**INSY5339 - Principles of Business Data Mining**

**Project Progress Report - Group 6**

**Project Title – Online Retail Customer Segmentation**

**Group Members**

|  |  |
| --- | --- |
| **Group Member** | **Student ID** |
| Ashwini Deshmukh | 1002125106 |
| Deepan Akash Don Bosco | 1002148522 |
| Nicholas Kingston | 1001523093 |
| Swathi Manjunatha | 1002162937 |

**Course Instructor: Jayarajan Samuel**

**Final Report Submission Date: 11/30/2024**

1. Introduction

* **Business Problem Statement:**

This project aims to segment customers based on purchasing patterns to enhance targeted promotions. By identifying customer groups, we can improve retention and boost sales.

* **Data Overview:**

Kaggle Link: <https://www.kaggle.com/datasets/yasserh/customer-segmentation-dataset/data>

The dataset used in this project contains 541,909 transactions from a UK-based online retail store. The data covers transactions over one year, with details such as invoice numbers, product code, product descriptions, quantities, Invoice dates unit prices, customer IDs and country. This data helps us analyze customer spending behavior and identify different customer segments.

A list of items on a white background

Description automatically generated

2. Data Preprocessing and Feature Engineering

* **Data Cleaning:**

To prepare the data, rows with missing *‘CustomerID’* values were removed to ensure each transaction could be linked to a customer. The *‘InvoiceDate’* column was converted to a datetime format for time-based analysis, and a new column, ‘*TotalPrice’*, was created by multiplying *‘Quantity’* by ‘*UnitPrice’* for each transaction.

* **Feature Engineering and Data Aggregation at the Customer Level:**

To gain insights into customer behavior, we aggregated the data at the **customer level**. This means that instead of analyzing each transaction individually, we summarized each customer’s activity over the entire dataset.

- **Recency**: Calculated as the number of days since the customer’s most recent purchase. This was done by finding the difference between a reference date (one day after the latest transaction date in the dataset) and each customer’s last purchase date.

- **Frequency**: Determined by counting the unique number of purchase instances (invoices) for each customer.

- **Monetary:** Calculated as the total spending for each customer by summing up the *‘TotalPrice’* (Quantity × UnitPrice) across all their transactions.

These aggregated RFM metrics (Recency, Frequency, and Monetary) form the foundation for customer segmentation in this project, enabling us to group customers based on their purchasing behavior.

* **Data Normalization:**

Before applying clustering algorithms, the RFM metrics were normalized using **StandardScaler** to ensure fair comparison. StandardScaler scales each feature to have a mean of 0 and a standard deviation of 1, so that larger ranges in one feature (such as Monetary) do not dominate the clustering process. This step allows each metric to contribute equally to the cluster formation.

* **Date Component Extraction:**

Additional features, including Month, Quarter, Day of the Week, and Hour, were extracted from *‘InvoiceDate’* to study trends in customer purchases over time.

3. Exploratory Data Analysis

* **Monthly and Quarterly Sales Trends:**

By plotting monthly and quarterly sales, we observed a significant increase in sales during October and November, suggesting a seasonal spike in customer purchases during the holiday period. This insight can help target seasonal promotions effectively.

A screenshot of a graph

Description automatically generated

* **Day-of-Week and Hourly Trends:**

Sales were analyzed by day of the week and hour of the day. The results showed that Thursday had the highest sales, and midday hours were peak times for purchases. This information is useful for timing promotions or planning staffing needs.

A graph of sales and sales

Description automatically generated with medium confidence

4. K-Means Clustering

* **Clustering Technique:**

We used **K-Means clustering** to group customers into different segments based on their RFM metrics. The **Elbow Method** helped determine the optimal number of clusters, which was found to be 4, we choose 3 for this project. These clusters represent different types of customers based on their spending and purchasing habits.

A graph with blue dots

Description automatically generated

**K-Means Clustering with 3 Clusters (before removal of outlier):**

* **Cluster Assignment:**

**A number of numbers and letters

Description automatically generated with medium confidence**

* **Centroids:** After clustering, we calculated the **centroids** for each cluster. These represent the average Recency, Frequency, and Monetary values for each group and help identify the primary characteristics of each segment.

