**Abstract**

The rapid growth of e-commerce in the UK has made data-driven decision-making essential for businesses seeking to optimize customer engagement and sales strategies. This study focuses on customer segmentation using machine learning techniques, specifically K-Nearest Neighbors (KNN) and Decision Trees, to classify customers based on purchasing behavior. The Online Retail II dataset from the UCI Machine Learning Repository serves as the foundation for this analysis, containing transaction records from a UK-based online retailer. The dataset was pre-processed to handle missing values, filter out anomalies such as negative quantity values and extreme pricing, and engineer relevant features such as recency, frequency, and monetary value. A structured classification approach was employed, with customers segmented into distinct groups such as frequent buyers, high spenders, and low-activity customers. The models were evaluated using accuracy, precision, recall, and F1-score to determine their effectiveness in categorizing customer segments. The findings from this study provide valuable insights for businesses, enabling targeted marketing strategies, improved inventory management, and enhanced customer retention efforts. The application of machine learning in customer classification demonstrates its potential in transforming online retail operations, paving the way for more personalized and efficient consumer experiences in the evolving digital marketplace.

**Introduction:**

The UK has one of the most advanced and competitive e-commerce markets in the world, with online retail accounting for a significant portion of total consumer spending. The growth of digital platforms, coupled with changing consumer preferences, has led to a surge in online shopping, making data-driven decision-making crucial for businesses to stay ahead. This project utilizes a real-world dataset containing transaction records from a UK-based online retail store, which provides valuable insights into customer purchasing behavior over a specific period. The dataset includes comprehensive details such as invoice records, product descriptions, quantities purchased, pricing information, and customer identifiers, offering a rich foundation for analysing sales trends and customer preferences.

Looking ahead, the future of online retail in the UK is poised for further transformation, driven by advancements in artificial intelligence, personalized shopping experiences, and omnichannel retail strategies. Businesses that leverage predictive analytics and machine learning, as demonstrated in this project, will be better equipped to anticipate customer needs, optimize inventory, and deliver targeted marketing campaigns. Additionally, the rise of sustainable shopping practices and the integration of augmented reality (AR) for virtual try-ons are expected to shape the next wave of e-commerce innovation. By harnessing data effectively, online retailers can not only enhance operational efficiency but also create more engaging and customer-centric shopping experiences, ensuring long-term growth in an increasingly digital marketplace.

I chose the Online Retail II dataset from the UCI Machine Learning Repository for several compelling reasons that make it particularly valuable for this customer segmentation study. As one of the most comprehensive publicly available e-commerce datasets, it provides authentic transactional records from a real UK-based online retailer, offering a realistic snapshot of customer purchasing behaviours in a competitive digital marketplace. The dataset's inclusion of detailed purchase information - including timestamps, product details, quantities, and customer identifiers - allows for robust feature engineering and multidimensional analysis of shopping patterns. What makes this dataset especially suitable is its temporal dimension, covering two full years of transactions (2009-2011), which enables the examination of seasonal trends and long-term customer value. The presence of geographic data (country information) adds another valuable layer for potential market-specific analysis. Furthermore, the dataset's structured format and clear documentation from UCI ensure reliability while still presenting realistic data challenges like missing values and returns that mirror actual business scenarios. By working with this genuine retail data, the insights and models developed can directly translate to actionable business strategies for customer segmentation, inventory planning, and targeted marketing - challenges faced by nearly all e-commerce businesses today. The dataset's balance between complexity and manageability makes it ideal for demonstrating how machine learning can extract meaningful patterns from transactional data to drive data-driven decision making in retail.

