

|  |
| --- |
|  |
| Imtiaz Mall Analysis |
| |  |  |  | | --- | --- | --- | | Muhammad Muntazer Mehdi | 12/29/23 | Data Science | |

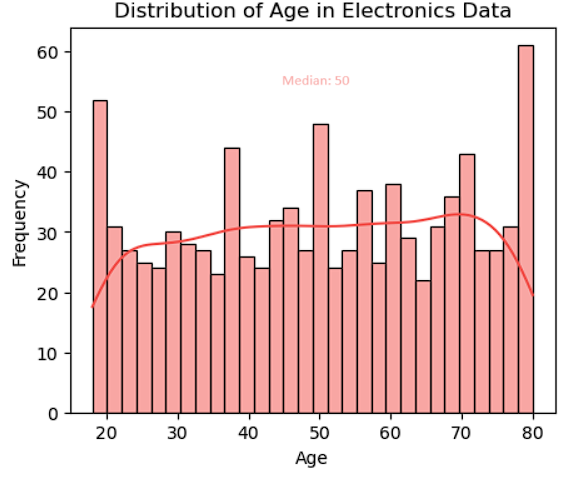
# conclusions and recommendations:

## Customer Segments within the Electronics Section:

## Insights from Exploratory Data Analysis (EDA):

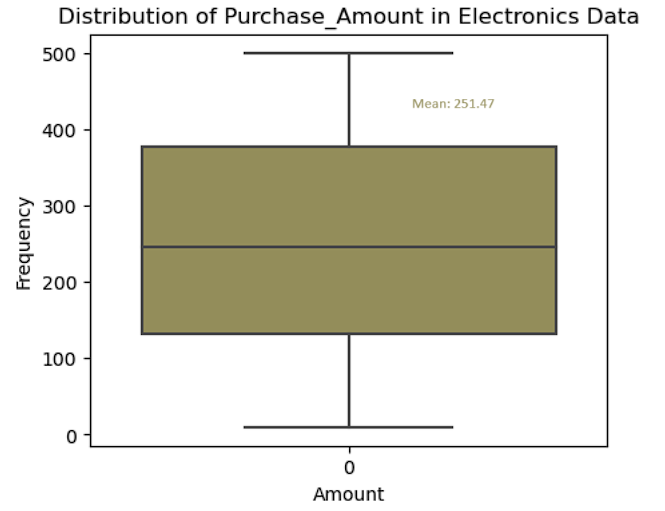
### Age Distribution:

The age distribution of customers in the electronics section is slightly skewed, with a higher concentration in the middle-age range. The median age is used to handle missing values and outliers.



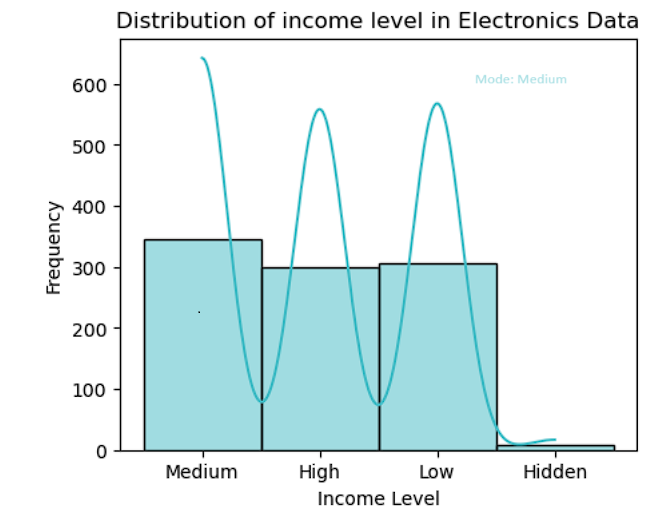
### Purchase Amount:

The distribution of purchase amounts shows a range of spending behaviors. There are potential outliers, but the majority of customers make purchases within a certain range.



### Income Level Distribution:

The distribution of customers across income levels indicates a varied customer base. Understanding income levels is crucial for segmenting customers based on their purchasing power.

