# K-MEANS CLUSTERING ALGORITHM FOR CUSTOMER SEGMENTATION

**Emmanuel, A. & Victor, S. O.**

Department of computer science, the federal polytechnic Ilaro Ogun State.
emmanuel.ayodele@federalpolyilaro.edu.ng victor.sodeinde@federalpolyilaro.end.ng

## Abstract

*Customer segmentation, a crucial tool in contemporary business, enables organizations to enhance their marketing campaigns, optimize resource allocation, and provide tailored experiences to diverse consumer groups. The K-means clustering method has shown to be an effective tool for identifying distinct client segments based on shared characteristics. This study applied K-means clustering algorithm's in customer segmentation, paying particular attention to the advantages, and potential downsides. The K-means approach reduces the sum of squared distances between data points and their associated cluster centroids in order to repeatedly divide a dataset into clusters. This technique helps discover homogeneous groups of customers who share similar behaviors, tastes, and traits in the context of customer segmentation. K-means accommodates multiple data kinds and shapes by using a variety of distance measurements and starting procedures, increasing its applicability to a variety of business settings. K-means reveals hidden patterns in client data through unsupervised learning, empowering businesses to decide on marketing tactics, product recommendations, and tailored communication. However, careful evaluation of potential difficulties is necessary for the effective deployment of the K-means algorithm for consumer segmentation. Due to the algorithm's sensitivity to initial centroid placements, using methods like K-means++ initialization may be necessary to avoid producing inferior results. The accomplishment of meaningful consumer segmentation by K-means clustering is made possible, despite hurdles, by smart parameter tweaking and validation procedures that reduce potential downsides. This research work is to implement K-means clustering algorithm for customer segmentation to predict the future consumption trend of customers.*

**Keywords:** *Customer segmentation, data analysis, unsupervised learning, marketing strategy, K-means clustering.*

## Introduction

Understanding and efficiently meeting the varied wants and preferences of consumers has become essential for long-term success in today's dynamic and data-driven corporate environment. Organizations now have an unparalleled chance to harness the power of cutting-edge analytical approaches to get useful insights thanks to the growth of digital platforms and the accumulation of enormous volumes of consumer data. (Lloyd, 1957)The K-means clustering algorithm stands out among these methods as a powerful and adaptable tool for client segmentation that enables organizations to draw insightful conclusions from their data. In order to achieve targeted measures for various consumers and provide the services that the client wants in the face of product rivalry, businesses should mine customer resources (Jagani & Chauhan, 2020). In order to develop an enterprise successfully, it is essential to first analyze the needs of the target market. This analysis should then be followed by the identification and analysis of various consumer groups within the system using customer segmentation.

At the core of contemporary marketing tactics is customer segmentation, which is the practice of grouping a heterogeneous client base into discrete and homogeneous groups based on similar traits. It gives businesses the ability to customize their interactions, communications, and services for various consumer segments, which ultimately results in increased customer happiness, brand loyalty, and revenue growth. Organizations may improve their targeting, create individualized marketing campaigns, and optimize resource allocation by finding clusters of customers who share similar habits, interests, and purchase patterns. The most significant unsupervised learning challenge is clustering. It focuses on identifying structure in a set of unlabeled data. Customer segmentation is one of the most effective business analytics approaches for analyzing consumer behavior and categorizing it (Chandrashekhar et al, 2020). Customers with comparable mean behavior are placed together into homogeneous clusters utilizing clustering algorithms (Monil, 2020). A common data mining technique, cluster analysis looks at the distribution features present in data sets to help achieve strategic objectives.

It is mostly used to analyze enterprise data information. With the aid of clustering, search results from a query can be organized into a few clusters, each of which focuses on a different component of the information being sought. The unsupervised machine learning discipline gave birth to the K-means clustering algorithm, which has achieved some notoriety for its ability to segment data into coherent groups. The basic idea behind it is to divide a dataset into K clusters, with each data point being given to the cluster with the centroid that is closest to it. The minimization of the sum of squared distances between the data points and the different cluster centroids serves as the process' guiding principle. Due to its mathematical foundation, K-means is particularly suited for jobs like customer segmentation, where the objective is to maximize the dissimilarity across clusters while grouping like consumers together (John MacQueen, 1967).

