CAPSTONE PROJECT

ANALYSIS OF ONLINERETAIL.CSV DATASET

Brendan C Rosario

TABLE OF CONTENTS

PROBLEM STATEMENT

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

PROBLEM OBJECTIVE

The objective of this project is to:

- **Perform customer segmentation** using Recency, Frequency, and Monetary (RFM) analysis to categorize customers based on their purchasing behavior.

- **Utilize machine learning models** to automate the classification of customers, allowing businesses to predict future customer actions and spending habits.

- **Identify high-value customers** who contribute the most revenue and implement targeted strategies to retain them.

- **Detect at-risk customers** who have reduced their purchasing frequency and develop re-engagement campaigns to bring them back.

- **Improve business decision-making** by providing actionable insights into customer trends, allowing companies to optimize inventory, pricing, and promotions.

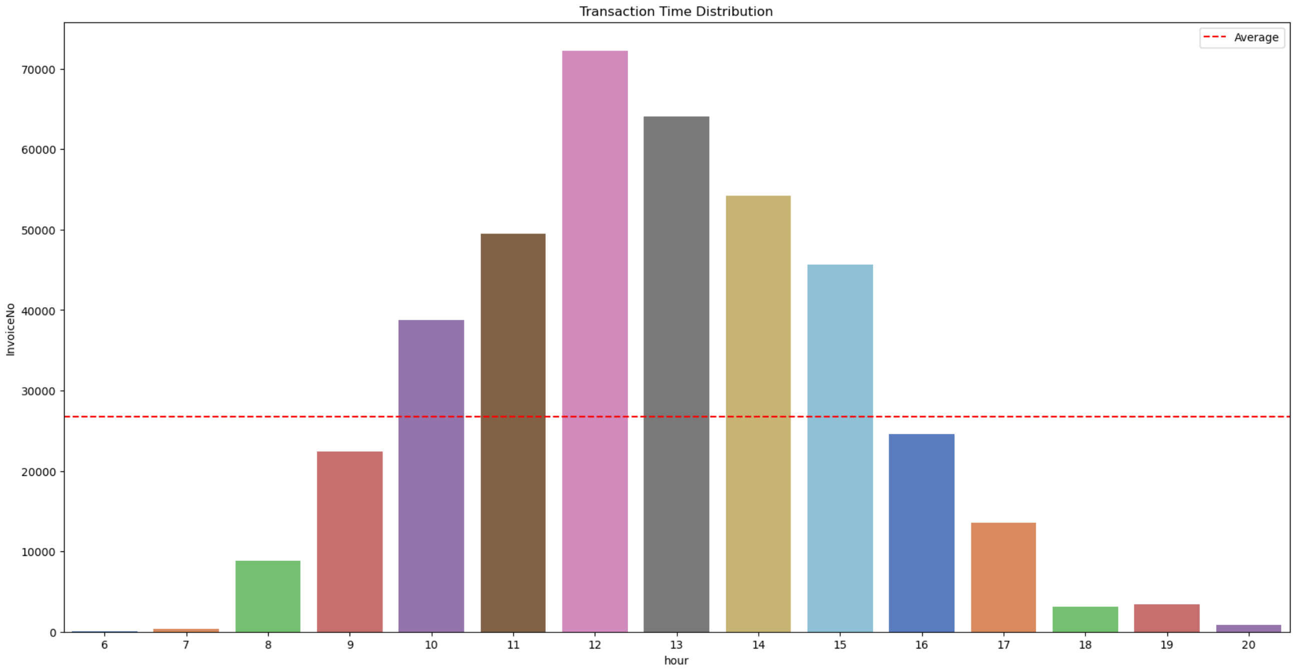
- **Enhance marketing effectiveness** by identifying different customer segments and creating personalized marketing strategies tailored to each group's behavior.