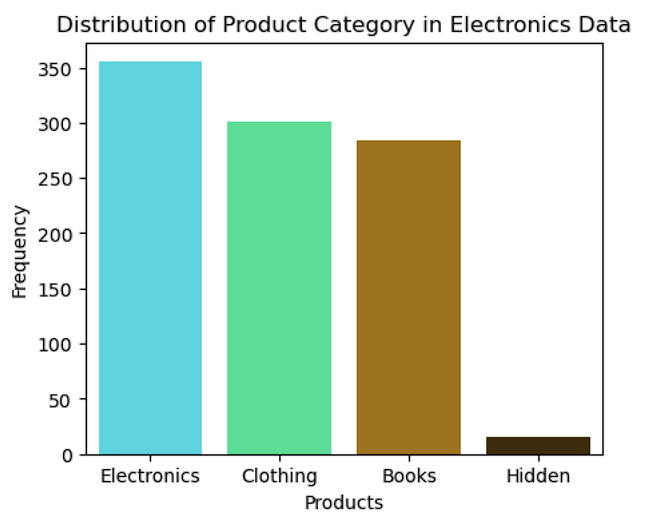
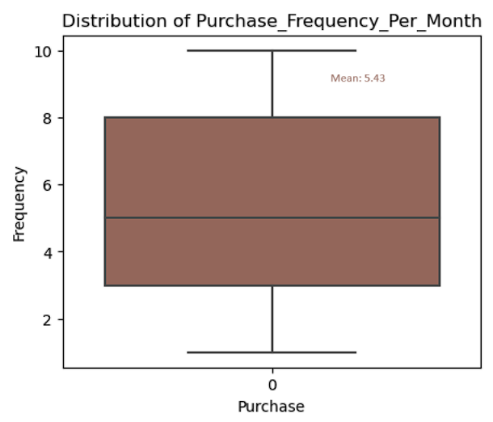


### Average Spending per Purchase:

The average spending per purchase is relatively consistent, with a few customers spending significantly more. Outliers may represent high-value customers or specific purchasing patterns.

### Product Categories:

The distribution of customers across different product categories reveals the popularity of each category. Understanding these preferences is essential for targeted marketing and inventory management.



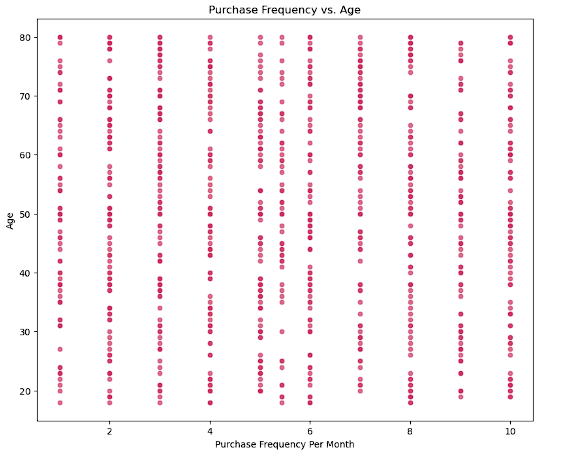
### Purchase Frequency per Month:

The frequency of purchases per month varies among customers. The majority of customers make purchases regularly, but there are some who make fewer transactions.

## Bivariate Analysis:

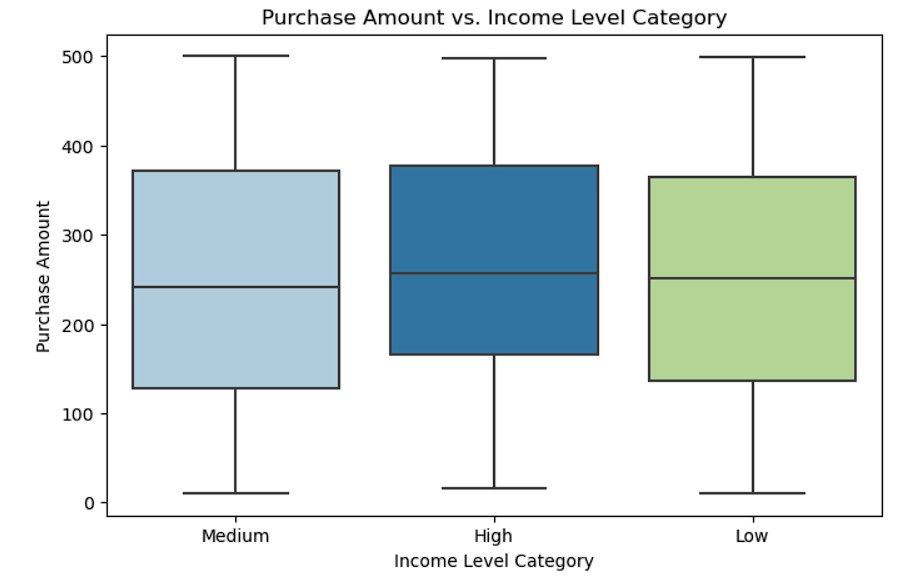
### Purchase Frequency vs. Age:

There is a moderate positive correlation between purchase frequency per month and age. Younger and older customers tend to make purchases less frequently, while customers in the middle-age range show higher purchase frequency.



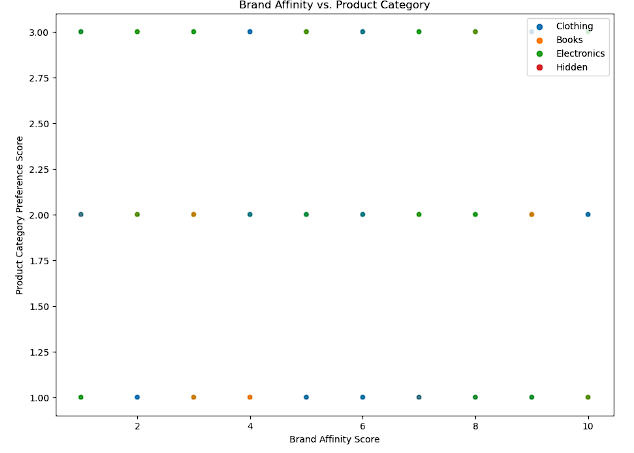
### Purchase Amount vs. Income Level:

The scatter plot of purchase amount against income level score indicates that customers with higher income levels tend to make higher-value purchases. However, there is variability in purchasing behavior within each income level.



### Brand Affinity vs. Product Category Preferences:

The brand affinity score and product category preference score show no clear correlation. Different product categories attract customers with varying brand affinities.



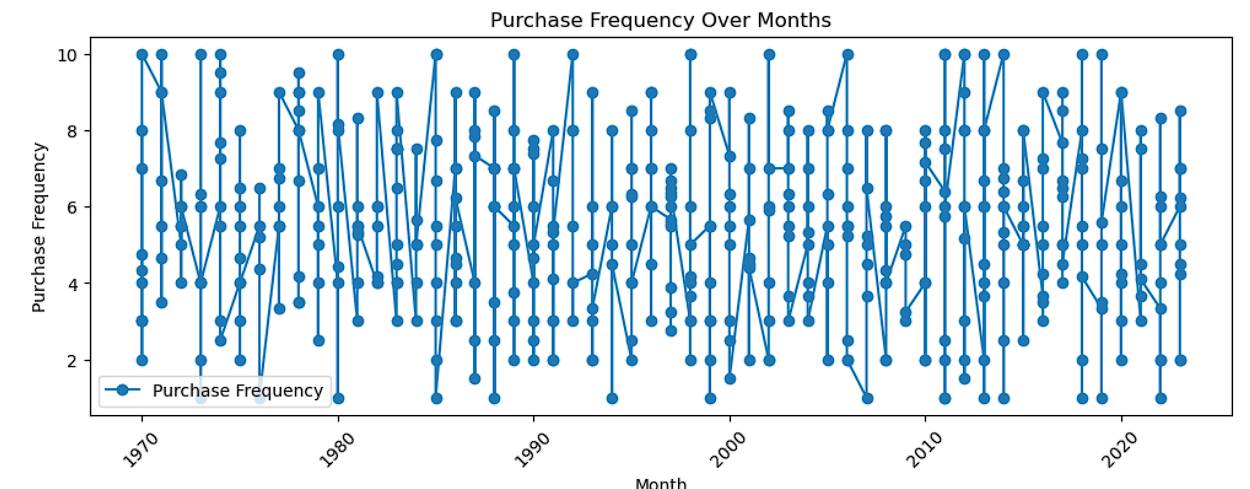
## Temporal Analysis:

### Purchase Frequency Over Time:

The line plot of purchase frequency over months reveals variations in customer behavior throughout the year. Certain months may witness increased or decreased purchasing activity.