The implementation of the K-means clustering method in the area of consumer segmentation is thoroughly explored in this work. It explores the algorithm's fundamental ideas, clarifies its iterative process, and explains the algorithm's applicability to various kinds of consumer data. Additionally, it clarifies the benefits of using K-means for customer segmentation, demonstrating how firms may use this strategy to improve their decision-making procedures and cultivate lasting client relationships. However, just like with other analytical instrument, effective use of the K-means clustering algorithm necessitates a thorough comprehension of its complexities and potential drawbacks. Key factors that necessitate careful study are the sensitivity of K-means to initial centroid placements and the inherent need to establish the ideal number of clusters. K-means++ initialization and validation procedures, for example, are crucial in addressing these issues and improving the algorithm's performance (Jain,et al 1999)

In the parts that follow, we will explore the K-means clustering algorithm's theoretical foundations, real-world applications, and implementation process as we travel the terrain of consumer segmentation. We seek to provide organizations and researchers with the insights essential to harness the power of K-means for customer segmentation and, in turn, change their marketing strategies and consumer interaction techniques by highlighting both its possibilities and limitations. For greater clustering effectiveness, the pro and con of the clustering technique have also been examined. When segmenting the client, certain factors are taken into account. Geographical, demographic, psychographic, and behavioral clustering factors can be extensively categorized (Monil, 2020).

Recent years have seen a significant increase in interest from both scholars and practitioners in the use of the K-means clustering method for consumer segmentation. This section provides a thorough summary of the associated research that examines the K-means clustering algorithm's use, extensions, difficulties, and improvements in the context of customer segmentation.

The foundation for using K-means in consumer segmentation was built by earlier studies. K-means were included in a seminal study of data clustering techniques by (Jain et al., 1999) that covered their uses, benefits, and drawbacks in the context of customer segmentation. The benefits of K-means-based segmentation in e-commerce were proved by (Ullah et al., 2020), who showed how the algorithm successfully divides online consumers into different segments depending on their purchasing habits.

Improvements to the conventional K-means algorithm have been looked upon for more successful client segmentation. To address initialization and local optimum problems, (Liu et al., 2009) developed a hybrid strategy combining K-means with a genetic algorithm. (Mukhopadhyay et al., 2015) added fuzzy clustering to K-means to solve consumer preference uncertainty and enable soft assignment to multiple clusters.

Researchers have looked into combining dimensionality reduction methods with K-means clustering to handle high-dimensional customer data. To increase the effectiveness and interpretability of consumer segmentation, (Lin et al., 2018) suggested a method that combines K-means with Principal Component Analysis (PCA).

A case study on airline passengers was used to discuss consumer segmentation based on self-organizing maps by (Serpil et al., 2020). English Customer segmentation is a method of classifying customers based on attributes in common, and it has a direct impact on how satisfied customers rate businesses. By getting to know the consumer better, it gives access to the correct customer with the proper ways. Airlines must reevaluate consumer segmentations in order to deal with market changes. This involves moving away from a social-demographic strategy and toward a more complex behavioral one that considers the complete travel experience as well as the way airlines operate at every touch point. In this study, a customer segmentation that centered on two ideas, such as customer loyalty and customer return, was carried out using data from the sales of airline tickets.

Dynamic fuzzy c-means clustering algorithm-application in dynamic consumer segmentation was modified by (Munusamy & Murugesan, 2020). Managers can utilize the dynamic customer segmentation (DCS) to develop marketing plans by tracking dynamic changes that occur over time in the customer segments.

Deep learning and PCA were used in (El-Bana et al., 2020) research on a comparative dimensionality reduction study in telecom consumer segmentation. The actions of customers are recorded by telecom companies, creating a vast amount of data that can yield valuable insights into customer behavior and demands. The two key attributes of such data are their abundance of features and their extreme sparsity, both of which present difficulties for the analytics processes. In order to produce higher-quality clustering results, this research explores dimensionality reduction on a real telecom dataset and compares customer clustering in reduced and latent space to original space. There are 220 features in the original dataset that correspond to 100,000 customers.

Research into dynamic segmentation algorithms has been sparked by the temporal dynamics of consumer behavior. In order to enable dynamic client segmentation across time, (Sheng et al., 2020) suggested a K-means-based methodology that takes into account both past and present purchase habits.

(Samber et al., 2020) investigated the use of data mining for clustering and client segmentation. This essay discusses the increased competition between businesses in an effort to keep clients. Data mining is effectively assisting in having a command in e-business and other industries. By doing various analyses on the data and condensing them into informative summaries. Data management in online stores is highly challenging because the databases are large and multifaceted. Withholding the customer is the main idea. To retain clients, a two-phase clustering technique is used. A heuristic method is employed in the first stage to modify the k-means algorithm. Outliers are found using aggregative clustering. Effective data analysis is provided by this procedure for the e-commerce industry to prevent customer failure.

K-means clustering has been shown to be beneficial for consumer segmentation in research related to the industry. (Li et al., 2016) used K-means to classify mobile customers in the telecoms industry based on call detail records, providing information for resource allocation and targeted marketing. K-means clustering for e-commerce client segmentation has drawn attention due to the development of online commerce platforms. The use of K-means in e-commerce consumer clustering was investigated by (Roshan & Chandrashekara, 2019), who also discussed the implications for individualized recommendations and marketing tactics.

