




DECEMBER 29, 2023

# CUSTOMER SEGMENTATION AND RETENTION STRATEGIES FOR IMTIAZ MALL'S ELECTRONICS SECTION: A COMPREHENSIVE DATA ANALYSIS APPROACH

UNVEILING PATTERNS, CLUSTERS, AND DATA-DRIVEN INSIGHTS

20F-0204 | 20F-0281 | 20F-0326



# Comparison and Conclusion Report

## Introduction:

This document provides a summary of the clustering analysis performed on Imtiaz Mall's customer data, with a focus on the electronics section. The objective was to identify distinct customer segments based on purchase behavior and preferences, enabling targeted marketing and improved customer retention strategies.

## Data Acquisition and Preprocessing:

First step to apply analysis to any kind of data is to clean and preprocess the data. Then for the categorical data we also need to encode it using some kind of encoding technique to convert it to numbers for further analysis. And last but no the least, we need to standardize the data for clustering. Following is the insight to all the cleaning, pre-processing, encoding and normalization techniques.

## Data Cleaning:

We have used the following techniques:

- **Drop empty rows:** We dropped the empty rows for column 'Product\_Category' because the analysis revolves around the 'Electronic category' and if the product category is not given then the data is vague.
- **Fill empty cells with column medians:** For rest of the columns on which we can apply statistical analysis, we fill the empty cell with the column medians.
- **Pandas F-fill method:** Lastly just to remove inconsistency from the data, we used fill-forward method for empty remaining cells.

Note: We have treated "Hidden" values as empty/Null values.

	Customer_ID	Age	Gender	Income_Level	...	Product_Category_Preferences	Month	Year	Season
0	b81ee6c9-2ae4-48a7-b283-220eaa244f43	40	Female	Medium	...	Low	01	2010	Winter
1	b81ee6c9-2ae4-48a7-b283-220eaa244f43	25	Male	High	...	Low	08	1989	Fall
2	fdf79bcd-5908-4c90-8501-570ffb5b7648	57	Other	Low	...	Low	08	1995	Winter
3	878dccba-893a-48f9-8d34-6ed394fa3c9c	38	Female	Medium	...	Low	09	2012	Fall
4	0af0bd81-73cc-494e-aa5e-75c6d0b6d743	68	Other	Medium	...	High	01	2010	Summer

[5 rows x 18 columns]

## Data-encoding:

In the initial stages of our data preprocessing, we applied the one-hot encoding method to transform categorical data into a numerical format. This technique is particularly valuable when dealing with variables such as gender, product categories, product category preferences and income levels, as it ensures that each category is represented as a binary column, effectively capturing the essence of these categorical features.

Unnamed: 0	Customer_ID	...	Product_Category_Preferences_Low	Product_Category_Preferences_Medium
0	b81ee6c9-2ae4-48a7-b283-220eaa244f43	...	1.0	0.0
1	b81ee6c9-2ae4-48a7-b283-220eaa244f43	...	1.0	0.0
2	fdf79bcd-5908-4c90-8501-570ffb5b7648	...	1.0	0.0
3	878dccba-893a-48f9-8d34-6ed394fa3c9c	...	1.0	0.0
4	0af0bd81-73cc-494e-aa5e-75c6d0b6d743	...	0.0	0.0

### Normalization:

To ensure consistency and comparability across different numeric features, we adopted the min-max normalization method. By scaling the values within a specific range, we aimed to prevent any particular feature from dominating others due to differences in magnitude. In this process, it set the maximum value to 1 and the minimum to 0, thereby transforming all values proportionally. This normalization step is crucial for enhancing the performance of clustering algorithms, such as K-Means and DBSCAN, by ensuring that each feature contributes equally to the analysis.

### Feature selection:

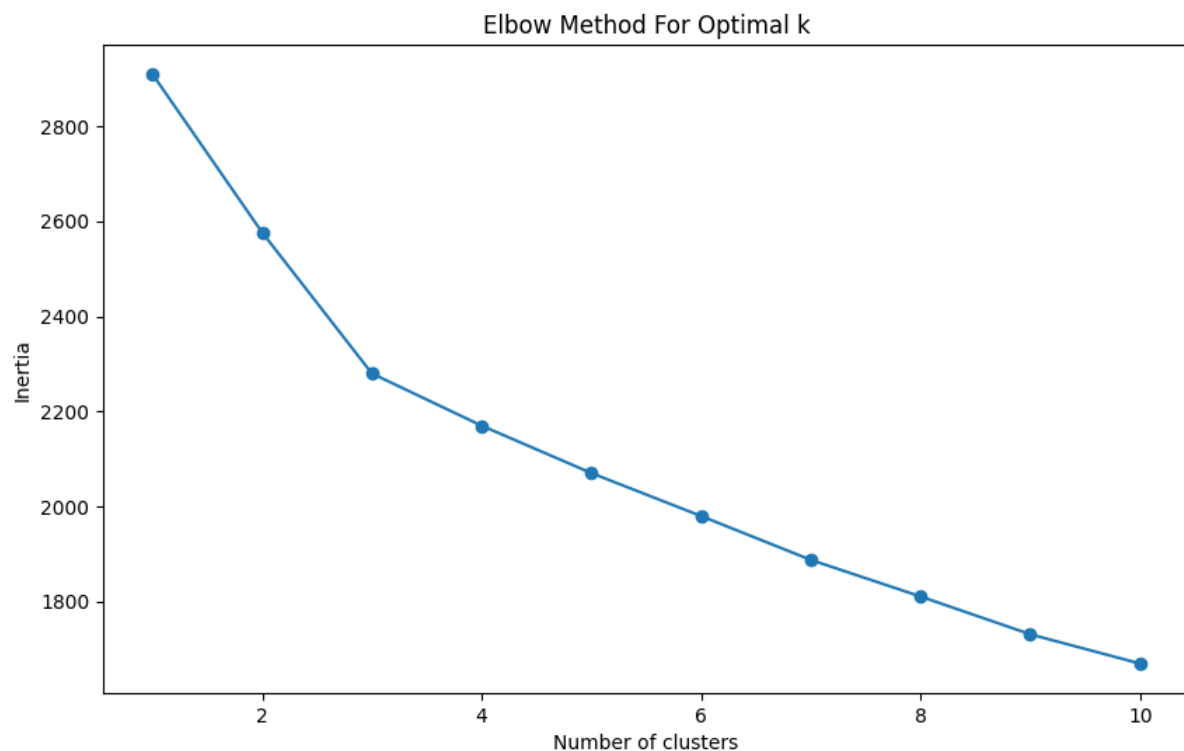
The consideration of feature selection played a pivotal role in optimizing the relevance and efficiency of our dataset. By conducting a comprehensive analysis of feature correlations, we identified and retained those attributes that contribute significantly to the overall understanding of customer behavior in the electronics section. By dropping non-numeric and non-relevant columns for clustering, we made sure that our clustering algorithms are provided with relevant information, ultimately leading to more meaningful and accurate segmentation of customer data.

# Comparison and Conclusion (Clustering Algorithms)

## 1. Comparison of Clustering Algorithms

K-Means, DBSCAN, and K-Means++

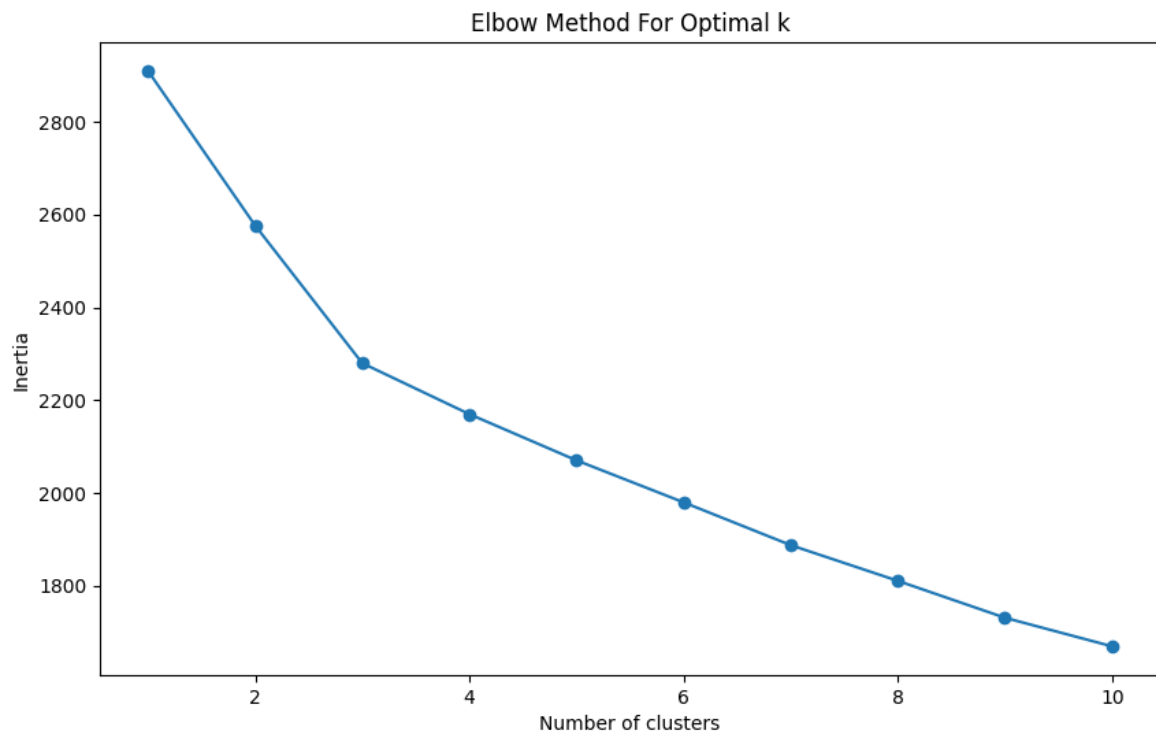
**K-Means** segmented customers into 5 distinct groups based on purchase behavior and preferences. The clusters showed clear differentiation in terms of purchasing patterns, brand affinity, and product preferences, making it suitable for broad market segmentation. We chose the number of clusters by using elbow method and cross test the values by using Silhouette Score and Visualization.



