# Shopping Mall Customers Clustering Based on LOF and K-means++

Yuhang Deng[1, *, †], Peiwei Wu[2, †]

[1]Department of Mechanical Engineering, Wuhan University of Technology, Wuhan, China

[2]Department of Data Science and Big Data Technology, Beijing Institute of Technology, Beijing, China

*Corresponding author: 325144@whut.edu.cn

†These authors contributed equally

**Abstract.** Effective customer analysis is vital for business success, and AI-based customer analysis can significantly improve its accuracy, particularly regarding customer grouping and labeling, enabling merchants to generate personalized marketing strategies. Theories on machine learning models for grouping data have been developed since the 1950s, including logistic regression, Support Vector Machine (SVM), decision tree, and random forest; however, computational limitations hindered their practical application. Recent advances in computer technology have led to the development of more accessible machine learning algorithms that generate high-value results. The K-means clustering algorithm is one such model that best fits the customer labeling requirements. As an unsupervised training model, the K-means algorithm clusters customer data into a predetermined number of clusters. In this paper, we apply the K-means algorithm to separately cluster data on male and female clients, while using the K-means++ model to keep initial cluster centers as far apart as possible. We also apply the LOF algorithm to remove any outliers and modify the dataset accordingly.

**Keywords:** K-means++, LOF, Customer segmentation.

## 1. Introduction

A deep understanding of consumers is the basis for forming a marketing strategy, because the financial response of consumers to marketing strategies determines the success of an enterprise. In the era of information, customer analysis is especially essential for merchants. Through analyzing the data of customers, companies can develop better strategies, design high-quality products and offer exceptional service to customers. Within the customer analysis system, there are two important factors: The value of customers and the demand of customers. The former one can contribute to labeling and grouping customers, while the latter one can be better predicted through a refined analysis of customers. For these two factors, the development of Artificial Intelligence (AI) will significantly contribute to high accuracy analysis and generate personalized strategies to customers, through building proper mathematical model based on various types of customer data [1]. Offering personalized service to specific customers can maximize the probability of stimulating consumption and improving overall profits.

In 1956, the term artificial intelligence was officially used at the Dartmouth College Artificial Intelligence Summer Seminar. This is the first discussion on artificial intelligence in human history, marking the birth of the discipline of artificial intelligence [2]. In several decades after that, scientists have been engaged in discussions regarding the concepts, applications, and mathematical models of Artificial Intelligence. However, due to the limited computing power, the relevant theories of AI have not practically implemented for an extended period. Over the past few years, with improvements of computer performance as well as the application of cloud computing, AI has achieved significant development and has been widely applied. Deep learning, a widely mentioned topic this year, is a rapidly developing subject within the field of AI. One of the most famous examples of deep learning is AlphaGo, which is a trained Go program. In 2017, Alphago beat Ke Jie in a Go Summit, which means an iconic progress in deep learning technology [3]. Another sensational deep learning

application is ChatGPT (generated pre-training), which was published at the end of 2022.The model's ability to communicate and handle complex tasks such as programming and creating table of data had made a great impression on most people. The basic technique of it is GPT-4, which is a language model trained by the database composed of huge amount of corpus. At the same time, a wide range of machine learning models was discovered and applied to solve specific decision and prediction tasks in real life. For example, many classification and decision problems are excellently solved by building a Random Forest model. Using the method, we can select deciding factors out from relatively unimportant ones and form a decision map to give predictions for specific cases [4].

For the problem of customer analysis, a machine learning model called K-means Clustering Algorithm can nicely fit the case. It is an unsupervised learning model, which gives classification results through iterative calculation [5]. Basically, the algorithm is applied to train and classify the processed data into preset number of groups. However, previous research on customer analysis basically ignored the consumption difference between men and women. Furthermore, because the original K-means algorithm chooses initial data points as the centroid in a random way, the output classification result can be highly sensitive to the initial centers. Therefore, the classification error could be large compared to the modified K-means ++ model. Meanwhile, we modify our dataset applying Local outlier factor Algorithm (LOF) to remove the divorced points [6].