The segmentation of indoor consumer journeys using intuitionistic fuzzy clustering, process mining visualization was studied by (Onur et al., 2020). To understand customer demands and behaviors from the journey, there are some research and methodologies in the literature. However, due to the numerous consumers' ability to take numerous distinct paths, path analysis has a complex structure.

The body of relevant research on the use of the K-means clustering method for consumer segmentation, in conclusion, illustrates its ongoing relevance and development. Through the use of hybrid methodologies, dimensionality reduction, temporal considerations, and applications tailored to particular industries, researchers have increased its capabilities. Collectively, these studies highlight the K-means algorithm's adaptability and efficiency in revealing customer insights, guiding marketing campaigns, and encouraging individualized customer encounters.

## Methodology

Utilizing a methodical process, the K-means clustering algorithm turns raw consumer data into insights that may be used to inform marketing strategies, provide individualized customer experiences, and improve customer engagement. This methodology includes cluster evaluation, algorithm execution, data preparation, and result interpretation.

The methodology workflow can be seen below

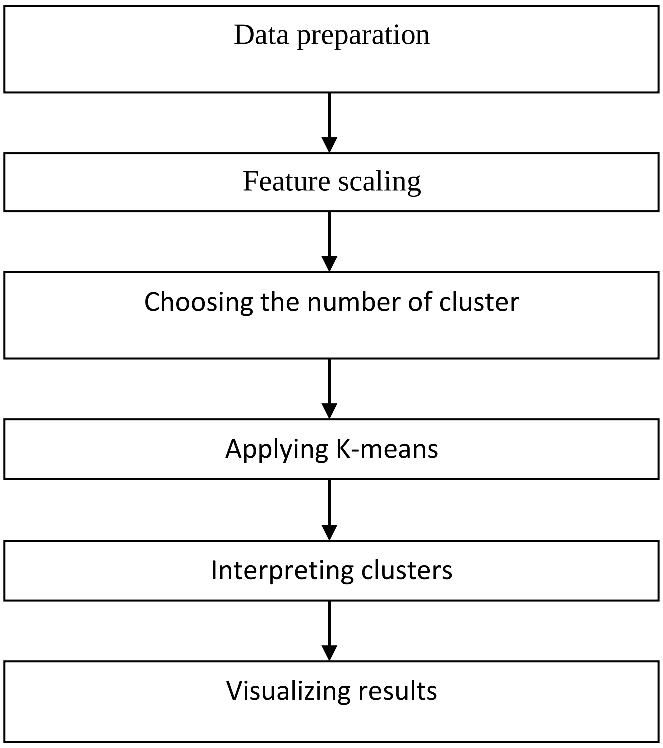


***Figure 1: Methodology Workflow***

Data preparation

- Obtain pertinent client information, including as demographics, purchasing patterns, and behavioral characteristics.
- Managing missing values and outliers, the data is cleaned.

Feature Scaling

- To ensure that the features are on the same scale, standardize or normalize them.

Choosing the Number of Clusters (k):

- Initialize K-means with the chosen value of k.
- Run the K-means algorithm on the scaled data.

Iterate through the following step until convergence or a maximum number of iterations

  - Assign each data point to the nearest centroid.
  - Recalculate the centroid based on the means of the assigned data points.

Interpreting clusters:

- Examine each cluster's attributes, such as the average feature values.
- Based on the traits of each cluster, give it a meaningful label.

Visualizing Results

- Create scatter plots to see how data points are clustered if your data is 2D or can be reduced to 2D.

