**Team Project: Mall Customer Segmentation**

Team 20 Members: [Roozbeh jafari](https://app.slack.com/team/U01B2SVLNFP), [Chang-Han Chen](https://app.slack.com/team/U01AZ7TJ2Q6), [Lequn Yu](https://app.slack.com/team/U01B2STLELD), [Jiahui Xu](https://app.slack.com/team/U01B5VBBD5G)

**Business Problem**

Market segmentation refers to that the enterprise divides the customers in the market into several customer groups according to a certain standard, and each customer group constitutes a sub-market. There are obvious differences in demand between different sub-markets.

After market segmentation, the sub-market is more specific, so it is easier to understand the needs of consumers and determine the target market. Aiming at the smaller target market, it is easy to formulate special marketing strategies. Through market segmentation, enterprises can analyze and compare the purchasing potential, satisfaction degree and competition of each market segment, explore the market opportunities that are beneficial to the enterprise, and work out the new product development plan. By segmenting the market and choosing the suitable target market, the enterprise can concentrate people, money, materials, and resources to strive for the advantages in the local market, and then occupy its own target market.

Through market segmentation, enterprises can face their own target market and produce marketable products, which can not only meet the needs of the market, but also increase the income of enterprises. To accomplish this task, machine learning has been applied in many stores. Shopping center uses customer data to develop ML model to locate the right customers. This not only reduces the cost of production and sales of enterprises, but also improves the quality of products and comprehensively improves the economic benefits of enterprises.

**Dataset**

* Mall customer segmentation data: <https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python>
* The dataset consists of 200 observations of mall customer data, 5 features.
* The data set we have chosen contains 2 types of input features there were collected at the mall:
* Objective: factual information.
* Subjective: calculated information given by the mall.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable name** | **Role** | **Data Type** | **Description** |
| CustomerID | Input | int | It is the unique ID given to a customer |
| Gender | Input | object | Gender of the customer |
| Age | Input | int | The age of the customer |
| Annual Income (k$) | Input | int | It is the annual income of the customer |
| Spending Score (1-100) | Input | int | It is the score (out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer. |

**Explanatory Data Analysis**

**Cleaning Dataset**

We start identifying and dropping uninformative column (CustomerID) and set it as index.

Moreover, there is no null and duplicate values, so we ensured the data is clean and ready to use.

A screenshot of a computer

Description automatically generated with low confidence

By exploring at AnnualIncome, it indicates there is a few outliers since there is a wide range from 75 percentile and max. By sorting AnnualIncome, it indicates that we have 6 observations that are much bigger than the remaining dataset.

In terms of male and female distribution the dataset is unbiased, and it is satisfying. 56% female and 44% male have been included in the dataset.

A picture containing chart

Description automatically generated**Exploratory data analysis and visualization**

There is no highly correlated feature. The highest correlation is between age and SpendingScore (negatively), which means we are not facing with some correlated features.

Diagram

Description automatically generated

Chart, histogram

Description automatically generated

* By analyzing distribution performances, on top left we see that the age is skewed toward right and is denser before 40.
* On center as mentioned earlier there is a clear evidence that there is a small distinct group which located between 120 and 137 including 3% of our dataset.
* On the right we see a normal distribution among the dataset in SpendingScore.

Chart, scatter chart

Description automatically generated

By plotting predictors in scatter form, we found two interesting points that we should consider in our models. First on center plot we can distinguish two potential clusters and then we can see a rabbit head shape with 5 clusters on the right plot.

Those all convey that we can find clusters with similar behaviors that we could design a recommendation approach to target customers in most efficient way.

