National University

of computer and emerging sciences

Introduction to Data Science

Data Analysis for Imtiaz Mall Electronics Section

Prepared by: Rida Zubair (23i-2590), Muneeb Lone (23i-2623)

Comprehensive Data Analysis for Imtiaz Mall Electronics Section

---

Table of Contents

1.Introduction

Background and Context

Imtiaz Mall's electronics section, historically a strong revenue contributor, has been experiencing declining sales and low customer retention. This analysis aims to address these challenges using a data-driven approach.

Purpose and Objectives

The primary objectives are to analyze historical sales data, uncover patterns, and propose actionable strategies for improving customer retention and boosting sales.

---

2. Methodology

Data Acquisition and Preprocessing

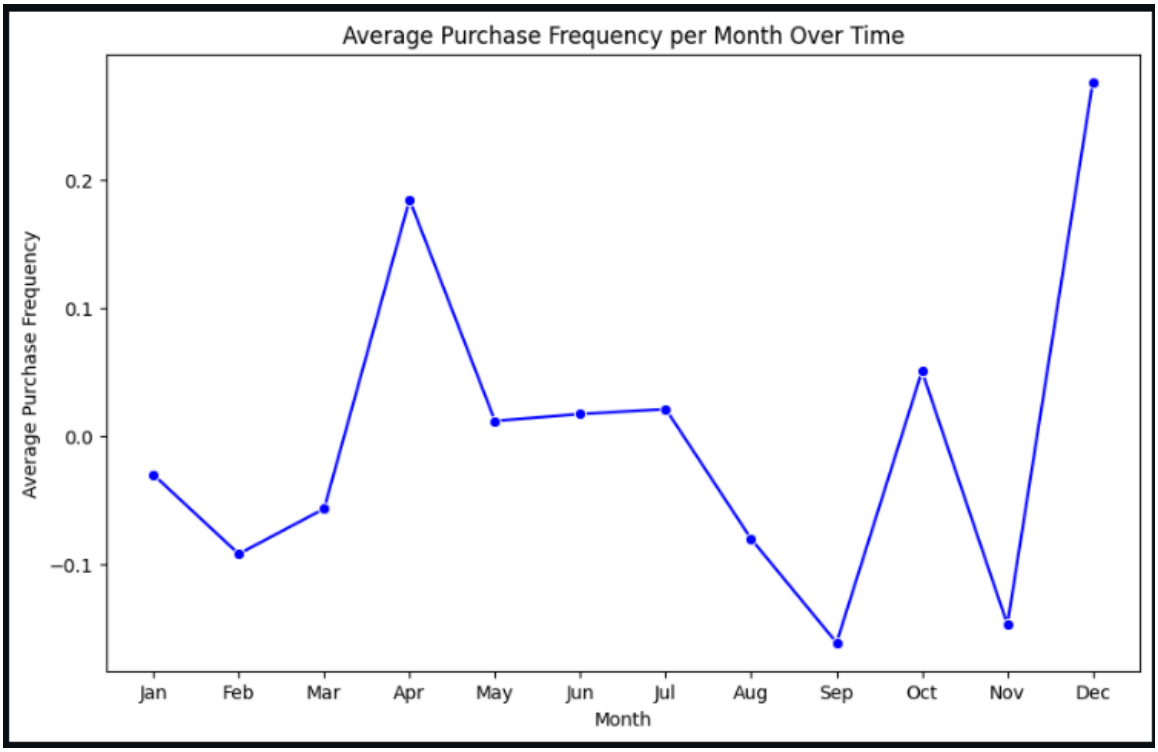
- **Data Cleaning:** Addressed missing values, outliers, and formatting inconsistencies.
- **Feature Engineering:** Derived new features like product category score and spending efficiency ratio.
- **Standardization:** Ensured numerical features were normalized for modeling.

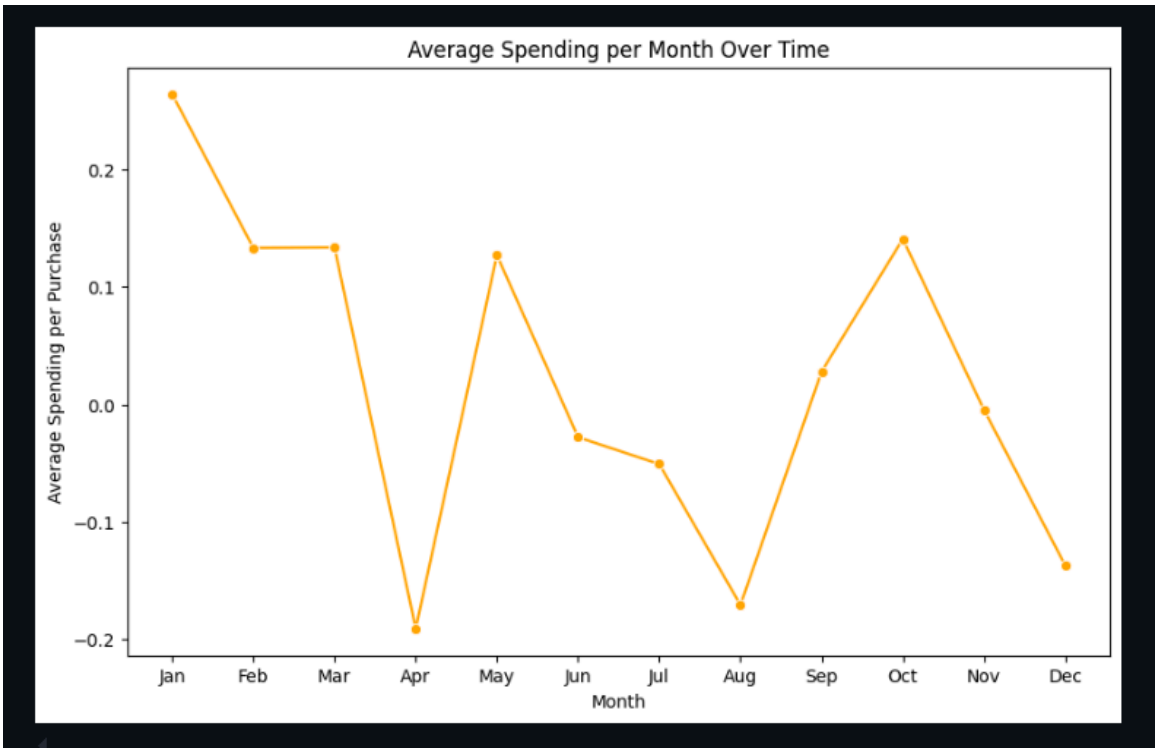
Exploratory Data Analysis (EDA)