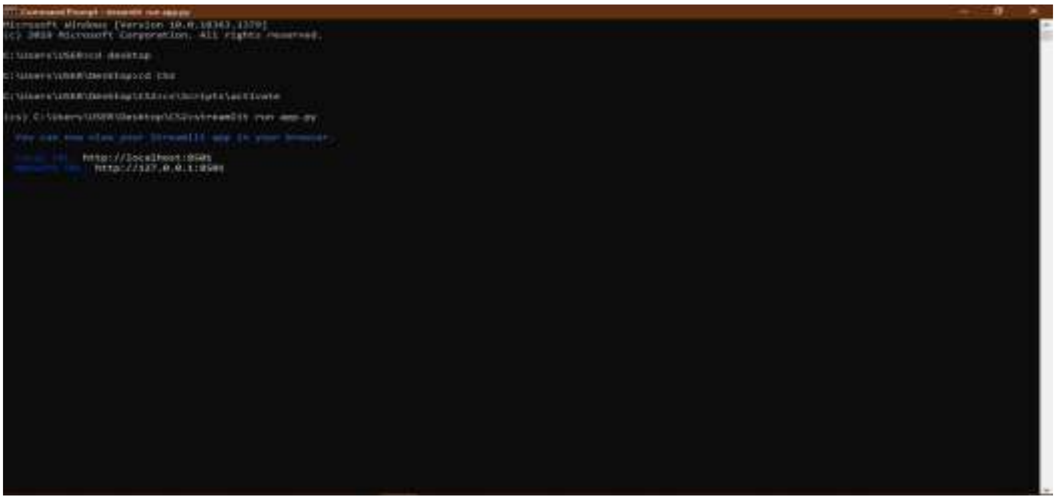
Business Insights

- To understand client preferences and behavior, interpret the cluster profiles.
- Use the segmented clusters to guide your product suggestions, marketing strategies, and personalized communications.

## Implementation

### Command Prompt

The figure below describes the starting up of the interface where the user needs to input few commands so as to bring up the interface. After inputting the commands the local URL and the Network URL comes up.

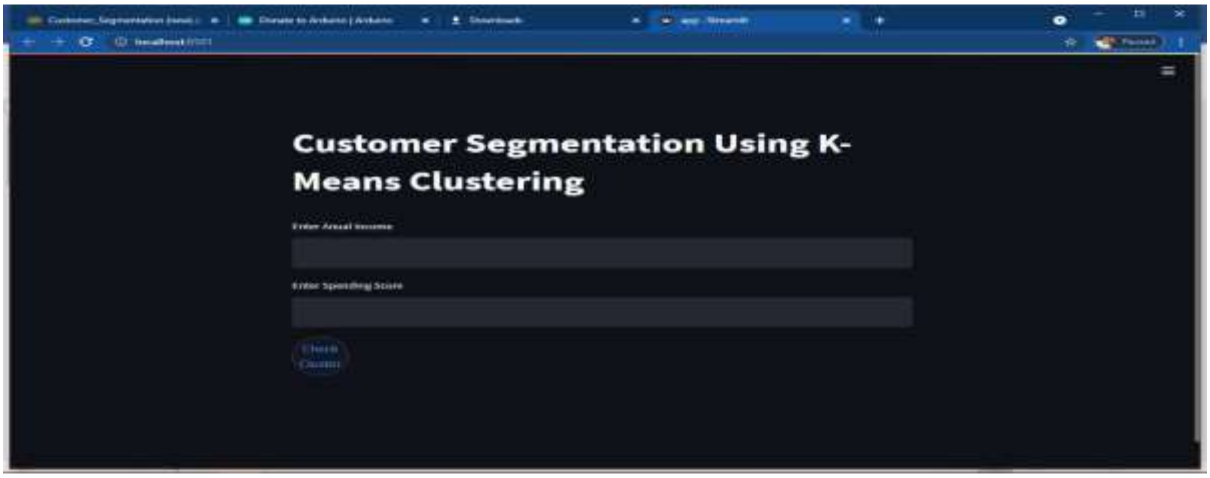


```
Command Prompt - streamlit run app.py
Microsoft Windows [Version 10.0.18363.1379]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\USER>cd desktop

C:\Users\USER\Desktop>cd CS2

C:\Users\USER\Desktop\CS2>cs\Scripts\activate

(cs) C:\Users\USER\Desktop\CS2>streamlit run app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://127.0.0.1:8501
```

*Figure 2: Command Prompt*

### Opening of the Interface

The interface displays with the help of the browser and this is after when the Local URL and Network URL comes up. This can be seen in the figure below.



