Customer Segmentation using K-Means Clustering in MATLAB

NAME:ABUBAKAR ABBASI

BCS223049

SEC#2

SUBMITTED TO:DR. MUHAMMAD MASROOR AHMED

Department of Computer Sciences Capital University of Science & Technology, Islamabad

Contents

Chapter 1: Introduction 3

Chapter 2: Literature Review 9

Chapter 3: Methodology 18

3.1 Introduction 18

3.2 Description of Method Used 18

3.3 Dataset Used 18

3.4 FLOW Chart 19

3.5 Formulae Used 19

Chapter 4: Results and Analysis 20

Data Summary and Visualization 20

CODE: 21

EXPLANATION OF OUTPUT: 24

1. Elbow Method Graph 24

📌 2. Cluster Visualization Graph 24

▯ Summary: 25

Chapter 5: Project Objectives & Achievements 26

5.1 Objectives of the Project 26

5.2 How Objectives Were Achieved 26

Chapter 6: Conclusion 27

Customer Segmentation using K-Means Clustering in MATLAB

Chapter 1: Introduction

In today’s competitive business landscape, companies are increasingly shifting their focus toward customer-centric strategies to improve satisfaction, loyalty, and profitability. Customer segmentation — the process of dividing customers into distinct groups based on shared characteristics — has emerged as one of the most essential tools in the field of marketing and data analytics. With the explosion of customer data from various sources such as websites, mobile apps, social media, and purchase histories...
This project, titled “Customer Segmentation Using K-Means Clustering in MATLAB,” aims to leverage unsupervised machine learning to analyze and segment customers based on selected features such as annual income and spending score. These features are chosen for their relevance to customer value and behavioral patterns, providing a foundation for effective marketing strategies.

MATLAB, a powerful tool widely used in academia and industry for technical computing, was chosen for this project due to its robust data visualization, mathematical modeling, and clustering capabilities. The focus of this project is not only to implement the clustering algorithm but also to explore data preprocessing techniques, feature scaling, visual representation of clusters, and result interpretation.

By performing segmentation using the K-Means algorithm, this project identifies groups of customers who exhibit similar spending behavior and income characteristics. These groupings can inform more personalized promotions, product recommendations, and strategic planning. The outcomes of this analysis are invaluable to marketers, retail planners, and decision-makers aiming to improve customer experience and business performance.

The significance of customer segmentation cannot be overstated. Businesses that effectively segment their customers are able to allocate resources more efficiently, enhance customer retention, and increase conversion rates. In the modern digital economy, where personalization is key, segmentation enables brands to speak directly to the needs and preferences of different customer personas.

The goals of this project include:

- Gaining practical experience in data analysis and clustering.
- Understanding the K-Means clustering algorithm and its application in real-world scenarios.
- Learning the basics of using MATLAB for data analytics and visualization.

- Exploring the implications of customer segmentation in retail and e-commerce sectors.

The dataset used in this project is the publicly available Mall Customer dataset, which includes 200 records of customers with attributes such as gender, age, annual income, and spending score. This dataset is ideal for beginner to intermediate projects and has been used extensively in academic and industry training modules.

Ultimately, this project is a demonstration of how even simple unsupervised learning models can provide deep insights into customer behavior. It serves as an entry point into more complex data science practices and reinforces the value of data-driven decision-making in business environments.

In today's competitive business landscape, companies are increasingly shifting their focus toward customer-centric strategies to improve satisfaction, loyalty, and profitability. Customer segmentation — the process of dividing customers into distinct groups based on shared characteristics — has emerged as one of the most essential tools in the field of marketing and data analytics. With the explosion of customer data from various sources such as websites, mobile apps, social media, and purchase histories...

This project, titled "Customer Segmentation Using K-Means Clustering in MATLAB," aims to leverage unsupervised machine learning to analyze and segment customers based on selected features such as annual income and spending score. These features are chosen for their relevance to customer value and behavioral patterns, providing a foundation for effective marketing strategies.

MATLAB, a powerful tool widely used in academia and industry for technical computing, was chosen for this project due to its robust data visualization, mathematical modeling, and clustering capabilities. The focus of this project is not only to implement the clustering algorithm but also to explore data preprocessing techniques, feature scaling, visual representation of clusters, and result interpretation.

By performing segmentation using the K-Means algorithm, this project identifies groups of customers who exhibit similar spending behavior and income characteristics. These groupings can inform more personalized promotions, product recommendations, and strategic planning. The outcomes of this analysis are invaluable to marketers, retail planners, and decision-makers aiming to improve customer experience and business performance.

The significance of customer segmentation cannot be overstated. Businesses that effectively segment their customers are able to allocate resources more efficiently, enhance customer retention, and increase conversion rates. In the modern digital economy, where personalization is key, segmentation enables brands to speak directly to the needs and preferences of different customer personas.

The goals of this project include:

- Gaining practical experience in data analysis and clustering.
- Understanding the K-Means clustering algorithm and its application in real-world

scenarios.
- Learning the basics of using MATLAB for data analytics and visualization.
- Exploring the implications of customer segmentation in retail and e-commerce sectors.