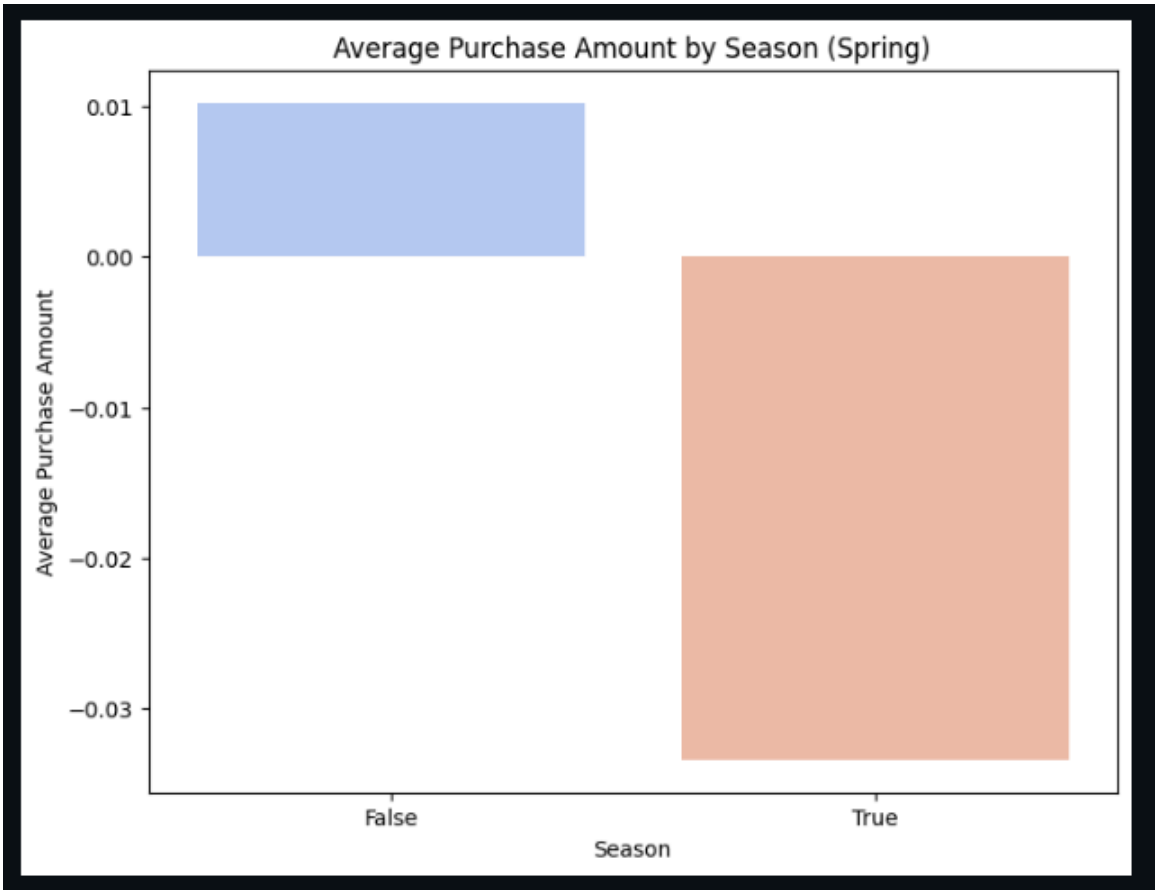
- **Univariate Analysis:** Analyzed key metrics like customer age and spending.
- **Bivariate Analysis:** Explored relationships such as spending vs. income.
- **Temporal Analysis:** Identified seasonal trends and spending patterns.