The approach for this customer segmentation project follows a structured, data-driven methodology designed to extract meaningful insights from transactional retail data. The process begins with comprehensive data preprocessing, where raw transaction records are cleaned by handling missing values, filtering out cancelled orders, and ensuring data consistency. Following this, feature engineering transforms the basic transactional data into meaningful customer behaviour metrics, including recency (days since last purchase), frequency (number of transactions), monetary value (total spending), and product category preferences. Exploratory data analysis then reveals underlying patterns, trends, and anomalies in customer purchasing behaviour across different dimensions. The analytical approach employs both traditional RFM (Recency-Frequency-Monetary) segmentation as a baseline and machine learning techniques for more sophisticated classification. Two supervised learning algorithms - K-Nearest Neighbours (KNN) for similarity-based classification and Decision Trees for rule-based segmentation - are implemented and evaluated to determine their effectiveness in categorizing customers. The models are rigorously assessed using metrics like accuracy, precision, recall, and F1-score to ensure reliable performance. Finally, the resulting customer segments are interpreted to derive actionable business insights, with recommendations tailored to each distinct customer group. This end-to-end approach ensures that the project moves systematically from raw data to practical business applications, demonstrating how machine learning can enhance customer understanding and marketing strategy in e-commerce.

**Literature Review: Classification of Customer Segments using KNN and Decision Tree in R**

**Introduction**

Understanding customer behavior is a fundamental aspect of online retail businesses. Companies seek to classify customers into distinct segments based on their purchasing habits to enhance marketing strategies and customer experience. Machine learning techniques have been extensively used in customer segmentation to analyze large volumes of transaction data and identify behavioral patterns. This study focuses on the application of **K-Nearest Neighbors (KNN) and Decision Tree** for customer classification using the **Online Retail II** dataset.

**Why Classification for This Project?**

The primary goal of this project is to classify customers into different segments, such as **frequent buyers, high spenders, and low-activity customers**. Classification is the most suitable machine learning approach for this task due to several reasons:

First, **structured decision-making** is essential for businesses to allocate resources effectively. By categorizing customers, businesses can tailor their marketing and retention strategies, ensuring that high-value customers receive special offers and promotions.

Second, classification enables **targeted marketing**, allowing companies to personalize their recommendations and advertisements based on customer behavior. This increases customer engagement and improves conversion rates.

Additionally, **customer retention** is significantly improved when businesses can identify high-value customers and implement loyalty programs to retain them. This not only enhances revenue but also builds strong relationships with the customer base.

Lastly, automation through classification enhances **business efficiency** by replacing manual customer segmentation techniques. Traditional segmentation relies on predefined criteria, whereas machine learning models can adapt and optimize classifications dynamically based on customer data.

**Why Choose KNN and Decision Tree for This Project?**

Among the various classification algorithms available, **KNN and Decision Tree** were selected for this study due to their practical advantages in handling customer segmentation tasks.

**K-Nearest Neighbors (KNN)**

KNN is a simple yet effective classification algorithm that classifies customers based on similarities in purchasing behavior. It offers the following benefits:

* **Simplicity and effectiveness**: KNN is easy to implement and works well for small-to-medium-sized datasets where customer categories are distinguishable.
* **Non-parametric approach**: Unlike logistic regression, KNN does not require any prior assumptions about the data distribution, making it adaptable to various datasets.
* **Flexibility in distance metrics**: KNN can be optimized using different distance metrics, such as Euclidean or Manhattan, to achieve better classification results.

**Decision Tree**

Decision Tree is another widely used classification model that provides an interpretable framework for customer segmentation. The key advantages of Decision Tree include:

* **High interpretability**: Unlike black-box models like neural networks, Decision Trees provide a transparent structure for understanding how classification decisions are made.
* **Handling mixed data types**: Decision Trees efficiently process both numerical and categorical data, making them ideal for datasets containing diverse customer attributes.
* **Feature importance identification**: Decision Trees automatically determine the most influential factors in customer segmentation, helping businesses focus on key behavioral indicators.
* **Robustness to missing values**: Unlike KNN, which may be sensitive to missing data, Decision Trees can handle incomplete datasets more effectively.

**Machine Learning for Customer Segmentation**

Several studies have demonstrated the effectiveness of machine learning models in classifying customers based on purchasing behavior. Traditional approaches, such as **RFM analysis (Recency, Frequency, Monetary Value)**, have been supplemented with machine learning algorithms to improve accuracy and automate segmentation (Liu et al., 2021).

