Analysis on Customer Data

SHOWCASING PATTERNS, TRENDS AND CLUSTERS

IMTIAZ MALL

20F1004 | 20F0101

Summary Report of Analysis

Introduction

This summary report has been generated after careful examination and analysis of the dataset given called electronics.json, the goal of analysis was to showcase distinctions between customers of Imtiaz Mall based on certain features and provide solutions or insights on how marketing can be improved for customer retention.

Module 1: Data Acquisition and Preprocessing

Data Acquisition:

The first step was to get the historical sales data and then separate the electronics section for analysis. Once loaded into python, we performed some manual examination of the given data by first creating a csv of the given json file. After examination, we noticed many empty (‘’) cells and ‘hidden’ cells. These cells were dealt with in the data cleaning process because of the inconsistencies they would generate in further analysis.

Data Cleaning:

After finding out the inconsistencies in data, we considered empty (‘’) and ‘hidden’ cells as Null values. The logic behind the merging of both into Null is that both are ultimately hidden and not provided hence we could either consider both as ‘hidden’ or both as Null which would mean the same thing.

Further cleaning was performed after examining the column that provides the initial segmentation which is **‘Product_Category’**. The column contained Null values (after merge). We dropped all the rows with null values to make the data consistent and keep the defined values which are **Electronics, Clothing and Books**.

We used the **forward fill** method for columns which contained categorical data as it seems to keep the change in difference between the values almost the same as without ffill. This seemed very appropriate for columns such as **‘brands’, ‘dates’, ‘gender’** etc.

We imputed numerical values using mean of the column. This is a safer option when the skew is balanced as the mean represents the central tendency of the data. We checked every numerical column skew before imputing it and we found out that the skewness of all columns is **0+- 10.**

```
...  <class 'pandas.core.frame.DataFrame'>
     Int64Index: 940 entries, 899 to 984
     Data columns (total 18 columns):
      #   Column                        Non-Null Count  Dtype
     ---  ------                        --------------  -----
      0   Customer_ID                   940 non-null    object
      1   Age                           940 non-null    float64
      2   Gender                        940 non-null    object
      3   Income_Level                  940 non-null    object
      4   Address                       940 non-null    object
      5   Transaction_ID                940 non-null    object
      6   Purchase_Date                 940 non-null    datetime64[ns]
      7   Product_ID                    940 non-null    object
      8   Product_Category              940 non-null    object
      9   Brand                         940 non-null    object
      10  Purchase_Amount               940 non-null    float64
      11  Average_Spending_Per_Purchase 940 non-null    float64
      12  Purchase_Frequency_Per_Month  940 non-null    float64
      13  Brand_Affinity_Score          940 non-null    Int64
      14  Product_Category_Preferences  940 non-null    object
      15  Month                         940 non-null    Int64
      16  Year                          940 non-null    Int64
      17  Season                        940 non-null    object
     dtypes: Int64(3), datetime64[ns](1), float64(4), object(10)
     memory usage: 142.3+ KB
```

	Customer_ID	Age	Gender	Income_Level	Address	Transaction_ID	Purchase_Date	Product_ID	Pr
899	a73774fe-d420-46ca-8a43-44eb51438f5e	48.0	Other	High	1412 Blake Parkway Apt. 316\nLake Rodneycheste...	c1cba058-2afd-41e4-826c-ea03c51afaad	2020-01-02	1c72a791-7b4d-4f7d-960e-7a611428a870	
788	8f25e25c-75c7-4eb7-b2e2-f708dee8ef13	39.0	Female	Low	414 Lauren Mountain Suite 243\nSouth Jessicabe...	638cded1-9504-4fc9-a1e1-09ee49388c8e	2020-01-03	495c76ec-35f1-4b80-86b5-b91558ffb2a5	
414	228febfa-bfb5-413a-ab8a-1eeb905b36fd	40.0	Female	Low	50568 Joseph Prairie\nPort Kimberlyview, ND 33279	96375f25-2e13-4e66-8e76-bbbf06760439	2020-01-04	aab09f53-a4a1-400e-932f-62120350545b	
160	09427631-943f-4427-80e6-79c9da0c2613	71.0	Other	High	4363 Leslie Hills\nLake Mary, FL 20948	be33a103-bf30-4787-ad68-54a3efc8d675	2020-01-05	6cb25dba-2dd1-4724-84d0-322497ead674	
389	d8fbc8d7-7b8a-4903-85b7-630519ab33d7	40.0	Male	Low	0114 Jacob Passage Suite 324\nAmandastad, NV 1...	fa0db7eb-2748-4e3d-95dc-8bf48c5dac5f	2020-01-07	878b2b79-6f19-4162-9164-953af3f6e903	

Data Transformation:

After cleaning the data, we performed data one hot encoding on 4 features which are **'Gender', 'Product_Category', 'Brand', 'Product_Category_Preferences'** and performed label encoding on **'Income_Level'**, later when it was time to perform clustering we dropped unnecessary columns such as **'Age_Group', 'Purchase_Amount_Binned', 'Customer_ID', 'Address', 'Transaction_ID', 'Purchase_Date', 'Product_ID', 'Season', 'Purchase_Date'** but Purchase_Date was extracted into year, month and day columns.

Afterwards the data was scaled using the MinMaxScalar.

	Age	Income_Level	Purchase_Amount	Average_Spending_Per_Purchase	Purchase_Frequency_Per_Month	Bran
0	0.483871	1.0	0.844898	0.400000	0.555556	
1	0.338710	0.0	0.853061	0.915789	0.000000	
2	0.354839	0.0	0.810204	0.473684	0.888889	
3	0.854839	1.0	0.757143	0.389474	0.444444	
4	0.354839	0.0	0.900000	0.505263	0.888889	
...	...	...	...	...	...	
935	0.322581	0.5	0.089796	0.768421	0.222222	
936	0.306452	1.0	0.530612	0.073684	0.222222	
937	0.419355	1.0	0.112245	0.073684	0.444444	
938	0.064516	1.0	0.506122	0.105263	0.555556	
939	0.677419	0.5	0.830612	0.694737	1.000000	

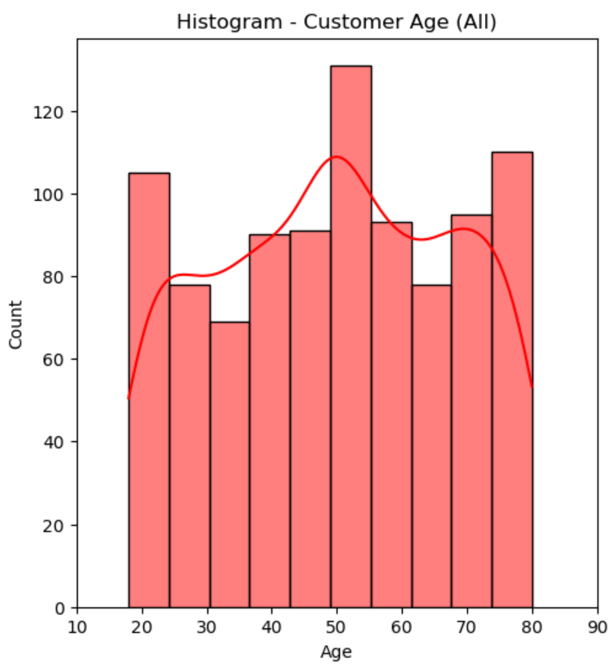
940 rows × 25 columns

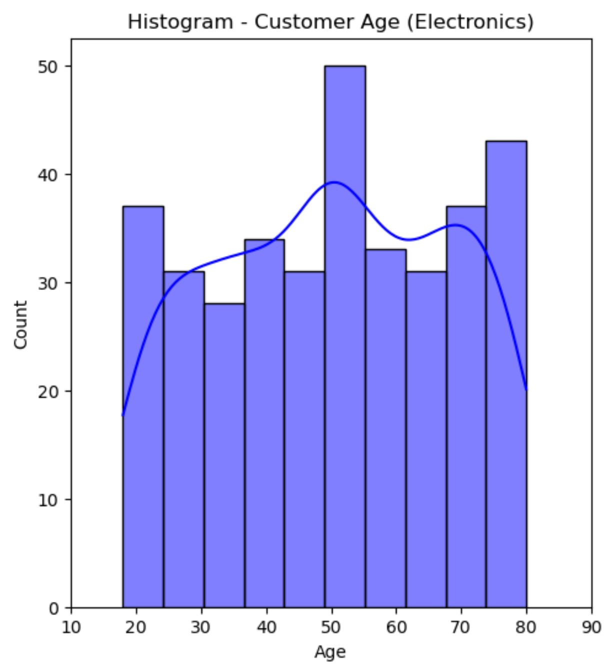
Module 2: Exploratory Data Analysis (EDA):

Univariate Analysis:

Note: All Data Analysis was performed on both all and only electronics category side by side.

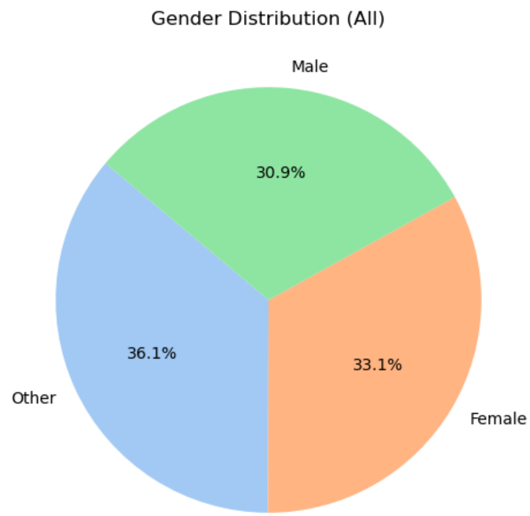
Age Distribution (All vs Electronics)

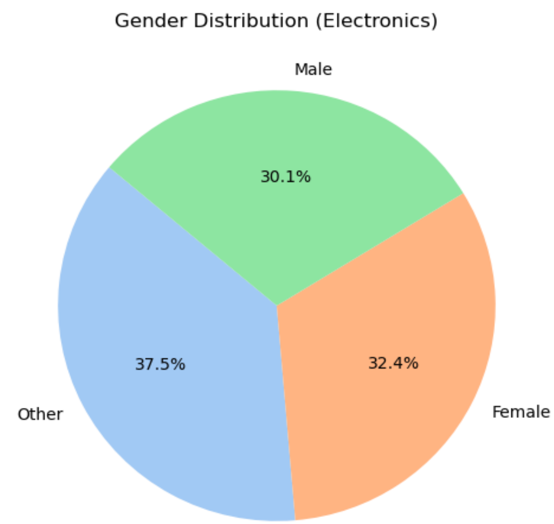




Histogram on the Age column are very similar for all products and only electronics product. The Central Tendency lies around 50 and skew is balanced for both data. The store can target the most occurring age groups in their marketing to boost their sales.

