Customer Segmentation and Lifetime Value Prediction

Santhosh BV

1 Objective

The objective of this project is to analyze customer purchase data and segment customers into distinct categories that can support the design of targeted marketing strategies. To effectively prioritize these segments, it is essential to have a metric that reflects the customers' future likelihood to purchase, based on their historical buying behavior. To achieve this, a predictive model is developed to estimate the purchase value of each customer in the upcoming quarter (four months). This estimate is referred to as the predicted Customer Lifetime Value (CLV), which is typically calculated for a defined period; here, the time horizon is four months. By identifying which customer segments are expected to generate the highest predicted CLV, businesses can optimize their marketing efforts, manage inventory more efficiently, and minimize missed opportunity costs.

2 Dataset

The dataset chosen for this experiment must include a unique customer identifier, purchase date, quantity, and total purchase amount. To meet these requirements, we use the E-Commerce dataset from the UCI Machine Learning Repository. This dataset contains transaction records from a UK-based retailer spanning one year, from December 1, 2010 to December 9, 2011.It consists of eight features: invoice number, stock code, product description, quantity purchased, invoice date, unit price, customer ID, and country. These features provide all the necessary information to analyze customer purchasing behavior, segment customers, and predict their future purchase value. Hence, this dataset is well-suited for our study and objectives.

3 Preprocessing

The dataset contains 541,910 records before any preprocessing. The InvoiceNo and StockCode columns are alphanumeric, Description and Country are strings, Quantity is an integer, UnitPrice is a real number in GBP, and InvoiceDate is in MM/dd/yyyy HH:MM format. There are 4,372 unique customer IDs and orders from 38 different countries. We cannot use the number of unique stock codes or invoice numbers as the number of transactions, because the data includes return orders and fees paid to third-party vendors. The quantities purchased range from -80,995 to 80,995, and unit prices range from -11,062 to £38,970. Negative quantities represent return orders or fees, and negative prices are either outliers or incorrect entries. Each invoice number had multiple entries for different products.

Column	Unique Count
InvoiceNo	25,900
StockCode	4,070
Description	4,223
Quantity	722
InvoiceDate	23,260
UnitPrice	1,630
CustomerID	4,372
Country	38

Table 1: Unique Count of Columns

From the missing values graph, we can see that 25% of the rows do not have a customer ID, so these rows are dropped, as the study requires a customer ID to proceed. Rows with negative quantities or prices are also removed, since we are only interested in actual purchases, not returns. Duplicate records are dropped, keeping a single copy. To facilitate further analysis of the data, revenue was derived from quantity and unit price. Sequentially, 25% of the data was removed due to missing customer IDs. 2.19% of the remaining data was dropped due to negative quantity, 0.01% due to zero or negative unit price, and 1.28% of duplicate entries were removed. After these steps, all remaining 392692 entries were valid based on their respective data types.

3.1 Feature Engineering

Our current preprocessed dataset contains transactions identified by invoice number. To segment customers based on their purchasing behavior and predict their future value (CLV), we need to create a dataset aggregated at the customer level. This enables features that express a customer's willingness to buy in the future.

The most commonly used features for this purpose are RFM features: Recency, Frequency, and Monetary. Additionally, to improve segmentation and predictions, we include Average Order Value (AOV) and Tenure.

