MERCER
UNIVERSITY

SCHOOL OF BUSINESS

Targeted Marketing: Customer Segmentation & Classification Strategies

Authors: Aakanksha Singh, Joyce Talluri, Srikar Nadimpalli

INTRODUCTION

In today's fast-paced and digitally driven e-commerce landscape, understanding customer behavior has emerged as a cornerstone for businesses seeking to thrive in a highly competitive environment. As online platforms become the primary touchpoints for consumer interaction, the complexity of customer behavior has increased manifold. Customers now engage with brands through various channels such as websites, mobile apps, social media, and personalized email campaigns, making it imperative for businesses to adopt innovative approaches to analyze and respond to these interactions. Traditional methods of customer segmentation and relationship management, while foundational, are no longer sufficient to address the intricate and dynamic nature of consumer behavior in modern marketplaces.

In this project, we address the need to evolve customer analytics by integrating classical methods with cutting-edge machine learning techniques. We seek to bridge the gap between traditional customer analytics frameworks and the transformative capabilities of advanced methodologies, enabling businesses to derive actionable insights and remain competitive. By leveraging a robust dataset sourced from Kaggle, this study encompasses detailed information on customer demographics, purchasing behavior, and engagement metrics, including transaction attributes such as purchase quantities, monetary values, discount usage, and marketing expenses. This comprehensive dataset serves as the foundation for combining traditional analysis methods like

Recency, Frequency, and Monetary (RFM) analysis with advanced machine learning techniques such as clustering and classification models.

The integration of these methodologies provides a powerful toolkit for uncovering nuanced behavioral patterns and refining customer segmentation strategies. Businesses can benefit from precise targeting, enhanced retention rates, and a deeper understanding of customer lifetime value (CLV). The ability to segment customers more effectively not only drives better engagement but also supports more informed decision-making in resource allocation and marketing strategies.

E-commerce, defined as the sale and purchase of goods and services through the internet, is at the forefront of reshaping traditional marketing strategies. With its vast capabilities for product information dissemination and data-driven decision-making, e-commerce enables businesses to align their strategies with consumer expectations. However, the growing volume of data generated through online transactions, including browsing histories, purchase patterns, and customer feedback, demands sophisticated tools for analysis. Modern marketing strategies must process this wealth of information to uncover insights into consumer needs, preferences, and behaviors.

The motivation for this project lies in addressing the inherent limitations of traditional customer analytics approaches by utilizing the advanced capabilities of machine learning. Techniques such as k-means clustering enable the creation of dynamic and actionable customer segments, while predictive models like Random Forest and XGBoost help businesses forecast customer behavior and identify key drivers of engagement and profitability. These advanced techniques empower businesses to adapt to the evolving demands of the e-commerce ecosystem while maintaining a customer-centric approach.

By leveraging a blend of classical and modern analytical methods, the aim of our project is to provide practical solutions for businesses and also contribute to the academic discourse on customer behavior analysis. It presents a scalable and adaptable framework for understanding consumer behavior, offering insights into how businesses can navigate the challenges and opportunities of a rapidly transforming digital marketplace. Through the effective application of data-driven insights, businesses can enhance their marketing strategies, deliver exceptional customer experiences, and achieve sustained growth in an ever-competitive e-commerce environment.

LITERATURE REVIEW

The competitive landscape today, makes it important to understand and enhance customer relationships that is critical for business success. Customer Lifetime Value (CLV) has emerged as a key metric for assessing the financial contribution of customers over time, helping businesses prioritize resources effectively. By combining CLV with Recency, Frequency, and Monetary (RFM) analysis, organizations can develop targeted marketing strategies and refine customer segmentation. This report explores advancements in CLV and RFM methodologies, emphasizing their practical applications, ongoing challenges, and future opportunities.

Evolution of CLV and RFM Frameworks

The traditional RFM model has long been a cornerstone of customer segmentation, analyzing when customers last interacted with a business (Recency), how often they engage (Frequency), and their spending patterns (Monetary). Despite its enduring relevance, researchers have worked to improve the model to account for the complexities of modern customer behavior.

Khajvand et al. enhanced the classic RFM approach by introducing additional dimensions like "Count Item" and applying weighted scoring methods. Their findings affirmed the robustness of traditional RFM metrics while highlighting the value of tailored enhancements for specific industries.

Yoseph and AlMalaily further developed RFM analysis by integrating it with advanced clustering algorithms such as fuzzy C-means and expectation-maximization. These methods allowed for dynamic customer segmentation, accommodating shifts in behavior over time and improving the identification of high-value customer groups. The inclusion of statistical outlier detection further bolstered the model's reliability.

Sun et al. (2023) incorporated machine learning techniques like random forests and support vector machines into CLV analysis, addressing the unique challenges of non-contractual customer relationships. These methods provided deeper insights into customer behavior and improved the predictive accuracy of CLV models, especially for e-commerce platforms.

Kim et al., (2006) introduced a comprehensive framework for the telecommunications industry that combined CLV with metrics for customer loyalty and potential value. Their work emphasized the importance of aligning marketing strategies with profitability and retention metrics to drive sustainable growth.

Methodological Innovations

1. Enhanced RFM Models: Modern RFM models now incorporate additional metrics, such as transaction diversity and behavioral trends, providing a more nuanced understanding of customer dynamics and enabling precise segmentation.

2. Machine Learning Integration: Advanced algorithms, including gradient boosting and neural networks, have revolutionized CLV analysis by identifying complex patterns within large datasets. These tools provide more accurate and actionable insights than traditional models.

