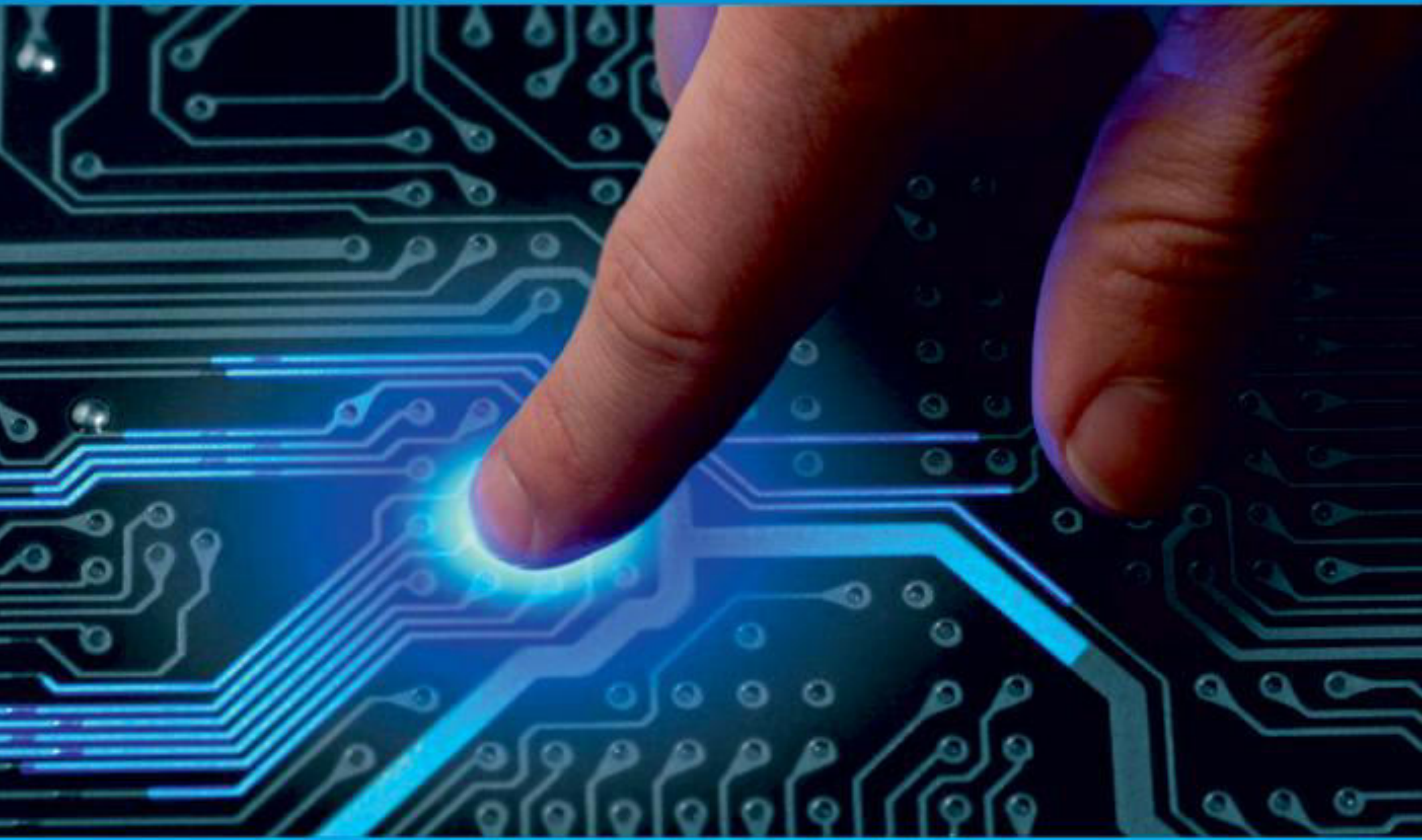
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906216 |

