E-Commerce Consumer Behavior Analysis report

FATMA HAMMAMI
M1 BDEEM

Table of Contents

E-Commerce Consumer Behavior Analysis

I. Introduction

E-commerce has altered the shopping patterns of consumers, and the degree of influence on these patterns should be understood by companies. This analytical project is supposed to answer the question: ***What are the key factors influencing consumer shopping behavior?***

This study deals with the above-discussed aspects by analyzing 3900 rows of transactional data, with about 19 variables. The study will focus on exploring the demographic, transactional, and behavioral drivers of purchase amounts.
Key methods under this research include regression analysis, clustering, and statistical tests. The outcome of these findings departs with effective conclusions towards marketing strategies, inventory management, and customer retention.

Major Insights
Customer price sensitivity: The customer is very price-sensitive to the product priced within the price brackets of $25 to $35.
The most buying segments: Regions such as the West along with T-shirts and the season "Fall" have the purchase amounts.
Model consequences: Demographics do not seem to influence much but transactional behaviors like previous purchases and mode of payment really make good predictors.
Strategic Recommendations: Targeted promotions to loyal customers, optimization of stocks for popular segments, and improved marketing during the Fall period would be some of the ways to go.

II. Exploratory Data Analysis and Preprocess

a. Data Cleansing

The dataset had undergone extensive cleaning to get precise and trustworthy analysis over the following processes :

Removal of Missing Value: Deleted the row with missing values in critical fields like Purchase Amount or Payment methods to keep data authenticity.

Identifying and Treatment of Outliers: As for Purchase Amount outliers were analyzed through the consideration of extreme value thresholding (95th percentile cap) for extreme high value and 5th percentile capping for extreme low values reducing skewness.

Dropping Unused Columns: Columns that said nothing more than the provision of unique transaction IDs or timestamps that were irrelevant to the analysis were dropped.

b. Data Preparation

The dataset was clean while it was ready for analysis. Encoding Categorical Variables: Region, Payment Method, and Gender were converted into numerical schemes of one hot encoding or dummy variables to use in statistical models.

Continual Feature Scaling for the continuous variable, for example, Purchase Amount and Previous Purchases were standardized, about the new variables engineered:

Loyalty Indicator: Segmenting customers based on previous purchase counts into "Low," "Medium," and "High" loyalty.

Discount Utilization: A binary variable shows whether a promo code was applied to the purchase.

c. Data Training

Designed the clean and processed dataset into a training and testing set for developing the model.

Splitting the Data:

Training Set- 80% of the data is used to train the regression and clustering models.

Test Set- This is the other 20% of data, kept aside to evaluate the model performance.

Cross-Validation: A technique of k-fold cross-validation is used to ensure strong performance and to avoid overfitting.

Dealing with Class Imbalances: The dependent variable Purchase Amount was stratified to ensure balanced representation of high- and low-end transactions within the sample data across training and testing datasets.

III. Exploratory Data Analysis (EDA)

The most salient patterns and relationships were discovered within the dataset with EDA. Some important findings include:

- Purchase Amount Distribution: Purchase amounts were skewed, with most transactions being relatively low-value, but some significant outliers representing high-value purchases.
- Previous Purchases Distribution: The previous purchases variable displayed a right-skewed distribution, with many customers making fewer than 25 purchases.
- Age and Previous Purchases: In the same manner, no significant correlation could be established between them, indicating that there would not be much difference in the buying frequencies across various age groups.

a. Key Insights:

- Price Sensitivity: A concentration of purchases in the $25 to $35 range suggests that customers are price-sensitive, preferring products within this "sweet spot."

- Target Market: The e-commerce platform primarily caters to mid-range consumers rather than luxury or bargain shoppers.

These findings reinforce the conclusions drawn from the linear regression model, further confirming that demographic variables like age and past purchasing behavior have limited influence on purchase amounts.

