International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

# A Study on Machine Learning Approach for Market Segmentation

[Mr. Prateek Dutta[1]]

Bachelors of Technology in Artificial Intelligence, India

## Abstract

Customer segregation can be a powerful way to identify unsatisfactory customer needs. This approach can be used by companies to outperform the competition by building more attractive products and services. Customer profile and purchase history were treated as internal data while server log, cookies and survey data were considered as external data. This information can be processed using one of several methods: Business Rule, Magento, Customer Performance, Quantile Membership, RFM Cell Separation Collection, Supervised Collection, Customer Identity Collection, Merge Purchase and Unchecked Mergers. Big data concepts and machine learning have encouraged greater acceptance of automated customer segregation methods by harvesting traditional market statistics that do not work whenever the customer base is too large. In this paper, the k-means clustering algorithm is used for this purpose. The Sklearn library is designed with the k-Means algorithm and the database available for commercial use. Features the average customer purchase number and the monthly customer average number.

***Keyword: Clustering, Customer Analytics, K-means, Python, Segmentation, Data Mining***

## 1. Introduction

The development of ecommerce began as the internet grew and continues to this day, especially in B2C ecommerce (Business to Customer). When shopping using ecommerce, the user finds it easier and faster [1]. Excessive information can be overcome by implementing personalization in ecommerce services such as providing product recommendations, linking recommendations, ads or text and graphics tailored to the user's features and needs [2]. In addition to solving the problem of over-information, customized services in ecommerce can maintain customer loyalty of existing customers, gain new customers by providing services to customers according to their needs and features [3].

Customer segregation is market segregation by different customer groups sharing the same characteristics. Customer segregation can be a powerful way to identify unsatisfactory customer needs. Using the above data companies can bypass the competition by creating more attractive products and services. Intelligence Segmentation Intelligence to improve marketing by providing products or services that meet the needs of each customer group. Collica-segregation is a process of classifying or classifying an object into a group with a similar feature and in the CRM (Customer Relationship Management) category used to classify customers according to a certain similarity by separating customer database records [4].

Boone and Roehm [5] studied Hopfield-Kagmar (HK) clustering method of customer segmentation using Hopfield's artificial neural network technology. The study has shown that, each neuron in HK clustering method is connected with other neurons , and information can flow between neurons in multiple directions, which is more suitable for customer segmentation than the K-means clustering method; Kim et al. [6] used neural network clustering method to segment the customers of tourism; contrasting K-means, self-organizing map neural network and particle swarm optimization for three kinds of clustering algorithm, Deng et al. proposed hybrid clustering algorithm which was used for segmentation problem of catering industry customer.

The ways in which businesses segments their customers information are:

1. **Demographic information**, such as gender, age, familial and marital status, income, education, and occupation.
2. **Geographical information**, which differs depending on the scope of the company. For localized businesses, this info might pertain to specific towns or counties. For larger companies, it might mean a customer's city, state, or even country of residence.
3. **Psychographics**, such as social class, lifestyle, and personality traits.
4. **Behavioural data**, such as spending and consumption habits, product/service usage, and desired benefits.

**Advantages of Customer Segmentation**

1. Determine appropriate product pricing.
2. Develop customized marketing campaigns.
3. Design an optimal distribution strategy.
4. Choose specific product features for deployment.
5. Prioritize new product development efforts.

## 2. Algorithm

### 2.1. K Means Clustering Algorithm

1. Specify number of clusters *K*.
2. By shuffling the dataset firstly, initialize centroids and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn't changing.

Figure:- 1

## 3. Comparative analysis

Customer segregation is a preliminary task of creating an effective and strategic marketing strategy for customer relationship management. Influenced by factors such as social environment and customer psychological performance, customer actions are often complex, increasing complexity, and acquiring key customer segregation qualities on the basis of not losing information [7]. K-means that the algorithm is one of the most popular category algorithms. This merging algorithm is centrally based, in which each data point is inserted into another spacing, pre-programmed into the K-algorithm [8]. The Customer Segmentation process in the Lieberman study [9] begins with determining business law, data collection disseminates the inquirer, and then corrects the data by order and falls and analyses statistical data. Birant [10] has a more complex process than Lieberman because it combines RFM Analysis with Data Mining to get product recommendations.

