Table of Contents

Summary

In this customer dataset, there are 200 instances with five features. First, I show the basic data structure and view statistical analysis of single attribute, 2-dimension attributes, and 3-dimension attributes. It is shown that there are some insights of the company's customers features and lay groundwork for the model building. Age has clear relations with SpendingScore. SpendingScore has a strong relationship with Income.From the statistical view, this company has a relatively stable customer base, and the overall development trend is stable. For future development, it is necessary to formulate different marketing and operation strategies for different groups of people.

In this report, I mainly use the K-means Cluster algorithm for customer segmentation. In this report, I build three models(based on Age-SpendingScore, based on Income-SpendingScore, based on Age-Income-SpendingScore). All three models have quite good Silhouette scores. But in different segmentation methods can be used for different purposes

First, use Age and SpendingScore to build a model to divide customers into 4 groups, which mainly focus on the Customers' Age and Purchasing Power. It is better to choose this segmentation when the company wants to launch and market some new products.

Second, use Income and SpendingScore to build a model to divide customers into 5 groups, which mainly focus on the Customers' Income level and Purchasing Power. It is better to choose this segmentation when the company intends to increase its GMV or Orders performance. This segmentation model also has the highest Silhouette score.

Third, use Age, Income and SpendingScore to build a model to divide customers into 6 groups, which mainly focus on the customers' comprehensive profile. It is better to choose this segmentation when the company wants to build detailed customer profiles to make better user management.

Business Problem Statement

A primary goal for this e-commerce company is to understand their customers. To explore their consumers' situations. The problem is to define the e-commerce company's consumers. In this Dataset, we have some basic customer data and we need to explore more business insights of the customers' features and to define the customer segmentation strategy to provide related marketing or management recommendations.

Usually in E-commerce companies, the key performance indicators are GMV, Orders, Active User Numbers. To make company performance better, it is important to focus on the customer segmentation and use different ways to stimulate customers to become active users and increase the GMV and Orders performance. For example, a company could launch some vouchers to attract those low income customers to purchase. In business view, the features of customers like Gender and Age are also important to analyze, for there are different types of products to recommend to different customers.

Except for the indicators listed in this dataset, the customer's behavior data(Page views, Add to cart, Purchasing Conversion Rate, Order Completion rate, Complaint Rate, etc.) can also be analyzed to contribute to customer segmentation. Common management and marketing method for customer segmentation are personalized recommendation, direct marketing(email, text, phone selling, etc)

All in all, customer segmentation is the most important part in ecommerce company, managing users not only to increase the daily KPI of a company, but also important for the company's future strategy management.

Exploratory Data Analysis

Dataset Dimensions and Summary of the Dataset

- Import pd.read_csv of the customer dataset
- Use "shape","head","info" method to check the basic information of Customer Dataset:

    - There are 200 instances and 5 attributes, 4 numerical variables and 1 character variables
    - Gender is object

```
customer.info() # use the info() method check the da
```

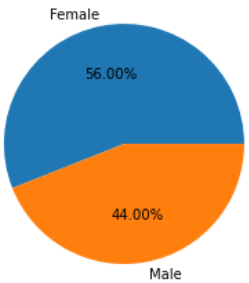
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   CustomerID      200 non-null    int64 
 1   Gender          200 non-null    object
 2   Age             200 non-null    int64 
 3   Income          200 non-null    int64 
 4   SpendingScore   200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

- Use describe() method
    - As for the statistical data values, their value ranges vary too much, need to standardize or normalize.
        - CustomerID: There are 200 customers
        - Age: Age of customers are ranging from 16 to 74
        - Income: Income ranges from 93 to 896
        - SpendingScore: SpendingScore ranges from 1 to 99
- Check there is no missing value and duplicated value