Customer Segmentation Analysis for Improving Sales Using Clustering

Dr.P.Srinivasa Rao [1], M. Santhosh Kumar [2], N. Sri Harsha[3], N. Laxmi Prasad[4],

T. Srihari [5]

Professor, Department of Computer Science Engineering, J B Institute of Engineering and Technology, Moinabad, Hyderabad, Telangana, India [1]

B. Tech Student, Department of Computer Science Engineering, J B Institute of Engineering and Technology, Moinabad, Hyderabad, Telangana, India [2,3,4,5]

ABSTRACT: Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers and potential customers, a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators. To scale efficiently and effectively, expansion stage companies need to focus their efforts not on a broad universe of potential customers, but rather on a specific subset of customers who are most similar to their best current customers. The key to doing so is through customer segmentation. The segmentation is based on customers having similar 'needs'(so that a single whole product can satisfy them) and 'buying characteristics'(responses to messaging, marketing channels, and sales channels, that a single go-to-market approach can be used to sell to them competitively and economically).

KEYWORDS: Clustering, data collection, Data pre-processing, Model training , Dataset preparation, Administrator, Customer.

I.INTRODUCTION

In the contemporary day and age, the importance of treating customers as the principal asset of an organization is increasing in value. Organizations are rapidly investing in developing strategies for better customer acquisition, maintenance and development. The concept of business intelligence has a crucial role to play in making it possible for organizations to use technical expertise for acquiring better customer insight for outreach programs. In this scenario, the concept of CRM garners much attention since it is a comprehensive process of acquiring and retaining customers, using business intelligence, to maximize the customer value for a business enterprise.

One of the two most important objectives of CRM is customer development through customer insight. This objective of CRM entails the usage of an analytical approach in order to correctly assess customer information and analysis of the value of customers for better customer insight. Keeping up with the changing times, organizations are modifying their business flow models by employing systems engineering as well as change management and designing information technology(IT) solutions that aid them in acquiring new customers, help retain the present customer base and boost the customers lifelongvalue.

Due to the diverse range of products and services available in the market as well as the intense competition among organizations, customer relationship management has come to play a significant role in the identification and analysis of a company's best customers and the adoption of best marketing strategies to achieve and sustain competitive advantage. One of the most useful techniques in business analytics for the analysis of consumer behavior and categorization is customer segmentation. By using clustering techniques, customers with similar means, end and behavior are grouped together into homogeneousclusters.

Customer Segmentation helps organizations in identifying or revealing distinct groups of customers who think and function differently and follow varied approaches in their spending and purchasing habits. Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behavior, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment

II.RELATEDWORK

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as *"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."*

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

The clustering technique is commonly used for statistical data analysis.

III.METHODOLOGY

algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinctnon-overlappingsubgroups(clusters)whereeachdatapointbelongstoonlyonegroup.It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that be long to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k means algorithm works is as follows:

Specify number of clusters*K*.

Initialize centroids by first shuffling the data set and then randomly selecting *K* datapoints for the centroids without replacement.

Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’tchanging.

Compute the sum of the squared distance between data points and all centroids.

Assign each data point to the closest cluster(centroid).

Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The approach k means follows to solve the problem is called Expectation-Maximization. The E- step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a break down of how we can solve it mathematically (feel free to skip it).

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \qquad (1)$$

where wik=1 for data point xi if it belongs to cluster *k*; otherwise, wik=0. Also, μk is the centroid of xi's cluster.

It's a minimization problem of two parts. We first minimize J w.r.t. wik and treat μk fixed. Then we minimize J w.r.t. μk and treat wik fixed. Technically speaking, we differentiate J w.r.t. wik first and update cluster assignments (*E-step*). Then we differentiate J w.r.t. μk and recompute the centroids after the cluster assignments from previous step (*M-step*). Therefore, E-step is:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \ \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from cluster's centroid.

And M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}}$$

Which translates to recomputing the centroid of each cluster to reflect the new assignments.