In this paper, we obtained the dataset on customer analysis provided by Kaggle. The main information about customers includes gender, annual income and spending score. We first divide the data into men group and women group, then apply LOF algorithm to remove the outliers. After that, we fit the improved K-means++ model on the divided data and generate the classification result. Finally, we analyze our pre-defined number of clusters and evaluate the best cluster number (K value), therefore forming an improved classification scheme for both genders.

## 2. Method

### 2.1 Dataset description and preprocessing

This project utilized the mall dataset obtained from Kaggle, comprising of 200 customer profiles, with specific information including ID, age, annual income, and spending scores. The whole dataset contained no NaN values and no need to manually remove rows and columns with Null values. Following the removal of the unnecessary ID information there were 200 rows $\times 4$ columns left in the dataset.

The preprocessing consisted of three parts. First, this project divided the dataset into two datasets about male and female. The number of male datasets was 88 rows $\times 4$ rows, and the number of female was 112 rows $\times$ 4 columns. The two groups were subsequently clustered separately, utilizing various parameters. Secondly, outliers were detected in the dataset, potentially affecting cluster center selection. Therefore, the Local Outlier Factor (LOF) algorithm was employed to eliminate such outliers, as depicted in Fig. 1, where the gray points represent the outliers. Thirdly, Data standardization which contributes to the model's speed of convergence is also utilized, and since K-means++ needs to compute the Euclid distance, standardization can improve the accuracy.

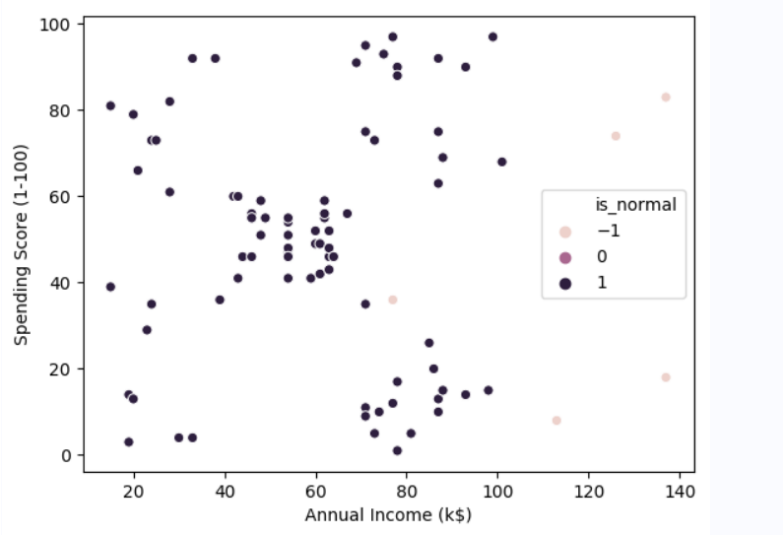


**Fig. 1** Outliers in dataset

In order to determine the factors that affect the spending scores of male and female customers, we use heatmap to analyze the strength of the correlation of each factor. Through heatmap, like Fig. 2, Fig. 3, spending scores of males are negatively correlated with age, while spending scores of females are negatively correlated with age and positively correlated with annual income.