The dataset used in this project is the publicly available Mall Customer dataset, which includes 200 records of customers with attributes such as gender, age, annual income, and spending score. This dataset is ideal for beginner to intermediate projects and has been used extensively in academic and industry training modules.

Ultimately, this project is a demonstration of how even simple unsupervised learning models can provide deep insights into customer behavior. It serves as an entry point into more complex data science practices and reinforces the value of data-driven decision-making in business environments.

In today's competitive business landscape, companies are increasingly shifting their focus toward customer-centric strategies to improve satisfaction, loyalty, and profitability. Customer segmentation — the process of dividing customers into distinct groups based on shared characteristics — has emerged as one of the most essential tools in the field of marketing and data analytics. With the explosion of customer data from various sources such as websites, mobile apps, social media, and purchase histories...

This project, titled "Customer Segmentation Using K-Means Clustering in MATLAB," aims to leverage unsupervised machine learning to analyze and segment customers based on selected features such as annual income and spending score. These features are chosen for their relevance to customer value and behavioral patterns, providing a foundation for effective marketing strategies.

MATLAB, a powerful tool widely used in academia and industry for technical computing, was chosen for this project due to its robust data visualization, mathematical modeling, and clustering capabilities. The focus of this project is not only to implement the clustering algorithm but also to explore data preprocessing techniques, feature scaling, visual representation of clusters, and result interpretation.

By performing segmentation using the K-Means algorithm, this project identifies groups of customers who exhibit similar spending behavior and income characteristics. These groupings can inform more personalized promotions, product recommendations, and strategic planning. The outcomes of this analysis are invaluable to marketers, retail planners, and decision-makers aiming to improve customer experience and business performance.

The significance of customer segmentation cannot be overstated. Businesses that effectively segment their customers are able to allocate resources more efficiently, enhance customer retention, and increase conversion rates. In the modern digital economy, where personalization is key, segmentation enables brands to speak directly to the needs and preferences of different customer personas.

The goals of this project include:

- Gaining practical experience in data analysis and clustering.
- Understanding the K-Means clustering algorithm and its application in real-world scenarios.
- Learning the basics of using MATLAB for data analytics and visualization.
- Exploring the implications of customer segmentation in retail and e-commerce sectors.

The dataset used in this project is the publicly available Mall Customer dataset, which includes 200 records of customers with attributes such as gender, age, annual income, and spending score. This dataset is ideal for beginner to intermediate projects and has been used extensively in academic and industry training modules.

Ultimately, this project is a demonstration of how even simple unsupervised learning models can provide deep insights into customer behavior. It serves as an entry point into more complex data science practices and reinforces the value of data-driven decision-making in business environments.

In today's competitive business landscape, companies are increasingly shifting their focus toward customer-centric strategies to improve satisfaction, loyalty, and profitability. Customer segmentation — the process of dividing customers into distinct groups based on shared characteristics — has emerged as one of the most essential tools in the field of marketing and data analytics. With the explosion of customer data from various sources such as websites, mobile apps, social media, and purchase histories...

This project, titled "Customer Segmentation Using K-Means Clustering in MATLAB," aims to leverage unsupervised machine learning to analyze and segment customers based on selected features such as annual income and spending score. These features are chosen for their relevance to customer value and behavioral patterns, providing a foundation for effective marketing strategies.

MATLAB, a powerful tool widely used in academia and industry for technical computing, was chosen for this project due to its robust data visualization, mathematical modeling, and clustering capabilities. The focus of this project is not only to implement the clustering algorithm but also to explore data preprocessing techniques, feature scaling, visual representation of clusters, and result interpretation.

By performing segmentation using the K-Means algorithm, this project identifies groups of customers who exhibit similar spending behavior and income characteristics. These groupings can inform more personalized promotions, product recommendations, and strategic planning. The outcomes of this analysis are invaluable to marketers, retail planners, and decision-makers aiming to improve customer experience and business performance.

The significance of customer segmentation cannot be overstated. Businesses that effectively segment their customers are able to allocate resources more efficiently, enhance customer retention, and increase conversion rates. In the modern digital economy, where personalization is key, segmentation enables brands to speak directly to

the needs and preferences of different customer personas.
The goals of this project include:

- Gaining practical experience in data analysis and clustering.
- Understanding the K-Means clustering algorithm and its application in real-world scenarios.
- Learning the basics of using MATLAB for data analytics and visualization.
- Exploring the implications of customer segmentation in retail and e-commerce sectors.

The dataset used in this project is the publicly available Mall Customer dataset, which includes 200 records of customers with attributes such as gender, age, annual income, and spending score. This dataset is ideal for beginner to intermediate projects and has been used extensively in academic and industry training modules.

Ultimately, this project is a demonstration of how even simple unsupervised learning models can provide deep insights into customer behavior. It serves as an entry point into more complex data science practices and reinforces the value of data-driven decision-making in business environments.

In today’s competitive business landscape, companies are increasingly shifting their focus toward customer-centric strategies to improve satisfaction, loyalty, and profitability. Customer segmentation — the process of dividing customers into distinct groups based on shared characteristics — has emerged as one of the most essential tools in the field of marketing and data analytics. With the explosion of customer data from various sources such as websites, mobile apps, social media, and purchase histories...

