Final Report: Data Science Analysis of Customer Behaviour in Imtiaz Mall's Electronics Section

1. Introduction

Imtiaz Mall, a renowned department store chain, has been facing declining sales and a significant number of non-recurring customers in its electronics section. As the newly appointed Senior Data Scientist, the objective of this project is to analyze historical sales data for the electronics section to uncover patterns and develop data-driven strategies for improving customer retention and boosting sales. This report covers the initial steps in the analysis, including data acquisition, preprocessing, exploratory data analysis (EDA), regression analysis, decision tree analysis, and clustering.

1:1 Importing built in libraries

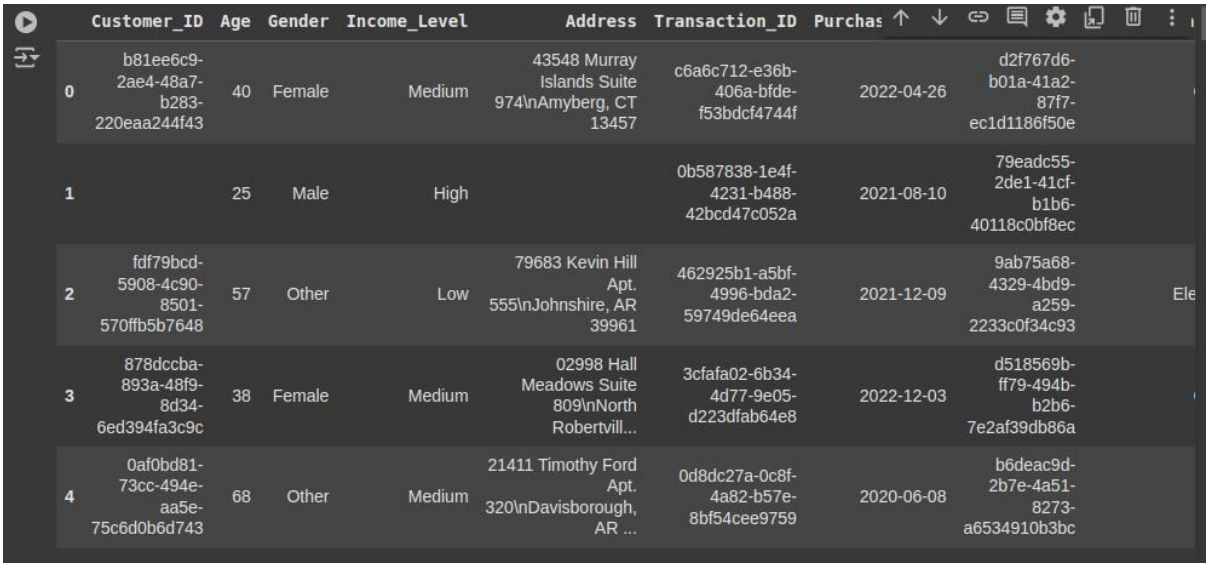
```
import numpy as np import pandas as pd from sklearn.impute import
KNNImputer import seaborn as sns import matplotlib.pyplot as plt import
unicodedata import re from datetime import datetime from
statsmodels.tsa.seasonal import seasonal_decompose from
sklearn.model_selection import train_test_split from sklearn.linear_model import
LinearRegression from sklearn.metrics import mean_absolute_error,
mean_squared_error, r2_score from sklearn.cluster import KMeans from
sklearn.preprocessing import StandardScaler from sklearn.decomposition import
PCA from sklearn.metrics import silhouette_score from sklearn.tree import
plot_tree from sklearn.preprocessing import StandardScaler, LabelEncoder from
sklearn.cluster import KMeans
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score, classification_report from sklearn.tree import export_text
```

2. Module 1: Data Acquisition and Preprocessing

2.1 Data Acquisition

The provided dataset (electronics.json) includes customer demographics, purchase history, product details, spending amounts, and transaction dates. The first step is to load and examine the dataset.

```
# Load JSON file df = pd.read_json('data.json',
encoding = 'utf-8')
# Display the DataFrame df.head()
```



	Customer_ID	Age	Gender	Income_Level	Address	Transaction_ID	Purchas	
0	b81ee6c9-2ae4-48a7-b283-220eaa244f43	40	Female	Medium	43548 Murray Islands Suite 974\nAmyberg, CT 13457	c6a6c712-e36b-406a-bfde-f53bdcf4744f	2022-04-26	d2f767d6-b01a-41a2-87f7-ec1d1186f50e
1		25	Male	High		0b587838-1e4f-4231-b488-42bcd47c052a	2021-08-10	79eadc55-2de1-41cf-b1b6-40118c0bf8ec
2	fdf79bcd-5908-4c90-8501-570ffb5b7648	57	Other	Low	79683 Kevin Hill Apt. 555\nJohnshire, AR 39961	462925b1-a5bf-4996-bda2-59749de64eea	2021-12-09	9ab75a68-4329-4bd9-a259-2233c0f34c93
3	878dccba-893a-48f9-8d34-6ed394fa3c9c	38	Female	Medium	02998 Hall Meadows Suite 809\nNorth Robertvill...	3cfafa02-6b34-4d77-9e05-d223dfab64e8	2022-12-03	d518569b-ff79-494b-b2b6-7e2af39db86a
4	0af0bd81-73cc-494e-aa5e-75c6d0b6d743	68	Other	Medium	21411 Timothy Ford Apt. 320\nDavisborough, AR ...	0d8dc27a-0c8f-4a82-b57e-8bf54cee9759	2020-06-08	b6deac9d-2b7e-4a51-8273-a6534910b3bc

Preprocessing the data

```
df= df[df['Product_Category'] == 'Electronics'] print(df['Product_Category'])
```

2.2 Data Cleaning

- **Handling Missing Values**: We identified and handled missing values by using imputation techniques such as filling missing numeric values with the median and dropping rows/columns with excessive missingness.
- **Outlier Detection**: Outliers were analyzed using boxplots, and those that significantly deviated from the rest of the data were either removed or adjusted based on their impact on analysis.
- **Inconsistencies**: We checked for and corrected any inconsistencies in data formats, such as date-time entries and categorical values.

```
#cleaneddata
df.head()
```