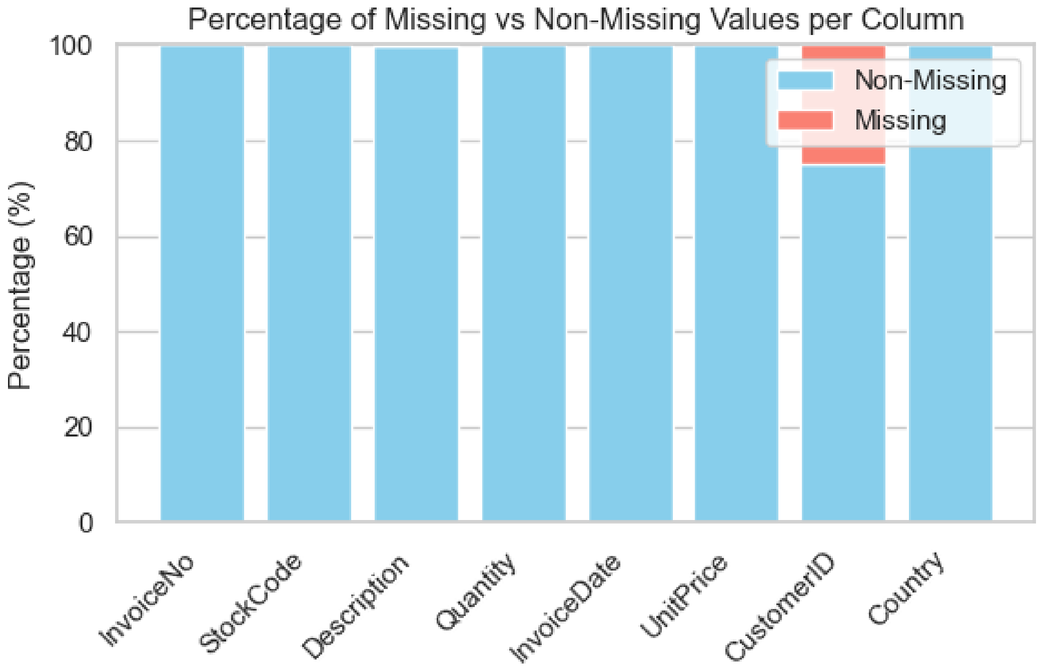

Figure 1: Visualization of Missing Data

Recency: The number of days since the customer's most recent purchase. It indicates customer engagement and activity. Recent purchasers are more likely to buy again.

Frequency: The total number of purchases made by the customer in the given period. It reflects the customer's willingness to do business with the company and is indicative of loyalty and future retention.

Monetary: The total value of purchases made by the customer. It highlights the spending behavior of the customer and their importance to the business.

Average Order Value (AOV): The value of purchase per order. AOV complements Monetary by differentiating between customers who buy infrequently but spend a lot per order versus those who buy frequently with lower spend per order.

Tenure: The number of days between the first and last purchase of the customer. It measures the length of the relationship with the company and provides insight into customer longevity. While it does not account for economic behavior directly, when combined with Frequency and Monetary, it adds an important dimension for segmentation and prediction.

The general preprocessing steps are completed and processes to handle data skew, varying range of values are dealt at the time of classifying or regressing.

4 Insights on Customer Behavior

To analyze and gain insights from the data, various visualizations were plotted. From these, we can infer that most of the company's revenue is generated domestically in the UK, while other European countries such as the Netherlands, Ireland, Germany, and France also contribute. Sales peaked during September, October, and November, with the average order value reaching its highest in December 2011. Most customers made fewer than 50 purchases during this period, though there is a long tail of high-frequency buyers.

Purchasing behavior was fairly consistent from Tuesday to Friday, peaking on Thursday, lowest on Sunday, and slightly improving on Monday, with no transactions recorded on Saturday. Most purchases occurred between 10:00 and 15:00, with the maximum at 12:00. Additionally, 65.6% of customers made more than one purchase, indicating a strong level of customer retention. This highlights the importance of segmenting customers, predicting future behavior, and devising strategies tailored to their purchasing patterns.

