INTRODUCTION TO DATA SCIENCE
DS-2001
SEMESTER PROJECT

Solution Designed By:

Muddassir Asghar
i23-2577

Muhammad Abdullah Ali
i23-2523

Problem statement:

Imtiaz Mall, a renowned department store chain, is experiencing declining sales and a significant number of non-recurring customers in its electronics section. To address this challenge, you, the newly appointed Senior Data Scientist, have been tasked with conducting a comprehensive analysis of the electronics section data and developing data-driven strategies for customer retention and sales growth. This project focuses on the initial steps of this analysis, specifically exploring the data through various techniques.

Module 1: Data Acquisition and Preprocessing

1. Data Acquisition:

The historical data provided for the electronics section of Imtiaz Mall was downloaded from GCR, then loaded appropriately into the program.

The loaded data certainly includes customer demographics, purchase history, product details, spending amounts, and dates of transactions.

Certain columns were transformed into numeric columns.

2. Data Cleaning:

Missing values

Univariate analysis was performed on the numeric columns to observe the distributions of the columns, it was observed that most of them have symmetric distributions, with a slight noticeable skew in purchase amount and average spending per purchase columns.

For the symmetrically distributed columns, we used mean imputation, and for those with skew, we used mode imputation.

For categorical data, mode imputation was performed.

The above strategies for missing values are appropriate since each column has <5% values missing.

Outliers

IQR detection was used, no outliers were observed in any column, so no action was taken.

Inconsistencies

Unnecessary columns such as customer ID, transaction ID, were dropped as they served no purpose for the analysis, as they were all unique values.

All the categorical data was then transformed into numerical data, e.g. Months were changed to represent January by '1'.

Categorical data was encoded using Hot-One-Encoding, which splits columns like Gender into gender male, gender female, etc. Then one of each encoded columns was dropped, as we do not need 3 categories to predict them, it suffices to know about 2.

3. Data Transformation:

New features

The features to be created, as mentioned in the project document, already existed inside the data provided. It was noticed that the data provided is inconsistent, but to stay true to the data, the existing data was not manipulated, e.g average spending per purchase already existed in the data, which could have otherwise been inferred from purchase frequency per month, and purchase amount columns and so on.

Standardized features

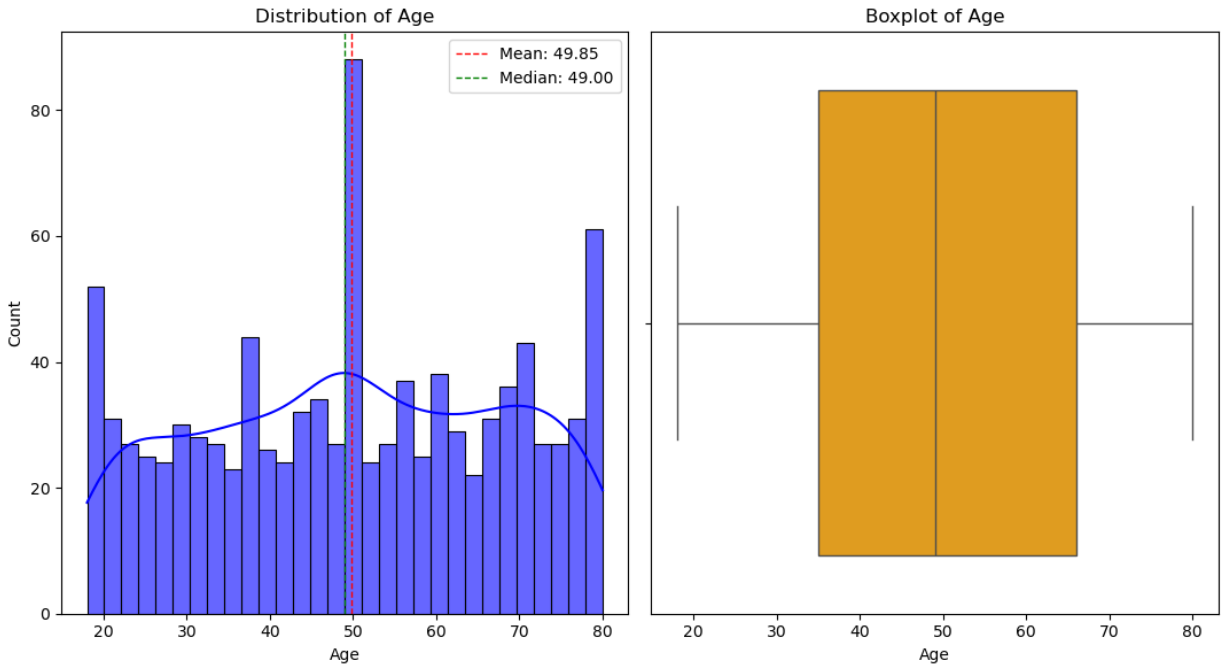
StandardScaler() was used, which transforms data to follow a normal distribution, where the mean is 0, and standard deviation is 1.