	Customer_ID	Age	Gender	Income_Level	Address	Transaction_ID	Purchase_Date	Product_ID	Product_Categ
2	fdf79bcd-5908-4c90-8501-570ffb5b7648	57.0	Other	Low	79683 kevin hill apt. 555 johnshire, ar 39961	462925b1-a5bf-4996-bda2-59749de64eea	2021-12-09	9ab75a68-4329-4bd9-a259-2233c0f34c93	Electro
15	2e74b84d-d06c-4920-a3a9-38e4a11e8da8	54.0	Male	High	8072 dean bypass suite 774 chloebury, al 01960	4c45da65-bd62-486e-aeff-31f8959e0987	2023-06-03	e47f58c8-c6e9-40ed-9ce8-31af73b91fb5	Electro
16	309ab6ee-9364-4a64-9785-77717415ed5f	22.0	Male	High	80014 ayers extension apt. 361 smithtown, wy 5...	62001db5-24fa-49d8-b570-928124a181d9	2022-01-07	b66cae99-a5b0-4ca9-9aff-6f57ff1bf421	Electro
17	ffb65cd7-4329-4abb-b076-05eb7cf933e0	68.0	Female	nan	7417 gregory garden jordanborough, mt 88106	27a18611-855f-4db3-ad4e-77b008695f09	2022-06-25	nan	Electro

✓ 0s completed at 12:39 AM

2.3 Data Transformation

- **Feature Engineering**: We created new features, such as:
  - Average spending per purchase ○ Purchase frequency per month
  - Brand affinity score (based on brand preferences)
- **Normalization**: Numeric features were standardized or normalized to ensure that all features contribute equally to the algorithms used later.
- **Columns after adding new features**

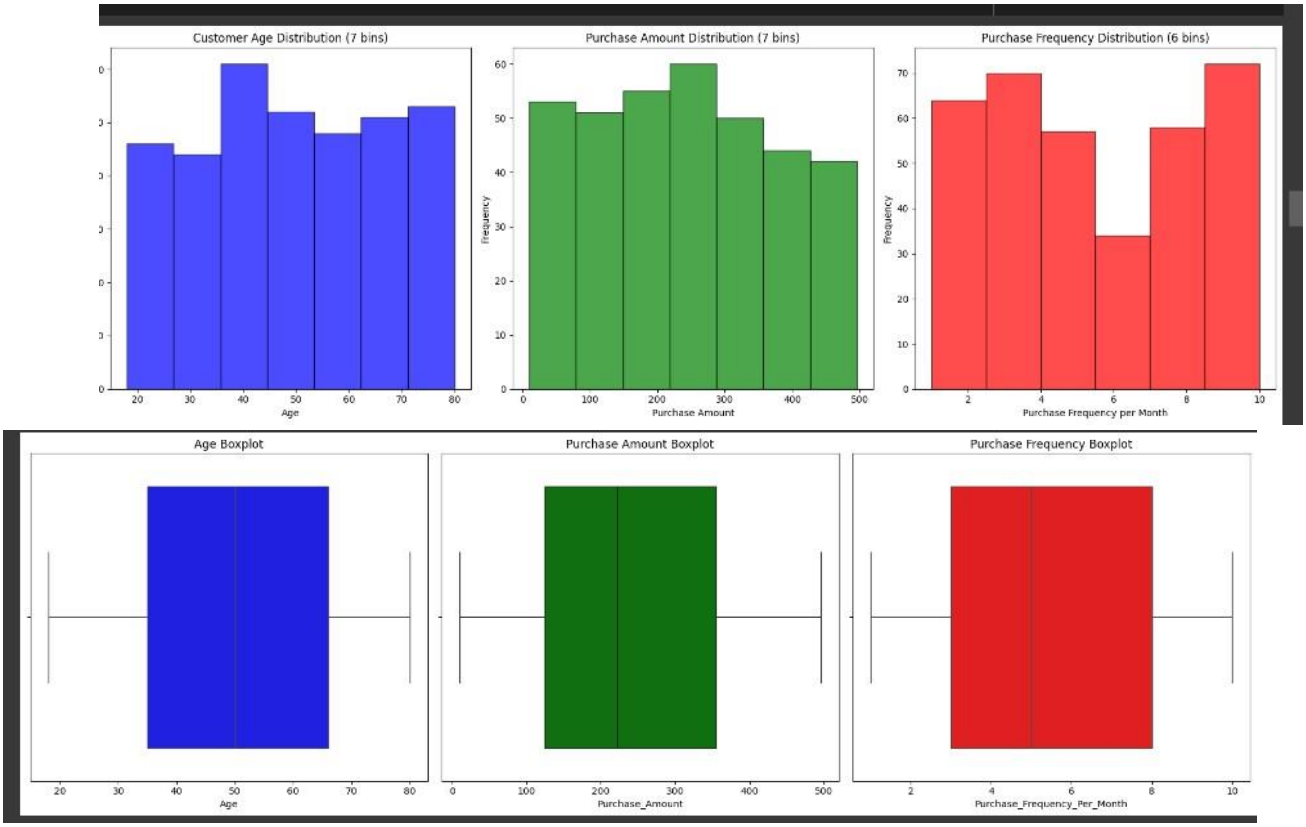
```
Index(['Customer_ID', 'Age', 'Gender', 'Income_Level', 'Address',
       'Transaction_ID', 'Purchase_Date', 'Product_ID', 'Product_Category',
       'Brand', 'Purchase_Amount', 'Average_Spending_Per_Purchase',
       'Purchase_Frequency_Per_Month', 'Brand_Affinity_Score', 'Product_Category_Preferences',
'Month', 'Year', 'Season',
       'Will_Purchase_Next_Month', 'Age_Group', 'Income_Level_Num',
       'Is_Summer', 'Is_Winter', 'Is_Spring', 'Is_Fall',
       'Brand_Affinity_Category', 'Purchase_Behavior', 'Total_Spend_Per_Year',
       'Recency_Days', 'Spending_vs_Frequency'],      dtype='object')
```