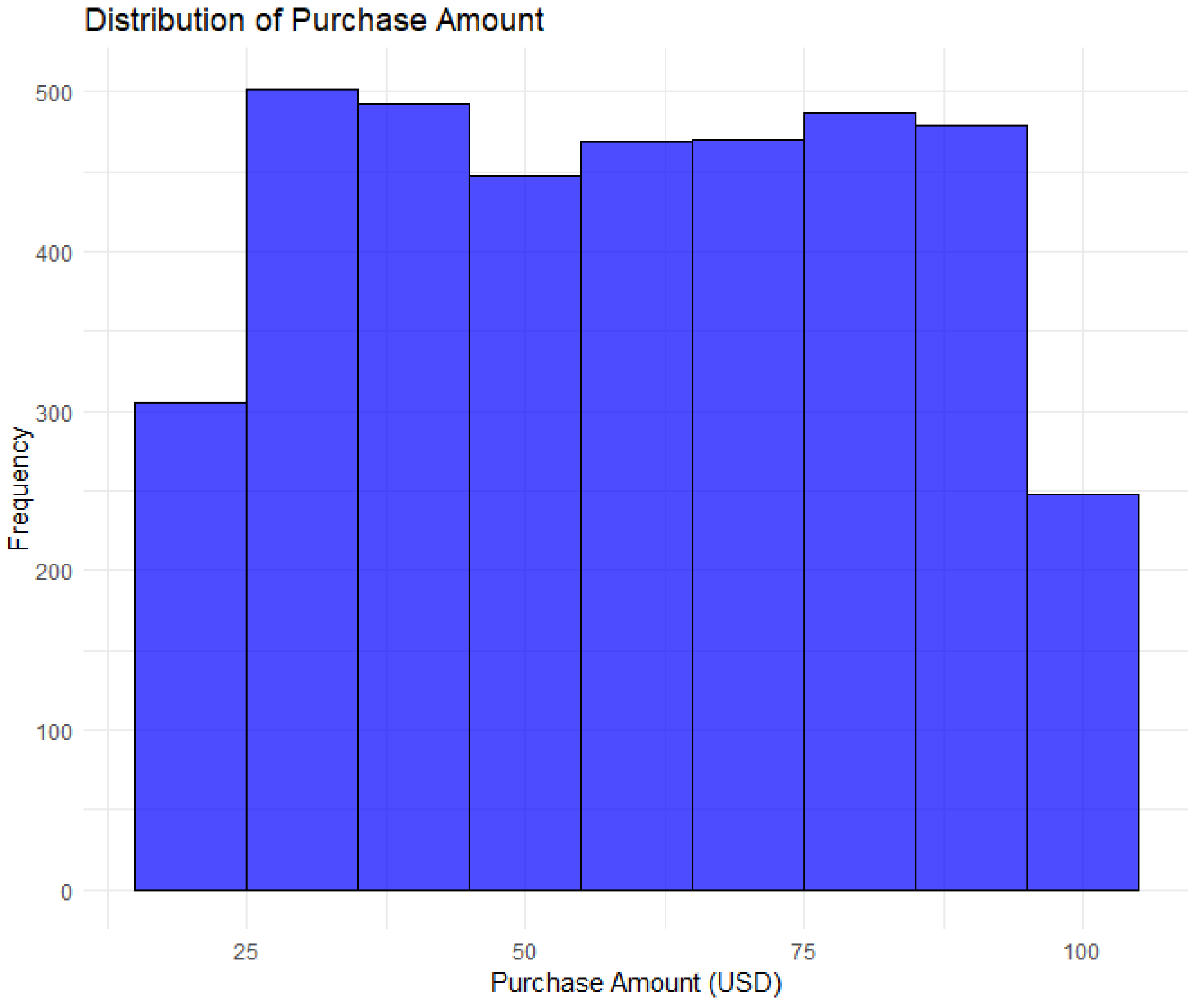
---

b. Visualizations

1. Purchase Amount Distribution :

- This distribution suggests that most customers tend to make purchases in the mid-range of $25 to $95, with a particular concentration in the $25 to $35 range. The lower frequency of very low-value and high-value purchases indicates that these are less common in this e-commerce dataset.

Key Insights from the Graph:

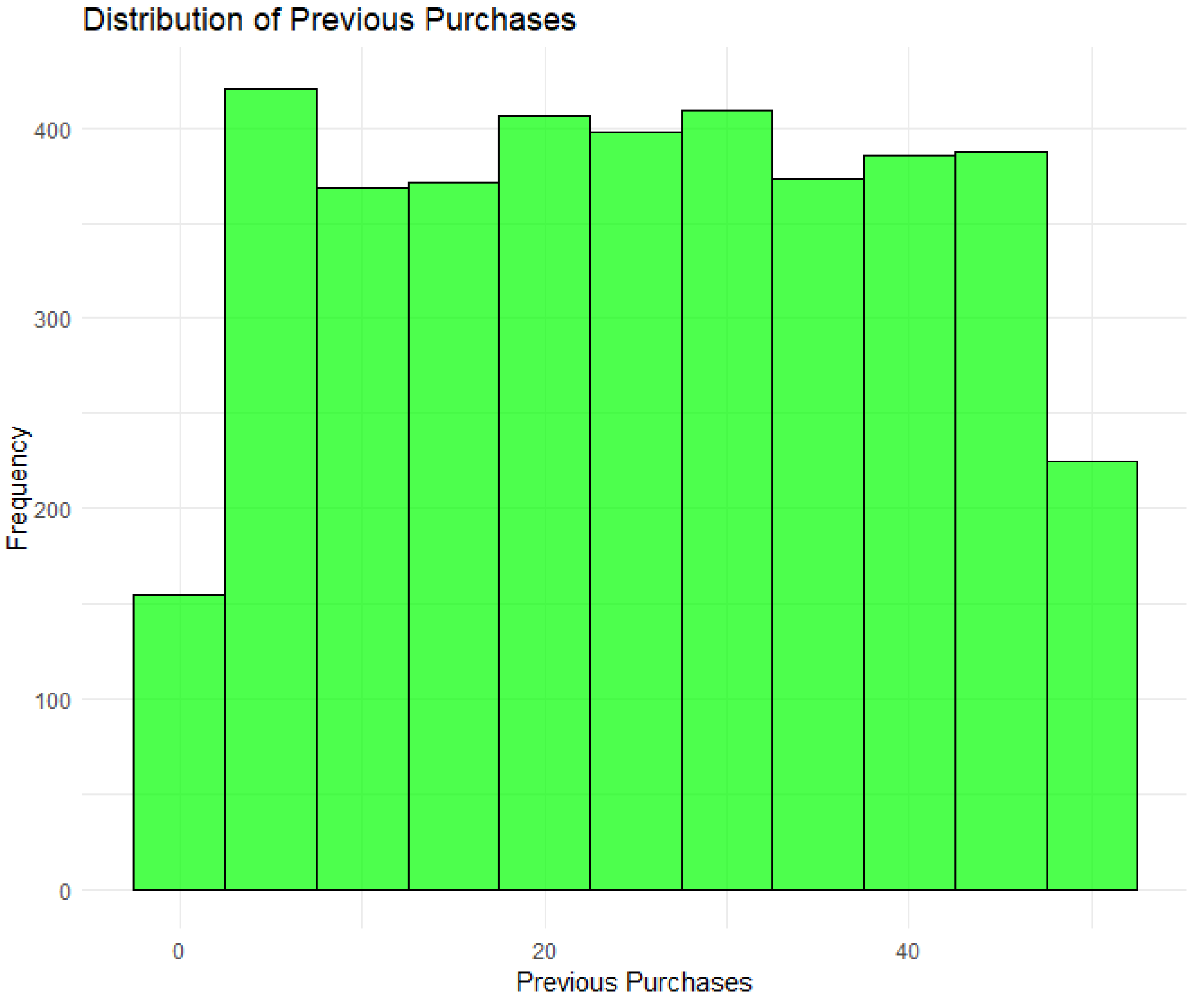


- Price Sensitivity: Most consumers prefer products priced between $25 and $35, indicating price sensitivity.
- Target Market: The platform primarily caters to middle-range consumers.
- High-Value Purchases: There is an opportunity to target customers for higher-value items to increase sales.

2. Previous Purchases Distribution :

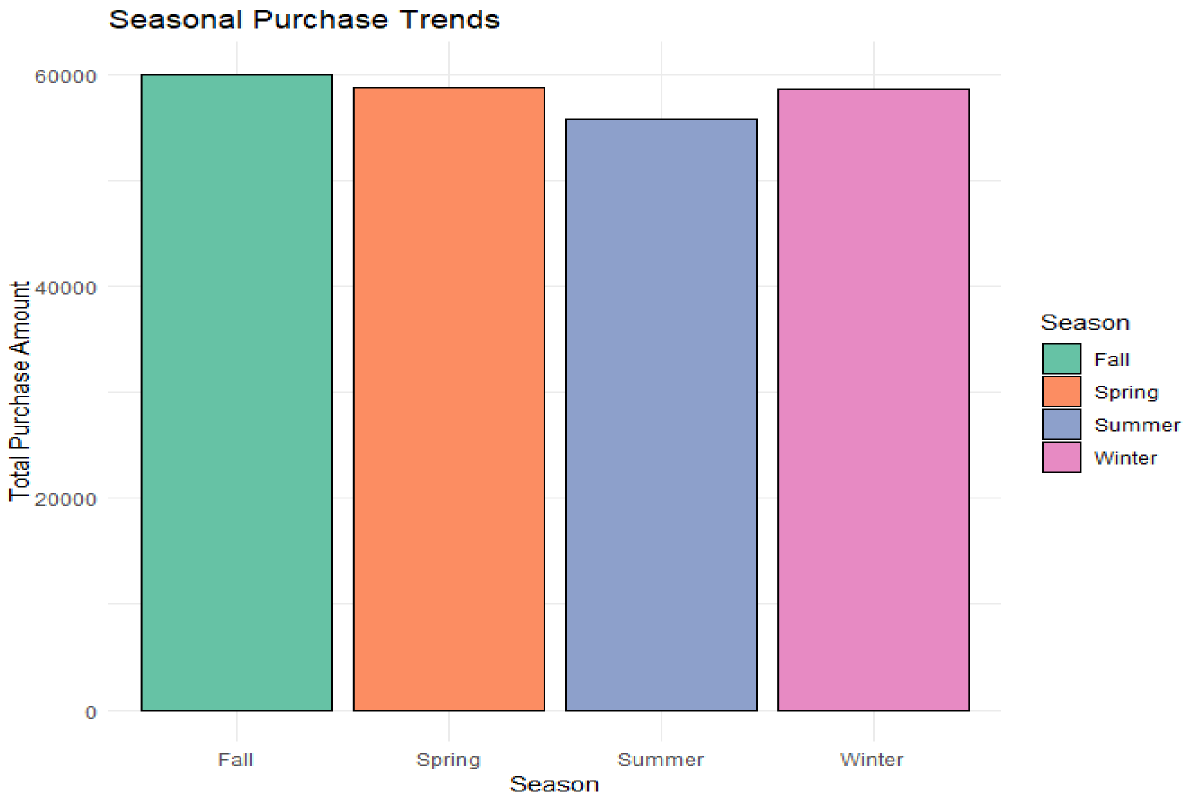
The previous purchases exhibit a right-skewed distribution, suggesting that many customers have made fewer than 25 previous purchases, while a small number have made significantly more.

