

**Group Members**

Waqas Dost 20F-0333

Junaid Hussain 20F-0338

Talha 20F-0248

**Numeric Columns and Distributions**

### Numeric Columns

### The following numeric columns have been considered for analysis:

### Age

### Purchase\_Amount

### Average\_Spending\_Per\_Purchase

### Purchase\_Frequency\_Per\_Month

### Brand\_Affinity\_Score

### Results

### Age Distribution:

### The age distribution appears to be relatively uniform, with a slight peak in the middle age range.

### Purchase Amount Distribution:

### The purchase amount distribution shows a right-skewed pattern, indicating a concentration of lower purchase amounts.

### Average Spending Per Purchase:

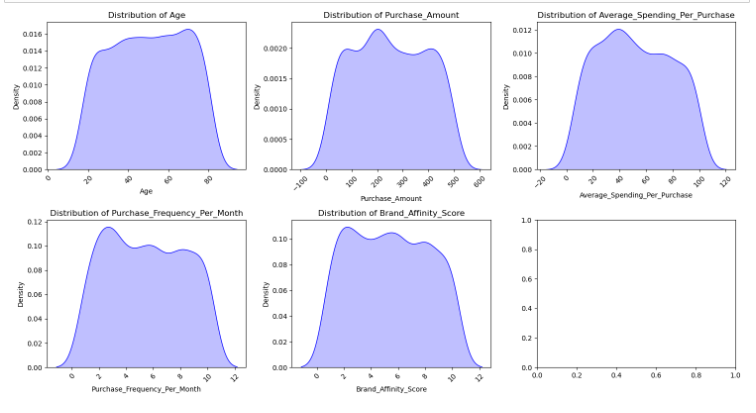
### The distribution of average spending per purchase suggests a concentration towards lower values, but with some variability.

### Purchase Frequency Per Month:

### The purchase frequency per month distribution exhibits multiple peaks, suggesting different purchasing patterns among the dataset.

### Brand Affinity Score:

### The distribution of brand affinity scores appears to be relatively uniform, with some concentration towards specific score ranges.



### Analyze outliers and determine whether to retain or remove them based on their impact on the analysis

**Outlier Detection and Visualization:**

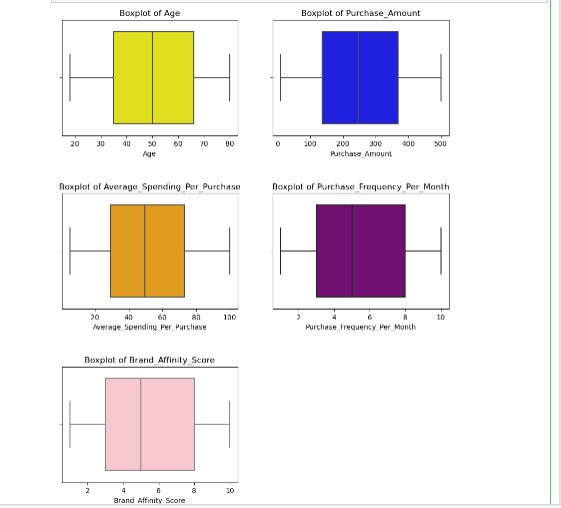
**Numeric Columns**

The following numeric columns are considered for outlier detection:

* Age
* Purchase\_Amount
* Average\_Spending\_Per\_Purchase
* Purchase\_Frequency\_Per\_Month
* Brand\_Affinity\_Score

**Results and Observations**

1. **Age:**
   * The boxplot for Age does not show any outliers, with the data points falling within the whiskers.
2. **Purchase Amount:**
   * There are several outliers present in the Purchase Amount column, indicated by data points beyond the whiskers.
3. **Average Spending Per Purchase:**
   * The boxplot for Average Spending Per Purchase reveals a few outliers on the higher end of the distribution.
4. **Purchase Frequency Per Month:**
   * Outliers are noticeable in the Purchase Frequency Per Month column, suggesting potential irregular purchasing patterns.
5. **Brand Affinity Score:**
   * The boxplot for Brand Affinity Score shows a few data points beyond the whiskers, indicating the presence of outliers



**Analysis of Age and Purchase Amount Distributions:**

**Age Distribution**

Histogram:

The histogram of the 'Age' column depicts a relatively uniform distribution.