**DBSCAN**, with  $\text{eps}=0.5$  and  $\text{min\_samples}=1$ , created a highly fragmented clustering pattern, forming numerous small clusters. This approach highlighted very specific customer behaviors and outliers, useful for detailed pattern recognition. Epsilons and minimum samples are selected through trial and error with the effective use of Silhouette Score for better clustering.

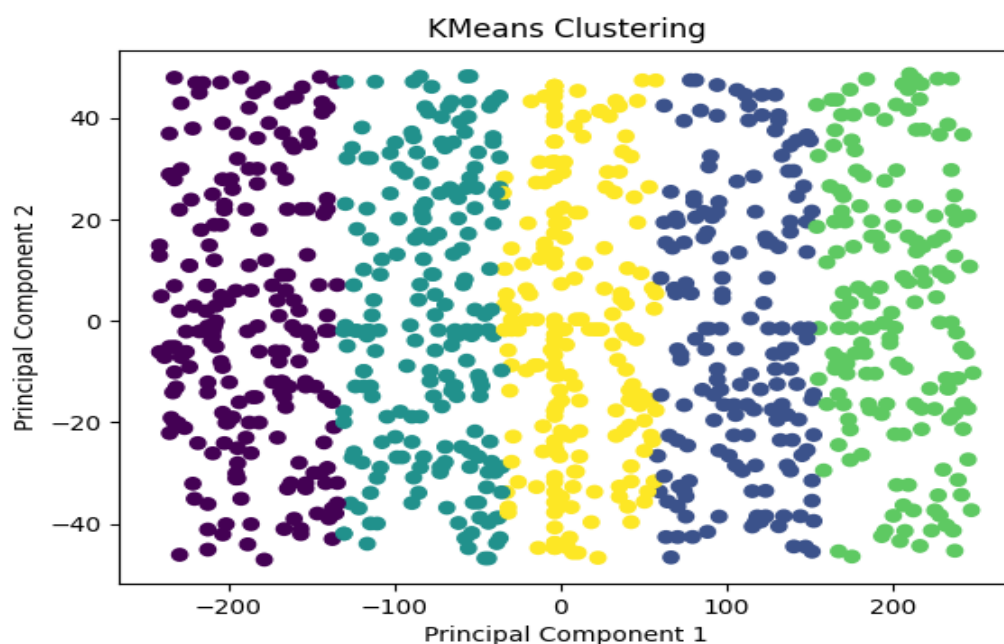
```
Sampling sample 0.1 1
DBSCAN Silhouette Score: 0.002007661628616272
Sampling sample 0.1 2
DBSCAN Silhouette Score: 0.015471143437235699
Sampling sample 0.3 1
DBSCAN Silhouette Score: 0.036288319845358165
Sampling sample 0.3 2
DBSCAN Silhouette Score: -0.36049709830409604
Sampling sample 0.5 1
DBSCAN Silhouette Score: 0.14639498839407505
Sampling sample 0.5 2
DBSCAN Silhouette Score: -0.07756016114895811
```

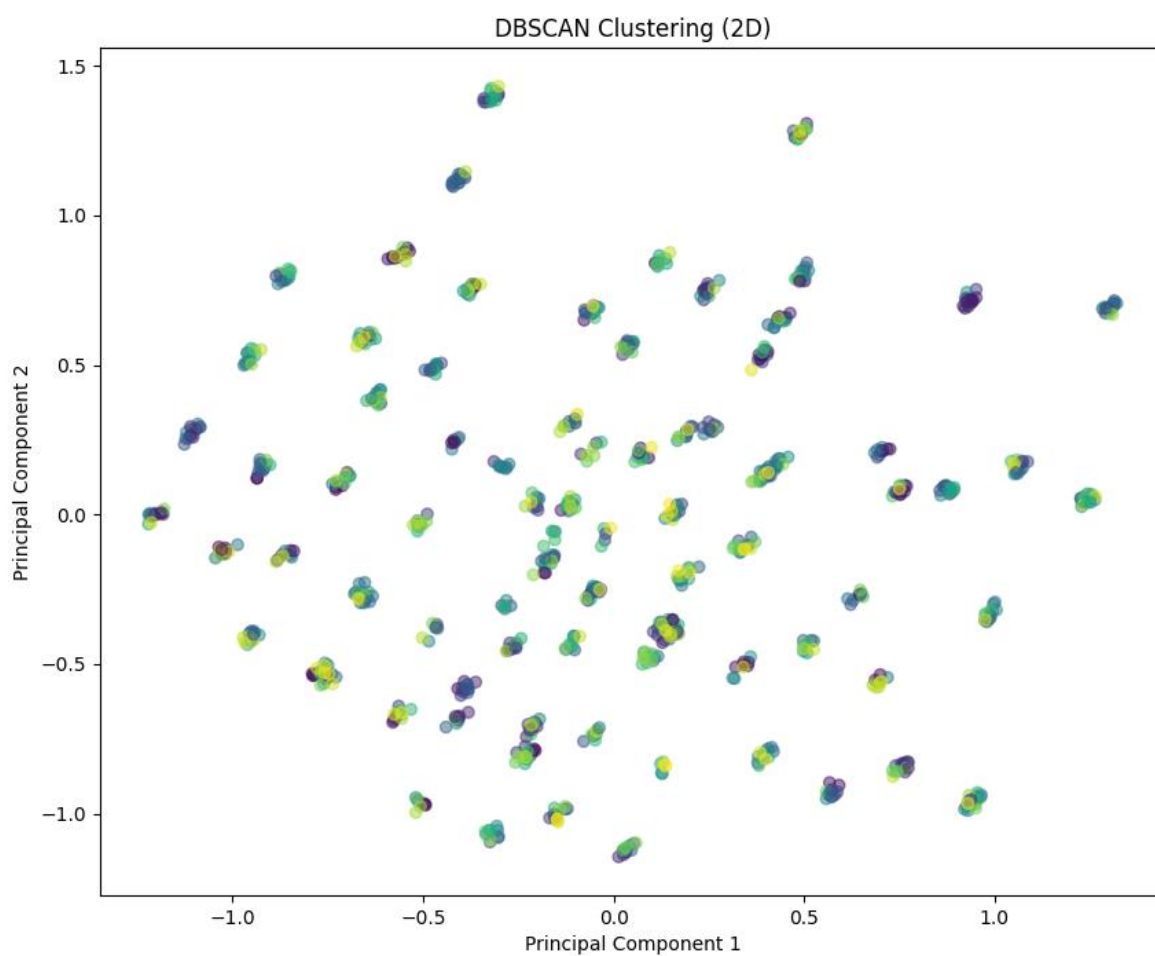
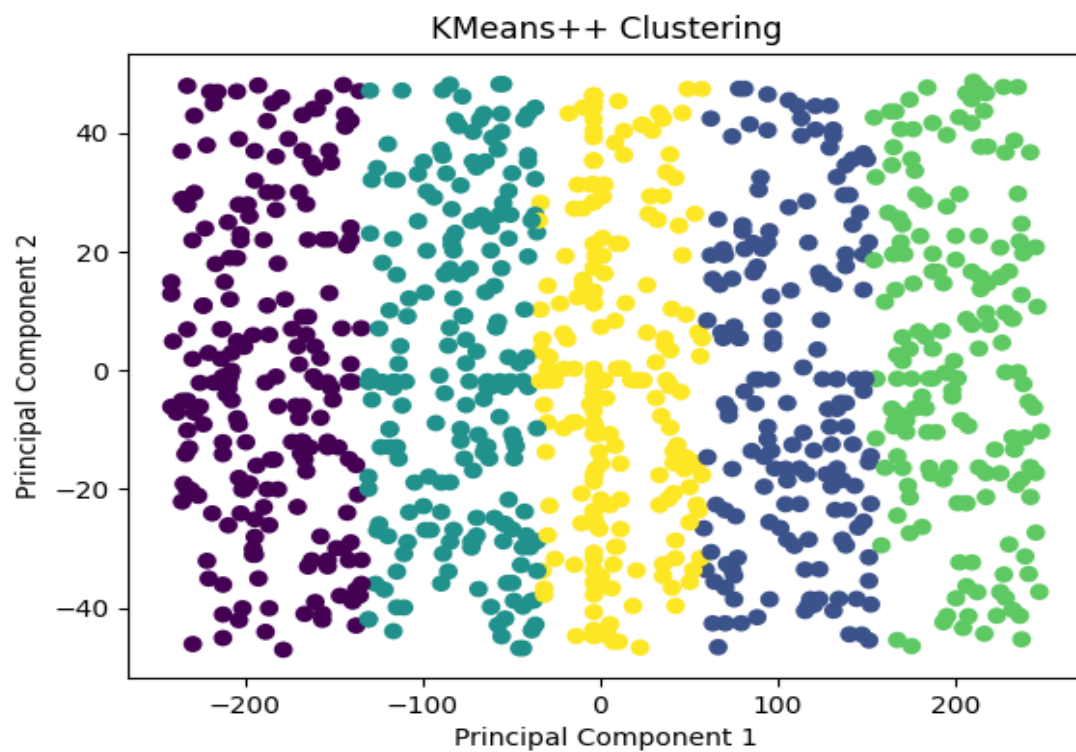
**K-Means++**, essentially an advanced version of K-Means with improved initialization, would generally provide similar segmentations to K-Means but potentially with more optimal initial cluster centers. It provides a better clustering. However, the results may not seem much different as it is because the traditional k-means method has converged into the k-means++