Key Insights from the Graph:



- Customer Loyalty: A significant proportion of customers have made between 4 and 46 previous purchases, indicating a loyal customer base.
- Retention Challenges: The platform faces challenges in retaining first-time or infrequent buyers (those with fewer than 4 purchases).
- VIP Customers: A small segment of highly loyal customers has made more than 46 purchases, representing potential VIP clients.

3. Seasonal Trends



The observed seasonal trends in the purchase amounts state that:

Fall is the First Season:

- Interpretation: Fall leads the total amount purchased, implying that customers spend more during this time. Several factors may explain this:

Preparations for Holidays: Fall routinely precedes other major holidays which, generally, include Thanksgiving, Halloween and the beginning of the winter holiday shopping season. Therefore, expenses increase, as customers anticipate the expenditure for celebration and other seasonal events.

New Product Launches: Most companies launch new collection or products at the start of fall, especially in the fashion or electronic sectors, resulting in large purchase volumes.

Buying Behavior: Customers start buying items for use during the colder season or turn to refresh their wardrobing or homesteading for autumn, with much more money spent.

Winter as Second Leading Season:

- Interpretation: Winter automatically gains second position as it hosts the major events of holiday shopping like Christmas, New Year's, and Black Friday, for example. Consumers spend more in winter due to:

Gifting: Holidays mean a lot of purchase boosts for people.

Price Cuts and Discounts: Another reason is that most retailers do "Final Clearance Sales" at the end of the year or after Christmas. This encourages most consumers to buy in bulk.

Because of the Activity During This Season: People also buy items associated with the cold, like coats, heaters, or holiday-related goods.

Spring as the Third Season

- Exposition: Moderate amounts of purchase generally characterizes spring and are likely due to:

Spring Cleaning: Many people would engage themselves in spring cleaning activities after the long winter and buying things for home improvement and organizing their houses. However, it does not amount to that huge spending, as that of the fall or winter.

Transition Seasons: People now will start purchasing light clothing; accessories However, purchase activity won't be as high as builds pre-purchases for the holiday season or even fall.

Outdoor Activities: Purchases for outdoor or gardening supplies may happen but would not be on par with how winter products sell or how people generally shop in the fall.

Summer as the Lowest Season

- Interpretation: Summer being the lowest purchase amount is explained as follows:

Vacation Mode: Most of the consumers spend their summer holidays vacationing and devoting their attention to travel and leisure rather than shopping.

Experiential Spending: In summer, people usually spend on experiences rather than products (holidays, outdoor activities, etc.), which further lessens the amount of purchase.

Very Little Shopping Events: Since summer does not have holidays with major shopping-related activities, there are very few sales events and promotions as compared to other seasons, thereby resulting in less spending**.**

4. Regional Trends:

- The West is multiyear in having the highest prediction of purchase amounts.

- There are clear differences in purchase amounts in several regions.

5. Product Categories:

- The predictions of purchase amounts are high for the most part associated with T-shirts and boots.

- Very different results occur within the same category of products.

6. Payment Methods:

- Credit card and Venmo seem to be the most probable payment methods with a value associated with higher predicted purchase amounts.

- In most instances, electronic payments are higher than cash payments in terms of purchase amounts.

7. Shipping Preferences:

- 2-day and express promised delivery have been tied with predicted purchases more robust than the average standard shipping.

- Standard shipping has the lowest predicted median purchase amount of all varieties.

8. Size Trends:

- For the most part, bigger sizes (XL, L) have slightly higher predictions in terms of purchase amounts.

- Compared to the other categories, the variation in purchase amounts across sizes is less standard.

Targeting specific customer needs should focus marketing strategies on that segment and inventory management with the aim of customer segmentation. The lack of association of region with subscription status suggests that subscription marketing strategies would not need to be different according to regions.

IV. Regression Models and Statistical Tests

i. Linear Regression Model:

This was the primary model that focused on predicting Purchase Amount using Age, Gender, and Previous Purchases.

1. Model Interpretation:

R-squared: 0.0016, which statistically denotes that the model explains very small part of the variance in purchase amounts.

2. Coefficients:

- **Age**: Indicates very low positive relationship between age and purchase amounts.