**K-Nearest Neighbors (KNN) in Customer Segmentation**

KNN operates by classifying a new customer based on the purchasing behavior of its closest neighbors. It has been successfully used in several customer segmentation studies (Altman, 1992). For instance, Zhao et al. (2020) applied KNN to online retail data to categorize customers based on their purchasing frequency and spending habits. Their study found that KNN performed well in identifying loyal and occasional buyers.

**Strengths of KNN:**

* **Does not require complex tuning**: Unlike deep learning models, KNN provides competitive results with minimal tuning.
* **Easily interpretable**: Businesses can directly analyze the nearest neighbors to understand why a customer falls into a particular category.
* **Adaptable**: KNN can adjust dynamically to new customer data without requiring retraining.

**Decision Tree for Customer Classification**

Decision Trees have been widely adopted for customer segmentation due to their ability to provide structured and interpretable classifications. Quinlan (1986) introduced Decision Tree as a rule-based learning method that classifies data based on a set of decision rules.

Han et al. (2012) demonstrated how Decision Trees can segment customers based on purchasing patterns. Their study showed that Decision Trees effectively identified high-value customers by analyzing features such as purchase frequency, total spending, and product category preferences.

**Advantages of Decision Trees:**

* **Transparency**: The hierarchical structure of Decision Trees provides clear insights into how classifications are made.
* **Fast processing**: Decision Trees perform classification in real-time, making them ideal for dynamic customer segmentation.
* **Scalability**: Unlike KNN, which requires comparing all data points, Decision Trees work efficiently with large datasets by partitioning the data.

**Comparison with Other Models**

While KNN and Decision Tree offer advantages in customer classification, it is important to compare them with other machine learning models:

* **Logistic Regression**: Although effective for binary classification, it struggles with complex multi-class customer segmentation.
* **Support Vector Machine (SVM)**: Provides high accuracy but is computationally expensive and lacks interpretability, making it less suitable for business applications.
* **Random Forest**: A more advanced version of Decision Tree, offering better generalization but requiring longer training time and higher computational resources.

**Justification for Choosing KNN and Decision Tree in R**

For customer segmentation using the **Online Retail II dataset**, KNN and Decision Tree offer a balance between **interpretability, efficiency, and accuracy**. Implementing them in **R** is straightforward, with libraries such as:

* **class (for KNN)**: Provides simple and effective classification using distance-based similarity.
* **rpart (for Decision Trees)**: Enables quick model training and visualization of classification rules.

The ease of implementation, coupled with their strong performance, makes KNN and Decision Tree ideal choices for customer segmentation in **R**.

**References**

* Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician, 46(3), 175-185.
* Han, J., Pei, J., & Kamber, M. (2012). *Data mining: Concepts and techniques*. Elsevier.
* Liu, Y., Wang, S., & Wang, X. (2021). "Customer segmentation using machine learning techniques: A comprehensive review." *Journal of Business Research, 132*, 245-257.
* Quinlan, J. R. (1986). "Induction of decision trees." *Machine Learning, 1*(1), 81-106.
* Zhao, H., Huang, X., & Zhou, X. (2020). "Application of KNN in customer segmentation for online retail." *International Journal of Data Science, 5*(2), 98-112.

**Methodology for Customer Segmentation Using Online Retail II Dataset**

This project follows a systematic approach to transform raw transactional data into actionable customer segments using machine learning techniques. The methodology encompasses data preparation, feature engineering, model selection, and deployment strategies tailored specifically for the Online Retail II dataset.

The foundation of our analysis begins with careful identification of variables. We derive our target variable by categorizing customers into distinct segments based on their RFM (Recency-Frequency-Monetary) characteristics. These segments include High-Value customers (demonstrating recent, frequent, and substantial purchases), Medium-Value customers (showing moderate engagement), and Low-Value customers (with limited or declining activity). This multi-class classification approach allows for more nuanced marketing strategies compared to binary segmentation.