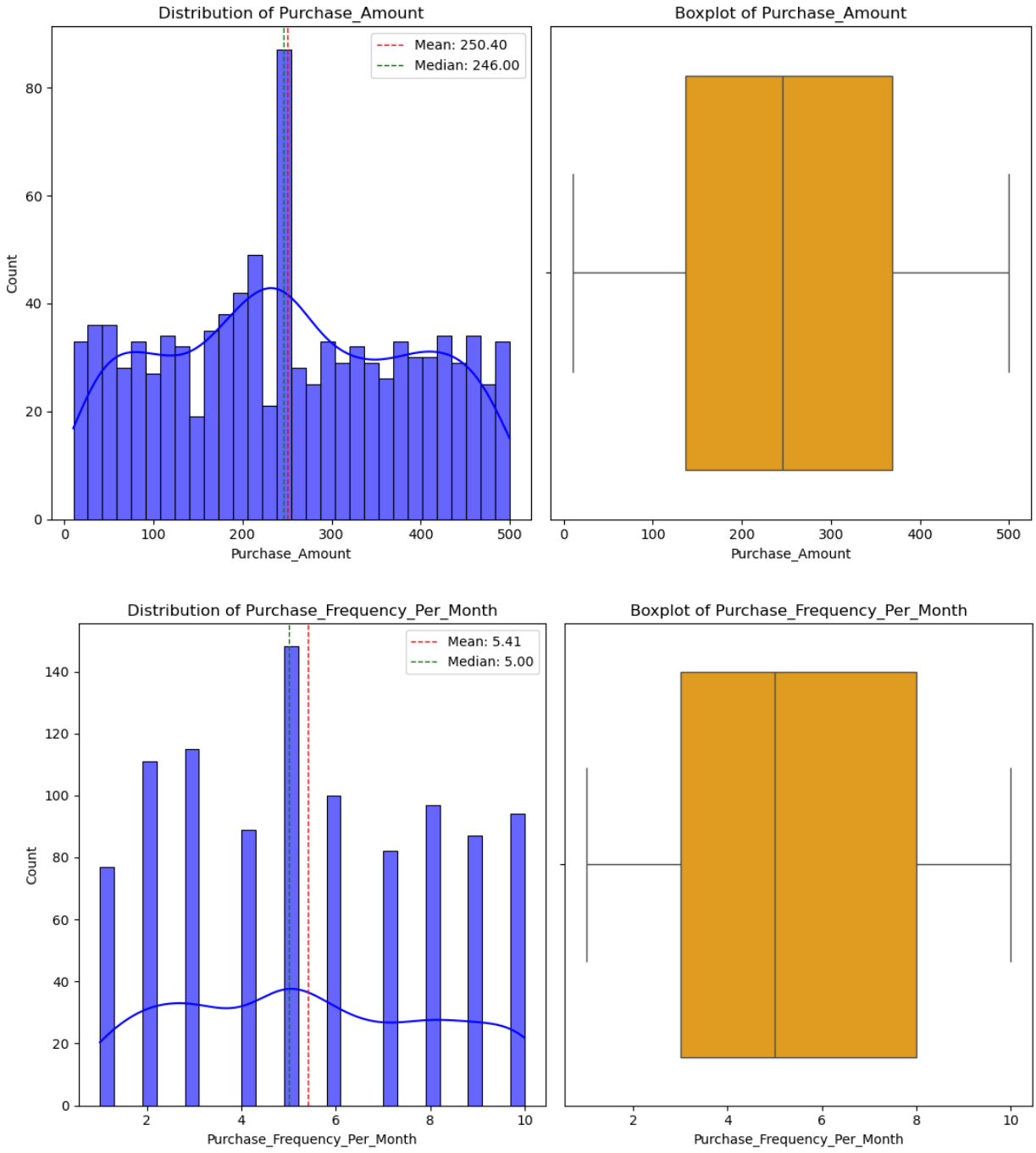
Module 2: Exploratory Data Analysis (EDA)

1. Univariate Analysis:

Univariate analysis shows that that age, purchase amount, purchase frequency per month, follow symmetric distributions. There are no outliers, seen via boxplots.

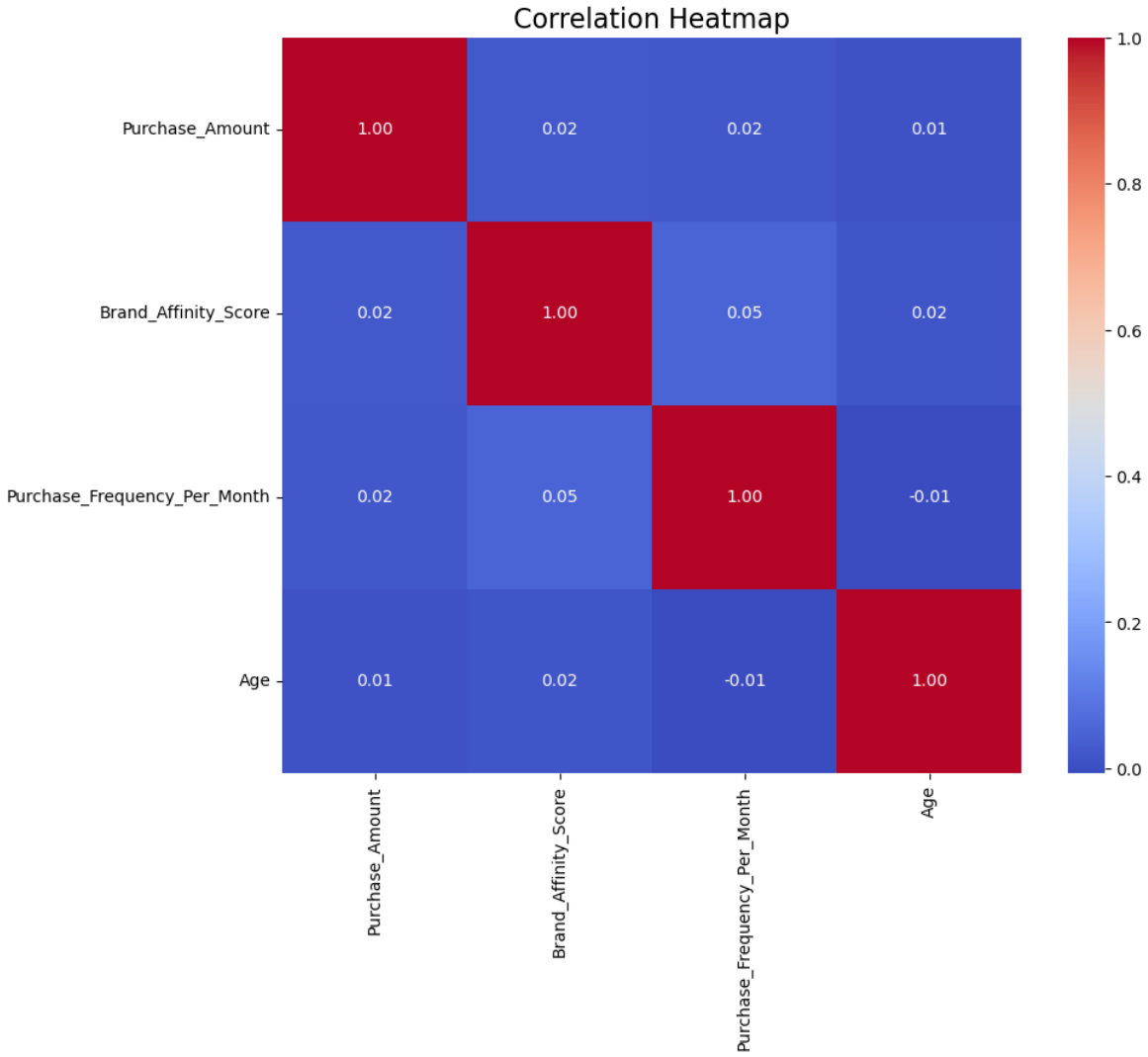
Observing the skew showed that there exists no significant skew in the mentioned columns.