It is noteworthy that the default clustering algorithm in the scikit-learn library is K-Means++, chosen for its superior initialization of centroids. While K-Means can be accessed as an alternative by specifying the `'init'` parameter as `'random'`, our analysis primarily focused on the default K-Means++ configuration.

Here is the visual representation of clusters made through each algorithm.





Effectiveness:

#### K-Means

**Effectiveness:** K-Means was effective in creating distinct, broad customer segments. It partitioned customers into 5 clear groups based on their purchasing behaviors and preferences. This is particularly useful for identifying overarching market segments.

**Strengths:** The algorithm is straightforward and efficient, especially on large datasets like retail customer data. It's excellent for revealing general trends and dominant market segments.

**Limitations:** K-Means assumes clusters to be of similar size and spherical shape, which might not always be true in real-world data. It's also sensitive to outliers and the initial placement of centroids.

```
Cluster 0 characteristics:  
Average Purchase Amount: 59.21739130434783  
Average Purchase Frequency Per Month: 5.3478260869565215  
Average Brand Affinity Score: 5.396739130434782  
  
Cluster 1 characteristics:  
Average Purchase Amount: 160.79787234042553  
Average Purchase Frequency Per Month: 5.601063829787234  
Average Brand Affinity Score: 5.202127659574468  
  
Cluster 2 characteristics:  
Average Purchase Amount: 452.45238095238096  
Average Purchase Frequency Per Month: 5.755952380952381  
Average Brand Affinity Score: 5.553571428571429  
  
Cluster 3 characteristics:  
Average Purchase Amount: 250.98648648648648  
Average Purchase Frequency Per Month: 5.162162162162162  
Average Brand Affinity Score: 5.441441441441442  
  
Cluster 4 characteristics:  
Average Purchase Amount: 356.2247191011236  
Average Purchase Frequency Per Month: 5.477528089887641  
Average Brand Affinity Score: 5.269662921348314
```

#### DBSCAN

**Effectiveness:** DBSCAN's effectiveness in this context was mixed. It identified numerous small clusters, highlighting very specific customer behaviors or outliers. This is useful for detailed pattern recognition but less so for broad market segmentation.

**Strengths:** DBSCAN is excellent for identifying outliers and does not require pre-specifying the number of clusters. It can also find clusters of arbitrary shapes.



**Limitations:** The algorithm is highly sensitive to its parameters (eps and min\_samples). The choice of eps=0.5 and min\_samples=1 led to a large number of small clusters, which might be less actionable for broad marketing strategies.

DBSCAN Silhouette Score: 0.14859438893407383

	Cluster	Average_Purchase	Average_Frequency	Average_Brand_Affinity
0	0	0.465986	0.185185	0.148148
1	1	0.642857	0.111111	0.000000
2	2	0.381633	0.888889	0.000000
3	3	0.855102	0.666667	0.111111
4	4	0.095918	0.111111	0.555556
...	...	...	...	...
634	634	0.387755	0.777778	0.000000
635	635	0.432653	0.000000	1.000000
636	636	0.791837	0.777778	1.000000
637	637	0.238776	1.000000	0.777778
638	638	0.346939	0.111111	0.444444

[639 rows x 4 columns]

## K-Means++

**Effectiveness:** K-Means++ is designed to optimize the K-Means approach by improving the initialization of centroids. It would generally provide similar segmentations to K-Means but potentially more optimal in terms of initial cluster centers.

**Strengths:** It shares the strengths of K-Means while reducing the likelihood of poor cluster initialization.

**Limitations:** Similar to K-Means, it assumes spherical clusters and is sensitive to outliers.

	Cluster	oPurchase_Amount	oPurchase_Frequency_Per_Month	oBrand_Affinity_Score
0	0	64.261084	5.315271	5.413793
1	1	359.789157	5.397590	5.259036
2	2	172.979899	5.597990	5.261307
3	3	452.452381	5.755952	5.553571
4	4	260.098039	5.240196	5.382353

## Conclusion:

In terms of effectiveness, both K-Means and K-Means++ demonstrated clear and distinct delineation of individual customer segments. The clusters were well-defined, allowing for straightforward interpretation of purchase behaviour patterns. On the contrary, DBSCAN returned a more scattered result, proving less effective in identifying distinct customer segments.

## Evaluation Metrics

**Silhouette Score:** Useful for assessing the compactness and separation of clusters. Higher values indicate better-defined clusters.

**Calinski-Harabasz Index:** Higher values suggest better-defined clusters.

**Davies-Bouldin Index:** Lower values indicate better clustering.



We applied three metrics on both of our algorithms and following are the results.

For **K-Means++** we have

Clusters = 5

```
n_clusters = 5 Silhouette score: 0.1413659211214313
```

```
Calinski-Harabasz Index: 94.76219593907132
```

```
Davies-Bouldin Index: 1.998286246146558
```

For **DBScan** we have

eps= 0.5

min\_samples = 1

```
Sampling sample 0.5 1  
DBSCAN Silhouette Score: 0.14639498839407505
```

```
Calinski-Harabasz Index: 32.684947372431324
```

```
Davies-Bouldin Index: 0.4138604733029717
```

## Findings

It is clear from the above comparison that the preferable clustering algorithm is K-means.

During the evaluation, we considered metrics such as cluster silhouette score, Calinski-Harabasz score, and Davies-Bouldin index to assess the overall quality of clustering results. The comparison of these metrics helped in identifying the strengths and weaknesses of each algorithm in capturing underlying patterns within the data.

## Individual Advantages and Disadvantages

### I. K-Means and K-Means++

The K-Means and K-Means++ algorithms showcased notable advantages in terms of effectiveness and clarity in segmenting customers. These algorithms performed well in identifying cohesive clusters, providing valuable insights into customer preferences and behaviour. The strengths of these methods lie in their simplicity, ease of interpretation, and consistency across multiple runs.

### II. DBSCAN

In contrast, DBSCAN exhibited limitations in effectively segmenting customers based on their purchase behaviour. The algorithm's scattered results may be attributed to its sensitivity to the density of data points and the challenges posed by varying data densities.

## Recommendations

Based on our findings, we recommend leveraging K-Means or K-Means++ for customer segmentation within the electronics section. These algorithms have proven effective in providing actionable insights for targeted marketing strategies and personalized customer engagement.

## Draw conclusions and recommendations:

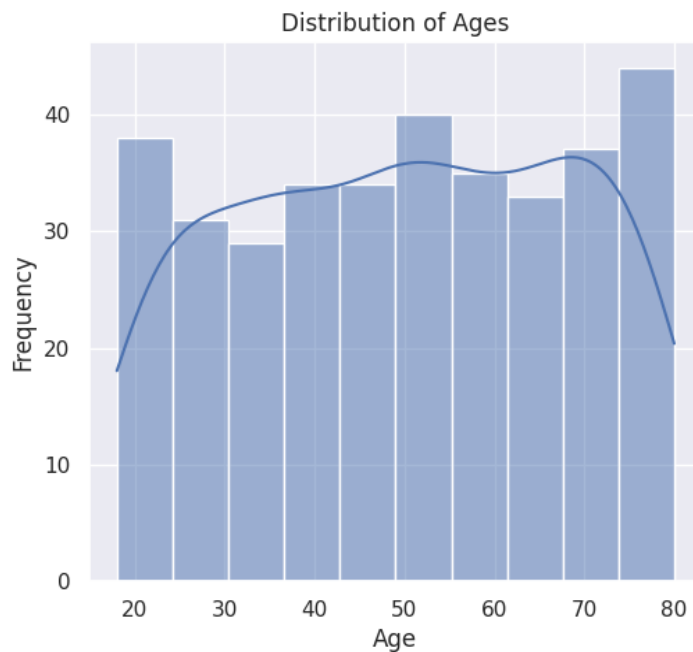
### Univariate Analysis:

Analyze the distribution of key features and identify potential outliers and skewness in data.