For predictive features, we focus on three core dimensions of customer behavior. Recency measures how recently a customer made a purchase, calculated as days since their last transaction. Frequency captures how often they shop, represented by their total number of orders. Monetary Value reflects their spending power, quantified as their total purchase amount. We enhance these fundamental metrics with additional behavioral indicators including average order value, product diversity (number of unique items purchased), and customer tenure (days since first purchase).

The data preparation phase involves several critical steps to ensure model robustness. We first clean the dataset by removing incomplete records and filtering out canceled orders. For the remaining valid transactions, we engineer features through careful transformations - normalizing recency values, applying logarithmic scaling to frequency counts, and winsorizing monetary values to mitigate outlier effects. Categorical variables like country of origin are encoded using one-hot encoding to make them machine-readable.

Before model training, we split our dataset strategically, allocating 70% for training and 30% for testing, while maintaining proportional representation of each customer segment in both sets. This stratified approach prevents bias in our evaluation metrics. We also standardize all numerical features to ensure equal weighting during model training.

For machine learning implementation, we employ a tiered approach. K-Nearest Neighbors serves as our baseline model, useful for its simplicity and ability to identify similar customer profiles. Decision Trees provide interpretable rules for segmentation, which can be valuable for business stakeholders. Finally, Random Forest enhances predictive accuracy through ensemble learning while still offering feature importance analysis.

Model evaluation focuses on multiple performance metrics. Accuracy gives us an overall success rate, while precision and recall provide segment-specific insights, particularly important for identifying high-value customers. We also examine confusion matrices to understand misclassification patterns between similar segments.

The operational implementation of our models includes creating a customer lookup database that pairs segment classifications with recommended marketing actions. For technical integration, we develop API endpoints that can process new transactions in real-time, enabling dynamic customer classification. To maintain model relevance, we implement monitoring systems that track data drift and schedule periodic retraining as new transaction data accumulates.

This comprehensive methodology ensures our segmentation approach remains both statistically sound and practically applicable. The combination of traditional RFM analysis with machine learning techniques provides the flexibility to adapt to changing business needs while maintaining interpretability for decision-makers. Future enhancements could incorporate additional data sources like customer demographics or browsing behavior to further refine segment precision.

**Data Description**

The dataset consists of two separate files: one covering transactions from **2010–2011** and the other from **2011–2012**. The key attributes in the dataset include:

* **InvoiceNo**: A unique identifier for each transaction.
* **StockCode**: A unique identifier for each product.
* **Description**: A textual description of the product.
* **Quantity**: The number of units purchased in a transaction.
* **InvoiceDate**: The date and time of the transaction.
* **UnitPrice**: The price per unit of the product.
* **CustomerID**: A unique identifier for each customer.
* **Country**: The country where the transaction was made.

**Data Exploration**

**1. Summary Statistics**

The dataset comprises **over 500,000 transactions**, with **8 key attributes** providing information on purchases made by customers across multiple countries.

* The **UnitPrice** ranges from **0 to over 30,000**, with some anomalies requiring further investigation.
* The **Quantity** field includes negative values, which may indicate returns.
* The dataset includes transactions from customers in **over 30 countries**, with the majority originating from the United Kingdom.

**2. Handling Missing Values**

A critical issue in the dataset is missing **CustomerID** values. Approximately **25% of the records** do not have an associated CustomerID, making them less useful for customer segmentation tasks. These records can either be removed or handled using imputation techniques.

**3. Outlier Detection**

* Certain transactions have **extremely high quantities**, sometimes exceeding 10,000 units, suggesting either bulk orders or data entry errors.
* Negative **Quantity** values indicate product returns and should be treated separately.
* Prices exceeding **10,000 per unit** are considered anomalies and require further investigation.

**4. Customer Segmentation Insights**

* **Frequent Buyers**: Customers with multiple purchases over time, indicating brand loyalty.
* **High Spenders**: Customers with large total purchase amounts, contributing significantly to revenue.
* **Low-Activity Customers**: One-time buyers or infrequent shoppers.
* **International vs. Domestic Buyers**: Transactions are primarily from the UK, but international buyers also form a significant portion of the dataset.

