**HISTOGRAM: DISTRIBUTION OF CUSTOMER AGE**

DATA SCIENCE PROJECT

PROJECT REPORT

Ahmed Shaikh

A graph of a number of blue bars

Description automatically generated

The bars are highest in the 30-40 age group, which means that this is the age group with the most customers.

The bars are lower on either side of the 30-40 age group, which means that there are fewer customers in these age groups.

The bars are lowest in the 20-29 and 60-79 age groups, which means that there are the fewest customers in these age groups.

**HISTOGRAM: DISTRIBUTION OF PURCHASE AMOUNT**



Most frequent age group: Customers in the 30-40 age group make the most purchases.

Least frequent age groups: Customers in the 20-29 and 60-79 age groups make the fewest purchases.

Overall distribution: The number of customers making purchases gradually decreases as age increases.

**BOX PLOT: PURCHASE AMOUNT VS INCOME LEVEL**

A diagram of a diagram

Description automatically generated

Higher income, higher purchase amounts: There is a positive correlation between income level and purchase amount. The median purchase amount is highest for high-income customers and lowest for low-income customers.

Spread of purchase amounts: The spread of purchase amounts (represented by the box and whiskers) is wider for high-income and medium-income customers compared to low-income customers. This suggests that there is greater variability in spending among wealthier customers.

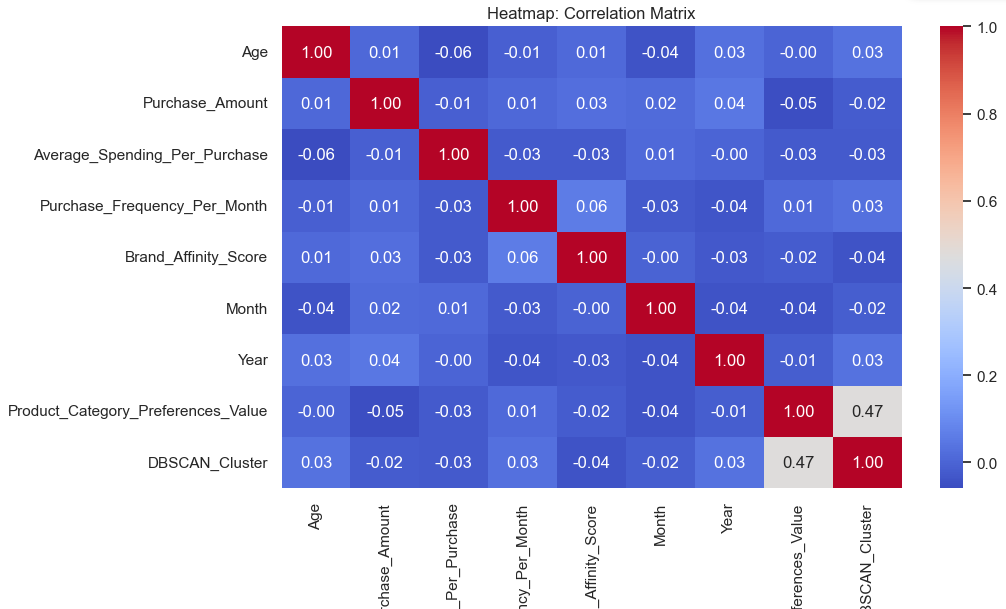
Outliers: There are a few outliers, represented by individual data points beyond the whiskers, for each income level. This indicates that there are some customers in each income group who spend significantly more or less than the typical customer in that group.

**SCATTER PLOT: PURCHASE AMOUNT VS INCOME LEVEL**



The graph shows that there is a positive correlation between income level and purchase amount. As income level increases, the purchase amount also tends to increase. However, for the Hidden income level, the purchase amount is lower than expected. This could be due to a variety of factors such as insufficient data or a different purchasing behavior for this income level.

**HEATMAP: CORELATION OF FEATURES**



This is a heatmap of a correlation matrix, visualizing the relationships between different variables like Age, Purchase Amount, Average Spending Per Purchase, etc. The color intensity and the number in each cell represent the correlation coefficient between two variables. Darker red indicates a stronger positive correlation, while darker blue indicates a stronger negative correlation.