- **Gender:** Gender coefficient denoted that both men and women do not show any difference in respect to purchase amounts.
- **Previous Purchases:** This particular variable has so small impact showing positive relationship with purchase amounts.

3. Diagnostic Tests:

Breusch-Pagan test: The result of 0.735 indicated the presence of no heteroskedasticity in the model thereby confirming the constant variance assumption in residuals**.**

- **Cross-validated linear model has been applied in order to improve the robustness of the model.**

4. Cross-validations:

- Root Mean Squared Error (RMSE): 23.70
- R-squared: 0.0016
- Mean Absolute Error (MAE): 20.61.

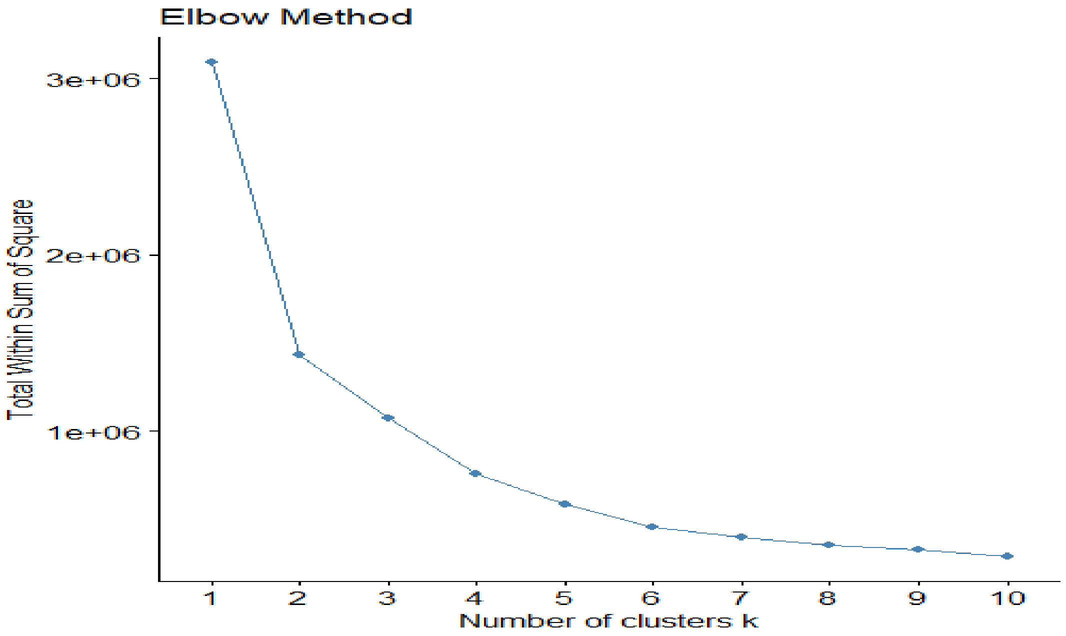
5. Interpretation:

- An RMSE of $23.70 shows that on average predictions are off by that much, which is really high considering the average purchase amount.
- R-squared is low indicating the independent variables used don't explain much of the variance in the purchase amounts, so something else drives consumer spending behavior.

ii. K-Means Clustering :

1. Elbow Method

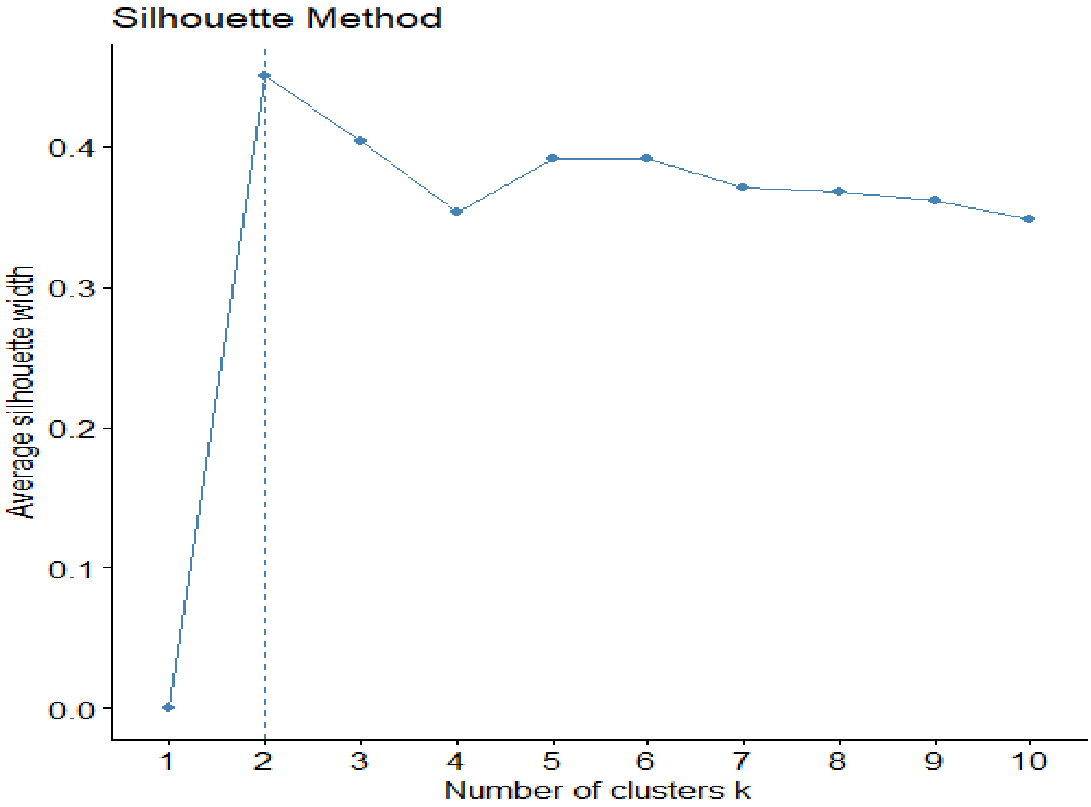
K-means clustering was used to segment customers based on Age and Purchase Amount. To determine the optimal number of clusters, the Elbow Method was employed by plotting the within-cluster sum of squares (WSS) for various values of k.



