Data Portfolio

E-commerce Customer Analytics for Increased Sales

Problem Statement

Malaysian e-commerce businesses are booming but struggle to understand customer behaviour and personalise experiences. This project aims to:

- Analyse customer segmentation and purchase patterns.
- Provide product recommendations.
- Predict customer lifetime value (CLV) to improve retention and sales.

Dataset

The Online Retail II UCI dataset was used for this study, which contains transactional data for a UK-based online retail store. The dataset includes:

- **InvoiceNo:** Invoice number (cancelled transactions start with 'C').
- **StockCode:** Product code.
- **Description:** Product description.
- **Quantity:** Quantity purchased.
- **InvoiceDate:** Date and time of the transaction.
- **UnitPrice:** Price per unit.
- **CustomerID:** Unique customer identifier.
- **Country:** Country of the customer.

Data Exploration and Cleaning

- Data Cleaning:
  - Cancelled transactions (invoices starting with 'C') were removed.
  - Missing Customer ID values were dropped.
  - A TotalAmount column was created (Quantity * Price).

- Observations:
  - The dataset has 525,461 rows, with some missing Customer ID values.
  - Negative quantities (e.g., -9600) indicate returns or cancelled orders, which were handled appropriately.

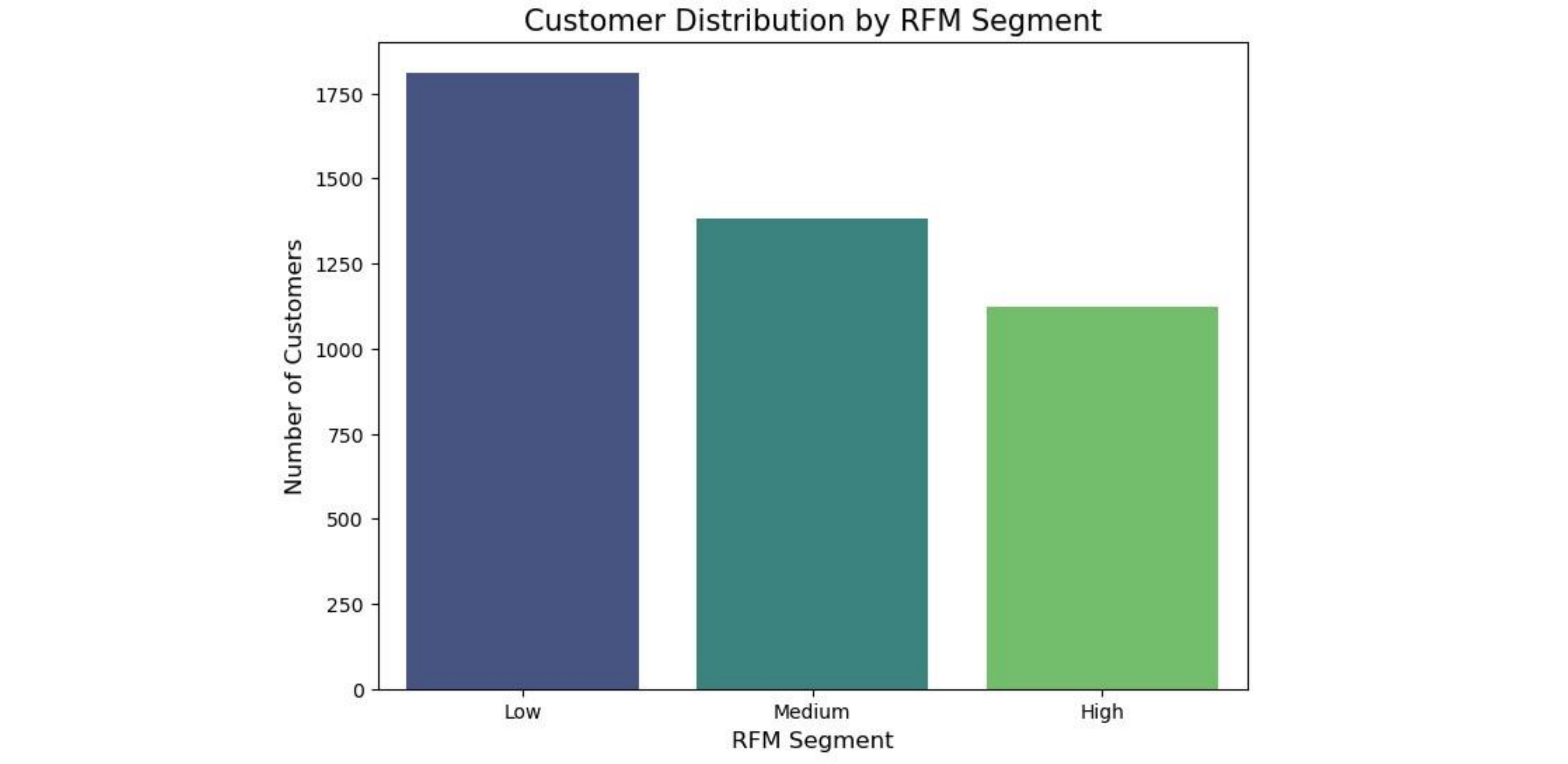
RFM Analysis

- RFM Scores:
  - **Recency:** Days since the last purchase.
  - **Frequency:** Number of unique invoices per customer.
