**Customer Behavior Insights: Segmentation and Predictive Analytics for E-Commerce Optimization**

***By Jasmine Cooper***

**Overview**

This project focuses on developing a customer segmentation model using Hierarchical Clustering techniques to analyze purchasing patterns and gain insights into customer behavior. Additionally, it aims to enhance targeted marketing by forecasting the lifetime value of each customer segment. These insights will empower stakeholders to make data-driven decisions regarding product development, inventory management, and seasonal marketing strategies.

*Tools:* Python and Tableau

*Skills Highlighted:* Clustering, predictive modeling, data visualization, and business insights

**Problem Statement and Objectives**

*Statement:*

A jewelry store seeks actionable insights into customer purchasing behaviors to optimize marketing strategies and enhance inventory management. By segmenting customers and predicting customer lifetime value, the store can target its marketing efforts more effectively and improve operational efficiency.

*Objectives:*

* Segment customers based on purchasing patterns and behaviors
* Predict customer segment lifetime value (CLV)
* Provide actionable recommendations for marketing and inventory strategies

**Data Collection and Preparation**

*Data Sources:*

The dataset was obtained from Kaggle:

[Jewelry Store E-Commerce Purchase History](https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-jewelry-store)

*Data Cleaning and Preprocessing:*

1. File Handling:

* Renamed the CSV file to jewelry\_proj1.csv, downloaded it, and uploaded it to Google Drive.
* Used the gdown library to access the file in Jupyter Notebook and loaded it into a pandas DataFrame.

1. Initial Exploration:
   * Identified missing values using missingno and visualized missingness with a bar chart. Additionally, potential correlations between missing values were visualized with a heatmap (matplotlib and seaborn).
   * Renamed columns for consistency and readability.
2. Imputation of Missing Values:
   * Brand ID: Used KNN imputation based on product ID and nearby rows to fill missing values.
   * Main Metal: Imputed missing values (~6%) with the most frequent metal, “gold”.
   * Price: Filled missing values using the average price for each main metal.
   * Gender: Imputed missing values proportionally (99% female, 1% male) based on available data.
   * Gems: Imputed based on similarity of price bins using available price and gem information,
   * Main Color: Filled with the mode of existing values.
   * Category: Removed invalid numerical entries and filled missing values randomly since less than 10% of the data was missing.
   * Category ID and User ID were dropped due to irrelevance for analysis.
3. Final Data Checks:
   * Ensured no missing values remained.
   * Converted Order Datetime column to datetime format for preprocessing.
4. Feature Engineering:
   * Applied one-hot encoding (pandas.get\_dummies) to categorical variables to prepare for hierarchical clustering.
   * Standardized numerical features using Standard Scaler to ensure uniformity across scales.

Libraries Used: pandas, missingno, matplotlib, seaborn, sklearn (KNNImputer and StandardScaler), and numpy

**Methodology**

*Customer Segmentation (Hierarchical Clustering)*

Hierarchical clustering was selected to analyze customer behavior and identify distinct purchasing patterns. The goal was to segment customers in a way that supports targeted marketing strategies and inventory optimization, ensuring that the jewelry store stocks products aligned with the preferences of its most revenue-generating customers. Additionally, this segmentation can help manage inventory for less active customer segments.

Given the dataset size (~95,000 rows), running hierarchical clustering on the full dataset caused computational issues. To address this, I sampled 25% of the data, a substantial portion sufficient for analysis.

The Ward method was employed because it minimizes variance within clusters, ensuring that customers in each group are as similar as possible. Initially, clustering was conducted on a sample size for exploratory analysis.

To determine the optimal number of clusters, I applied the elbow method, which suggested 4 or 5 clusters as optimal. Also, silhouette scores were applied, and 4 clusters yielded the highest score.

Based on these analyses, hierarchical clustering was finalized with the Ward method and 4 clusters.

*Predictive Modeling*

Churn Prediction

Predicting customer churn was the first step in modeling customer lifetime value. Churn prediction was based on customer purchase frequency (intervals between purchases) and assumed each Order ID represented a unique customer, though some customers may have multiple IDs.