#### Based On Electronics

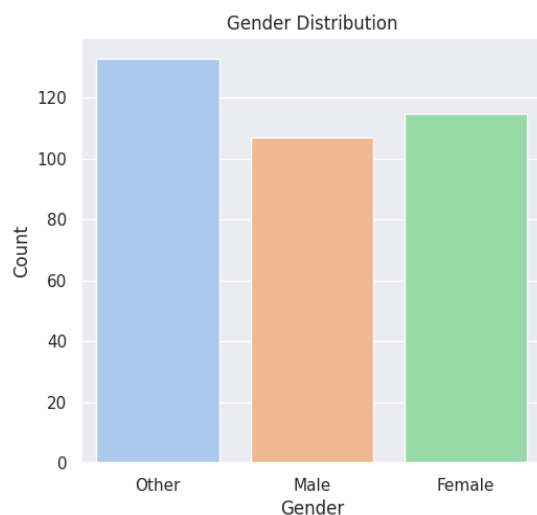
##### *Age Distribution:*

The age distribution for electronics purchases shows the range and concentration of ages among customers buying electronics. This can inform targeted marketing and product offerings. Store can design targeted market campaigns for particular age groups.



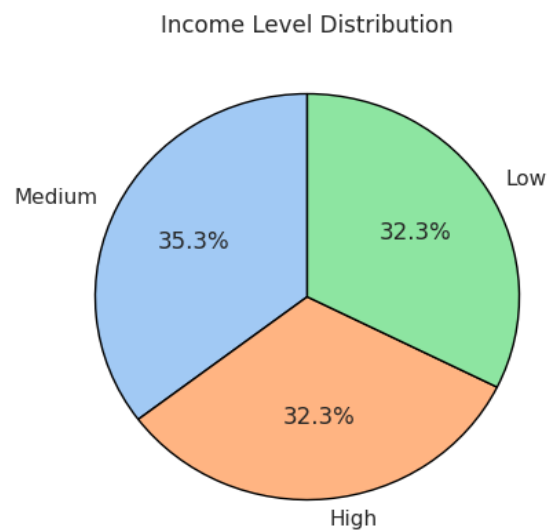
##### *Gender Roles:*

Gender analysis indicates no significant dominance in purchases between males and females in the electronics segment. Although females lead in comparison to males, the difference is not substantial. Plus, the notable thing is, there is no significant difference between males and females. This is a very interesting analysis as it is a general perception that males tends to purchase more electronics. This can leads to gender-specific marketing strategies.



### *Income Level Distribution:*

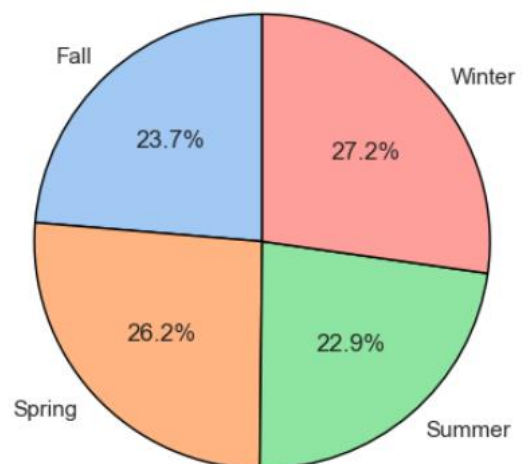
The distribution of income levels among electronics customers is relatively balanced, with similar percentages for low, medium, and high-income levels. What the strategists can do is, they can make promotional offers targeting a particular income level.



### *Seasonal Distribution:*

All seasons exhibit comparable purchase amounts in electronics, with a marginal edge for winter (27.2%).

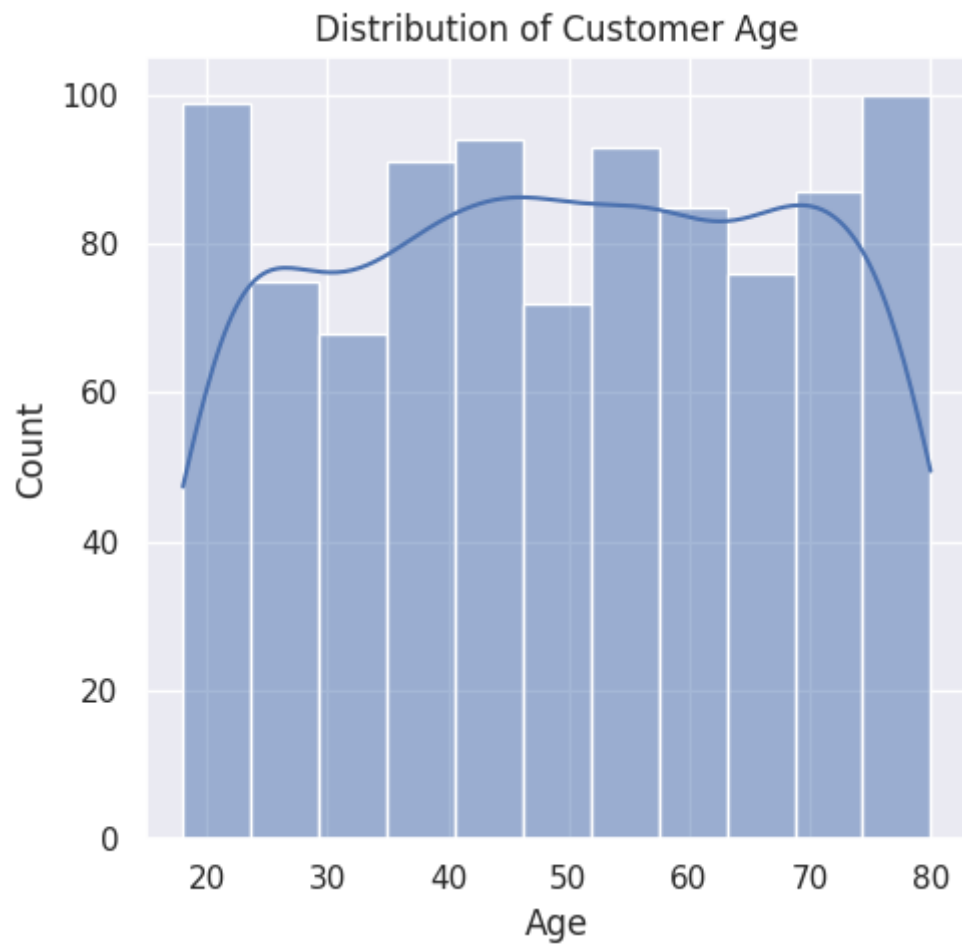
Seasonal Distribution of Purchase Amounts in Electronics



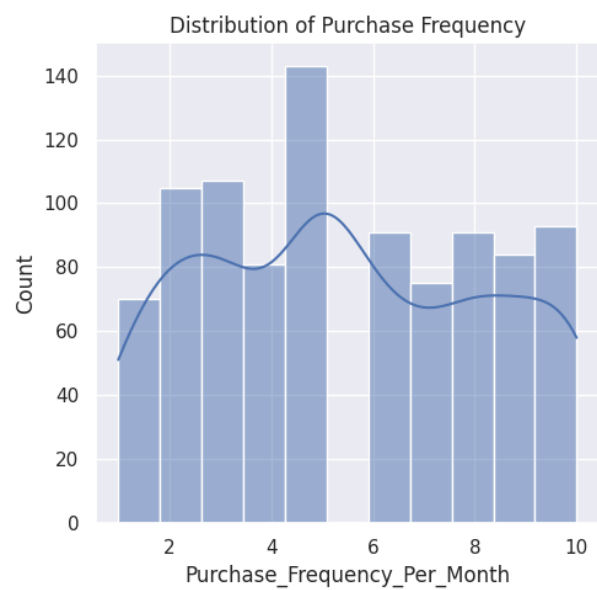
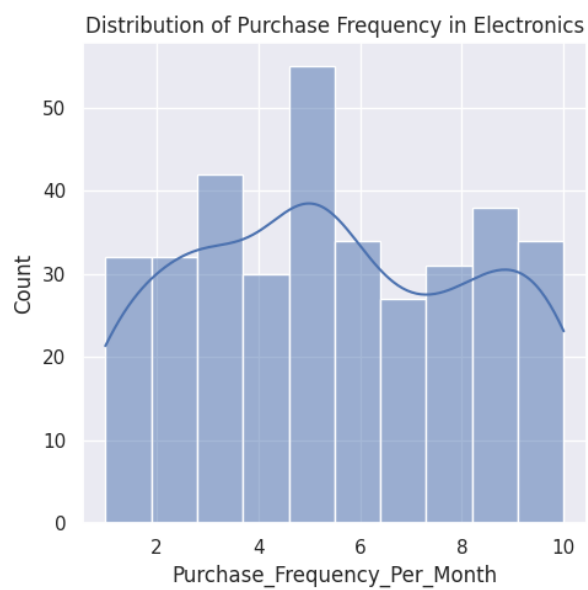
### *General Trends*

Understanding general customer behaviour on a broader scale is also very important if you want to attract or target a certain number of people. Here are few things can help you in this regards

We have studied the age distribution of customer for electronic purchaser but here is a look at overall age distribution of buyers that can be our particular customers. Compare this graph with electronics purchasers and then try to target people who are general buyers but don't buy electronics so much.