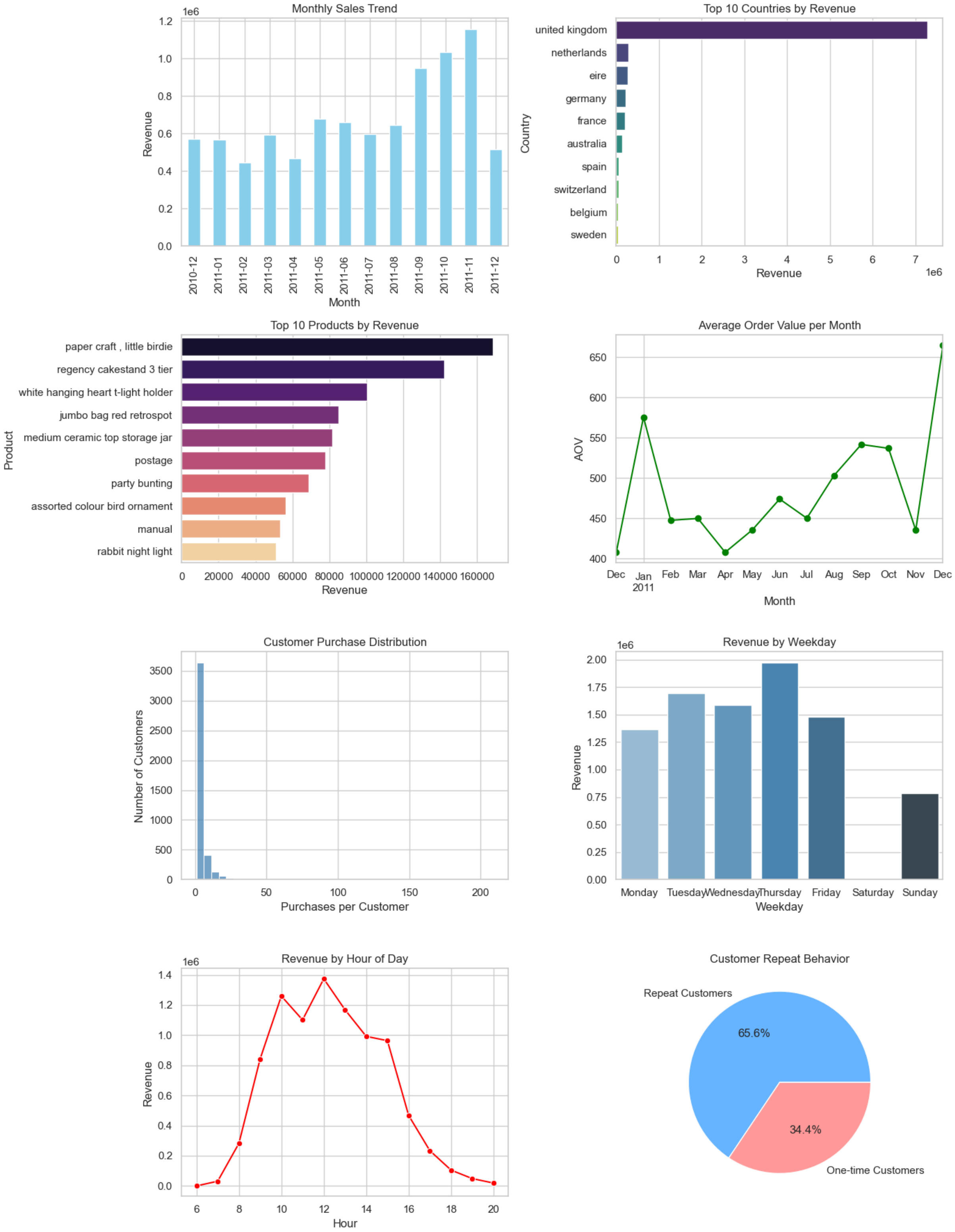

Figure 2: Insights on Customer Behavior

5 Customer Segmentation

To understand the distribution of values across features, histograms were initially plotted. The visualizations revealed a pronounced right skew in several features. To mitigate the influence of these long-tailed higher values, a log transformation was applied. However, since some features, such as tenure, contain zero values, a direct log transform was not feasible. Therefore, a log(1 + x) transformation was employed to handle zeros effectively.