3. Hybrid Clustering Techniques: Combining traditional clustering methods with advanced approaches like fuzzy C-means has improved segmentation accuracy. These methods are particularly effective in identifying subtle behavioral differences among diverse customer populations.

4. CRM Applications: Integrating CLV models into customer relationship management (CRM) systems has enabled organizations to adjust marketing strategies in real-time, enhancing customer retention and satisfaction.

Practical Applications

- Retail: Retailers have employed enhanced RFM and clustering models to identify high-value customers, allowing for more targeted and effective marketing campaigns.

- Telecommunications: By incorporating CLV, customer loyalty, and potential value, telecommunications companies have successfully tailored strategies for retaining and upselling to their most profitable customers.

- E-commerce: Machine learning-powered CLV models have proven particularly effective for e-commerce, helping businesses predict purchase behaviors and design better retention strategies.

DATA

Data Collection:

The Marketing Insights for E-Commerce Company dataset from Kaggle was utilized for this research. This dataset contained detailed customer data, including demographic information, purchasing behaviour, and engagement metrics, providing insights into customer segmentation and marketing strategies. It served as a foundation for exploring relationships between customer attributes and e-commerce performance, enabling the identification of key drivers of customer acquisition, retention, and revenue growth. Furthermore, the dataset allowed for predictive modelling and optimization of marketing campaigns to enhance overall e-commerce efficiency. Kaggle: Marketing Insights for E-Commerce Company Dataset

Data Dictionary:

Online_Sales.csv	This file contains actual orders data (point of Sales data) at transaction level with below variables
CustomerID	Customer unique ID
Transaction_ID	Transaction Unique ID
Transaction_Date	Date of Transaction
Product_SKU	SKU ID – Unique Id for product
Product_Description	Product Description
Product_Cateogry	Product Category
Quantity	Number of items ordered
Avg_Price	Price per one quantity
Delivery_Charges	Charges for delivery
Coupon_Status	Any discount coupon applied
Customers_Data.csv	This file contains customer's demographics.
CustomerID	Customer Unique ID
Gender	Gender of customer
Location	Location of Customer
Tenure_Months	Tenure in Months
Discount_Coupon.csv	Discount coupons have been given for different categories in different months

Month	Discount coupon applied in that month
Product_Category	Product category
Coupon_Code	Coupon Code for given Category and given month
Discount_pct	Discount Percentage for given coupon
Marketing_Spend.csv	Marketing spend on both offline & online channels on day wise.
Date	Date
Offline_Spend	Marketing spend on offline channels like TV, Radio, NewsPapers, Hordings etc.
Online_Spend	Marketing spend on online channels like Google keywords, facebook etc.
Tax_Amount.csv	GST Details for given category
Product_Category	Product Category
GST	Percentage of GST

Data Cleaning:

For this research, which focused on analysing marketing insights for e-commerce companies and enhancing customer engagement strategies, we took deliberate steps to ensure the integrity and clarity of the data.

Missing data was handled with care to maintain the accuracy and reliability of the analysis. By addressing issues such as missing values (NAs) and undefined entries (NaNs) through targeted strategies, we ensured the dataset was complete and ready for meaningful exploration.

These data preparation efforts laid a strong foundation for the research, enabling us to derive actionable insights and support our overarching goal of refining marketing strategies to drive better customer acquisition, retention, and overall business growth in the e-commerce sector.

Calculated Metrics:

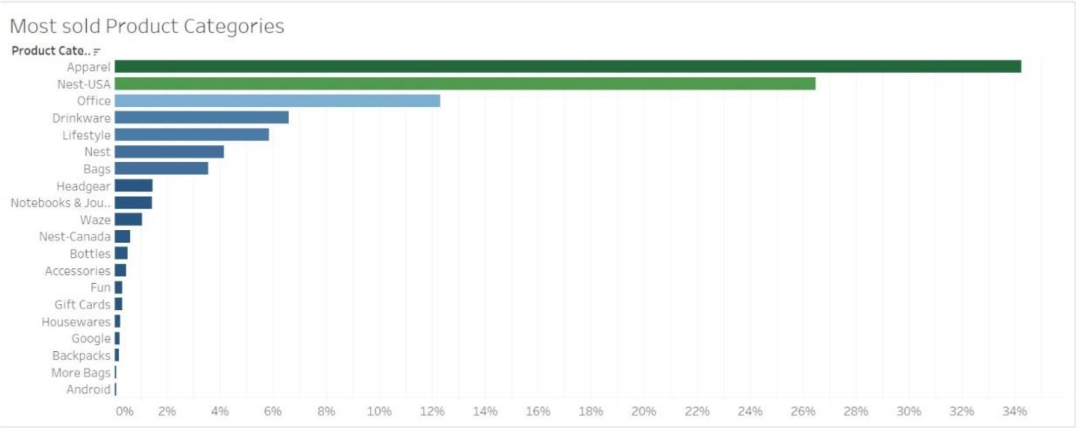
Metric	**Formula**	**Description**
Discounted Price	Discounted Price = Average Price – (Average Price * Discount Rate)	Price after applying any applicable discount rate.

GST Price	GST Price = Discounted Price * GST	Calculates tax on the discounted price using the GST rate.
Price	Price = (Discounted Price * Coupon Used) + GST Price	Final price after accounting for discounts, coupons, and GST.
Total Purchase Amount	Total Purchase Amount = Quantity * Price + Delivery Charges	Total spending for a transaction, factoring in quantity and delivery charges.
Average Purchase Value	Average Purchase Value = Total Purchase Amount for Customer / Number of Transactions	Average spending per purchase by a customer.
Purchase Frequency	Purchase Frequency = Tenure (in Months) / Number of Transactions	Measures how often a customer makes purchases over their tenure.
Customer Value per Month	Customer Value per Period (monthly) = Total Purchase Amount / Tenure (in months)	Average monetary value generated per month by each customer
Customer Lifetime Value (CLV)	CLV = Average Purchase Value * Purchase Frequency * Average Customer Lifespan	Estimated total value a customer generates during their relationship with the business.
Customer Value	Customer Value = Average Purchase Value * Average Frequency Rate	Total monetary value a customer brings based on their spending and purchasing frequency.
Average Purchase Value (APV)	APV = Total Revenue over a Time Frame / Total Number of Purchases	Represents the average amount of money spent per transaction over a specific period.