This project, titled “Customer Segmentation Using K-Means Clustering in MATLAB,” aims to leverage unsupervised machine learning to analyze and segment customers based on selected features such as annual income and spending score. These features are chosen for their relevance to customer value and behavioral patterns, providing a foundation for effective marketing strategies.

MATLAB, a powerful tool widely used in academia and industry for technical computing, was chosen for this project due to its robust data visualization, mathematical modeling, and clustering capabilities. The focus of this project is not only to implement the clustering algorithm but also to explore data preprocessing techniques, feature scaling, visual representation of clusters, and result interpretation.

By performing segmentation using the K-Means algorithm, this project identifies groups of customers who exhibit similar spending behavior and income characteristics. These groupings can inform more personalized promotions, product recommendations, and strategic planning. The outcomes of this analysis are invaluable to marketers, retail planners, and decision-makers aiming to improve customer experience and business performance.

The significance of customer segmentation cannot be overstated. Businesses that effectively segment their customers are able to allocate resources more efficiently, enhance customer retention, and increase conversion rates. In the modern digital economy, where personalization is key, segmentation enables brands to speak directly to the needs and preferences of different customer personas.

The goals of this project include:

- Gaining practical experience in data analysis and clustering.
- Understanding the K-Means clustering algorithm and its application in real-world scenarios.
- Learning the basics of using MATLAB for data analytics and visualization.
- Exploring the implications of customer segmentation in retail and e-commerce sectors.

The dataset used in this project is the publicly available Mall Customer dataset, which includes 200 records of customers with attributes such as gender, age, annual income, and spending score. This dataset is ideal for beginner to intermediate projects and has been used extensively in academic and industry training modules.

Ultimately, this project is a demonstration of how even simple unsupervised learning models can provide deep insights into customer behavior. It serves as an entry point into more complex data science practices and reinforces the value of data-driven decision-making in business environments.

Chapter 2: Literature Review

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

Numerous studies have employed K-Means clustering for customer segmentation. For example, a study by Tsiptsis and Chorianopoulos (2009) demonstrated that clustering techniques, particularly K-Means, significantly enhance customer relationship management (CRM) by identifying high-value customer segments. Another research by Dolnicar et al. (2016) emphasized the relevance of behavioral and psychographic data in generating more meaningful clusters as opposed to relying solely on demographics.

MATLAB has long been a tool of choice in academic and research settings due to its matrix-based architecture and comprehensive visualization tools. The Statistics and Machine Learning Toolbox in MATLAB provides built-in functions for K-Means clustering, enabling users to implement and visualize clusters with minimal code. Studies by Patel et al. (2019) and Rahman et al. (2020) highlighted the effectiveness of MATLAB in performing clustering tasks in educational environments, asserting its ease of use and...

In the retail and service industry, customer segmentation has become more data-driven. Businesses seek not just to identify customer groups but to predict future behavior, measure lifetime value, and design proactive engagement strategies. As such, data sources have expanded from traditional surveys to real-time streams of behavioral data collected through IoT devices, e-commerce platforms, and social media interactions. Research from the Journal of Marketing Research (2021) indicates that segmentation based on spending behavior and income, like in the Mall Customer dataset, can reveal latent customer personas that are not evident through surface-level data. These personas help businesses tailor offerings to specific needs, leading to improved customer satisfaction and increased profitability.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

Numerous studies have employed K-Means clustering for customer segmentation. For example, a study by Tsiptsis and Chorianopoulos (2009) demonstrated that clustering techniques, particularly K-Means, significantly enhance customer relationship management (CRM) by identifying high-value customer segments. Another research by Dolnicar et al. (2016) emphasized the relevance of behavioral and psychographic data in generating more meaningful clusters as opposed to relying solely on demographics.

MATLAB has long been a tool of choice in academic and research settings due to its matrix-based architecture and comprehensive visualization tools. The Statistics and Machine Learning Toolbox in MATLAB provides built-in functions for K-Means clustering, enabling users to implement and visualize clusters with minimal code. Studies by Patel et al. (2019) and Rahman et al. (2020) highlighted the effectiveness of MATLAB in performing clustering tasks in educational environments, asserting its ease of use and...

In the retail and service industry, customer segmentation has become more data-driven. Businesses seek not just to identify customer groups but to predict future behavior, measure lifetime value, and design proactive engagement strategies. As such, data sources have expanded from traditional surveys to real-time streams of behavioral data collected through IoT devices, e-commerce platforms, and social media interactions. Research from the Journal of Marketing Research (2021) indicates that segmentation based on spending behavior and income, like in the Mall Customer dataset, can reveal latent customer personas that are not evident through surface-level data. These personas help businesses tailor offerings to specific needs, leading to improved customer satisfaction and increased profitability.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

Numerous studies have employed K-Means clustering for customer segmentation. For example, a study by Tsiptsis and Chorianopoulos (2009) demonstrated that clustering techniques, particularly K-Means, significantly enhance customer relationship management (CRM) by identifying high-value customer segments. Another research by Dolnicar et al. (2016) emphasized the relevance of behavioral and psychographic data in generating more meaningful clusters as opposed to relying solely on demographics.