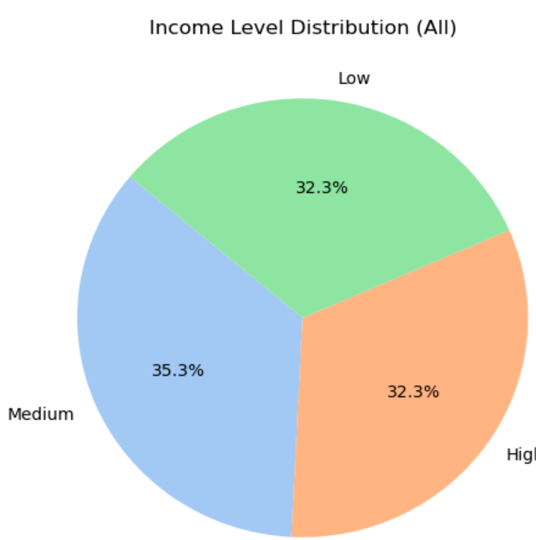
Gender Distribution:

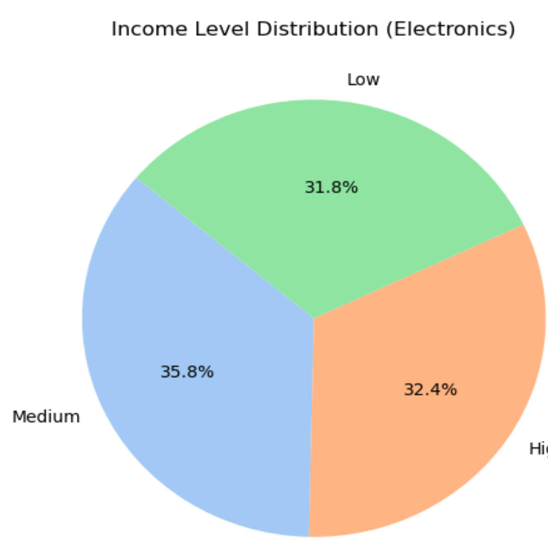




The Pie charts for Gender Distribution shows no particular imbalance between males and females. The notable fact from these pie charts are that females lead males by a small margin for all products and electronics. It comes as a surprise that females happen to purchase more electronics than males. The other category consists of 36.1% of the distribution in all product categories where as it is 37.5% for electronics.

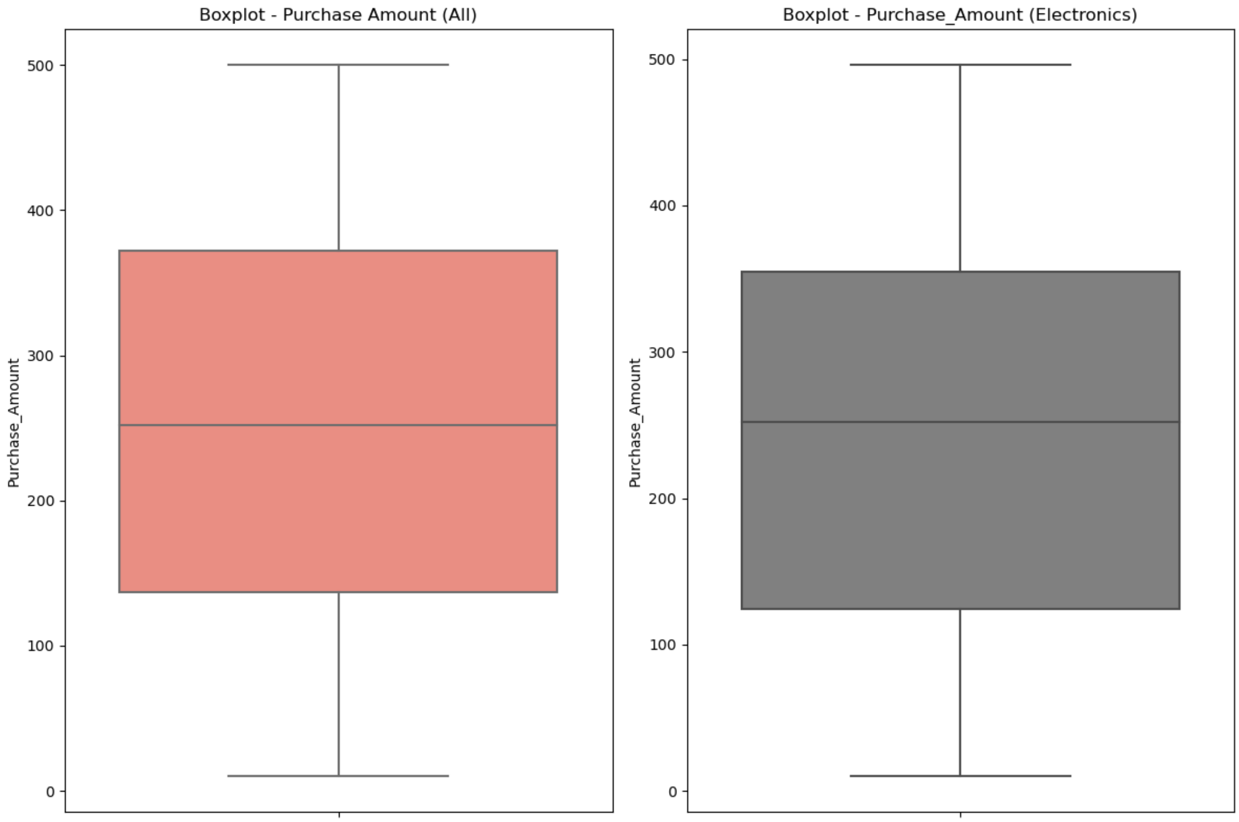
Income Level Distribution:





The Income Level distribution is also nearly equal. There is no such marginal difference between the Income levels of the customers for electronics and all products. Medium Income Level leads in both distributions at 35.3% and 35.8%. Since there does exist a relatively small gap for medium compared to the two other. The store can provide sales on products which are normally bought by High Income Level Customers so that the Medium Customers can have more opportunities to purchase products out of their range ultimately bringing in more Medium Customers.

Purchase Amount Distribution:

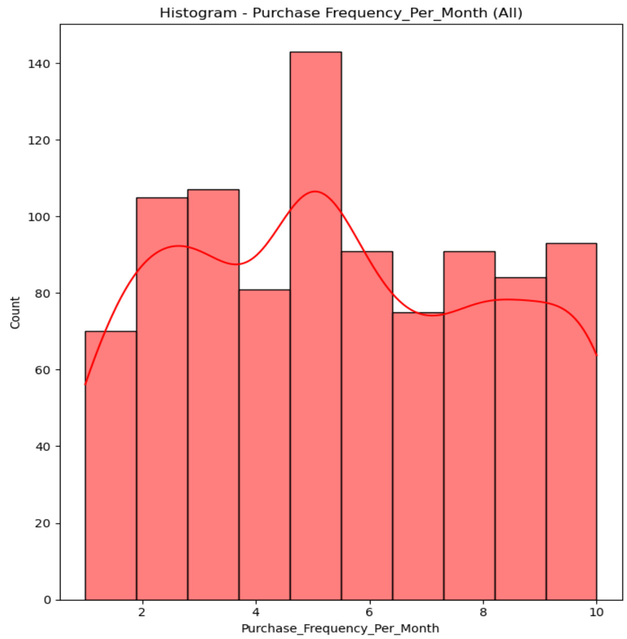


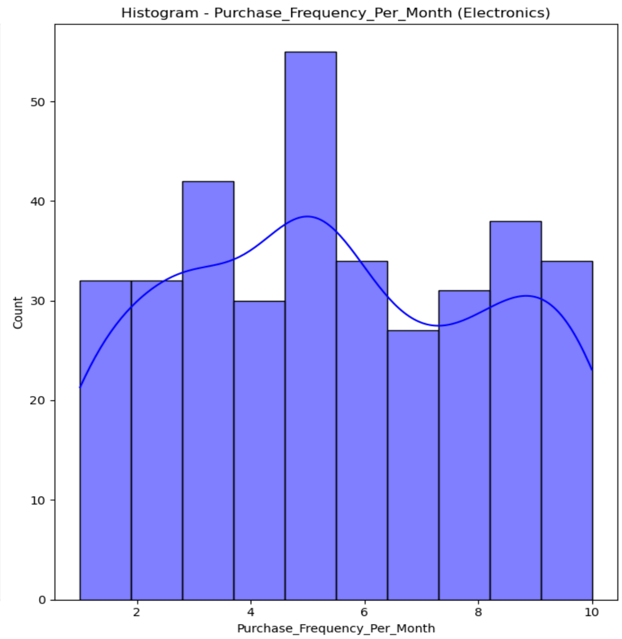
In the Box plot for All Product Categories the median purchase amount is around the lower mid-range of the data, indicating a skew towards lower purchase amounts. The interquartile range (IQR) spans from low to mid-range, suggesting moderate variability in purchase amounts among customers.

In the Box plot for Electronics the median here appears slightly higher than the median for all customers, suggesting that electronics purchases tend to be somewhat more expensive. The IQR is also broader, which implies greater variability in the amount spent on electronics.