DATASET DESCRIPTION

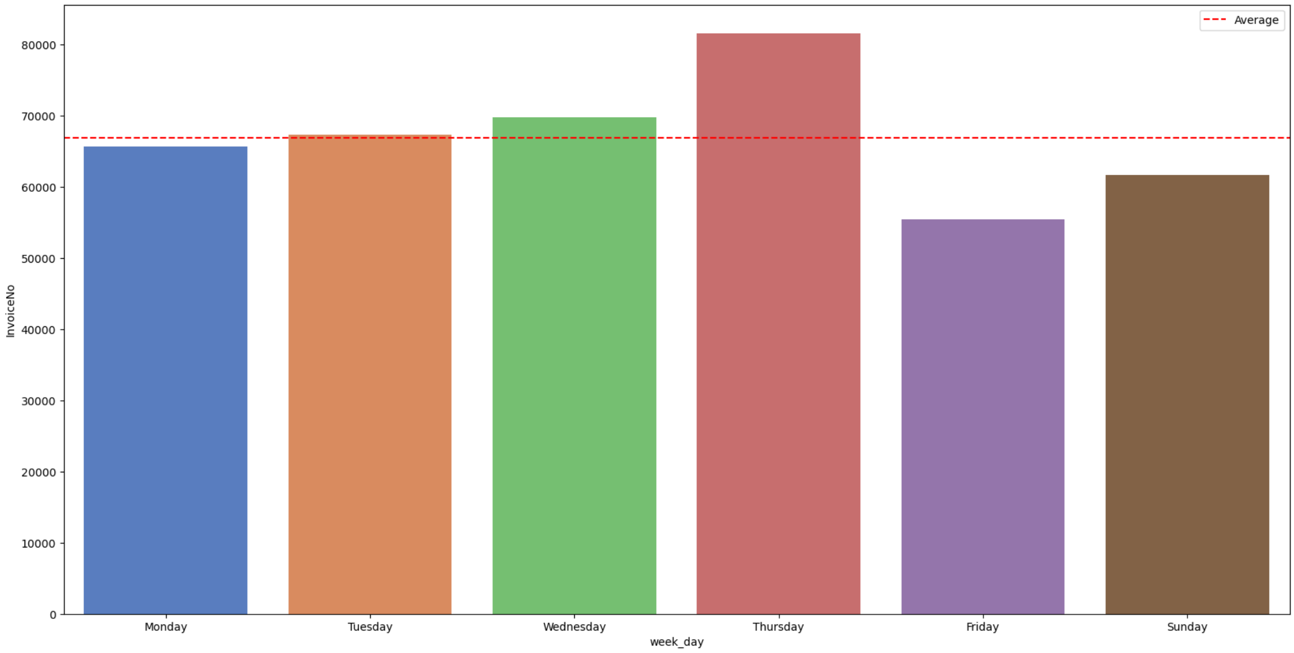
- **InvoiceNo:** Unique invoice identifier.
- **StockCode:** Unique product code.
- **Description:** Product description.
- **Quantity:** Number of products purchased.
- **InvoiceDate:** Date and time of purchase.
- **UnitPrice:** Price per unit.
- **CustomerID:** Unique customer identifier.
- **Country:** Country of the customer.