Attributes Analysis and Insights

Categorical Attribute `Gender`

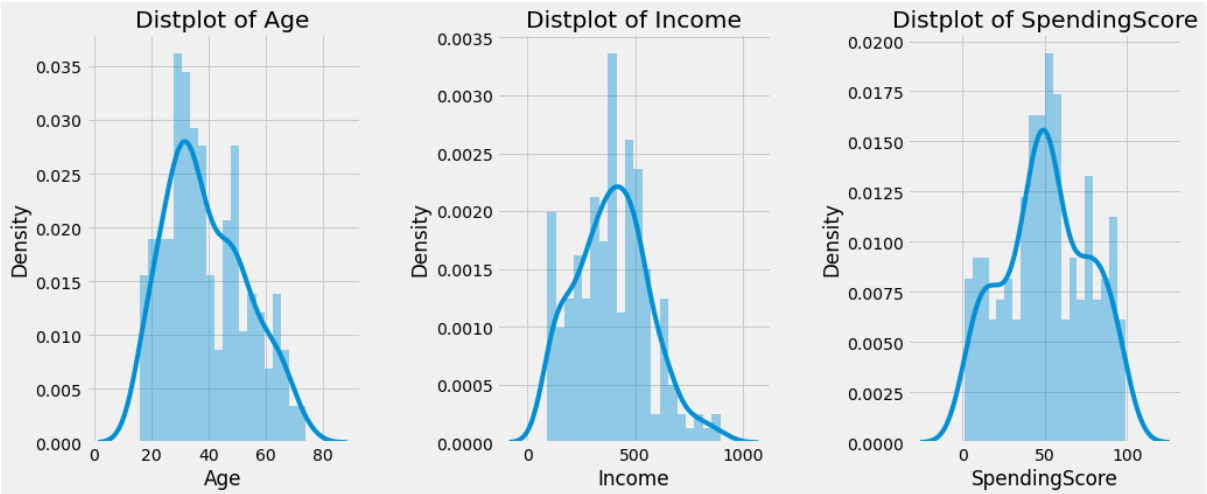
- Use Pie Chart to show that 'Female' accounts for 56%, 'Male' accounts for 44%
- In this e-commerce company, there are more females shopping than males



Single Attribute Distribution Insights(Age, Income,SpendingScore)

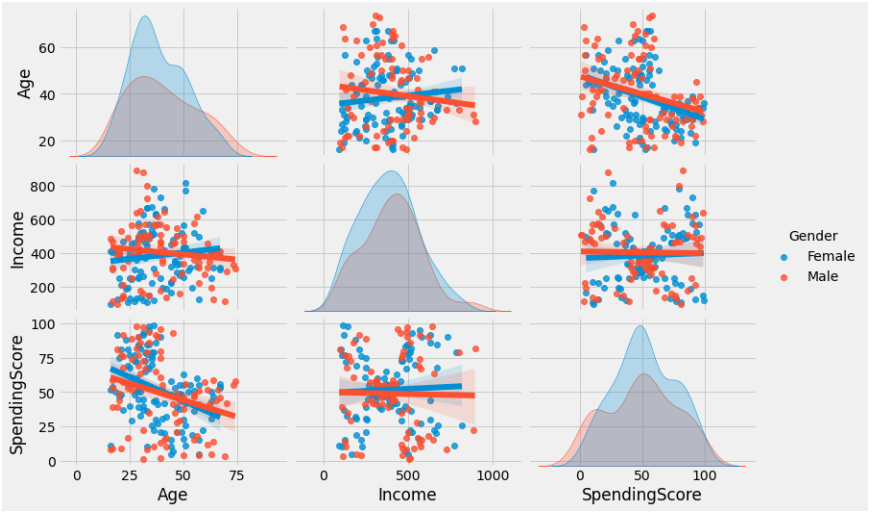
- The three attributes' shapes are similar to the normal distribution, indicating that the amount of data is sufficient and the distribution of data sampling is also relatively ideal

- **Age:** Customers' Age in the [30,40] range are the most, and there are also many people in the [20,30] range, but the elderly over 60 are the least frequent consumers, and the distribution is irregular in the [40,60] range. It is speculated that there are other factors that affect the consumption of people in this age group
- **Income:** Most of the customer income is concentrated in the range of [400,550]
- **SpendingScore**: Spending Score is concentrated between [40,55]