  - **Monetary:** Total amount spent by each customer.

- RFM Segmentation:
  - Customers were segmented into **Low, Medium**, and **High** based on their RFM scores.
  - Manual binning was used for Frequency and Monetary due to issues with pd.qcut().

- Observations:
  - The RFM analysis successfully segmented customers, but there were some NaN values in the MonetaryScore column that were dropped.
  - The final RFM table contains 4,312 rows (customers) with scores and segments.

RFM Analysis



RFM Analysis

- **High-value customers:**
  - **Insights:** Most loyal and profitable customers. They purchase frequently, spend a lot, and have made recent purchases.
  - **Action:** Focus on retaining these customers through loyalty programs, personalised offers, and exclusive perks.

- **Medium-value customers:**
  - **Insights:** Have potential to become high-value customers but may need encouragement to increase their spending or purchase frequency.
  - **Action:** Target them with upselling and cross-selling campaigns and encourage repeat purchases through discounts or rewards.

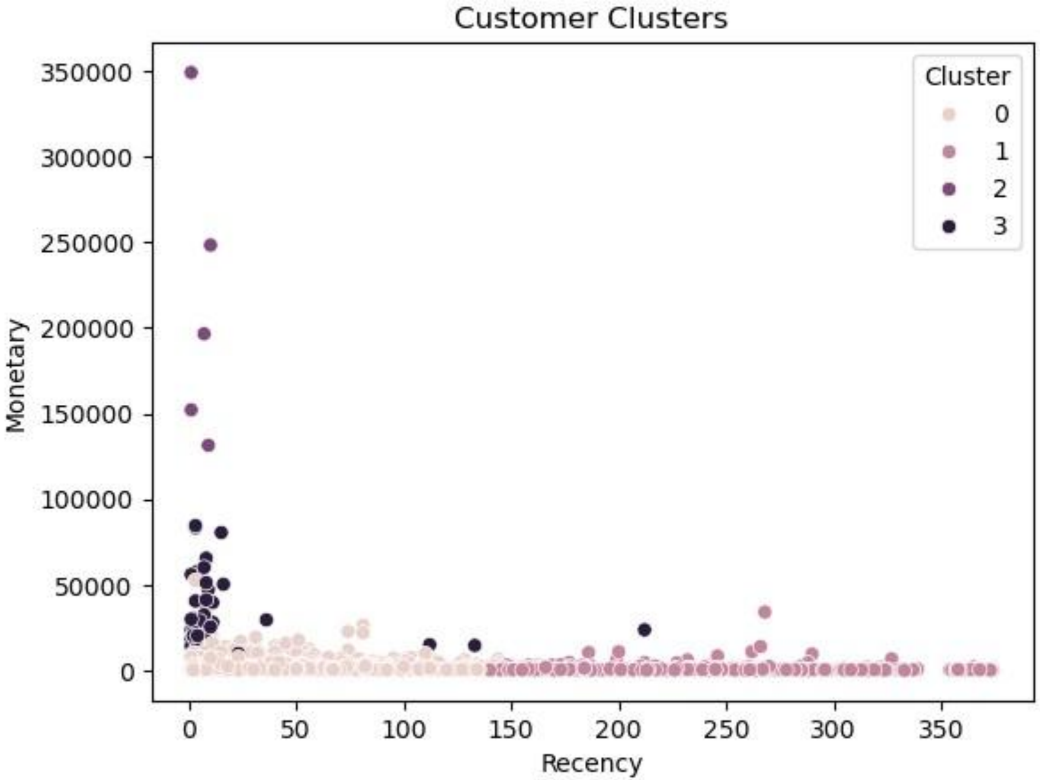
- **Low-value customers:**
  - **Insights:** These customers are either inactive or make infrequent low-value purchases.
  - **Action:** Implement re-engagement campaigns (e.g., win-back offers, personalised emails) to bring them back. If re-engagement fails, consider deprioritizing these customers to focus on more profitable segments.