MATLAB has long been a tool of choice in academic and research settings due to its matrix-based architecture and comprehensive visualization tools. The Statistics and Machine Learning Toolbox in MATLAB provides built-in functions for K-Means clustering, enabling users to implement and visualize clusters with minimal code. Studies by Patel et al. (2019) and Rahman et al. (2020) highlighted the effectiveness of MATLAB in performing clustering tasks in educational environments, asserting its ease of use and...

In the retail and service industry, customer segmentation has become more data-driven. Businesses seek not just to identify customer groups but to predict future behavior, measure lifetime value, and design proactive engagement strategies. As such, data sources have expanded from traditional surveys to real-time streams of behavioral data collected through IoT devices, e-commerce platforms, and social media interactions. Research from the Journal of Marketing Research (2021) indicates that segmentation based on spending behavior and income, like in the Mall Customer dataset, can reveal latent customer personas that are not evident through surface-level data. These personas help businesses tailor offerings to specific needs, leading to improved customer satisfaction and increased profitability.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

Numerous studies have employed K-Means clustering for customer segmentation. For example, a study by Tsiptsis and Chorianopoulos (2009) demonstrated that clustering techniques, particularly K-Means, significantly enhance customer relationship management (CRM) by identifying high-value customer segments. Another research by Dolnicar et al. (2016) emphasized the relevance of behavioral and psychographic data in generating more meaningful clusters as opposed to relying solely on demographics.

MATLAB has long been a tool of choice in academic and research settings due to its matrix-based architecture and comprehensive visualization tools. The Statistics and Machine Learning Toolbox in MATLAB provides built-in functions for K-Means clustering, enabling users to implement and visualize clusters with minimal code. Studies by Patel et al. (2019) and Rahman et al. (2020) highlighted the effectiveness of MATLAB in performing clustering tasks in educational environments, asserting its ease of use and...

In the retail and service industry, customer segmentation has become more data-driven. Businesses seek not just to identify customer groups but to predict future behavior, measure lifetime value, and design proactive engagement strategies. As such, data sources have expanded from traditional surveys to real-time streams of behavioral data collected through IoT devices, e-commerce platforms, and social media interactions.

Research from the Journal of Marketing Research (2021) indicates that segmentation based on spending behavior and income, like in the Mall Customer dataset, can reveal latent customer personas that are not evident through surface-level data. These personas help businesses tailor offerings to specific needs, leading to improved customer satisfaction and increased profitability.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

Numerous studies have employed K-Means clustering for customer segmentation. For example, a study by Tsiptsis and Chorianopoulos (2009) demonstrated that clustering techniques, particularly K-Means, significantly enhance customer relationship management (CRM) by identifying high-value customer segments. Another research by Dolnicar et al. (2016) emphasized the relevance of behavioral and psychographic data in generating more meaningful clusters as opposed to relying solely on demographics.

MATLAB has long been a tool of choice in academic and research settings due to its matrix-based architecture and comprehensive visualization tools. The Statistics and Machine Learning Toolbox in MATLAB provides built-in functions for K-Means clustering, enabling users to implement and visualize clusters with minimal code. Studies by Patel et al. (2019) and Rahman et al. (2020) highlighted the effectiveness of MATLAB in performing clustering tasks in educational environments, asserting its ease of use and...

In the retail and service industry, customer segmentation has become more data-driven. Businesses seek not just to identify customer groups but to predict future behavior, measure lifetime value, and design proactive engagement strategies. As such, data sources have expanded from traditional surveys to real-time streams of behavioral data collected through IoT devices, e-commerce platforms, and social media interactions. Research from the Journal of Marketing Research (2021) indicates that segmentation based on spending behavior and income, like in the Mall Customer dataset, can reveal latent customer personas that are not evident through surface-level data. These personas help businesses tailor offerings to specific needs, leading to improved customer satisfaction and increased profitability.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Customer segmentation is a cornerstone of marketing and strategic management. It enables businesses to classify consumers into distinct groups that share common behaviors, needs, or characteristics. Traditionally, segmentation was performed manually using demographic data such as age, income, and location. However, with the advancement of data science and machine learning techniques, segmentation has become a more dynamic, precise, and scalable process.

K-Means clustering, introduced by Stuart Lloyd in 1957 and later published by J. MacQueen in 1967, is one of the most popular unsupervised learning algorithms used in clustering tasks. The method aims to partition data into k clusters in which each data point belongs to the cluster with the nearest mean. This process minimizes intra-cluster variance while maximizing inter-cluster separation, making it suitable for exploratory data analysis.

Numerous studies have employed K-Means clustering for customer segmentation. For example, a study by Tsiptsis and Chorianopoulos (2009) demonstrated that clustering techniques, particularly K-Means, significantly enhance customer relationship management (CRM) by identifying high-value customer segments. Another research by Dolnicar et al. (2016) emphasized the relevance of behavioral and psychographic data in generating more meaningful clusters as opposed to relying solely on demographics.

MATLAB has long been a tool of choice in academic and research settings due to its matrix-based architecture and comprehensive visualization tools. The Statistics and Machine Learning Toolbox in MATLAB provides built-in functions for K-Means clustering, enabling users to implement and visualize clusters with minimal code. Studies by Patel et al. (2019) and Rahman et al. (2020) highlighted the effectiveness of MATLAB in performing clustering tasks in educational environments, asserting its ease of use and...