Average Frequency Rate	Frequency Rate = Total Number of Purchases over a Period / Total Number of Customers	Measures how frequently an average customer makes purchases within a specific period.
Average Customer Lifespan	Avg Customer Lifespan = Average Number of Days Customers Stay Active / Total Number of Customers	Typical duration a customer continues purchasing before churning.

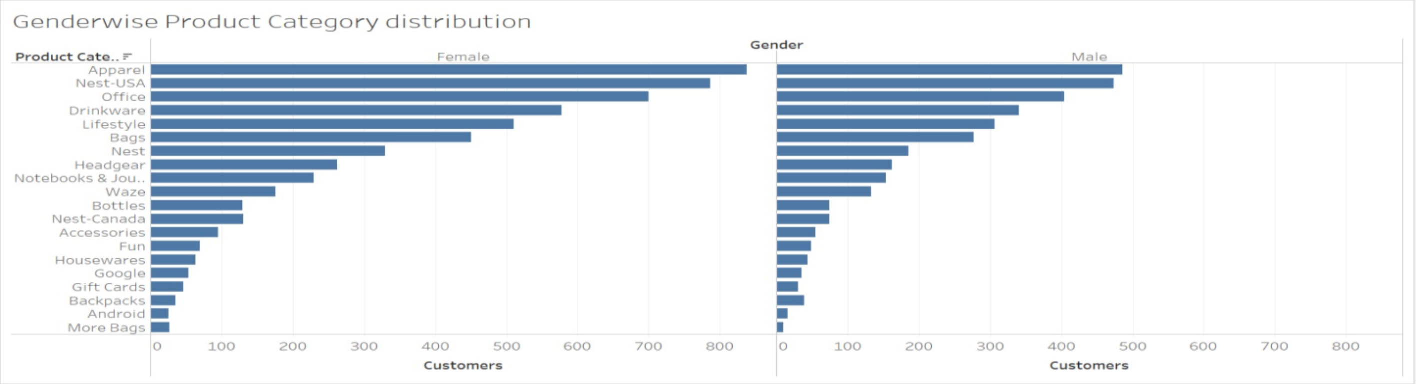
EXPLORATORY DATA ANALYSIS:

Most Sold Product Categories



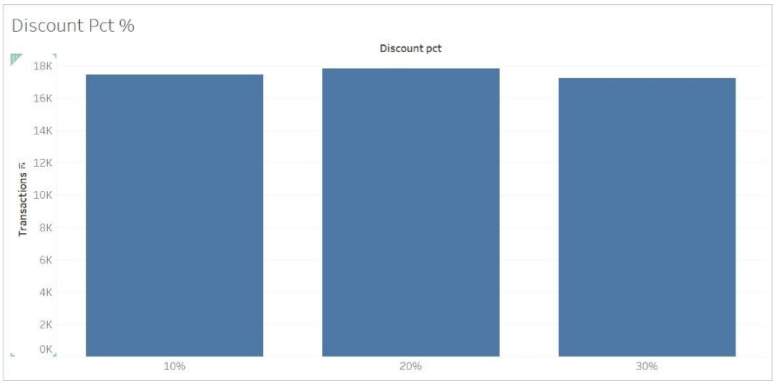
This bar chart highlights the most popular product categories based on sales volume. Apparel is the top-selling category, accounting for nearly 35% of total sales, followed by Nest-USA and Office products. Categories such as Drinkware, Lifestyle, and Bags are moderately popular, while less frequently purchased items include Notebooks & Journals, Bottles, and Accessories. This chart can help prioritize inventory stocking and promotional efforts on high-demand products.

Gender wise Product Category Distribution



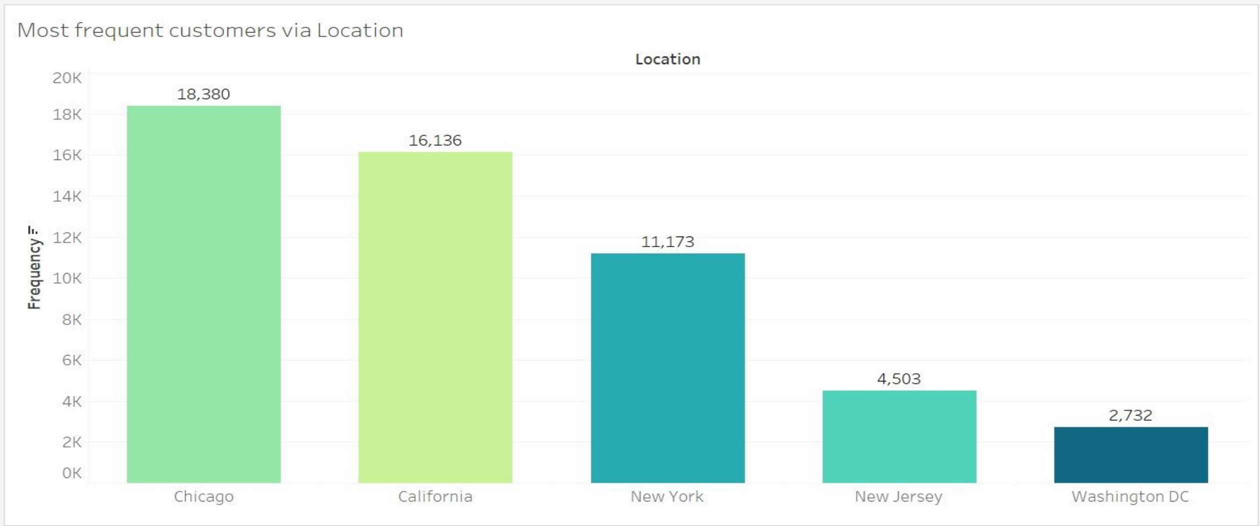
This graph compares product preferences between male and female customers. Both genders exhibit similar preferences for the top categories, such as Apparel, Nest-USA, and Office, although women tend to buy more Apparel and Lifestyle items, while men purchase slightly more Office products. Additionally, categories like Drinkware and Bags have balanced gender appeal. This data can guide targeted marketing campaigns, ensuring product recommendations align with gender-specific buying patterns.