Customer Segmentation Using K-Means Clustering

Enter Annual Income

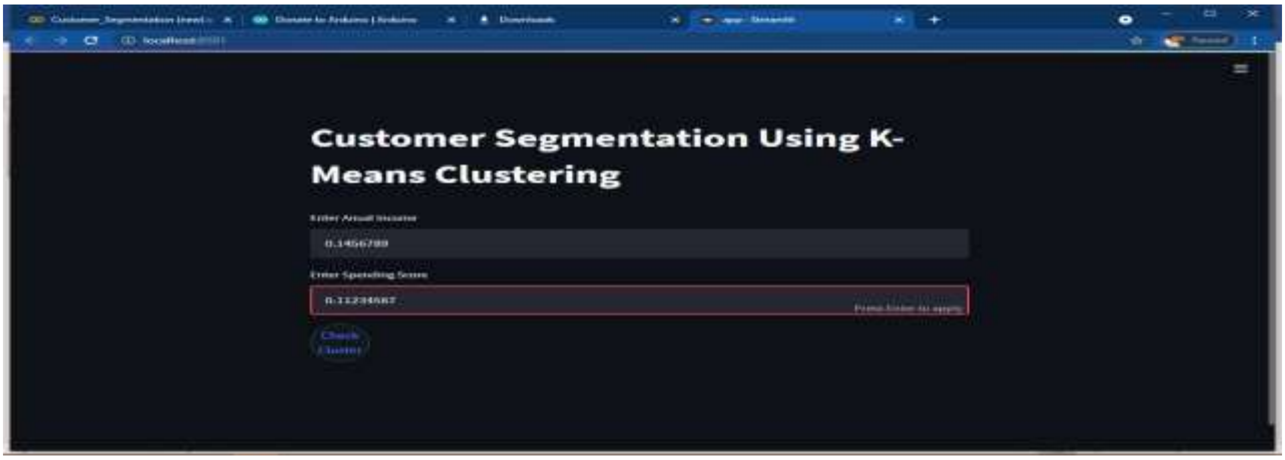
Enter Spending Score

- *Figure 3: Opening of the interface*

### Input Interface

This phase or page allows users to enter parameters which can be seen in the figure below, the parameters to input includes the annual income and the spending score.

*Figure 4: Input Interface*



Customer Segmentation Using K-Means Clustering

Enter Anual Income

0.1456789

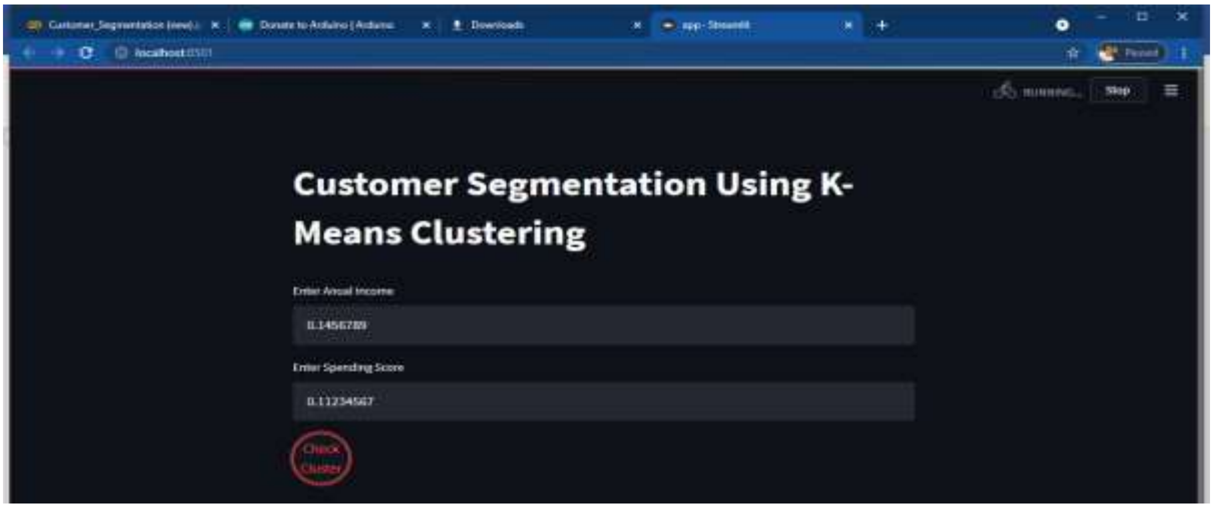
Enter Spending Score

0.11234567

Press Enter to apply

### Processing Interface

The figure below shows that the result of the parameters entered is in process after clicking the check cluster button, then the text “running” pops up.