The gap is not that significant but It can be taken that electronics tend to have a higher median and greater variability in amounts spent.

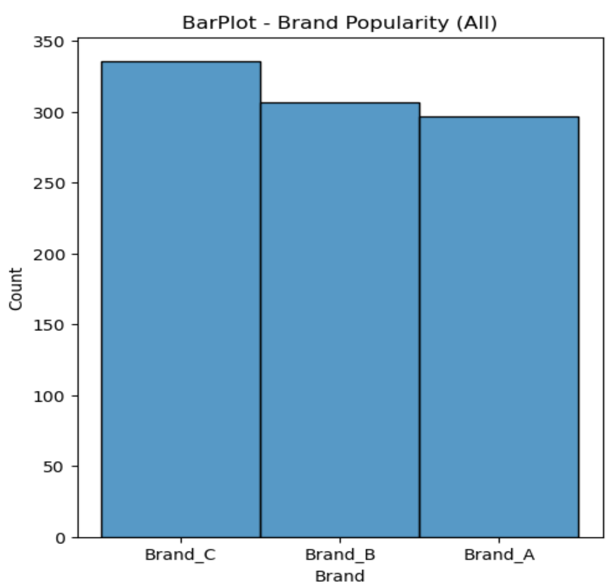
Purchase Frequency Distribution:

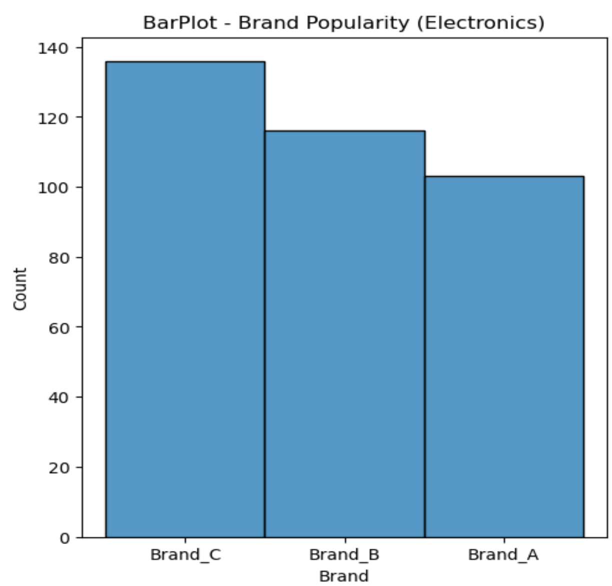




The above two plots display the Purchase Frequency Per Month for All Product Categories and Electronics only. The skew for both plots is balanced, the 5th month show a peak due it being the mean and imputation performed during the cleaning phase has boosted the mean more while keeping the rest of the data relatively same. The mall should target to boost up the lower frequency months and try to keep the peaks balanced.

Brand Frequency Distribution:





From the Brand Frequency Distribution, we can see that both plots indicate that Brand C is the most popular for All Product Categories as well as Electronics only, Followed by Brand B and then Brand A.

Descriptive Statistics:

For All Data:

	Age	Purchase_Amount	Purchase_Frequency_Per_Month
count	940.000000	940.000000	940.000000
mean	49.639362	251.537234	5.452128
std	18.054548	137.804378	2.780743
min	18.000000	10.000000	1.000000
25%	35.000000	136.750000	3.000000
50%	50.000000	252.000000	5.000000
75%	65.250000	372.000000	8.000000
max	80.000000	500.000000	10.000000

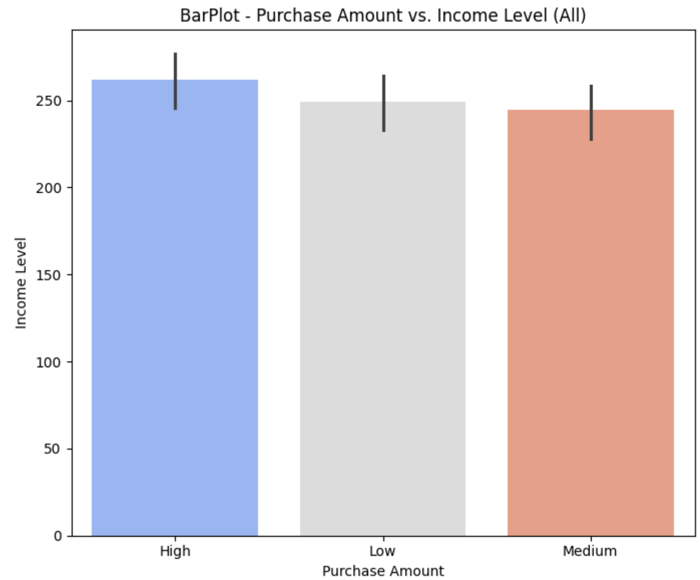
For Electronics only:

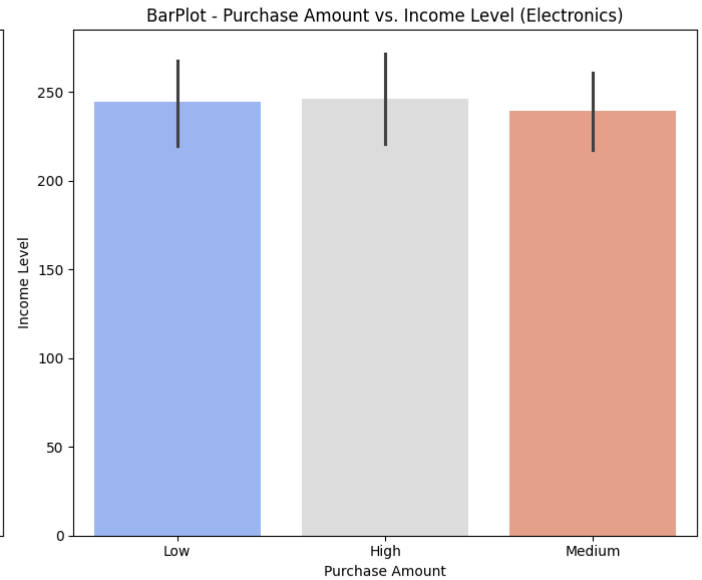
	Age	Purchase_Amount	Purchase_Frequency_Per_Month
count	355.000000	355.000000	355.000000
mean	49.870423	243.171831	5.464789
std	18.109993	136.427735	2.812558
min	18.000000	10.000000	1.000000
25%	35.000000	124.500000	3.000000
50%	50.000000	252.000000	5.000000
75%	66.000000	354.500000	8.000000
max	80.000000	496.000000	10.000000

The above descriptions describe the quartiles, max, min, standard deviation, mean and count for all products and only electronic products respectively.

Bivariate Analysis:

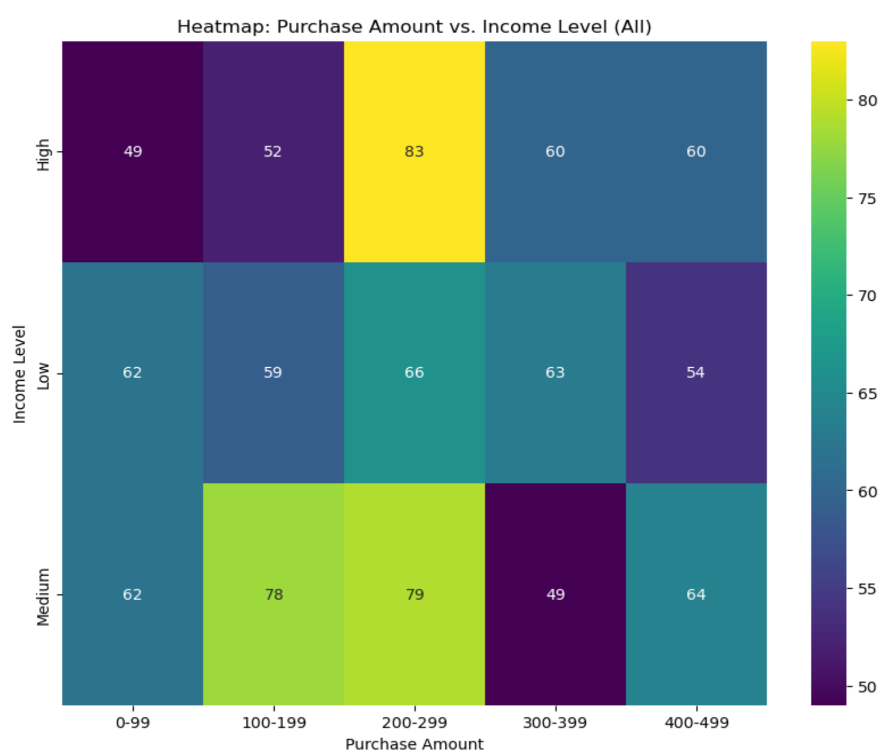
Purchase Amount against Income Level:





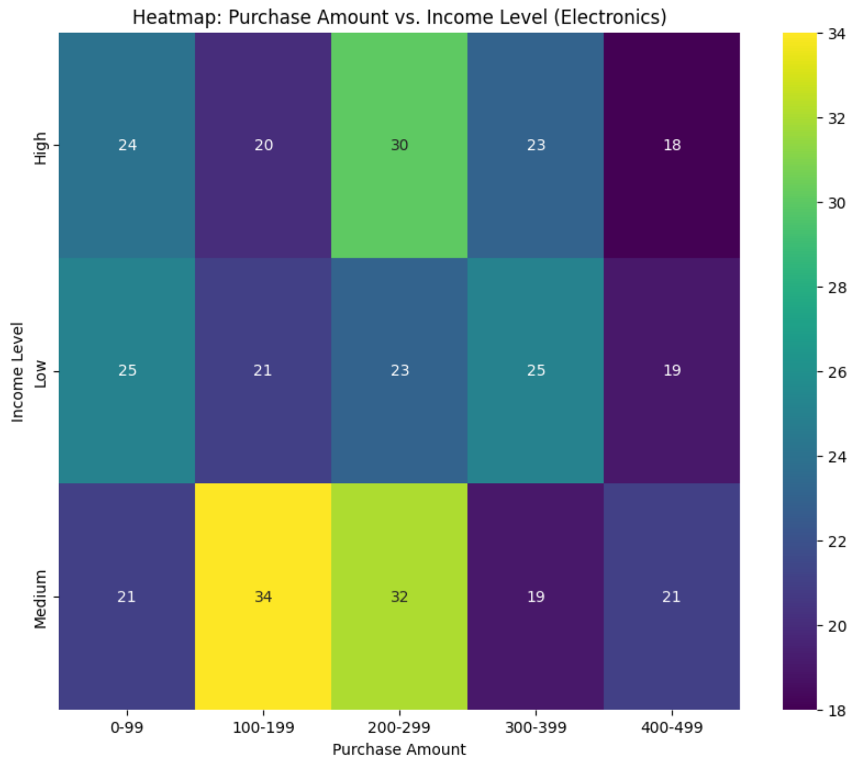
The left side plot represents the relationship between the Purchase Amount and Income Level for all purchases. There is a general trend of increasing purchase amounts with higher income levels. Customer with higher purchase tend to make larger purchase.