Discount Percentage and Transactions



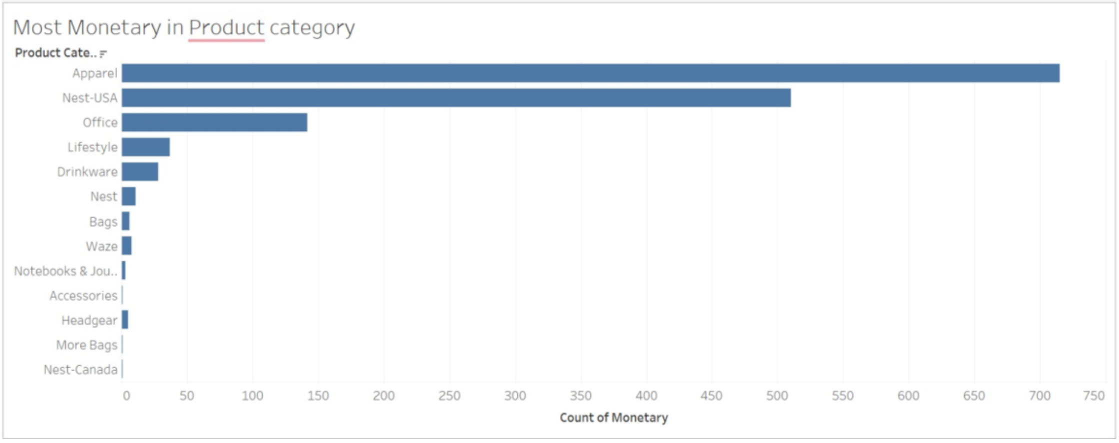
This chart shows the distribution of transactions across different discount percentages (10%, 20%, and 30%). The uniformity across the three discount levels suggests that offering discounts in this range has a consistent impact on driving transactions. Businesses can use this insight to maintain or optimize their discount strategies, ensuring competitiveness without overly reducing margins.

Most Frequent Customers by Location



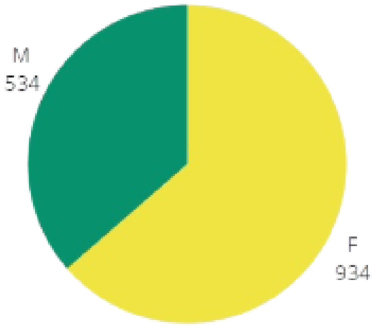
This bar chart illustrates the frequency of customers across different locations. Chicago has the highest number of frequent customers, followed by California and New York. New Jersey and Washington D.C. have comparatively fewer frequent customers. This information highlights key geographic markets, with Chicago and California being top priorities for targeted campaigns and inventory planning.

Most Monetary Contribution by Product Category



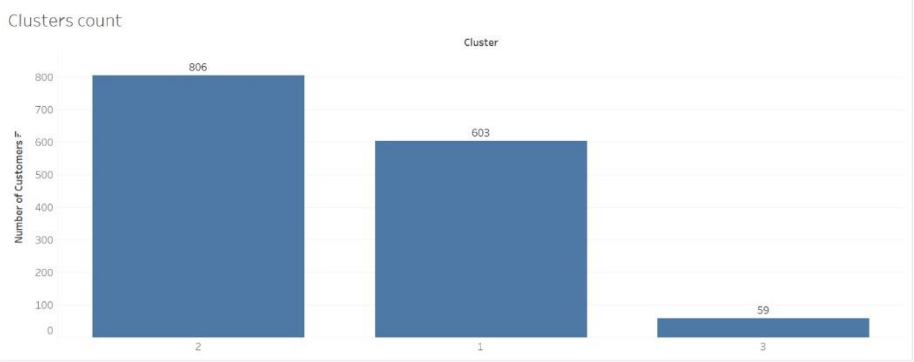
This chart displays the product categories generating the most revenue. Apparel and Nest-USA lead in monetary contribution, followed by Office products. Categories like Lifestyle and Drinkware contribute moderately, while others like Notebooks & Journals and Accessories show minimal revenue impact. This suggests a focus on top performing categories to maximize revenue.