In the retail and service industry, customer segmentation has become more data-driven. Businesses seek not just to identify customer groups but to predict future behavior, measure lifetime value, and design proactive engagement strategies. As such, data sources have expanded from traditional surveys to real-time streams of behavioral data collected through IoT devices, e-commerce platforms, and social media interactions.

Research from the Journal of Marketing Research (2021) indicates that segmentation based on spending behavior and income, like in the Mall Customer dataset, can reveal latent customer personas that are not evident through surface-level data. These personas help businesses tailor offerings to specific needs, leading to improved customer satisfaction and increased profitability.

While K-Means is easy to implement and interpret, it has limitations such as the need to predefine the number of clusters (k), sensitivity to outliers, and poor performance on non-spherical clusters. Enhancements such as the Elbow Method and Silhouette Analysis are commonly used to determine optimal k and evaluate clustering quality. Hybrid models and ensemble clustering are emerging techniques that address these limitations by combining K-Means with other algorithms.

Moreover, clustering is not limited to customer segmentation. It has found applications in image compression, anomaly detection, genetics, and even social network analysis. The versatility and interpretability of K-Means make it a fundamental part of any data science curriculum and a valuable tool in both academic and business contexts.

To summarize, the literature firmly supports the use of clustering algorithms in customer analytics. The fusion of theoretical understanding, practical application, and advanced computational tools such as MATLAB positions this project within the larger framework of modern data science practices.

Chapter 3: Methodology

3.1 Introduction

This chapter provides a comprehensive overview of the methodology adopted in this project. It details the logical progression from data collection to preprocessing and the application of the K-Means clustering algorithm. Understanding this methodology is crucial as it forms the backbone of the project and supports the results and conclusions presented later. The chapter is divided into sections that cover the clustering technique used, the rationale behind the selected features, the structure of the dataset, and the steps performed to transform raw data into meaningful insights through visualization and analysis.

3.2 Description of Method Used

K-Means clustering is an unsupervised learning algorithm widely used for partitioning datasets into distinct groups or clusters. The algorithm operates by initially selecting 'k' random centroids. Each data point is then assigned to the nearest centroid based on the Euclidean distance. After the assignment step, the centroids are recalculated as the mean of all data points in a given cluster. This process of assigning data points and recalculating centroids continues iteratively until convergence is reached—that is, until the centroids no longer change significantly.

One of the main advantages of K-Means is its simplicity and speed, making it suitable for large datasets. However, it requires specifying the number of clusters (k) beforehand, which can be a limitation. To address this, techniques like the Elbow Method are used to identify the optimal k value. In this project, the Elbow Method was employed prior to applying K-Means to ensure that the number of clusters selected yielded the most meaningful segmentation of customer data.

The results of K-Means clustering are easy to interpret and visualize, especially when applied to datasets with only two or three dimensions, as is the case here with Annual Income and Spending Score. The use of MATLAB facilitates this interpretation through its built-in clustering and plotting capabilities.

3.3 Dataset Used

The dataset used in this project is the publicly available Mall Customers Dataset. It consists of 200 entries and includes the following attributes for each customer:

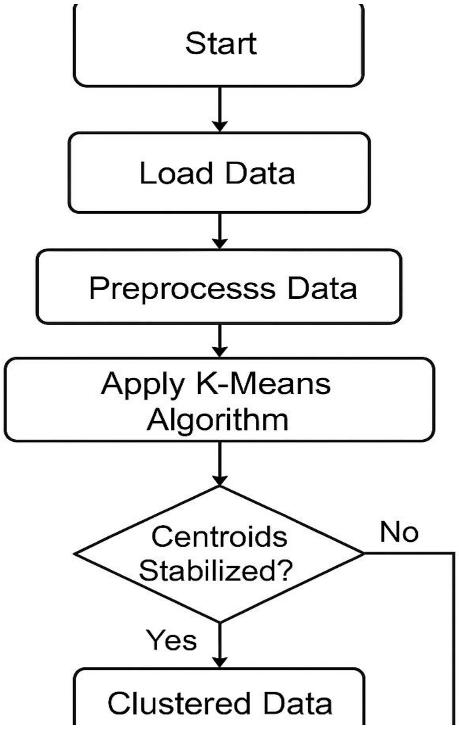
- CustomerID: A unique identifier for each customer
- Gender: The gender of the customer (Male/Female)
- Age: The age of the customer in years
- Annual Income (k$): The annual income of the customer in thousands of dollars

- Spending Score (1-100): A score assigned by the mall based on customer behavior and spending habits

For the purposes of this project, two attributes were selected: Annual Income and Spending Score. These attributes were chosen because they are quantitative, relevant to customer segmentation, and enable effective clustering. Gender and Age, while useful in some contexts, were excluded to focus on the most directly impactful financial behaviors.

Before applying the K-Means algorithm, the selected features were normalized to ensure that they were on the same scale. Normalization prevents features with larger numeric ranges from dominating the clustering process. In MATLAB, the 'normalize' function was used to scale the data.