The right side plot focuses on the subset of Purchases related to the Electronics. Here, we can see the Purchase Amount of Low, High, and Medium Income Levels differ with lesser amounts. Especially, low and high amounts have very less difference.



Here, we have the heat map for the Purchase Amount vs Income Level for all (general all categories purchases). Here, it illustrates the distribution of the different income levels, and binned purchased amounts.

Significant frequency of purchases occurs in the High Income Level for range (200-299). Where as the Medium shows relatively less distribution across purchase amounts, 200-299 with a slight preference for the 100-199 also.

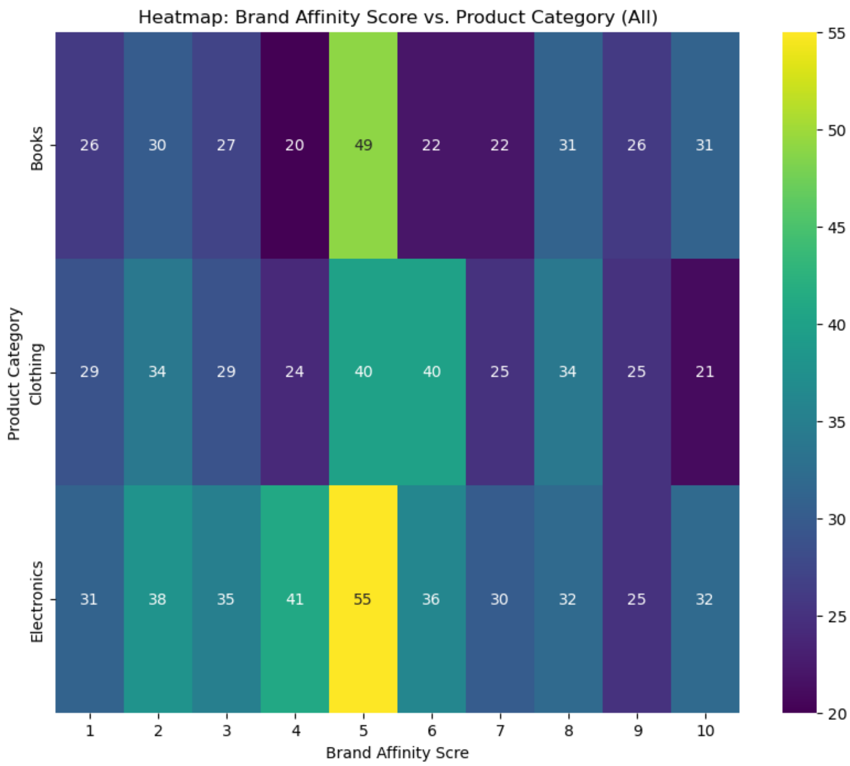


This heat map focuses specially on the purchases related to the Electronics. Providing the insights into the relationship between the income levels, and purchase amounts for electronics.

Medium Income Level were the most frequent purchaser in the 100-199 range. High-income. Here the High Income Level also shows the fairly even purchases across all amount ranges. But, with fewer transactions than the All heat map, where it was dominant in 200-299 range.

In both heat maps, the spending does not drastically increase with the Income Level. Suggesting that all income levels prioritize the purchases accordingly, or there is price that appeals across the segment.

Brands Affinity Score vs Product Category:

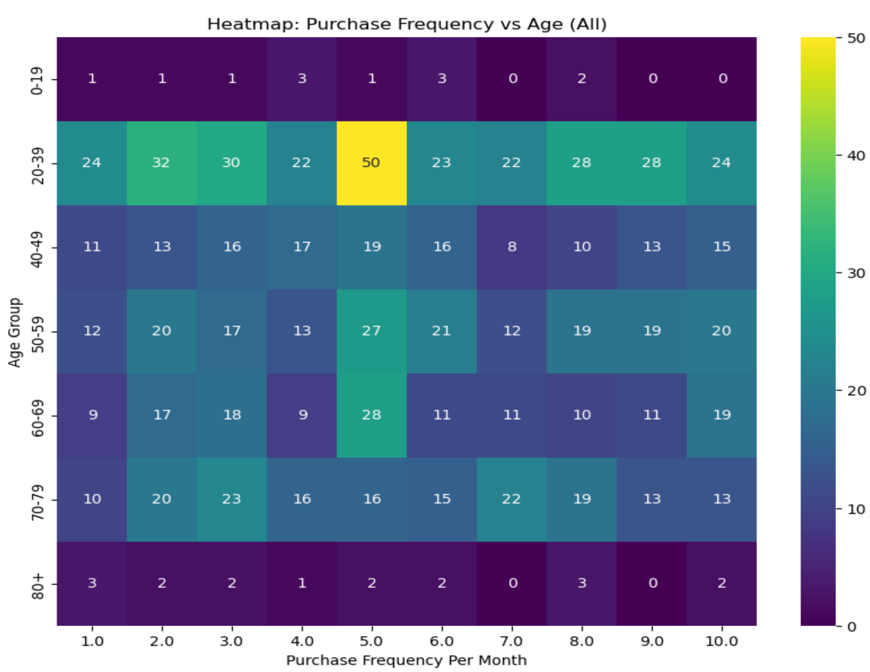


The heat map for Brand Affinity vs Product Category suggests there is a varying degree of brand affinity across the different product category

A standout observation here is that within Electronics, the brand affinity is 5 has a high score of 55. This means that customer have strong preference towards that one brand for electronics.

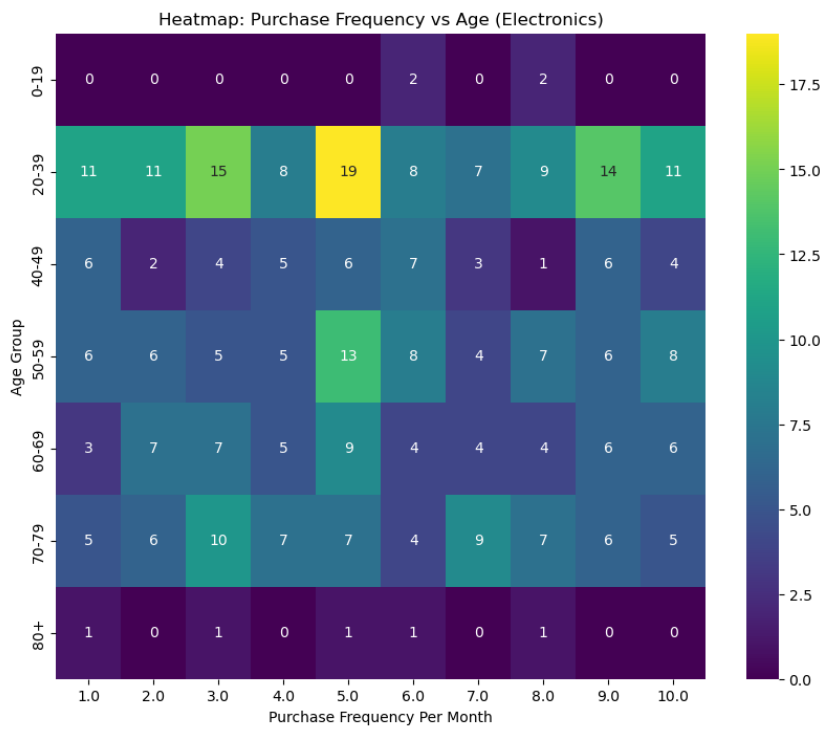
While, other categories like Clothing, and Books show a more moderate distributed brand affinity, resulting in less customer preference for certain brands in these 2 categories.

Purchase Frequency vs Age:



This heat map compares the frequency of purchases across all the categories product for all the age groups. The color intensity represents how often different age groups make purchase.

Here, the heat map focuses on the Purchase Frequency vs Age. So, the most frequent purchase activities are among 20-39 age, at an age frequency of 5.0. It is the peak in the dateset.

