Unsupervised Learning: Mall Customer Segmentation

Zixin Huang

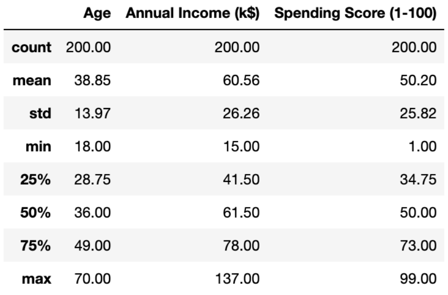
July 16, 2021

1 Objective and Summary of Data

The goal of this project is to help the owner of a mall to identify different groups of customers, given some basic information of each customer. We use several different unsupervised learning models to recognize clusters within customers. Our main purpose, therefore, is clustering, though we also apply method of dimensionality reduction for visualization.

The data set[1] consists of 200 observations, with 5 columns representing customer ID, gender, age, annual income, and spending score, which range from 1 to 100. There is no missing value. However, we may encounter difficulties since the data set is very small, and features of customers are limited. The summary statistics of numerical features is shown in Figure 1.

Figure 1: Summary Statistics



	Age	Annual Income (k$)	Spending Score (1-100)
count	200.00	200.00	200.00
mean	38.85	60.56	50.20
std	13.97	26.26	25.82
min	18.00	15.00	1.00
25%	28.75	41.50	34.75
50%	36.00	61.50	50.00
75%	49.00	78.00	73.00
max	70.00	137.00	99.00

2 EDA and Feature Engineering

2.1 Gender

The distribution of gender is this data set is almost balanced. As we can see from Figure 2, there are slightly more female than male. To be more precise, 56% of customers are female and 44% of customers are male. Table 1 shows that, on average, female customers are slightly younger, tend to have lower income, but higher spending than male customers.

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

Figure 2: Gender Distribution

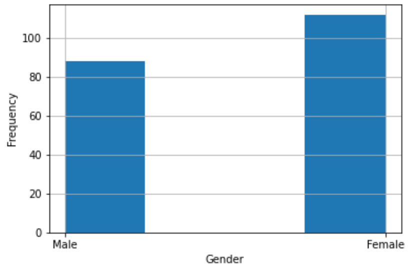


Table 1: Average Attributes of each Gender Group

Gender	Age	Annual Income (k$)	Spending Score (1-100)
Female	38.10	59.25	51.53
Male	39.81	62.23	48.51

2.2 Age

The distribution of customer age is shown in Figure 3. Although the range of age of customers are from 18 to 70, most of customers are around 20 and 30, less than 50% of the customers are older than 45.

The relationship between age, income, and spending can be seen in Figure 4. We can see that customers who have highest income are around age 30 and customers who have the lowest income are around age 18. Income level among customers younger than 40 are more diverse than income level among customers older than 40. The spending score are, in general, higher among customers younger than 40 compared to the spending among customers older than 40.

Figure 3: Age Distribution

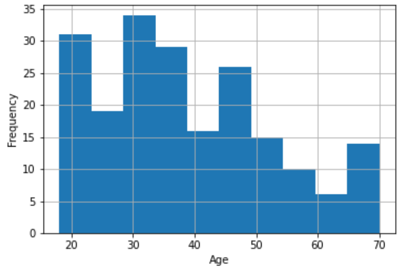
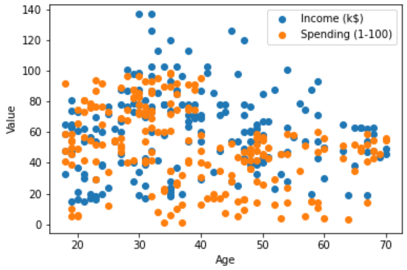