## 4. Data

Data collection is the process of collecting and measuring data in relation to targeted changes in an established system, which enables one to answer relevant questions and evaluate results [11]. Data collection is part of research in all fields of study including physical and social sciences, humanities and business [12]. The purpose of all data collection is to obtain quality evidence that leads to analysis and constructs concrete and misleading answers to the questions presented. The project data has been extracted from the Mall Customer Segmentation Data competition held in Kaggle. The database can be downloaded from the Kaggle website.

## 5. Environment and tools

1. scikit-learn
2. seaborn
3. numpy
4. pandas
5. matplotlib

We started with loading all the libraries and dependencies. The features in the data are customer id, gender, age, income and spending score.

| | CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|---|
| 0 | 1 | Male | 19 | 15 | 39 |
| 1 | 2 | Male | 21 | 15 | 81 |
| 2 | 3 | Female | 20 | 16 | 6 |
| 3 | 4 | Female | 23 | 16 | 77 |
| 4 | 5 | Female | 31 | 17 | 40 |

Figure:- 2

We plotted the age frequency of customers.

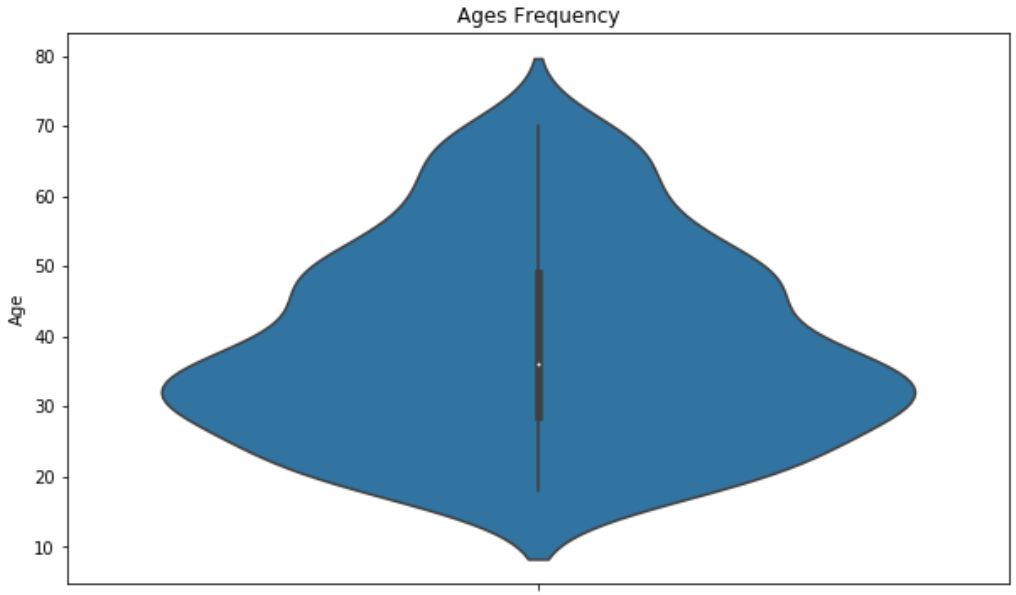


Figure:- 3

Next we made a box plot of spending score and annual income to better visualize the distribution range. The annual income range is clearly less than the range of spending score.

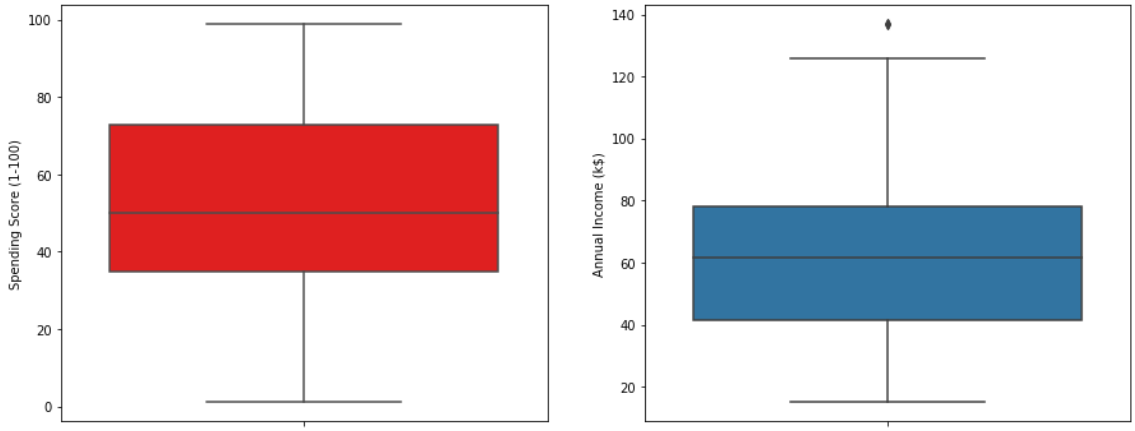


Figure:- 4

We made a bar plot to check the distribution of male and female population in the dataset. The female population clearly outweighs the male counterpart.