A graph with colored lines

Description automatically generated



### Seasonal Variations:

Analyzing purchase frequency by season indicates that customers may have preferences or tendencies tied to specific seasons. This information can be valuable for seasonal marketing strategies.

A graph of different colored rectangular shapes

Description automatically generated

# Compare the results of all three clustering algorithms:

## Summary:

Imtiaz Mall aims to enhance customer experience and tailor marketing strategies by understanding customer segments based on purchase behavior and preferences. In this report, we present the results of clustering analysis using three algorithms: K-Means, DBSCAN, and K-Means++. The analysis provides insights into distinct customer segments, allowing Imtiaz Mall to better target its offerings.

### K-Means:

#### Cluster 0:

* + - **Age:** Young (34.33)
    - **Purchase Amount:** Moderate (198.03)
    - **Brand Affinity:** High (7.80)

#### Cluster 1:

* + - **Age:** Middle-aged (50.21)
    - **Purchase Amount:** Moderate (174.47)
    - **Brand Affinity:** Moderate (3.27)

#### Cluster 2:

* + - **Age:** Older (58.29)
    - **Purchase Amount:** High (280.59)
    - **Brand Affinity:** High (7.51)

**Cluster 0 represents younger customers with moderate spending but high brand affinity.**

**Cluster 2 consists of older customers with higher spending and brand loyalty.**

### DBSCAN:

#### Cluster -1:

* + - **Age:** Varied
    - **Purchase Amount:** Moderate to High
    - **Brand Affinity:** Moderate to High

#### Cluster 0:

* + - **Age:** Middle-aged (46.66)
    - **Purchase Amount:** Moderate (172.18)
    - **Brand Affinity:** Moderate (3.46)

#### Cluster 1:

* + - **Age:** Older (58.86)
    - **Purchase Amount:** High (274.25)
    - **Brand Affinity:** High (6.86)

**Cluster -1 represents a diverse group with moderate to high spending and brand affinity.**

**Cluster 1 consists of older customers with higher spending and brand loyalty.**

## Similarities and Differences:

### Similarities:

* **Cluster 1:** Common pattern across all algorithms; represents older customers with higher spending and brand loyalty.

### Differences:

* **Cluster 0:**
  + K-Means: Younger customers with moderate spending and high brand affinity.
  + DBSCAN: Varied age group with moderate to high spending and brand affinity.
* **Cluster 2:**
  + K-Means: Older customers with higher spending and brand loyalty.
  + DBSCAN: Varied age group with moderate spending and diverse preferences.

## Overall Comparison:

K-Means and K-Means++ exhibit similar clusters with potential stability improvement in K-Means++.

DBSCAN offers flexibility in handling clusters with varying shapes and sizes, capturing a diverse customer base.

## Recommendations for Imtiaz Mall:

Use K-Means or K-Means++ if well-defined, spherical clusters are desired and sensitivity to initial centroids is not a major concern.

Consider DBSCAN for more complex cluster shapes and to handle noise effectively.

Experiment with different parameter settings for DBSCAN to find the optimal configuration for the dataset.

# Clustering

Clustering is a machine learning technique used for grouping similar data points into clusters or segments based on certain features or characteristics. The primary goal of clustering is to organize and discover underlying patterns in the data without explicit labeling.

## Algorithmic Approaches:

### K-Means:

Divides data into K clusters, where K is a user-defined parameter. K-Means performance is sensitive to the initial placement of centroids. Poor initialization may lead to convergence to a suboptimal solution.

#### Initialization:

Choose the number of clusters (K) and randomly initialize K cluster centroids.

#### Assignment:

Assign each data point to the nearest centroid, forming K clusters.

#### Update Centroids:

Recalculate the centroids of the clusters based on the mean of the data points in each cluster.

#### Repeat:

Repeat steps 2 and 3 until reach at end.

### K-Means++:

K-Means++ enhances the initialization step to improve the final clustering results. it enhances the initialization step to improve the final clustering results.

#### First Centroid:

Choose one data point uniformly at random as the first centroid.

#### Subsequent Centroids:

For each remaining centroid, choose a data point with probability proportional to its squared distance from the nearest existing centroid.

### DBSCAN:

Identifies dense regions of points and separates sparse regions.