A number of numbers and symbols

Description automatically generated with medium confidence

* **Cluster Visualization:** We used PCA to reduce the data to two dimensions for visualization. The 2D scatter plot showed three distinct clusters, confirming that K-Means separated customers effectively.

A diagram of a graph

Description automatically generated with medium confidence

* Identified initial rectangular edges in PCA plot, to address this, we removed outliers **(~692 customers)** using the Interquartile Range (IQR) method. The revised data was then re-clustered, resulting in smoother and more cohesive clusters.

A screenshot of a graph

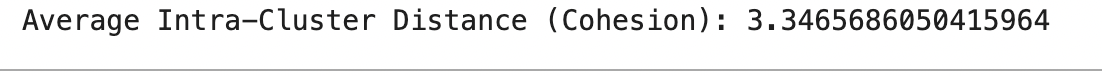
Description automatically generated

A diagram of a multicolored diagram

Description automatically generated

* **Cohesion, Separation, and Quality Ratio and Silhouette score:**

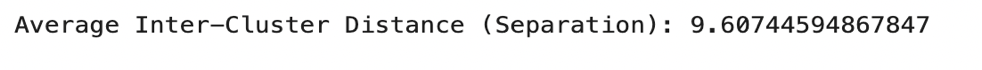
- **Cohesion:** The average intra-cluster distance (3.35) suggests that customers within each cluster have similar purchasing behaviors, as indicated by relatively tight cluster formations. Cluster 2 has the largest cohesion value, likely due to its high spread in monetary values.



A graph of a graph

Description automatically generated with medium confidence

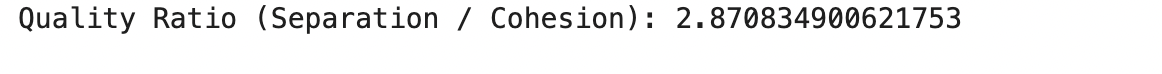
- **Separation**: The average inter-cluster distance (9.61) reflected how distinct each cluster was from others.



A graph with a red and orange bar

Description automatically generated with medium confidence

- **Quality Ratio:** A separation-to-cohesion ratio of 2.87 suggested that clusters were well-separated and compact, indicating good clustering quality.



- **Silhouette Score:** improved from **0.332** to **0.435** post-outlier removal, indicating better-defined clusters.



5. K- Means Cluster Analysis and Interpretation

* **Summary Statistics per Cluster:**

We calculated detailed statistics (mean, standard deviation, min, max, median) for each RFM metric within each cluster. This analysis helps us understand the key differences among the clusters.

A screenshot of a computer

Description automatically generated

**Detailed Cluster Profiles:**

**Cluster 0: Regular Customers**

* **Recency**: Moderate (average of 40.09 days), indicating that customers have engaged relatively recently but not immediately.
* **Frequency**: Moderate (average of 5.60 purchases), suggesting that these customers make purchases at a steady, average rate.
* **Monetary**: Moderate spending (average of 1,821.83), highlighting that the purchases made by these customers are of average value, though they may not make large purchases consistently.

**Cluster 0** represents regular customers who engage moderately with the business. This segment may benefit from consistent engagement through promotions, special offers, or reminders to encourage them to spend slightly more or purchase more frequently.

**Cluster 1: Inactive/Low-Value Customers**

* **Recency**: High (average of 246.37 days), indicating long periods of inactivity and low recent engagement.
* **Frequency**: Low (average of 1.85 purchases), showing that these customers make few purchases overall.
* **Monetary**: Low spending (average of 459.54), which suggests these customers have a low overall spending history.

**Cluster 1** represents inactive or low-value customers. To re-engage these customers, campaigns such as discount offers, reminders, or personalized communication may help increase their purchasing frequency and spending.

**Cluster 2: High-Value, Loyal Customers**

* **Recency**: Very low (average of 6.09 days), showing frequent and recent engagement with the brand.
* **Frequency**: Very high (average of 86.87 purchases), indicating that this group of customers is highly engaged and frequently makes purchases.
* **Monetary**: High spending (average of 81,835.86), indicating that these customers spend significantly on their purchases, making them highly valuable to the business.