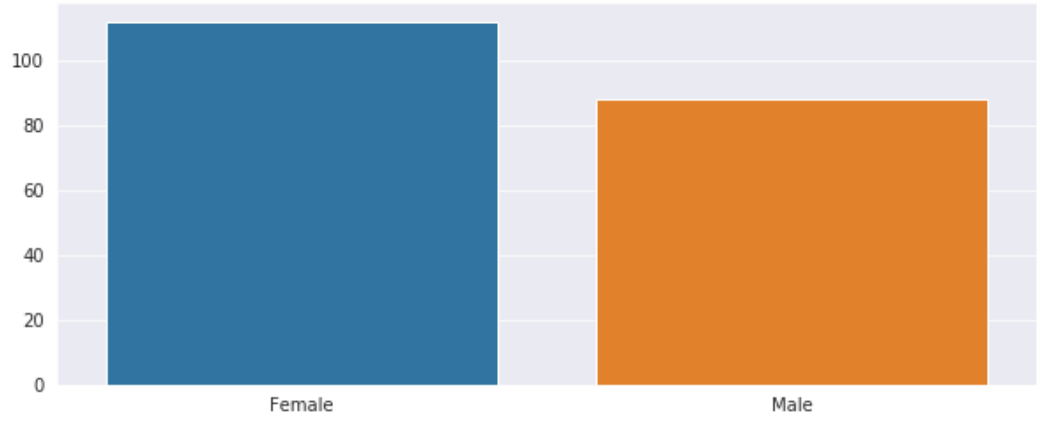


Figure:- 5

Next we made a bar plot to check the distribution of number of customers in each age group. Every other age group are clearly outweighed by 26-35 age group.

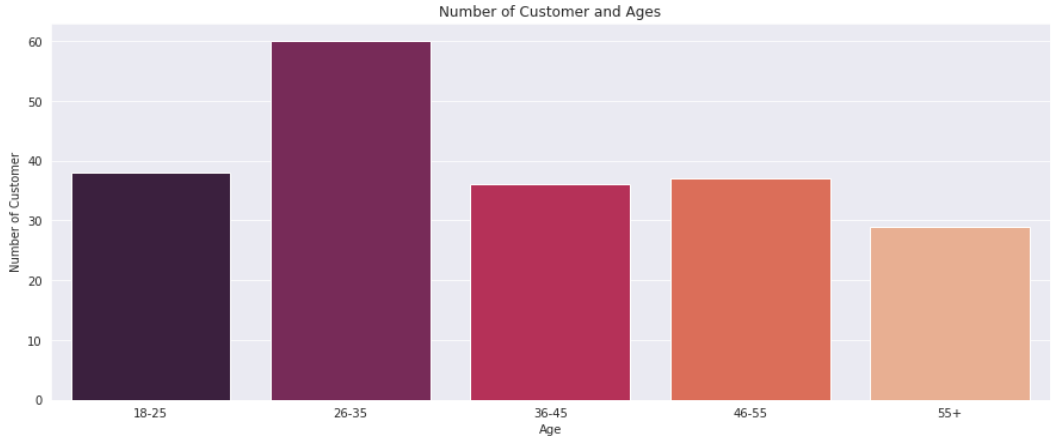


Figure:- 6

We continued with making a bar plot to visualize the number of customers according to their spending scores. The maximum customers are spending score in the range 41–60.