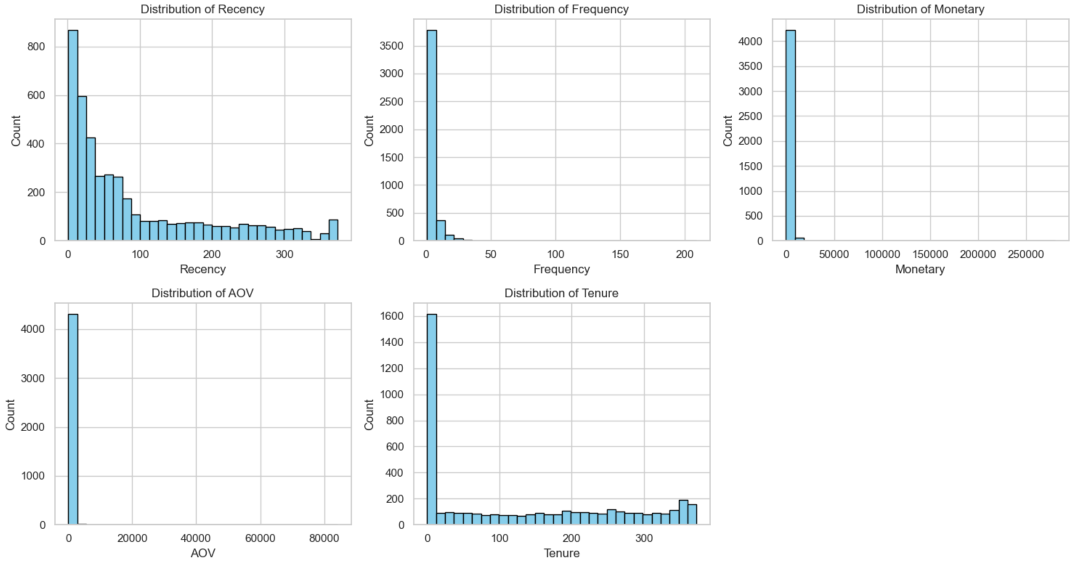

Figure 3: Histogram of features before preprocessing

After addressing skewness within features, it was essential to compare values across features. Features like Monetary and AOV were on a much larger scale compared to others. To prevent the model from overemphasizing these high-scale features, standard scaling was applied to bring all features to a comparable scale.

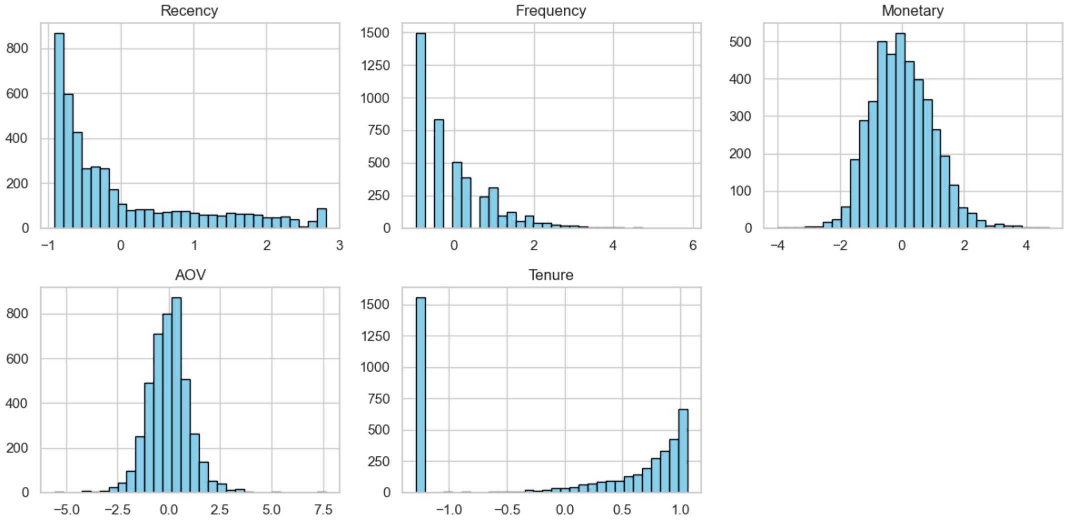

Figure 4: Histogram of features sfter preprocessing

For visualizing the overall structure of the data, Principal Component Analysis (PCA) was performed, and the first two principal components (PC1 and PC2) were plotted. The plots indicated the presence of two distinct clusters. To gain deeper insights, clustering methods were applied: KMeans, which is distance-based, and Gaussian Mixture Model (GMM), which is probability-based.