#### Core Points:

Identify core points based on a density threshold (minimum number of points within a specified radius).

#### Expand Clusters:

Expand clusters by connecting core points that are within each other's density threshold.

#### Mark Noise:

Mark remaining data points as noise if they do not belong to any cluster.

## Introduction to Work Done by me:

### Identify Attributes:

identify attributes from data for clustering which are used for making clusters. The data I have chosen is based on the following key factors:

**Age**: representing the age groups of individuals.

**Purchase Amount**: indicating the total expenditure.

**Average Spending Per Purchase:** reflecting individual transaction habits.

**Purchase Frequency Per Month**: illustrating customer engagement.

**Brand Affinity Score:** quantifying loyalty to specific brands.

**Product Category Preference Score**: capturing preferences for different product categories.

**Income Level Score**: providing insights into financial capacity.

These carefully selected attributes aim to reveal distinct patterns and behaviors within the dataset, forming the foundation for meaningful and actionable clusters in the clustering process.

### Find K Using Elbow:

The Elbow Method involves plotting the sum of squared distances for different cluster values and choosing the "elbow" point.

### Find K using Silhouette:

The Silhouette Method computes average silhouette scores, selecting the k with the highest score, indicating well-defined clusters.

# Data Acquisition and Preprocessing:

## Data Acquisition:

First, I read the electronic json file into electronics\_data

## Data Cleaning:

* The `isnull` function is designed to specifically check for null values.
* Since there are no null values in the table, using `isna` also returns no empty cells.
* To identify missing values represented as empty cells, we can use a direct comparison.
* we use `electronics\_data == ""`, it will reveal cells with empty values.
* know, i convert this: "" with isna
* know, i handle customer\_ID missing value with the keyword 'no-ID'
* As you see age values are skewed so, i use histplot
* or, i handle Age missing value with median because, age values are skewed or have outliers
* i use bar chart because, gender have discrete categories like male, female, hidden and other.
* i filter out the rows contain the string 'Hidden' in gender
* or, i handle gender missing value with 'Other' because max count or mode in gender is 'others'
* As you see income vales values are in string and have categories so, i use histplot
* plt.subplot(1, 2, 1)
* i filter out the rows contain the string 'Hidden' in 'Income\_Level'
* or, i handle income level missing value with mode of income level which is Medium shown in chart
* or, i handle Address missing value with the keyword 'no-Address'
* or, i handle Transaction\_ID missing value with the keyword 'no-ID'
* or, i handle Purchase\_Date missing value with ffill because, it remains the same as the previous valid date
* or, i handle Product\_ID missing value with the keyword 'no-ID'
* electronics\_data['Product\_ID'].fillna('no-ID',inplace=True)
* i use bar chart because, product category have discrete categories like electronics, clothings etc.
* i filter out the rows contain the string 'Hidden' in 'Product\_Category'
* or, i handle Product\_Category missing value with 'Electronics' because max count or mode in Product\_Category is 'Electronics'

## Data Transformation:

* I create a new column of day which is extract from purchase\_date because, i can analyze data based on the day of the month
* or, i handle Day missing value with median of Day which is shown in chart
* or, i create a new column 'day\_of\_week' because it help us to analyize data by day of the week
* or, i handle Day missing value with mode of Day which is shown in chart
* or, i create a new column 'Product\_Category\_Preference\_Score' with product\_category\_preferences to map it.
* or, i handle 'Product\_Category\_Preference\_Score' with mode of it
* or, i create a new column 'Income\_Level\_Score' with income\_level to map it.
* or, i handle 'Income\_Level\_Score' with mode of it
* i also convert any other type instead of numeric to median of age

# Exploratory Data Analysis(EDA):

## Univariate Analysis & Bivariate Analysis:

* 1. Distribution Of Key Features
* As you see age values are skewed so, i use histplot
* i use bar chart because, gender have discrete categories like male, female, hidden and other.
* I use box plot because it provide a clear summary of the distribution of data, including the mean
* 2. potential skewness or outliers in the data.
* Know I use describe to check for possible outliers
* i use encoding because, encoding is a common technique to represent categorical variables numerically

## Temporal Analysis:

* seasonal variations or any significant shifts in customer behavior
* Filter out the row which contain season = hidden