segmentation

09.03.2021

by Tom K. Walter

1. Dataset and Goal

For my final project on unsupervised machine learning, I have chosen the Mall Customer Dataset.[1] This dataset contains 200 rows of observations that correspond to individual customers; and 5 columns containing information about the customers' age, gender, annual income, and spending score. A more detailed statistical breakdown of the data follows in the next section.

Table 1:

Name	Description	Type
CustomerID	index number assigned to each customer	useless for ML
Gender	gender given as Male or Female	categorical
Age	age in years	continuous
Annual Income (k$)	annual income in $1,000s	continuous
Spending Score	customer's spending score ranging from 1 to 100	continuous

The goal of unsupervised learning is to find hidden patterns such as clusters among unlabelled data. Hence, the goal of my project is to find "natural" groupings of customers among the mall customers (where no group label yet exists). This process is called "customer segmentation" and can help to identify homogeneous groups of customers from a larger set of heterogeneous customers. Rather than grouping customers just by a single attribute such as their age alone, clustering allows to group customers by all their given features. In turn, this can help business owners within the mall to answer key questions about their customer base, for instance:

- Who is the most valuable customer group, and what is their demographic make-up?
- What customer segments should new or existing stores target to be successful at this mall?
- Are there under-serviced customer groups, i.e. customers with high incomes but a low spending?

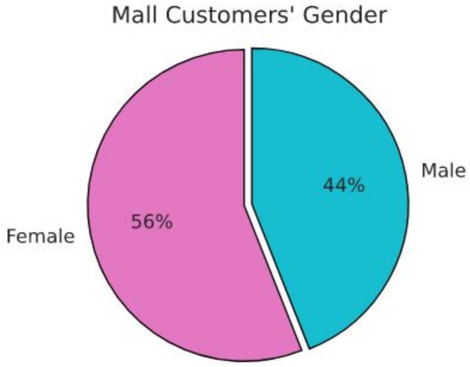
In the next steps of my project, I will explore the individual features of the dataset before applying 3 selected clustering algorithms. I will then compare the performance of these clustering algorithms to find the best performing model, and finally analyze the customer segments it uncovered.

[1] https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python.

2. Exploratory Data Analysis and Preprocessing

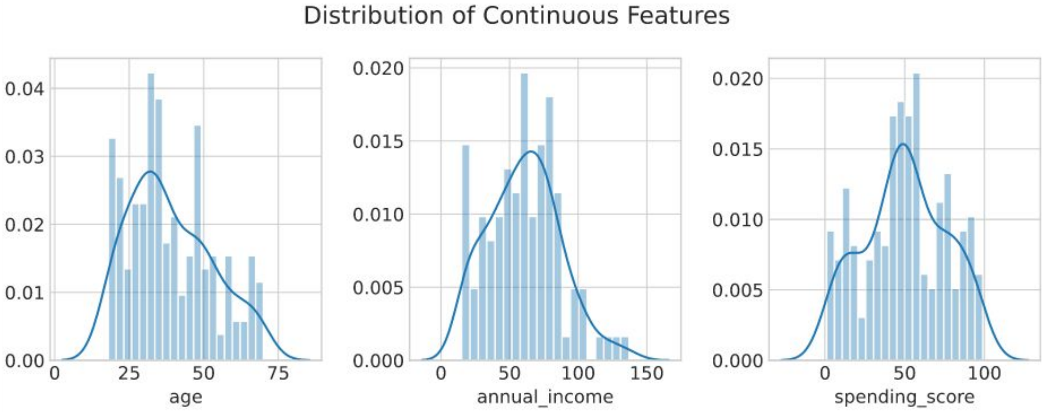
The dataset contains 1 categorical feature and 3 continuous features that can be used to perform clustering. In terms of gender, the mall customer dataset is relatively balanced with 112 women and 88 men (cf. Image 1).