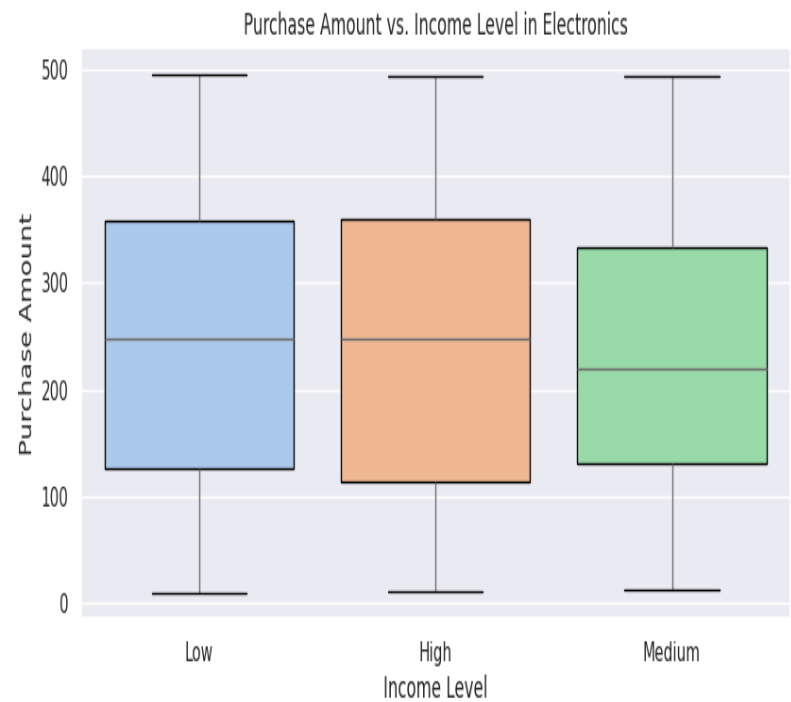
Similarly, this is the purchase frequency of electronics alone vs overall purchase frequency over the months. What we can do is, we can place sales or offers based on the low frequency months to encourage buyers to buy more.



Bivariant Analysis

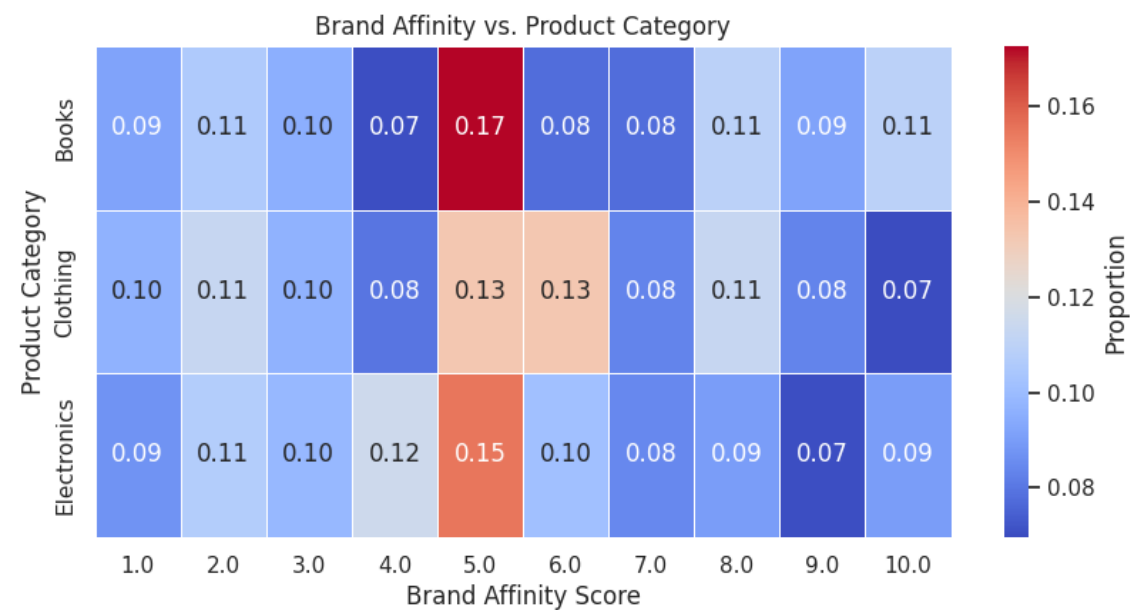
Purchase Amount vs. Income Level:

The boxplot shows variation in purchase amounts across different income levels. It appears that customers with higher income levels tend to have a wider range of purchase amounts, possibly indicating more significant spending.



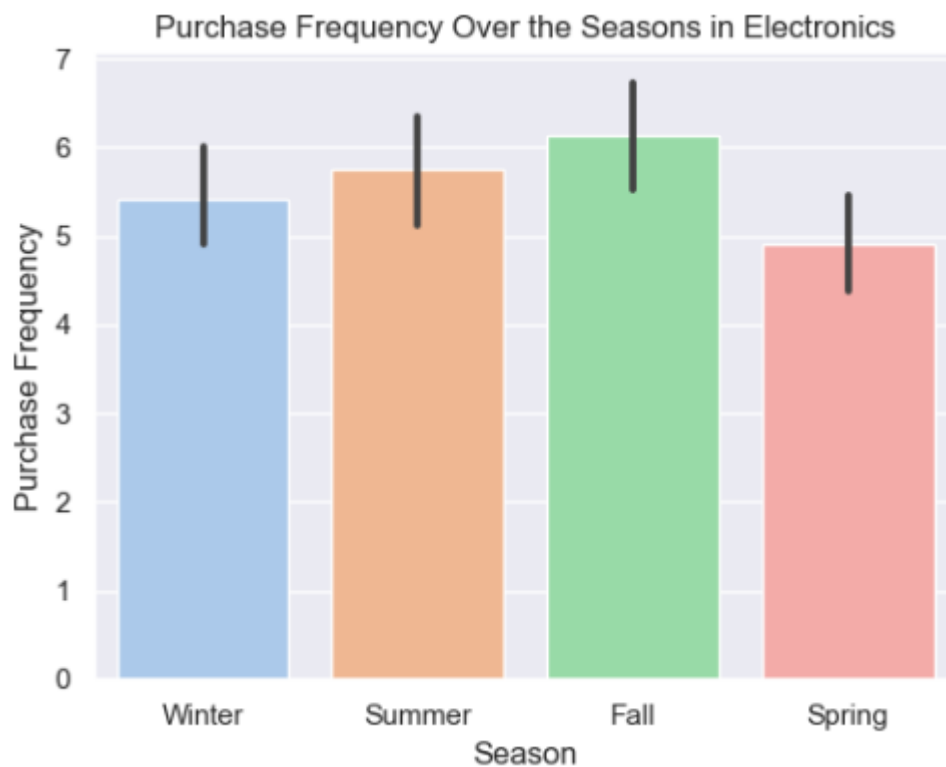
Brand Affinity vs. Product Category Preferences:

The heatmap suggests that electronics is the least liked product category in terms of brand affinity, with others being the more popular. There is a reason that only a handful of people are recurring customers and brand affinity score clearly show that.



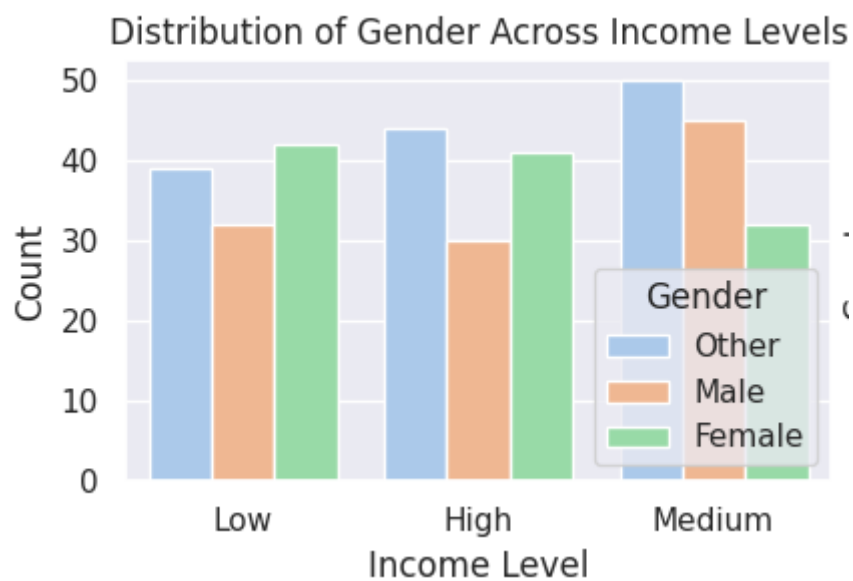
### Purchase Frequency Over Seasons:

The bar chart indicates that fall is the peak season for purchases in electronics, with spring being the least. Summer and winter exhibit comparable purchase frequencies.