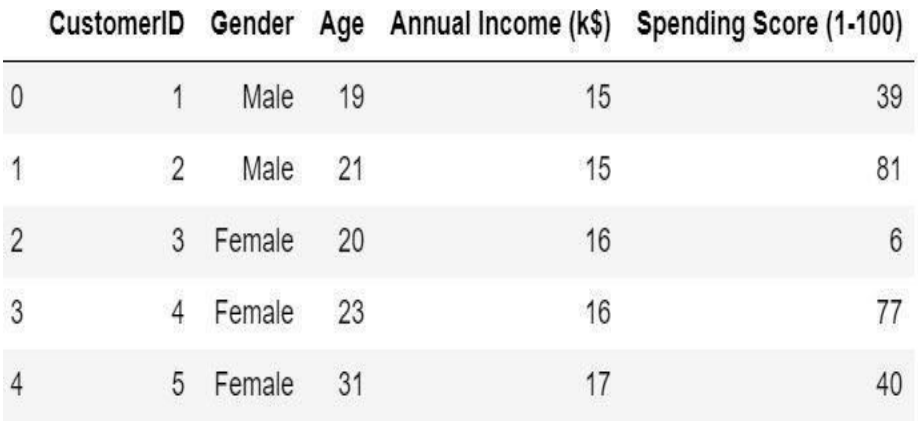
K-means is an unsupervised clustering algorithm designed to partition unlabelled data into a certain number (thats the " *K"*) of distinct groupings. In other words, k-means finds observations that share important characteristics and classifies them together into clusters. A good clustering solution is one that finds clusters such that the observations within each cluster are more similar than the clusters themselves.

There are countless examples of where this automated grouping of data can be extremely useful. For example, consider the case of creating an online advertising campaign for a brand new range of products being released to the market. While we could display a single generic advertisement to the entire population, a far better approach would be to divide the population into clusters of people who hold shared characteristics and interests displaying customised advertisements to each group. K-means is an algorithm that finds these groupings in big datasets where it is not feasible to be done by hand.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906216 |



	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

The intuition behind the algorithm is actually pretty straight forward. To begin, we choose a value for *k* (the number of clusters) and randomly choose an initial *centroid* (centre coordinates) for each cluster. We then apply a two step process:

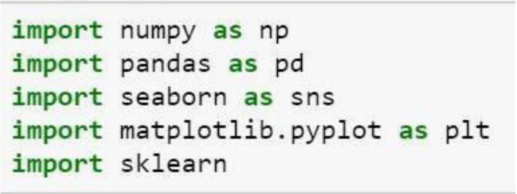
Assignment step — Assign each observation to it's nearestcentre.
Update step — Update the centroids as being the centre of their respectiveobservation.

We repeat these two steps over and over until there is no further change in the clusters. At this point the algorithm has converged and we may retrieve our final clusterings

One final key aspect of k-means returns to this concept of convergence. We previously mentioned that the k-means algorithm doesn't necessarily converge to the global minima and instead may converge to a local minima (i.e. k-means is not guaranteed to find the *best* solution). In fact, depending on which values we choose for our initial centroids we may obtain differing results.

IV. EXPERIMENTAL RESULTS

Step 1: Import packages and libraries

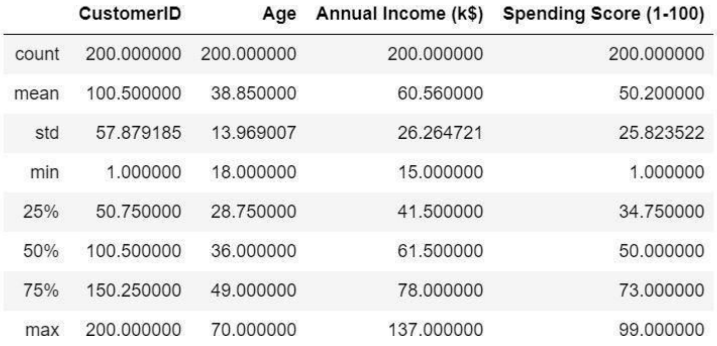


```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn
```