Image 1:



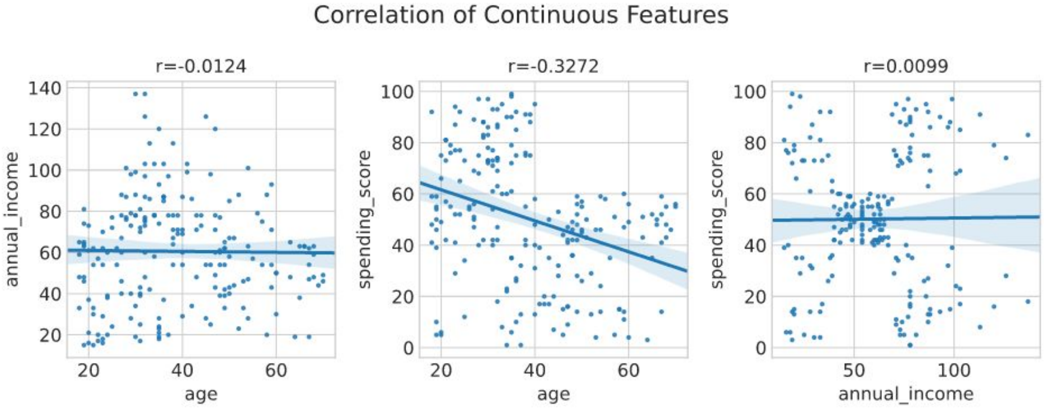
As Image 2 shows, all of the continuous features are relatively normally distributed. The age range of mall customers is between 18 and 70 years old with a mean age of 38.85 years. The mean income among mall customers is $61,500 per year, but the overall distribution ranges from $15,000 to $137,000. Lastly, the spending score is the most normally distributed feature, ranging from 1 point to 99 points with a mean of 50.2.

Image 2:



To gain a better understanding of the relationships among the continuous features, Image 3 depicts the correlation of the continuous features against each other. These correlation plots suggest that there is no strong correlation among any of the continuous features. Although normal distributed and uncorrelated features are considered ideal for supervised learning as it requires no further transformation, the absence of any relations may suggest no clusters. However, a closer look at the 3rd correlation plot that compares the annual income versus the spending score reveals that there are some natural groupings among the mall customers. In order to uncover these clusters or customer segments, the use of clustering algorithms becomes important.

Image 3:



Before clustering, the dataset must be quality checked and pre-processed. No duplicate or missing values have been found. The continuous features have been standard-scaled and the gender variable has been binary encoded (with 0 = female, 1 = male).

3. Clustering Models and Evaluation

To find the optimal amount of clusters, I have employed 3 different clustering algorithms that take a predetermined number of clusters *k* as input. This makes it easier to evaluate the performance of these different algorithms by initiating them with different k-values and then comparing it against 3 selected error metrics. Thus, the algorithms can not only be compared against themselves but also the model that has produced the most optimal clustering can be determined (see Table 2). The 3 clustering methods are K-Means, Hierarchical Agglomerative Clustering (HAC) with ward linkage, and HAC with average linkage.[2]

Moreover, the 3 error metrics that I have selected are inertia, distortion, and silhouette score. Inertia and distortion are similar measurements, as inertia measures the sum of squared distances of all points within a cluster to the cluster's mean, and distortion is simply the average of these distances. Using inertia or distortion, the best number for k is usually determined by the elbow method, which means a point in the curve where adding more clusters leads to diminishing improvements on accuracy. Since this elbow point is not always visually obvious in the graphs below, the silhouette score has also been plotted (cf. Image 4). The silhouette score measures how similar points are within their own cluster and dissimilar to other clusters. It ranges from -1 to 1, where a high value indicates that observations are well matched to their own cluster.