INSIGHTS FROM THE DATA

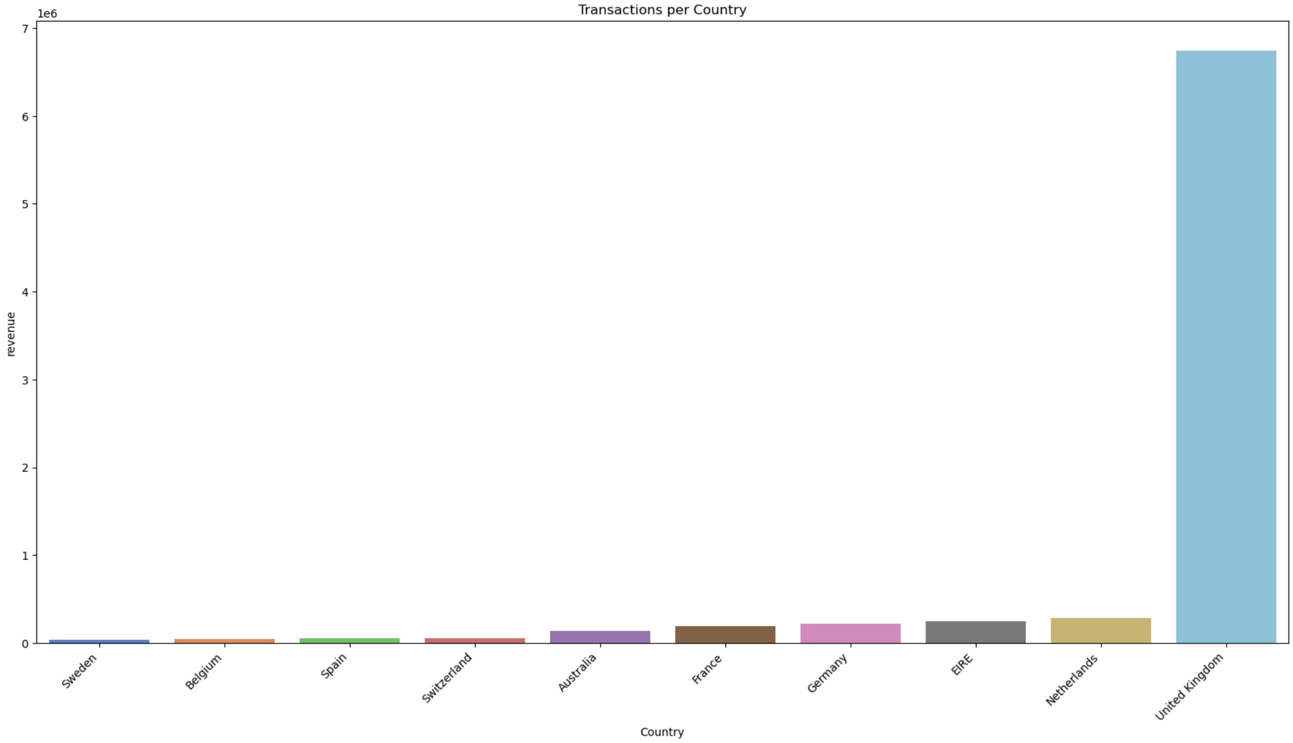
- The majority of transactions occur during mid-day.



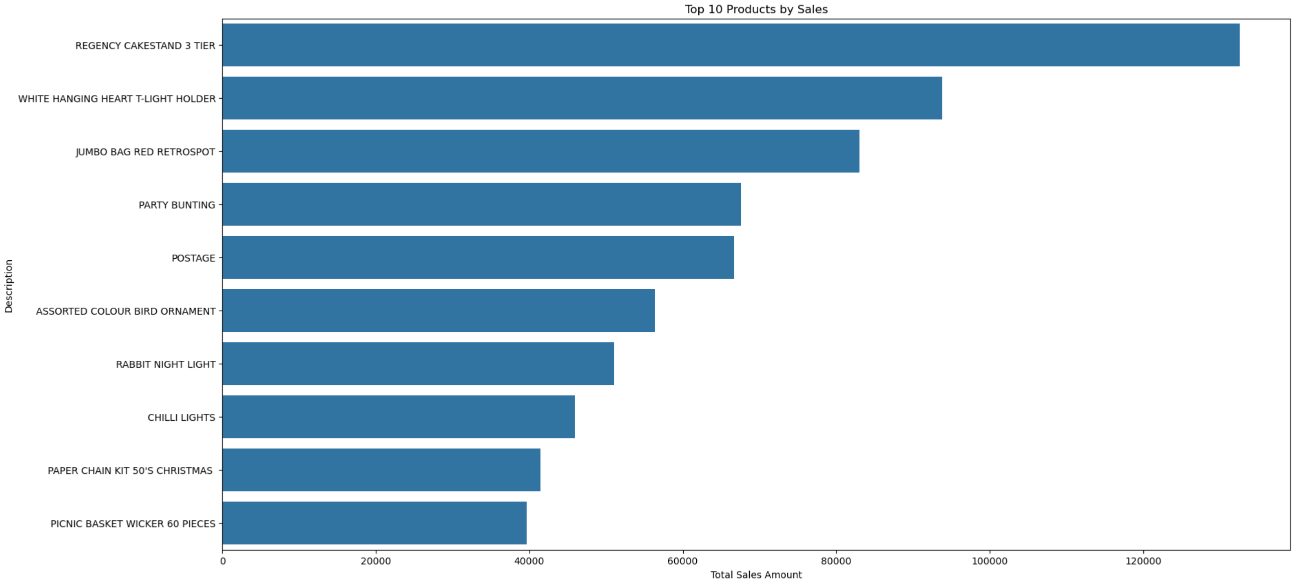
- Sales or transactions peak on Thursday, while Friday experiences the lowest activity, indicating a mid-week surge in business and a decline towards the weekend.