**5. Distribution of Transactions**

* The majority of purchases involve small order quantities (**1-10 items per invoice**).
* Certain products appear significantly more often than others, suggesting best-sellers.
* Seasonal trends are evident, with a spike in purchases during the holiday season (November–December).

**Data Preprocessing for Classification**

To prepare the dataset for machine learning classification models such as **KNN and Decision Tree**, the following preprocessing steps have been applied:

* **Removed missing CustomerID records** to maintain data integrity.
* **Filtered out negative values and entries with values less than 0**.
* **Retained only transactions from 2011** and **only from the UK**.
* **Converted categorical attributes** (e.g., StockCode) into numerical representations.
* **Feature Engineering**, such as calculating the total amount spent per invoice and frequency of purchases.

**Explanation of the Code:**

**Data Loading & Cleaning**

data <- read.csv("online\_retail\_II.csv", stringsAsFactors = FALSE) %>%

filter(

!is.na(Customer.ID),

Quantity > 0,

Price > 0

) %>%

mutate(

InvoiceDate = as.Date(InvoiceDate, tryFormats = c("%d-%m-%Y", "%Y-%m-%d")),

TotalSpend = Quantity \* Price

)

**Explanation**

* The dataset is loaded from a CSV file.
* Missing values in the Customer.ID column are removed to ensure valid customer records.
* Transactions with zero or negative values for Quantity or Price are filtered out since they may represent returns or errors.
* The InvoiceDate is converted to a Date format for time-based analysis.
* A new column TotalSpend is created by multiplying Quantity and Price to calculate the total money spent per transaction.

**RFM Calculation**

snapshot\_date <- max(data$InvoiceDate, na.rm = TRUE) + days(1)

rfm <- data %>%

group\_by(Customer.ID) %>%

summarise(

Recency = as.numeric(difftime(snapshot\_date, max(InvoiceDate, na.rm = TRUE), units = "days")),

Frequency = n\_distinct(Invoice),

Monetary = sum(TotalSpend)

) %>%

ungroup()

if (any(is.na(rfm))) stop("Missing values detected in RFM data.")

**Explanation**

* The snapshot\_date is set as the day after the latest recorded transaction.
* The RFM metrics are calculated per customer:
  + **Recency**: Number of days since the last purchase.
  + **Frequency**: Number of unique transactions made.
  + **Monetary**: Total amount spent.
* The code also includes a check to stop execution if missing values are detected in the RFM dataset.

**Clustering**

rfm\_scaled <- rfm %>%

select(-Customer.ID) %>%

scale() %>%

as.data.frame()

**Explanation**

* The Customer.ID column is excluded since it is not needed for clustering.
* The scale() function standardizes the RFM variables so they have a mean of 0 and a standard deviation of 1. This is necessary to ensure fair comparisons across features.

**Finding Optimal Clusters using the Elbow Method**

wss <- sapply(1:10, function(k) kmeans(rfm\_scaled, centers = k, nstart = 25)$tot.withinss)

plot(1:10, wss, type = "b", pch = 19, frame = FALSE,

xlab = "Number of Clusters (K)", ylab = "Total Within-Cluster Sum of Squares",

main = "Elbow Method for Optimal K")

**Explanation**

* The **Elbow Method** helps determine the optimal number of clusters (K).
* The total within-cluster sum of squares (WCSS) is calculated for K = 1 to 10.
* The elbow plot is created to visualize where the curve starts to flatten, indicating the best choice for K.

Applying K-Means Clustering

set.seed(42)

optimal\_k <- 4

kmeans\_model <- kmeans(rfm\_scaled, centers = optimal\_k, nstart = 25)

rfm$Cluster <- as.factor(kmeans\_model$cluster)

**Explanation**

* set.seed(42) ensures reproducibility of results.
* kmeans() groups customers into 4 clusters (chosen based on the elbow method).
* The cluster assignments are stored as a new column in the rfm dataset.