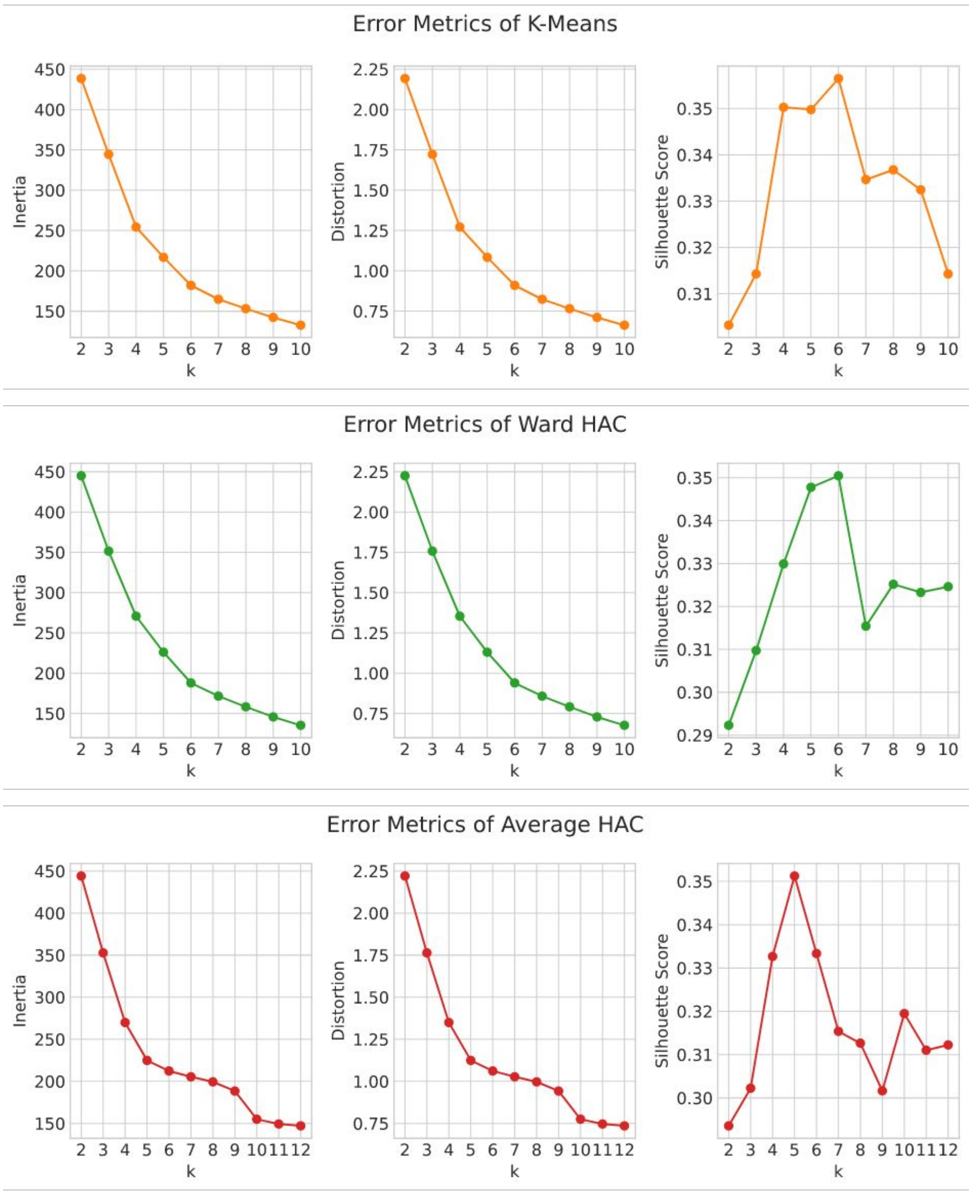
$$Inertia\ =\ \sum_{i=1}^{n}(x_i - C_k)^2\ ;\quad Distortion\ =\ \frac{1}{n}\sum_{i=1}^{n}(x_i - C_k)^2\ ;\quad Silhouette\ Score\ =\ \frac{(b-a)}{max(a,\, b)}\ .$$ [3]

The K-Means and the Ward HAC have been initiated for k ranging from 2 to 10, but for the Average HAC the range has been extended to 12, because the inertia and distortion take a dip at k=10. A cursory look at the all models' error metrics, on Image 4, surprisingly shows very similar performances in terms of inertia and distortion. Also the silhouette scores peak at either k=5 or k=6.

[2] https://en.wikipedia.org/wiki/K-means_clustering;
https://en.wikipedia.org/wiki/Hierarchical_clustering#Agglomerative_clustering_example.

[3] https://scikit-learn.org/stable/modules/clustering.html#k-means;
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.

Image 4:



In Table 2, I have picked the models with the best silhouette score from each clustering algorithm for closer evaluation. Only this closer comparison reveals that among them, the K-Means model performed best in terms of minimizing inertia and distortion, while maximizing the silhouette score. Thus, for the final clustering of the mall customer dataset and an analysis of the uncovered clusters, the K-Means model with k=6 will be chosen.

Table 2:

	K-Means	Ward HAC	Average HAC
k	6	6	5
Inertia	181.9514	187.9196	224.863
Distortion	0.9098	0.9396	1.1243
Silhouette	0.3565	0.3504	0.3512

4. Cluster Analysis: Customer Segments

This section serves to analyze the clusters produced by the best K-Means model and discuss what customer segments it has uncovered. After applying the clusters as labels to the mall customer dataset, I have summarized the insights in Table 3 and visualized them in Image 5.