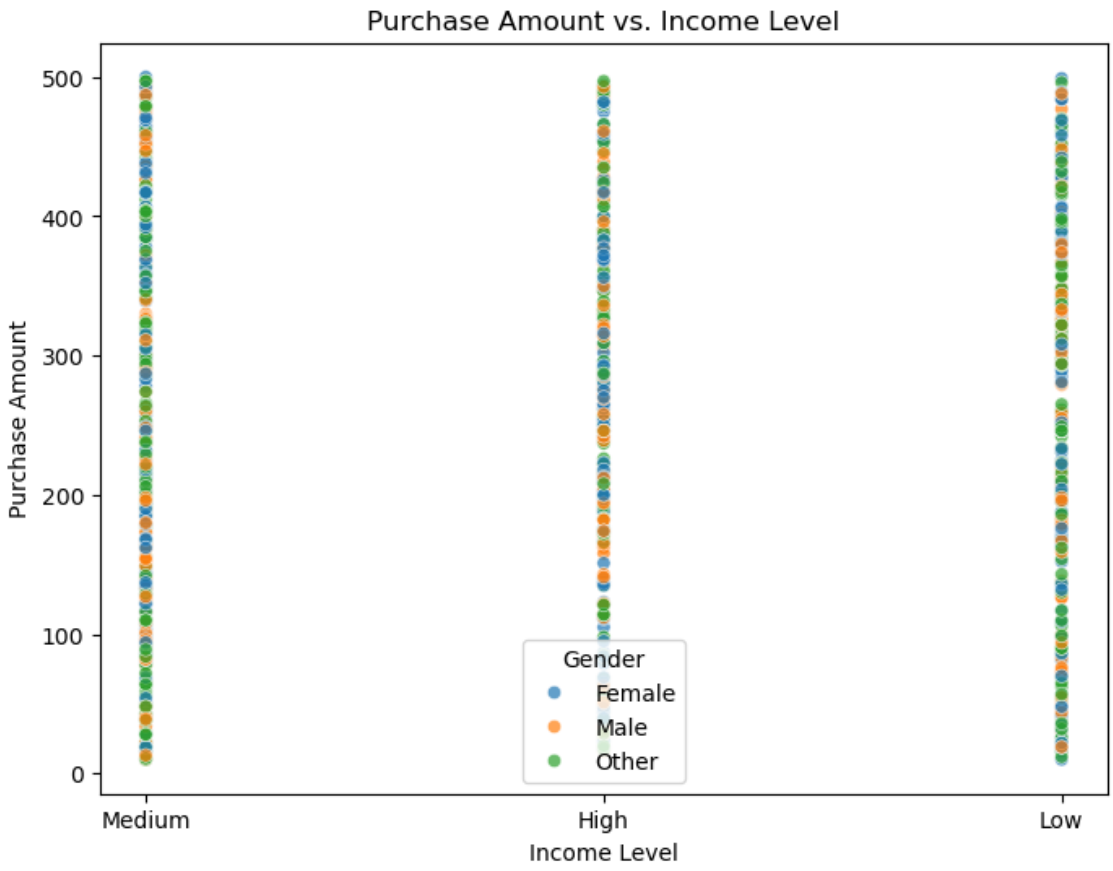


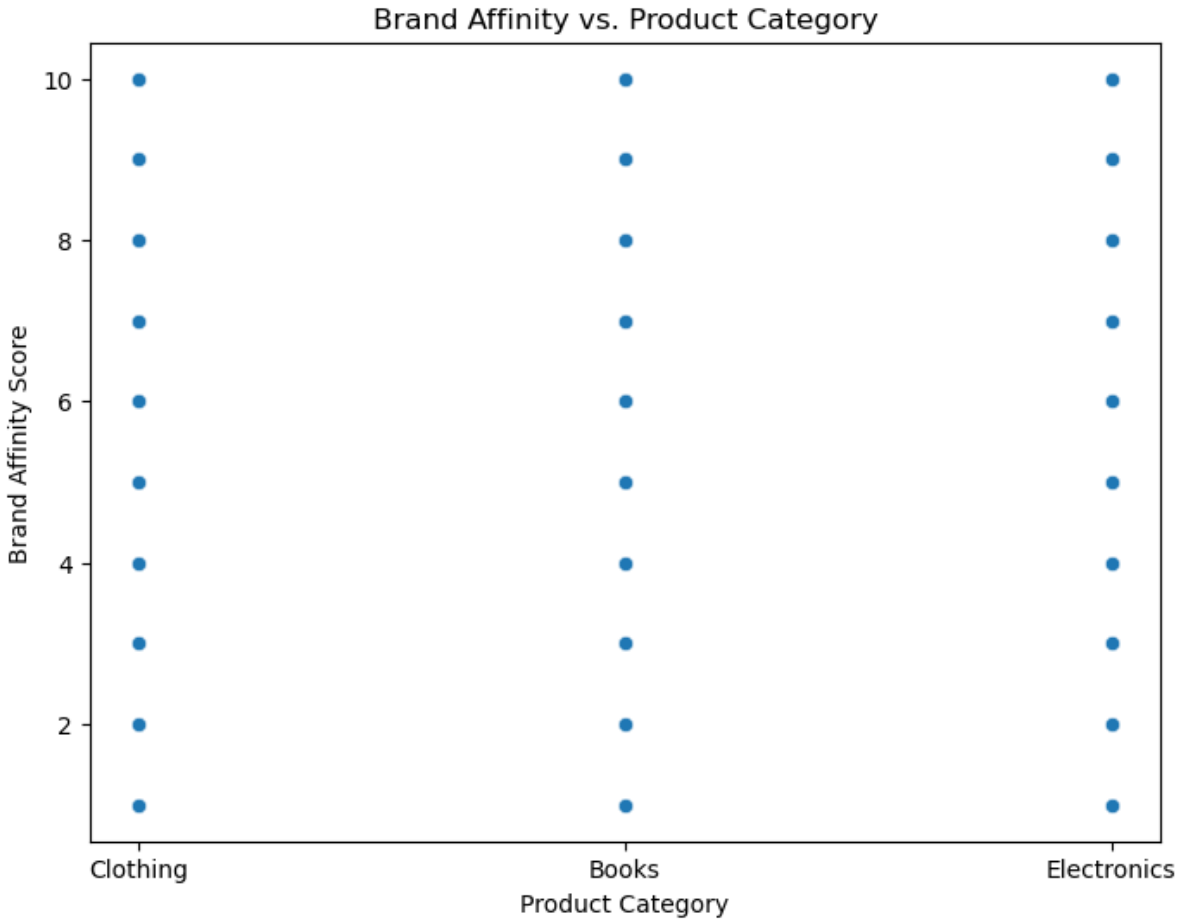
2. Bivariate Analysis:

Bivariate analysis shows, through heatmaps, that there exists no significant correlation between the columns purchase amount, brand affinity score, purchase frequency per month, and age.



The scatter plots do not reveal meaningful information due to the nature of the data, but through close inspection it is observed that the data is spread randomly, further suggesting a lack of correlation between the mentioned columns.


Purchase Amount vs. Income Level
Purchase Amount
500
400
300
200
100
0
Medium
High
Low
Income Level
Gender
Female
Male
Other

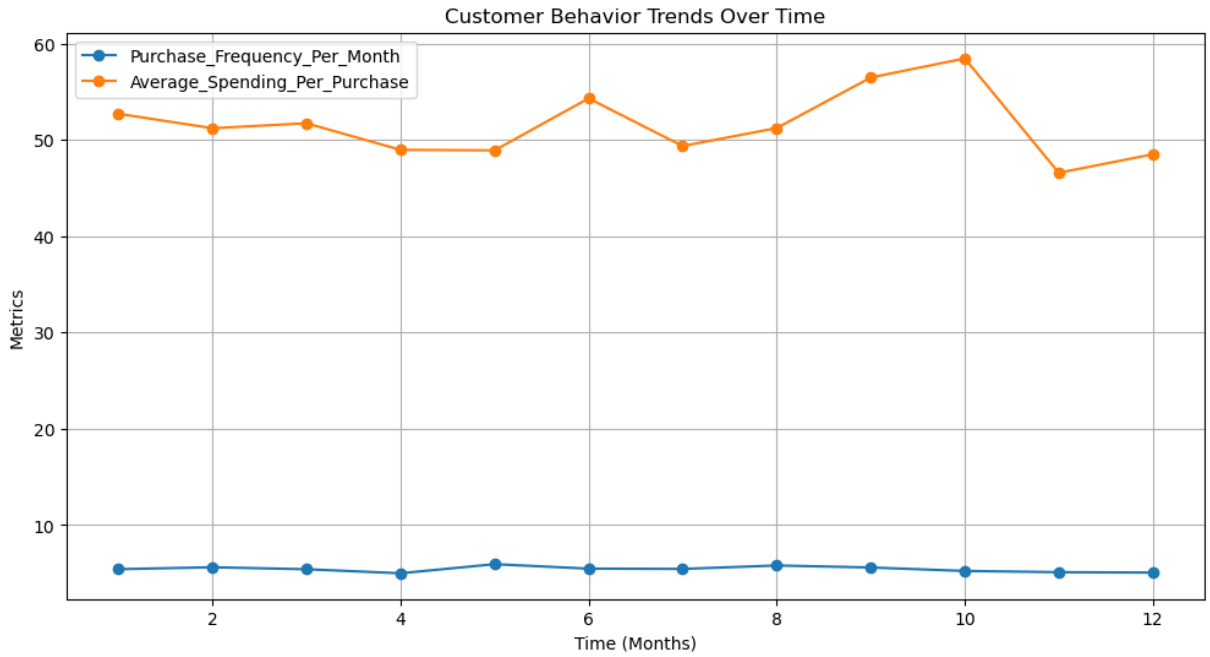

Brand Affinity vs. Product Category
Brand Affinity Score
10
8
6
4
2
Clothing
Books
Electronics
Product Category

3. Temporal Analysis:

Temporal analysis by plotting average spending per purchase and purchase frequency per month shows that there is a high in October, during Fall, for average spending per purchase by the customers, and the least values exist towards the last two months of the year.

For purchase frequency per month, there exist no significant things to note, it is flat throughout the year.

This temporal analysis is performed throughout the history provided in the data, and the is data grouped into months.



Module 3: Regression and Decision Tree Analysis

A. Linear Regression Analysis:

1. Problem Definition:

The goal of the linear regression analysis is to predict the average spending per purchase based on customer demographics and purchase history.

2. Model Building:

The relevant numerical/categorical variables were chosen as age, genders, income levels, purchase amounts, purchase frequency per month, and product category preferences.

The target variable was average spending per purchase.

3. Implementation:

The regression model was trained appropriately, using a 70-30 train-test split:

```
model = LinearRegression().fit(X_train, y_train)
```