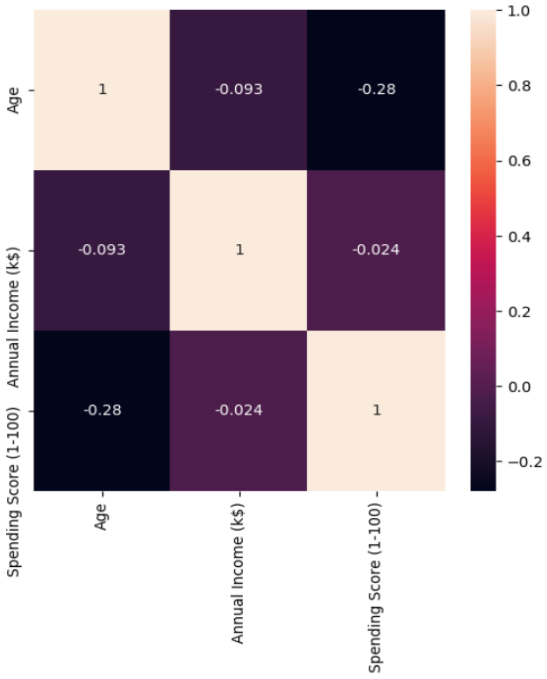


**Fig. 2** Heat map of male customers

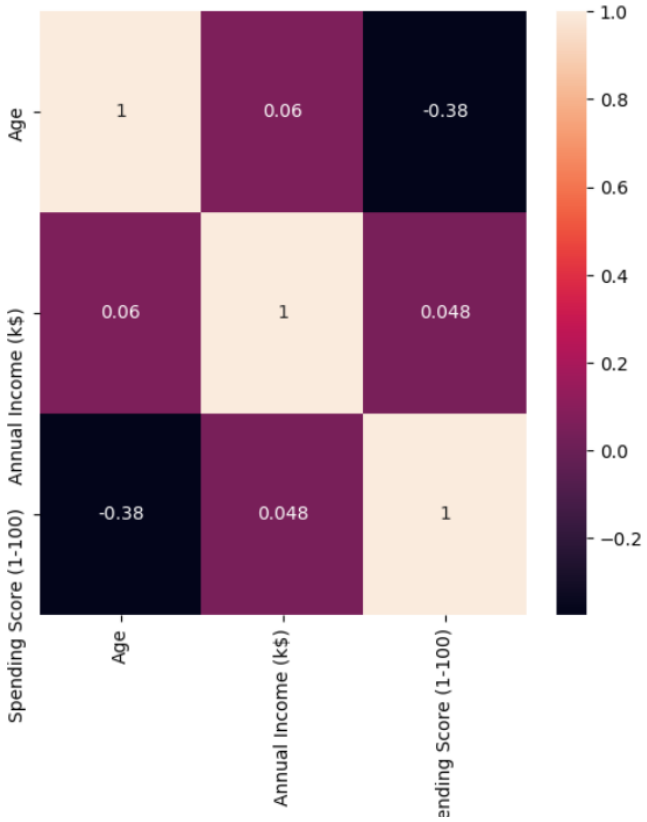


**Fig. 3** Heat map of female customers

### 2.2 Proposed approach – K-means++

Since the results of K-means algorithm are easily affected by the selection of initial points, an improvement of this algorithm called K-means++ can be considered [7]. The difference between them is the method of choosing initial clustering centroids. The fundamental principle of K-means++ algorithm in the selection of the initial cluster centroid is: the distance between initial cluster centroids should stay as far as possible.

Steps: 1) Randomly select a point from the data set as the first center point; 2) Calculate the shortest distance D(x) between each sample and the existing cluster centroid, that is, the distance from the nearest one. The larger the value, the more likely it is to be chosen as a clustering center; The likelihood for each sample to be chosen as the next cluster centroid is computed $\frac{D(x)^2}{\sum D(x)^2}$, and the following cluster center is chosen by the Roulette Wheel Selection. 3) Repeat step (2) until k clustering centers are obtained. 4) Calculate Euclidean Distance between other samples and k numbers of clustering centers and compare. And group the samples with the nearest central point 5) Repeat step (2) to (4) until the position of the cluster centroids reaches convergence.

In K-means++ method, Roulette Wheel Selection is mainly introduced to screen cluster center points [8]. In this method, the likelihood to be selected for each individual is positively correlated to its fitness value, and the better the fitness is, the higher the selection possibility. Steps: 1) Calculate the fitness of each individual $f(i = 1, 2, \ldots, m)$ in the population, where m is the total number of customers; 2) Compute the likelihood of each individual being passed on to the next generation:

$$P(x_i) = \frac{f(x_i)}{\Sigma_{j=i}^{N} f(x_j)} \tag{1}$$

3) Calculate the cumulative probability for each person:

$$q_i = \sum_{j=1} ip(x_j) \tag{2}$$

Like most clustering algorithms, K-means++ algorithms also need the user to define the value of cluster(k). In this paper, we apply the elbow method and contour coefficient method to determine the optimum value of K [9, 10]. After comparison, as shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7, the optimal value of k in male classification is 5, while that in female classification is 6.