Customer Segmentation (Clustering)

- K-Means Clustering:
  - Customers were clustered into 4 groups based on their RFM values (Recency, Frequency, Monetary).
  - The clusters were visualised using a scatter plot (Recency vs. Monetary).

- Observations:
  - The clusters show distinct groups of customers based on their purchasing behaviour.
  - For example:
    - Cluster 0: Low recency, high monetary (recent high spenders).
    - Cluster 3: High recency, low monetary (inactive low spenders).

K-Means Clustering



K-Means Clustering

- **Cluster 0 (Recent High Spenders):**
  - **Insights:** These customers have made recent purchases and spent a significant amount.
  - **Action:** Reward them with loyalty programs, exclusive offers, or early access to new products to maintain their engagement.

- **Cluster 1 (Frequent Low Spenders):**
  - **Insights:** Purchase frequently but spend less per transaction. They may be price-sensitive or buying low-cost items.
  - **Action:** Encourage them to spend more by offering discounts on higher-value items or bundling products

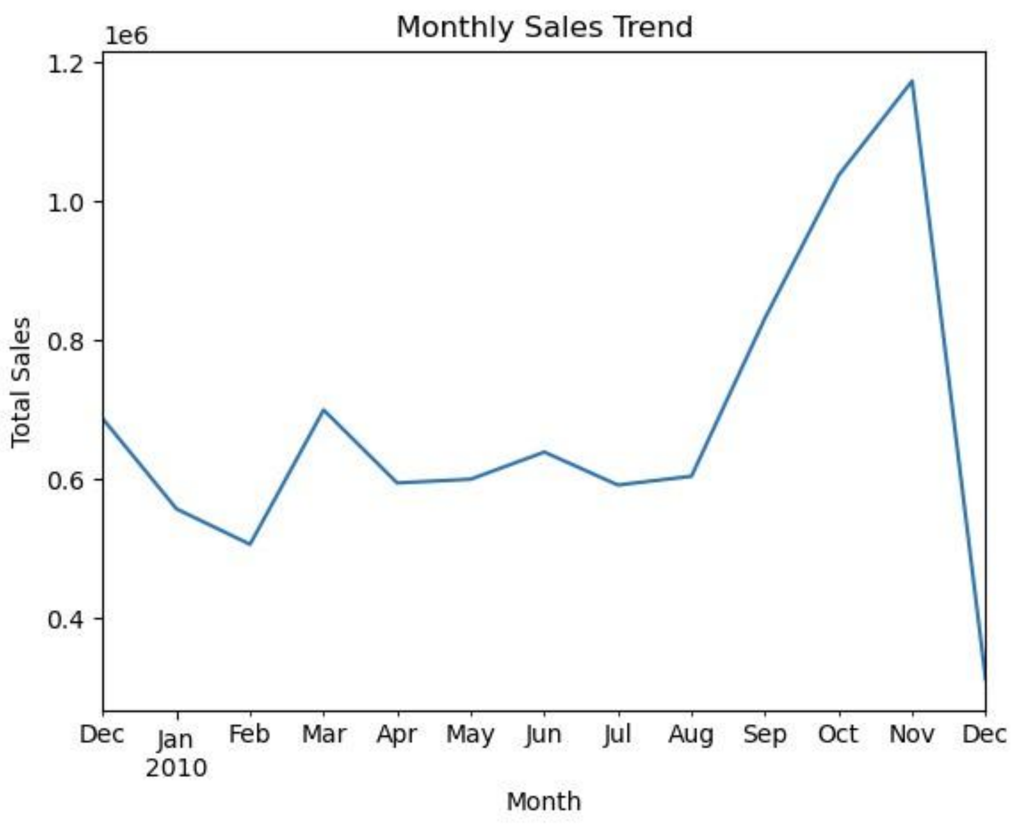
- **Cluster 2 (Inactive Low Spenders):**
  - **Insights:** Either dormant or have churned. They have not made recent purchases and have historically spent little.
  - **Action:** Implement re-engagement campaigns to bring them back. If re-engagement fails, consider deprioritizing these customers.