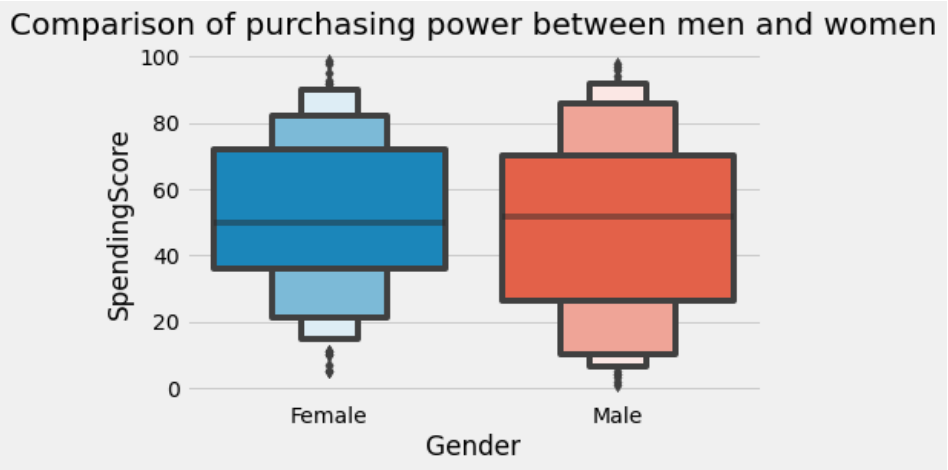
Two-Dimension Attributes Analysis

- Use pairplot to show the each 2 attributes relationships
- **Age** has negatively relations with **SpendingScore**
- **SpendingScore** has strong relationship with **Income**
- In marketing view, the age is the target group for the different marketing intentions



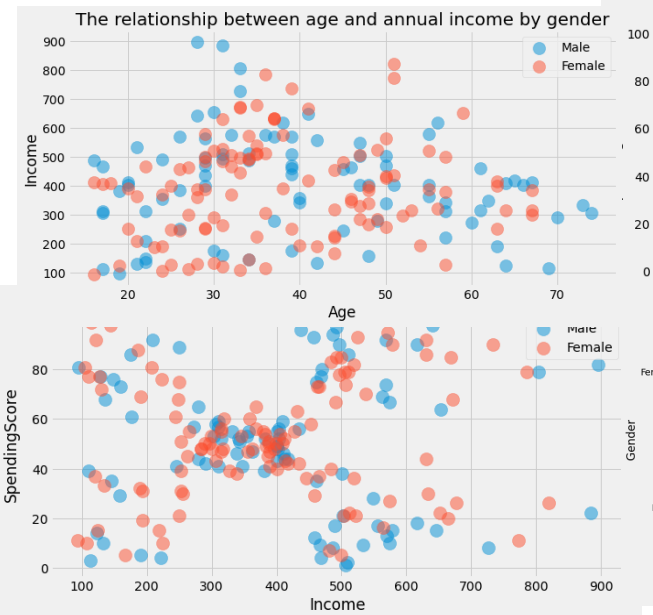
Comparison of purchasing power between men and women

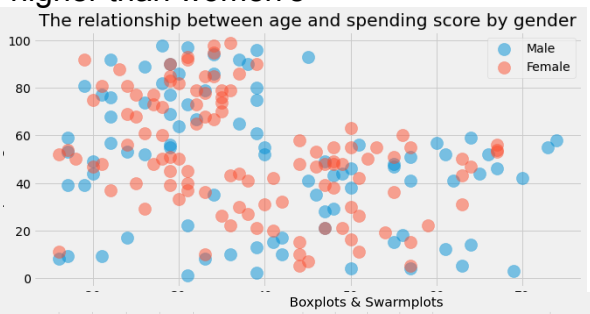
- Man's Spending Scores are concentrated in [25, 70], while women's Spending Scores are concentrated in [37, 75], which to some extent shows that women perform better than men in purchasing power

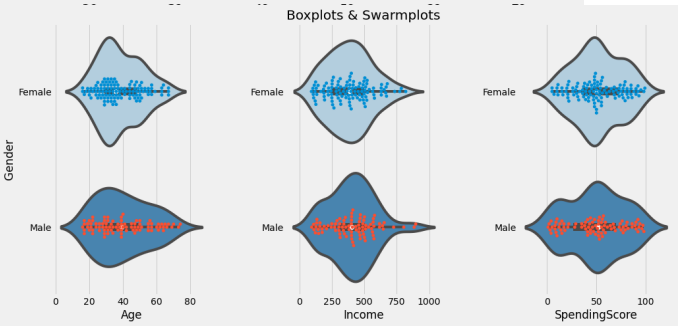


Three-Dimension Attributes Analysis: Check the relationship with different gender