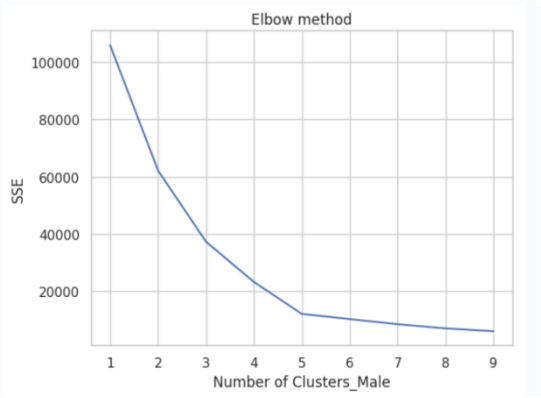


**Fig. 4** Elbow method of Male


**Fig. 5** Contour coefficient method of male

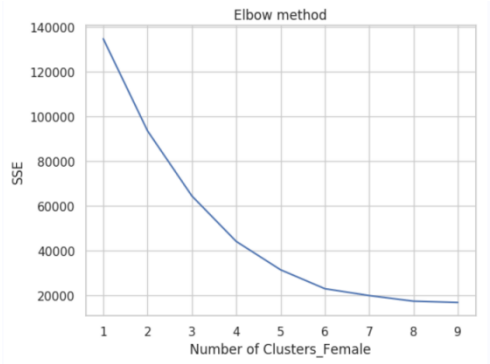


**Fig. 6** Elbow method of Female

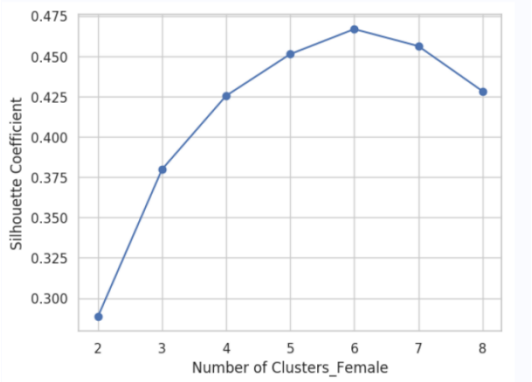


**Fig. 7** Contour coefficient method of female

## 3. Results and Discussion

Upon application of the K-means ++ and K-means clustering algorithms on customer segmentation, it was found that K-means ++ algorithm outperformed K-means algorithm in the segmentation of female customers. The empirical results presented in Fig. 8 and Fig. 9 demonstrated that, despite an initial definition of cluster center K set to 6, K-means algorithm was only able to classify female customers into 5 clusters. On the other hand, K-means ++ algorithm has the ability to generate segmentation for female customers in better accuracy.

The final outcome of the customer segmentation process using K-means ++ algorithm resulted in the dividing male customers into 5 categories and female customers into 6 categories. The clustering results of male and female customers were further illustrated in Fig. 10 and Fig. 11, which depicted the scatter plot of the customer clustering results.