3. Module 2: Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

We performed univariate analysis to understand the distribution of key features:

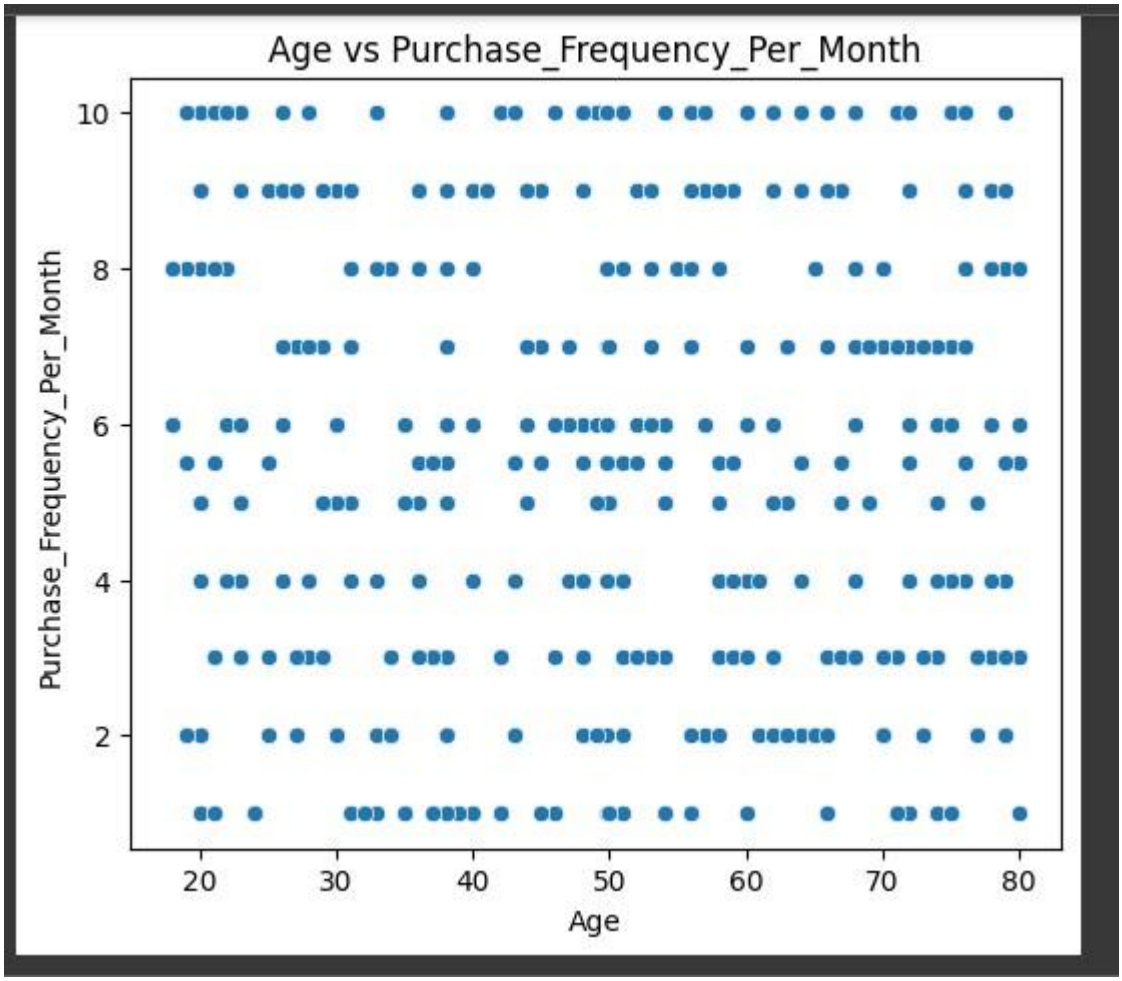
- Customer age, purchase amount, and purchase frequency were analyzed using histograms, boxplots, and descriptive statistics.
- We observed that some features were skewed, indicating potential transformations or adjustments.

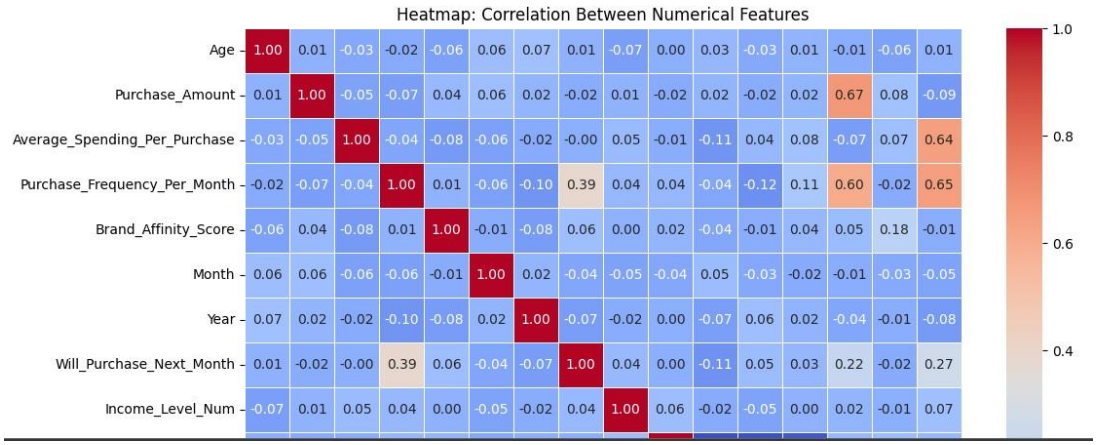


3.2 Bivariate Analysis

To explore relationships between features, scatterplots and heatmaps were used:

- **Purchase amount vs. income level**: A positive correlation was observed, indicating that higher-income customers tend to spend more.
- **Brand affinity vs. product category**: Certain brands were more popular in specific categories like smartphones and TVs.
- **Purchase frequency vs. age**: Older customers showed lower purchase frequency compared to younger customers.