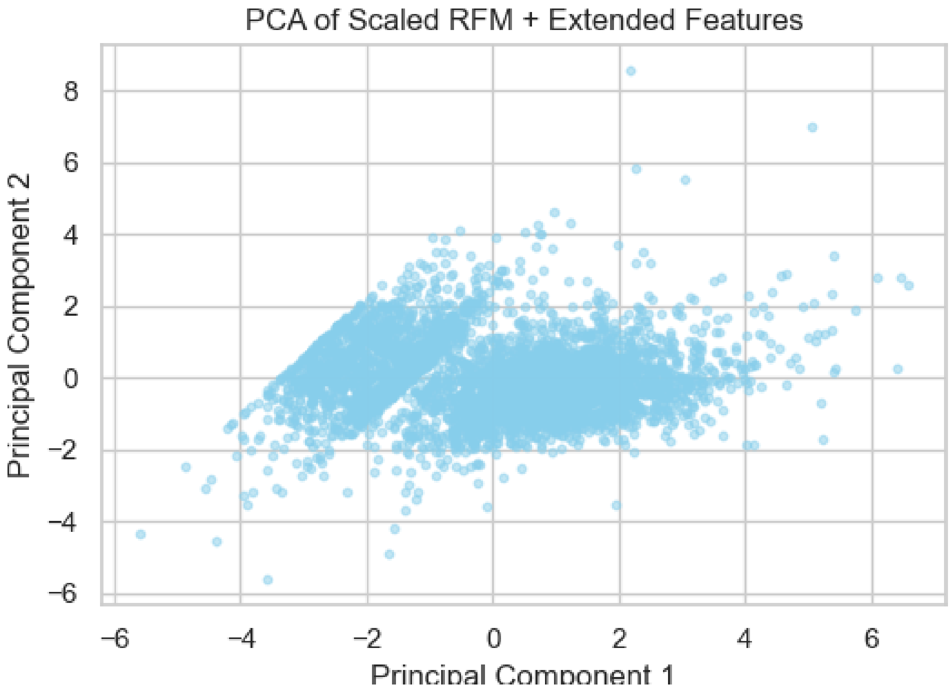

Figure 5: Visualization using PCA

The optimal number of clusters for KMeans was determined using the silhouette score, which evaluates intra-cluster cohesion and inter-cluster separation. In contrast, for GMM, the Bayesian Information Criterion (BIC) was used, which measures the likelihood of points belonging to clusters while penalizing overly complex models. A lower BIC indicates better clustering performance. KMeans suggested two clusters, whereas GMM identified seven. While the silhouette score favors the two clusters from KMeans, using only two clusters would not provide sufficient granularity for effective customer segmentation.

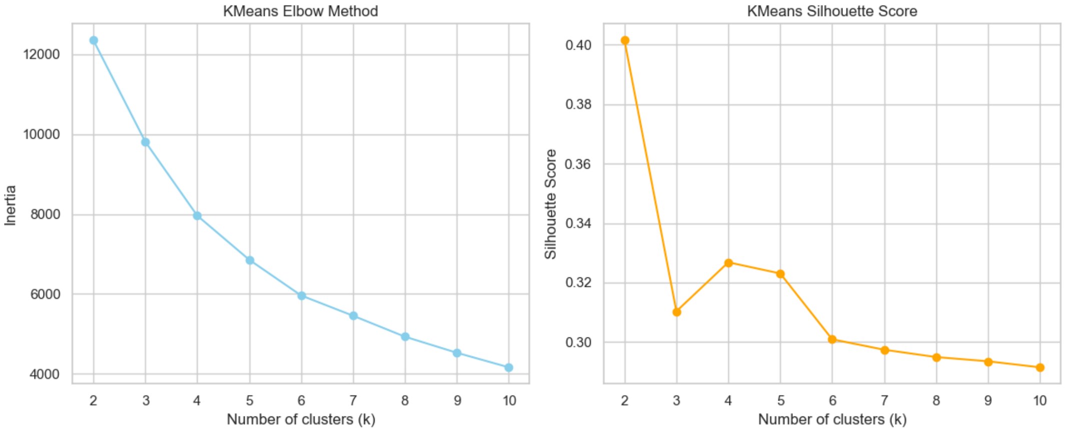

Figure 6: Optimal cluster analysis for K-Means

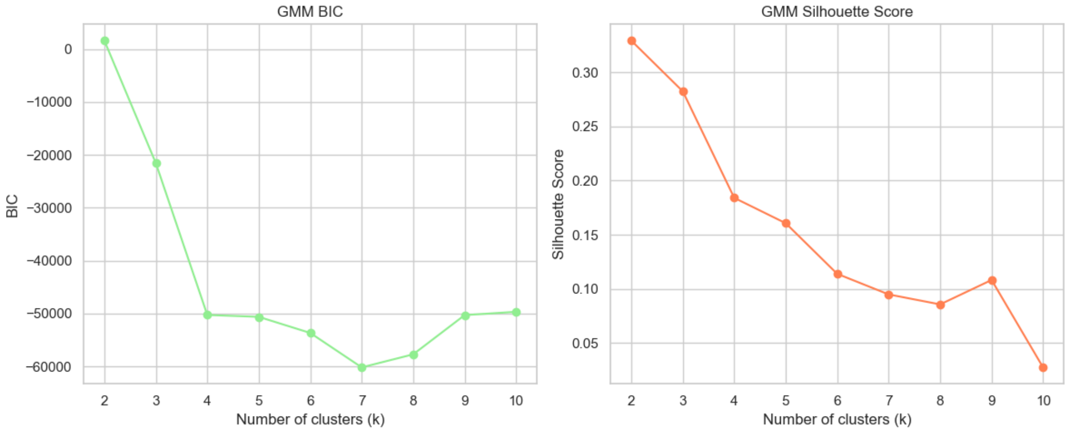

Figure 7: Optimal cluster analysis for GMM

PCA plots with clusters from both KMeans and GMM were examined. It was evident that GMM captured more enhanced clustering structure, effectively distinguishing different customer behaviors. Based on this analysis, seven clusters from GMM were selected for further examination.