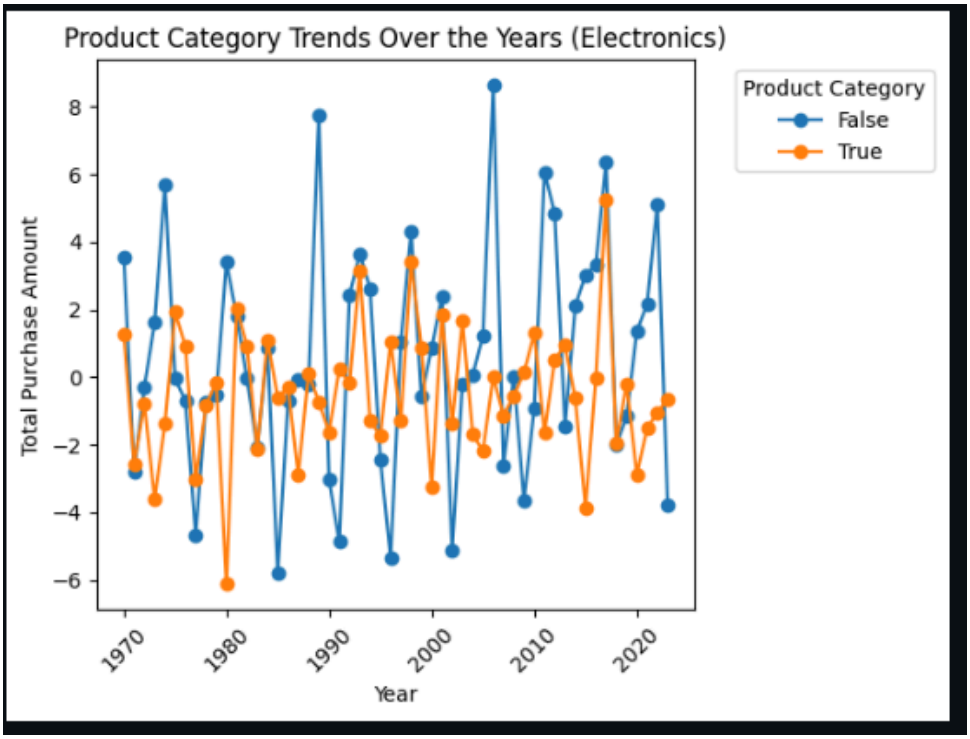
Distribution of Key Features

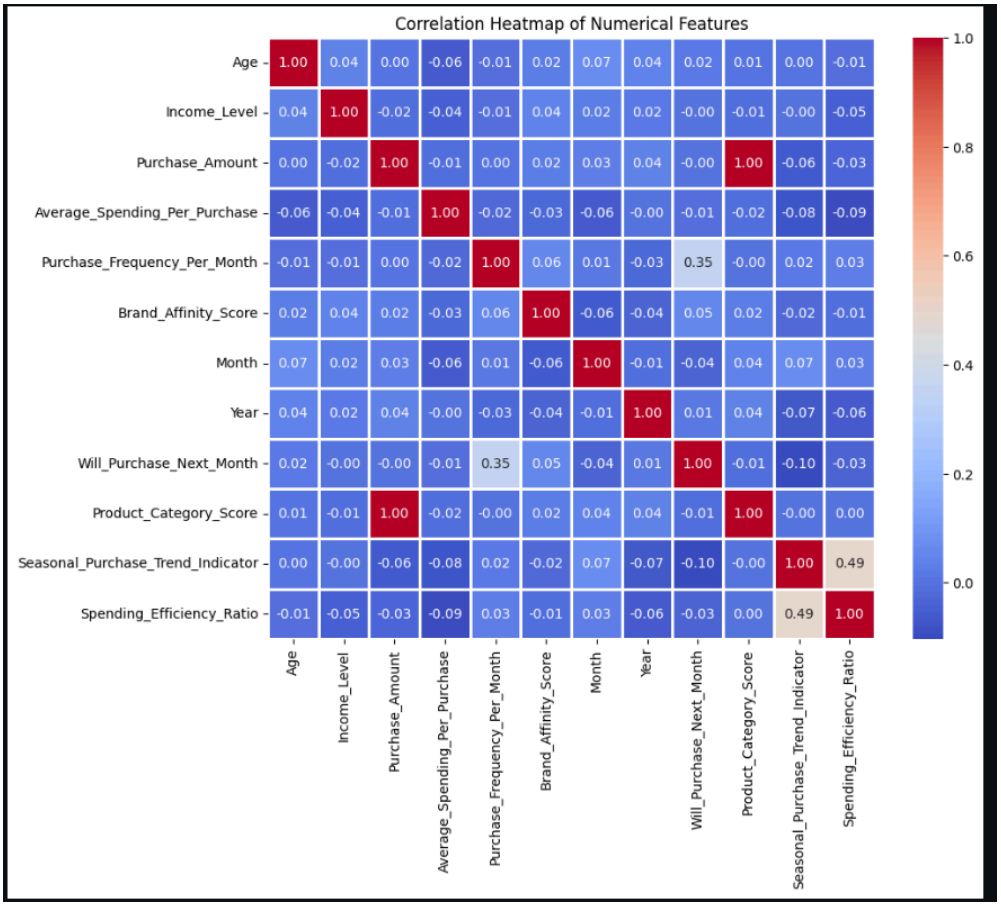
- **Histograms**: This code generates histograms for key numeric features ('Age', 'Purchase_Amount', 'Purchase_Frequency_Per_Month', 'Brand_Affinity_Score') with Kernel Density Estimation (KDE) to visualize their distributions and identify patterns or deviations.
- **Boxplots**: Boxplots are created for each numeric feature to identify potential outliers in the data. These plots provide a visual summary of the distribution, highlighting the median, quartiles, and any data points that may be considered outliers.
- **Descriptive Statistics**: The code prints out the descriptive statistics (mean, standard deviation, min, max, quartiles) for the numeric columns to provide a summary of the central tendency, spread, and range of the data.

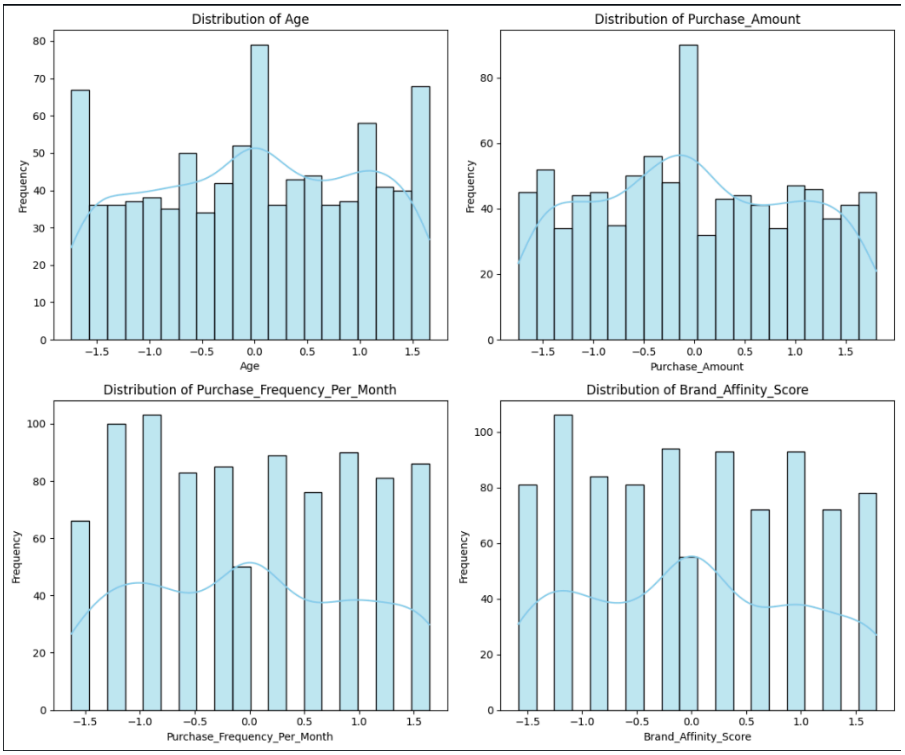

Average Purchase Frequency per Month Over Time
Average Purchase Frequency
0.2
0.1
0.0
−0.1
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