Boxplot:

The boxplot for 'Age' further confirms the absence of outliers, as all data points fall within the whiskers.

**Purchase Amount Distribution**

Histogram

* The histogram of the 'Purchase Amount' column shows a right-skewed distribution, indicating a concentration of lower purchase amounts

**Boxplot**

* The boxplot reveals the presence of outliers in the 'Purchase Amount' column, with data points beyond the whiskers.

A diagram of a number of age distribution

Description automatically generated with medium confidence

**Distribution of Age Groups**

**Introduction**

This analysis focuses on the distribution of individuals across different age groups in the dataset. The visualization employs a bar chart to illustrate the counts of each age group, providing insights into the demographic composition of the dataset.

A graph of different colored rectangles

Description automatically generated

**Distribution of Time Since Last Purchase**

**Introduction**

This analysis focuses on visualizing the distribution of the variable 'Time\_Since\_Last\_Purchase' in the dataset. The histogram, created using the seaborn library, provides insights into the frequency of different intervals since the last purchase.

**Histogram:**

* The histogram illustrates the frequency of different intervals since the last purchase.
* Each bar corresponds to a bin, and the height of the bar indicates the frequency of individuals falling within that time interval.

A graph showing a number of days

Description automatically generated with medium confidence

**Purchase Analysis: Income Level vs. Purchase Amount**

**Introduction**

This analysis explores the relationship between income levels and purchase amounts, considering the gender distribution within each income category. The scatterplot visualizes the variation in purchase amounts across different income levels, providing insights into consumer spending patterns.

**Scatterplot:**

* Each point on the scatterplot represents an individual in the dataset, positioned based on their income level and corresponding purchase amount.
* Gender is distinguished by different colors, providing additional insights into the composition of the data.

A chart of income level

Description automatically generated

**Brand Affinity vs. Product Category**

**Introduction**

This analysis investigates the relationship between brand affinity scores and product categories, considering the gender distribution. The scatterplot visually represents the brand affinity scores across different product categories, offering insights into consumer preferences.

**Scatterplot:**

* Each point on the scatterplot represents an individual in the dataset, positioned based on their brand affinity score and the corresponding product category.
* Gender is distinguished by different colors, providing additional context.

A chart with different colored dots

Description automatically generated

**Correlation Heatmap Analysis**

**Introduction**

This analysis visualizes the correlation matrix of numeric variables within the dataset using a heatmap. The heatmap provides a graphical representation of the relationships between different numerical features, allowing for a quick assessment of the strength and direction of correlations.

A screenshot of a graph

Description automatically generated

**Analysis of Purchase Trends Over Months**

**Average Purchase Frequency Over Months**

This analysis explores the average purchase frequency over different months, providing insights into the seasonal patterns of customer buying behavior. The line plot visualizes the trends, highlighting fluctuations in purchase frequency across the months.

A graph with blue lines and dots

Description automatically generated

**Average Spending Per Purchase Over Months**

This analysis examines the average spending per purchase across different months. The line plot visualizes variations in spending behavior, helping identify months with higher or lower average spending per transaction.

A graph with blue lines and numbers

Description automatically generated

**Product Category Preferences Over Months**

This analysis investigates product category preferences over different months. The countplot displays the number of purchases for each product category within each month, offering insights into changing consumer preferences.

A graph of different colored bars

Description automatically generated

**Average Purchase Amount by Income Level**

This analysis focuses on understanding the relationship between income levels and the average purchase amount. The line plot visualizes the average purchase amount for different income levels, providing insights into how spending behavior varies across income categories.

A graph with a line

Description automatically generated

**Purchase Frequency Across Brands:**

**Line Plot Analysis**

This analysis focuses on visualizing the variation in purchase frequency across different brands. The line plot illustrates how 'Purchase\_Frequency\_Per\_Month' varies for each brand, providing insights into consumer preferences and brand loyalty.

A graph showing a line of purchase

Description automatically generated

# .

# K-Means Clustering:

**Objective**

Apply K-Means clustering to identify distinct customer segments based on 'Average\_Spending\_Per\_Purchase,' 'Purchase\_Frequency\_Per\_Month,' and 'Brand\_Affinity\_Score.'