### Gender Distribution Across Income Levels:

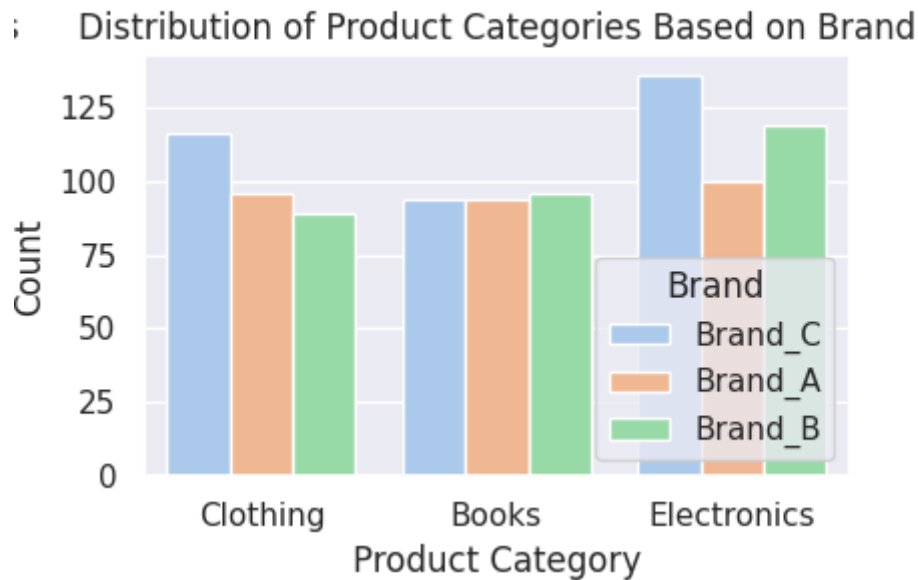
In the bar graph, medium-income males outweigh females significantly, while high-income levels show the opposite. Low incomes have comparable gender distributions.





Brands Based on Product Category:

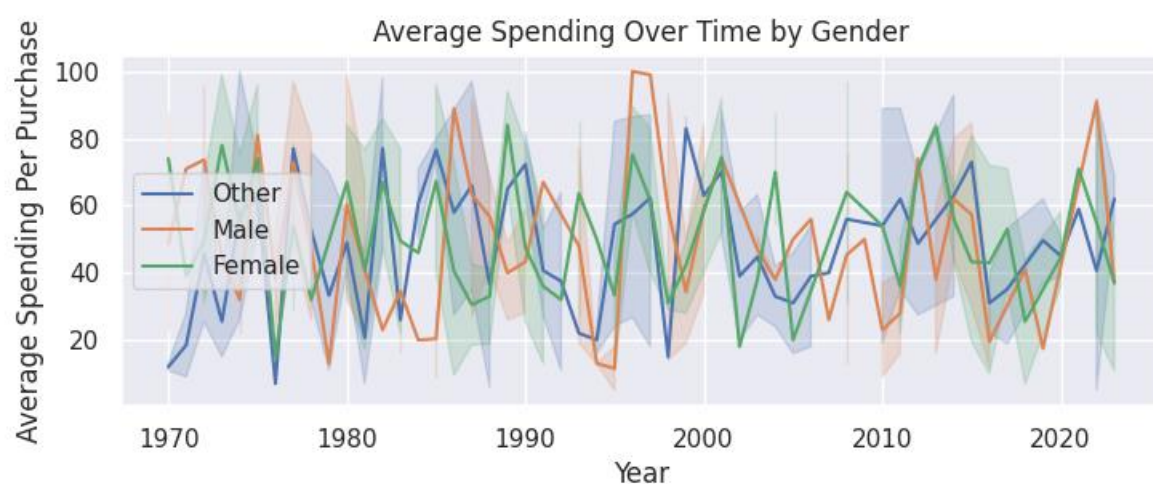
Brand\_C is the most popular in electronics, while Brand\_A is the least. Other brand categories are relatively comparable.



Temporal Analysis:

Average Spending Over Time by Gender:

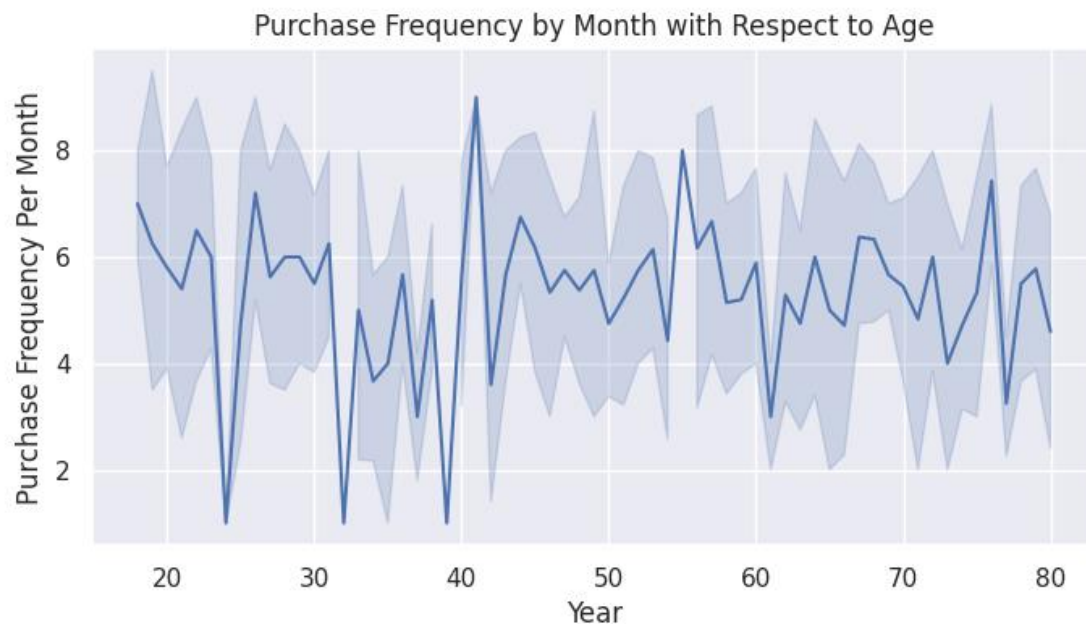
Males exhibit more fluctuation in purchases over time compared to females. Marketing strategies could play a role in observed spikes and drops. Also we can see that there is a sudden decrease in male electronics purchases after 2021. So the store should revise their policies they have made around 2021 regarding electronics.



Purchase Frequency by Month vs. Age:

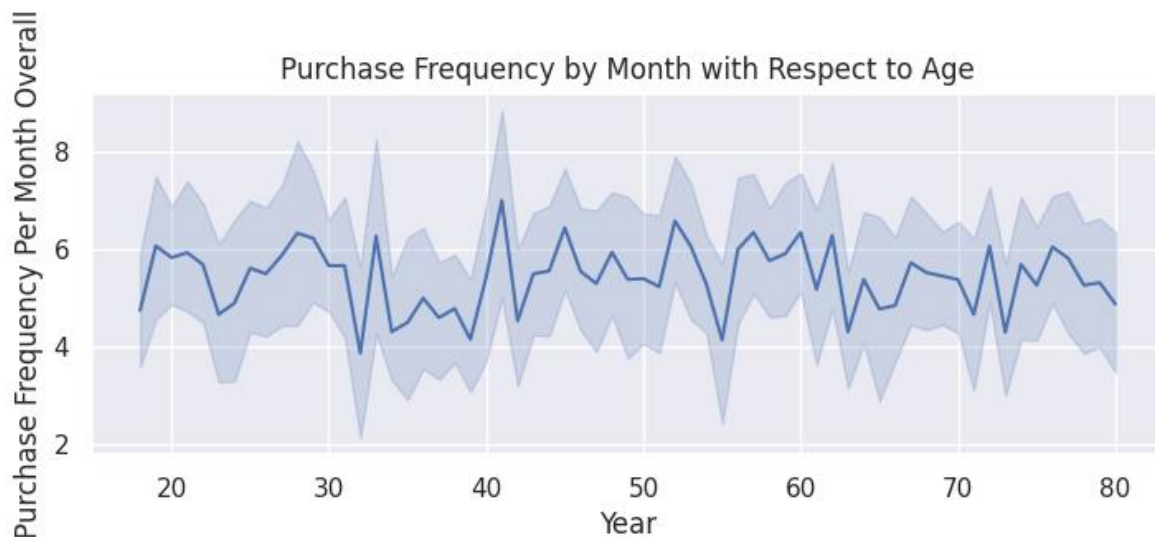
Below is the line chart of Purchase Frequencies Vs Age. The first one is for electronics section and the other one is overall. Individuals aged 40 and above consistently make more purchases than those

below 40. However, the notable thing is, the purchase frequency of age before 40 of electronics is significantly down, which is effecting the overall sales as well.