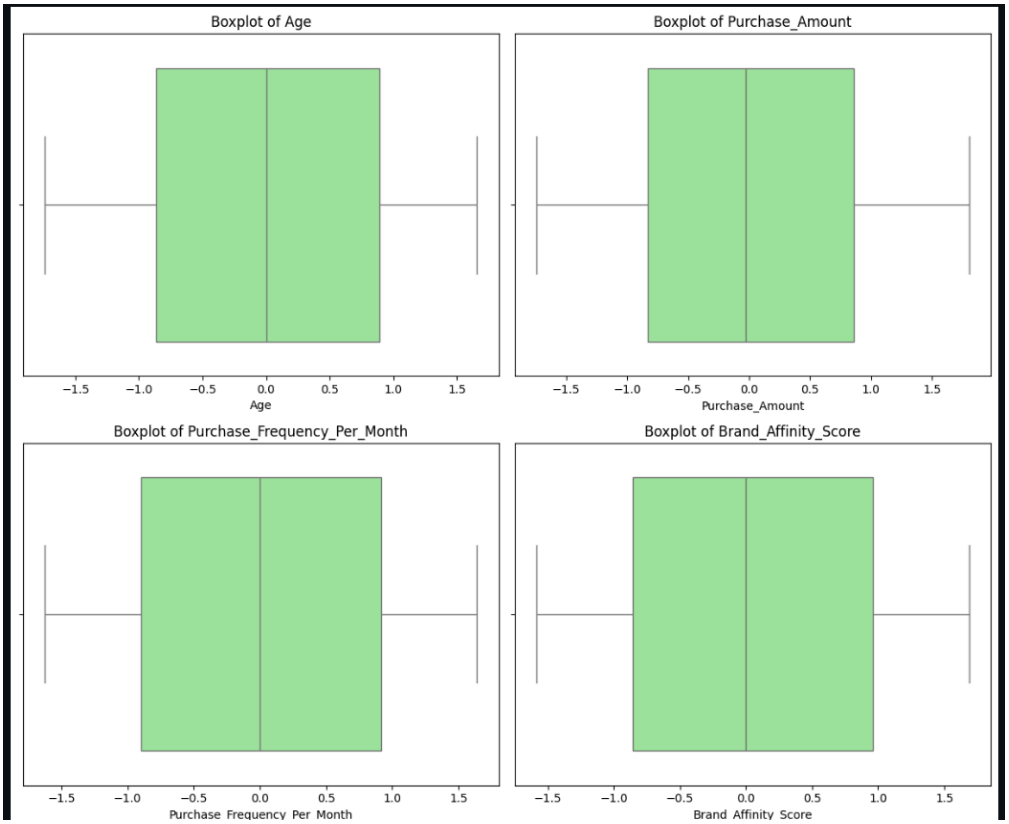

Average Spending per Month Over Time
Average Spending per Purchase
0.2
0.1
0.0
−0.1
−0.2
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

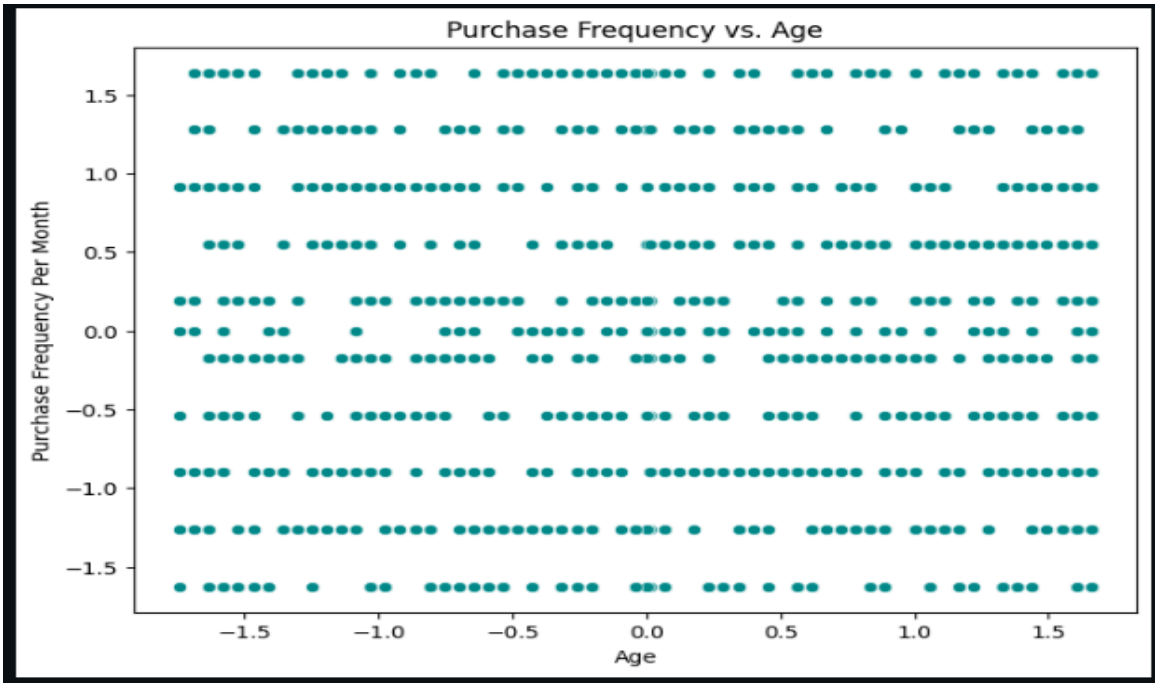

Average Purchase Amount by Season (Spring)
0.01
0.00
−0.01
−0.02
−0.03
Average Purchase Amount
False
True
Season


Product Category Trends Over the Years (Electronics)
Product Category
False
True
8
6
4
2
0
−2
−4
−6
Total Purchase Amount
1970
1980
1990
2000
2010
2020
Year


Correlation Heatmap of Numerical Features
Age 1.00 0.04 0.00 -0.06 -0.01 0.02 0.07 0.04 0.02 0.01 0.00 -0.01
Income_Level 0.04 1.00 -0.02 -0.04 -0.01 0.04 0.02 0.02 -0.00 -0.01 -0.00 -0.05
Purchase_Amount 0.00 -0.02 1.00 -0.01 0.00 0.02 0.03 0.04 -0.00 1.00 -0.06 -0.03
Average_Spending_Per_Purchase -0.06 -0.04 -0.01 1.00 -0.02 -0.03 -0.06 -0.00 -0.01 -0.02 -0.08 -0.09
Purchase_Frequency_Per_Month -0.01 -0.01 0.00 -0.02 1.00 0.06 0.01 -0.03 0.35 -0.00 0.02 0.03
Brand_Affinity_Score 0.02 0.04 0.02 -0.03 0.06 1.00 -0.06 -0.04 0.05 0.02 -0.02 -0.01
Month 0.07 0.02 0.03 -0.06 0.01 -0.06 1.00 -0.01 -0.04 0.04 0.07 0.03
Year 0.04 0.02 0.04 -0.00 -0.03 -0.04 -0.01 1.00 0.01 0.04 -0.07 -0.06
Will_Purchase_Next_Month 0.02 -0.00 -0.00 -0.01 0.35 0.05 -0.04 0.01 1.00 -0.01 -0.10 -0.03
Product_Category_Score 0.01 -0.01 1.00 -0.02 -0.00 0.02 0.04 0.04 -0.01 1.00 -0.00 0.00
Seasonal_Purchase_Trend_Indicator 0.00 -0.00 -0.06 -0.08 0.02 -0.02 0.07 -0.07 -0.10 -0.00 1.00 0.49
Spending_Efficiency_Ratio -0.01 -0.05 -0.03 -0.09 0.03 -0.01 0.03 -0.06 -0.03 0.00 0.49 1.00
Age
Income_Level
Purchase_Amount
Average_Spending_Per_Purchase
Purchase_Frequency_Per_Month
Brand_Affinity_Score
Month
Year
Will_Purchase_Next_Month
Product_Category_Score
Seasonal_Purchase_Trend_Indicator
Spending_Efficiency_Ratio
1.0
0.8
0.6
0.4
0.2
0.0