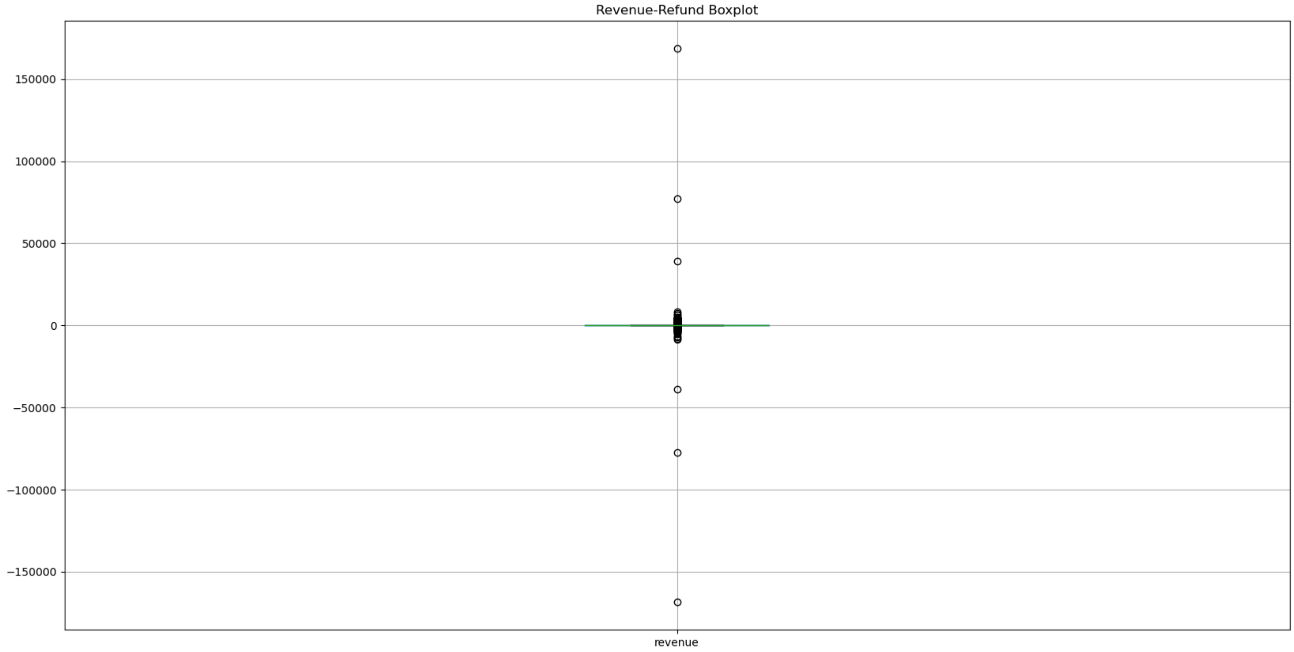
- The highest sales come from the United Kingdom.



- Certain Products show repeated purchases, while others only make few transaction.



- Negative revenue values indicate refunds or cancellations (Refund and Cancellation Transactions Plot).



DATA PRE-PROCESSING STEPS AND INSPIRATION

To ensure high-quality input data for the machine learning models, the following pre-processing steps were applied:

- **Datetime Feature Engineering:** Extracted the year, month, day, and hour from the InvoiceDate column to analyze seasonal trends and purchasing behavior over time.

- **Revenue Calculation:** Created a Revenue column by multiplying Quantity and UnitPrice to measure total transaction value and help assess customer spending patterns.

- **Handling Missing Values:** Identified and removed transactions with missing CustomerID values to ensure data integrity in customer segmentation.

- **Duplicate Removal:** Identified and dropped duplicate entries to prevent data redundancy and bias in analysis.