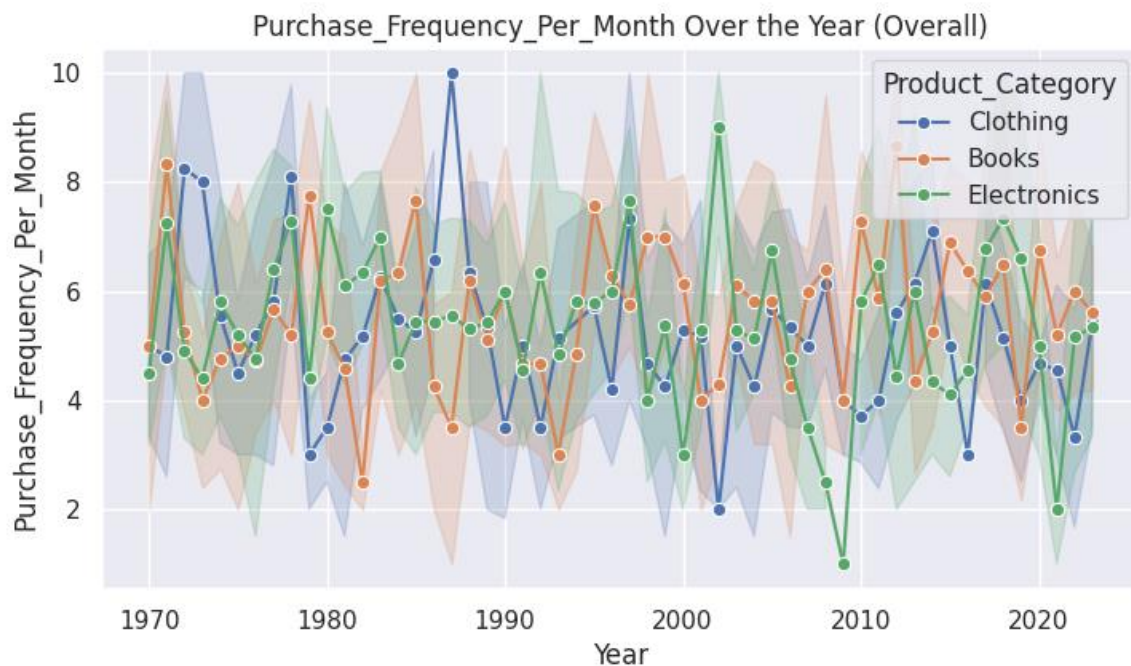


Purchase Frequency per Month Over Years:

Purchase frequency in electronics peaked between 2000 and 2005, decreased significantly in 2010, and continues to decline post-2020. What the store need to do is to reflect their decisions from 2000 to 2005, check what went wrong after 2005 and then also take a close look on the data after 2018.

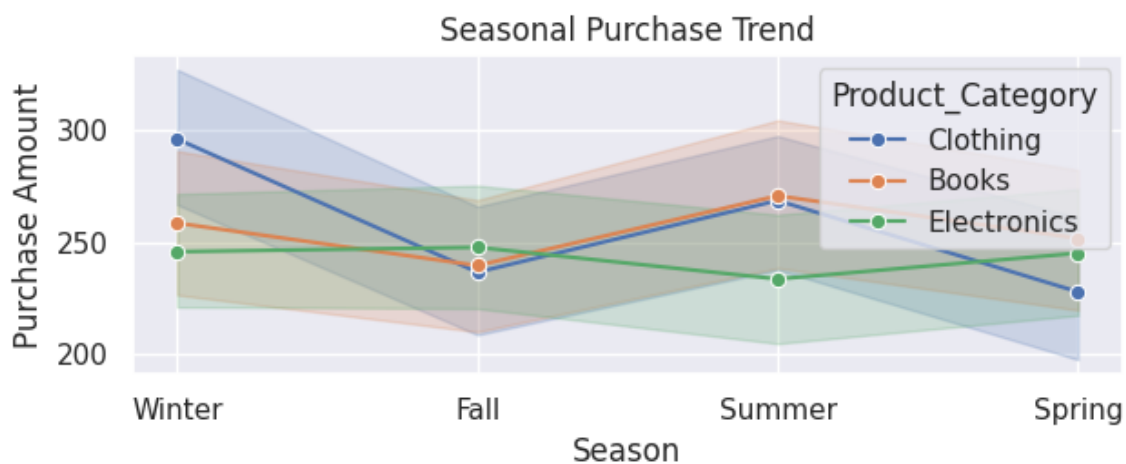


From the below given all category results, we can see from the indications that purchase frequency of others seems almost normal with a little fluctuation but with electronics, it's a constant rollercoaster ride.



#### Seasonal Purchase Trend:

The line chart indicates that winter is the peak season for electronics purchases, with spring selling the least. We can see that the overall sales are low in fall and spring. So we need to start making policies to attract customers in that season.

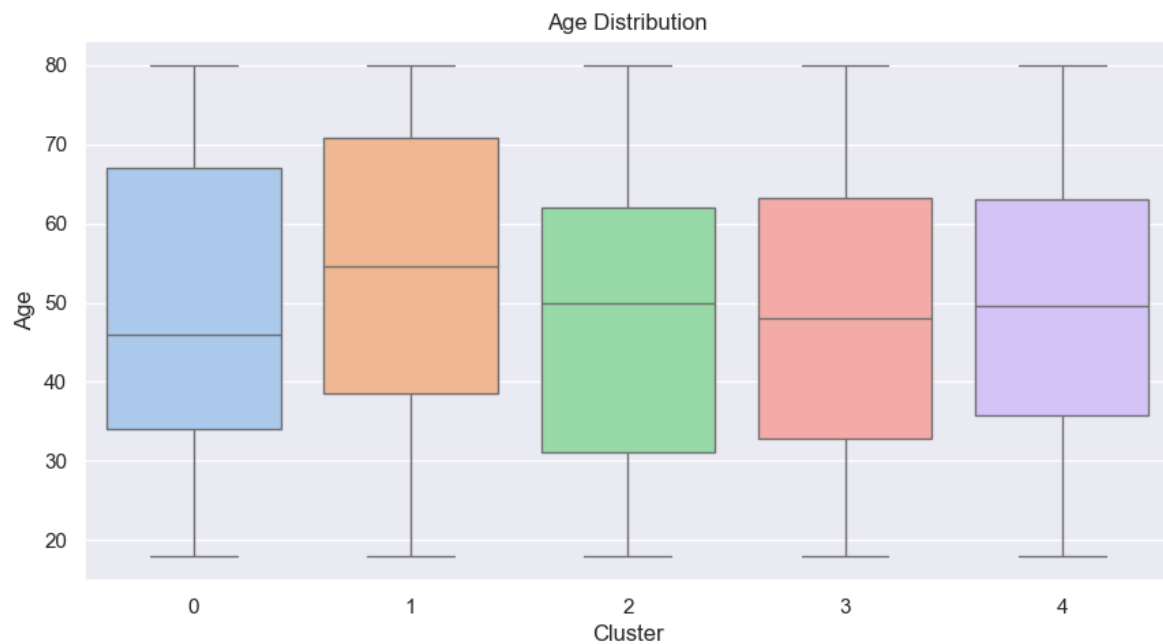


#### Cluster Based Analysis

Cluster based analysis gives us insights about the clusters, their purchasing trends, their characteristics, and trends. The following are the insights that we have found. This helps us in understanding how segments of customers behave. Through market segmentation we can target specific customer groups, to advertise and to attract them towards shopping in the electronics section.

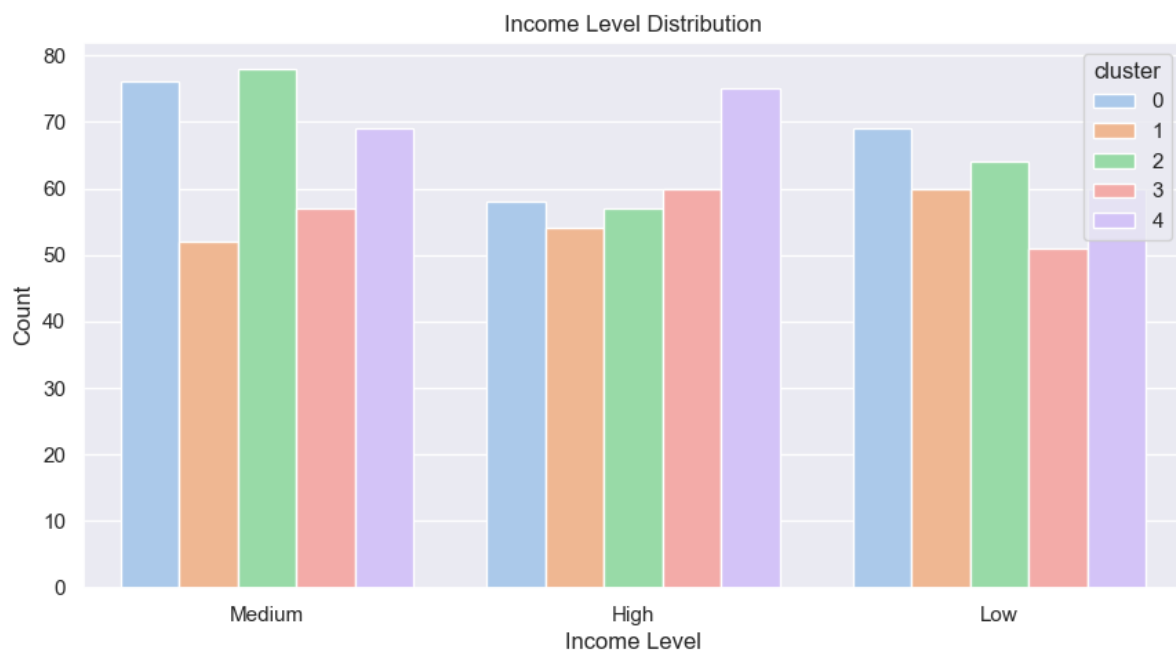
### Age Distribution Among Clusters:

The box plot graphs depicts the age distribution about the clusters, which would help us draft offers for the targeted clusters.