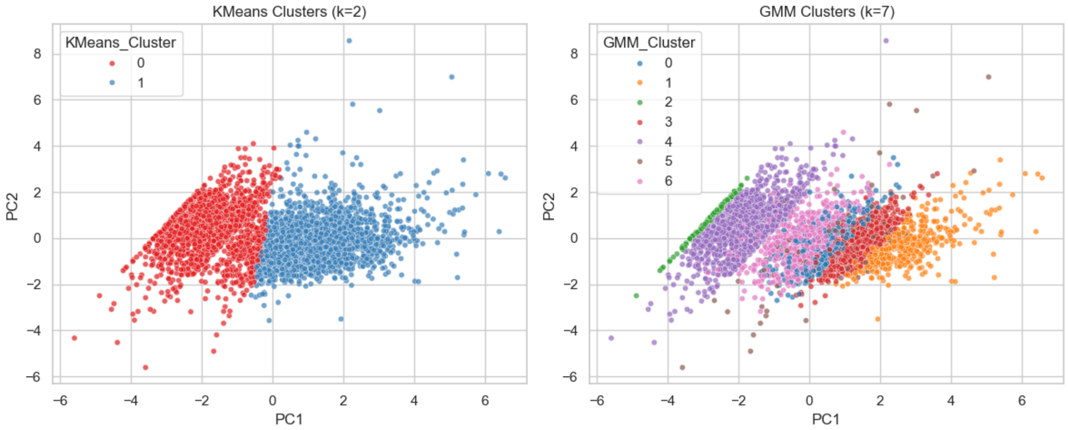

Figure 8: Cluster Visualization using PCA

5.1 Cluster Analysis

- Cluster 1 consists of long-term, frequent, and high-value customers. While not having the highest AOV, they are the second most valuable group and represent a minority class at 11.78% of the customer base.
- Cluster 5 represents relatively new customers with high frequency and the highest AOV. Despite being only 2.4% of the total customers, they contribute significantly to revenue, making them another high-value segment.
- Cluster 2 comprises one-time customers who placed multiple orders on a single day but did not return, suggesting potential dissatisfaction. These customers also account for 2.4% of the base, and their feedback and return data should be analyzed to improve satisfaction.
- Cluster 4 includes customers who made recent purchases but never returned, forming the largest one-time buyer group at 32%. Overall, clusters 2 and 4 together account for approximately 34% of the customers.
- Clusters 0, 3, and 6 represent medium-value customers, ranked in descending order based on recency, frequency, monetary value, and tenure. These clusters share similar AOV levels and collectively account for 50% of the customers.

GMM_Cluster	Recency	Frequency	Monetary	AOV	Tenure	CustomerCount	CustomerPct
0	70.26	3.00	1102.58	367.53	164.07	500	11.53
1	14.78	16.74	9482.13	467.68	325.71	511	11.78
2	366.91	1.00	239.79	239.79	0.00	103	2.37
3	37.70	5.08	1939.87	378.85	234.41	899	20.72
4	141.10	1.00	423.95	423.95	0.00	1390	32.04
5	122.09	7.27	5420.59	1442.20	125.31	104	2.46
6	94.13	2.00	684.57	342.29	112.69	831	19.16

Table 2: GMM Cluster Statistics

In summary, high-value customers (clusters 1 and 5) constitute 15% of the base, one-time buyers (clusters 2 and 4) make up 34%, and medium-value customers (clusters 0, 3, and 6) account for 50%. This segmentation provides actionable insights for targeted marketing, retention strategies, and revenue optimization.