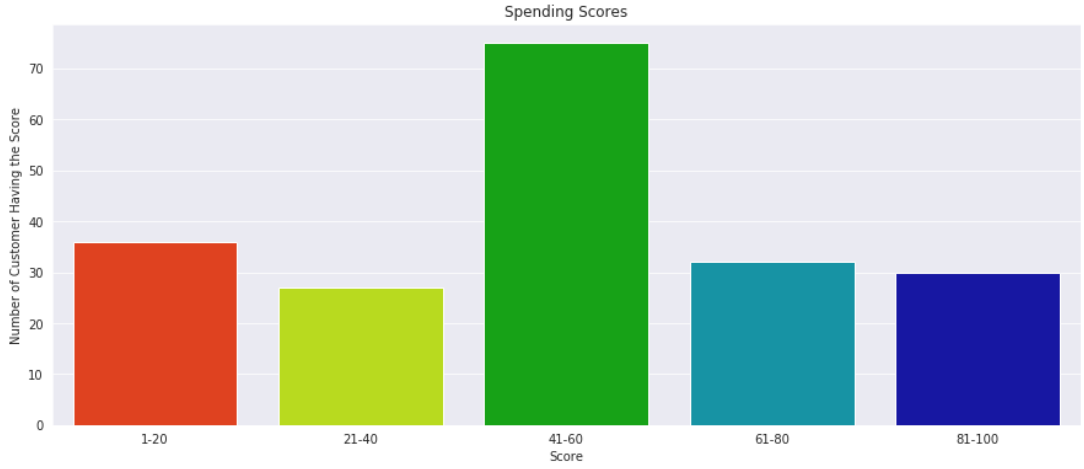


Figure:- 7

Also, we made a bar plot to visualize the number of customers according to their annual income. Maximum customers have annual income in the range 60000 and 90000.

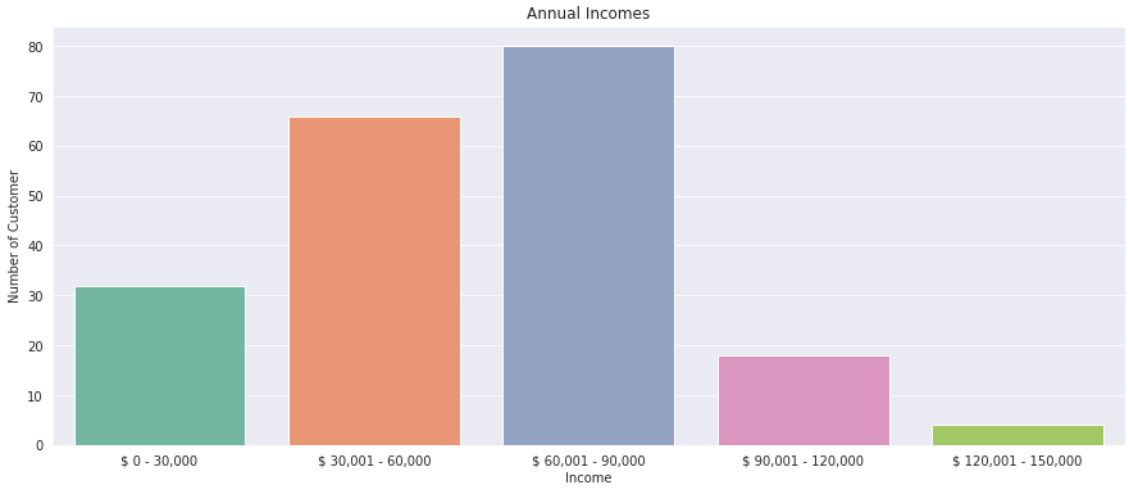


Figure:- 8

Next we plotted Within Cluster Sum Of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. The sum of distances of observations from their cluster centroids is measured by WCSS, which is given by the below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where *Yi* is centroid for observation *Xi*. The main aim would be to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