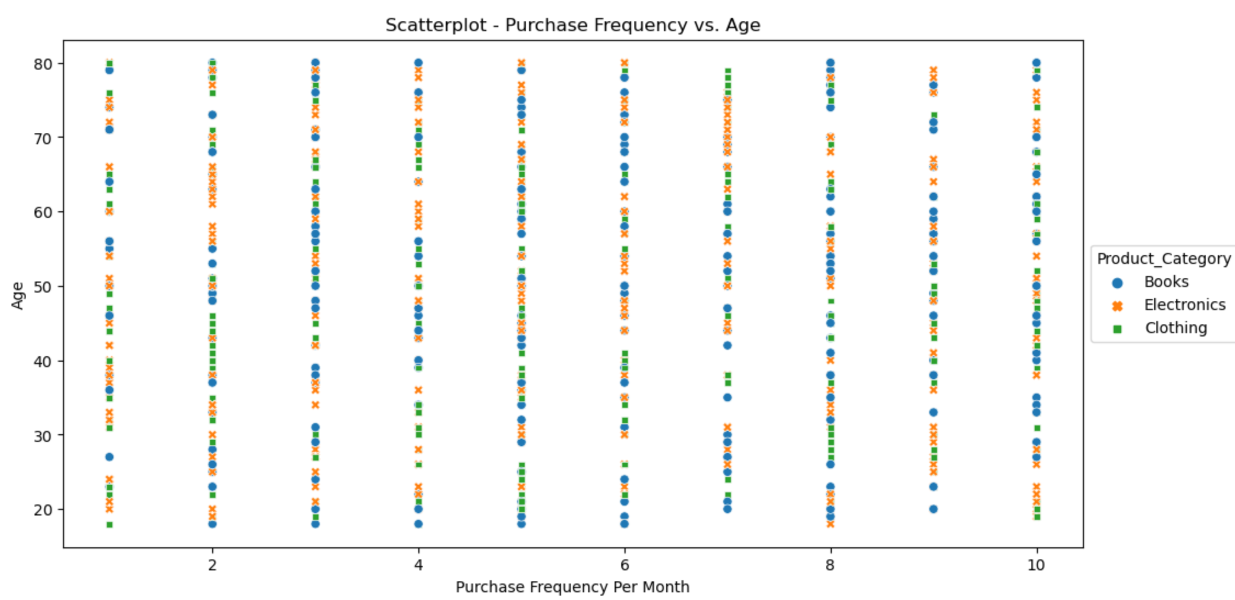


This heat map compares the frequency of purchases of the Electronics across all the age groups. Here also the color intensity represents the how often the age group make the purchase.

The heat map shows the spike in the 20-39 age group at a purchase frequency of 5.0 per month.

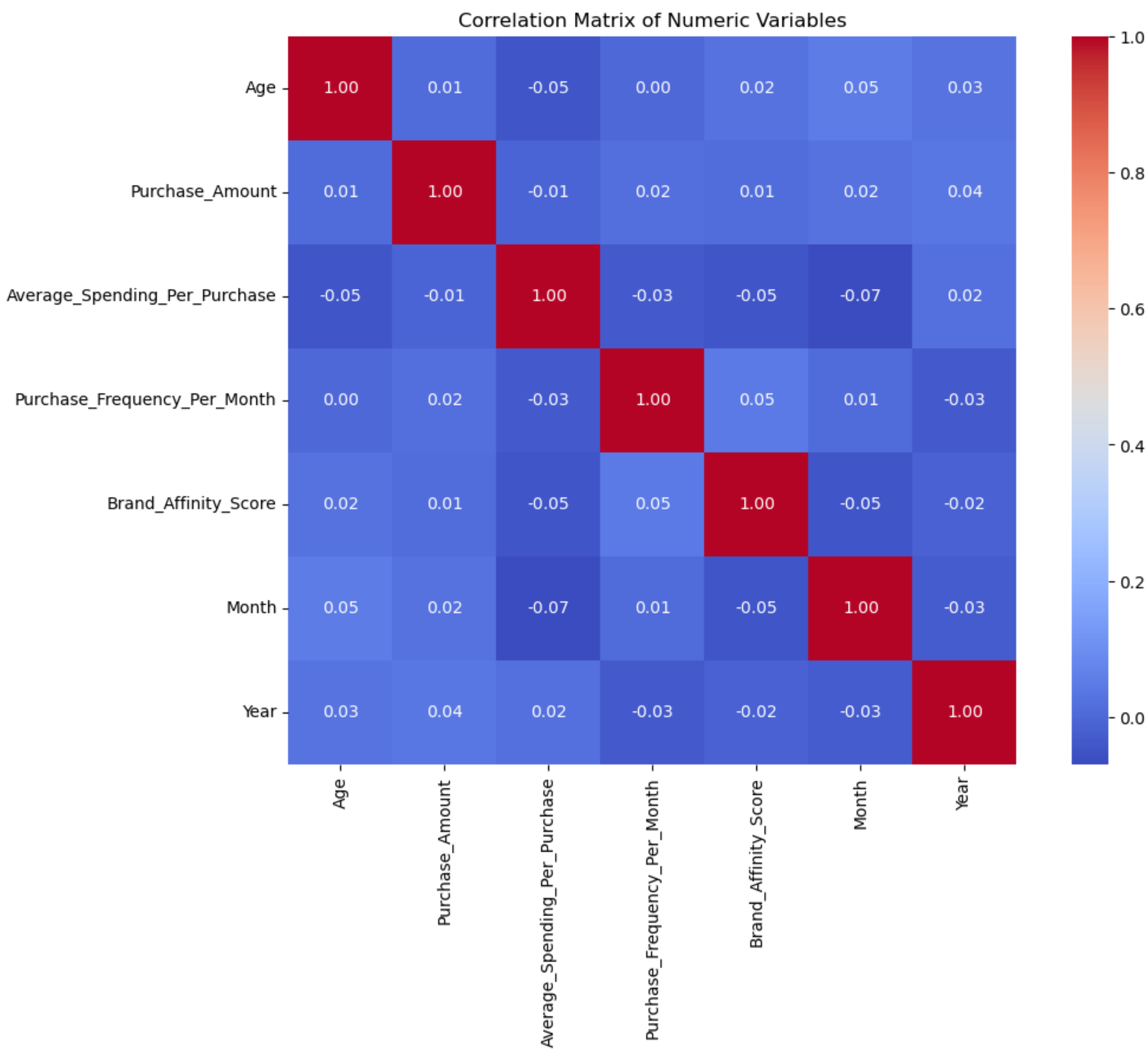
Thus, it shows this age group is prominent in buying the electronics more often than other age groups.

Across both the heat maps, the younger age groups exhibits more higher purchase frequencies, which ultimately decreases with age. Thus, reflecting higher engagement with the market among younger customers.



Here, the scatter plot for the "Purchase Frequency vs Age" reveals the distribution of purchase frequencies across different age groups, differentiated by product category. Purchase Frequency is relatively the consistent across all the ages, with out the clear trend that one age group occurs more, or less frequently than the other. All three product categories: Books, Electronics, and Clothing are represented across the ages. But there is not a strong age specific preference for any particular product.

Presence of Correlations:



Here, the correlation matrix visualize the pairwise relationships between the various numerical variables. The values ranges from -1 t0 1, where 1 represents the positive correlation, -1 a perfect negative correlation, and 0 no correlation.

So, it is evident that there is no strong correlation between any of the pairs. As all are close to zero. The Purchase Amount and Average_Spending_Per_Purchase have a small positive correlation. Which is intuitive since higher purchase amounts might result in higher spending.

Where as Brand_Affinity does not show the a significant correlation with the Purchase Amount and Average Spending Per Purchase. So, it does not lead to higher spending.

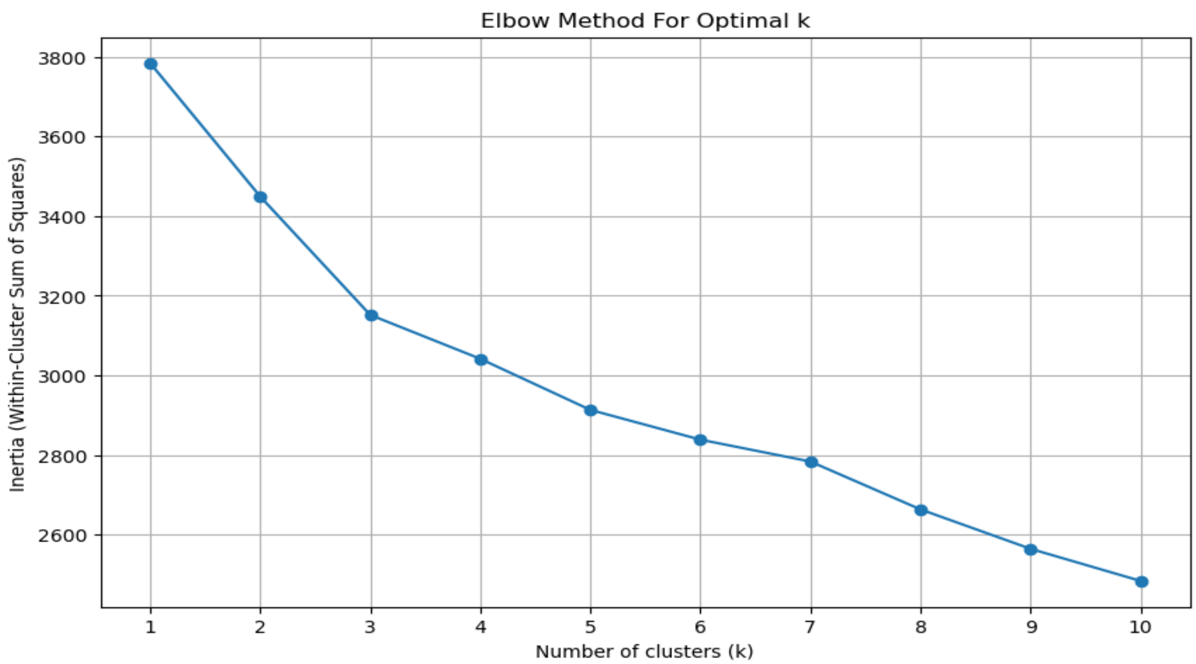
Month and Year have no meaningful correlation with other variables, indicating no significant changes or trends in the data based on time variables provided in here.