**Cluster 2** represents loyal, high-value customers who demonstrate frequent and significant spending behavior. These customers are ideal for loyalty programs, exclusive offers, and personalized communication to further strengthen retention and maximize long-term value.

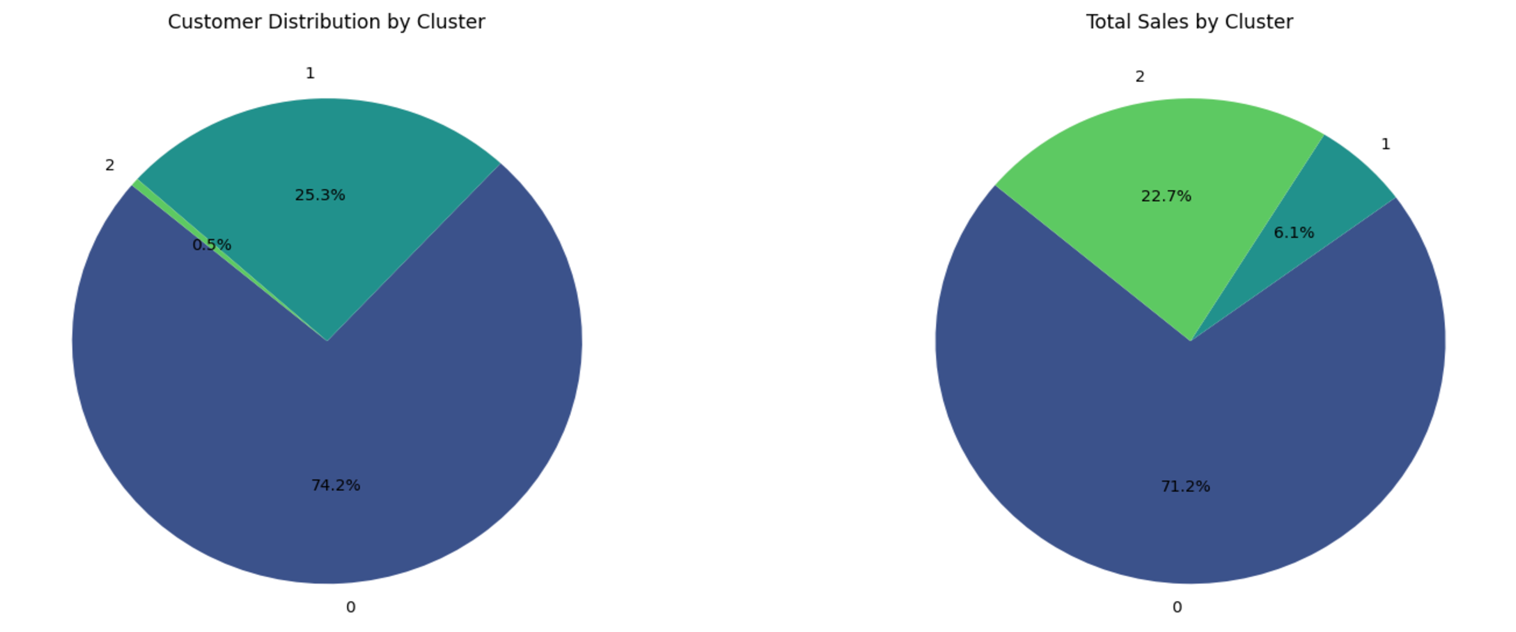
**Box Plot of Recency, Frequency and Monetary across all 3 clusters:**

**A graph with a blue and black line

Description automatically generated with medium confidence**

* **Cluster Distribution:**

These pie charts reveal that, with 74.2% of the customer base, **Cluster 0** contributes 71.2% of total sales, indicating it as the primary revenue driver. **Cluster 2**, despite comprising just 0.5% of customers, accounts for 22.7% of total sales, showcasing its high value. Meanwhile, **Cluster 1** represents 25.3% of the customer base but contributes only 6.1% of sales, suggesting that targeted engagement campaigns could boost their sales potential.



