Capstone 3: Customer Segmentation

Problem Statement

- How many distinct customer groups are needed to segment the customers into to allow for targeted marketing campaigns based on purchasing patterns and behavior?

Context

- Customer segmentation helps businesses tailor marketing strategies to different groups of customers based on their behavior. The focus will be on RFM model-based clustering techniques using K-means to group customers to better target specific customer segments depending on the product or products we are trying to sell.

Criteria for Success

- Identify customer segments using a K-means clustering technique based on purchasing patterns and behavior.

Data Sample

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

** Data was gathered from Online Retail Data from UC Irvine Machine Learning Repository* - *[Retail Transactions]*

Data Wrangling

- **Dropped rows missing CustomerID -** 135,080 rows

- **Removal of cancellations and duplicates -** In RFM analysis, the goal is typically to understand a customer's value and engagement level. Cancellations are not representative of active engagement and can skew the understanding of customer loyalty or spending power. Removing cancellations ensures that only positive engagements (purchases) are considered, which provides a more accurate picture of a customer's contribution to the business. (~14,000 rows)

- **Filtered out rows** containing *'M'= Manual, 'POST'= Postage, 'PADS'= Pads to match all cushions, 'DOT'= Dotcom postage, 'CRUK'= CRUK Commission.* This removed UnitPrice outliers

- **Created Recency, Frequency, and Monetary** columns for later analysis.

What is RFM modeling?

RFM Modeling is a marketing analysis technique used to evaluate and segment customers based on their purchasing behavior. RFM stands for **Recency, Frequency, and Monetary**—three key metrics that provide insights into customer engagement and value. It is commonly used in customer segmentation, retention strategies, and personalized marketing

Recency (R):

- **Definition**: Measures how recently a customer made a purchase.
- **Purpose**: More recent purchases indicate higher customer engagement, meaning the customer is more likely to respond to future marketing campaigns.

Frequency (F):

- **Definition**: Measures how often a customer makes purchases.
- **Purpose**: Customers who purchase frequently are often more loyal, and they may generate steady revenue for the business.

Monetary Value (M):

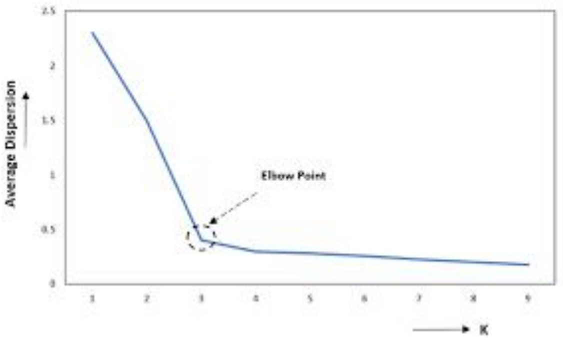
- **Definition**: Measures the total amount spent by a customer.
- **Purpose**: Customers with high monetary value are more valuable to the business, contributing more to the overall revenue.

What is K-means clustering?

- **K-means clustering** is an unsupervised machine learning algorithm used to group similar data points into **clusters**. The goal of K-means is to divide the data into **K** clusters such that data points within each cluster are more similar to each other than to those in other clusters.

Key Concepts

- Centroids
- Number of Clusters (K)
- Elbow method



Strengths

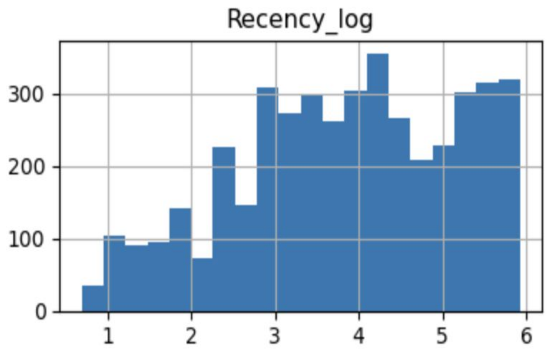
- Simple and easy to implement.
- Scales to a large number of data points.
- Works well when clusters are clearly separated and have a spherical shape