To classify customer churn, I tested Logistic Regression and Random Forest Classifier. Both models achieved similar cross-validation and AUC scores, but I selected the Random Forest Classifier for its ability to capture the non-linear relationship between purchase frequency and churn. The model predicted a 31% churn rate, a reasonable estimate given the sampled data.

However, this analysis had the following limitations:

* The 25% sampling may introduce bias, as the dataset could overrepresent certain behaviors.
* Jewelry purchases are often seasonal or event-driven, meaning churn might not indicate permanent disengagement but rather periodic inactivity.

Customer Lifetime Value Prediction

Using the churn predictions, I calculated Customer Lifetime Value (CLV) to identify which customer segments provide the highest revenue potential. Predicting CLV enables the jewelry store to prioritize high-value customer segments for inventory optimization and marketing efforts.

Further improvement in predictive performance could be achieved by analyzing the full dataset, enabling a more comprehensive understanding of churn and its impact on CLV.

**Analysis and Findings**

The analysis revealed distinct customer profiles across four clusters, each with unique purchasing behaviors and value contributions. Cluster 4 holds the highest Customer Lifetime Value (CLV) at $15.31M, largely due to its significantly larger customer base of 13,643 individuals. In comparison, Cluster 1, the smallest group with 1,338 customers, has the lowest CLV of $1.58M. Clusters 2 and 3, with CLVs of $6.8M and $4.22M respectively, are closer in size and proportion, making their purchasing behaviors more comparable.

Across all clusters, the average purchase price is $361, with most transactions occurring around 11 am on Wednesdays. Diamonds and gold jewelry are the most popular materials, with earrings and rings leading as the most sought-after items. Seasonality plays a significant role in purchasing behavior, with peak sales observed in November, likely driven by holiday preparations and gift purchases, while December experiences cluster-specific spikes, particularly in Cluster 2. August, also sees a surge in sales, aligning with back-to-school shopping, summer vacations, and weddings.

Cluster-specific insights highlight variations in preferences and behaviors. Cluster 1 customers, who favor necklaces and brooches, tend to make purchases early in the week, with peak activity on Mondays between 9 am and 11 am. This group also shows a notable spike in November. Cluster 2 demonstrates the highest average sales price at $392, even surpassing Cluster 4’s average of $347, indicating a greater willingness to spend on jewelry items. This cluster peaks in November and December, with consistent midweek purchases, particularly on Wednesday s between 10 am and 12 pm. Similarly, Cluster 3 mirrors some of Cluster 2’s behaviors but peaks in February, likely due to Valentine’s Day, with customers purchasing earrings, rings, diamonds, and gold during late morning hours on Wednesdays. Cluster 4, despite its lower average sales price, represents the bulk of the customer base and exhibits peak sales in November, focusing on similar popular items like diamonds, gold, earrings, and rings.

Overall, the data reveals that most purchases occur midweek, around lunchtime, with consistent demand for high-value items like diamonds and gold across all clusters. Seasonal trends significantly influence sales, with notable peaks in February, August, and November. Additionally, Cluster 2’s higher average sales price and December peak suggest that customers in this group are more inclined to make premium purchases, while Cluster 4 represents frequent but lower-cost transactions.

These insights inform strategies for customer engagement and potential purchases. For example, during peak periods like February, August, November, and December can maximize revenue, while focusing on smaller clusters such as Cluster 2 could help retain high-spending customers. Additional churn prediction models may also help prioritize clusters with fewer customers to ensure sustained engagement and profitability.