Gender Distribution of Customers



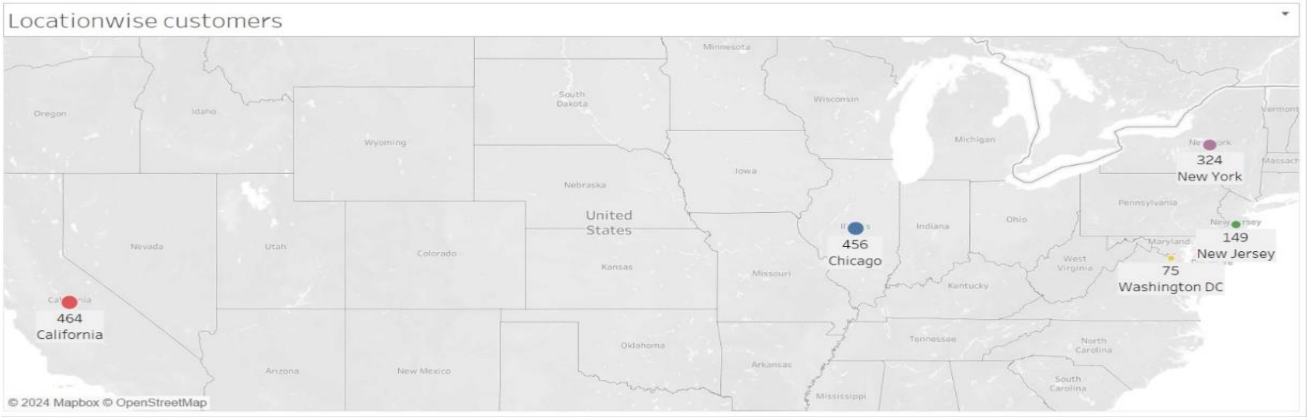
The pie chart shows the gender distribution of customers, with **females** (934) making up a larger portion of the customer base compared to **males** (534). This indicates an opportunity to further engage female customers through targeted marketing while identifying strategies to increase male customer participation.

Cluster Count of Customers



The bar chart shows the distribution of customers into three clusters. Cluster 2 has the largest group with 806 customers, followed by Cluster 1 with 603 customers, and Cluster 3 being the smallest with 59 customers. This segmentation can guide tailored strategies for different clusters to optimize engagement and sales.

Location-Wise Customer Distribution Map



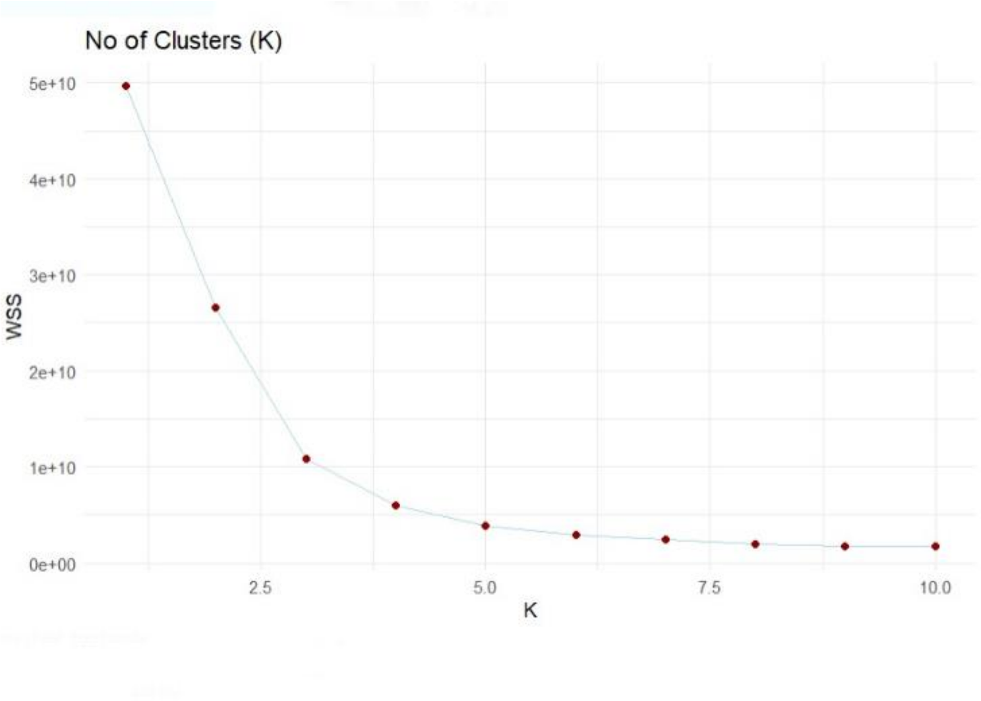
This geographic map visualizes customer distribution by location. **California** has the highest concentration of customers (464), followed by **Chicago** (456) and **New York** (324). Other locations, such as **New Jersey** (149) and **Washington D.C.** (75), have smaller customer bases. This map emphasizes the importance of focusing resources on top locations while exploring growth opportunities in smaller regions.

METHODOLOGIES

Obtaining the dataset from Kaggle, we compiled and transferred the data into Excel, and then imported the dataset into RStudio for further analysis and refinement. Our team implemented various testing methods, including Multinomial Logistic Regression, Classification and Regression Trees, Random Forest, and XGBoost models in order to compare results across models and determine the rationale behind our testing outcomes. This multi-stage approach enabled a thorough evaluation of model performance under different parameter configurations.