- In terms of age: the distribution of men is relatively even, and there are more in their 20s; the age of women is mostly concentrated in the range of 20+~30+, and they are generally younger
- In terms of Income: Man's income is higher than women's





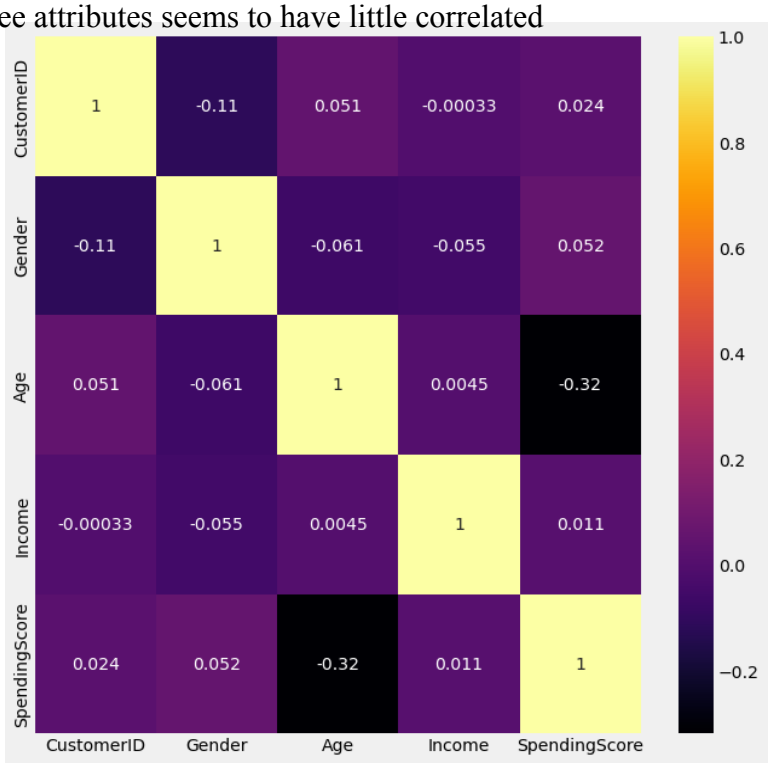


Build the Model

- In this case, I use the K-means Cluster algorithm to build the unsupervised machine learning model
- K-Means clustering is an efficient machine learning algorithm to solve data clustering problems. It's an unsupervised algorithm that's quite suitable for solving customer segmentation problems.
- The reason why using K-means clustering for customer segmentation
    - K-means algorithm discovers groups (clusters) in the data, where the number of clusters is represented by the K value. The algorithm acts iteratively to assign each input data to one of K clusters, as per the features provided. All of this makes k-means quite suitable for the customer segmentation problem.
    - Given a set of data points are grouped as per feature similarity. The output of the K-means clustering algorithm is: The centroids values for K clusters, Labels for each input data point.
- Before using the K-means to cluster, it needs to scale the variables and to look at a scatter plot or data table to estimate the number of cluster centers to set for the k parameter in the model

Preprocess the Categorical Data

- To convert the categorical variable "Gender" into a numerical variable by using map method{'Male':0,'Female':1}
- After converting all the attributes into the numericals, use the heatmap to check the correlation among all attributes
- The Result shows that `Age` is negatively correlated with `SpendingScore`, corr score is -0.32
- Other three attributes seems to have little correlated



Scale the Data Before Building the Model

- Use the StandardScaler to Process all the attributes before feeding the model

```
from sklearn.preprocessing import StandardScaler
# Scale the numerical columns
scaler = StandardScaler()
df[['Age', 'Income', 'SpendingScore']] = scaler.fit_transform(df[['Age', 'Income', 'SpendingScore']])
```