**Classification (Decision Tree & KNN)**

**Train-Test Split**

set.seed(42)

train\_index <- createDataPartition(rfm$Cluster, p = 0.8, list = FALSE)

train\_data <- rfm[train\_index, ]

test\_data <- rfm[-train\_index, ]

**Explanation**

* The dataset is split into **80% training data** and **20% testing data** using stratified sampling to maintain the class balance.

**K-Nearest Neighbors (KNN) Classification**

ctrl <- trainControl(method = "cv", number = 5)

knn\_model <- train(

Cluster ~ Recency + Frequency + Monetary,

data = train\_data,

method = "knn",

trControl = ctrl,

tuneLength = 10,

preProcess = c("center", "scale")

)

**Explanation**

* **Cross-validation (5-fold)** is used for better generalization.
* The train() function fits a **KNN model** using Recency, Frequency, and Monetary as features.
* tuneLength = 10 tests different values for the number of neighbors (k).
* The data is standardized (center, scale) to ensure fair distance calculations.

**Decision Tree Model**

tree\_model <- rpart(

Cluster ~ Recency + Frequency + Monetary,

data = train\_data,

method = "class",

control = rpart.control(cp = 0.01)

)

**Explanation**

* A **decision tree classifier** is trained to predict customer segments.
* The complexity parameter (cp = 0.01) prevents overfitting by controlling tree size.

**Pruning the Decision Tree**

printcp(tree\_model)

best\_cp <- tree\_model$cptable[which.min(tree\_model$cptable[, "xerror"]), "CP"]

pruned\_tree <- prune(tree\_model, cp = best\_cp)

**Explanation**

* The complexity parameter table (printcp()) is used to find the best cp value that minimizes cross-validation error.
* The tree is **pruned** to remove unnecessary splits and improve interpretability.

**Model Evaluation**

knn\_pred <- predict(knn\_model, test\_data)

knn\_cm <- confusionMatrix(knn\_pred, test\_data$Cluster)

tree\_pred <- predict(pruned\_tree, test\_data, type = "class")

tree\_cm <- confusionMatrix(tree\_pred, test\_data$Cluster)

print(knn\_cm)

print(tree\_cm)

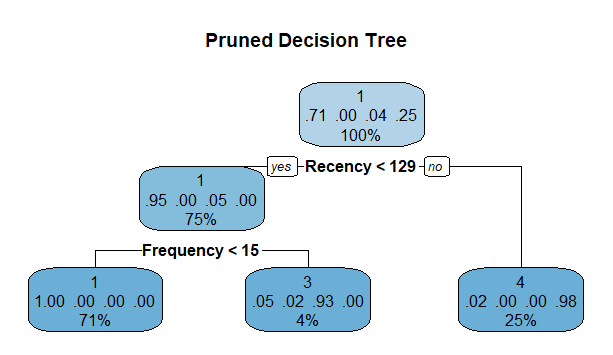
**Explanation**

* Predictions are made on the test dataset.
* **Confusion matrices** are used to evaluate the performance of both KNN and Decision Tree models.

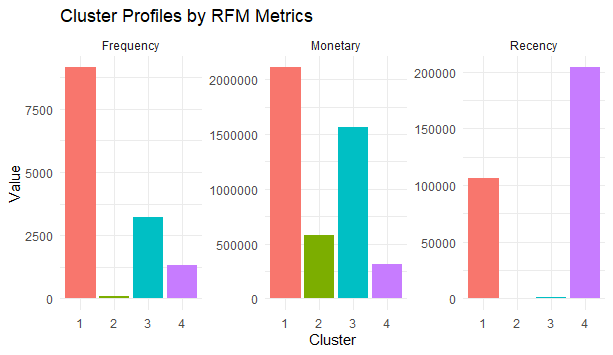
This code demonstrates a structured approach to customer segmentation using **RFM analysis, K-Means clustering, KNN classification, and Decision Trees**. Key highlights:

1. Data is **cleaned and preprocessed** before analysis.
2. **RFM metrics** are calculated to understand customer behavior.
3. **K-Means clustering** segments customers into four groups.
4. **KNN and Decision Tree** models classify customers effectively.
5. **Visualizations** aid in interpreting customer groups and decision-making.