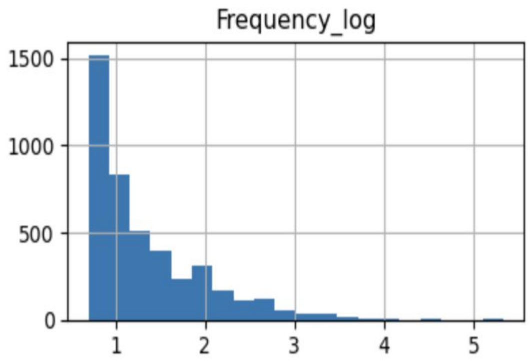
Weaknesses

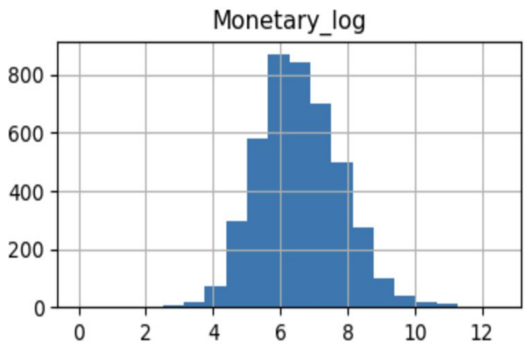
- Sensitive to Initialization
- Number of Clusters (K) Must Be Specified
- Not Ideal for Non-spherical Clusters
- Sensitive to Outliers