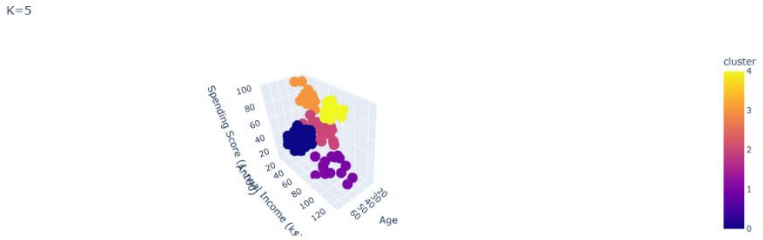


**Fig. 8** The visualized results for K=5

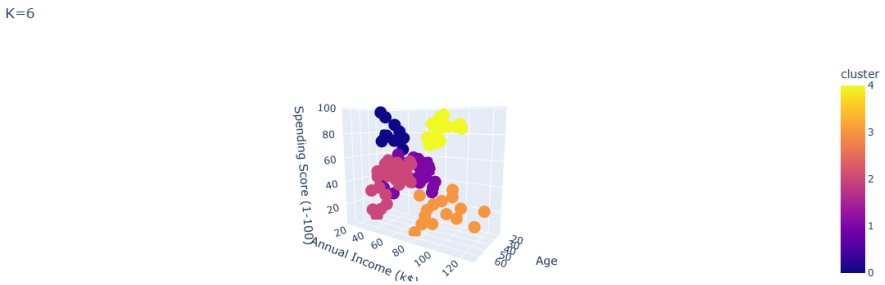


**Fig. 9** The visualized results for K=6

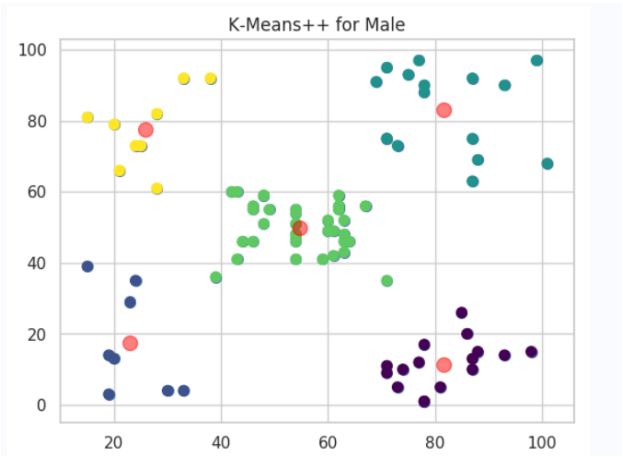


**Fig. 10** The clustering results for male based on the K-means++

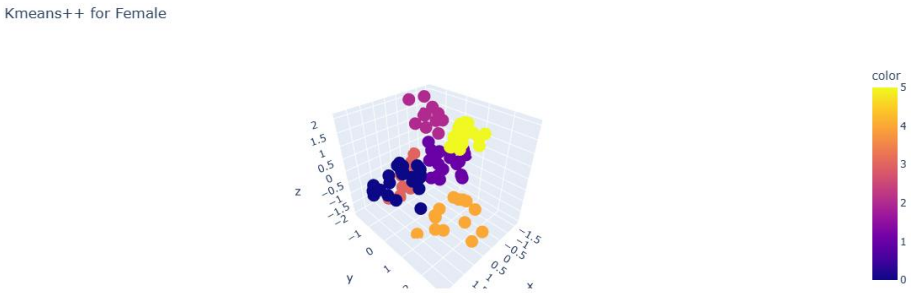


**Fig. 11** The clustering results for female based on the K-means++

**Table 1**. Parameter correlation and algorithm comparison

| Gender | Correlation | | | Method | |
|---|---|---|---|---|---|
| | | Age | Annual Income | K-means | K-means++ |
| Total (number=200) | Age<br>Annual Incom<br>Spending Score | -0.012<br>-0.33 | 0.0099 | K=5 | K=5 |
| Male (number=88) | Age<br>Annual Incom<br>Spending Score | -0.093<br>-0.28 | -0.024 | K=5 | K=5 |
| Female (number=112) | Age<br>Annual Incom<br>Spending Score | 0.06<br>-0.38 | 0.048 | K=5 | K=6 |

The findings from Table 1 indicate that K-means ++ algorithm performs better than K-means algorithm in terms of clustering accuracy. Moreover, the K-means++ algorithm selects the optimal

value of K for the cluster centroids based on the data distribution, thereby enhancing the effectiveness of the clustering process. In addition, it is noteworthy that the correlation among factors influencing customer Spending Score has improved to a certain degree. Specifically, the most notable change observed was the relationship between consumption points and annual income, which shifted from a positive correlation coefficient of 0.0099 to a negative correlation coefficient of -0.024 for male customers, and from 0.0099 to 0.048 for female customers.