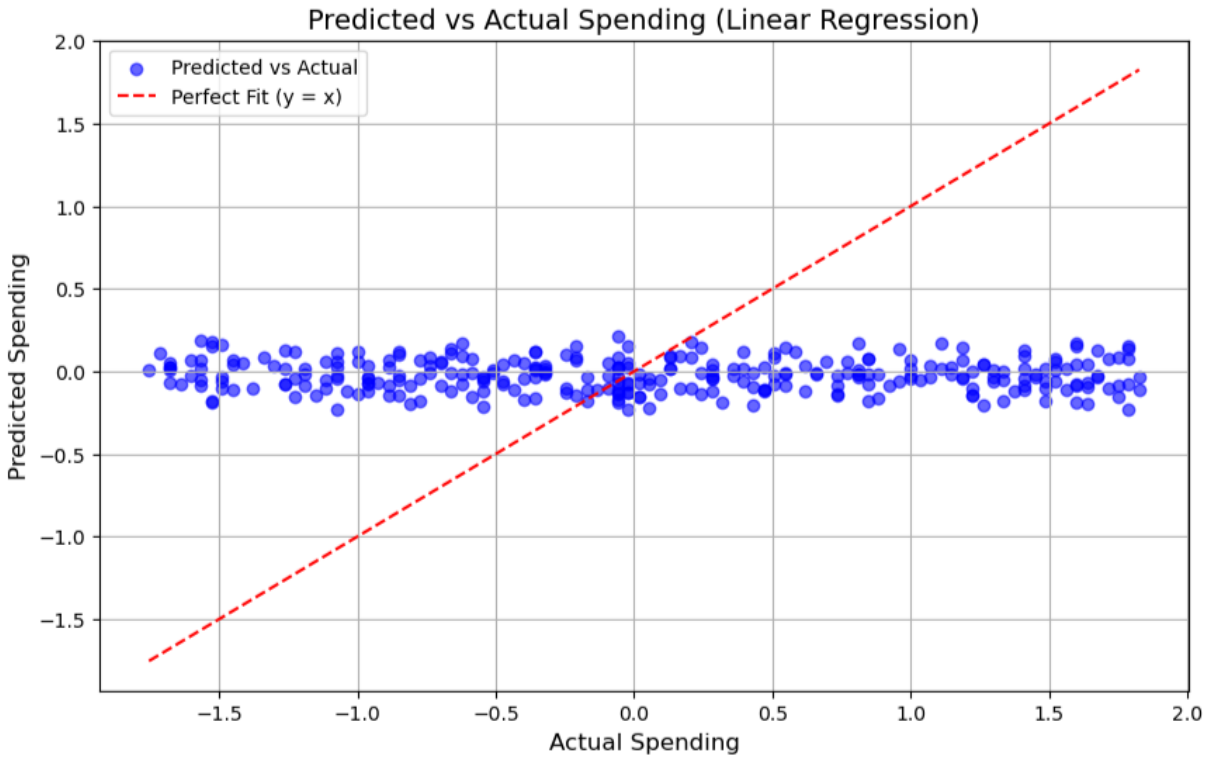
Evaluation of the regression model gave us values for R^2 and Mean Absolute Error and Mean Squared Error.

R^2 = -0.0249

MAE = 0.8629

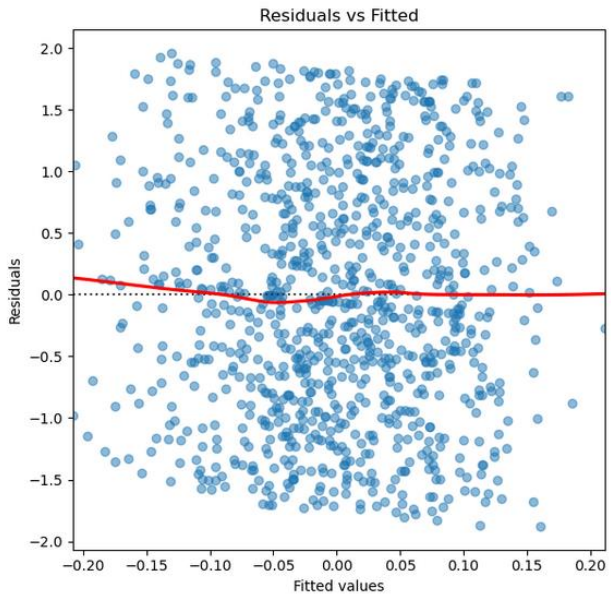
MSE = 1.0447

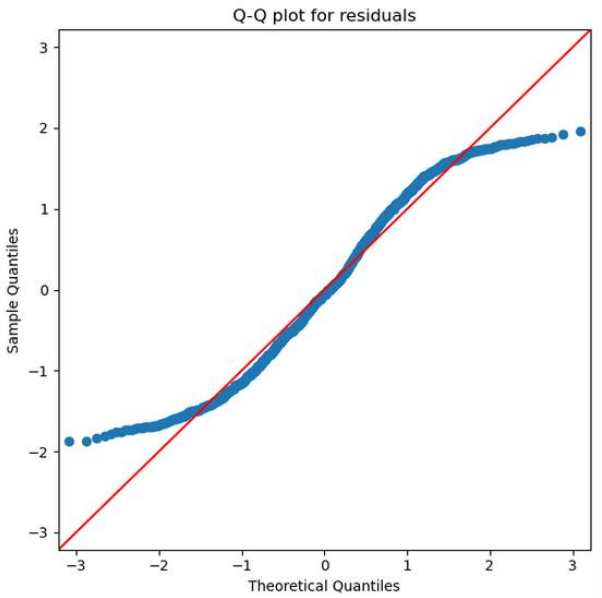
4. Visualization:

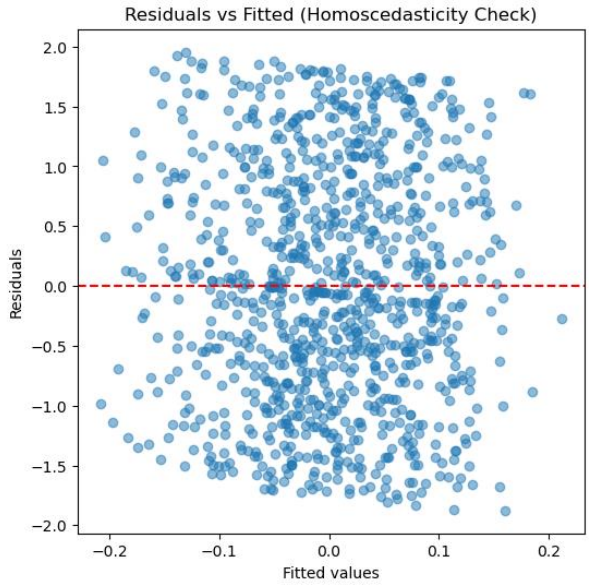


5. Interpretation

Due to the nature of the data, linear regression is not a good choice to predict the selected target variable. From the calculated summary statistics of the model the model performs very poorly, even when the data used follows normal distributions, and assumptions for a MLRM are satisfied.