To build the First Clustering Model Based on Age and SpendingScore

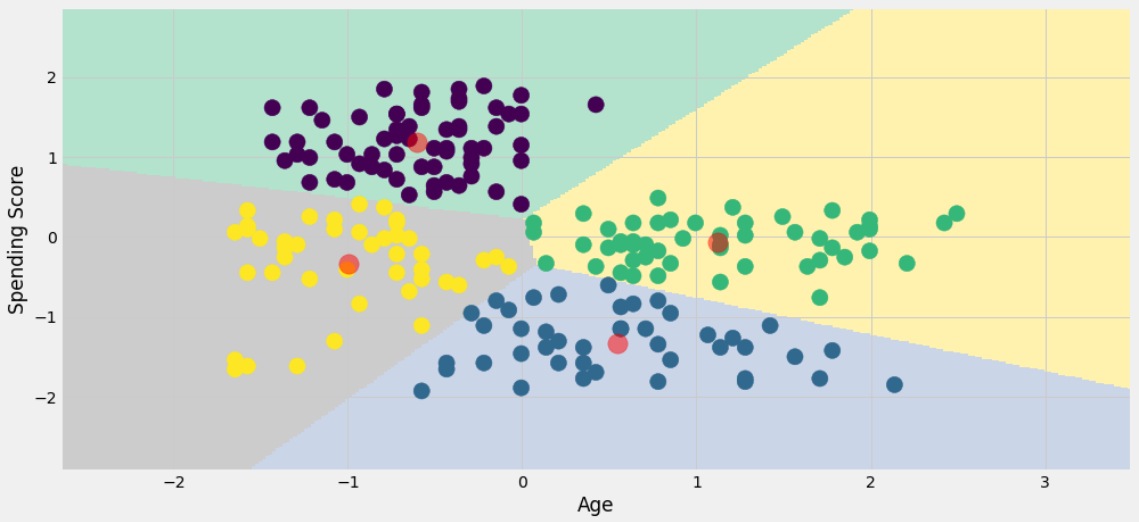
- Based on the Attributes Analysis Insights before, we can see the linear relationship between `Age` and `SpendingScore`
- First to use the two Attributes to build the model

```
'''Age and spending Score'''
df1 = df[['Age' , 'SpendingScore']].iloc[: , :].values
inertia = []
for n in range(1 , 11):
    km1 =(KMeans(n_clusters = n , # define the cluster numbers
                 init='k-means++', #inital centroid
                 n_init = 10 , #default is 10
                 max_iter=300, #max iteration times defined to be 300
                 tol=0.0001, #Tolerance minimum error
                 random_state= 42 , #define the random_state to 42
                 algorithm='elkan'#Using elkan K-Means algorithm
            ) )
    km1.fit(df1)
    inertia.append(km1.inertia_)
```

```
plt.figure(1 , figsize = (15 ,6))
plt.plot(np.arange(1 , 11) , inertia , 'o')
plt.plot(np.arange(1 , 11) , inertia , '-' , alpha = 0.5)
plt.xlabel('Number of Clusters') , plt.ylabel('Inertia')
plt.show()
```

- Through Inertia numbers to define the n_clusters number is 4

- Use the scatter figure to show the clusters and distribution of the first model



- Based on `Age` and `SpendingScore`, Customer can be divided into four customer segmentation
- The Silhouette score km1_best model(Based on SpendingScore and Age, n_clusters=4) is 0.45
- The effect of km1_best_model is good

```
score1 = metrics.silhouette_score(df1,labels1)
print('Four Customer Segmentation based on SpendingScore and Age,Silhouette_score is:' , score1)
```

Business Insights on the first model

cluster_1	CustomerID	Age	Income	SpendingScore
0	101.666667	30.634921	415.984127	80.873016
1	98.976744	46.720930	429.348837	16.418605
2	102.058824	54.803922	357.745098	48.843137
3	98.465116	25.139535	367.720930	41.930233

- Based on the Age and SpendingScore two attributes, customers can be divided into four segments
- Group 0 is Young People but with High Spending Score
- Group 1 is Middle Aged People with Low Spending Score
- Group 2 is Old People with medium Spending Score
- Group 3 is Young People with Low Spending Score
- For Group 0 Young people with High Spending Score, they are compulsory customers, companies should use some new and trendy products to attract them. For Group 3 Young People with Low Spending Score, company should

use more discount methods to stimulate purchasing orders. For Group 1&2, they are middle contribution customers, it is important to maintain them.