**Methodology**

1. **Data Selection:**
   * Extract relevant columns for clustering.
2. **Standardization:**
   * Standardize data to ensure uniform scaling.
3. **Optimal Cluster Selection:**
   * Determine optimal clusters using the elbow method and silhouette analysis.
4. **K-Means Clustering:**
   * Perform clustering with the chosen optimal k.
5. **Cluster Summary:**
   * Generate a summary with mean values and common product category preferences for each cluster.
6. **Visualization:**
   * Create scatter plot with cluster centroids and bar plot for cluster summaries.

**Results**

* Optimal clusters (k) determined.
* Customer segments characterized by spending, frequency, and brand affinity.
* Visualizations aid interpretation of cluster characteristics.

A graph with blue dots

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph of a bar chart

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

# B. DBSCAN Clustering:

**Objective**

Apply DBSCAN clustering to identify customer segments based on 'Average\_Spending\_Per\_Purchase,' 'Purchase\_Frequency\_Per\_Month,' and 'Brand\_Affinity\_Score.'

**Methodology**

1. **Data Selection:**
   * Extract relevant columns for clustering.
2. **Standardization:**
   * Standardize data to ensure uniform scaling.
3. **Parameter Exploration:**
   * Experiment with different **eps** and **min\_samples** values to find the optimal configuration.
4. **Optimal Parameters:**
   * Determine the optimal **eps** and **min\_samples** values resulting in the highest number of clusters.
5. **DBSCAN Clustering:**
   * Perform DBSCAN clustering using the optimal parameters.
6. **Cluster Summary:**
   * Generate a summary with mean values and common product category preferences for each DBSCAN cluster.
7. **Visualization:**
   * Create scatter plots using Matplotlib and Plotly Express to visualize the clustering results. Also, generate a bar plot for the cluster summary.

**Results**

* Optimal DBSCAN parameters identified.
* Customer segments characterized by spending, frequency, and brand affinity.
* Visualizations aid interpretation of DBSCAN clusters.

A chart with many colored dots

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

# C. K-Means++ Clustering:

**K-Means Clustering Analysis with Optimal Parameters**

**Objective**

Apply K-Means clustering to segment customers based on 'Average\_Spending\_Per\_Purchase,' 'Purchase\_Frequency\_Per\_Month,' and 'Brand\_Affinity\_Score' with optimal parameters.

**Methodology**

1. **Data Selection:**
   * Extract relevant columns for clustering.
2. **Standardization:**
   * Standardize data to ensure uniform scaling.
3. **Optimal Cluster Selection:**
   * Utilize the elbow method and silhouette analysis to determine the optimal number of clusters.
4. **K-Means Clustering:**
   * Perform K-Means clustering using the chosen optimal number of clusters and KMeans++ initialization.
5. **Cluster Summary:**
   * Generate a summary with mean values and common product category preferences for each K-Means cluster.
6. **Visualization:**
   * Create scatter plots using Matplotlib and Plotly Express to visualize the clustering results. Also, generate a bar plot for the cluster summary.

**Results**

* Optimal number of clusters identified.
* Customer segments characterized by spending, frequency, and brand affinity.
* Visualizations aid interpretation of K-Means clusters.

A graph of a number of clysters

Description automatically generated

A chart with different colored dots

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated

**Discuss the advantages and disadvantages of each algorithm in the context of Imtiaz Mall's specific needs and data characteristics**

**K-Means:**

**Advantages:**

1. **Simple and Fast:**
   * K-Means is computationally efficient and easy to understand, making it suitable for large datasets.
2. **Scalability:**
   * Works well with a large number of variables and clusters.
3. **Versatility:**
   * Can be used for a variety of data types and shapes of clusters.

**Disadvantages:**

1. **Sensitivity to Initial Points:**
   * Sensitive to the initial placement of centroids, which can lead to suboptimal solutions.
2. **Assumption of Spherical Clusters:**
   * Assumes that clusters are spherical and equally sized, which may not be suitable for all types of data.
3. **Requires Pre-specification of Clusters:**
   * Requires prior knowledge or estimation of the number of clusters, which might not be known in advance.

**K-Means++:**

**Advantages:**

1. **Improved Initialization:**
   * Addresses the sensitivity to initial points by using a smarter initialization method.
2. **Better Convergence:**
   * Tends to converge faster and to a better final solution compared to the basic K-Means.