6. Hierarchical Clustering

We used **Agglomerative Clustering** with the Ward linkage method to segment customers based on their purchasing behavior. The Ward method minimizes variance within clusters, producing well-defined groups.

* **Cluster Determination**: A **dendrogram** was generated to visualize the clustering process. By placing a cut line at a height of 35, we identified **three main clusters**.

A diagram of a city

Description automatically generated

* **Cluster Assignment**:

We set the number of clusters to 3 based on the dendrogram analysis and applied **Agglomerative Clustering** to assign each customer to one of the three clusters.

Cluster labels were assigned as **0, 1, and 2** to ensure consistency with our K-Means clustering for comparison purposes.

A table of numbers and a few words

Description automatically generated with medium confidence

* **Hierarchical Clusters Characteristics in Boxplot:**

**A graph with a blue and black box

Description automatically generated**

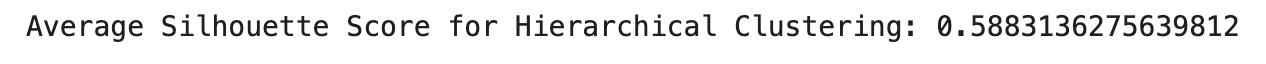
**Cluster 0**: High-value, high-frequency customers with low recency, indicating loyal customers.

**Cluster 1**: Moderate frequency and spending customers, with varying recency.

**Cluster 2**: Low frequency and monetary value customers, with high recency, indicating less-engaged customers.

* **Silhouette Score:**

The silhouette score for hierarchical clustering is 0.588, indicating that the clusters have a good level of separation. This means the clusters are distinct enough to provide useful insights.



* **Cluster Summary Statistics**:  
  The hierarchical clusters were analyzed using summary statistics for **Recency, Frequency, and Monetary (RFM)** metrics:

**Cluster 0**: Avg. Recency: 6.4, Avg. Frequency: 82.6, Avg. Monetary: 78,233.

**Cluster 1**: Avg. Recency: 48.5, Avg. Frequency: 5.3, Avg. Monetary: 1,710.

**Cluster 2**: Avg. Recency: 270.1, Avg. Frequency: 1.7, Avg. Monetary: 450.

A screenshot of a computer

Description automatically generated

These values highlight the distinct purchasing patterns across clusters, with Cluster 0 representing high-value customers and Cluster 2 capturing less-engaged ones.

7. Project Conclusion and Recommendations

This project effectively segmented customers using **EDA** , K-means, hirarchical clustering, providing clear strategies for targeted promotions, boost retention, and increase sales.

* **High variance in Recency, Frequency, and Monetary (RFM) metrics** among customers, indicating diverse purchasing patterns.
* **Seasonal Trends**: Sales peak in November and December, especially in Q4, suggesting high potential for targeted holiday promotions.
* **Timing Patterns**: Sales are highest on Fridays and Saturdays, with afternoon/evening spikes, indicating optimal times for weekend deals or flash sales.

**Cluster Insights:** The table below summarizes the key customer segments identified through both clustering methods and provides actionable recommendations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Segment** | **K-Means Cluster** | **Hierarchical Cluster** | **Insights** |
| **High-Value Frequent Buyers** | Cluster 2 | Cluster 0 | Loyal, high-spending customers: ideal for loyalty programs, and exclusive offers to retain engagement and enhance satisfaction. |
| **Moderate Buyers** | Cluster 0 | Cluster 1 | Regular buyers: Targeted promotions or personalized product recommendations could encourage these customers to increase purchase frequency. |
| **Inactive/Low-Value Buyers** | Cluster 1 | Cluster 2 | Disengaged customers: Re-engagement campaigns, such as win-back offers or reminders, can help reactivate this group. |

Model Selection:

* Although Hierarchical Clustering, while offering slightly better silhouette scores, is less suitable for large datasets and lacks adaptability for real-time updates.
* Therefore, **K-Means** was chosen for its scalability, computational efficiency, and clear centroid-based clusters align with business strategies, offering actionable insights.