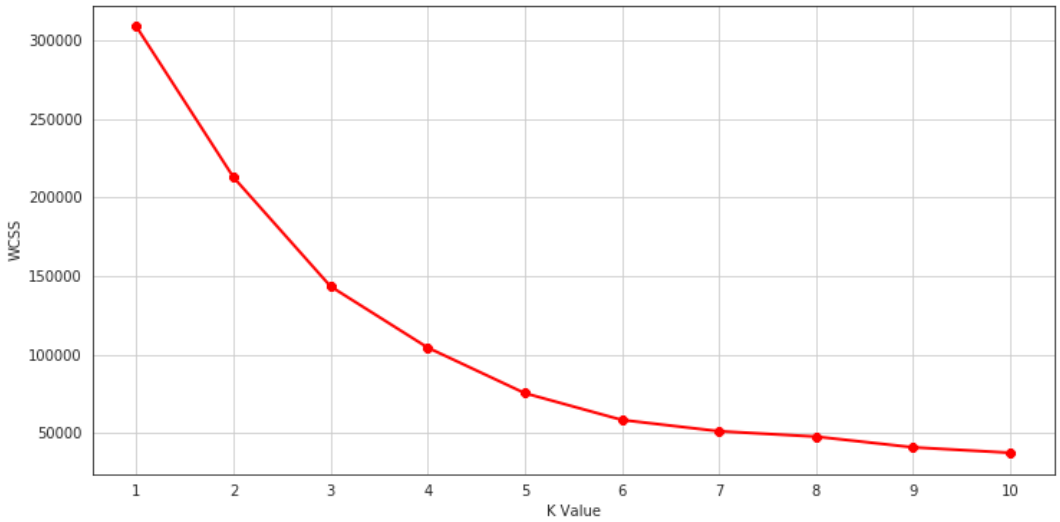


Figure:- 9

## 6. Results & Conclusion

Here, the result suggests that the orange collection as customers have the highest value, green as the lowest value customers, and blue and red as the highest potential customers. Customer segregation is associated with the purpose of the business. The first step in the division is to determine the purpose of the business.

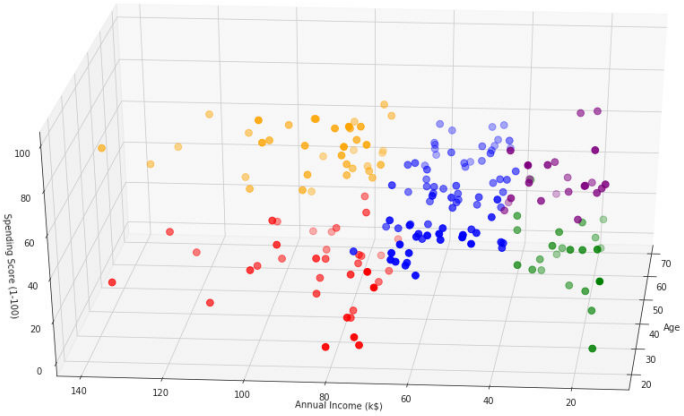


Figure:- 10

K means clustering is one of the most common algorithms and is often the first thing that works when solving clustering tasks to get an idea of a database structure. The purpose of the K methods is to collect data points into subdivisions. One of the major applications of K means customer segmentation to better understand those that can be used to maximize company revenue. Throughout the research customer segmentation has been demonstrated with machine learning algorithms. The graphical approach has been used to showcase the result in better and understandable way. Clustering approach has been implemented finely to group the data categorically and represent it in more efficient way.

## 7. Reference

*[1] Juni Nurma Sari, Lukito Edi Nugroho, Ridi Ferdiana, P. Insap Santosa. "Review on Customer Segmentation Technique on Ecommerce". American Scientific Publisher. Adv. Sci. Lett. 4, 400–407, 2011.*

*[2] Mobasher B, Cooley R, Srivastava J. "Automatic Personalization Based on Web Usage Mining". Commun ACM. 2000;43(8).*

*[3] Cherna Y, Tzenga G. "Measuring Consumer Loyalty of B2C e-Retailing Service by Fuzzy Integral: a FANP-Based Synthetic Model". In: International Conference on Fuzzy Theory and Its Applications iFUZZY.; 2012:48-56.*

*[4] Colica R. "Customer Segmentation and Clustering Using SAS Enterprise Miner". Part I The Basics. 2011:1-14.*

*[5] D.S. Boone, M. Roehm, Retail segmentation using arti-ficial neural networks, International Journal of Research in Marketing, 19 (2002), 287-301*

*[6] J Kim, et al. Segmentation the market of West Australi-an senior tourist using artificial neural network [J]. Tourism Management, 2003, 24 (1): 25-34.*

*[7] Wei Gao, Huiting Jia, Ruzhen Yan. "Customer segmentation model based on two-step optimization in big data era". International Conference on Information Technology and Management Innovation (ICITMI 2015)*

*[8] AMAN BANDUNI, Prof ILAVENDHAN A. "Customer Segmentation using Machine Learning". IJCRT_196519*

*[9] Lieberman M. Target " golden egg " consumer to achieve maximum ROI. 2009;(May):50-51.*
*[10] Birant D. "Data Mining Using RFM Analysis". Knowledge Oriented Appl Data Min. 2011;(iii):91-108. doi:10.5772/13683.*
*[11] A.K. Jain, M.N. Murty and P.J. Flynn. "Data Integration: A Review". ACM Computer Research. 1999. Vol. 31, No. 3.*
*[12] Prateek Dutta. "A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION". International Research Journal of Modernization in Engineering Technology and Science. Volume-3, Issue-11. January 2021.*