Clustering

Elbow Method:



The chart illustrates the Elbow Method used to determine the optimal number of clusters (kkk) for k-means clustering. The Within-Cluster Sum of Squares (WSS), representing the variance within each cluster, decreases sharply as the number of clusters increases from k=1k = 1k=1 to k=3k = 3k=3, indicating significant improvements in cluster cohesion. Beyond k=3k = 3k=3 or k=4k = 4k=4, the rate of decrease in WSS slows considerably, forming an "elbow" point. This suggests that dividing the data into 3 or 4 clusters provides the best balance between minimizing variance and maintaining simplicity. Choosing k=3k = 3k=3 allows for clear segmentation of customer groups, while k=4k = 4k=4 offers more granular insights. This analysis highlights the optimal cluster count for effective data-driven decision-making.

k-means Clustering:

The k-means clustering analysis grouped customers into three distinct clusters based on their Recency, Frequency, and Monetary (RFM) and CLV values. Cluster 1 includes high-value, loyal customers who purchase frequently and spend significantly, like Customer 12347, with 60 transactions worth 15,686.84 and a Recency of just 7 days. Cluster 2 represents low-value or inactive customers with fewer transactions and longer gaps, such as Customer 12346, with 2 transactions worth 174.98 and a Recency of 16 days. Cluster 3 comprises occasional customers with moderate activity. These insights help businesses reward loyal customers, re-engage inactive ones, and nurture occasional buyers, driving stronger relationships and growth.

Profiling

Cluster 1:

- Characteristics: Longest customer relationship length and highest AOV.
- Engagement: Most frequent and highly recent purchases.
- Value: Contributes the highest monetary value, AOV, and CLV.
- Demographics: Located in Chicago, predominantly female.
- Focus: Retain and reward these high-value customers with exclusive programs and premium experiences to maintain loyalty.

Cluster 2:

- Characteristics: Short customer relationship length and lowest average order value (AOV).
- Engagement: Least recency and infrequent purchase behaviour.
- Value: Minimal monetary contribution and lowest customer lifetime value (CLV).
- Demographics: Located in California, predominantly female.
- Focus: Implement re-engagement strategies and targeted promotions to boost activity and

Cluster 3:

- Characteristics: Moderate customer relationship length and average AOV.
- Engagement: Somewhat recent and moderate purchase frequency.
- Value: Contributes moderately to overall monetary value and CLV.
- Demographics: Located in Chicago, predominantly female.
- Focus: Strengthen loyalty through personalized offers and nurture them toward higher engagement levels.

Classification

Model Building:

The dataset provides a detailed classification of customers based on transaction history, demographics, and derived metrics such as Recency, Frequency, Monetary value, and Customer Lifetime Value (CLV), along with cluster assignments from k-means clustering. Transactional attributes like the number of transactions, total purchase amount, and average purchase value highlight spending behaviors, while demographics such as Gender and Location provide context for targeted marketing. Derived metrics like purchase frequency, repeat rate, and churn rate measure customer engagement and loyalty. For example, high-value customers, such as Customer 1 from Cluster 1, exhibit high Recency (7 days), frequent purchases (60 transactions), and significant CLV (34.39), indicating strong engagement. In contrast, occasional buyers, like Customer 3 from Cluster 1, show low Frequency (3 transactions) and minimal CLV (0.83), suggesting the need for reactivation efforts. Moderate customers, such as Customer 5 from Cluster 1, with a Frequency of 77 and a CLV of 18.26, demonstrate steady engagement with potential for upselling. Cluster analysis reveals distinct behavioral patterns, such as high-value customers favoring categories like Nest-USA and Apparel, which inform targeted marketing strategies. Businesses can leverage these insights to reward loyal customers, re-engage inactive ones, and maximize lifetime value through personalized campaigns, ultimately driving growth and retention.

Multinominal Regression:

The multinomial logistic regression model predicts cluster membership (Cluster 2 or Cluster 3) based on Quantity, Product Category, Gender, Location, and Tenure Months. For Cluster 2, higher quantities slightly decrease the likelihood of belonging to the cluster, and categories such as Apparel, Bags, and Drinkware significantly reduce the log-odds of membership. Gender (male) and location have minimal impact on Cluster 2. For Cluster 3, higher quantities, purchasing Apparel or Nest-USA products, and being male strongly increase the likelihood of membership, while locations such as New York and Washington, DC, significantly reduce the likelihood of being in this cluster. Tenure Months has a negligible effect on Cluster 2 but negatively influences Cluster 3 membership. The model demonstrates that Cluster 3 represents customers with high engagement and product-specific preferences, particularly for Apparel and Nest-USA, while

Cluster 2 customers exhibit less distinct patterns. These insights can guide targeted marketing strategies, such as focusing on high-value products for Cluster 3 and re-engagement efforts for Cluster 2. The model achieves a Residual Deviance of 1988.042 and an AIC of 2068.042, indicating a reasonable fit.

The model's performance across the three clusters demonstrates varying levels of accuracy. For **Cluster 1**, 84 out of 149 instances were correctly predicted, indicating moderate performance. Similarly, for **Cluster 2**, the model accurately classified 96 out of 161 instances, reflecting a comparable level of effectiveness. However, for **Cluster 3**, the model achieved perfect classification, correctly predicting all 161 instances, showcasing exceptional performance for this cluster.

The class-specific statistics highlight the model's varying performance across clusters. Sensitivity, or the true positive rate, indicates that 56.3% of Cluster 1 instances and 57.1% of Cluster 2 instances were correctly classified, while Cluster 3 achieved an impressive 96.9% sensitivity, demonstrating strong performance for this cluster. Specificity, or the true negative rate, shows that the model avoids incorrectly predicting Cluster 1 with 76.9% specificity and Cluster 2 with 79.4% specificity, while Cluster 3 achieves a perfect 100% specificity, effectively identifying all non-Cluster 3 instances. The balanced accuracy, averaging sensitivity and specificity, further underscores these findings, with Cluster 1 at 66.6%, Cluster 2 at 68.3%, and Cluster 3 excelling at 98.4%, making it the most reliably predicted cluster.