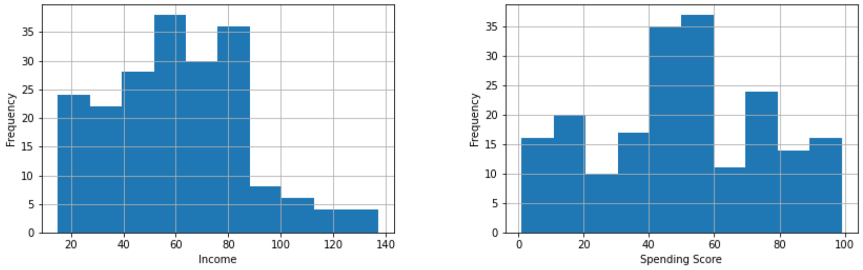
Figure 4: Age, Income, and Spending



2.3 Income and Spending

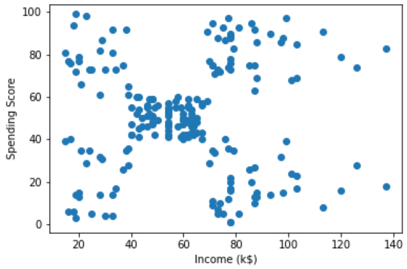
The distribution of income and spending score can be seen in Figure 5. We can see that most of the customers have annual income around \$60k, many have income between \$15k and \$55k, and few have income more than \$90k.
The distribution of spending score seems more symmetric. The score of most of customers are around 50, and around 7.5% of customers have the lowest and highest spending score respectively.

Figure 5: Distributions of Income and Spending



The relationship between income and spending is shown is Figure 6. We can see there are 5 groups of customer in terms of these two dimensions. For customers who have medium income, which range from \$40k to \$70k, their spending score concentrate between 40 and 60. For customers who have higher income, nearly half of them have spending score higher than 60 and the other half of them have spending score lower than 40. The spending among customers have lower income is similar to that among customers have higher income.

Figure 6: Relationship between Income and Spending



2.4 Prepare Data

To prepare the data for training, we first binarize gender column, using 0 to represent female customers and 1 to represent male customers. Then we scale all the columns into range 0 and 1 to avoid the imbalanced influence of magnitudes on our models.

3 Models

3.1 K Means

We first use K Means model with parameter k ranging from 3 to 11, and plot the inertia related to each different k (see Figure 7). We cannot clearly observe the optimal k, so we use TSNE plot to visualize the clusters for k equals to 4, 5, and 6. The visualization is shown in Figure 8.

Figure 7: Inertia for each K

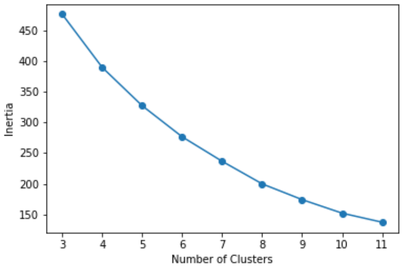
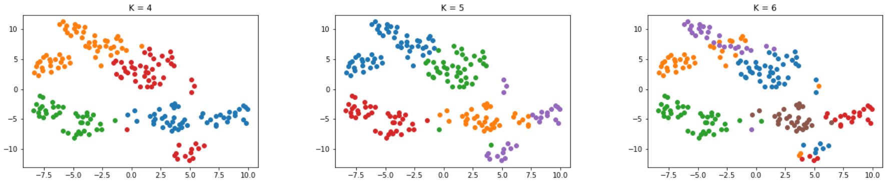


Figure 8: TSNE Plot for K Means Clusters



3.2 DBSCAN (OPTICS)

The second model we use is OPTICS, which is a version of DBSCAN but it trains and allows for different ε, thus performs better on data with different density. We try the hyperparameter from 4 to 14, and plot the number of clusters related to the number of sample in neighbourhood (N) in Figure 9.

Since we do not want to have more than 8 groups, we visualize clusters with N equals to 10, 11, and 12 in Figure 10. We can see that, unlike K Means, OPTICS leaves some of the customers as outliers in each result, and the groups appear to be more tight compared with the groups we have using K Means. In addition, as the number of neighbors (N) increases, more customers are labeled as outliers.