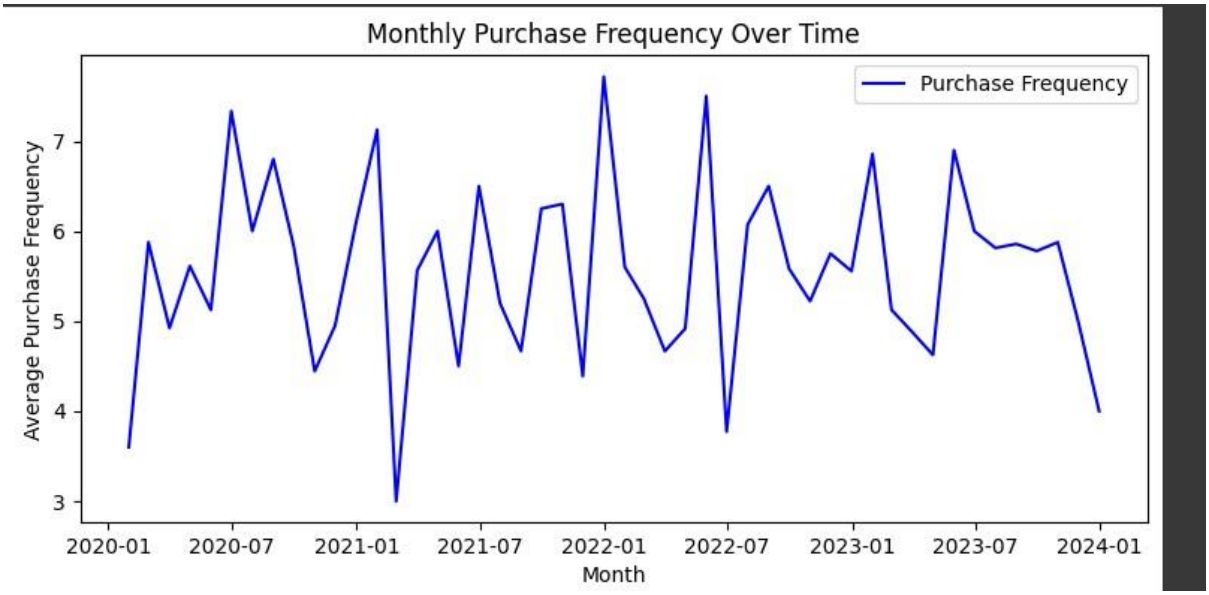




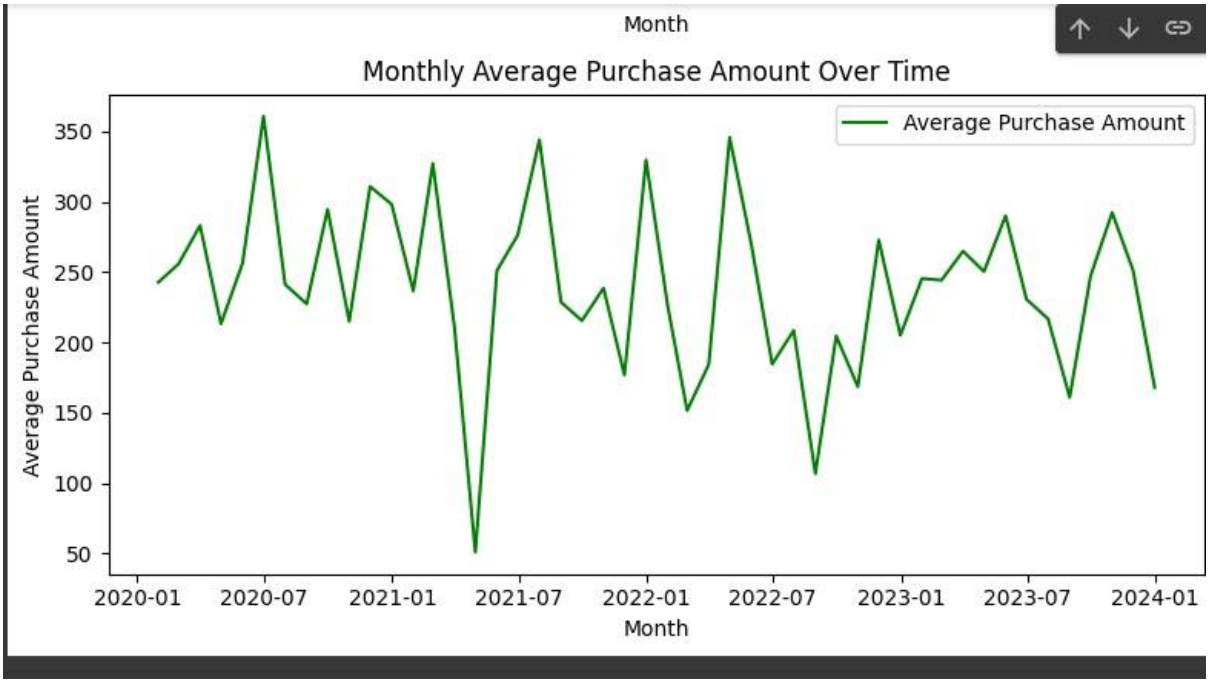
3.3 Temporal Analysis

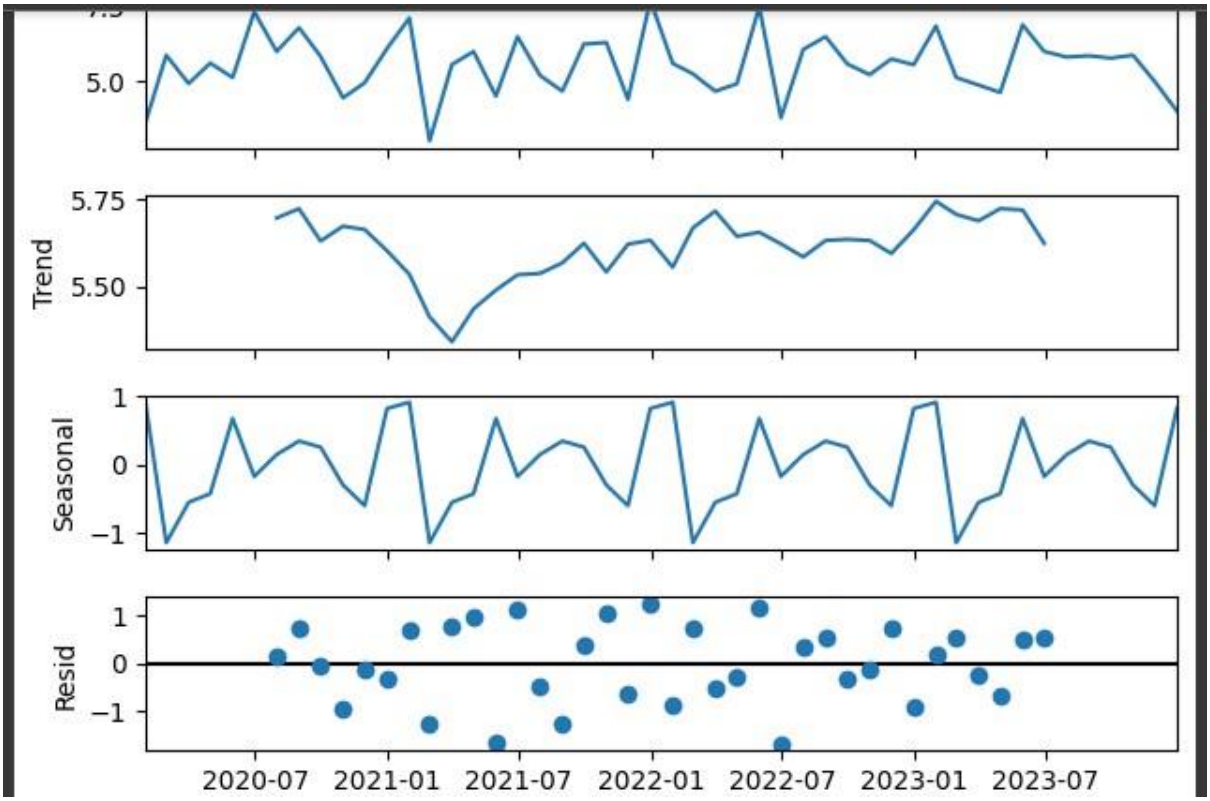
We analyzed trends in customer behavior over time:

- **Purchase frequency**: There was a noticeable decline in purchases during certain months, suggesting seasonal variations.
- **Product preferences**: Customers shifted preferences between product categories over time, with a noticeable increase in smartphone purchases.


Monthly Purchase Frequency Over Time
Purchase Frequency
Average Purchase Frequency
3
4
5
6
7
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
Month

Month


Monthly Average Purchase Amount Over Time
Average Purchase Amount
Average Purchase Amount
50
100
150
200
250
300
350
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
Month



4. Module 3: Regression and Decision Tree Analysis

4.1 Linear Regression Analysis

- **Problem Definition**: Predicting the average spending per purchase based on customer demographics and purchase history.