- **Zero and Negative Values Treatment:** Transactions with UnitPrice = 0 or negative values were removed since they indicate corrections, refunds, or data errors that could mislead model predictions.

- **Outlier Detection and Removal:** Extreme values in revenue and quantity were analyzed and capped to prevent model distortion caused by anomalies.

- **RFM Analysis:** Created three new features:
  - Recency (R): Days since the last purchase to assess customer activity.
  - Frequency (F): Total number of transactions per customer.
  - Monetary (M): Total spending per customer.

- **Scaling and Normalization:** Applied Min-Max Scaling for machine learning models that require normalized input features.

- Data Splitting: The dataset was divided into training and testing sets using an 80-20 split to evaluate model performance effectively.

INSPIRATION FOR PREPROCESSING STEPS

- **RFM Segmentation Theory:** Used in marketing and CRM strategies, RFM analysis helps businesses identify valuable customers by analyzing purchasing behavior. Implementing this method allowed us to quantify customer engagement levels effectively.

- **Exploratory Data Analysis (EDA) Findings:** The presence of missing values, duplicate records, and outliers required careful handling to avoid skewing the model results.

- **Retail Industry Best Practices:** Handling refunds, zero-priced items, and extreme values is critical in sales data preprocessing to maintain data accuracy and reflect real purchasing behavior.

- **Feature Engineering Best Practices:** Extracting meaningful insights from date-based variables improves predictive performance, as seasonality and recency impact customer retention.

- **Scalability Considerations:** Applying data normalization ensures that models such as K-Means clustering perform optimally without being influenced by large-scale monetary differences across customers.

- **Model Performance Optimization:** Data splitting ensures fair model evaluation, preventing overfitting and allowing generalization to unseen customer data.

The preprocessing pipeline was crucial in transforming raw data into structured, meaningful insights that support effective customer segmentation and predictive modeling.

CHOOSING THE ALGORITHM FOR THE PROJECT

We first use RFM Analysis to understand and segment the customers based on the Three models were selected for the prediction task:

- **K-Means Clustering:** For segmenting customers into different clusters.
- **Logistic Regression:** For classifying customers based on their RFM scores.
- **Decision Tree Classifier:** For high-accuracy classification of customers into different segments.