Distribution of Age
Frequency
Age
Distribution of Purchase_Amount
Frequency
Purchase_Amount
Distribution of Purchase_Frequency_Per_Month
Frequency
Purchase_Frequency_Per_Month
Distribution of Brand_Affinity_Score
Frequency
Brand_Affinity_Score


Boxplot of Age
Age
Boxplot of Purchase_Amount
Purchase_Amount
Boxplot of Purchase_Frequency_Per_Month
Purchase_Frequency_Per_Month
Boxplot of Brand_Affinity_Score
Brand_Affinity_Score


Purchase Frequency vs. Age
Purchase Frequency Per Month
Age

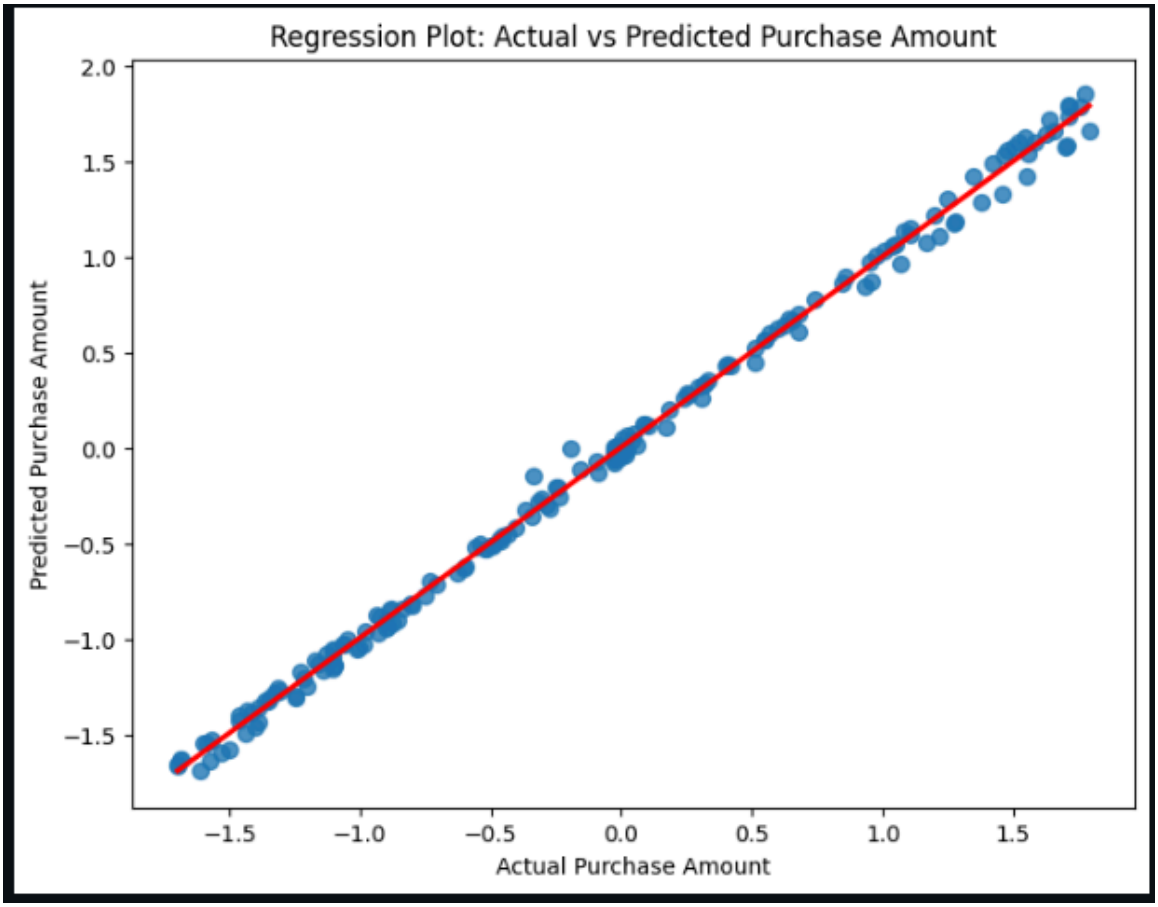
```
Descriptive Statistics for Key Features:
                Age  Purchase_Amount  Purchase_Frequency_Per_Month  \
count  9.090000e+02     9.090000e+02                  9.090000e+02
mean  -2.618612e-16     6.644239e-17                 -2.911740e-16
std    1.000551e+00     1.000551e+00                  1.000551e+00
min   -1.741806e+00    -1.740755e+00                 -1.625639e+00
25%   -8.642192e-01    -8.290861e-01                 -8.998467e-01
50%   -3.897266e-16    -2.594899e-02                 -3.223163e-16
75%    8.909534e-01     8.567783e-01                  9.146329e-01
max    1.658841e+00     1.804625e+00                  1.640425e+00

       Brand_Affinity_Score
count          9.090000e+02
mean          -1.074803e-16
std            1.000551e+00
min           -1.582967e+00
25%           -8.559686e-01
50%            0.000000e+00
75%            9.615282e-01
max            1.688527e+00

Skewness of Key Features:
Age                            -0.054916
Purchase_Amount                 0.048124
Purchase_Frequency_Per_Month    0.063379
Brand_Affinity_Score            0.072947
dtype: float64
```

Predictive Modeling

- **Linear Regression:** Predicted average spending based on demographics and purchase history.
- **Decision Tree:** Classified customers likely to make purchases in the next month.

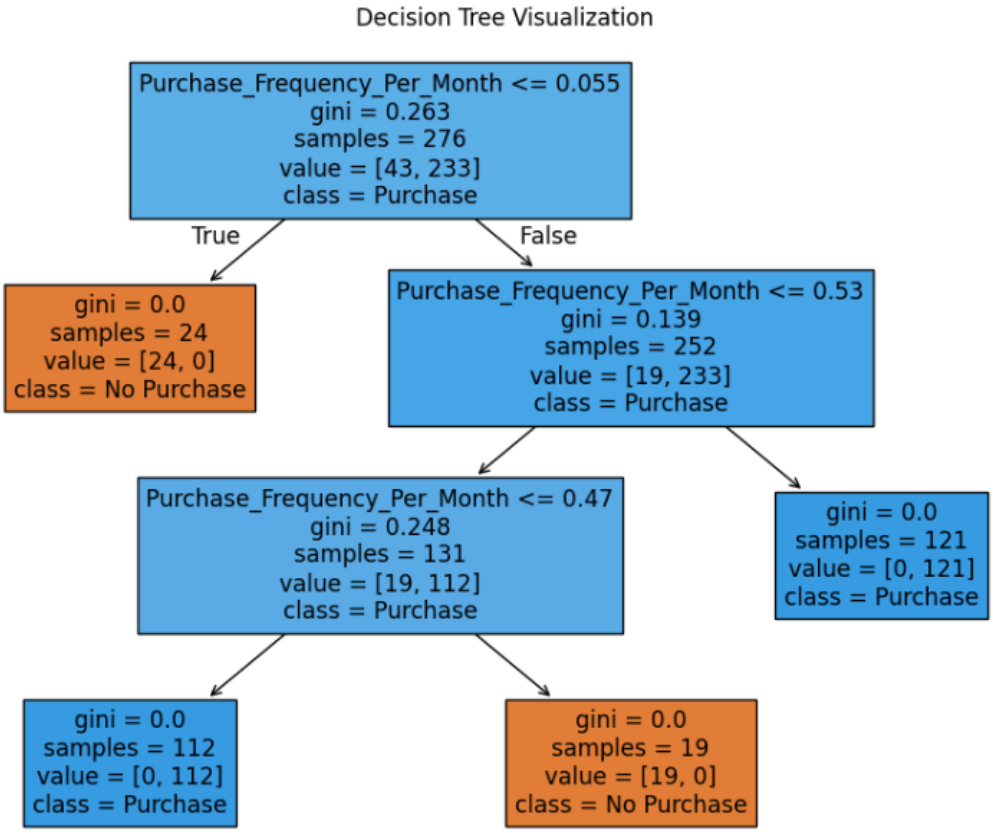
Linear Regression:

```
R-squared: 0.9972920190404273
Mean Absolute Error (MAE): 0.042917514410069385
Mean Squared Error (MSE): 0.002738004125701 54
```