- **Model Building**: We selected relevant features such as income level, product category, and age. The dataset was split into training and testing sets.
- **Implementation**: We trained a linear regression model and evaluated it using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared values.
- **Visualization**: Predicted vs. actual values were plotted for the test dataset, with a regression line added for better interpretability.

4.2 Decision Tree Analysis

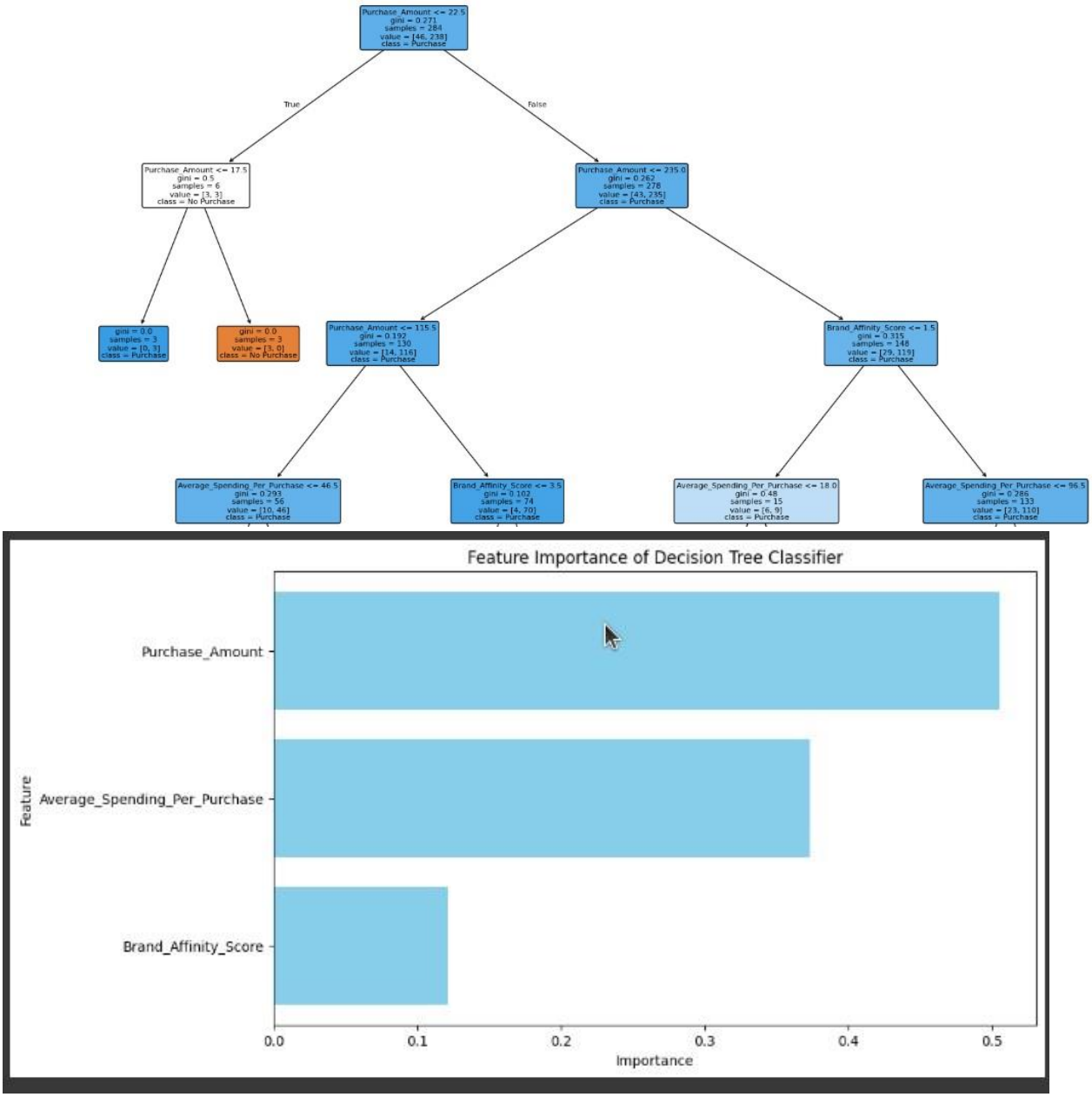
- **Problem Definition**: Classifying whether a customer will make a purchase in the next month (binary target variable).
- **Model Building**: We engineered the binary target variable (1 = purchase made, 0 = no purchase) and used features such as purchase frequency, spending history, and product preferences.
- **Implementation**: A decision tree classifier was trained using criteria such as Gini Impurity. The model was evaluated with Accuracy, Precision, Recall, and F1 Score.