- **Cluster 3 (Occasional High Spenders):**
  - **Insights:** Make occasional purchases but spend a lot when they do. They may be seasonal or event-driven buyers.
  - **Action:** Target them with personalised offers during peak seasons or special events to encourage repeat purchases.

Purchase Pattern Analysis

- **Monthly Sales Trend:**
  - Sales were aggregated by month, and a line plot was created to visualise trends.

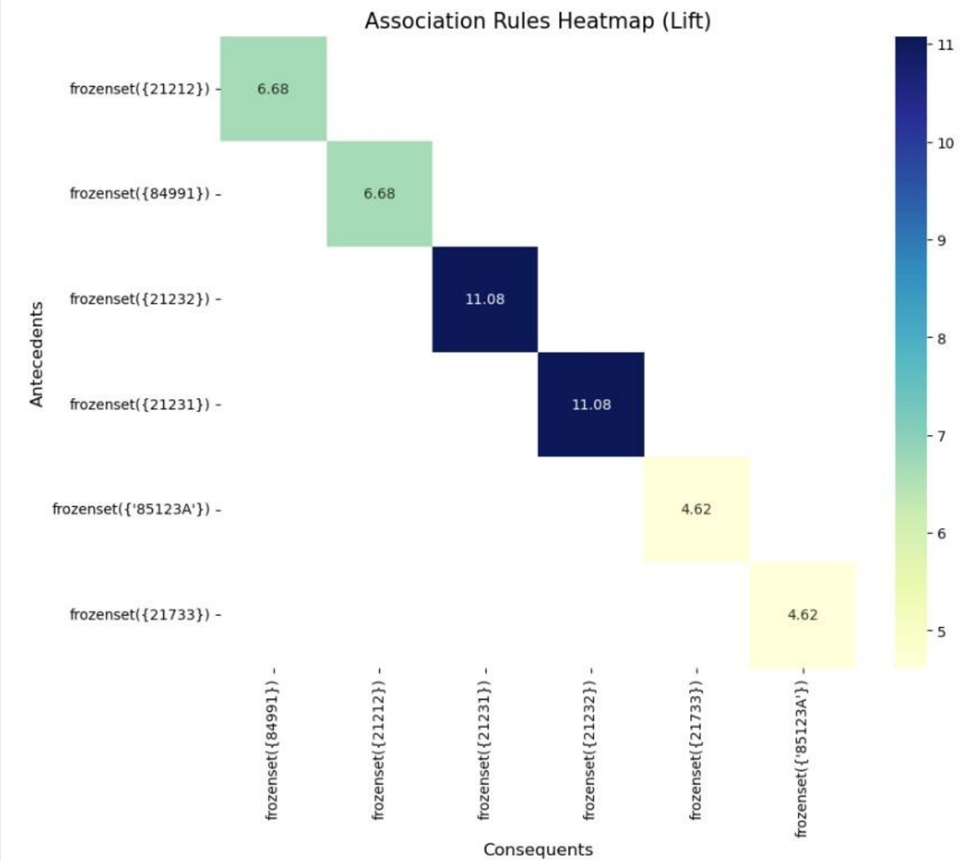
- **Observations:**
  - The plot shows fluctuations in monthly sales with potential peaks during certain months (e.g., holiday seasons).
  - This analysis can help identify seasonal trends and plan marketing campaigns.



Product Recommendations (Association Rules)

- **Apriori Algorithm:**
  - Association rules were generated to identify frequently co-purchased products.
  - A transaction matrix was created, and itemsets with a minimum support of 0.03 were identified.

- **Observations:**
  - Strong associations were found between certain products (e.g., 21212 and 84991 with a lift of 6.68).
  - These rules can be used for cross-selling and product bundling strategies.