To improve the identification of Clusters 1 and 2, efforts should focus on exploring additional features or fine-tuning the model parameters, as well as addressing potential class imbalances that may be contributing to the performance disparity between clusters. For Cluster 3, the strong predictive performance should be leveraged in marketing and engagement efforts, as the predictions for this cluster are highly reliable. Overall, the model shows solid performance, particularly for Cluster 3, while presenting opportunities for refinement in enhancing the accuracy for Clusters 1 and 2.

Classification Trees:

The CART model's decision tree is structured around two key splits based on the Quantity variable. The first split divides observations into class 1 (if Quantity >= 776) and other classes (if Quantity

< 776). The second split further categorizes the data into class 2 (if Quantity < 26) and class 3 (if Quantity >= 26). This simple yet interpretable structure highlights the relationship between Quantity and the target classes, enabling clear decision-making.

The model achieved an overall accuracy of 70.83%, with a kappa statistic of 0.5625, indicating moderate agreement between predictions and true classifications. Class 1 was classified with high sensitivity (95.2%), specificity (100%), and precision, while classes 2 and 3 showed moderate performance. Class 2 had a sensitivity of 62.2% and a positive predictive value (PPV) of 42.5%, while class 3 demonstrated sensitivity of 55.26% and a PPV of 70%. These results suggest the model performs well for class 1 but struggles to some extent with distinguishing between classes 2 and 3.

Overall, the CART model provides a solid baseline with high interpretability and strong performance for class 1. The lower metrics for classes 2 and 3 indicate the need for further improvement, such as incorporating additional features or leveraging ensemble methods like Random Forests. These enhancements could help achieve a better balance across all classes and improve the model's overall performance.

Random Forest:

The Random Forest model effectively classifies customers into three clusters based on transactional data and customer attributes. The confusion matrix reveals high precision for class 1, with 240 correct predictions and minimal misclassifications. Class 2 achieved moderate accuracy, with 174 correct predictions but significant misclassifications into class 3. Class 3 exhibited the weakest performance, with 107 correct predictions and frequent misclassifications into classes 1 and 2. Overall, the model achieved an accuracy of 72.36%, significantly outperforming random guessing, as confirmed by the p-value (<2.2e-16). The kappa statistics of 0.5854 indicates moderate agreement between predictions and actual values.

Class-wise metrics show that class 1 performed exceptionally well, with 100% sensitivity and 98.75% specificity, highlighting the model's ability to identify this class with minimal errors. Class 2 demonstrated moderate performance, with a sensitivity of 72.5% but a lower positive predictive value (57.81%), suggesting issues with precision. Class 3 showed the most challenges, with a sensitivity of 44.58% and a positive predictive value of 61.85%, reflecting difficulty in distinguishing it from other classes. These results suggest that while the model handles class 1 effectively, it struggles with differentiating classes 2 and 3.

To enhance performance, especially for classes 2 and 3, additional techniques such as feature engineering, hyperparameter tuning, or using ensemble methods can be explored. Balancing the dataset and analyzing feature importance might also improve classification. Despite these challenges, the model provides valuable insights into customer segmentation, offering a solid foundation for targeted strategies.

KNN:

The KNN model achieves an overall accuracy of 69.58%, with a kappa statistic of 0.5438, reflecting moderate agreement between predicted and actual classifications. The confusion matrix reveals excellent performance for class 1, with 240 correct predictions and minimal misclassifications. However, the model struggles with class 2 and class 3, as a significant number of observations for these classes are misclassified, particularly between these two classes. The McNemar's Test p-value (0.1287) suggests no significant differences in misclassification patterns.

Class-wise metrics highlight the strengths and weaknesses of the model. Class 1 exhibits perfect sensitivity (100%) and high specificity (98.96%), ensuring accurate classification with minimal errors. In contrast, class 2 and class 3 show moderate sensitivity (57.5% and 51.25%, respectively) and lower precision, as evidenced by positive predictive values of 54.98% and 54.91%. Balanced accuracy for class 2 (66.98%) and class 3 (65.10%) indicates the need for improvement in correctly identifying and differentiating these classes.

While the model performs exceptionally well for class 1, its difficulties with classes 2 and 3 suggest room for optimization. Adjustments such as tuning the number of neighbors (k), refining feature selection, or rescaling the dataset could enhance performance. Overall, the KNN model provides a strong foundation but requires further refinement to achieve balanced accuracy across all classes.

XGBoost:

The XGBoost model was trained using parameters optimized for multi-class classification, including a maximum tree depth of 6, a learning rate of 0.3, and sampling ratios of 0.8 for both columns and rows. The model was evaluated for 100 iterations using the multi-class log-loss metric, achieving an overall accuracy of 72.78% with a confidence interval of 69.37% to 76.00%. The kappa statistic of 0.5917 reflects moderate agreement between predicted and actual classifications, and the statistically significant p-value (<2e-16) confirms the model's strong performance compared to random guessing.

Class-wise performance reveals exceptional results for class 1, with perfect sensitivity (100%), high specificity (98.75%), and a positive predictive value (PPV) of 97.56%. For class 2, the sensitivity was moderate at 67.5%, with a specificity of 76.67% and a PPV of 59.12%, indicating challenges in precision. Class 3 performed the weakest, with sensitivity of 50.83% and a PPV of 61.00%, highlighting difficulty in distinguishing this class from others. While the model performs well overall, the lower metrics for class 2 and class 3 suggest room for improvement in distinguishing between these classes.

The XGBoost model demonstrates strong overall performance, particularly for class 1, but further refinement is necessary to improve its handling of classes 2 and 3. Potential enhancements include hyperparameter tuning, feature engineering, or addressing class imbalance to reduce misclassification. Exploring feature importance could also provide insights into improving the separation of these classes, leading to a more balanced and accurate model.