Exploratory Data Analysis

RFM Distributions

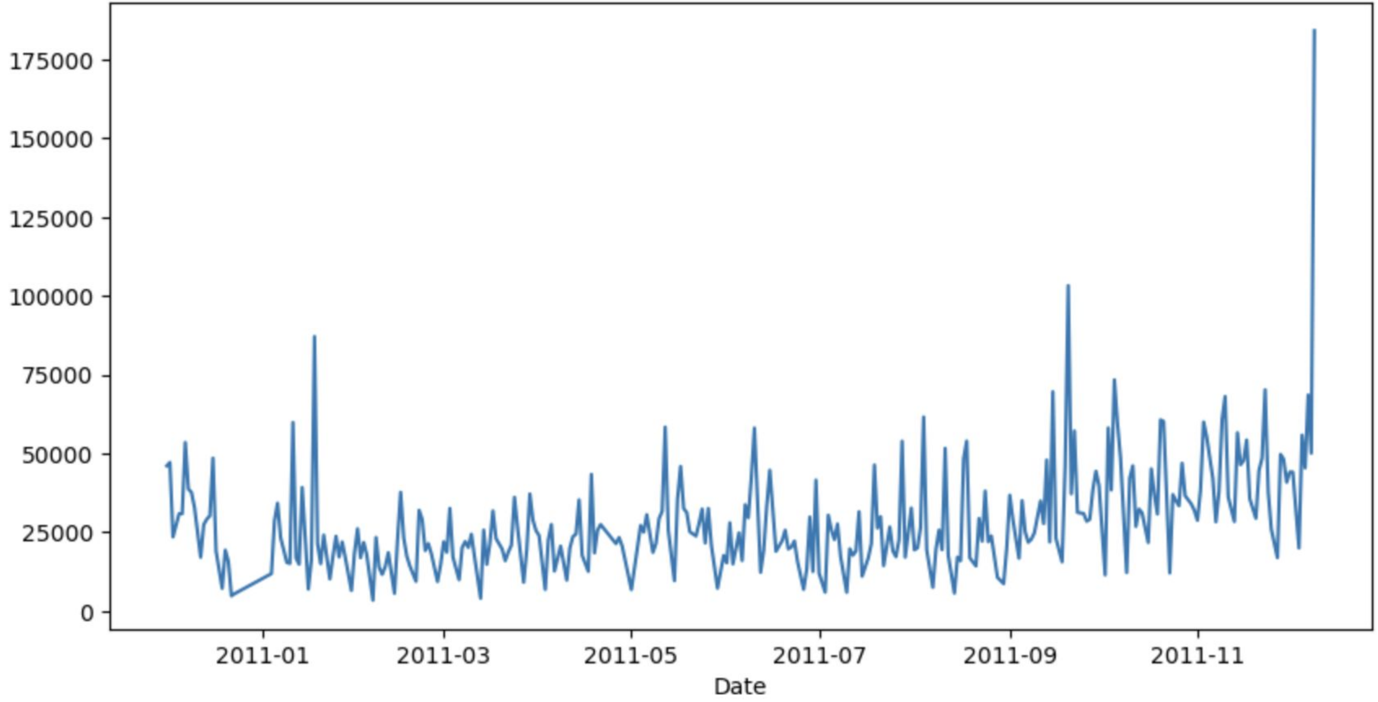






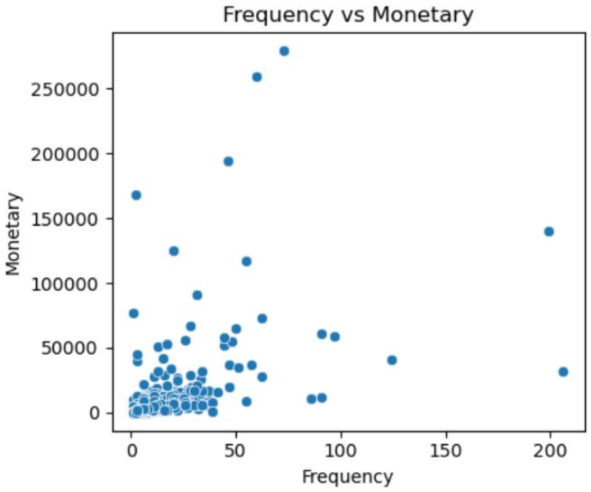
Exploratory Data Analysis

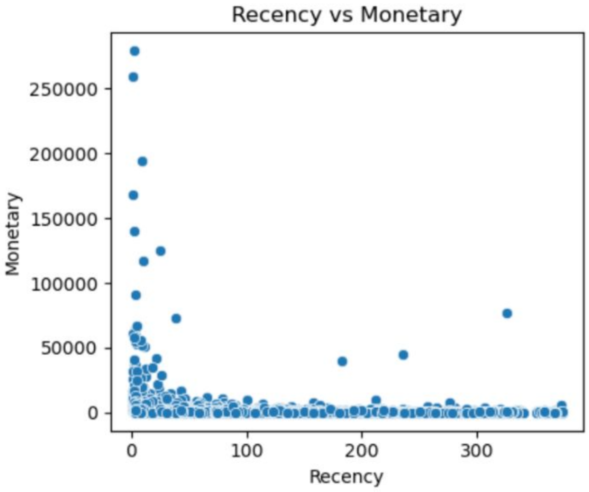
Time Series View

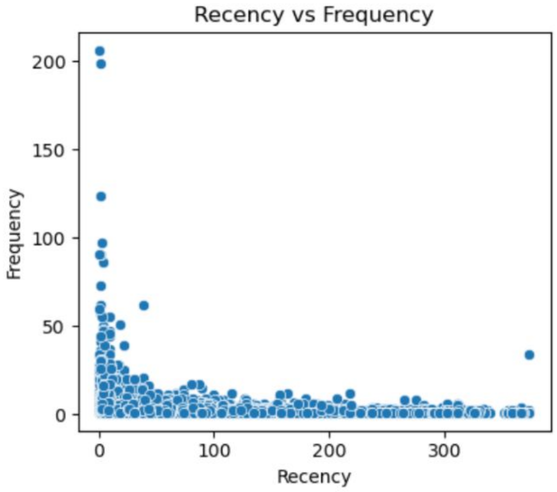


- *Increased purchases during holiday season (e.g. Christmas)*