### Income Level Distribution:

This shows the income level distribution among the clusters, which could be another important thing to consider. The differences are shown as follows:



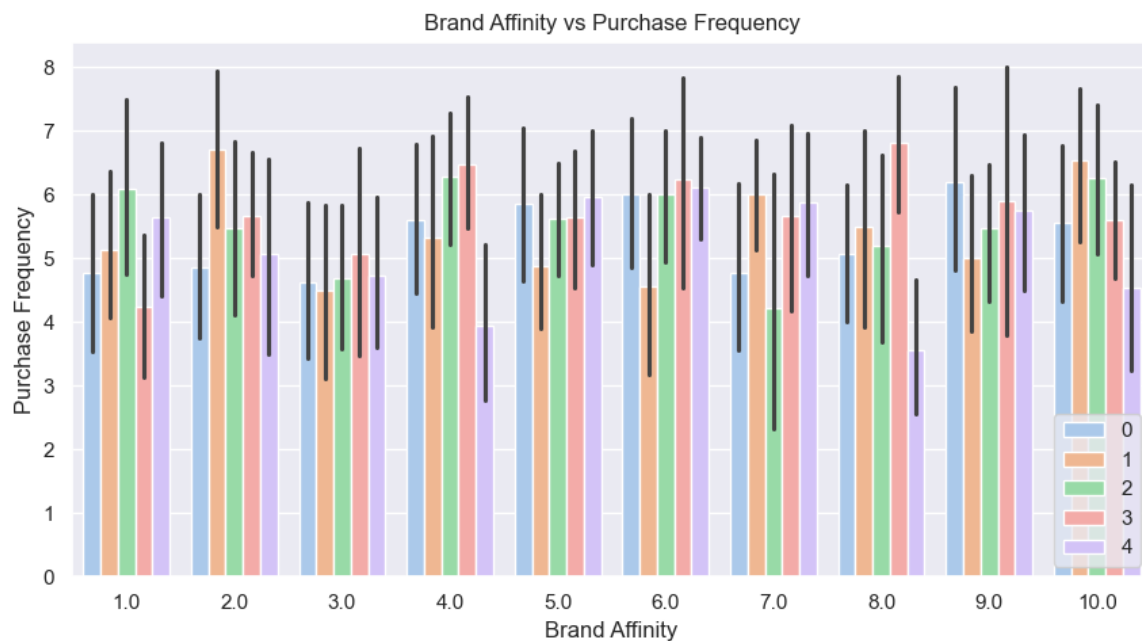
### Seasonal Purchase Trends:

This graph shows us the seasonal purchase trends of the different clusters. It would be an important thing to consider to multiple customer growth and help in customer retention.



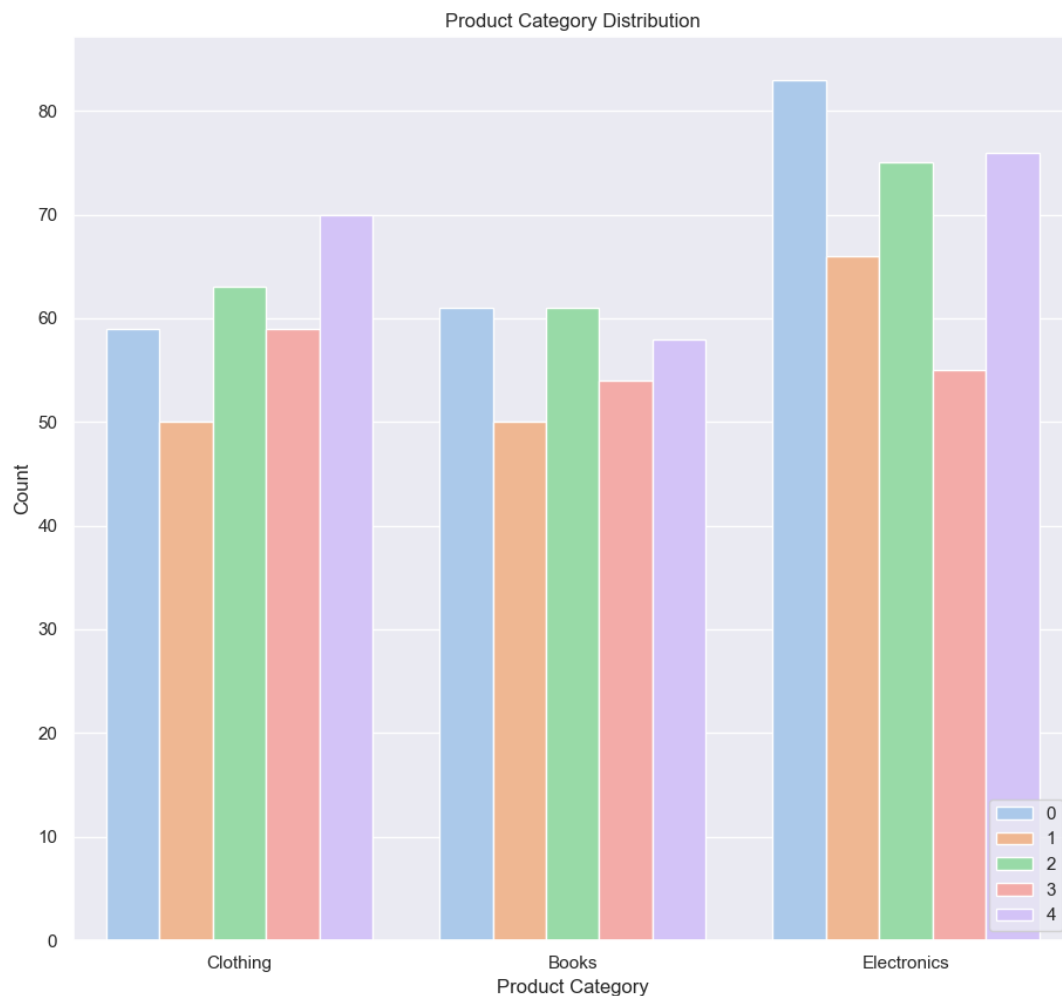
### Brand Affinity VS Purchase Frequency:

Another insightful thing is to watch the purchase frequency trends along with the brand affinity scores to depicts the likelihood of the customer review.



### Product Category Distribution:

Lastly, product category distribution is a very important insight to have. From this what we can do is to target clusters or groups with low purchases in the electronics category.



### Insights into Customer Segments:

#### Key Factors Differentiating Segments:

The age, gender, income level, and seasonal preferences contribute to the differentiation of customer segments. From the data we have seen how the season and months have significant impact on the sales. Similar, how people from different age groups, income level and genders tends to do purchasing.

#### Purchasing Behavior Patterns:

By digging into how people shop, we see that certain age groups dominate, gender doesn't strongly influence purchases, and spending habit stays almost consistent throughout all the income levels, though the spread is different. Similarly, season and months plays a big role and could be very crucial for targeting customers.

## Data-Driven Strategies:

### Customer Retention and Sales Growth:

To foster customer loyalty and boost sales in the electronics department, we leverage data to create personalized marketing strategies. This involves tailoring our approach to align with the preferences of different age groups, seasonal trends, and brand choices. Provide value to the new customer somehow which could results in customer retention.

### Potential Applications:

#### *Dynamic Pricing Strategies:*

Utilize data insights to implement dynamic pricing strategies. Adjust product prices based on seasonal demand, customer demographics, and brand popularity. This ensures competitive pricing and maximizes revenue during peak purchasing periods.

#### *Inventory Optimization:*

Employ data analytics to optimize inventory management specifically for the electronics section. By understanding the seasonal variations and popular product categories, we can fine-tune stock levels, reduce excess inventory, and ensure timely availability of in-demand products.

#### *Behavioral Email Campaigns:*

Implement targeted email campaigns based on customer behavior. Leverage insights into age-specific purchasing patterns, seasonal preferences, and brand affinity to craft personalized emails. These campaigns can include exclusive offers, product recommendations, and seasonal promotions tailored to individual customer segments.

#### *Social Media Engagement:*

Enhance social media strategies by aligning content with identified customer preferences. Share content related to popular electronics brands, seasonal promotions, and age-specific interests. Engage with customers through targeted social media campaigns to build a community around the electronics department.

## Further Analysis and Investigations:

- Investigate reasons for one-time visits, especially in high-value segments.
- Conduct time-series analysis to understand purchasing trends over time.
- Use advanced machine learning techniques for predictive modeling of customer behavior.
- Collect and analyze customer feedback specifically related to the electronics department. Use this information to address any pain points, enhance product offerings, and refine strategies based on direct insights from customers.

## Final Note

This analysis provides a foundation for understanding the diverse customer base at Imtiaz Mall. By leveraging these insights, the mall can enhance customer satisfaction, improve retention, and ultimately drive sales growth, especially in the electronics section.