The XGBoost model was ultimately selected as our final model due to its strong overall performance and ability to effectively capture the relationships within the data. Despite some challenges with classifying classes 2 and 3, the model's superior accuracy of 72.78% and its well-balanced metrics compared to alternative approaches made it the most reliable choice. Additionally, its flexibility for further optimization and robust handling of feature importance provided a solid foundation for addressing the classification problem. This final model's results will serve as the basis for actionable insights and recommendations moving forward.

MARKET BASKET ANALYSIS:

The market basket analysis reveals significant associations among product categories, offering valuable insights into customer purchasing behavior. For example, the rule {Bags, Office} => {Drinkware} has a support of 1.43%, indicating that this combination appears in 1.43% of all transactions. Its confidence of 54.24% means that when customers purchase {Bags, Office}, there is a 54.24% likelihood they will also purchase {Drinkware}. Additionally, the lift of 5.385 shows that this association is over 5 times more likely than random chance, making it a strong candidate for cross-selling opportunities.

Similarly, the rule {Bags, Lifestyle} => {Office} demonstrates a confidence of 68% and a lift of 4.833, suggesting a strong relationship between these products. Another notable rule, {Drinkware, Lifestyle} => {Office}, has a confidence of 64.66% and a lift of 4.595, further highlighting meaningful associations that can guide marketing strategies.

In contrast, rules like {Lifestyle} => {Office} and {Headgear} => {Apparel} have lower lifts, at 3.653 and 1.985, respectively, indicating weaker but still noteworthy relationships. These findings can be used to design targeted promotions, improve product placements, and drive cross-selling initiatives. The analysis emphasizes high-confidence, high-lift rules as key opportunities to enhance sales strategies and improve customer satisfaction. Marketing campaigns should target top-selling categories like **Apparel** and frequently associated items like {Office} and {Drinkware}. Additionally, ensuring adequate inventory for linked products will help meet customer demand effectively and maximize sales potential.

CHALLENGES AND LIMITATIONS

Despite significant advancements, several challenges persist:

- Data Quality: Missing data, noise, and outliers can undermine the reliability of segmentation and prediction models, requiring robust data preprocessing methods.
- Adapting to Change: Traditional RFM models often struggle to capture rapidly evolving customer behaviors, particularly in fast-moving industries like e-commerce.
- Non-Contractual Relationships: Predicting CLV for customers without ongoing contractual ties remains challenging due to intermittent engagement and limited behavioral consistency.

Future Research Directions

Several promising areas for further exploration include:

- Hybrid Modeling Approaches: Combining traditional probability models with machine learning can balance interpretability and predictive accuracy, offering the best of both worlds.
- Real-Time Analytics: Developing systems capable of processing live data streams will enable dynamic segmentation and more responsive marketing efforts.
- Industry-Specific Models: Expanding CLV applications to sectors such as healthcare and financial services could unlock new insights and opportunities for customer engagement.

Conclusion

Via this project we highlight the transformative power of integrating traditional customer analytics methods with advanced machine learning techniques to drive actionable insights in a competitive e-commerce landscape. By leveraging a comprehensive dataset, we effectively combined Recency, Frequency, and Monetary (RFM) analysis with modern clustering and classification models, enabling precise customer segmentation and prediction of behavior.

The segmentation process identified three distinct customer clusters, each representing unique behavioral patterns. Cluster 1 included high-value, loyal customers who contribute significantly to revenue, while Clusters 2 and 3 comprised moderate and low-engagement customers, offering opportunities for re-engagement and nurturing. These insights provide a foundation for personalized marketing strategies, such as premium loyalty programs for high-value customers and targeted promotions for less active segments.

Advanced models, including Random Forest and XGBoost, demonstrated strong performance in predicting customer behavior, particularly for high-value clusters. The XGBoost model emerged as the most reliable, offering a balanced accuracy of 72.78% and actionable insights for future marketing initiatives. Additionally, market basket analysis revealed meaningful associations between product categories, such as {Bags, Office} => {Drinkware}, which can guide cross-selling strategies and inventory planning.

Despite the promising results, challenges such as data quality and adapting to rapidly changing customer behaviors still remain. Addressing these limitations through real-time analytics and hybrid modeling approaches can enhance the scalability and adaptability of these methods. Future research could explore the application of these techniques to other industries, such as healthcare or financial services, to uncover further opportunities for innovation.

Overall, this project demonstrates how data-driven insights can enhance customer engagement, improve resource allocation, and drive sustained business growth in an ever-evolving digital marketplace.

References

Sun, Y., Liu, H., & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. Heliyon, 9(2), e13384.

Kumar, V., Ramani, G., & Bohling, T. (2004). Customer lifetime value approaches and best practice applications. Journal of interactive Marketing, 18(3), 60-72.

Lewaaelhamd, I. (2023). Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis. Journal of Data Science and Intelligent Systems, 2(1), 29-36.

Choi, J. A., & Lim, K. (2020). Identifying machine learning techniques for classification of target advertising. ICT Express, 6(3), 175-180.

Gonzalez, M., & Rabbi, F. (2023). Evaluating the Impact of Big Data Analytics on Personalized E-commerce Shopping Experiences and Customer Retention Strategies. Journal of Computational Social Dynamics, 8(2), 13–25.

F. Yoseph and M. AlMalaily, “New market segmentation methods usingenhanced (RFM), CLV, modified regression and clustering methods,” Int.J. Comput. Sci. Inf. Technol., vol. 11, no. 1, pp. 43–60, Feb. 2019.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia computer science, 3, 57-63.

Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications, 31(1), 101-107.