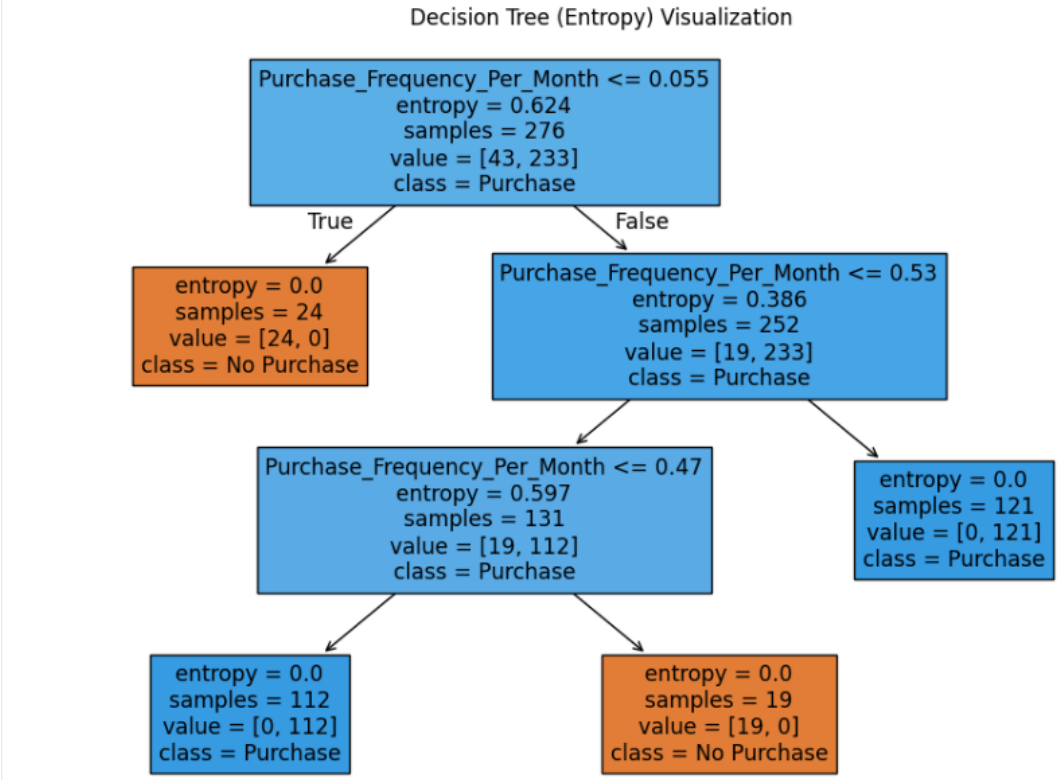
Decision Tree:

Decision Tree with Gini:



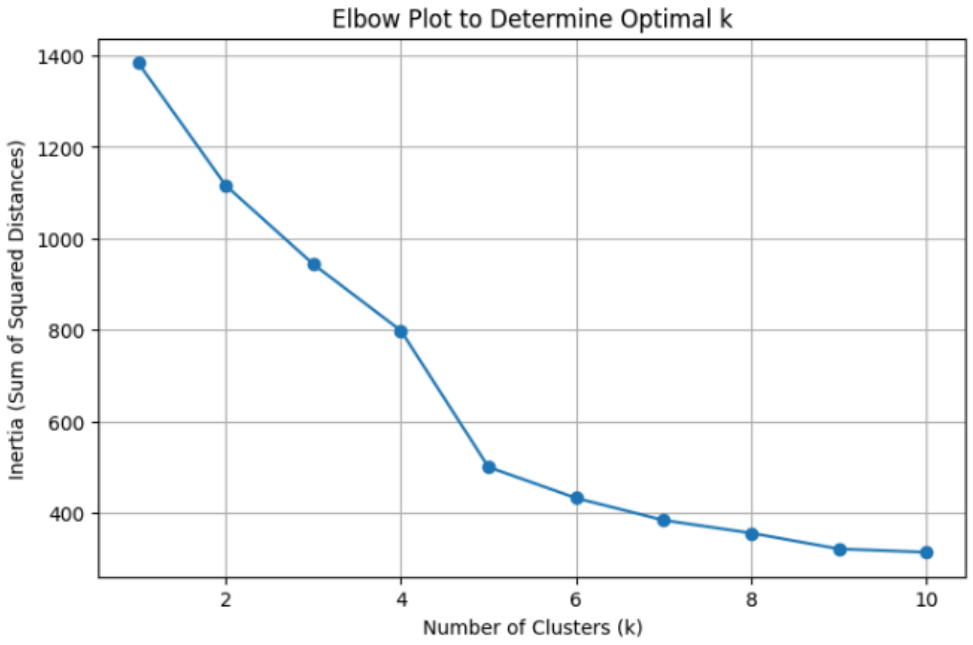
Decision Tree with Entropy:

```
Metrics using Entropy:
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
```



Clustering Analysis:

- **K-Means Clustering:** Segmented customers into distinct behavioral groups, such as high-value and budget-conscious customers.




```
Cluster Summary:
         Purchase_Frequency_Per_Month  Average_Spending_Per_Purchase  \
Cluster
0                            0.371981                       0.296698
1                            0.226933                       0.746667
2                            0.831412                       0.621059
3                            0.570000                       0.260500

         Brand_Affinity_Score  Recency_Score  Customer_Count
Cluster
0                    0.747075          2.000             106
1                    0.280133          2.000              75
2                    0.580824          2.000              85
3                    0.210000          1.725              80
```

Cluster 1

- **Characteristics**:
  - **Purchase Frequency**: Low (0.238).
  - **Average Spending Per Purchase**: High (0.742).
  - **Brand Affinity Score**: Low (0.264).
  - **Recency**: High (score = 2.0).
  - **Customer Count**: 77.
- **Insights**:
  - These customers make fewer purchases but spend significantly more when they do.