Exploratory Data Analysis

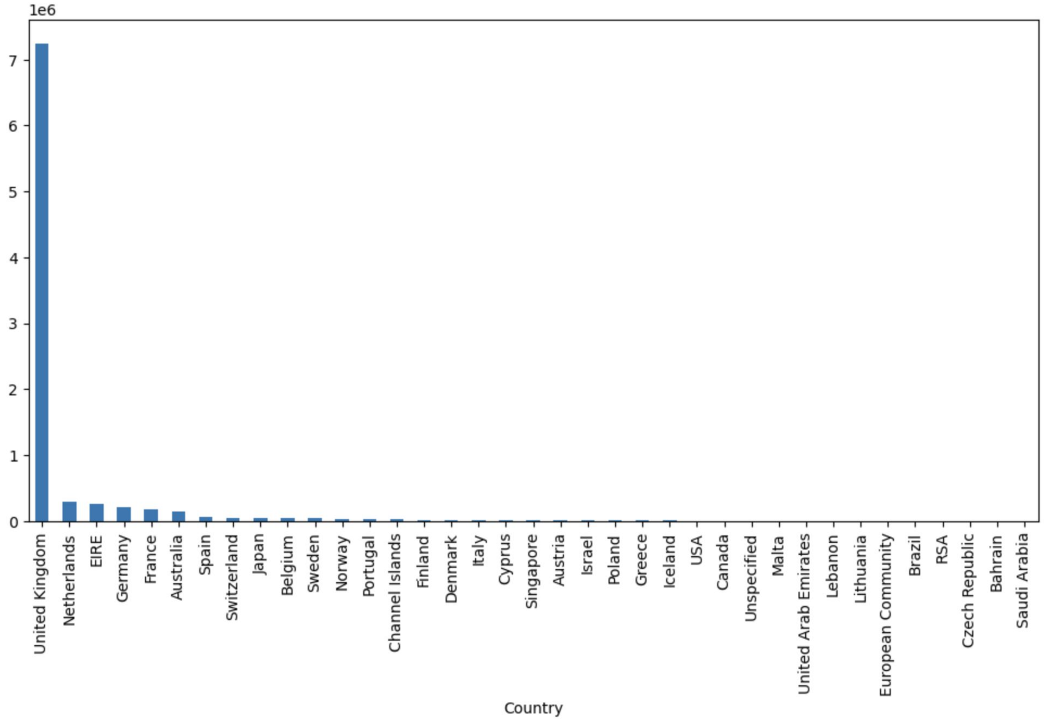


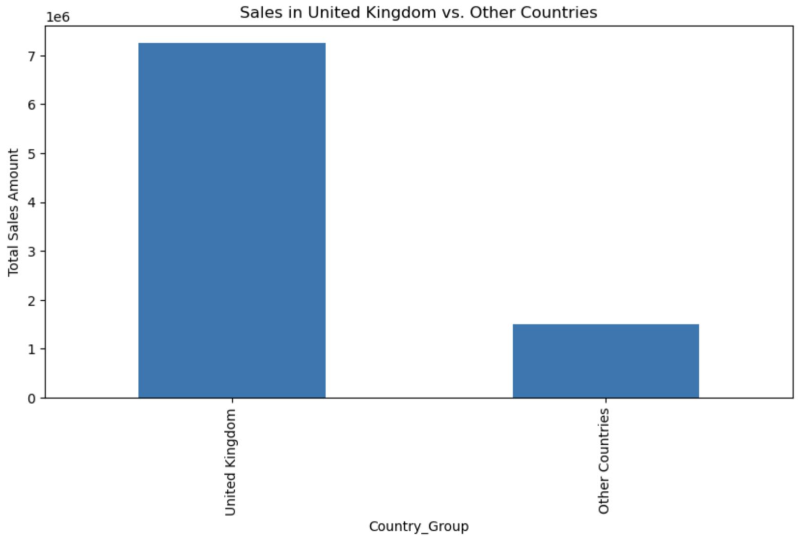




- *Slight positive trend between Frequency and Monetary*
- *Negative relationship between Recency and Monetary*
- *No strong correlation between Recency and Frequency*