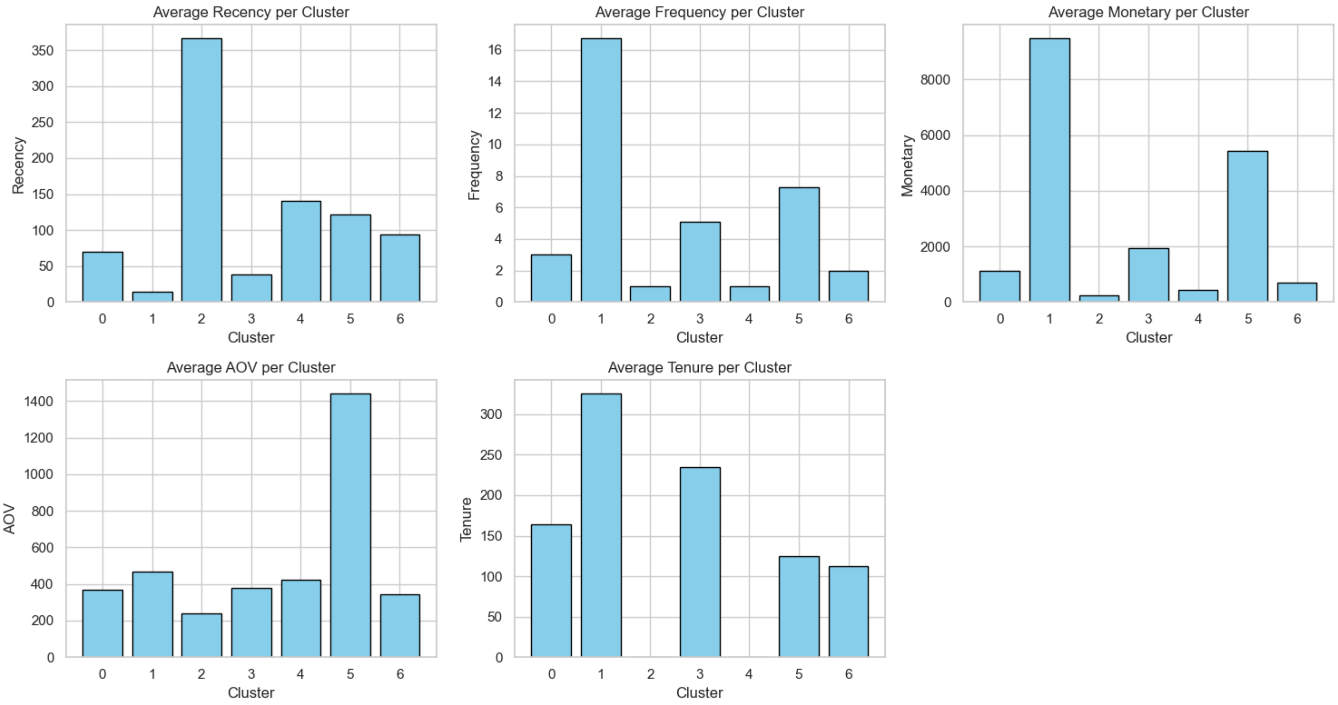

Figure 9: Visualization of GMM clustering statistics

6 CLV Prediction

The Customer Lifetime Value (CLV) for the next quarter is to be predicted using historical customer purchase data. The dataset has been preprocessed to include features such as Recency, Frequency, Monetary value, Average Order Value (AOV), and Tenure.

To train the model, a training and testing dataset is required. However, since future purchases of customers are not available at present, a proxy approach is used in which a model is developed to predict the CLV for the last quarter of 2011 based on data from the previous eight months. This trained model is then applied to predict CLV for the next quarter using the full 12 months of data.

Similar preprocessing steps as used in classification are applied to remove skewness in the data and standardize the features. The RFM metrics, AOV, and Tenure are calculated for the first eight months, and the customer purchase behavior for the next four months is recorded. A new dataset is created where 36% of customers did not make future purchases, and their CLV is set to zero.

The dataset is split into 80% training data and 20% testing data. Multiple regression models, including Linear Regression, Random Forest, Gradient Boosting, and XGBoost, are trained with varying hyperparameters. Grid search is used to identify the best hyperparameters for each model. Model performance is evaluated on the testing set using R^2 and RMSE metrics. Among all models, the Gradient Boosting model performs the best with **$\mathbf{R^2}$ of 0.93 and**

RMSE of 1565 on the training dataset.