Subsequently, the male and female customer clustering graphs were utilized to segment the customers into distinct categories using the RFM model, which took into account factors such as consumer level consumption points, annual income, and age. Based on these reference factors, the appropriate marketing strategies were identified for each customer category shown in Table 2.

**Table 2**. Customer value

| Gender | | Annual income | Spending scores | Type |
|---|---|---|---|---|
| Male | Customer 0 | Low income | Low spending | Low value customers |
| | Customer 1 | Low income | High spending | Important keep customers |
| | Customer 2 | Middle income | Middle spending | Average value customers |
| | Customer 3 | High income | Low spending | Important development customers |
| | Customer 4 | High income | High spending | Important value customers |

| | | Age | Annual income | Spending scores | Type |
|---|---|---|---|---|---|
| Female | Customer 0 | Youth | Low income | Low spending | Low value customers |
| | Customer 1 | Youth | Low income | High spending | Important keep customers |
| | Customer 2 | Youth | High income | High spending | Important value customers |
| | Customer 3 | Middle age | Low income | Low spending | Low value customers |
| | Customer 4 | Middle age | Middle income | Middle spending | Average value |
| | Customer 5 | Middle age | High income | Low spending | Important development customers |

Therefore, malls can specify corresponding marketing suggestions for different types of customers. For important value customers, malls should invest resources on them to maintain the loyalty of such customers. For important development customers, malls can strengthen the satisfaction of such customers to the mall, improves the attraction, and makes them gradually become loyal customers. For important keep customers, malls should strengthen the promotion with such customers and extend the life cycle of customers. For average value customers and average development customers, malls should strengthen their contact with such customers. For low value customers, malls can use vouchers to attract customers to trade.

## 4. Conclusion

In this paper, we employed K-means ++ model for the customer segmentation. In this study, of particular importance is the correlation discrepancy between spending score and annual income for males and females. Gender differences in customer analysis are often overlooked, resulting in inaccurate customer labeling based on gender. In this study, we address this issue by conducting a separate analysis of male and female datasets and developing an improved clustering system or method for both genders. Using the RFM model, we generate the optimal cluster number for men and women and adjust our division logic to fit the best K value. These refined methods enable us to better cluster and label customers based on their gender, thereby improving the accuracy of our customer analysis. The scope of future work may lie in training the K - means model for customer data in more dimensions, which we have made a step on it (taking age into consideration for female data), such as category-based spending score for each customer for enterprises that sell multi-category products.

## References

[1] Tanveer M, Khan N, Ahmad A R. AI Support Marketing: Understanding the Customer Journey towards the Business Development. 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA). IEEE, 2021: 144-150.

[2] Moor J. The Dartmouth College artificial intelligence conference: The next fifty years. Ai Magazine, 2006, 27(4): 87-87.

[3] Chao X, Kou G, Li T, et al. Jie Ke versus AlphaGo: A ranking approach using decision making method for large-scale data with incomplete information. European Journal of Operational Research, 2018, 265(1): 239-247.

[4] Liaw A, iener M. Classification and Regression by random Forest. R News, 2002, 23(23).

[5] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 1979, 28(1): 100-108.

[6] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density-Based Local Outliers. Acm Sigmod International Conference on Management of Data. ACM, 2000.

[7] Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding. Proc. Of the Eighteenth Annual ACM – SIAM Symposium on Discrete Algorithms (SODA)Society for Industrial and Applied Mathematics, Philadelphia, 2007:1027-1035.

[8] Yu Q, Chen P, Lin Z, et al. Clustering Analysis for Silent Telecom Customers Based on K-means++. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 1023-1027.

[9] Jenssen R, Eltoft T. A new information theoretic analysis of sum-of-squared-error kernel clustering. Neurocomputing, 2008, 72(1-3): 23-31.

[10] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 1987, 20: 53-65.