Customer Segmentation Using K-Means Clustering

Enter Anual Income

0.1456789

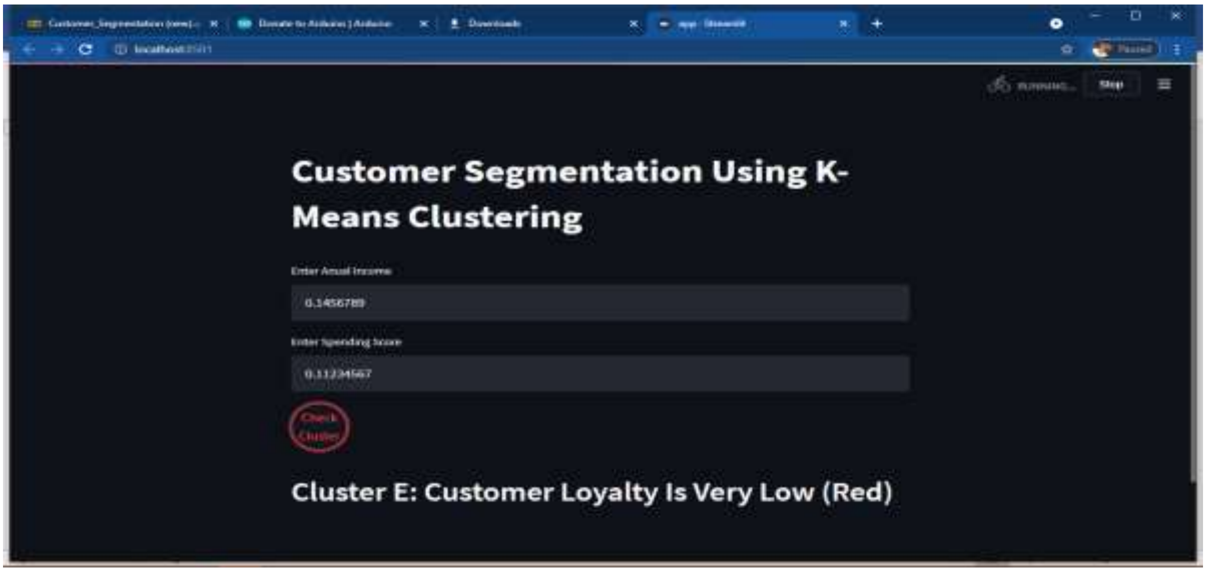
Enter Spending Score

0.11234567

*Figure 5: Processing Interface*

### Output Interface

The figure below shows the output interface, this interface is meant to display the clustered data entered by the users that is the end result of the model.



Customer Segmentation Using K-Means Clustering

Enter Anual Income

0.1456789

Enter Spending Score

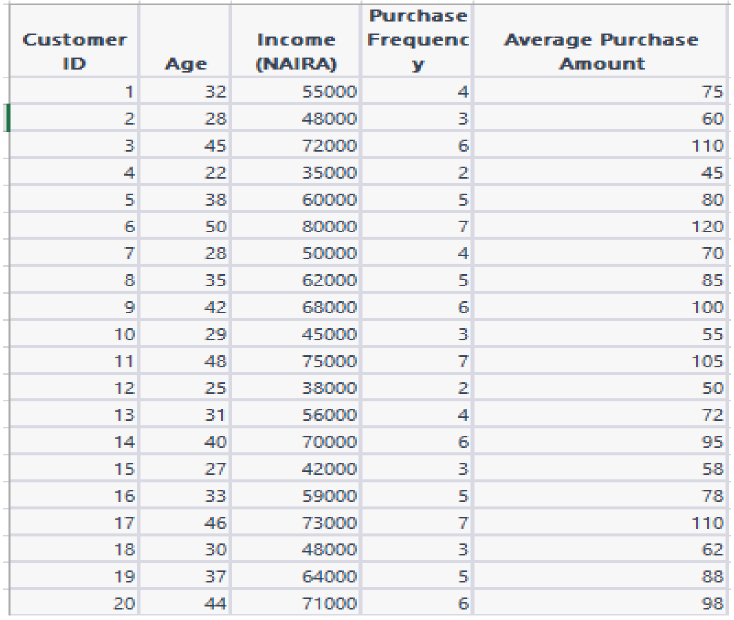
0.11234567

Cluster E: Customer Loyalty Is Very Low (Red)