MOTIVATION AND REASONS FOR CHOOSING THE ALGORITHM

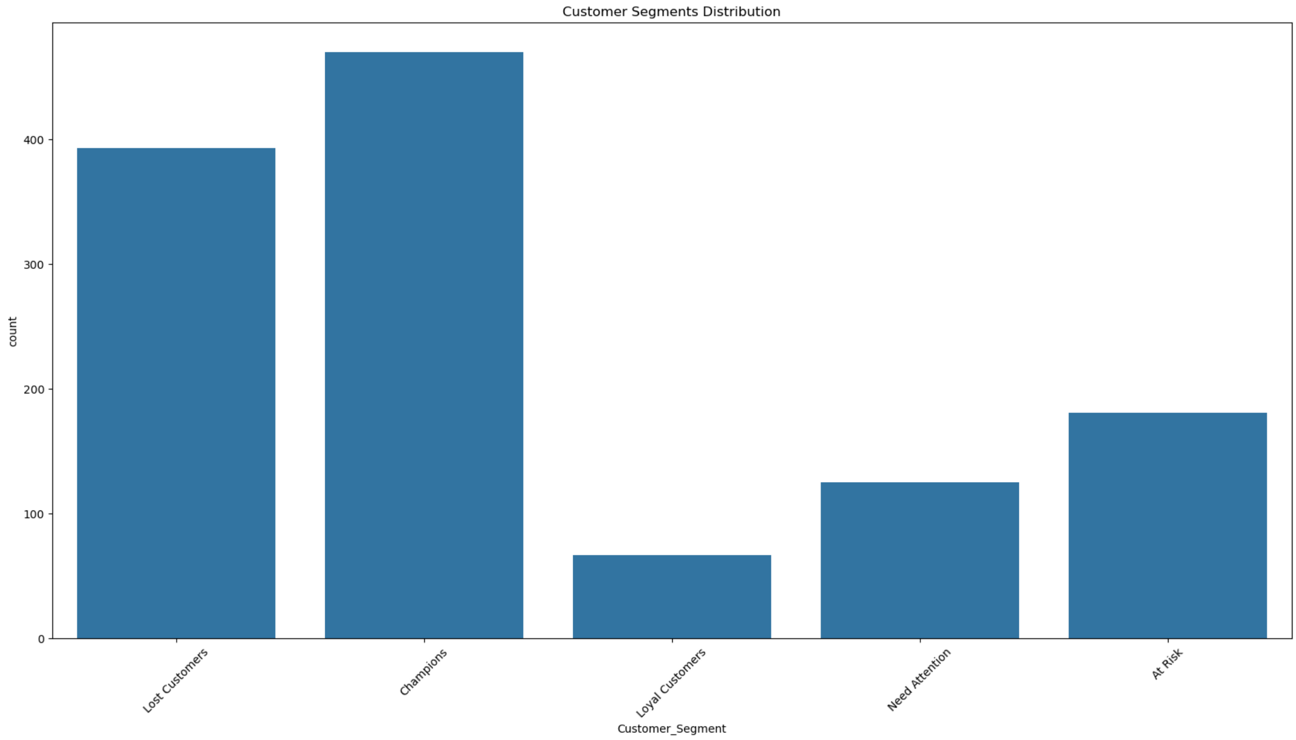
- **K-Means Clustering:** K-Means is an effective unsupervised learning algorithm used for customer segmentation. It groups customers into clusters based on similar purchasing behaviors, making it useful for identifying patterns and trends in customer data. It was chosen for its simplicity and efficiency in handling large datasets.
- **Logistic Regression:** This model was selected for its interpretability and effectiveness in binary and multi-class classification problems. Since we aim to classify customers based on their RFM scores, logistic regression provides a straightforward and computationally efficient approach.
- **Decision Tree Classifier:** Decision trees were chosen because they handle non-linearity well and provide high accuracy. They also allow for easy visualization of decision-making processes, making them useful for understanding which factors contribute most to customer segmentation.

ASSUMPTIONS

- Customers with recent purchases are more likely to buy again.
- Frequent customers are more valuable to the business.
- Higher monetary value transactions indicate high-value customers.
- Customer behavior follows historical purchasing patterns.

MODEL EVALUATION AND TECHNIQUES

- **The RFM analysis** RFM analysis segments customers based on Recency (last purchase date), Frequency (purchase count), and Monetary value (total spend) to identify high-value customers and optimize marketing strategies. The segments it has classified the data into are:



And the count for each segment is:

RFM_Segment	Customer_Segment	Count
111	Champions	470
444	Lost Customers	393
211	At Risk	181
311	Need Attention	125
411	Loyal Customers	67

- **Silhouette Score (for K-Means Clustering):** Measures how well each data point fits within its assigned cluster. A higher silhouette score (closer to 1) indicates well-separated clusters, while a lower score suggests overlapping or poorly defined clusters.

- **Accuracy Score (for classification models):** Measures the proportion of correctly predicted customer segments out of the total predictions. Higher accuracy signifies better classification performance.