**Disadvantages:**

1. **Complexity:**
   * Slightly more complex than the basic K-Means algorithm.
2. **Not Always Optimal:**
   * While it improves convergence, it may not always find the globally optimal solution.

**DBSCAN:**

**Advantages:**

1. **Automatic Cluster Detection:**
   * Automatically detects the number of clusters and identifies noise points.
2. **Handles Irregular Shapes:**
   * Effective in identifying clusters of arbitrary shapes and sizes.
3. **No Assumptions on Cluster Shape:**
   * Does not assume any specific shape for clusters, making it more flexible.

**Disadvantages:**

1. **Sensitivity to Parameters:**
   * Requires careful tuning of parameters such as **eps** and **min\_samples**, which may not be straightforward.
2. **Difficulty with Varying Densities:**
   * May struggle with clusters of varying densities.
3. **Not Suitable for High-Dimensional Data:**
   * Performance may degrade with high-dimensional data.

**Considerations for Imtiaz Mall's Needs:**

1. **Data Characteristics:**
   * The nature of the data, such as the shape and density of clusters, can influence the choice of algorithm. DBSCAN might be advantageous if clusters have irregular shapes.
2. **Interpretability:**
   * Consider the interpretability of the clusters. K-Means and K-Means++ may provide more interpretable results with clearly defined centroids.
3. **Noise Handling:**
   * If there is a possibility of noise points or outliers, DBSCAN's automatic noise handling might be beneficial.
4. **Computational Resources:**
   * Consider the available computational resources. K-Means is computationally efficient, making it suitable for large datasets.
5. **Cluster Size and Shape:**
   * If the clusters have varying sizes and shapes, DBSCAN might be more suitable.

**MODULE 4: Conclusions and Recommendations**

Identification of Customer Segments in the Electronics Section:

The clustering analysis has successfully unveiled distinct customer segments within the electronics section. These segments have been delineated based on a myriad of factors, encompassing purchase behavior, brand affinity, and preferences for product categories.

Crucial Factors Setting Apart Customer Segments:

The pivotal factors that serve as distinguishing features among customer segments encompass age, income levels, brand loyalty, and proclivities towards specific product categories. These elements collectively contribute to the emergence of distinct purchasing patterns observed within each segment.

Observations on Purchasing Behavior:

Varied purchasing behavior patterns are discernible across customer segments. For instance, one segment may exhibit a predilection for upscale electronics, showcasing a penchant for higher average spending. Conversely, another segment may demonstrate a proclivity for budget-friendly options, indicative of a more price-sensitive orientation.

Strategies Driven by Data for Customer Retention and Sales Growth:

• Roll out personalized marketing initiatives meticulously crafted for each customer segment. This may involve targeted promotions, discounts, or loyalty programs attuned to the distinct preferences of each segment.

• Scrutinize seasonal fluctuations in customer behavior to optimize product offerings and tailor promotions during peak seasons.

Potential Applications of Clustering Outcomes:

• Personalized Product Suggestions: Leverage the delineated customer segments to offer bespoke product recommendations, augmenting the overall shopping experience.

• Dynamic Pricing Approaches: Integrate dynamic pricing strategies based on the identified customer segments. Provide adaptable pricing structures aligned with the price sensitivity characteristic of each segment, fostering competitiveness and catering to diverse customer budgets.

• Targeted Marketing Endeavors: Devise marketing campaigns specifically tailored to each segment, addressing their unique preferences and requirements.

• Exploitation of Cross-Selling Prospects: Identify cross-selling opportunities within and between segments. Propose complementary products or accessories based on the purchase history of each segment, enticing customers to explore and make additional purchases.

• Customized Loyalty Programs: Formulate loyalty programs delivering rewards and perks in harmony with the preferences distinctive to each segment.

Further Analysis and Explorations:

• Undertake a comprehensive analysis to gauge the influence of external factors (e.g., economic conditions, technological trends) on customer behavior within the electronics section.

• Scrutinize customer feedback and reviews to comprehend sentiment and satisfaction levels within each segment.

• Maintain a continuous monitoring mechanism, updating customer segments as preferences and trends undergo evolutionary shifts over time.

Optimizing the Electronics Section:

• Regularly assess and revise the product offerings in response to the evolving preferences within each segment.

• Implement mechanisms for soliciting feedback directly from customers, aiding in the refinement of product selection and overall enhancement of customer satisfaction.