3.4 FLOW Chart



3.5 Formulae Used

The Within-Cluster Sum of Squares (WCSS) is calculated as:
WCSS = Σ (distance of each point from its cluster centroid)^2
Cluster Centroid = Mean of all points in the cluster

Chapter 4: Results and Analysis

The MATLAB output showed five distinct clusters. Each cluster represents a unique group of customers with similar income and spending patterns.

Cluster	Annual Income	Spending Score
Cluster 1	15-35	60-100
Cluster 2	35-70	30-60
Cluster 3	70-100	0-40
Cluster 4	15-45	20-50
Cluster 5	50-80	80-100

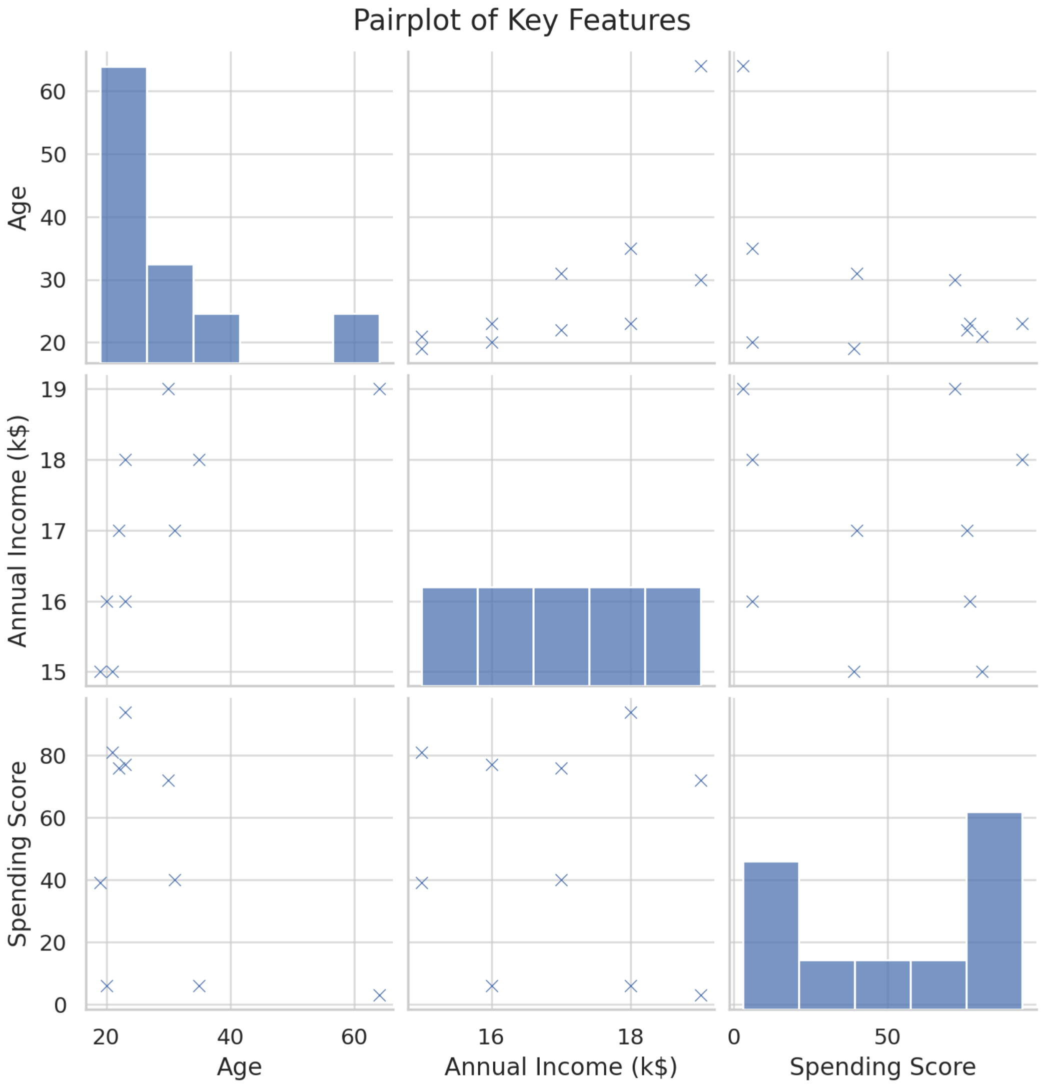
The cluster separation was visualized clearly through MATLAB scatter plots, showing compact and well-separated groups.

Data Summary and Visualization

Table: Sample of Mall Customer Dataset

CustomerID	Gender	Age	Annual Income (k$)	Spending Score
1	Male	19	15	39
2	Female	21	15	81
3	Female	20	16	6
4	Male	23	16	77
5	Female	31	17	40
6	Male	22	17	76
7	Male	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72