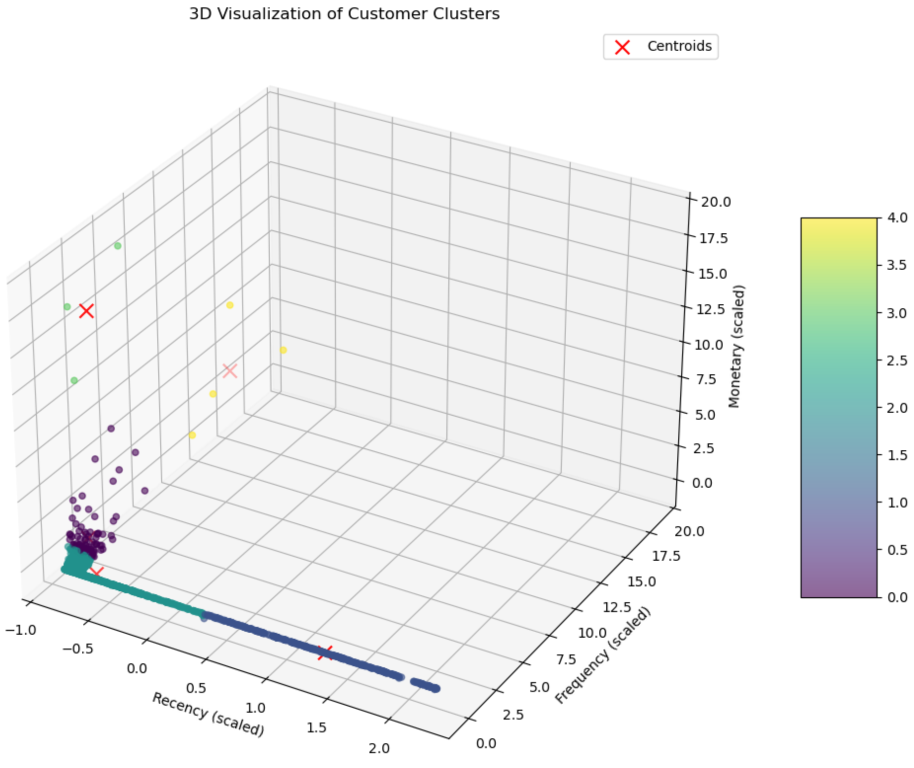
- **Confusion Matrix:** Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives, helping to analyze misclassification rates and understand model reliability.

Model	Silhouette Score / Accuracy
K-Means (Clustering)	0.60
Logistic Regression	0.92
Decision Tree	1.00

INFERENCES DRAWN FROM THE MODEL EVALUATION

- **K-Means Clustering** produced a silhouette score of 0.60, indicating reasonably well-separated customer segments, though further tuning of cluster numbers could enhance segmentation accuracy.

- The 3D visualization of customer clusters from the RFM analysis using K-Means clustering reveals distinct customer segments, with a majority of low-value customers and a few high-value clusters, highlighting opportunities for targeted retention, upselling, and reactivation strategies.



- **Logistic Regression** demonstrated a high 92% accuracy but fine-tuning the data may lead to better segmentation insights.

- **Decision Tree Classifier** performed well, achieving 100% accuracy, precision, recall, and F1-score, indicating that the model perfectly classified customers. However, decision trees can sometimes overfit the data, so future evaluations with pruning techniques could provide a more generalized solution.

- Recency and frequency played a crucial role in customer classification, while monetary value had a moderate impact.

- The store has very few loyal customers, indicating a need for better retention strategies such as loyalty programs and personalized promotions

FUTURE POSSIBILITIES OF THE PROJECT

- **Predicting Customer Churn:** Identifying customers who are likely to stop purchasing and implementing proactive retention strategies.

- **Personalized Marketing Strategies:** Using machine learning to recommend products based on past purchase behavior.

- **Real-Time Customer Insights:** Integrating real-time data streams for dynamic segmentation and instant decision-making.

- **Deep Learning Models**: Exploring neural networks and LSTMs for advanced customer behavior prediction.

- **Automated Customer Retention Programs:** Developing AI-powered systems to automatically identify and engage high-value or at-risk customers.

- **Developing Dynamic Pricing Models:** Implementing machine learning techniques to optimize pricing based on demand patterns.

- **Sentiment Analysis on Customer Reviews:** Using NLP techniques to analyze customer feedback and improve service offerings.

CONCLUSION

This project successfully analyzed customer purchasing behavior and classified customers into 5 meaningful segments.

Machine learning models, particularly the the Logistic Regression, provided highly accurate results in predicting customer categories. The insights gained can help businesses optimize their marketing strategies, improve customer retention, and drive revenue growth.

The Decision Tree also offered excellent accuracy, however due to the possibility of overfitting, we would have to further evaluate the model before using it in the real world.

Future enhancements could involve real-time analytics, deep learning techniques, and integrating more external data sources such as customer reviews and social media interactions. By leveraging AI-driven strategies, businesses can enhance their ability to attract, retain, and engage customers more effectively.

REFERENCES

- Scikit-learn documentation: Linear Regression, Random Forest Regressor