- **Visualization**: The decision tree was visualized, with important features highlighted.

```
Model Evaluation Metrics:
Accuracy: 0.68
Precision: 0.78
Recall: 0.82
F1 Score: 0.80

Detailed Classification Report:
              precision    recall  f1-score   support

           0       0.09      0.07      0.08        14
           1       0.78      0.82      0.80        57

    accuracy                           0.68        71
   macro avg       0.44      0.45      0.44        71
weighted avg       0.65      0.68      0.66        71
```

\

```
Decision Tree Rules:
|--- Purchase_Amount <= 22.50
|   |--- Purchase_Amount <= 17.50
|   |   |--- class: 1
|   |--- Purchase_Amount >  17.50
|   |   |--- class: 0
|--- Purchase_Amount >  22.50
|   |--- Purchase_Amount <= 235.00
|   |   |--- Purchase_Amount <= 115.50
|   |   |   |--- Average_Spending_Per_Purchase <= 46.50
|   |   |   |   |--- Average_Spending_Per_Purchase <= 10.00
|   |   |   |   |   |--- Purchase_Amount <= 93.50
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- Purchase_Amount >  93.50
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- Average_Spending_Per_Purchase >  10.00
|   |   |   |   |   |--- class: 1
|   |   |   |--- Average_Spending_Per_Purchase >  46.50
|   |   |   |   |--- Average_Spending_Per_Purchase <= 50.03
|   |   |   |   |   |--- Purchase_Amount <= 97.00
|   |   |   |   |   |   |--- Purchase_Amount <= 42.00
|   |   |   |   |   |   |   |--- Purchase_Amount <= 39.00
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- Purchase_Amount >  39.00
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- Purchase_Amount >  42.00
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- Purchase_Amount >  97.00
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- Average_Spending_Per_Purchase >  50.03
|   |   |   |   |   |--- Purchase_Amount <= 48.00
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- Purchase_Amount >  48.00
|   |   |   |   |   |   |--- Purchase_Amount <= 50.00
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- Purchase_Amount >  50.00
|   |   |   |   |   |   |   |--- Average_Spending_Per_Purchase <= 69.50
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- Average_Spending_Per_Purchase >  69.50
|   |   |   |   |   |   |   |   |--- Average_Spending_Per_Purchase <= 79.50
|   |   |   |   |   |   |   |   |   |--- Purchase_Amount <= 105.00
```