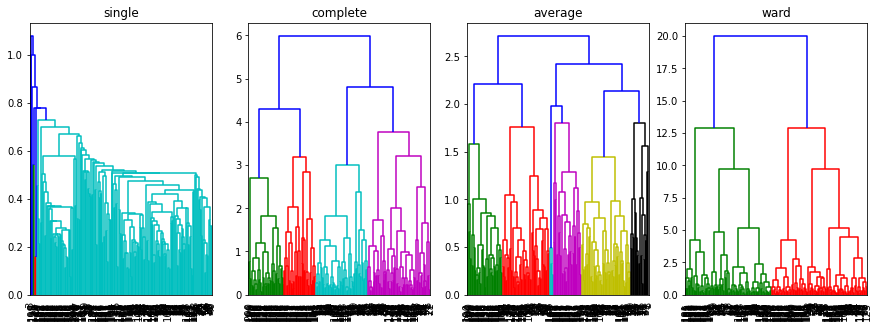
**Purpose of Analysis**

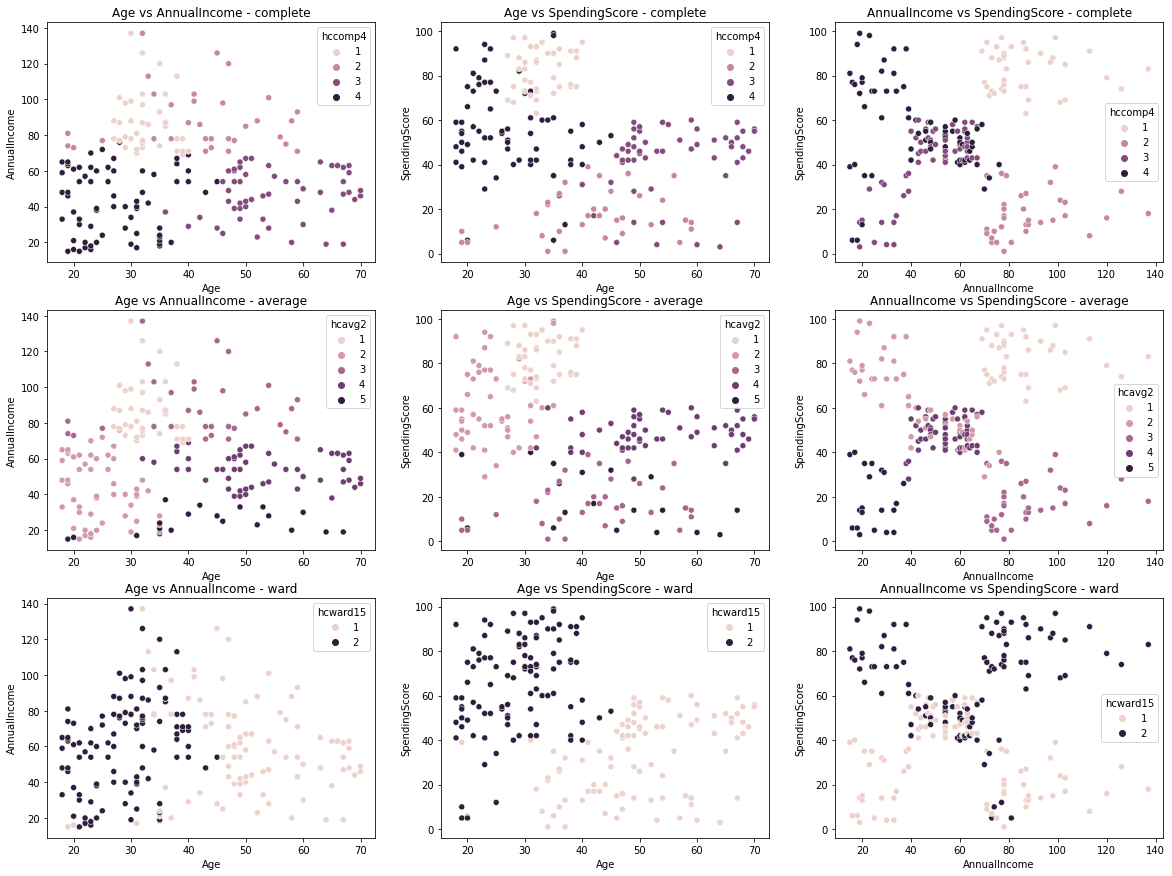
By analyzing and creating models in next part, we aim that our work allows operator of mall for effective allocation of marketing resources and cross and up-selling opportunities. When a group of customers is sent an email that is specific to their needs, it is easier for companies to send those customer special offers.

