**Team Project: mall customer segmentation**

Team member: [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)

**Project motivation and 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. In order 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. |

**Steps to solve the problem**

Exploration of data

Data cleaning will entail checking missing values in the dataset, using dummy variables and dropping irrelevant features.

Data visualization

Observe the distribution of each features and to see whether there is a relationship between different variables.

Selection of clusters

Do each segment between different feature and see try to find the best number of clusters.

Visualization of Cluster Result

Because this is an unsupervised problem, we can't determine which is the best in real life. We can only look at the data and synthesize all the pictures to see which cluster the best result should be.

**Proposed analysis methodology:**

1. Cleaning Data and basic EDA

2. Using different UML algorithms, Clustering, KMeans

3. Identify the best model and interpret the result.