Figure 9: Number of Clusters for each N

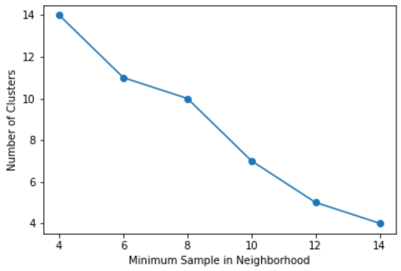
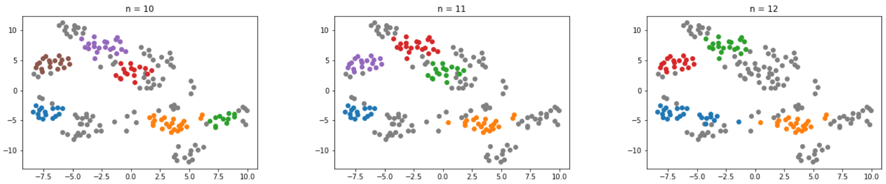


Figure 10: TSNE Plot for OPTICS Clusters


n = 10

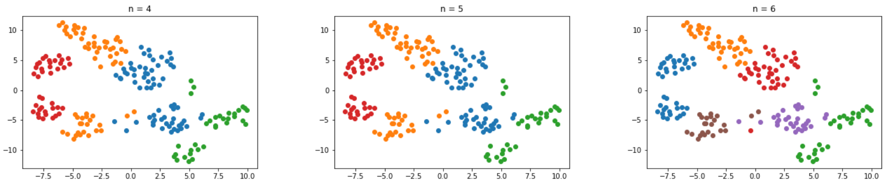
n = 11

n = 12

3.3 Agglomerative Clustering

Finally, we use Agglomerative Clustering with ward linkage, and number of clusters equals to 4, 5, and 6. The visualization is shown in Figure 11.

Figure 11: TSNE Plot for Agglomerative Clusters


n = 4

n = 5

n = 6

4 Compare Results

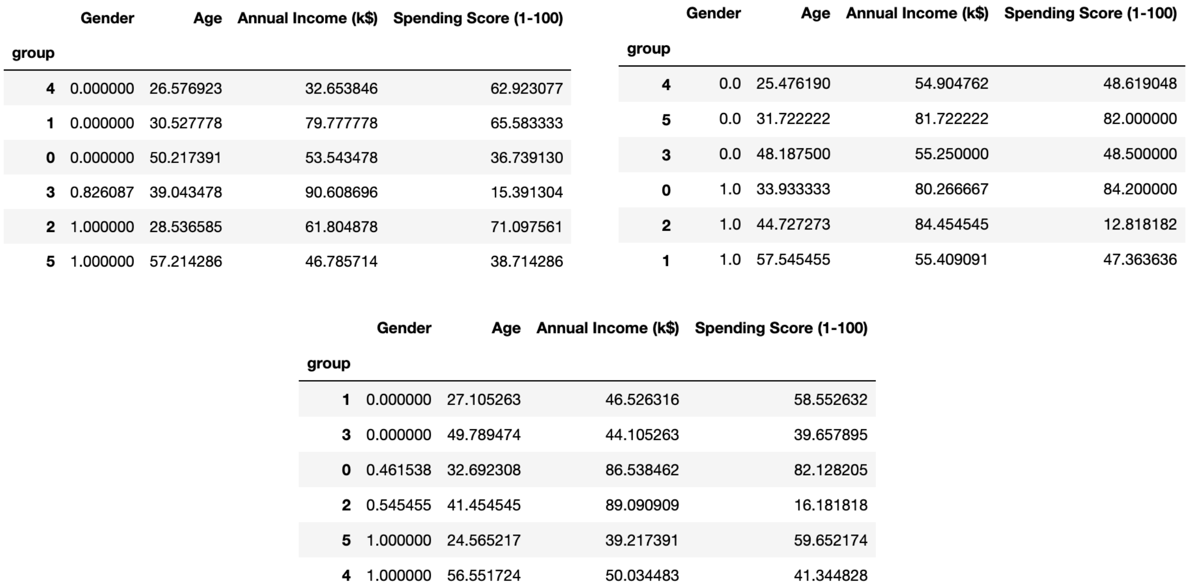
To compare the resulting clusters we select the result from each algorithm with 6 clusters. The average attributes for each group of each classification can be seen in Figure 12.