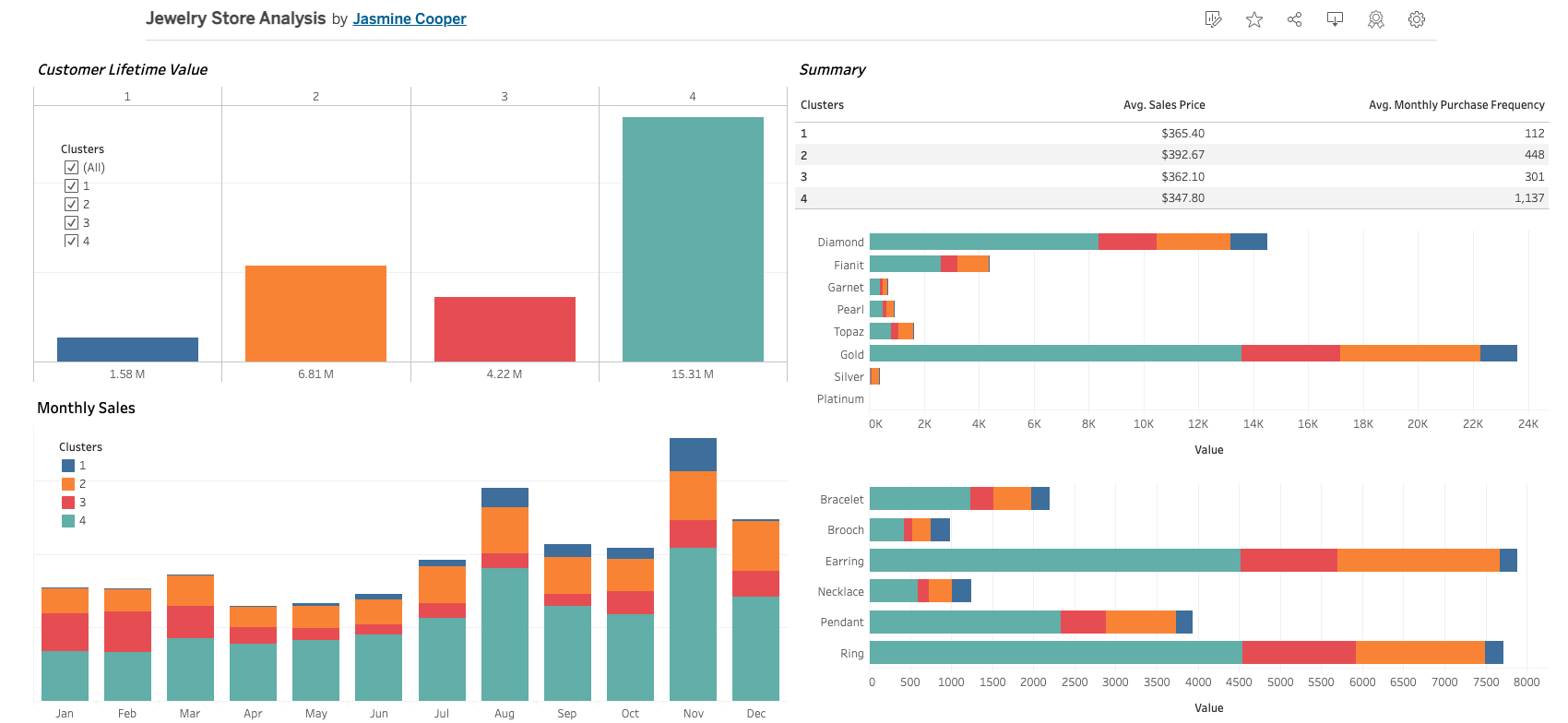
**Dashboard Design and Visualization**

The dashboard is designed to provide jewelry stakeholders with a comprehensive snapshot of customer cluster information, enabling them to analyze purchasing behaviors, identify trends, and develop targeted seasonal marketing strategies. It offers a structured view of key insights, including customer distributions, purchasing patterns, and seasonal sales trends, segmented by cluster to highlight differences. This segmentation allows stakeholders to conduct detailed analyses of Customer Lifetime Value (CLV) and seasonal demand.

Interactive features enhance usability, allowing stakeholders to explore specific clusters or compare clusters side by side. Filters and interactive elements help users uncover patterns, such as peak purchasing times, top-selling items, and seasonal sales spikes. For example, stakeholders can identify that Cluster 4 dominates in terms of customer count and CLV, suggesting that store layouts, inventory management, and marketing efforts could primarily target Cluster 4, with smaller-scale strategies tailored to Clusters 1-3.