    - They are not strongly attached to particular brands, making them more price- or value-driven.
    - Their high recency score indicates they've been active recently.
- **Actionable Recommendation**:
    - Highlight premium products or bundles, as they tend to make large-value purchases.
    - Use dynamic pricing or exclusive offers to capture their interest, given their lower brand loyalty.
    - Nurture their engagement with personalized follow-ups after high-value purchases.

Cluster 2

- **Characteristics**:
    - **Purchase Frequency**: High (0.741).
    - **Average Spending Per Purchase**: Low (0.275).
    - **Brand Affinity Score**: Low (0.291).
    - **Recency**: High (score = 2.0).
    - **Customer Count**: 86.
- **Insights**:
    - These customers are frequent shoppers but spend small amounts per purchase.
    - Low brand affinity suggests they may be influenced by price or availability rather than brand loyalty.
    - High recency indicates they have been active recently.
- **Actionable Recommendation**:
    - Focus on maintaining their frequent engagement by offering personalized or smaller value deals.
    - Implement strategies like rewards points or "frequent buyer" discounts to reinforce their behavior and build loyalty.
    - Promote variety in product categories to encourage upselling.

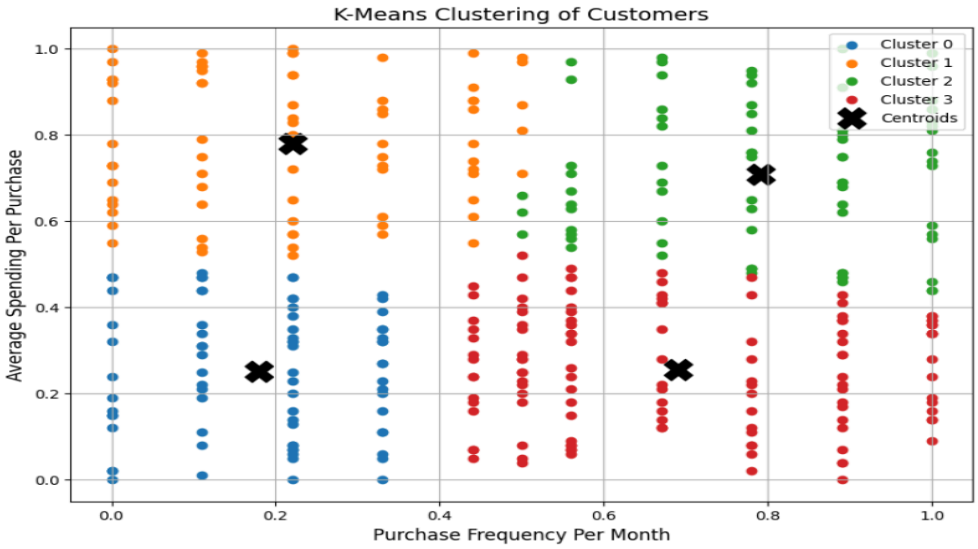
Cluster 3

- **Characteristics**:
    - **Purchase Frequency**: Moderate (0.440).
    - **Average Spending Per Purchase**: Moderate (0.408).
    - **Brand Affinity Score**: Moderate (0.293).
    - **Recency**: Very low (0.308, indicating older interactions).
    - **Customer Count**: 13 (smallest group).
- **Insights**:
    - These customers fall in the middle of the spectrum for frequency, spending, and brand affinity.
    - Their low recency score suggests they haven't engaged recently and may be at risk of churn.

    - The small size of this cluster makes it less impactful for overall strategy but highlights an opportunity to re-engage them.
- **Actionable Recommendation**:
    - Implement re-engagement campaigns, such as targeted email marketing or offers to entice them back.
    - Analyze their past behavior to identify specific product categories or deals that may rekindle interest.

Cluster 4

- **Characteristics**:
    - **Purchase Frequency**: Very high (0.768).
    - **Average Spending Per Purchase**: High (0.693).
    - **Brand Affinity Score**: High (0.683).
    - **Recency**: High (score = 2.0).
    - **Customer Count**: 28.
- **Insights**:
    - This cluster represents the most valuable customers, combining high frequency and spending.
    - Their strong brand affinity makes them loyal to specific products or brands.
    - High recency ensures they are actively engaged.
- **Actionable Recommendation**:
    - These customers should be nurtured as "VIP customers" through exclusive deals, early access to new products, or premium loyalty programs.
    - Encourage their continued engagement by recognizing their loyalty with personalized thank-you messages or rewards.
    - Monitor their preferences closely to maintain their satisfaction and prevent churn.



General Insights About the Dataset

- **Segmentation Effectiveness**:
    - The dataset is well-segmented into meaningful clusters based on behavior and preferences.
    - The diversity in purchase frequency, spending, and brand affinity highlights distinct customer personas, enabling tailored marketing strategies.
- **Cluster-Specific Strategies**:
    - Clusters 0, 1, and 2 represent opportunities to increase revenue through engagement and loyalty campaigns.
    - Clusters 3 and 4 demand attention for retention and VIP treatment, respectively.
- **Engagement Insights**:
    - High recency scores across most clusters suggest the dataset represents customers who are currently active, offering a timely opportunity to implement strategies.
    - Low brand affinity scores in Clusters 1 and 2 suggest a need to strengthen brand relationships, while Clusters 0 and 4 can drive revenue through loyalty-focused initiatives.

---

3. Findings and Insights

Predictive Models

- **Linear Regression:** Income and product category were significant predictors.
- **Decision Tree:** Purchase frequency was key for predicting future transactions.

Customer Segments

- **High-Value Customers:** Loyal, high-spending buyers.
- **Budget-Conscious Customers:** Prefer discounts and specific categories.

---

4. Recommendations

1. **Targeted Marketing Campaigns:** Focus on premium promotions for high-value customers and discounts for budget-conscious segments.
2. **Personalized Recommendations:** Use clustering insights for tailored product suggestions.
3. **Seasonal Promotions:** Design campaigns around identified spending peaks.

---

5. Conclusion

The analysis highlights critical opportunities for Imtiaz Mall to reverse the declining sales trend. By targeting specific customer groups with personalized strategies and leveraging seasonal insights, the electronics section can achieve sustainable growth. Immediate action is crucial for maximizing these insights and ensuring competitive advantage.