4.1 K Means

For K Means, there are 3 male groups, 2 female groups, and 1 group consists mainly male. The younger male group (group 4), with average age of 26, has relatively low income but high spending. The middle-aged female group (group 1) has relatively high income and high spending. The older female group (group 0) has medium income and lower spending.
The customers in group that consists of mainly males (group 3) are on average middle aged and have highest income, but lowest spending.

The younger male group (group 2) has relatively high income and high spending, whereas the older male group (group 5) has relatively low income and lower spending.

Figure 12: Clusters using K Means, OPTICS, and Agglomerative Clustering



group	Gender	Age	Annual Income (k$)	Spending Score (1-100)
4	0.000000	26.576923	32.653846	62.923077
1	0.000000	30.527778	79.777778	65.583333
0	0.000000	50.217391	53.543478	36.739130
3	0.826087	39.043478	90.608696	15.391304
2	1.000000	28.536585	61.804878	71.097561
5	1.000000	57.214286	46.785714	38.714286

group	Gender	Age	Annual Income (k$)	Spending Score (1-100)
4	0.0	25.476190	54.904762	48.619048
5	0.0	31.722222	81.722222	82.000000
3	0.0	48.187500	55.250000	48.500000
0	1.0	33.933333	80.266667	84.200000
2	1.0	44.727273	84.454545	12.818182
1	1.0	57.545455	55.409091	47.363636

group	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	0.000000	27.105263	46.526316	58.552632
3	0.000000	49.789474	44.105263	39.657895
0	0.461538	32.692308	86.538462	82.128205
2	0.545455	41.454545	89.090909	16.181818
5	1.000000	24.565217	39.217391	59.652174
4	1.000000	56.551724	50.034483	41.344828

4.2 OPTICS

The OPTICS generates 3 male groups and 3 female groups. The younger female group (group 4) has medium income and medium spending. The middle-aged female group (group5) has relatively high income and high spending. And the slightly older female group (group 3) has medium income and medium spending.
The middle-aged male group (group 0) with high income has the highest spending. The slightly older male group (group 2) with the highest income has the lowest spending. And the senior male group (group 1) has medium income and medium spending.

4.3 Agglomerative Clustering

This algorithm produces 2 male groups, 2 female groups, and 2 groups with about half of male and female customers. The younger female group (group 1) has medium income and medium spending. The middle-aged female group (group 3) has medium income but lower spending.
One of the mixed group (group 0) is slightly younger, with high income and high spending. Another mixed group (group 2) are older, with high income but low spending.
The younger male group (group 5) has relatively low income but medium spending. And the older male group (group 4) has medium income and spending.

5 Findings and Insights

If the marketing team of this mall wants to reach out to customers and send advertisements, their targeting customers should be young or middle aged, with high income as well as spending score.

This is because younger people may be more easily to be attracted to new things and high income and spending score means they have potential and willingness to buy.
For example, we might want to target customers in group 1 and group 2 using K Means, or customers in group 5 and group 0 using OPTICS, or customers in group 0 using agglomerative clustering.

6 Next Steps

For the next steps in analysis we may want to acquire more observations and more attributes to improve our model. And if we can have more features we may need to use dimensionality reduction tools to avoid the curse of dimensionality. Then we can use the reduced data to fit our previous models and come up with a better result.