Figure 8.1 data collection

Step 2: collect the dataset

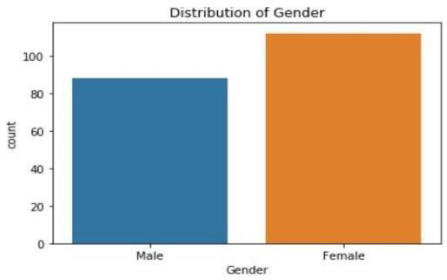
Step 3: Describe the dataset



	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Figure 8.2 describing dataset

Step 3: Plot the counter-plot



Step 4:Plotting Histograms

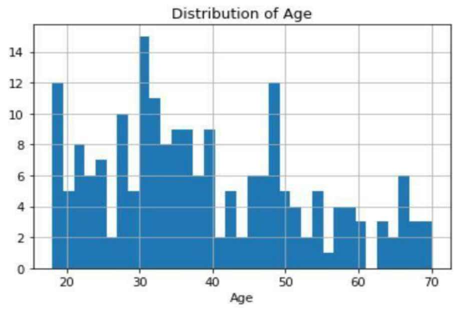


Figure 8.4:Distribution of Age histogram

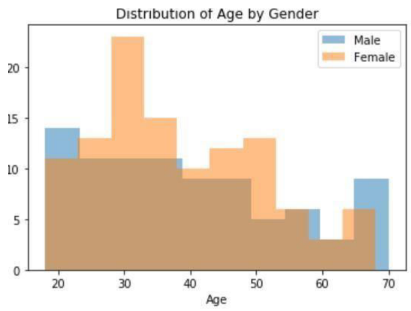


Figure 8.4.1:Distribution of Age by Gender histogram

Step 5:Plotting Scatterplots

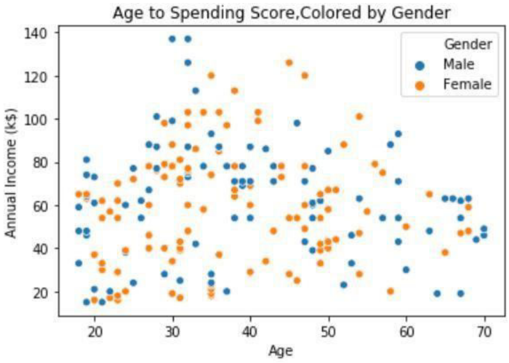


Figure 8.5:Age to Spending Score by GenderScatterplot

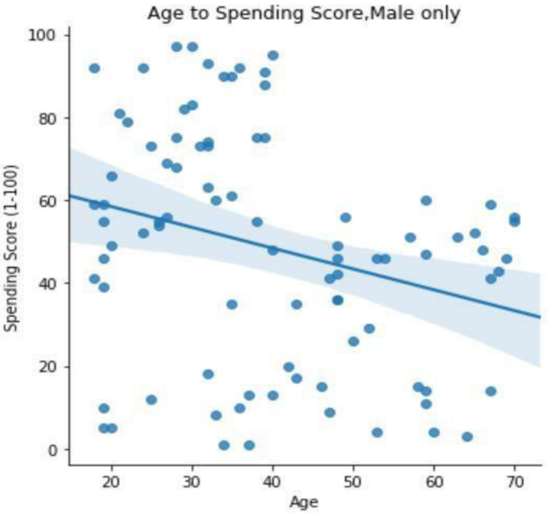


Figure 8.5.1:Age to Spending Score by Male Scatterplot

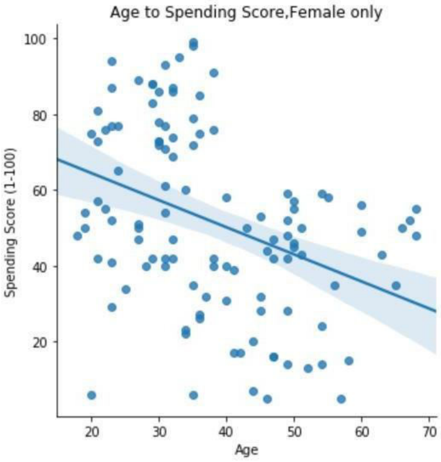


Figure 8.5.2:Age to Spending Score by Female scatterplot

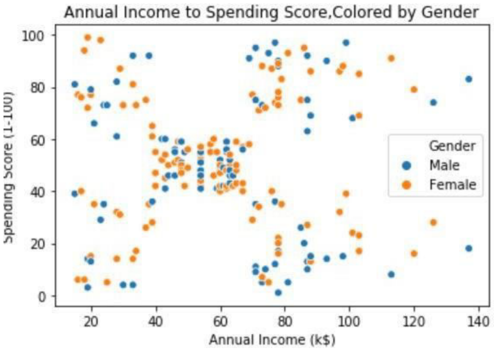
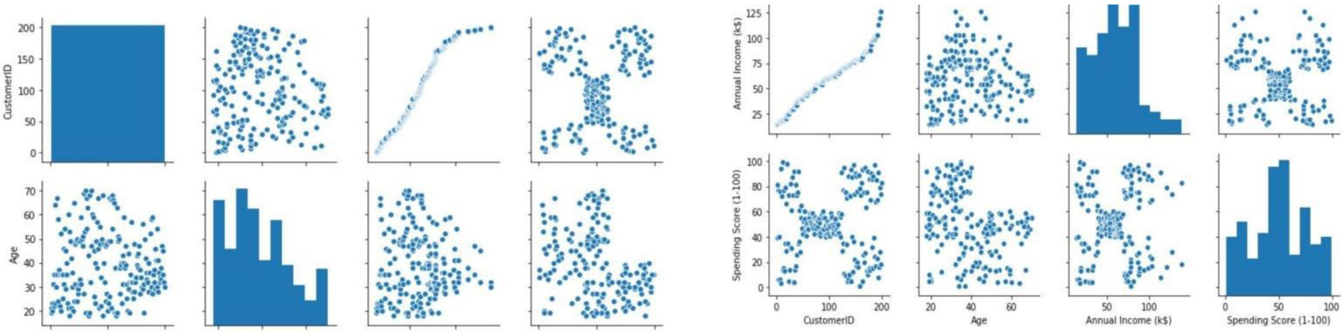


Figure 8.5.3:Annual Income to Spending Score scatterplot

Step 6:Plotting pairplots



VI. CONCLUSION

Due to increasing commercialization, consumer data is increasing exponentially. When dealing with this large magnitude of data, organizations need to make use of more efficient clustering algorithms for customer segmentation. These clustering models need to possess the capability to process this enormous data effectively. Each of the above discussed clustering algorithms come with their own set of merits and demerits. The computational speed of K-Means clustering algorithm is relatively better as compared to the hierarchical clustering algorithms as the latter require the calculation of the full proximity matrix after each iteration . K-Means clustering gives better performance for a large number of observations while hierarchical clustering has the ability to handle fewer datapoints

REFERENCES

[1] E. Ngai, L. Xiu and D. Chau, “Application of data mining techniques in customer relationship management: A literature review and classification”, Expert Systems with Applications, vol. 36, no. 2, pp. 2592-2602,2009.

[2] J. Peppard, “Customer Relationship Management (CRM) in financial services”, European Management Journal, vol. 18, no. 3, pp. 312-327,2000.

[3] A. Ansari and A. Riasi, “Taxonomy of marketing strategies using bank customers clustering”, International Journal of Business and Management, vol. 11, no. 7, pp. 106-119,2016.

[4] M. Ghzanfari, et al., “Customer segmentation in clothing exports based on clustering algorithm”, Iranian Journal of Trade Studies, vol. 14, no. 56, pp. 59-86,2010.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details