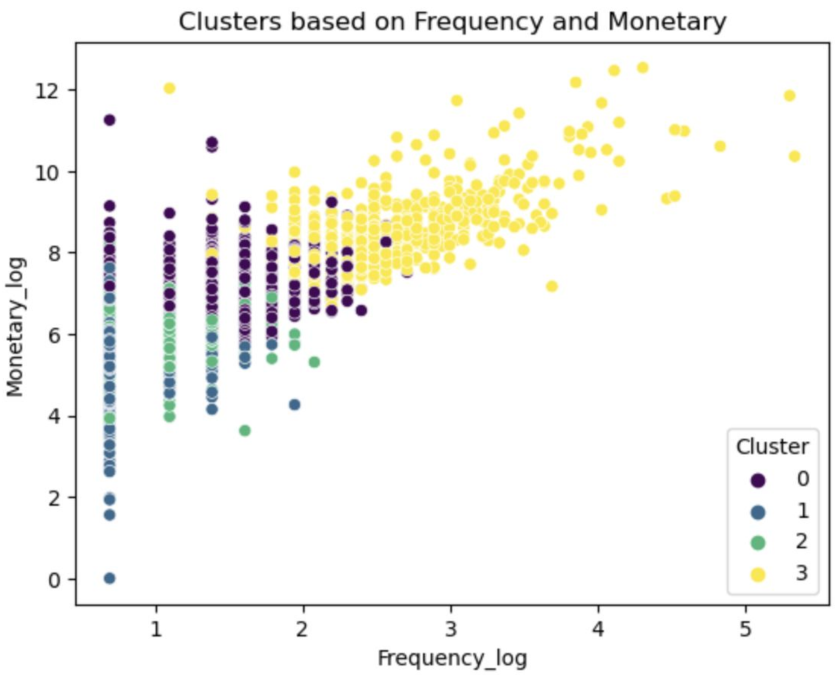
Exploratory Data Analysis

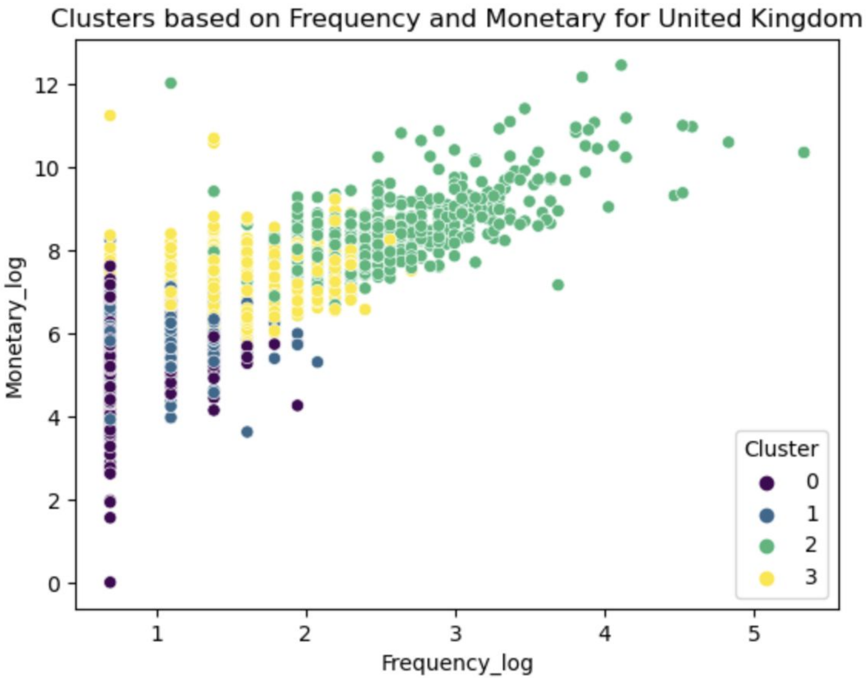




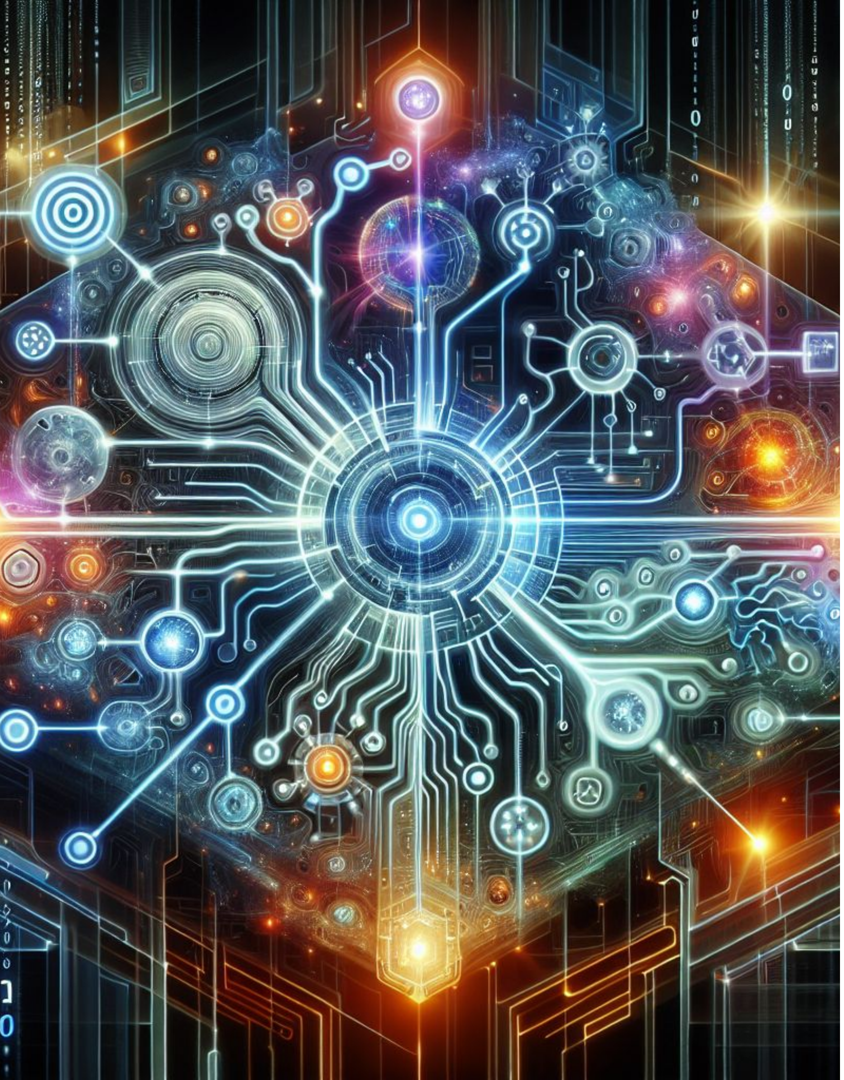
- *The UK makes up the ~83% of the overall sales with all other countries totaling ~17%*

Exploratory Data Analysis





- *The scatter plot for the entire dataset shows that clusters are more mixed, especially for higher values of frequency and monetary, with notable overlap among Clusters 0, 1, and 3.*
- *When focusing on the UK alone, there appears to be a more distinct grouping of the data points, especially with Cluster 2 (green), which is more defined and distinct from other clusters.*



Methodology

Three classification models were employed:

- Random Forest
- XGBoost
- Ensemble of Random Forest, XGBoost, & Logistic Regression

Why these models?

Random Forest

- **Interpretability**: Random Forest is an ensemble of decision trees, making it easier to interpret relative to more complex algorithms. You can understand the feature importance, which is useful for understanding which features contribute most to the clusters.
- **Handling Non-Linearity and Complex Relationships**: Random Forest can *capture complex, non-linear relationships in the data*, which is especially useful in customer segmentation where interactions between features like income, behavior, and demographics may be intricate.

XGBoost (Extreme Gradient Boosting)

- **Handles Complex Patterns**: Like Random Forest, XGBoost *captures complex, non-linear relationships and interactions*, which makes it well-suited for modeling customer behavior and the factors driving different segments.
- **Feature Importance**: XGBoost also provides *insights into feature importance*, which can help you understand the main characteristics that define each customer segment.
- **Speed and Efficiency**: It is optimized for both *speed and performance*, allowing it to handle large datasets efficiently, which is often the case in customer segmentation.

Ensemble of Random Forest, XGBoost, & Logistic Regression

- **Combining Strengths**: The ensemble *combines the strengths of each model*—Random Forest for robust feature handling, XGBoost for accuracy, and Logistic Regression for interpretability. This often results in a more **balanced** and **generalizable** model that captures diverse aspects of the data.
- **Reduced Bias and Variance**: By combining multiple models, an ensemble helps reduce **bias** and **variance**, making it less likely to overfit and better at generalizing to unseen data.
- **Logistic Regression for Linear Components**: Including Logistic Regression allows the ensemble to *capture linear relationships* that Random Forest and XGBoost might overlook. It helps provide a **simpler interpretation** of the decision boundaries in the segmentation.