Second Clustering Model Based on Income and SpendingScore

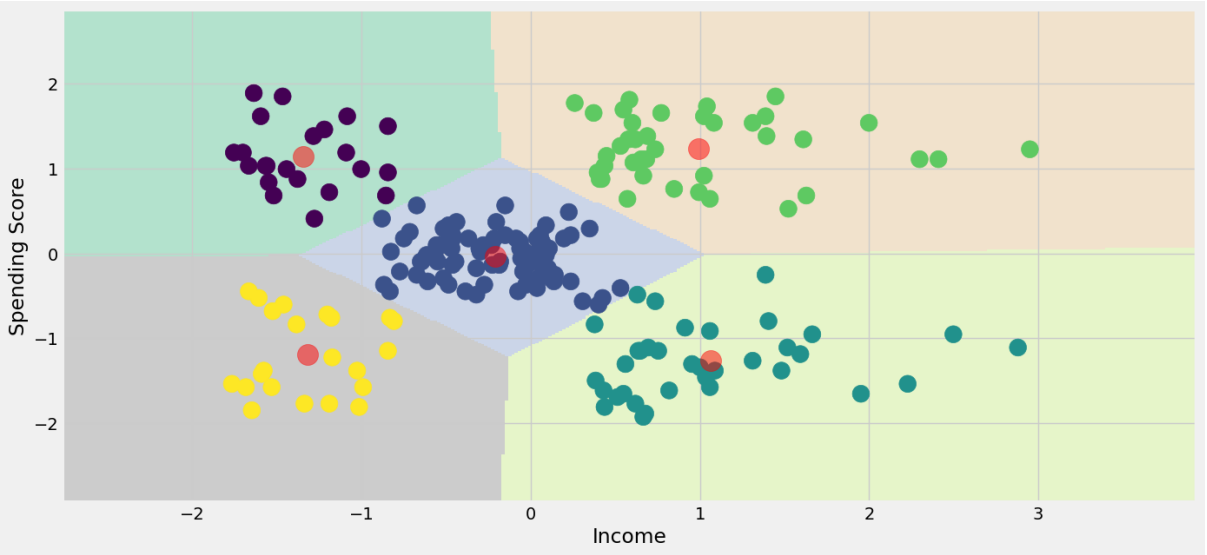
- Based on the Attributes Analysis Insights before, we can see the distribution gap line between `Income` and `SpendingScore`
- Secondly to use the two Attributes to build the model

```
'''Income and spending Score'''
df2 = df[['Income' , 'SpendingScore']].iloc[: , :].values
inertia_2 = []
for n in range(1 , 11):
    km2 =(KMeans(n_clusters = n , # define the cluster numbers
                 init='k-means++', #inital centroid
                 n_init = 10 , #default is 10
                 max_iter=300, #max iteration times defined to be 300
                 tol=0.0001, #Tolerance minimum error
                 random_state= 42 , #define the random_state to 42
                 algorithm='elkan'#Using elkan K-Means algorithm
           ) )
    km2.fit(df2)
    inertia_2.append(km2.inertia_)
```

```
plt.figure(1 , figsize = (15 ,6))
plt.plot(np.arange(1 , 11) , inertia_2 , 'o')
plt.plot(np.arange(1 , 11) , inertia_2 , '-' , alpha = 0.5)
plt.xlabel('Number of Clusters') , plt.ylabel('Inertia')
plt.show()
```

- Through Inertia numbers to define the n_clusters number is 5

- Based on `Income` and `SpendingScore`, Customer can be divided into five customer segmentation
- The Silhouette score km2_best model(Based on `SpendingScore` and `Income`, n_clusters=5) is 0.55
- The effect of km2 best model is good and better than the model1



Business Insights on the second model

cluster_2	CustomerID	Age	Income	SpendingScore	(
0	99.272727	26.045455	165.136364	79.818182	
1	99.560976	42.963415	357.853659	49.524390	
2	98.885714	41.428571	574.428571	18.000000	
3	103.820513	33.025641	562.230769	82.282051	
4	101.909091	44.590909	168.954545	19.954545	