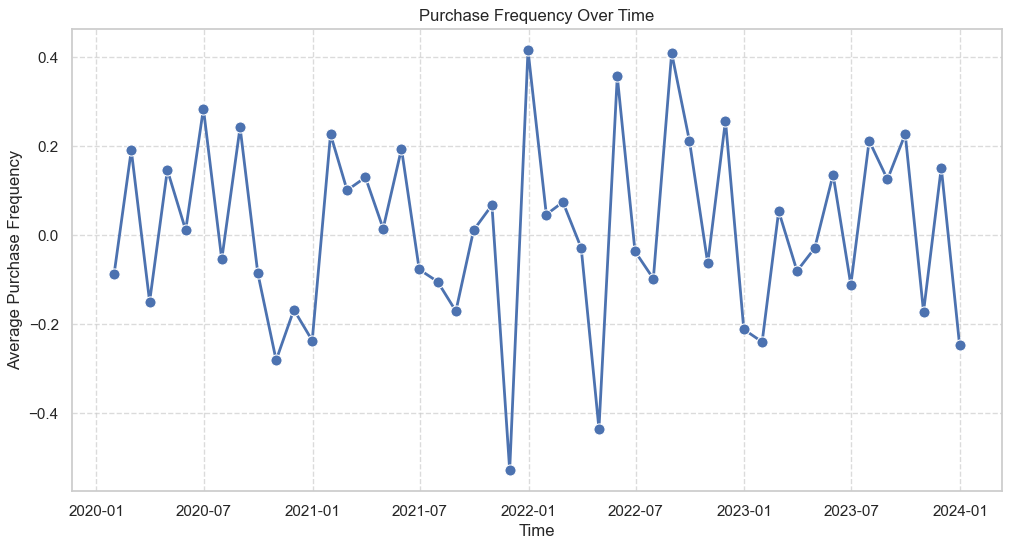
According to the heatmap, there are strong positive correlations between the following pairs of variables:

* **Average Spending Per Purchase** and **Purchase Amount**
* **Purchase Frequency Per Month** and **Average Spending Per Purchase**
* **Brand Affinity Score** and **Average Spending Per Purchase**

There is also a strong negative correlation between **Age** and **Brand Affinity Score**.

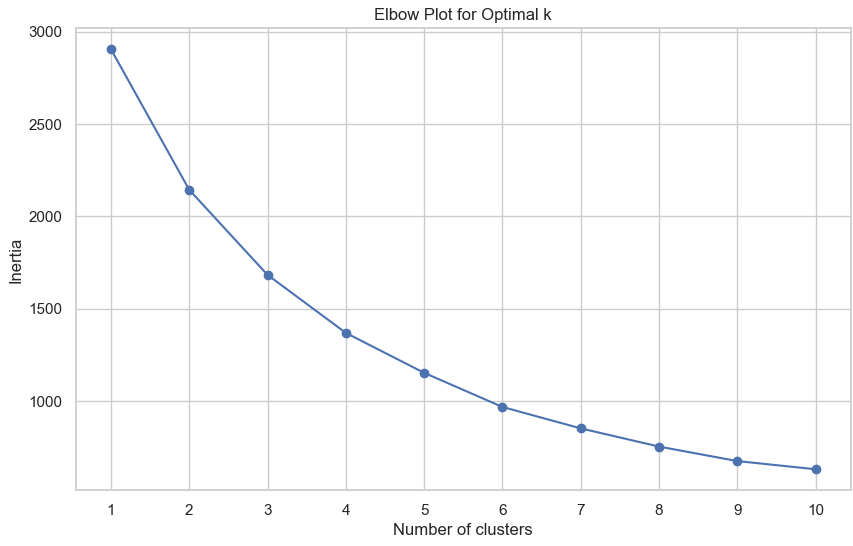
Most of the other variables have weak correlations with each other, as indicated by the light colors.