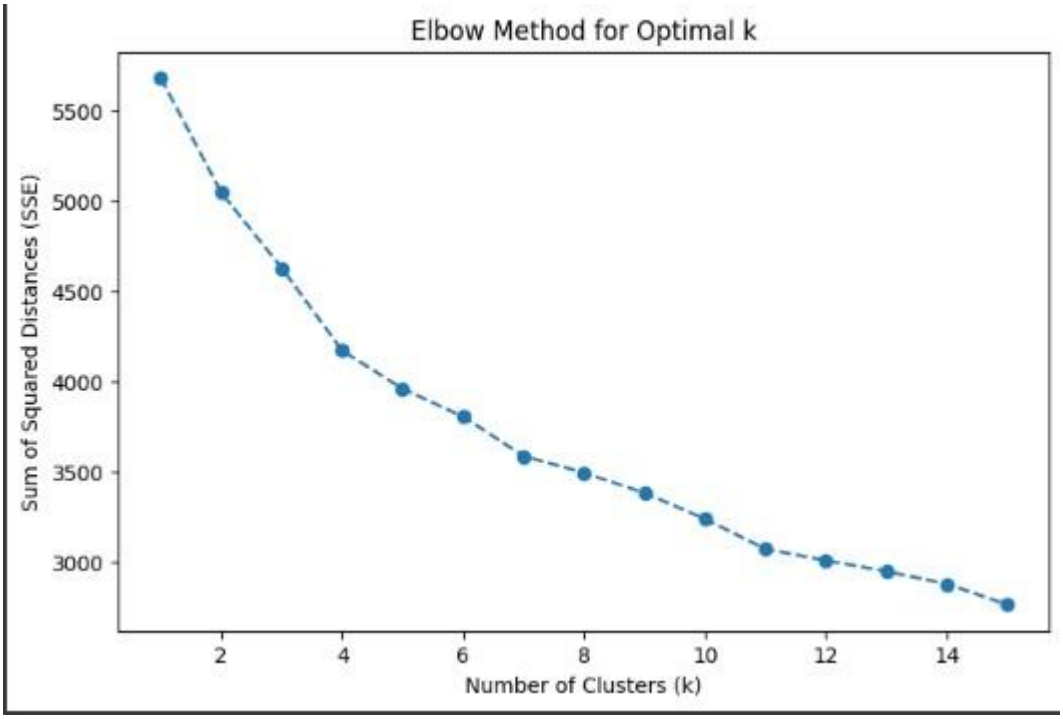


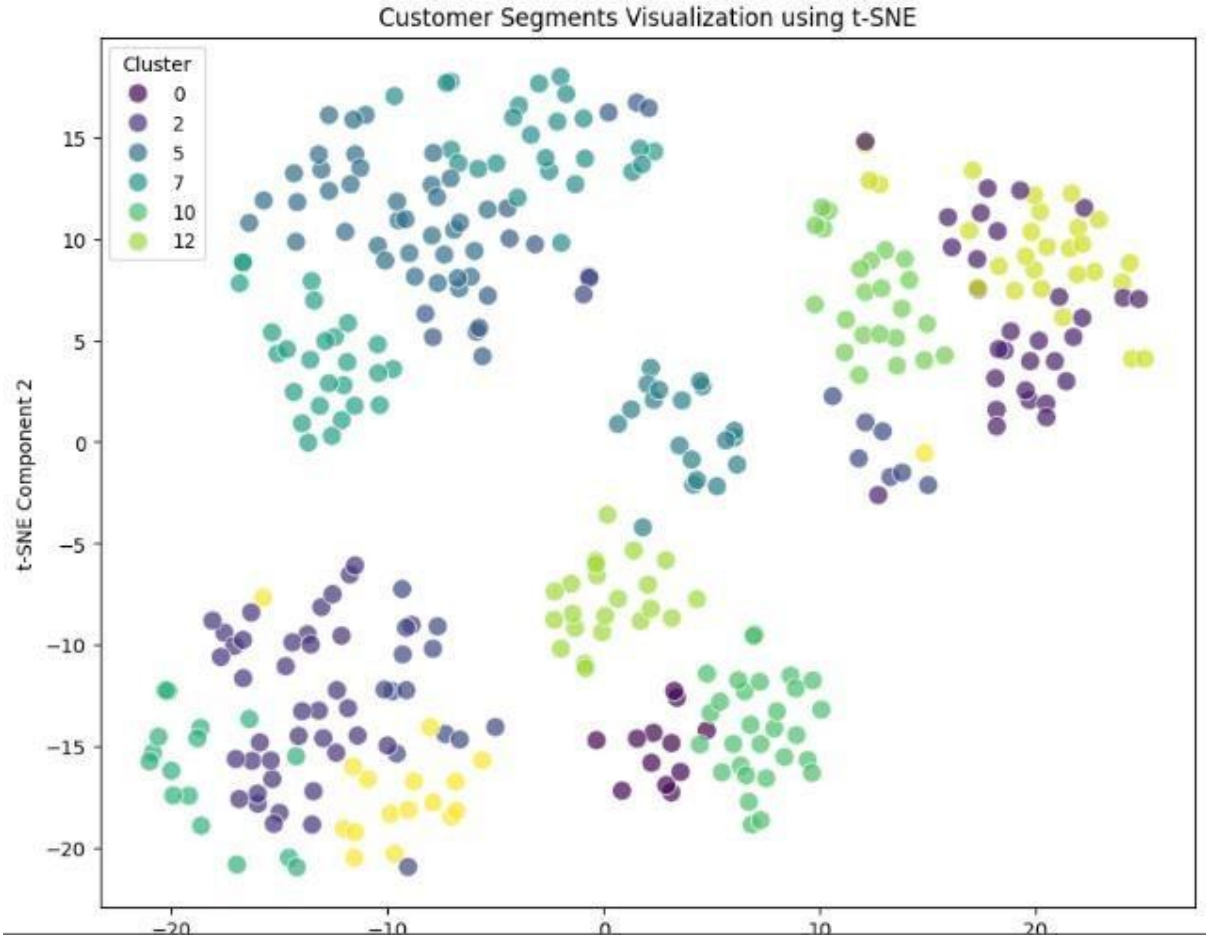
5. Module 4: Clustering Analysis

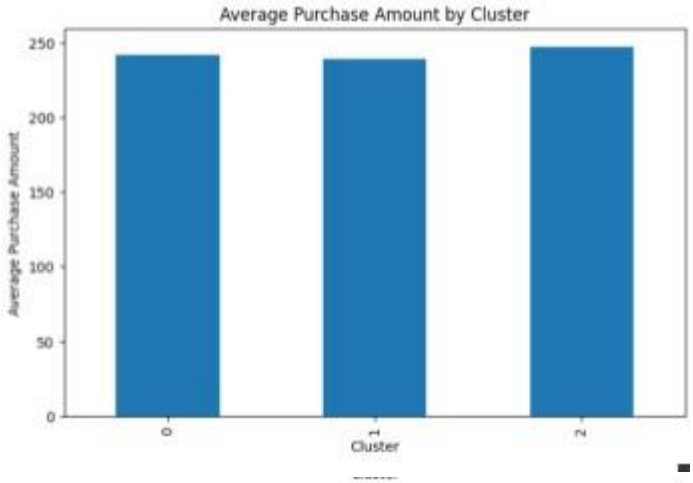
5.1 K-Means Clustering

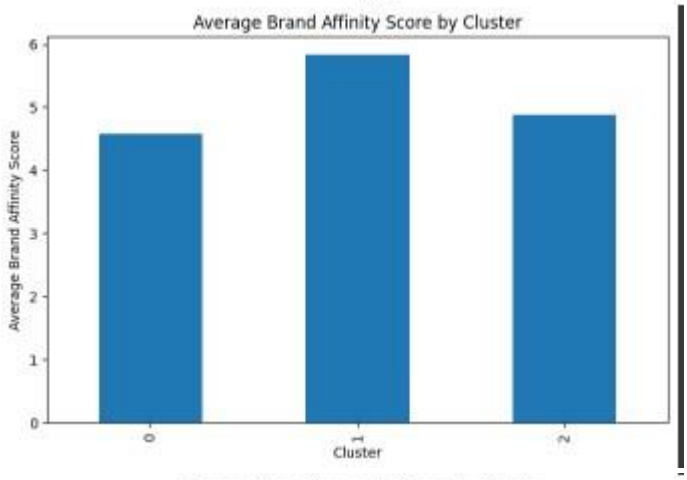
- **Defining the Number of Clusters (k)**: We used the elbow method to determine the optimal number of clusters. The elbow plot indicated that 3 clusters would be the most appropriate.
- **Clustering Implementation**: We applied K-Means clustering with k=3 to segment customers based on their purchase behavior and preferences.
- **Cluster Characteristics**: We investigated key features of each cluster, including average purchase amount, brand affinity, and product category preferences.

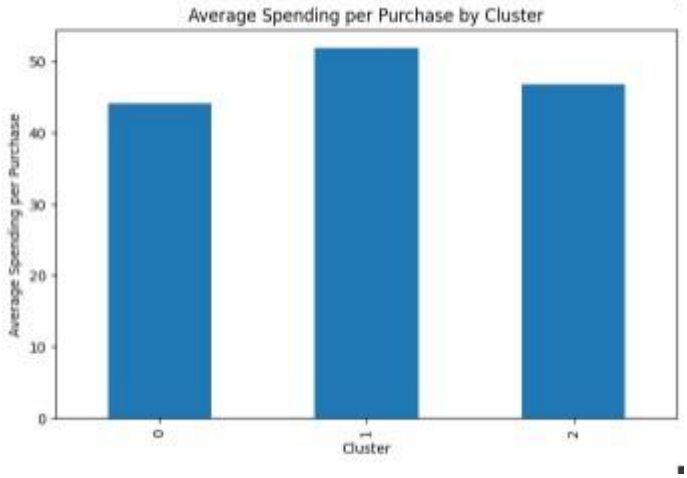
Each cluster displayed distinct characteristics in terms of spending behavior and product preferences.

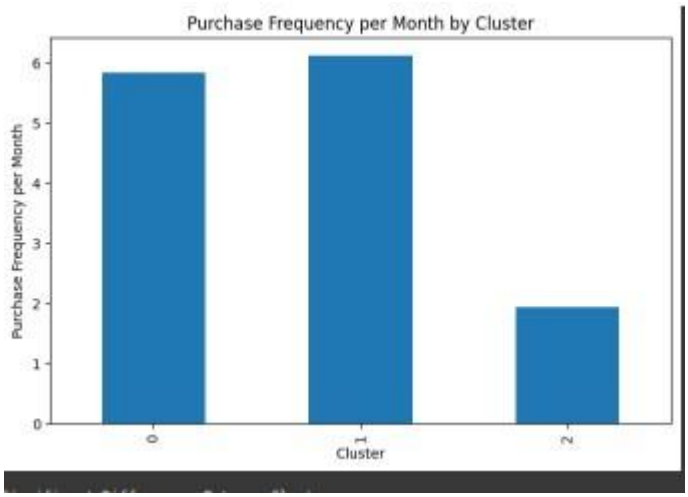





Average Purchase Amount by Cluster
Average Purchase Amount
250
200
150
100
50
0
0
1
2
Cluster


Average Brand Affinity Score by Cluster
Average Brand Affinity Score
6
5
4
3
2
1
0
0
1
2
Cluster


Average Spending per Purchase by Cluster
Average Spending per Purchase
50
40
30
20
10
0
0
1
2
Cluster




```
Significant Differences Between Clusters:
Cluster with highest average purchase amount: Cluster 2
Cluster with highest brand affinity score: Cluster 1
Cluster with highest average income level: Cluster 2

High Spending Customers:
         Purchase_Amount  Average_Spending_Per_Purchase  \
Cluster
2             247.133333                      46.816667

         Purchase_Frequency_Per_Month  Brand_Affinity_Score  Income_Level  \
Cluster
2                            1.933333              4.883333      1.233333

        Gender
Cluster
2        Other

Low Spending Customers:
         Purchase_Amount  Average_Spending_Per_Purchase  \
Cluster
0             242.153846                      44.123077