**Analytical Findings: Hierarchical Clustering**

In hierarchical clustering, four different methods (single, complete, average, and ward) are used to find out if there exists obvious pattern to divide observations into groups with similarities. All except the “single” method indicates pattern and these methods can separate the customers into balanced groups. For example, with “complete” method, if we slice the dataset into 4 groups, we have 69, 57, 39, 35 data in each group, respectively.

Next, we plot paired features and use different colors to distinguish clusters. After trying 4 different approaches, we found that the dataset could be divided into 4 or 5 clusters. For the “ward” method, we simply segmented the dataset into 2 groups. The bottom 3 plots show that the segmentation is somewhat rough, though this method separates the dataset well. However, the “complete” and “average” methods show clear patterns, for example, when looking at the scatterplot between AnnualIncome versus SpendingScore, the pattern of 4 corners is well captured by clustering.





**Analytical Findings: K-Means Clustering**

In K-Means clustering, the first goal is to find a proper k (number of clusters). According to the elbow plot and the silhouette plot, it is more appropriate to use 5 clusters to fit the K-Means model, as inertia decreases slower when k is greater than 5 and the Silhouette Score arrives the peak when k equals to 5.

Compared to the result from hierarchical clustering, the 5 clusters contain 28(average), 56(High), 39(Highest), 48(low), and 29(lowest) spender customers, respectively. Hence with K-Means clustering method, the customers are evenly distributed among the 5 clusters. Also, we observed from the scatterplot between features and used different colors to identify clusters, the K-Means method captures similar pattern comparing to the hierarchical clustering method.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, bubble chart

Description automatically generatedChart, bubble chart

Description automatically generatedChart, bubble chart

Description automatically generated

* Left Figure:
* Cluster1: High income low spending = Careful
* Cluster2: Medium income medium spending = Standard
* Cluster3: High Income and high spending = Target
* Cluster4: Low income and high spending = Careless
* Cluster5: Low income and low spending = sensible
* Center Figure: We cluster the data into four group.
* Low spenders
* Young High Spenders
* Young Average Spenders
* Old Average spenders

We can clearly see that Only young people (18-40 age group) are involved in High Spending. As age increases people fall into average or Low spending category.

* Right Figure: We can see people in age group 0f 30-40 have high number of high-income people.

**Analytical Findings: Principle Component Analysis**

Based on what we have seen from data behavior and similarity and as expected, the model correctly formed 5 clusters without large number of outliers. By assigning 90% of variance hyperparameter, our model stayed with 4 components. By running the model, we labeled the dataset and then named them based on mean of spending score.

Table

Description automatically generatedChart, bar chart

Description automatically generated

**Conclusion & Business Suggestions**

Basically, our team aims to use unsupervised ML to group customers into clusters with other customers that are similar in terms of their measured attributes like Annual Income, Spending Score, etc. Recognizing customers by their similarity will increase efficiency of money allocation for target customers and then mall operator can allocate the budget in more efficient way and avoid wasting money for example in advertising and campaigns.

* Running different campaigns for each targeted group.
* People aged from 21 to 35 are the most spenders and almost 50% of this range are in high and highest groups, so allocating resources on this range could be priority.

According to our analysis, the mall is recommended to develop marketing strategies with groups of similar customers. For example, research from Deutsche Bank in 2017 shows that 40% of American low-income families tend to consume goods on luxuries, which corresponds to one of our clusters—low income and high spending. Though it should be better if the consumers fall in this clusters can make ends meet.

Graphical user interface

Description automatically generated with medium confidence