The plot revealed an elbow point at k = 3, indicating that three clusters provide the best fit.

2. Silhouette Analysis:

To validate the cluster quality, the silhouette method was used. The average silhouette width for k = 2 confirmed that the clustering structure is strong and well-separated.

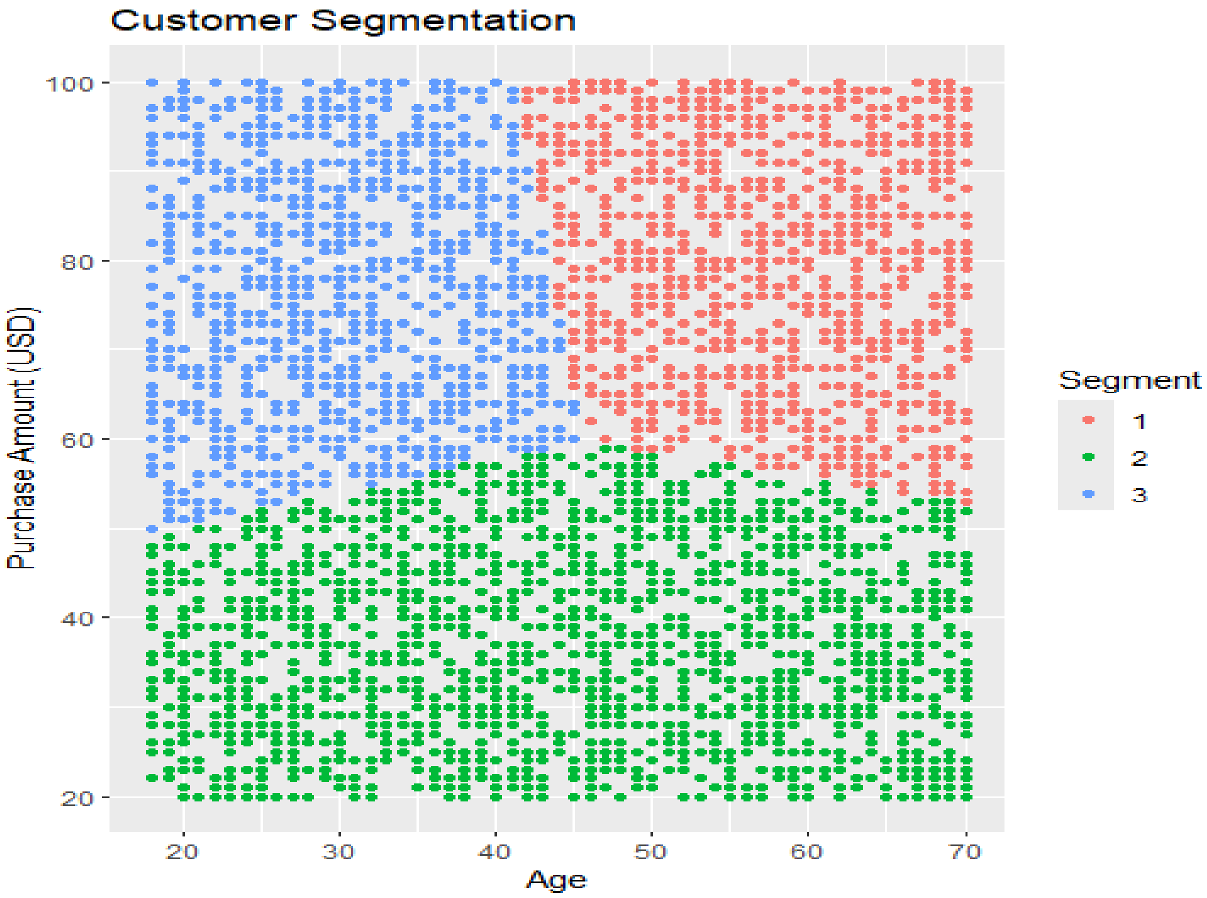


3. Cluster Insights:

- **Segment 1 (Aged 45-70)**: Moderate spenders, with a purchase peak around $58.
- **Segment 2 (All Ages)**: A diverse segment, with a broad range of purchase amounts.
- **Segment 3 (Aged up to 45)**: Younger customers who tend to spend more, starting at $60.