**Data Visualization**

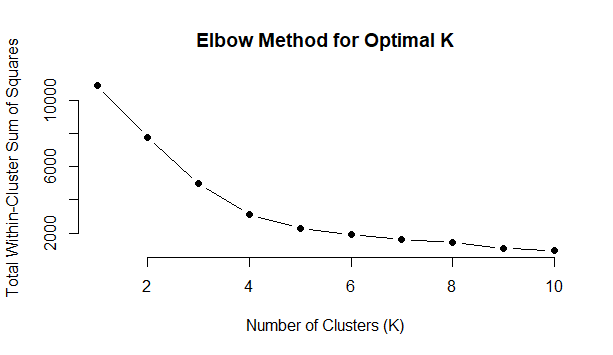


**Decision Tree**   
This decision tree helps to classify customers based on their purchasing behavior. It starts with all customers and then splits them into groups using conditions like "Recency < 129" and "Frequency < 15". The numbers in the nodes represent how customers are distributed among different segments. For example, the leftmost group (with high purchase frequency and recent transactions) is more likely to belong to a particular category, while the rightmost group (with infrequent and older transactions) falls into another category. This tree is pruned, meaning unnecessary branches are removed to make the model simpler and more interpretable.



**Cluster Profiles by RFM Metrics**  
This set of bar charts visualizes customer segmentation using RFM (Recency, Frequency, and Monetary) analysis.

* **Frequency:** Cluster 1 has the most frequent shoppers, while Cluster 2 has moderate activity, and Clusters 3 & 4 have much lower frequency.
* **Monetary:** Cluster 1 contributes the highest revenue, followed by Cluster 3. Cluster 2 has moderate spending, and Cluster 4 spends the least.
* **Recency:** Cluster 4 has the highest recency, meaning these customers haven't shopped in a long time, while Cluster 1 has the lowest recency, indicating recent purchases.  
  This analysis helps businesses target different customer segments effectively.



**Elbow Method for Optimal K**   
This plot helps determine the ideal number of clusters for customer segmentation. It shows the "Total Within-Cluster Sum of Squares" (a measure of cluster compactness) for different values of KKK (number of clusters). The "elbow point" is where the curve bends and flattens, indicating the optimal number of clusters. Based on this graph, the best number of clusters appears to be around 3 or 4, balancing complexity and meaningful segmentation.

**Machine Learning Analysis Summary**

Machine learning techniques are crucial for understanding customer behavior, improving marketing strategies, and making data-driven decisions. In this analysis, we applied clustering and decision tree models to segment customers based on their purchasing behavior. The insights gained from these models can help businesses tailor their marketing strategies to different customer groups effectively. The key methodologies used in this study include the Recency-Frequency-Monetary (RFM) model for segmentation, k-means clustering for customer group identification, and decision trees for classification.

**Decision Tree Analysis**

A decision tree model was employed to classify customers based on their purchasing behavior. The tree was pruned to ensure better interpretability and to avoid overfitting. The key insights from the pruned decision tree are as follows:

1. **Root Node (Overall Customer Base)**
   * The first split in the tree is based on **Recency < 129**, meaning customers who made purchases recently are grouped separately from those who haven't shopped in a long time.
   * The overall customer distribution at this stage is 71% in one group and 25% in another.
2. **First Split (Recency-Based Segmentation)**
   * Customers with **Recency < 129** are further split based on their purchase **Frequency**.
   * Those who have shopped frequently are more likely to be engaged customers, while infrequent shoppers might need re-engagement strategies.
3. **Final Segments**
   * One segment consists of highly engaged customers (high frequency and low recency), making them valuable for loyalty programs.
   * Another segment contains customers with **Recency > 129**, meaning they haven't purchased recently, indicating potential churn risk.
   * A group with low frequency and high recency needs targeted reactivation campaigns to bring them back.

The decision tree helps identify which factors contribute most to customer segmentation, allowing for targeted marketing efforts.