1             239.442424                      51.818182

         Purchase_Frequency_Per_Month  Brand_Affinity_Score  Income_Level  \
Cluster
0                            5.838462              4.584615      1.215385
1                            6.115152              5.830303      0.945455

         Gender
Cluster
0        Female
1         Other
```

6. Module 5: Comparison and Conclusion

6.1 Model Comparison

- **Regression vs. Decision Tree**:
    - The regression model performed well in predicting continuous values like spending per purchase but lacked the ability to classify customers into specific groups.
    - The decision tree classifier was more effective in predicting binary outcomes (e.g., whether a customer will make a purchase), with high accuracy and precision.
- **Clustering**: K-Means clustering provided valuable insights into customer segments, allowing for targeted marketing strategies based on distinct behavioural patterns.

6.2 Actionable Recommendations

Based on the analysis, the following recommendations are made for the electronics section:

- **Targeted Marketing**: Use the results from clustering to create targeted marketing campaigns for each customer segment.
- **Seasonal Promotions**: Address seasonal declines in sales by offering promotions during low-sales periods.
- **Increase Customer Retention**: Develop loyalty programs or personalized offers for customers who are identified as high-value or at-risk for leaving.

7. Conclusion

The analysis of the electronics section data has provided valuable insights into customer behaviour, allowing for data-driven recommendations that could significantly enhance sales and customer retention. The use of linear regression, decision trees, and K-Means clustering has proven effective in uncovering trends and patterns that are actionable for improving business strategies.