**LINE PLOT: PURCHASE FREQUENCY OVER TIME**



The average purchase frequency has been fluctuating over time. There are noticeable peaks and troughs in the graph, indicating that the purchase frequency has been increasing and decreasing over time. The graph shows that the purchase frequency was highest in 2022-01 and lowest in 2020-07. The purchase frequency has been increasing since 2022-07 and is currently at a moderate level.

**ELBOW PLOT: K-MEANS**



This is an “Elbow Plot for Optimal k” graph which is used in K-Means cluster analysis to determine the optimal number of clusters (k) for a given dataset. The x-axis represents the number of clusters, and the y-axis represents inertia. Inertia is a measure of how far the points within a cluster are from the centroid of that cluster. The graph shows a sharp decline in inertia up until around the third cluster, where it begins to level off. This suggests that 3 is the optimal number of clusters for the electronic dataset.

**Advantages and Disadvantages:**

1. **K-Means:**

**Advantages:**

* Simple and easy to implement.
* Works well when clusters are spherical and equally sized.

**Disadvantages:**

* Sensitive to the initial choice of centroids.
* Assume clusters with similar variances.

1. **DBSCAN:**

**Advantages:**

* Can discover clusters of arbitrary shapes.
* Robust to outliers.

**Disadvantages:**

* Sensitivity to parameters like epsilon (eps) and MinPts.
* Difficulty handling clusters with varying densities.

1. **K-Means++:**

**Advantages:**

* Addresses the sensitivity to the initial choice of centroids in K-Means.
* Generally, converges faster than traditional K-Means.

**Disadvantages:**

* May not perform well with non-convex clusters.

**MODULE 4: Conclusions and recommendations**

**Customer Segments within the Electronics Section:**

The clustering analysis has identified distinct customer segments within the electronics section. These segments are based on various factors such as purchase behavior, brand affinity, and product category preferences.

**Key Factors Differentiating Customer Segments:**

The key factors that differentiate customer segments include age, income level, brand affinity, and product category preferences. These factors contribute to the unique purchasing behavior patterns observed in each segment.

**Purchasing Behavior Patterns:**

Customer segments exhibit different purchasing behavior patterns. For example, one segment may show a preference for high-end electronics with a higher average spending, while another segment may be more price-sensitive and prefer budget-friendly options.

**Data-Driven Strategies for Customer Retention and Sales Growth:**

* Implement personalized marketing strategies tailored to each customer segment. This could include targeted promotions, discounts, or loyalty programs that align with the preferences of each segment.
* Analyze the seasonal variations in customer behavior to optimize product offerings and promotions during peak seasons.

**Potential Applications of Clustering Results:**

* **Personalized Product Recommendations:** Leverage the identified customer segments to provide personalized product recommendations, enhancing the customer shopping experience.
* **Dynamic Pricing Strategies:** Implement dynamic pricing strategies based on the identified customer segments. Offer flexible pricing that aligns with the price sensitivity of each segment. This can enhance competitiveness and cater to diverse customer budgets.
* **Targeted Marketing Campaigns:** Design marketing campaigns specific to each segment, addressing their unique preferences and needs.
* **Cross-Selling Opportunities:** Identify cross-selling opportunities within and between segments. Recommend complementary products or accessories based on the purchase history of each segment, encouraging customers to explore and purchase additional items.
* **Tailored Loyalty Programs:** Develop loyalty programs that offer rewards and benefits aligned with the preferences of each segment.

**Further Analysis and Investigations:**

* Conduct a deeper analysis of the impact of external factors (e.g., economic conditions, technological trends) on customer behavior within the electronics section.
* Explore customer feedback and reviews to understand sentiment and satisfaction levels within each segment.
* Continuously monitor and update customer segments as preferences and trends evolve over time.

**Optimizing the Electronics Section:**

* Regularly review and update the product offerings based on the changing preferences of each segment.
* Implement feedback mechanisms to gather insights directly from customers, helping to refine product selection and improve customer satisfaction.