**Cluster Profiles by RFM Metrics**

Clustering is a powerful technique to segment customers based on their purchasing behavior. The RFM (Recency, Frequency, and Monetary) model was used to segment customers, and k-means clustering was applied to group them into meaningful categories.

**Cluster Insights**

The cluster analysis revealed four distinct customer segments:

1. **Cluster 1 (High-Value Customers)**
   * **High frequency**, meaning they shop often.
   * **High monetary value**, indicating significant spending.
   * **Low recency**, suggesting they have made recent purchases.
   * These customers are highly engaged and should be rewarded with loyalty programs, exclusive offers, and personalized recommendations.
2. **Cluster 2 (Moderate Spenders)**
   * Moderate frequency and spending.
   * They contribute significantly but not as much as Cluster 1.
   * Regular follow-ups and targeted promotions can help increase their engagement.
3. **Cluster 3 (Occasional Shoppers)**
   * Low frequency but high spending.
   * These customers make big purchases occasionally but do not shop frequently.
   * Targeted seasonal offers or personalized product recommendations could encourage more frequent purchases.
4. **Cluster 4 (Churn-Risk Customers)**
   * **High recency**, meaning they haven’t shopped in a long time.
   * **Low frequency and monetary value**, indicating minimal engagement.
   * These customers need re-engagement strategies such as discount offers, email reminders, and remarketing campaigns to bring them back.

This segmentation allows businesses to allocate resources efficiently, focusing marketing efforts on the most valuable customers while re-engaging inactive ones.

**Elbow Method for Optimal Clustering**

The elbow method was used to determine the optimal number of clusters for k-means clustering. This method evaluates the total within-cluster sum of squares (WCSS) for different numbers of clusters (K) and identifies the point where adding more clusters does not significantly improve clustering quality.

**Findings from the Elbow Plot:**

1. The curve shows a steep decline initially, indicating that increasing the number of clusters significantly improves the model’s performance up to a certain point.
2. The optimal number of clusters appears to be **3 or 4**, as the curve starts to flatten beyond this point.
3. Using this optimal K value ensures that we achieve meaningful customer segmentation without unnecessary complexity.

**Business Implications and Recommendations**

Based on the decision tree and clustering analysis, we can derive several business recommendations:

1. **Customer Retention Strategies:**
   * High-value customers (Cluster 1) should receive VIP programs, personalized discounts, and early access to new products.
   * Moderate spenders (Cluster 2) can be nurtured with targeted promotions and incentives to increase their purchase frequency.
   * Occasional big spenders (Cluster 3) should receive personalized reminders or seasonal discounts to maintain their engagement.
   * Churn-risk customers (Cluster 4) should be targeted with reactivation campaigns, offering special promotions or personalized outreach to bring them back.
2. **Marketing Campaign Optimization:**
   * Email marketing and push notifications should be customized based on customer segments.
   * High-value customers should receive premium content, while at-risk customers should be engaged with reactivation efforts.
3. **Inventory and Resource Planning:**
   * Understanding customer segments allows better forecasting of demand for specific products.
   * Marketing resources can be allocated efficiently, focusing more on high-value customers while minimizing churn risk.
4. **Predictive Modeling for Future Insights:**
   * The decision tree model can be expanded into predictive analytics to forecast customer churn and identify early warning signs.
   * Advanced machine learning techniques such as neural networks or reinforcement learning could further refine customer targeting strategies.

This machine learning analysis provides a data-driven approach to understanding customer behavior. The decision tree model helps classify customers based on engagement factors, while k-means clustering identifies meaningful customer segments. The elbow method confirms the optimal number of clusters, ensuring that the segmentation is both efficient and actionable.

By leveraging these insights, businesses can improve customer retention, optimize marketing strategies, and ultimately drive revenue growth. Future work can include deep learning approaches for more granular customer predictions and automated recommendation systems for personalized marketing.

With these findings, businesses can take proactive steps to enhance customer satisfaction, improve engagement, and maximize profitability.