*Figure 6: Output Interface*

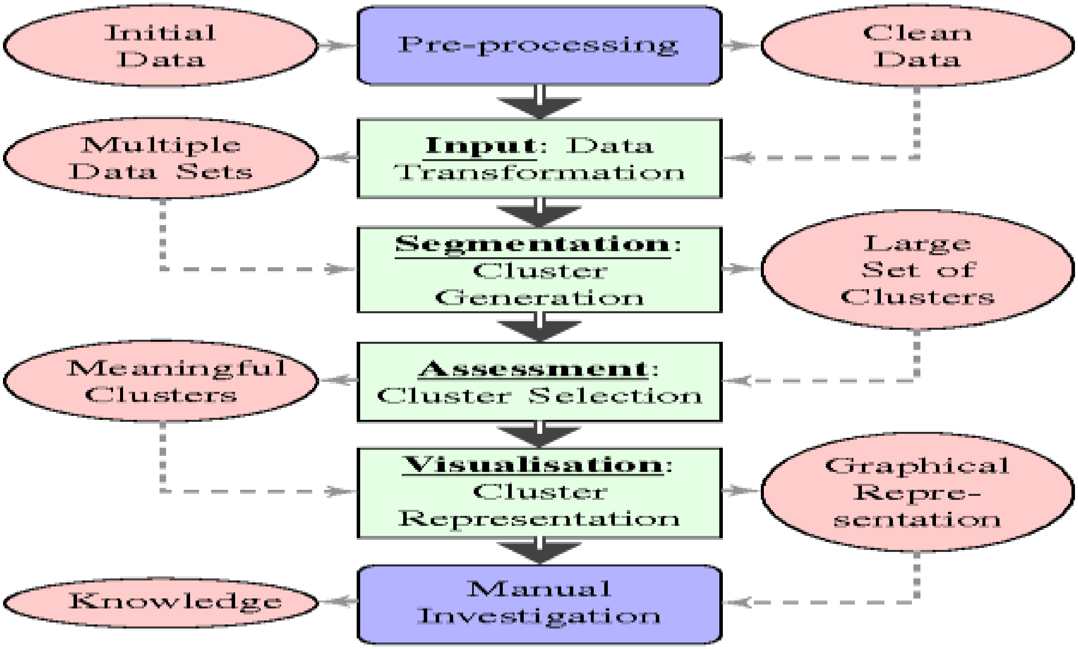
**Databases design**

The dataset to be produced for the research work was captured from the attributes shown in the database design. In particular, the developed database design is implemented using Microsoft Excel (as a CSV file). The "Mall Customers" dataset has that name. As a result, a few of the properties used are displayed below.



| Customer ID | Age | Income (NAIRA) | Purchase Frequency | Average Purchase Amount |
|---|---|---|---|---|
| 1 | 32 | 55000 | 4 | 75 |
| 2 | 28 | 48000 | 3 | 60 |
| 3 | 45 | 72000 | 6 | 110 |
| 4 | 22 | 35000 | 2 | 45 |
| 5 | 38 | 60000 | 5 | 80 |
| 6 | 50 | 80000 | 7 | 120 |
| 7 | 28 | 50000 | 4 | 70 |
| 8 | 35 | 62000 | 5 | 85 |
| 9 | 42 | 68000 | 6 | 100 |
| 10 | 29 | 45000 | 3 | 55 |
| 11 | 48 | 75000 | 7 | 105 |
| 12 | 25 | 38000 | 2 | 50 |
| 13 | 31 | 56000 | 4 | 72 |
| 14 | 40 | 70000 | 6 | 95 |
| 15 | 27 | 42000 | 3 | 58 |
| 16 | 33 | 59000 | 5 | 78 |
| 17 | 46 | 73000 | 7 | 110 |
| 18 | 30 | 48000 | 3 | 62 |
| 19 | 37 | 64000 | 5 | 88 |
| 20 | 44 | 71000 | 6 | 98 |

Sample of the dataset use for the research work


Initial Data
Pre-processing
Clean Data
Multiple Data Sets
Input: Data Transformation
Segmentation: Cluster Generation
Large Set of Clusters
Meaningful Clusters
Assessment: Cluster Selection
Visualisation: Cluster Representation
Graphical Repre-sentation
Knowledge
Manual Investigation

System architecture

## Results and Discussion

When the K-means clustering method is used for customer segmentation, it produces insightful results that enable organizations to comprehend their consumer base in detail. We explore the results, ramifications, and potential difficulties of utilizing K-means for consumer segmentation in this section.

### Segmentation

K-means clustering is excellent in dividing customer data into multiple clusters, each of which corresponds to a different customer group. Customers are grouped by the algorithm according to shared characteristics, tendencies, or preferences. The resulting clusters give firms a comprehensive breakdown of their consumer base, allowing them to see trends and distinguish between various demographics. Through K-means clustering, a merchant could, for instance, identify consumer segments that are discount-focused, occasional purchases, and frequent shoppers.

### Marketing Implications

The segmented clusters formed using K-means clustering have important marketing strategy implications. For each cluster, individual marketing strategies can be created that cater to the needs and tastes of its members. By making the experience more personalized, businesses may increase conversion rates. For instance, a travel agency might offer outdoor excursions to adventure-seeking clients in one cluster while offering relaxation-oriented discounts to those in a different cluster.

### Resource Allocation Optimization

K-means aids in resource allocation optimization by segmenting clients. Segments having a higher potential for conversion can be targeted with marketing resources more effectively. This wise distribution maximizes profits while minimizing waste. For instance, a telecoms corporation might assign customer support agents to the group of clients who inquire about services frequently.

Most machine learning algorithms perform well when analyzing the results from the above analyses, however the k-means approach had the greatest probable cluster accuracy rate of 94.5%. Another algorithm for consumer grouping is not quite perfect.

When applied to marketing problems, machine learning techniques become effective instruments for data mining in big noisy databases. With the help of these techniques, researchers have more opportunities to learn about consumer preferences while also increasing the precision of prospective and predictive models used to draw audiences to the performing arts.

The most accurate statistical representation may offer significant, statistically recognizable, stable, and responsive segments; nevertheless, those same segments may not necessarily be reachable or actionable. This will take place anytime the model suggests that it is challenging to appropriately categorize specific customers to those segments. In other words, marketable segments that are statistically recognizable may not always be fully identifiable. In these situations, the study may offer a statistical description of a sizable, precisely defined, and significant category that is also responsive to particular marketing initiatives, but to which it is challenging to assign specific customers with accuracy. When group-random effects are required by the best scientific model, this is frequently the case.