Residuals vs Fitted
Residuals
Fitted values


Q-Q plot for residuals
Sample Quantiles
Theoretical Quantiles


Residuals vs Fitted (Homoscedasticity Check)
Residuals
Fitted values

```
Variance Inflation Factor (VIF) for each feature:
                          feature       VIF
0                           const  1.000000
1                             Age  1.008227
2                   Gender_Female  1.259987
3                     Gender_Male  1.263930
4                Income_Level_High  1.394273
5              Income_Level_Medium  1.393352
6                  Purchase_Amount  1.006993
7    Purchase_Frequency_Per_Month  1.005668
8  Product_Category_Preferences_High  1.261525
9  Product_Category_Preferences_Medium  1.268554
                            OLS Regression Results
==============================================================================
Dep. Variable:     Average_Spending_Per_Purchase   R-squared:                       0.006
Model:                                     OLS   Adj. R-squared:                 -0.003
Method:                          Least Squares   F-statistic:                     0.6517
Date:                         Thu, 05 Dec 2024   Prob (F-statistic):              0.753
Time:                                 23:20:21   Log-Likelihood:                -1416.0
No. Observations:                         1000   AIC:                             2852.
Df Residuals:                              990   BIC:                             2901.
Df Model:                                    9
Covariance Type:                     nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                           -4.51e-17      0.032  -1.42e-15      1.000      -0.062       0.062
Age                               -0.0485      0.032     -1.523      0.128      -0.111       0.014
Gender_Female                      0.0058      0.036      0.164      0.869      -0.064       0.076
Gender_Male                       -0.0330      0.036     -0.926      0.355      -0.103       0.037
Income_Level_High                  0.0001      0.037      0.004      0.997      -0.073       0.074
Income_Level_Medium               -0.0264      0.037     -0.706      0.481      -0.100       0.047
Purchase_Amount                   -0.0171      0.032     -0.537      0.591      -0.079       0.045
Purchase_Frequency_Per_Month      -0.0245      0.032     -0.771      0.441      -0.087       0.038
Product_Category_Preferences_High -0.0225      0.036     -0.632      0.528      -0.092       0.047
Product_Category_Preferences_Medium -0.0004    0.036     -0.010      0.992      -0.070       0.070
==============================================================================
Omnibus:                      290.232   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.324
Skew:                           0.105   Prob(JB):                     3.21e-11
Kurtosis:                       1.944   Cond. No.                         1.83
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

B. Decision Tree Analysis:

1. Problem Definition:

We need to predict if a customer will make a purchase in the next month using the binary target variable “Will purchase next month”.

2. Model Building:

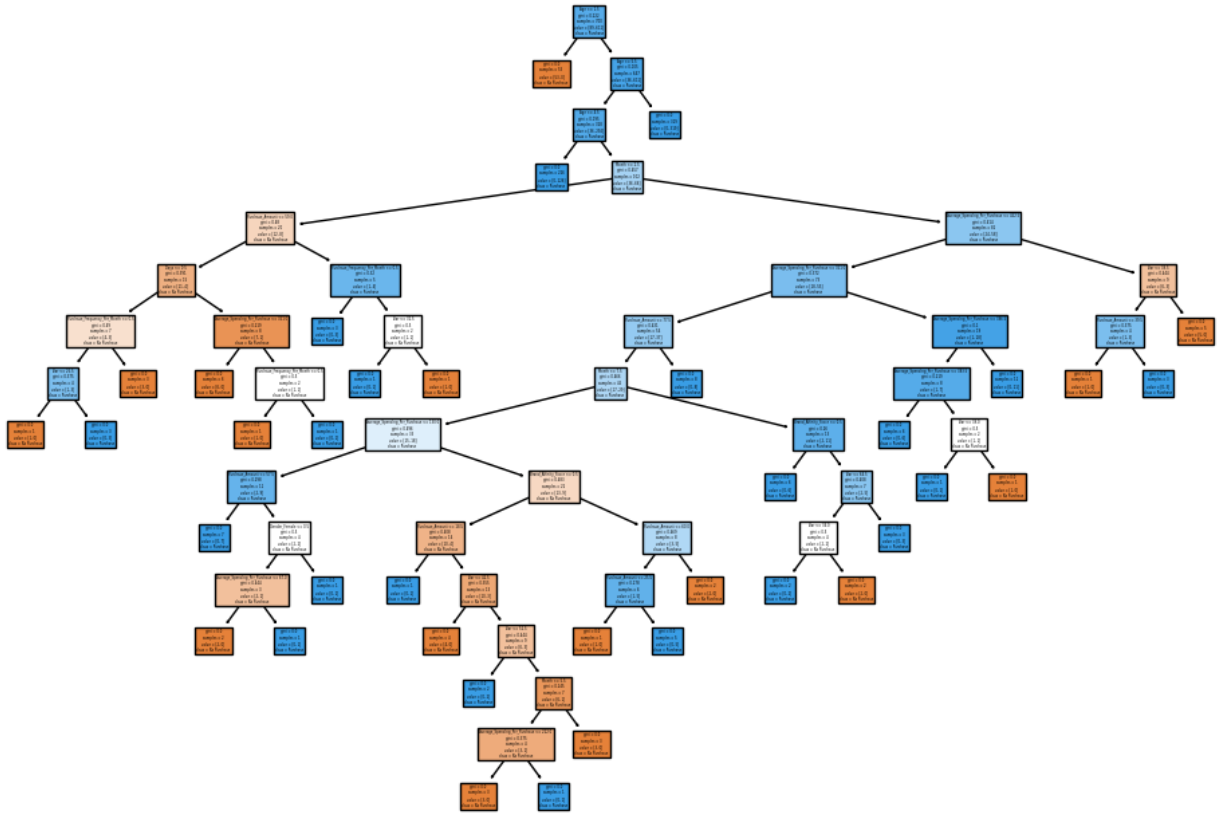
Predictors were chosen as purchase frequency per month, average spending per purchase, purchase amount, product category references, brand affinity score, age, season.

3. Implementation:

The decision tree was classified appropriately using Gini Impurity:

```
clf = DecisionTreeClassifier(criterion='gini', random_state=18)
clf.fit(X_train, y_train)
```

4. Visualization:



5. Interpretation

The model is highly effective at identifying buyers, although it struggles to predict non-buyers accurately.

Module 4: Clustering Analysis

1. Define the number of clusters(k):

The number of classes were decided by using elbow method, and then choosing the number based on the sum of squared distances within each other.

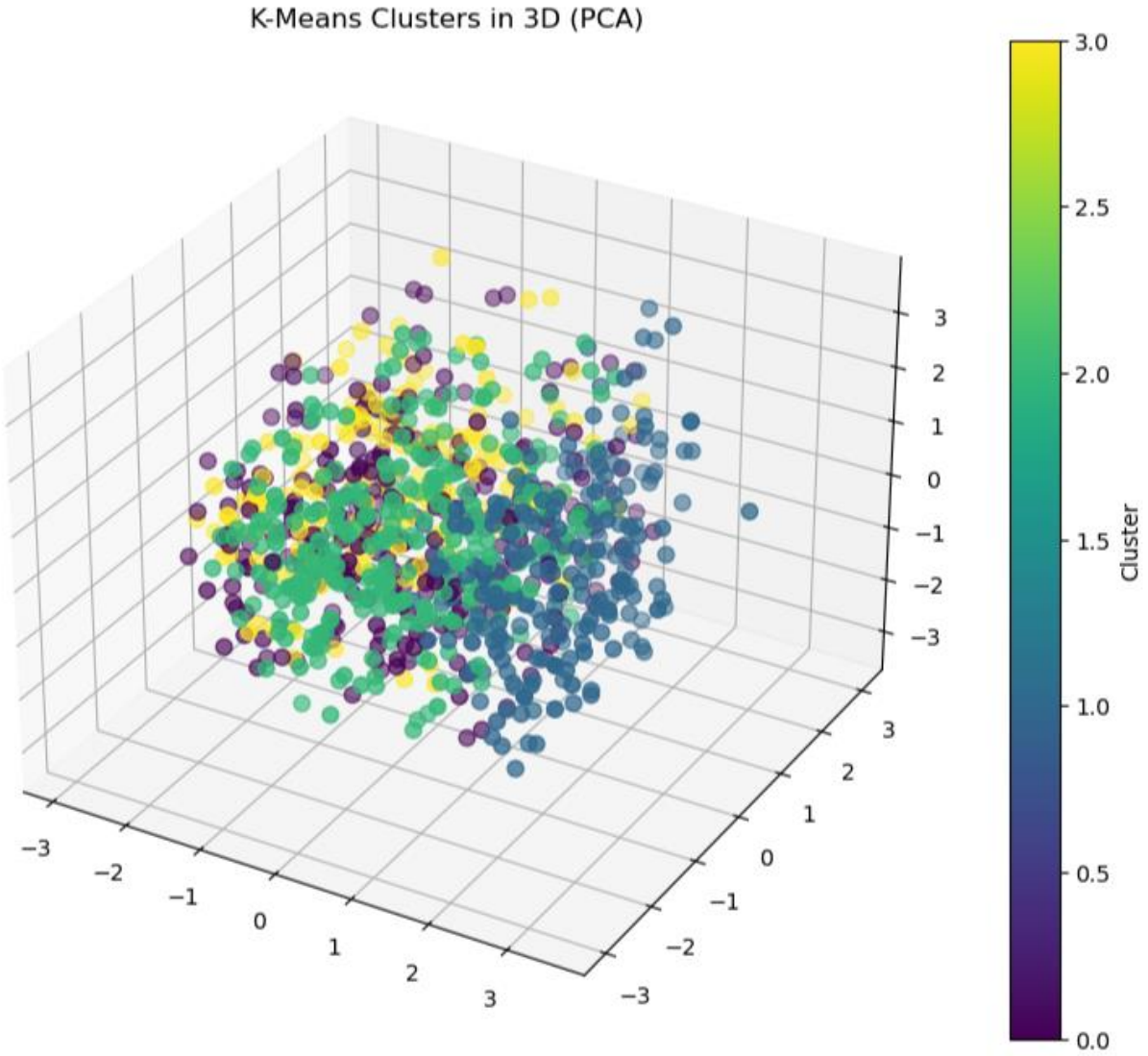
2. Apply K-Means Clustering:

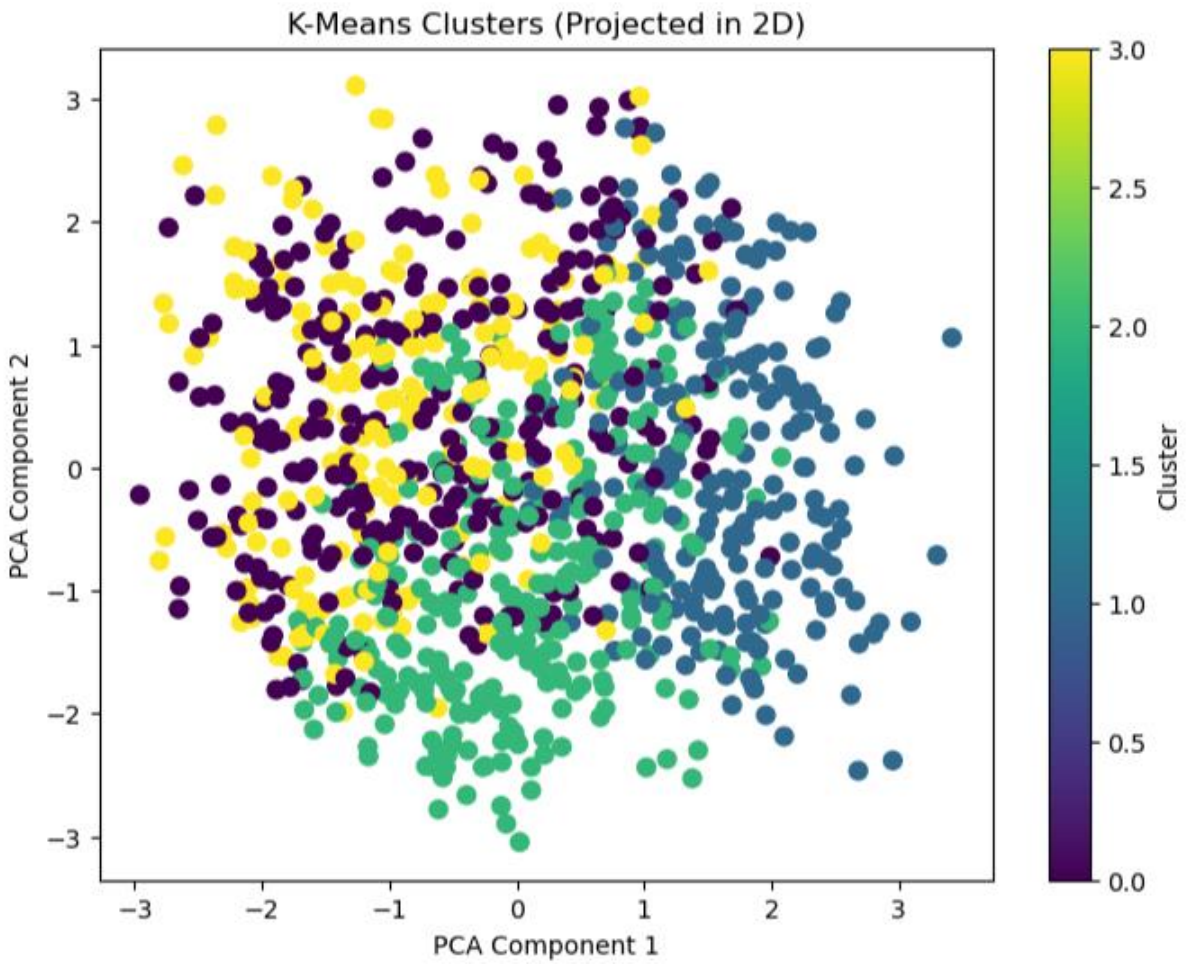
K-means was applied by:

```
kmeans = KMeans(n_clusters=optimal_k, random_state=18)
clusters = kmeans.fit_predict(X)
```

3. Analyze cluster characteristics:

Cluster analysis reveals:





PCA components were used to reduce dimensionality of the data for visualization.

Significant Differences:

- **Age**: Cluster 3 is notably distinct from others with the largest differences (e.g., Cluster 2 vs. 3 = 10.2).
- **Purchase Amount**: Largest differences are observed between Cluster 2 and 3 (31.5) and Cluster 0 and 3 (24.9).
- **Average Spending Per Purchase**: Cluster 2 and 3 differ significantly (14.1), showing distinct spending behaviors.
- **Days**: High differences, especially between Cluster 1 and 3 (141.9).
- **Gender**: Female/Male composition varies significantly between most clusters (e.g., Female: 0.6 for Cluster 2 vs. 3).

Similarities:

- **Will Purchase Next Month**: No significant difference across clusters.

- **Income Levels (High)**: No significant differences.
- **Seasonal Preferences**: Similar for most features except minor differences in Spring and Fall.

Key Insights:

- Clusters differ most in spending habits, demographics (e.g., gender), and brand affinity.
- Behavioral consistency is observed for future purchase likelihood and income level preferences.

Module 5: Comparison and Conclusion

1. Comparison:

1. Regression

R^2 Score: -0.0249

A negative R^2 score indicates that the model performs poorly, worse than simply predicting the mean value.

Mean Absolute Error (MAE): 0.8629

This error shows the average magnitude of the errors between predicted and actual values.

Mean Squared Error (MSE): 1.0447

This measures the average squared differences between predicted and actual values, indicating how far off the predictions are.

2. Decision Tree

Accuracy: 92.67%

The accuracy is high, meaning that the decision tree model makes correct predictions on the test set about 92.67% of the time.

Precision: 96.50%

Precision indicates that when the model predicted a positive outcome, it was correct 96.50% of the time.

Recall: 95.02%

Recall shows that the model correctly identified 95.02% of the actual positive outcomes.

F1 Score: 95.75%

The F1 score balances precision and recall, showing that the model performs very well in identifying positive outcomes.

3. K-Means Clustering

Cluster Analysis: The significant differences between clusters, such as the features "Age," "Purchase Amount," and "Days," highlight patterns in the data, but K-Means is an unsupervised learning algorithm that does not predict a target variable directly. The differences between cluster centroids indicate that the clusters exhibit significant variability based on different features, especially "Days" and "Brand Affinity."

Key Insights:

1. Regression:

The regression model shows a poor fit, with a negative R^2 score and high errors. It may not be capturing the underlying patterns effectively.

2. Decision Tree:

The decision tree performs excellently, with high accuracy, precision, recall, and F1 score, indicating it is well-suited for classification tasks.

3. K-Means Clustering:

While it does not directly predict a target variable, the significant differences between clusters could reveal insights into the distribution of data and the presence of inherent patterns, but it's more of a data exploration tool rather than a predictive model.

2. Actionable Recommendations:

1. High-Impact Features:

Purchase Frequency & Spending: Customers with different purchase frequencies and higher spending patterns (e.g., Cluster 0 vs. Cluster 3) should be targeted with tailored offers and promotions.

Brand Affinity: Use brand loyalty data to create personalized experiences or loyalty programs.

2. Marketing Refinements:

Age & Purchase Amount: Tailor offers based on both customer age and spending habits (e.g., higher-value electronics for those with higher spending).

Product Preferences: Introduce category-specific offers based on cluster preferences (e.g., electronics promotions for clusters preferring electronics).

3. Cluster-Specific Recommendations:

Provide personalized electronics recommendations based on cluster preferences (e.g., high-end electronics for certain clusters).

4. Seasonal Promotions:

Adjust promotions based on seasonality (e.g., discounts or bundles for clusters engaged during Fall or Summer).

5. Customer Retention:

Create gender-specific marketing campaigns to retain customers and increase engagement.

6. Offer Optimization:

Optimize offers for customers with lower purchase frequency but higher spending, such as offering high-ticket items or personalized bundles.

7. Purchase Intent:

Since future purchase intent doesn't vary significantly across clusters, focus on boosting customer engagement to drive repeat purchases.