The dashboard also supports inventory optimization by identifying trends in sales volumes. For instance, while the store offers 30 different types of gemstones, only 5 exceed 500 sales over a 12-month period. This insight allows stakeholders to prioritize stocking higher quantities of popular gemstones and reduce inventory for lower-demand items, optimizing revenue and reducing overstock.

This dynamic and user-friendly design ensures stakeholders can make informed decisions to enhance marketing efforts, streamline inventory, and maximize revenue.



**Conclusion and Recommendations**

Conclusion

This project provided valuable insights into customer purchasing behaviors at the jewelry store, achieving objectives of segmenting customers and predicting Customer Lifetime Value (CLV). Hierarchical clustering identified four distinct customer groups with unique purchasing patterns and revenue contributions. Cluster-specific analyses revealed variations in jewelry preferences, peak purchasing times, and seasonal trends. The predictive CLV model highlighted the revenue potential of each cluster, guiding the store in prioritizing marketing and inventory strategies. These insights empower stakeholders to optimize inventory, enhance targeted marketing, and improve customer retention, driving overall business efficiency and profitability.

Recommendations

1. Tailored Marketing Campaigns
   * Cluster 1: Launch targeted campaigns for necklaces and brooches during November to capture this group’s seasonal preferences.
   * Cluster 2: Focus on high-value customers by promoting premium products during November and December.
   * Cluster 3: Align campaigns with Valentine’s Day to maximize revenue in February.
   * Cluster 4: Implement broad marketing strategies to engage the largest customer segment, emphasizing discounts on popular items like gold diamond earrings and rings.
2. Inventory Adjustments:
   * Prioritize stocking high-demand items such as gold and diamond jewelry, particularly earrings and rings.
   * Optimize inventory by changing low-demand gemstones to custom orders only and reallocating resources to popular gemstones with high sales volumes.
   * Prepare for seasonal demand surges in February, August, November, and December by adjusting stock levels and ERP forecasts proactively.
3. Retention Strategies:
   * Encourage repeat purchases by offering exclusive early access to new collections and discounts for frequent buyers.
4. Operational Enhancements:
   * Use the dashboard’s interactive features to continuously monitor sales trends and adjust strategies dynamically.
   * Explore cross-selling opportunities by analyzing purchasing patterns within clusters. For example, bundling popular items like earrings and rings or offering buy one, get one (BOGO) deals.

**Limitations and Future Work**

Limitations

* Data Constraints: The analysis relied on 25% sample of the dataset, which may not fully represent the diversity of customer behaviors.
* Seasonal and Event-Driven Purchases: Jewelry purchasing patterns are often influenced by specific events like weddings and holidays, which may not indicate long-term customer preferences.
* Model assumptions: Predictive models assumed that purchase frequency is a reliable indicator of churn, which might not account for all or the main reason for infrequent purchases and churn.
* Cluster Size Variability: Smaller clusters may have introduced bias, particularly in calculating CLV and designing targeted strategies.

Future Enhancements

1. Incorporate Additional Data Sources:
   * Enrich analysis with demographic data, online browsing behavior, and social media interactions to enhance customer segmentation.
   * Include competitor data to benchmark performance and identify market opportunities.
2. Improve Predictive Modeling:
   * Develop and refine demand forecasts using time-series analysis such as the ARIMA model in churn and CLV predictions.
3. Real-Time Dashboard Updates:
   * Integrate live sales data to provide stakeholders with real-time insights and actionable metrics for immediate decision-making.
4. Personalized Cluster Recommendations:
   * Develop a recommendation engine to offer product suggestions tailored to each cluster’s preferences, enhancing the shopping experience, and driving sales.

By addressing these limitations and implementing future enhancements, the jewelry store can achieve a deeper understanding of customer behaviors and further optimize its marketing and inventory strategies.

Overall, this analysis successfully uncovered actionable insights into customer purchasing behaviors, enabling segmentation and the prediction of customer lifetime value to inform strategic decision-making. These findings empower the jewelry store to optimize inventory, enhance targeted marketing efforts, and address customer retention, ultimately driving operational efficiency and maximizing revenue potential.