4. Visualization:

this plot clearly distinguishes the three segments based on age and spending, helping to inform targeted marketing strategies.



iii. Random Forest Analysis

A Random Forest model was built to predict purchase amounts using multiple variables, offering a non-linear approach to explore the dataset.

1. Model Insights:

- **Variable Importance:** The Random Forest identified Previous Purchases and Payment Method as the most significant predictors of Purchase Amount.
- **Accuracy:** The model achieved higher predictive performance compared to linear regression, though it remains limited by the features available in the dataset.

Top Performing Segments

Category	Top Performer
Region	West
Item	T-shirt
Payment Method	Credit Card
Shipping Type	Store Pickup
Season	Fall
Size	L

These top performers represent the categories associated with the highest predicted purchase amounts.

---

iv. Statistical Test Results

- Treatment of the Breusch-Pagan Test (Heteroscedasticity Test) :
  - The result of the BP test on heteroscedasticity showed the presence of no significant heteroscedasticity at the p-value of 0.735 on the residual of the linear regression model. This gives credence to the fact that the error assumption will therefore have constant variance.
- ANOVA Result :
  - The ANOVA examination considered the differences of purchase amounts from various regions and genders. The p-values do not prove that any are markedly differing cases, and thus, it is not asserted that purchase behavior is to any extent reliant upon gender or region.
- Interpretation of Results of Tukey's HSD Pairwise Comparisons:
  - The most comparisons show that the differences in means between regions are less, with p-values going well above the threshold set for significance, $p < .05$.
- Confidence Intervals:
  - The confidence intervals (lwr and upr) contain zero for all pairwise comparisons; that is, there are no price category differences made by purchase between any two regions.
- Conclusion:
  - Hence, by both ANOVA and Tukey's HSD tests, there is no significant difference between the purchase amounts across different regions.
- Results of Chi-square test:

- X-Squared=3.5666
- degrees of freedom=5
- p-value=0.6133

- ***Interpretation:***

The p-value (0.6133) is above the common threshold significance level of 0.05. Thus, it indicates the absence of any statistically significant association between Region and Subscription Status. In other words, customers do not seem likely to subscribe based on their region or factor.

- In synthesis, the analysis indicates that purchase amounts within this database are not primarily affected by region, thereby reinforcing that marketing strategies and price points may not need to have varied effects due to geographical distinctions since the behavior of customers relating to expenditures is uniform across regions. Thus, the results should sustain decision-making related to the efficient use of resources and marketing strategies that do not necessitate regional differentiation on the basis of purchase amounts alone.

To further contextualize the findings of this analysis, a SWOT table has been prepared highlighting the strengths, weaknesses, opportunities, and threats of the e-commerce platform. It summarizes the internal and external factors influencing strategic decision-making simply.

SWOT Analysis	**Details**
Strengths	- **High Predictive Accuracy with Random Forest**: Offers actionable insights into consumer behavior.
	- **Loyal Customer Base**: VIP customers can be leveraged for exclusive offers and loyalty programs.
	- **Clear Segment Identification**: K-means clustering segments customers, enabling targeted marketing.
Weaknesses	- **Limited Predictive Power of Demographic Variables**: Age and gender have minimal impact on purchase amounts.
	- **Promo Code Effectiveness**: Promo codes do not significantly influence purchase amounts.
	- **Lack of Significant Regional Differences**: No regional effect on consumer behavior limits regional strategies.
Opportunities	- **Targeted Marketing for Younger Segments**: Focus on younger consumers with targeted promotions.
	- **Optimization of Store Pickup Options**: Leverage store pickup preference to optimize logistics and reduce costs.
	- **Personalized Discounts and Loyalty Programs**: Experiment with personalized promotions and loyalty rewards.
Threats	- **Price Sensitivity**: Concentration of purchases in the $25-$35 range suggests challenges in offering premium products.

SWOT Analysis	Details
	- **Retention of First-Time Buyers**: Difficulty in retaining customers who make few purchases.
	- **Competition**: Strong competitors with better personalization or promotional strategies could attract more customers.

V. Recommendations

Based on the findings of this analysis, several actionable recommendations can be made for improving marketing strategies and targeting consumer segments:

1. Segmented Marketing Campaigns:

- ✓ Develop targeted campaigns for each segment identified through K-means clustering.
- ✓ For Segment 3 (young consumers), focus on high-value products and exclusive deals to increase spending.
- ✓ For Segment 1, promote products that cater to older consumers' preferences.