Table 3:

Cluster	Mean Age	Mean Annual Income	Mean Spending Score	# of Women	# of Men	Total # of Customers
1	32.69	86.54	82.13	21	18	39
5	25	25.26	77.61	13	10	23
2	27	56.66	49.13	25	13	38
0	56.33	54.27	49.07	26	19	45
3	45.52	26.29	19.38	13	8	21
4	41.26	88.5	16.76	14	20	34

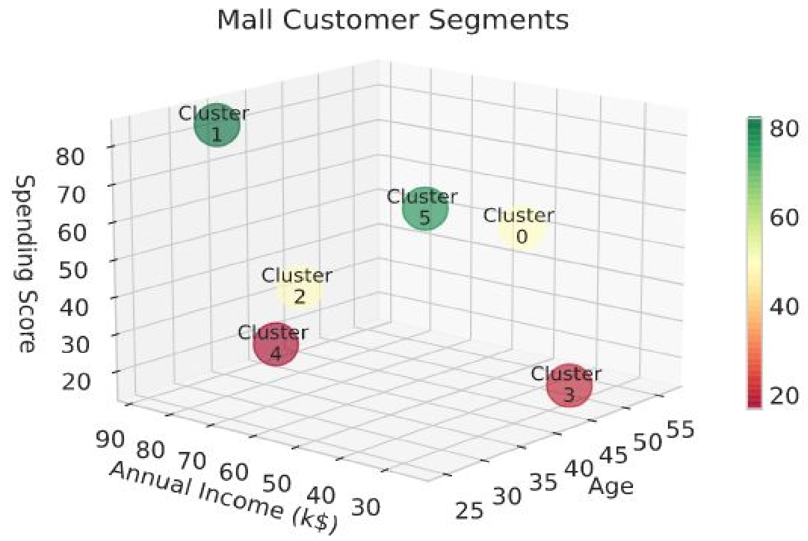
Table 3 is ranked by the mean spending score of each customer segment in descending order and displays the mean age, mean income, mean spending score, number of men, number of women, and total number of customers per each cluster. The color-scale in Image 5 also follows the idea of ranking customer segments by spending, where green stands for high spending segments, yellow for middle spending segments, and red for low spending segments.

Now, this can help to answer the key questions about the mall customer base raised in the beginning. For instance, customer segments can be ranked by how valuable they are to the mall and what characteristics they have:

- Cluster 1: high spending, high income, young-ish aged → highest value.
- Cluster 5: high spending, low income, young age → high value.
- Cluster 2: moderate spending, moderate income, young age → middle value.
- Cluster 0: moderate spending, moderate income, old-ish age → middle value.
- Cluster 3: low spending, low income, middle aged → low value.
- Cluster 4: low spending, high income, middle aged → lowest value.

For any new or existing store to be successful at this mall, it should probably target younger customers of all income groups. Given that there are 3 customer segments that are young on average but belong to 3 different income classes, mall stores also do not need to compete for the same segment but could differentiate themselves along these income lines. Another important question was whether there is an under-serviced customer group. Indeed clustering found that currently the lowest spending customer segment conversely has the second highest income on average. Moreover, this customer segment (Cluster 4) stands out as it is the only group with more men than women. This suggests high income men are under-serviced at this mall and there may be opportunities for stores to increase their revenue by targeting this segment.

Image 5:



5. Conclusion

Overall, this exercise has revealed that spending behavior is not tied to the gender or the income of customers alone but that customers can be segmented into different groups. Although exploratory data analysis revealed that there were no strong (cor-)relations among the features in general, the use of 3 different clustering algorithms has proven that there were natural groupings among the mall customers. Moreover, the highly similar performance of these algorithms reinforces the idea that natural segments existed but these patterns were hidden from exploratory methods. However, it needs to be added that the dataset was relatively small, both in terms of observations and number of features. To gain yet a deeper understanding of mall customers' shopping and spending behavior, future surveys should ask for further information, e.g.:

- at which type of stores customers shop and how much money they spend there;
- demographic data such as marital status and race (if this is not considered insensitive);
- customer satisfaction with the mall, its individual stores and services.

All these additional features could lead to more precise customer segmentation from clustering algorithms and provide more actionable insights for mall management and store owners. Although data beyond 3 dimensions cannot be visualised in the same way as given in Image 5, more advanced clustering algorithms such as DBSCAN could still manage to find meaningful clusters in high dimensional data (if K-Means and HAC would fail).