Temporal Analysis:

Module 3: Clustering Analysis

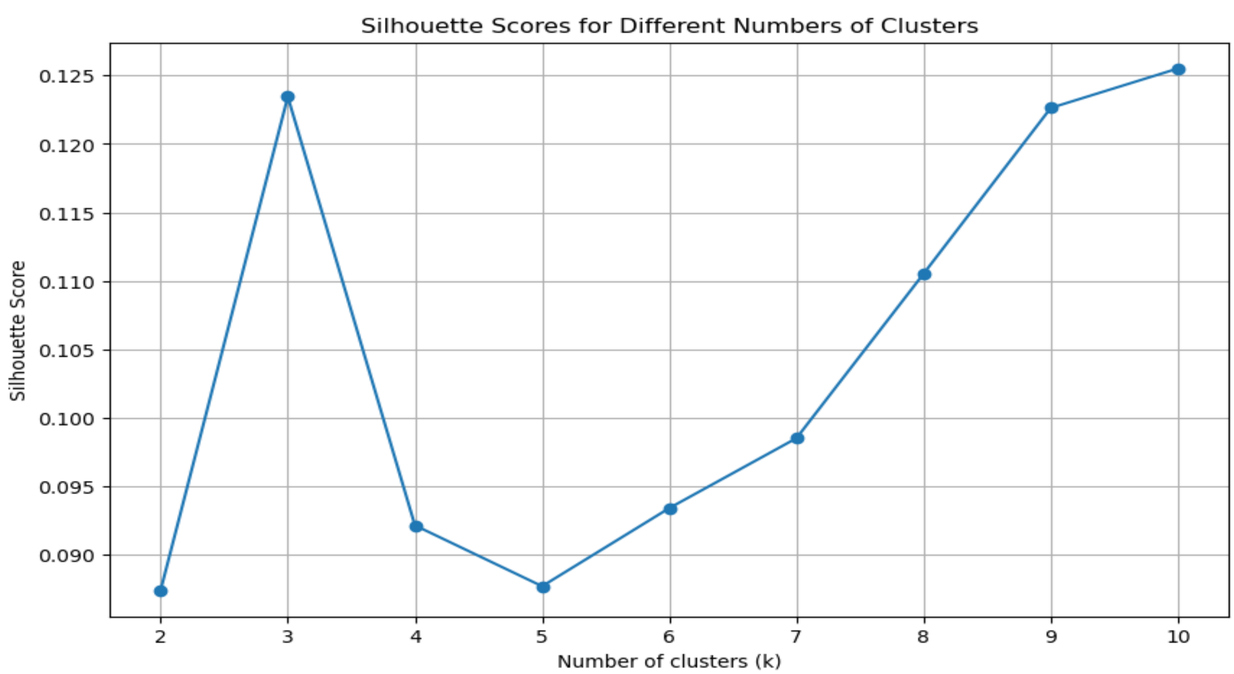
K-Means Clustering (Finding Optimal k)

Elbow plot:



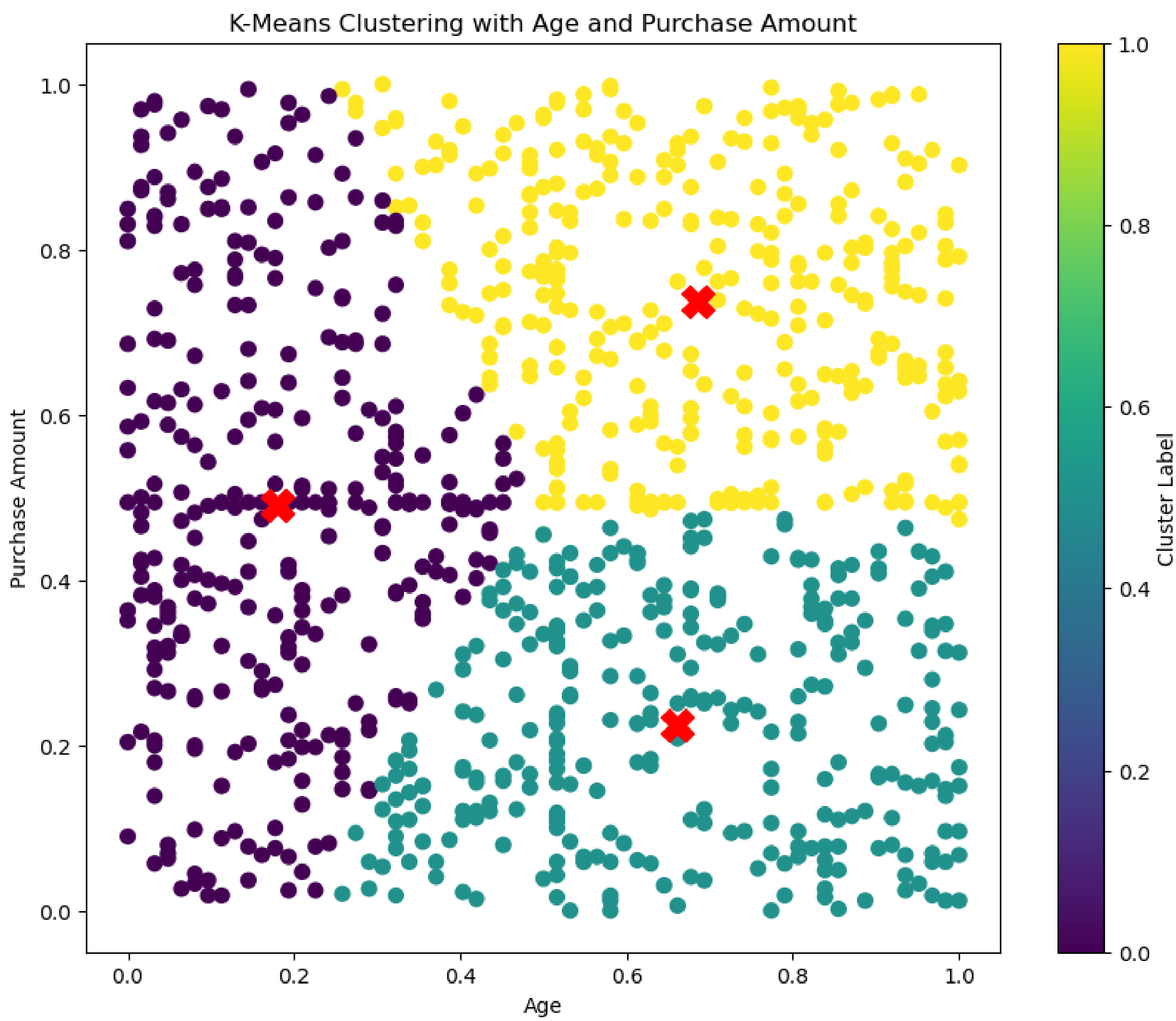
Elbow methods indicates a good k value to be 3.

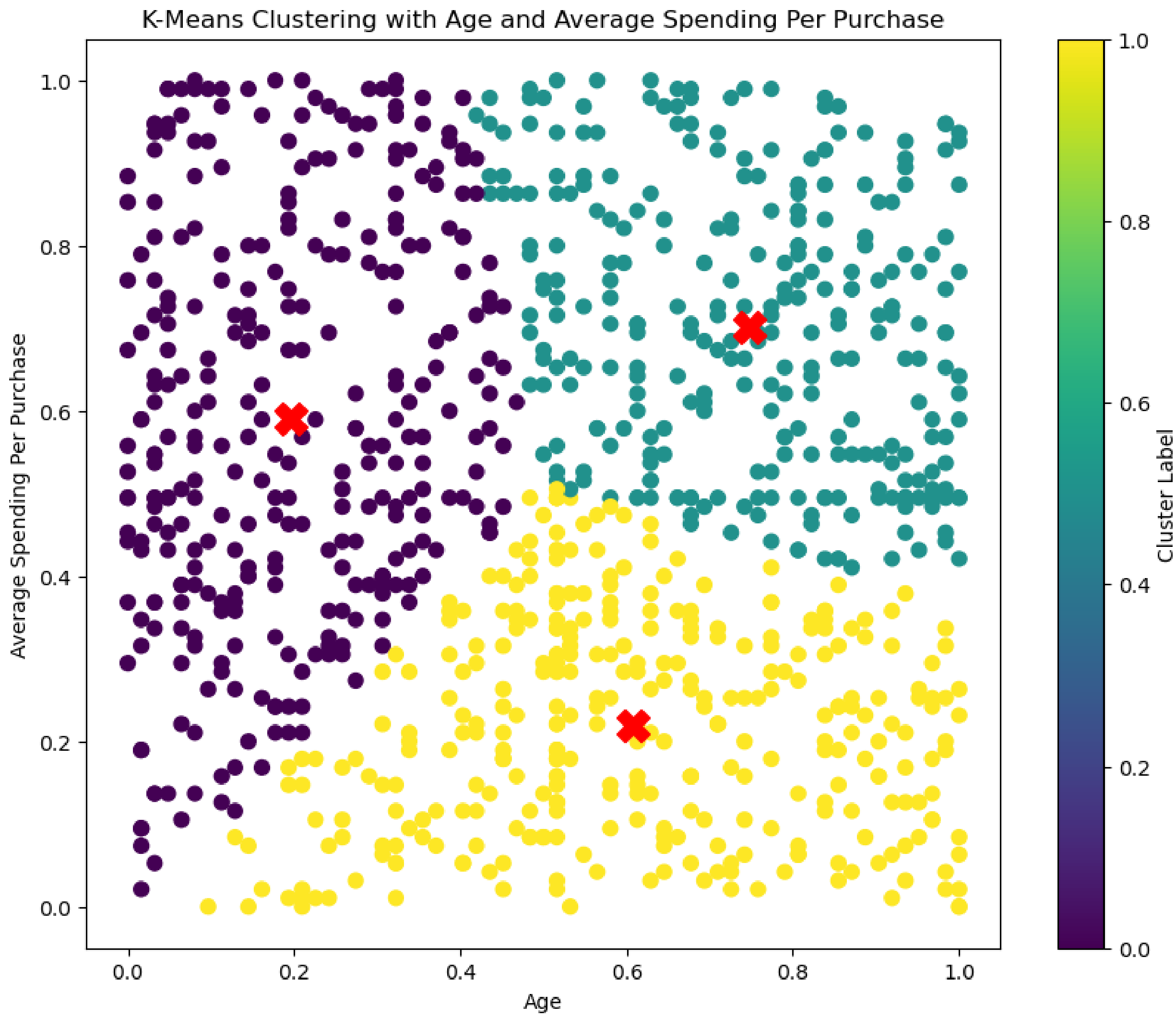
Silhouette Scores:



Silhouette Scores indicate that a good k value can be 3 and 10, we will choose 3 as our k for all clustering.