Finally, a statistical definition of a subset of clusters that are not fully recognizable or actionable (due to, for example, substantial random effects) may simultaneously reveal another group of clusters that do fit these requirements.

## CONCLUSION

The use of the K-means clustering algorithm for client segmentation represents a critical paradigm change in contemporary commercial tactics. It is clear from a thorough examination of K-means' approach, benefits, drawbacks, and related studies that it has the ability to fundamentally alter how businesses perceive, interact with, and serve their wide range of clientele. By effortlessly dividing enormous customer datasets into coherent clusters, K-means, a reliable unsupervised machine learning technique, emerges as a key enabler of this transition. These clusters act as access points to client identities, habits, preferences, and buying patterns that would otherwise be hidden in the complexities of big data.

The adaptability and simplicity of the algorithm are crucial for ensuring its wide acceptance, it becomes clear throughout the discussion of K-means clustering for consumer segmentation. Its adaptability to various data kinds, sizes, and shapes ensures its use in a wide range of businesses and situations. The algorithm is also a versatile tool for both short-term and long-term decision-making processes due to its computational effectiveness and simplicity.

The growth of the field's study and application shows how the algorithm is still relevant and evolving today. The body of related work highlights the adaptability of K-means in handling complicated segmentation circumstances, ranging from fundamental studies that established the foundations to modern inquiries into dynamic and industry-specific applications.

## References

Chandrashekhar, Y., Alexander, T., Mullasari, A., Kumbhani, D. J., Alam, S., Alexanderson, E. ... & Narula, J. (2020). Resource and infrastructure-appropriate management of ST-segment elevation myocardial infarction in low-and middle-income countries. *Circulation*, *141*(24), 2004-2025.

El-Bana, S., Al-Kabbany, A., & Sharkas, M. (2020). A multi-task pipeline with specialized streams for classification and segmentation of infection manifestations in COVID-19 scans. *PeerJ Computer Science*, *6*, e303.

Jagani, K., Oza, F. V., & Chauhan, H. (2020). Customer Segmentation and Factors Affecting Willingness to Order Private Label Brands: An E-Grocery Shopper's Perspective. In *Improving Marketing Strategies for Private Label Products* (pp. 227-253). IGI Global.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

John MacQueen (1967). A few techniques for categorizing and analyzing multivariate observations. Pages. 281-297 in Volume 1, No. 14, "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability."

Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, *113*, 107924.

Lin, L., Wang, W., & Chen, B. (2018). Leukocyte recognition with convolutional neural network. *Journal of Algorithms & Computational Technology*, *13*, 1748301818813322.

Liu, D. R., Lai, C. H., & Lee, W. J. (2009). A hybrid of sequential rules and collaborative filtering for product recommendation. *Information Sciences*, *179*(20), 3505-3519.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129- 137.

Monil, P., Darshan, P., Jecky, R., Vimarsh, C., & Bhatt, B. R. (2020). Customer segmentation using machine learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *8*(6), 2104-2108.

Mukhopadhyay, S., Paul, M., Pal, R., & De, D. (2021). Tea leaf disease detection using multi-objective image segmentation. *Multimedia Tools and Applications*, *80*, 753-771. Munusamy, S., & Murugesan, P. (2020). Modified dynamic fuzzy c-means clustering algorithm– Application in dynamic customer segmentation. *Applied Intelligence*, *50*(6), 1922-1942.

Onur Dogan, Basar Oztaysi and Carlos Fernandez-Llatas (2020). "Segmentation of Indoor Customer Paths Using Intuitionistic Fuzzy Clustering", Process Mining Visualization; Journal of Intelligent & Fuzzy Systems 38 (1), 675-684.

Samber, D. D., Ramachandran, S., Sahota, A., Naidu, S., Pruzan, A., Fayad, Z. A., & Mani, V. (2020). Segmentation of carotid arterial walls using neural networks. *World Journal of Radiology*, *12*(1), 1.

Serpil Ustebay, İlkay Yelmen and Metin Zontul (2020). "Customer Segmentation Based onSelf-Organizing Maps: A Case Study on Airline Passengers", Journal of Aeronautics & Space Technologies/Havacilik ve Uzay Teknolojileri Dergisi 13 (2).

Sheng, W., Wang, X., Wang, Z., Li, Q., Zheng, Y., & Chen, S. (2020). A differential evolution algorithm with adaptive niching and k-means operation for data clustering. *IEEE Transactions on Cybernetics*, *52*(7), 6181-6195.

Ullah, I., Boreli, R., & Kanhere, S. S. (2020). Privacy in targeted advertising: A survey. *arXiv preprint arXiv:2009.06861*.