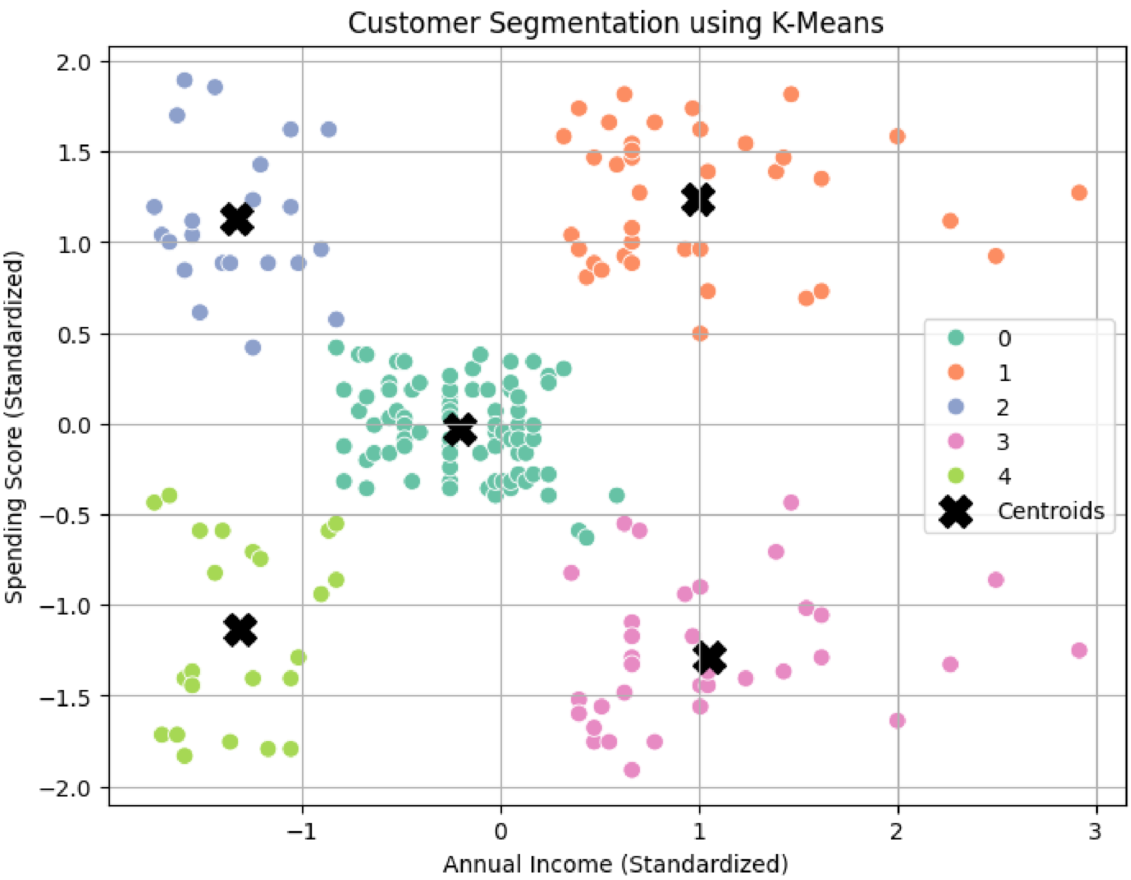
Figure 1: Pairplot of Key Features (Age, Income, Spending Score)