Performing K-Means:





Here we have taken Age and Average Spending Per Purchase as our x and y axis respectively. There are 3 clusters which have separated young, old and young+old.

The young cluster 20-50 age have been placed under 70 avg spending per purchase meanwhile the from 45-max are part of old which also have almost the same threshold on avg spending per purchase. The last cluster composes of all age groups who have an average spending of above 50.

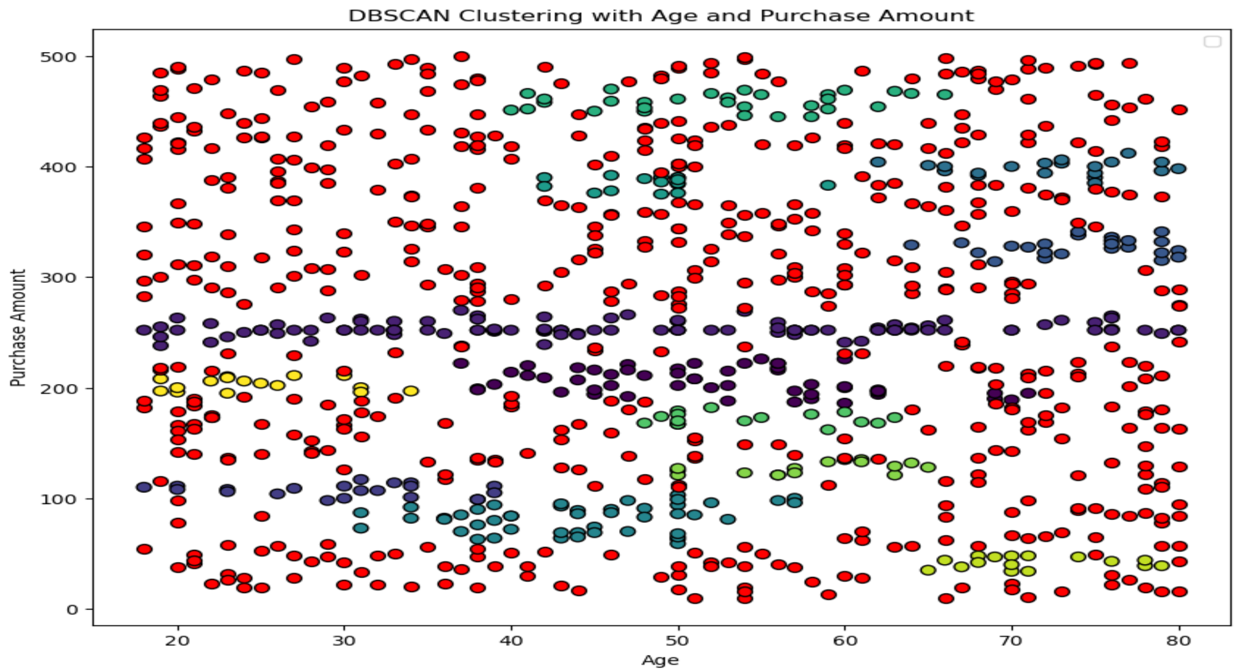
Analyzing cluster characteristics:

```
Cluster 0 characteristics:
Age                                       0.516978
Income_Level                              0.462406
Purchase_Amount                           0.493555
Average_Spending_Per_Purchase             0.471349
Purchase_Frequency_Per_Month              0.311612
Brand_Affinity_Score                      0.791562
Month                                     0.487013
Year                                      0.489715
Year_PD                                   0.506266
Month_PM                                  0.531784
Gender_Female                             0.323308
Gender_Male                               0.300752
Gender_Other                              0.375940
Product_Category_Books                    0.293233
Product_Category_Clothing                 0.357143
Product_Category_Electronics              0.349624
Brand_Brand_A                             0.353383
Brand_Brand_B                             0.300752
Brand_Brand_C                             0.345865
Product_Category_Preferences_High         0.285714
Product_Category_Preferences_Low          0.360902
Product_Category_Preferences_Medium       0.353383
Extract_Date                              0.427444
Extract_Month                             0.531784
...
Extract_Year                              0.523810
Cluster_DBSCAN                            0.000000
Name: 2, dtype: float64
```

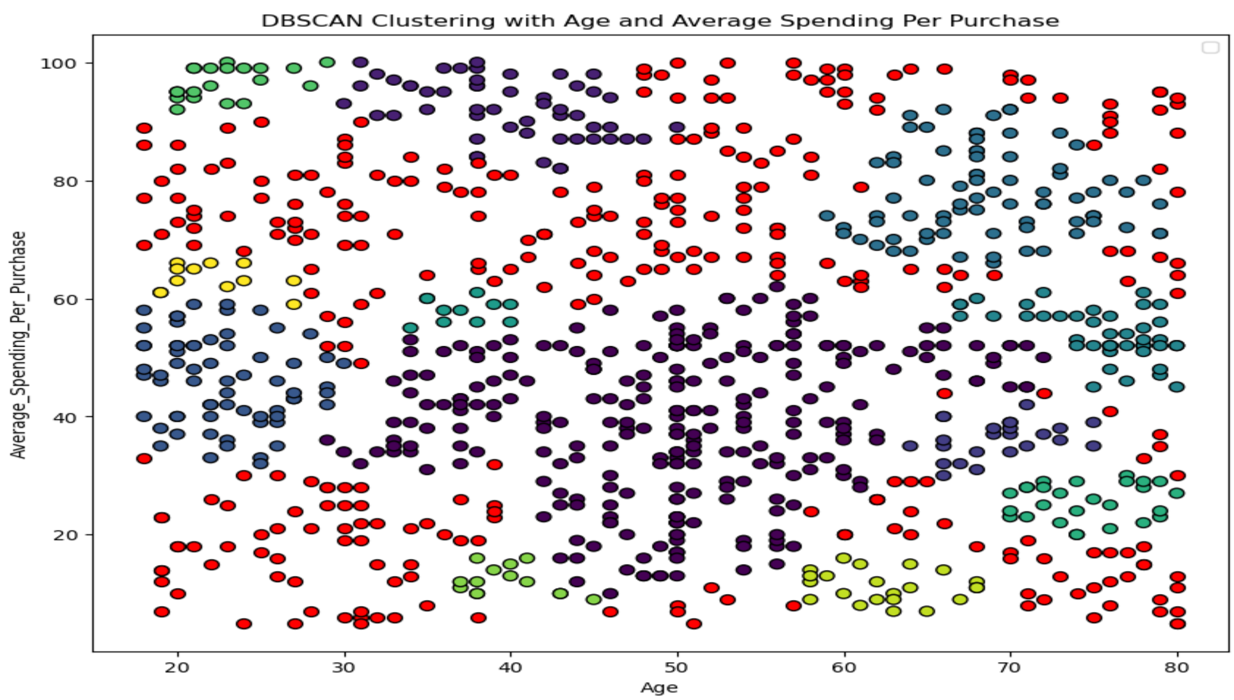
DBSCAN Clustering:

Applying DBSCAN Algorithm:

Eps=10, min_samples=15



Eps=5, min_samples=16

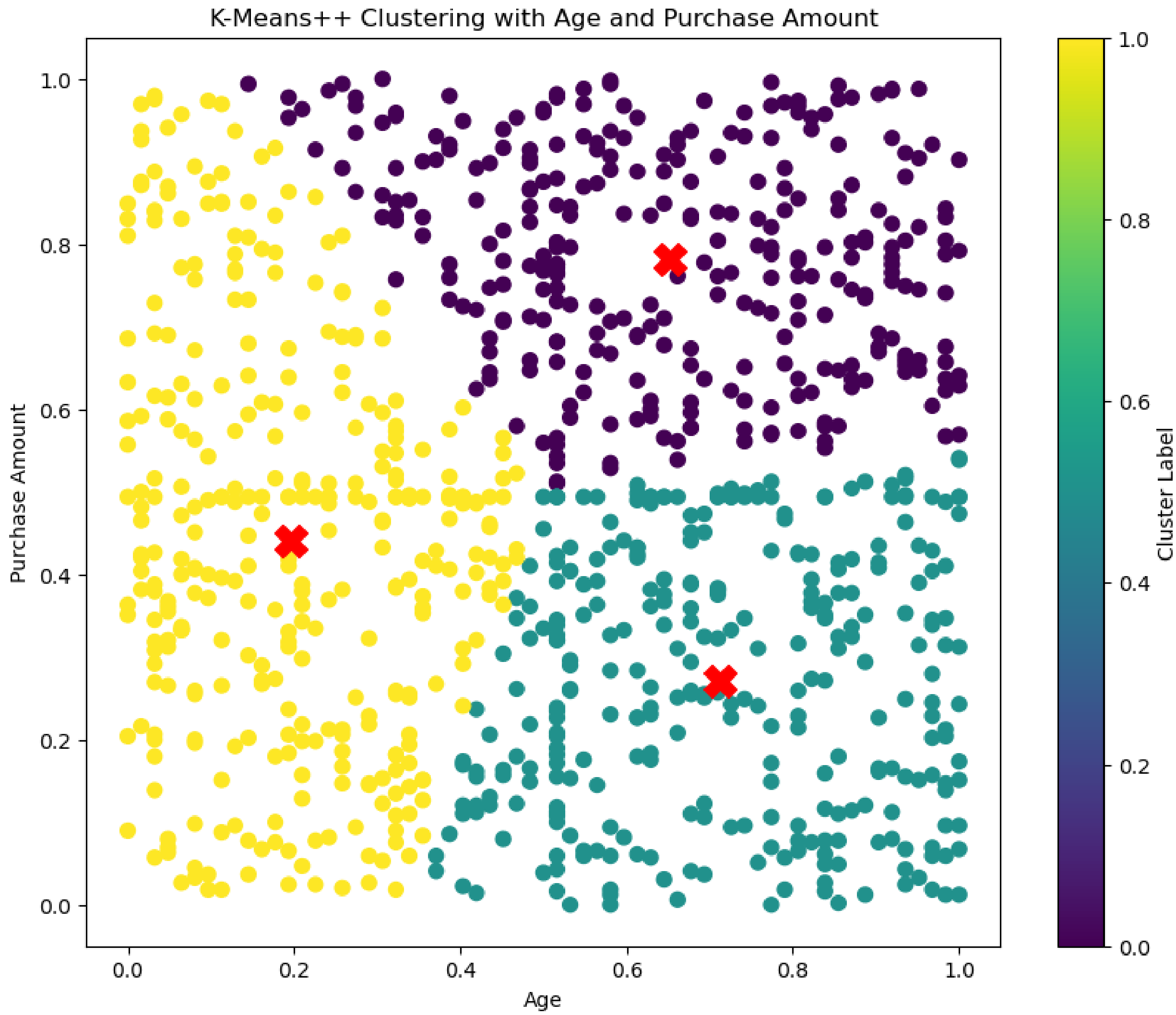


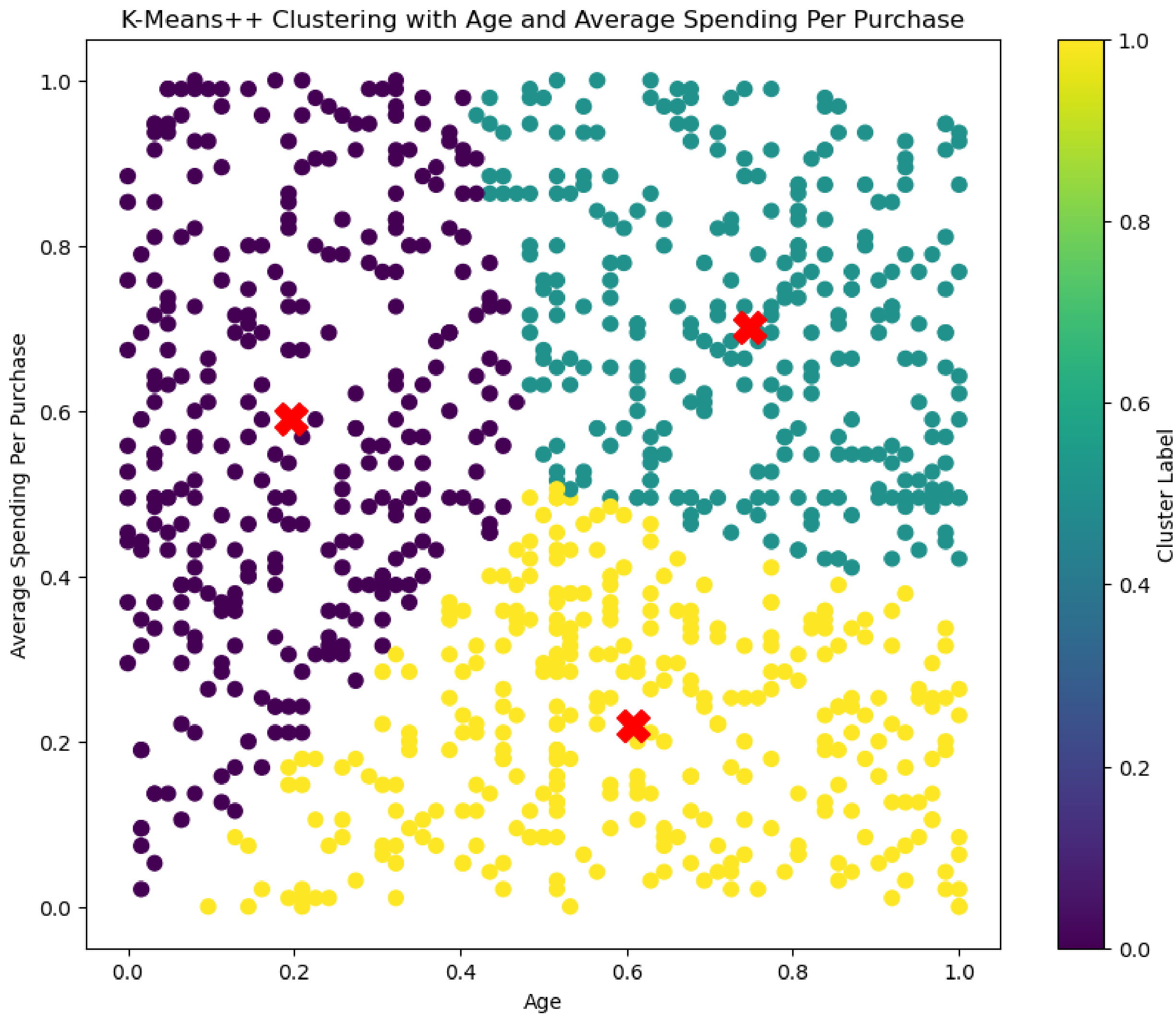
DBScan has not performed well in our dataset, the clusters are mixed and scatters and the eps and min values are difficult to optimize.

Analyzing Cluster Characteristics:

```
... Cluster 0 characteristics:
    Age                                      0.510312
    Income_Level                             0.500000
    Purchase_Amount                          0.492933
    Average_Spending_Per_Purchase            0.492385
    Purchase_Frequency_Per_Month             0.494681
    Brand_Affinity_Score                     0.485816
    Month                                    0.507544
    Year                                     0.509956
    Year_PD                                  0.519149
    Month_PM                                 0.526402
    Gender_Female                            0.330851
    Gender_Male                              0.308511
    Gender_Other                             0.360638
    Product_Category_Books                   0.302128
    Product_Category_Clothing                0.320213
    Product_Category_Electronics             0.377660
    Brand_Brand_A                            0.315957
    Brand_Brand_B                            0.326596
    Brand_Brand_C                            0.357447
    Product_Category_Preferences_High        0.320213
    Product_Category_Preferences_Low         0.345745
    Product_Category_Preferences_Medium      0.334043
    Extract_Date                             0.451950
    Extract_Month                            0.526402
    Extract_Year                             0.519149
    Cluster                                  2.127660
```

K-Means++ Clustering:





K-Means++ produced great results similar to K-Means. The clustering is almost the same.

Analyzing Cluster Characteristics:

```
Cluster 0 characteristics:
Age                                       0.508723
Income_Level                              0.528912
Purchase_Amount                           0.490629
Average_Spending_Per_Purchase             0.481239
Purchase_Frequency_Per_Month              0.830688
Brand_Affinity_Score                      0.383220
Month                                     0.523500
Year                                      0.506225
Year_PD                                   0.515873
Month_PM                                  0.519481
Gender_Female                             0.336735
Gender_Male                               0.326531
Gender_Other                              0.336735
Product_Category_Books                    0.289116
Product_Category_Clothing                 0.323129
Product_Category_Electronics              0.387755
Brand_Brand_A                             0.295918
Brand_Brand_B                             0.357143
Brand_Brand_C                             0.346939
Product_Category_Preferences_High         0.346939
Product_Category_Preferences_Low          0.336735
Product_Category_Preferences_Medium       0.316327
Extract_Date                              0.456576
Extract_Month                             0.519481
...
Extract_Year                              0.508718
Cluster_DBSCAN                            0.000000
Name: 2, dtype: float64
```

Module 4: Comparison and Conclusion:

Compare the results of all three clustering algorithms:

K-Means and K-Means++ have worked way better than DBSCAN for our data. The Clustering and Central Tendencies of the clusters are placed well to allow for a simple easy visual representation of the clusters made. The K-Means and K-Means++ have separated the customers based on their purchase behavior and preferences in 3 clusters.

```
Cluster 0 characteristics:
Age                                      0.508723
Income_Level                             0.528912
Purchase_Amount                          0.490629
Average_Spending_Per_Purchase            0.481239
Purchase_Frequency_Per_Month             0.830688
Brand_Affinity_Score                     0.383220
Month                                    0.523500
Year                                     0.506225
Year_PD                                  0.515873
Month_PM                                 0.519481
Gender_Female                            0.336735
Gender_Male                              0.326531
Gender_Other                             0.336735
Product_Category_Books                   0.289116
Product_Category_Clothing                0.323129
Product_Category_Electronics             0.387755
Brand_Brand_A                            0.295918
Brand_Brand_B                            0.357143
Brand_Brand_C                            0.346939
Product_Category_Preferences_High        0.346939
Product_Category_Preferences_Low         0.336735
Product_Category_Preferences_Medium      0.316327
Extract_Date                             0.456576
Extract_Month                            0.519481
...
Extract_Year                             0.508718
Cluster_DBSCAN                           0.000000
Name: 2, dtype: float64
```

The numbers are relatively same for both K-Means and K-Means++ while DBSCAN lacks behind.

K-Means/ K-Means++:

```
Silhouette Coefficient: -0.01
Calinski-Harabasz Score: 1.07
Davies-Bouldin Index: 41.62
```

The scores of K-Means and K-Means++ are better than the ones by DBScan

Draw conclusions and recommendations:

The evaluation suggests that K-Means is the superior clustering technique for this particular dataset. By examining measures such as the silhouette score, the Calinski-Harabasz index, and the Davies-Bouldin index, we have gauged the effectiveness of each algorithm in delineating the data's inherent structures.

Advantages and Disadvantages of Each Algorithm:

1. K-Means and K-Means++:

- These algorithms excel in forming distinct and meaningful clusters, making them highly suitable for discerning customer preferences and behaviors. Their main benefits include straightforward implementation, interpretability, and reliable results upon repeated applications.

2. DBSCAN:

- On the other hand, DBSCAN did not perform as well in segregating customers by their purchasing patterns. This could be due to its inherent design, which is highly sensitive to data point densities and can struggle with data that has variable densities.

Conclusion and Suggestions:

Given the analysis, it is advisable to adopt K-Means or K-Means++ for segmenting customers in the electronics domain. Their efficacy in yielding insightful and actionable data positions them as valuable tools for crafting targeted marketing initiatives and enhancing customer relations.