Model Performance

Predictive Modeling: To predict cluster assignments for new customers, I trained a Random Forest classifier using the RFM features. The initial model achieved an accuracy of **97%**. Subsequent optimization efforts included:

- **GridSearchCV:** Used to fine-tune hyperparameters, but resulted in in no improvements of model performance.
- **Bayesian Optimization**: Used to fine-tune hyperparameters, which resulted in minor improvements in model performance.
- **XGBoost**: Evaluated for its ability to handle complex relationships within the data, achieving similar accuracy to the Random Forest model but with increased computational cost.
- **Ensemble Methods**: Combined multiple models, including Random Forest and XGBoost, to improve overall accuracy. The ensemble model achieved an accuracy of **98%**, showing a slight improvement over the individual models.

Model Performance

RANDOM FOREST				
customer segment	precision	recall	f1-score	support
0	0.97	0.98	0.98	442
1	0.97	0.96	0.96	236
2	0.98	0.97	0.98	187
3	0.97	0.96	0.96	311
accuracy			0.97	1176
macro avg	0.97	0.97	0.97	1176
weighted avg	0.97	0.97	0.97	1176

RANDOM FOREST w/ GridSearchCV				
customer segment	precision	recall	f1-score	support
0	0.97	0.98	0.98	442
1	0.97	0.96	0.96	236
2	0.98	0.97	0.98	187
3	0.97	0.96	0.96	311
accuracy			0.97	1176
macro avg	0.97	0.97	0.97	1176
weighted avg	0.97	0.97	0.97	1176

RANDOM FOREST w/ Bayesian Optimization				
customer segment	precision	recall	f1-score	support
0	0.98	0.98	0.98	442
1	0.95	0.97	0.96	236
2	0.98	0.97	0.97	187
3	0.97	0.95	0.96	311
accuracy			0.97	1176
macro avg	0.97	0.97	0.97	1176
weighted avg	0.97	0.97	0.97	1176

RANDOM FOREST w/ XGBoost				
customer segment	precision	recall	f1-score	support
0	0.97	0.98	0.98	442
1	0.97	0.96	0.97	236
2	0.98	0.99	0.99	187
3	0.97	0.96	0.96	311
accuracy			0.97	1176
macro avg	0.97	0.97	0.97	1176
weighted avg	0.97	0.97	0.97	1176

Best overall performance

RANDOM FOREST w/Ensemble Model				
customer segment	precision	recall	f1-score	support
0	0.98	0.99	0.98	442
1	0.97	0.96	0.96	236
2	0.98	0.99	0.98	187
3	0.97	0.96	0.97	311
accuracy			0.98	1176
macro avg	0.97	0.97	0.97	1176
weighted avg	0.98	0.98	0.98	1176

Conclusion

- The segmentation of customers using RFM and machine learning provided valuable insights into customer behavior and spending patterns. With effective marketing strategies tailored to each customer segment, the client can enhance customer retention, maximize revenue, and allocate resources more efficiently.

Future Work

- **Behavioral Analysis**: Conduct additional research into the purchasing behavior of each cluster, including product preferences and buying triggers. This analysis could support more refined targeting strategies and improve the effectiveness of cross-selling.
- **Seasonality Analysis**: Investigate purchasing patterns based on seasonality to identify peak times for different customer segments and align marketing campaigns with those periods.
- **Customer Lifetime Value (CLV) Prediction**: Develop a model to predict CLV for each customer segment, allowing for a better understanding of the long-term value of each cluster and helping prioritize marketing resources effectively.

Recommendations

1. **Target High-Value Customers (Cluster 2)**: Use loyalty programs, exclusive deals, and early access to products to maintain engagement and retain these valuable customers.
2. **Grow Mid-Value Customers (Cluster 0)**: Encourage higher spending by offering personalized promotions, bundle offers, or cross-selling opportunities.
3. **Re-engage Low-Value Customers (Clusters 1 and 3)**: Use discounts, incentives, and educational content to encourage repeat purchases and drive higher engagement.