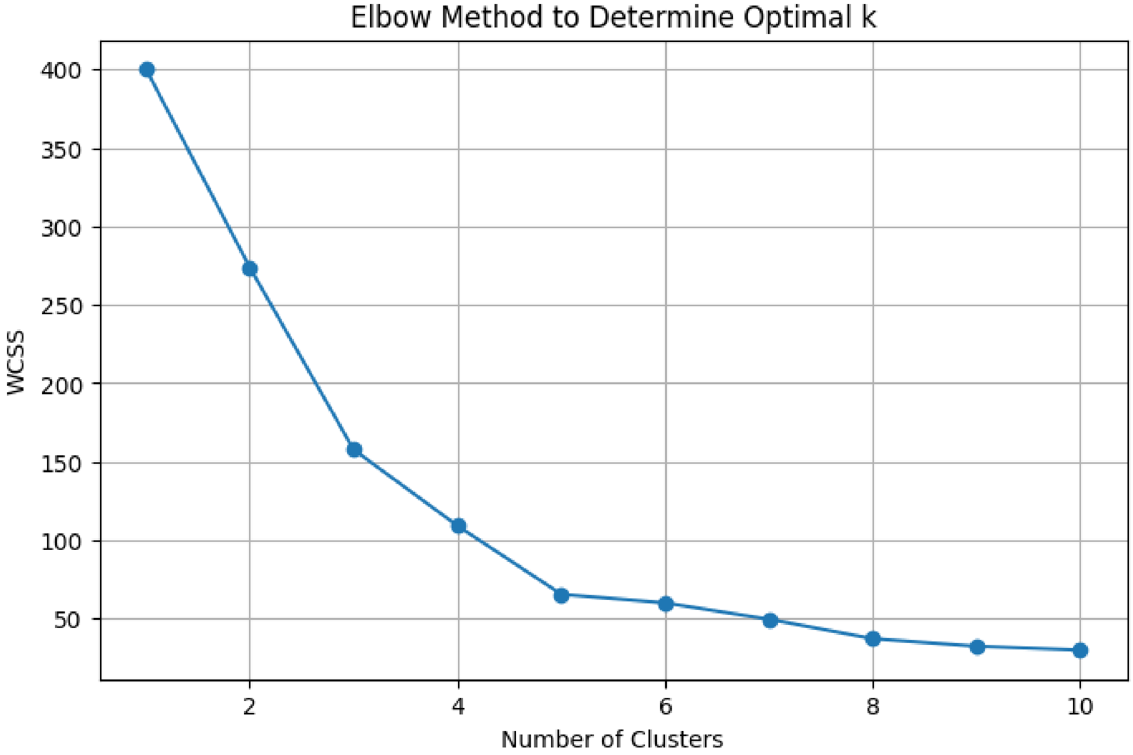


CODE:

```
% Customer Segmentation using K-Means in MATLAB

% Load dataset (assuming it's in CSV format)

data = readtable('Mall_Customers.csv');

% Extract relevant features: Annual Income and Spending Score

X = data{:, {'AnnualIncome','SpendingScore'}};

% Feature Scaling (optional)

X = normalize(X);

% Determine optimal number of clusters using Elbow Method
```

```
wcss = zeros(10,1);

for k = 1:10

    [~, ~, sumd] = kmeans(X, k);

wcss(k) = sum(sumd);

end

figure;

plot(1:10, wcss, '-o');

xlabel('Number of Clusters');

ylabel('Within-Cluster Sum of Squares');

title('Elbow Method to Determine Optimal k');

% Apply K-means clustering with optimal k (e.g., 5)

k = 5;

[idx, C] = kmeans(X, k);

% Visualize the clusters

figure;

gscatter(X(:,1), X(:,2), idx);

hold on;

plot(C(:,1), C(:,2), 'kx', 'MarkerSize', 15, 'LineWidth', 3);

xlabel('Annual Income (Normalized)');

ylabel('Spending Score (Normalized)');

title('Customer Segmentation using K-Means');

legend('Cluster 1','Cluster 2','Cluster 3','Cluster 4','Cluster 5','Centroids');

hold off;
```


Customer Segmentation using K-Means
Spending Score (Standardized)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3
Annual Income (Standardized)
0
1
2
3
4
Centroids


Elbow Method to Determine Optimal k
WCSS
400
350
300
250
200
150
100
50
2
4
6
8
10
Number of Clusters

EXPLANATION OF OUTPUT:

1. Elbow Method Graph

Explanation:

- **X-axis:** Number of Clusters (k)
- **Y-axis:** WCSS (Within-Cluster Sum of Squares)

Purpose:

This graph helps determine the **optimal number of clusters** for your data.

How it works:

- As you increase the number of clusters (`k`), the total WCSS decreases. This is because each point is closer to its cluster center.
- However, the decrease in WCSS starts to slow down at a certain point — this point is called the **"elbow"**.
- The **"elbow point"** represents the ideal number of clusters, beyond which adding more clusters doesn't significantly improve the model.

Interpretation of Your Graph:

- From the graph, the **elbow appears around `k = 5`**, which means **5 clusters** is the optimal choice for your dataset.

📌 2. Cluster Visualization Graph

⍰ Explanation:

- **X-axis:** Standardized Annual Income
- **Y-axis:** Standardized Spending Score
- Colored dots represent customer data points.
- Each **color** represents one of the **5 clusters**.
- The **black 'X' markers** represent the **centroids** (center points) of each cluster.

What This Shows:

- The K-Means algorithm successfully grouped customers into **five clusters** based on similarities in income and spending behavior.
- Each cluster shows a distinct grouping — for example:
  - One group may represent **high income & high spending**

    - Another may represent **low income & low spending**, etc.
- The centroids help visualize the average characteristics of each cluster.

Summary:

Graph	What it shows	Insight
Elbow Method	Optimal number of clusters (k)	Ideal `k = 5`
Cluster Plot	Segmented customer groups	5 meaningful groups based on income & spending

Chapter 5: Project Objectives & Achievements

5.1 Objectives of the Project

The main objective of this project was to implement customer segmentation using the K-Means clustering algorithm in MATLAB. By analyzing customer behavior, specifically their annual income and spending scores, the goal was to divide the customer base into distinct and meaningful groups to support data-driven marketing strategies. The specific objectives were:

- To understand the concept of customer segmentation and its importance in business.
- To apply the K-Means clustering algorithm using MATLAB.
- To identify meaningful customer clusters based on Annual Income and Spending Score.
- To visualize the segmentation results using MATLAB graphs.
- To determine the optimal number of clusters using the Elbow Method.

5.2 How Objectives Were Achieved

All the defined objectives were successfully achieved during the course of this project:

- A clear understanding of customer segmentation was developed through literature review and research.
- The K-Means algorithm was implemented effectively using MATLAB, demonstrating the power of unsupervised learning.
- The Mall Customers dataset was used and filtered based on relevant features (Annual Income & Spending Score).
- The Elbow Method was applied, and k = 5 was identified as the optimal number of clusters.
- MATLAB visualizations clearly showed five well-separated clusters along with centroids, confirming the success of clustering.
- The segmentation revealed meaningful insights into customer behavior, such as identifying high-income low-spending clusters, etc.

Chapter 6: Conclusion

In conclusion, this project successfully applied the K-Means clustering algorithm to achieve effective customer segmentation using MATLAB. By analyzing the spending score and income of mall customers, five distinct clusters were identified, each representing a different customer profile. The project not only met all of its objectives but also demonstrated the real-world applicability of clustering techniques in business environments. Future improvements could include using additional features like age and gender or experimenting with other clustering algorithms for deeper insights.