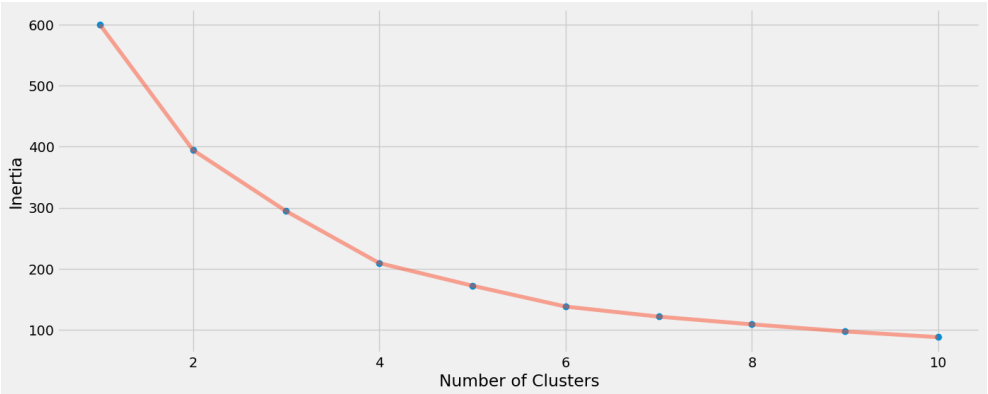
- Based on the Income and SpendingScore two attributes, customers can be divided into five segments
- Group 0 is Low Income but with High Spending Score
- Group 1 is Middle Income with Middle Spending Score
- Group 2 is High Income with Low Spending Score
- Group 3 is High Income with High Spending Score
- Group 4 is Low Income and with Low Spending Score
- All High Spending Score customers, companies should launch some marketing method with more high price products to attract people to buy and therefore to increase the company's GMV. In terms of Income features of customers, the most important considerations is the company’s ATV(Average Transaction Value)

Third Clustering Model Based on Income and SpendingScore and Age

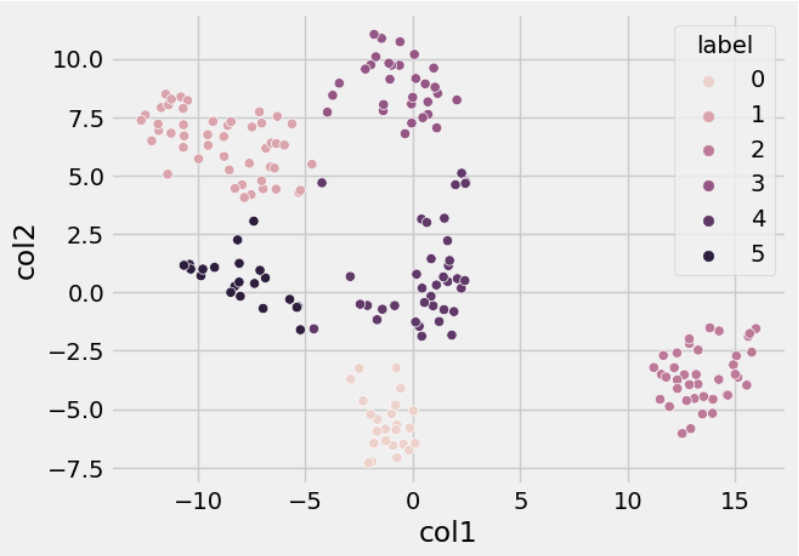
```
# select the three attributes to feed the model

'''Income and spending Score'''
df3=df[['Age','Income','SpendingScore']].iloc[: , :].values
inertia_3 = []
for n in range(1 , 11):
    km3 =(KMeans(n_clusters = n , # define the cluster numbers
                 init='k-means++', #inital centroid
                 n_init = 10 , #default is 10
                 max_iter=300, #max iteration times defined to be 300
                 tol=0.0001, #Tolerance minimum error
                 random_state= 42 , #define the random_state to 42
                 algorithm='elkan'#Using elkan K-Means algorithm
            ) )
    km3.fit(df3)
    inertia_3.append(km3.inertia_)
```

```
plt.figure(1 , figsize = (15 ,6))
plt.plot(np.arange(1 , 11) , inertia_3 , 'o')
plt.plot(np.arange(1 , 11) , inertia_3 , '-' , alpha = 0.5)
plt.xlabel('Number of Clusters') , plt.ylabel('Inertia')
plt.show()
```