Predictive Modelling (Customer Lifetime Value)

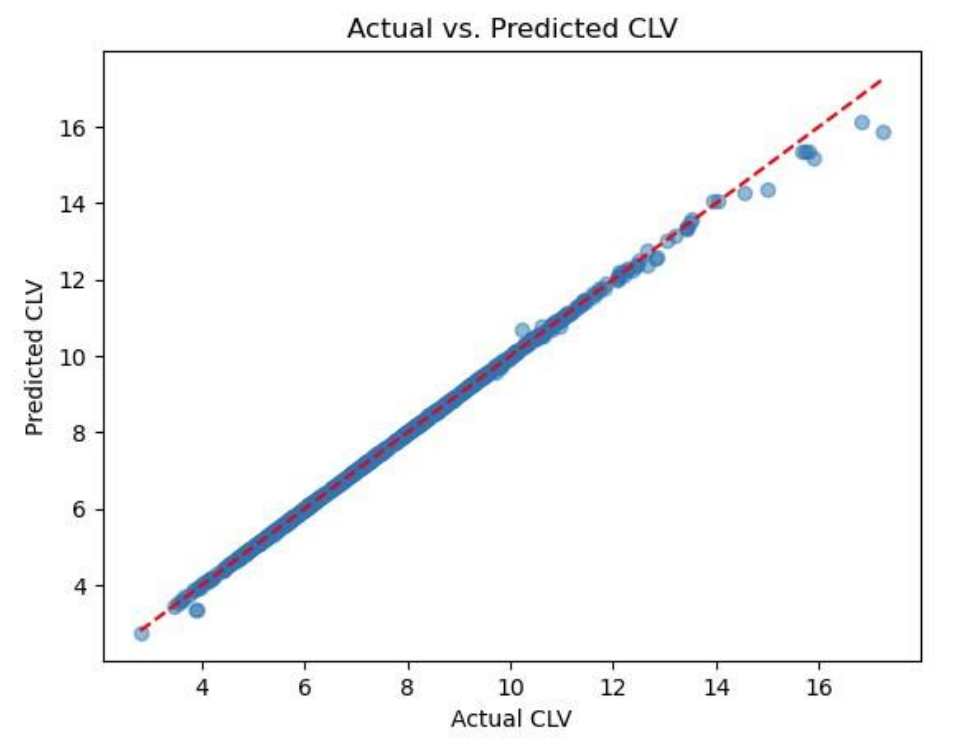
- **CLV Calculation:**
  - CLV was calculated as Monetary * Frequency.
  - A log transformation was applied to handle the large scale of CLV values.

- **Random Forest Model:**
  - The model was trained on features like Recency, Frequency, Monetary, AvgOrderValue, and Tenure.
  - The initial model had a very high MSE (569,572,054,196.656), indicating poor performance.
  - After log transformation and feature engineering, the MSE improved significantly (0.006).

Predictive Modelling (Customer Lifetime Value)

- **Visualisation:**
  - A scatter plot of actual vs. predicted CLV values shows a strong correlation, indicating good model performance.

- **Observations:**
  - The log transformation and additional features significantly improved the model.
  - The model can now be used to predict CLV and identify high-value customers.



Key Insights & Recommendations

- **Customer segmentation**
  - High-value customers:
    - Focus on retaining customers in the High RFM segment and Cluster 0 (recent high spenders).
    - Offer personalised discounts, loyalty programs, or exclusive offers.
  - At-risk customers:
    - Customers in the Low RFM segment or Cluster 3 (inactive low spenders) may need re-engagement campaigns.
    - Send targeted emails or promotions to encourage repeat purchases.

Key Insights & Recommendations

- **Purchase patterns**
  - Seasonal trends:
    - Use the monthly sales trend to plan inventory and marketing campaigns during peak seasons.
  - Product recommendations:
    - Leverage association rules to create product bundles or recommend complementary products (e.g., 21212 and 84991).

Key Insights & Recommendations

- **Customer Lifetime Value (CLV)**
    - High CLV customers:
        - Identify customers with high predicted CLV and focus on increasing their lifetime value through upselling and cross-selling.
    - Low CLV customers:
        - Analyse the behaviour of low CLV customers and implement strategies to improve their engagement and spending.

Areas for Improvement

1. **Handling NaN Values:**
   - The RFM analysis had NaN values in the MonetaryScore column, which were dropped. Consider imputing these values instead of dropping them.

2. **Model Evaluation:**
   - While the MSE improved after log transformation, additional metrics like MAE or R-squared could provide a more comprehensive evaluation.

3. **Scalability:**
   - The association rule mining step could be optimised for larger datasets by using more efficient algorithms (e.g., FP-Growth).