Model	Train R^2	Train RMSE	Test R^2	Test RMSE
Linear Regression	0.156	5415.153	0.207	5180.519
Random Forest	0.821	2489.222	0.610	3574.333
Gradient Boosting	0.985	699.155	0.930	1565.064

Table 3: Model Performance Metrics

6.1 CLV Predictions per Cluster

The trained Gradient Boosting model is used to predict the CLV for the next quarter for customers in each cluster. The predicted total revenue per cluster is shown below, and these predictions align with the previous customer segmentation analysis:

- Class 1: High-value customers and a significant portion of total customers. Their contribution to total revenue is the highest at 53%.
- Class 3: Medium-value customers with a higher proportion of the customer base, contributing 23%.
- Class 2: Mostly single-time buyers expected to make few future purchases, contributing the least at 0.2%.
- Other classes follow expected patterns based on their segmentation, creating a clear hierarchy for prioritizing customer segments.

GMM Cluster	Total CLV	Revenue per Person	% Contribution
0	5.284e+05	1056.949	5.910
1	4.536e+06	3876.432	53.590
2	1.992e+04	184.671	9.220
3	1.951e+06	2146.976	22.810
4	4.973e+05	357.216	5.870
5	4.936e+05	3861.625	4.720
6	5.302e+05	661.886	6.500

Table 4: Cluster wise predicted CLV

The average CLV per cluster is also explainable based on segmentation. Class 5 has fewer high-value customers, so its average CLV is the second highest after Class 1.

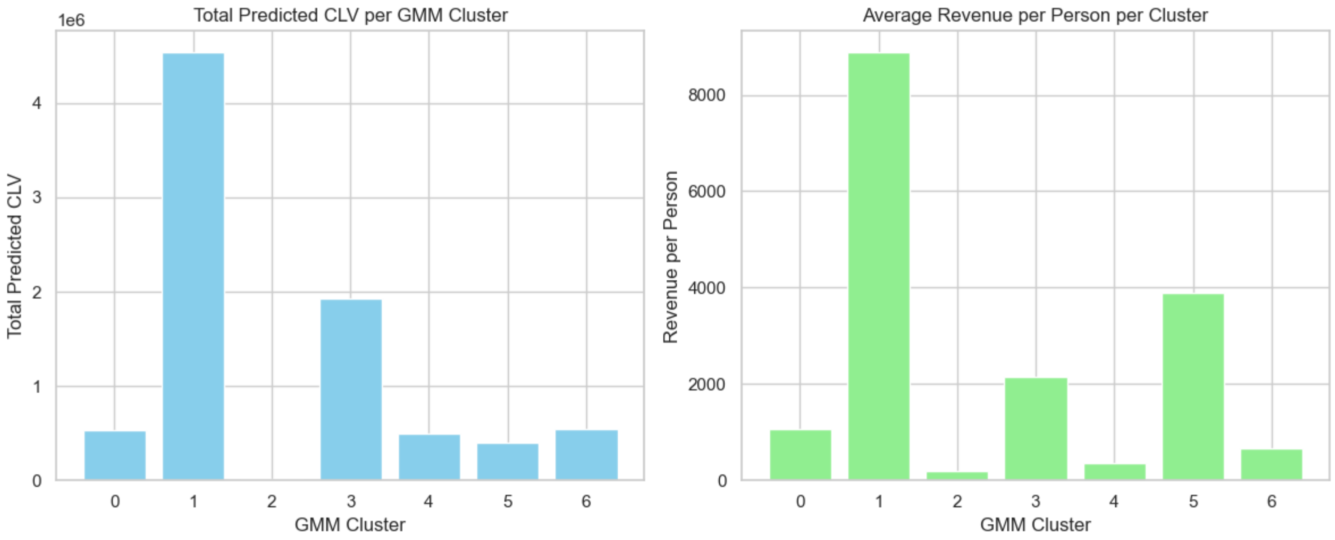

Figure 10: Visualization of predicted revenue and average CLV per cluster

7 Business Implications

This analysis provides a basis for classifying customers based on Recency, Frequency, Monetary value, AOV, and Tenure. Using this classification, the company can:

- Identify which types of goods are preferred by different customer segments and in what quantities.
- Optimize warehouse inventory through country-wise and segment-wise analysis.
- Detect customer dissatisfaction and take corrective measures using returns data and feedback.
- Target customers more effectively with personalized advertisements and discounts, ensuring potential buyers are not missed.
- Develop strategic marketing and retention plans based on predicted CLV to maximize long-term revenue.

This structured approach enables the company to focus on high-value customers, improve customer retention, and optimize resource allocation for maximum business impact.