2. Promotions and Discounts:

Since promo codes did not significantly influence purchase amounts, consider experimenting with other types of promotions, such as loyalty programs or personalized discounts based on browsing history.

3. Product and Service Development:

Create new product lines or services that specifically appeal to the spending habits of Segment 1 (moderate spenders aged 45-70). Segment 2 represents a broad customer base, so marketing campaigns should be more inclusive, catering to a wide range of preferences.

4. Seasonal Marketing Strategy:

Leverage the identified seasonal trends to adjust marketing strategies and promotions

5. Logistics and Shipping Optimization:

Focus on optimizing store pickup options as this was identified as a preferred shipping type, which can also reduce logistics costs. • Consider promotional campaigns that highlight the convenience and cost-effectiveness of store pickups.

6. Invest in Predictive Analytics:

While Random Forest gave better predictive performance, the features in dataset limited its accuracy. Increase the data collection on browsing habits, customer reviews, and product preferences for future prediction improvements.

7. Customize Regionally:

Although the regional influences were not very significant, performance among the West region was slightly better. Regional data for tailored marketing and product availability aligned with localized market demand be used.

8. Increased Loyalty:

For a small segment of highly loyal customers (VIPs), develop special loyalty plans in order to reward them for frequent purchases. A personalized offer for these customers can lead to increased revenues.

VI. Conclusion and Perspectives

The analysis reveals that price sensitivity, product preference, shipping method, previous purchase behavior, and other behavioral variables play a significant role in determining consumer shopping patterns. However, regional differences appear not to matter in terms of purchase amounts. Businesses could potentially invest in segmented marketing, exceptional conditioning of logistics, and adopt such promotional strategies to achieve the right fit with customer preference and improved sales performance.

Employed thus was a complete picture approach for this study, exploratory analysis of data, statistical testing, clustering, and predictive modeling to unravel shopping behavior's main drivers in terms of price. There were minimal predictive capabilities for amount spent on purchases regarding demographic variables; conversely, behavioral variables were of more significance. K-means clustering and Random Forest modeling provided actionable insights for customer segmentation and preferences while at the same time emphasizing the need for advanced analytical methods in understanding the consumer behavior phenomenon.

Perspectives :

Digital Integration: The creation of a mobile app or website that enables the establishment of shopping experiences with consumer segments resulting in targeted promotions and individualized recommendations.

Integration of Algorithms: The adoption of trading algorithms for superior stock optimization coupled with dynamic pricing strategies.

Advanced Analysis: Apply more stringent analysis using the latest machine learning (ML) and deep learning (DL)-based algorithms for better prediction and identifying industry-specific insight and intricate patterns in consumer behavior.

references and sources for my report

Academic and General References:

1. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.
    - For insights on statistical tests, linear regression, and Random Forest methods.
2. **Kaufman, L., & Rousseeuw, P. J.** (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley-Interscience.
    - For methodologies related to K-means clustering and silhouette analysis.
3. **Lander, J. P.** (2014). *R for Everyone: Advanced Analytics and Graphics.* Addison-Wesley.
    - Relevant for data manipulation and visualization using tidyverse and ggplot2.
4. **Kabacoff, R. I.** (2015). *R in Action: Data Analysis and Graphics with R.* Manning Publications.
    - For guidance on exploratory data analysis and statistical tests.

Online Resources:

5. **Kaggle Dataset: Customer Shopping Latest Trends**
   Source: Kaggle (Dataset link, if applicable)
    - Original dataset used for the analysis, consisting of 3,900 rows and 19 variables.
6. **R Documentation:**
    - *Base R and tidyverse documentation:* https://cran.r-project.org/
    - *ggplot2 documentation:* https://ggplot2.tidyverse.org/
    - *dplyr documentation:* https://dplyr.tidyverse.org/
7. **UCLA Statistical Consulting Group**
   Website: https://stats.oarc.ucla.edu/r/
    - A valuable resource for implementing regression and ANOVA in R.

Specific Tests and Methodologies:

8. **Tukey's HSD Test:**
    - Details on post hoc analysis and interpretation of ANOVA results.
      Source: https://www.statisticshowto.com/tukeys-test/
9. **Breusch-Pagan Test:**
    - For testing heteroscedasticity in regression models.
      Source: https://www.statology.org/breusch-pagan-test-r/

Statistical Analysis Tools:

10. **R Studio:**

- Integrated Development Environment (IDE) for R used throughout the analysis.
  Website: https://posit.co/products/open-source/rstudio/

11. **R Packages:**

  - tidyverse: Data manipulation and visualization.
  - car: For statistical tests like Breusch-Pagan.
  - cluster: For K-means and silhouette analysis.
  - randomForest: For building Random Forest models.