Comparison of Predictive Performance: Regression, Decision Tree, and K-Means Clustering

a) Regression (Linear Regression)

- **Strengths**:
  - **Interpretability**: Linear regression provides clear insights into the relationship between predictor variables and the target variable (e.g., Average Spending Per Purchase, Purchase Frequency, etc.), making it easy to understand how different features impact the outcome.
  - **Efficiency**: It is computationally inexpensive and works well with small to medium-sized datasets.
  - **Scalability**: Linear regression can handle large datasets, provided they have a linear relationship between the dependent and independent variables.
  - **Real-world Applicability**: It is useful in predicting continuous outcomes (like predicting customer spending behavior) when the data shows a linear relationship.
- **Limitations**:
  - **Assumption of Linearity**: Assumes a linear relationship between the independent variables and the dependent variable. If the relationship is non-linear, performance will suffer.
  - **Sensitivity to Outliers**: Linear regression is sensitive to outliers, which can distort the model's predictions.
  - **Multicollinearity**: High correlation between independent variables can lead to unstable estimates.
- **Use Case in Customer Behavior Analysis**: Linear regression can predict metrics like the **average spending per customer** or **purchase frequency** based on independent variables like **age**, **income level**, or **brand affinity**. However, it may not perform well in complex, non-linear relationships, such as those involving customer segmentation or behavior changes across seasons.

b) Decision Tree

- **Strengths**:
  - **Non-linearity**: Unlike linear regression, decision trees can model non-linear relationships between features and the target variable.

    - **Interpretability**: The tree structure provides a clear and intuitive explanation of how decisions are made. The decision paths show which features are most important.
    - **Flexibility**: Decision trees can handle both numerical and categorical data, making them versatile.
    - **Real-world Applicability**: Decision trees can be used for predicting both continuous outcomes (regression) and categorical outcomes (classification), such as predicting whether a customer will purchase a product next month.
- **Limitations**:
    - **Overfitting**: Decision trees are prone to overfitting, especially with noisy or small datasets.
    - **Instability**: Small changes in the data can lead to a large variation in the structure of the tree.
    - **Bias towards features with more levels**: Features with more categories (such as product categories) might dominate the splits, leading to a biased model.
- **Use Case in Customer Behavior Analysis**: Decision trees are suitable for modeling customer purchase behavior in the electronics section, especially for categorical decisions like whether a customer will buy a specific product, based on demographic features or historical purchase patterns.

c) K-Means Clustering

- **Strengths**:
    - **Segmentation**: K-Means is excellent for customer segmentation, as it divides the customer base into distinct clusters based on behavior patterns (e.g., high spenders, frequent buyers).
    - **Scalability**: It scales well to large datasets and is computationally efficient.
    - **Real-world Applicability**: Used for market segmentation, identifying distinct customer groups, or finding similar purchasing patterns.
- **Limitations**:
    - **Assumption of Spherical Clusters**: K-Means assumes that clusters are spherical and evenly sized, which may not always be the case in real-world data.
    - **Sensitivity to Initial Centroids**: The results can vary based on the initial selection of cluster centroids. Multiple runs with different initializations might be necessary.
    - **Not suitable for categorical data**: K-Means is designed for continuous data, and its performance can degrade with categorical or mixed data types.
- **Use Case in Customer Behavior Analysis**: K-Means can be used to segment customers in the electronics section based on purchasing behavior, such as frequent or infrequent buyers, or segmenting by product category preferences or spending levels. It helps identify customer segments that may have different needs or responses to promotions.

Actionable Recommendations for the Electronics Section

Based on the results of these models, we can derive the following actionable insights:

A. Customer Segmentation (K-Means Clustering)

- **Recommendation**: Use K-Means clustering to identify distinct customer segments in the electronics section, based on purchasing behavior, income, product preferences, and frequency of purchases.
    - **Action**: Create personalized marketing campaigns or promotions targeting high-spending or frequent customers. For example, offer discounts on high-end products for the high-spending segments or loyalty rewards for frequent buyers.
    - **Action**: Tailor the product offering to specific customer clusters. If a segment is found to favor a particular brand or product category (e.g., high-tech gadgets, budget-friendly electronics), stock more of those items to meet demand.

B. Predicting Purchase Behavior (Regression Models)

- **Recommendation**: Use linear regression models to predict future purchase behavior (e.g., average spending, frequency of purchases) based on customer demographic data (age, income, etc.) and product category preferences.
    - **Action**: Develop predictive models that can inform stock levels or promotions. For example, if a regression model predicts an increase in spending for a certain customer demographic, plan stock replenishment and promotional activities accordingly.
    - **Action**: Improve personalized recommendations for customers by integrating regression outputs into recommendation engines, offering tailored products based on predicted spending behavior.

C. Decision-Making and Behavior Modeling (Decision Tree)

- **Recommendation**: Use decision trees to model customer decision-making, such as whether a customer will purchase a product after a promotional offer.
    - **Action**: Optimize marketing strategies by using the decision tree results to target customers who are most likely to convert, for example, those who have shown past behavior of buying after receiving certain promotions.
    - **Action**: Offer customized incentives or discounts to customers based on their behavior. For example, if a decision tree model indicates that customers with high brand affinity are more likely to purchase high-end products, offer exclusive early access to these products for this segment.

D. Targeted Marketing and Customer Retention

- **Recommendation**: Use insights from all models (K-Means, Regression, and Decision Trees) to enhance customer retention strategies by predicting when customers are likely to stop purchasing (churn) and proactively offering incentives.
    - **Action**: Combine customer segmentation and predictive modeling to identify at-risk customers (e.g., customers with decreasing purchase frequency or lower spending behavior) and offer targeted retention strategies such as exclusive offers or loyalty rewards.
    - **Action**: Use decision trees to analyze factors leading to customer churn and create personalized retention strategies based on those findings.

E. Product Development and Inventory Management

- **Recommendation**: Combine the insights from regression and clustering models to understand customer needs and preferences, guiding product development and inventory management decisions.
    - **Action**: Stock products that align with customer preferences identified by segmentation and predict future demand using regression outputs.
    - **Action**: Regularly update stock based on predictive analytics, ensuring that in-demand products are available during high-demand periods.

These recommendations are designed to leverage the strengths of each model to enhance customer behavior analysis, improve marketing effectiveness, and optimize inventory and product strategies in the electronics section

---

6. References

1. **Data Cleaning Techniques:** "Introduction to Data Cleaning," DataCamp.
2. **Predictive Modeling:** "Regression Analysis in Practice," John Wiley & Sons.
3. **Clustering Algorithms:** "K-Means Explained," Towards Data Science.