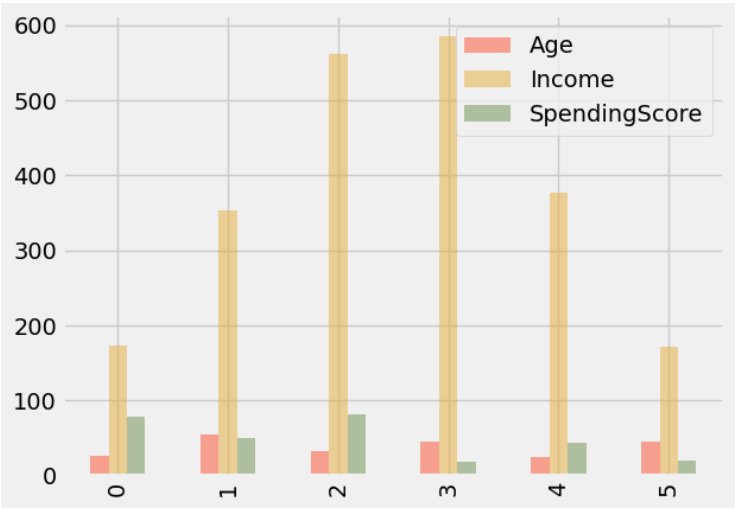


- Through Inertia numbers to define the n_clusters number is 6
- To make clustering more directly seen, I use the manifold to reduce the dimension, clustering is shown below.



- Based on `Income` and `SpendingScore` and `Age`, Customer can be divided into six customer segmentation
- The Silhouette score km3_best model(Based on `SpendingScore` and `Income` and `Age`, n_clusters=6) is 0.43
- Through dimensionality reduction, the three-dimensional stereogram is reduced to a plane diagram. The 6 different colors in the above picture represent 6 different groups of people. Because K-Means is unsupervised learning, it is mainly responsible for classifying users with obvious characteristics into one category. Specifically, each We need to analyze what groups each class represents, which will be described below.

Business Insights on the third model



cluster	CustomerID	Age	Income	SpendingScore
0	100.291667	26.500000	173.166667	78.416667
1	101.437500	55.104167	353.729167	49.500000
2	103.820513	33.025641	562.230769	82.282051
3	104.774194	44.903226	584.935484	19.129032
4	95.459459	24.540541	376.486486	43.918919
5	95.000000	45.047619	171.476190	19.523810

- Group 0 is Impulse Consumers: Average age is 27 years old, Income 173, SpendingScore is 78. This is mainly young people with average income but high consumption score
- Group 1 is Important to Keep Consumers: Average age is 55 years old, Income 354, SpendingScore is 50. This is mainly middle-aged and elderly people with middle income and consumption scores

- Group 2 is Important Value Consumers: Average age is 33 years old, Income 562, SpendingScore is 82. This is mainly middle-aged people, with high income and consumption scores, belonging to the optimal customer group.
- Group 3 is Prudent Consumers: Average age is 45 years old, Income 585, SpendingScore is 19. This is mainly middle-aged people with high income but low consumption score.
- Group 4 is Important Development Consumers: Average age is 25 years old, Income 376, SpendingScore is 44. This is mainly young people, with middle income and consumption scores, they have great potential.
- Group 5 is Average Value Consumers : Average age is 45 years old, Income 171, SpendingScore is 20. This is mainly middle-aged people, but with low income and consumption scores. They are the least valuable customers

```
customer['cluster'].value_counts(1)
```

```
Maintenance Consumers    0.240
Value Consumers          0.195
Development Consumers    0.185
Prudent Consumers        0.155
Impulse Consumers        0.120
Average Consumers        0.105
Name: cluster, dtype: float64
```

- Maintenance customer is the biggest customer segmentation, which means the company has the most important to keep the customers.
- Average customer is the least customer segmentation, which means the company has the least average value consumers

Conclusion on the customer segmentation

- From the statistical view, this company has a relatively stable customer base, and the overall development trend is stable. For future development, it is necessary to formulate different marketing and operation strategies for different groups of people.
- In this report, first use Age and SpendingScore to build a model to divide customers into 4 groups, which mainly focus on the Customers' Age and Purchasing Power. It is better to choose this segmentation when the company wants to launch and market some new products.
- Second, use Income and SpendingScore to build a model to divide customers into 5 groups, which mainly focus on the Customers' Income level and Purchasing Power. It is better to choose this segmentation when the company intends to increase its GMV or Orders performance. This segmentation model also the highest Silhouette score.
- Third, use Age, Income and SpendingScore to build a model to divide customers into 6 groups, which mainly focus on the customers' comprehensive profile. It is better to choose this segmentation when the company wants to build a detailed customer profiles to make better user management.