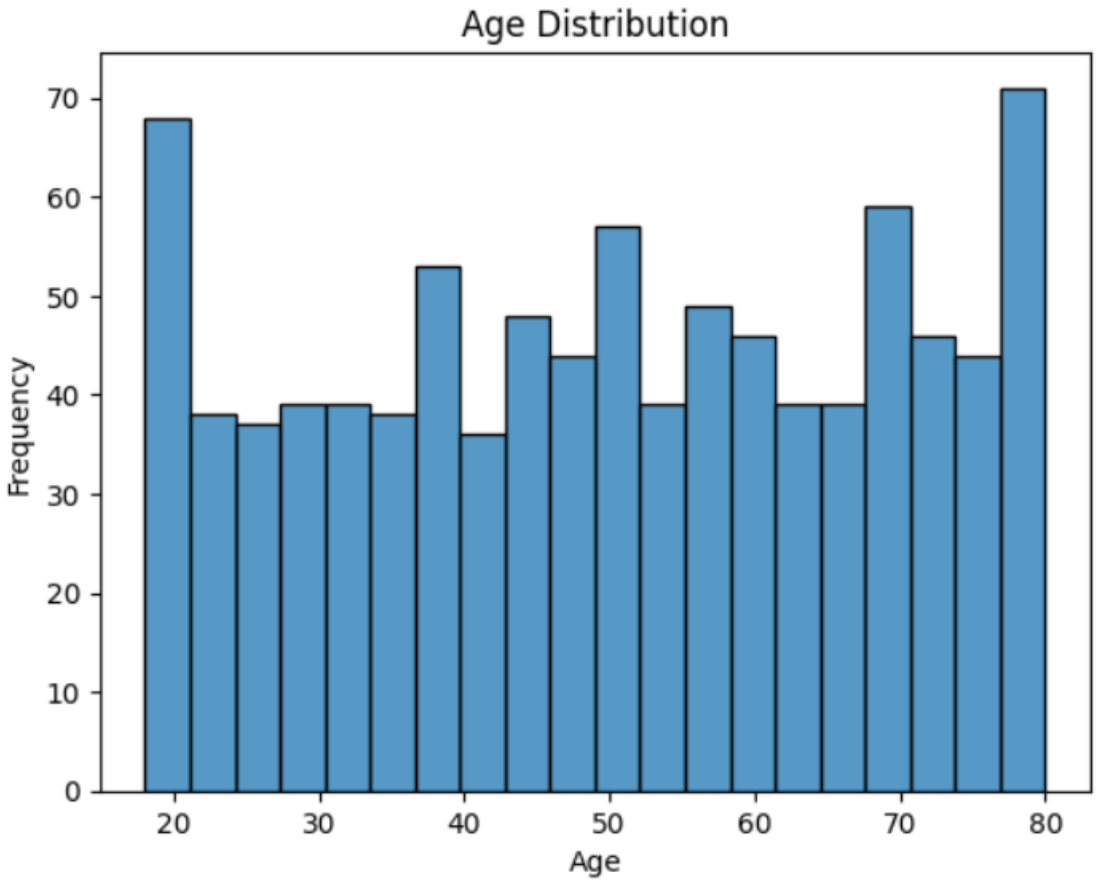
DATA SCIENCE REPORT

ZAIN UL ABIDEEN 20F-0277
AHMAD RAZA 20F-0109
AHMAD

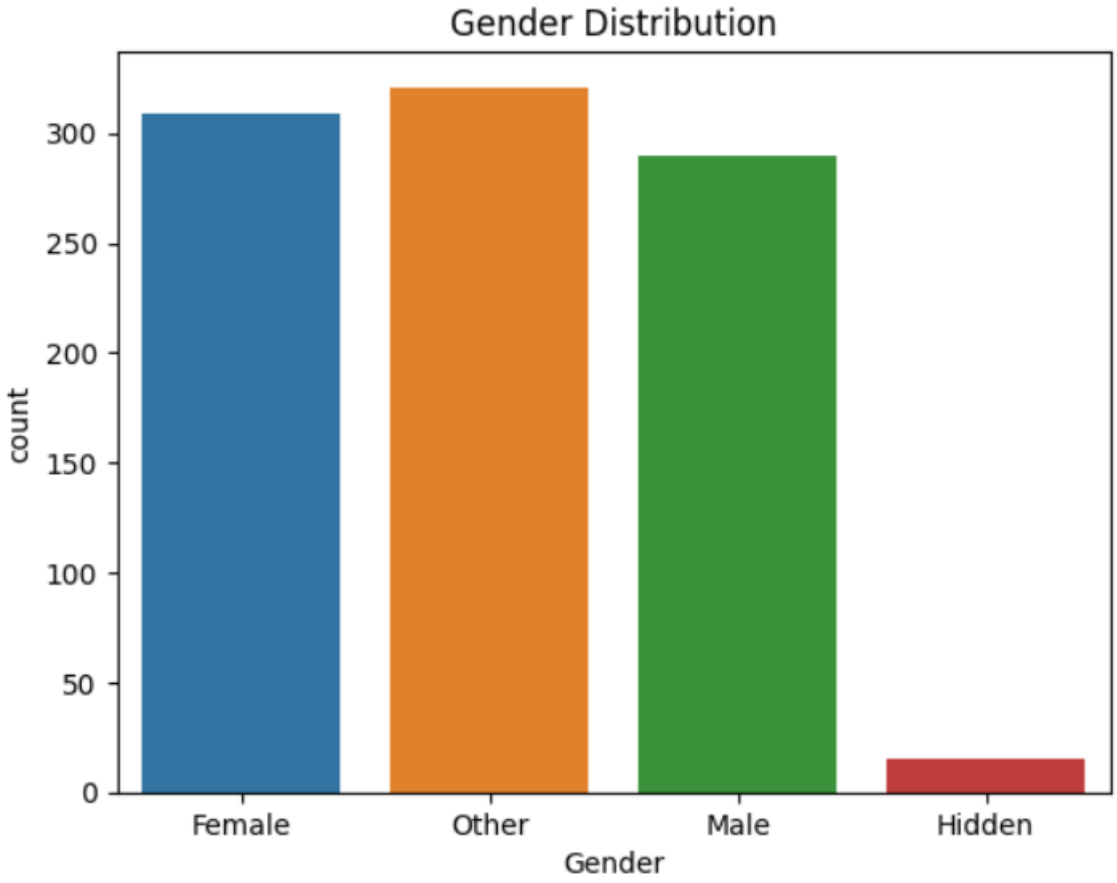
HISTOGRAM: DISTRIBUTION OF CUSTOMER AGE



Key insights:

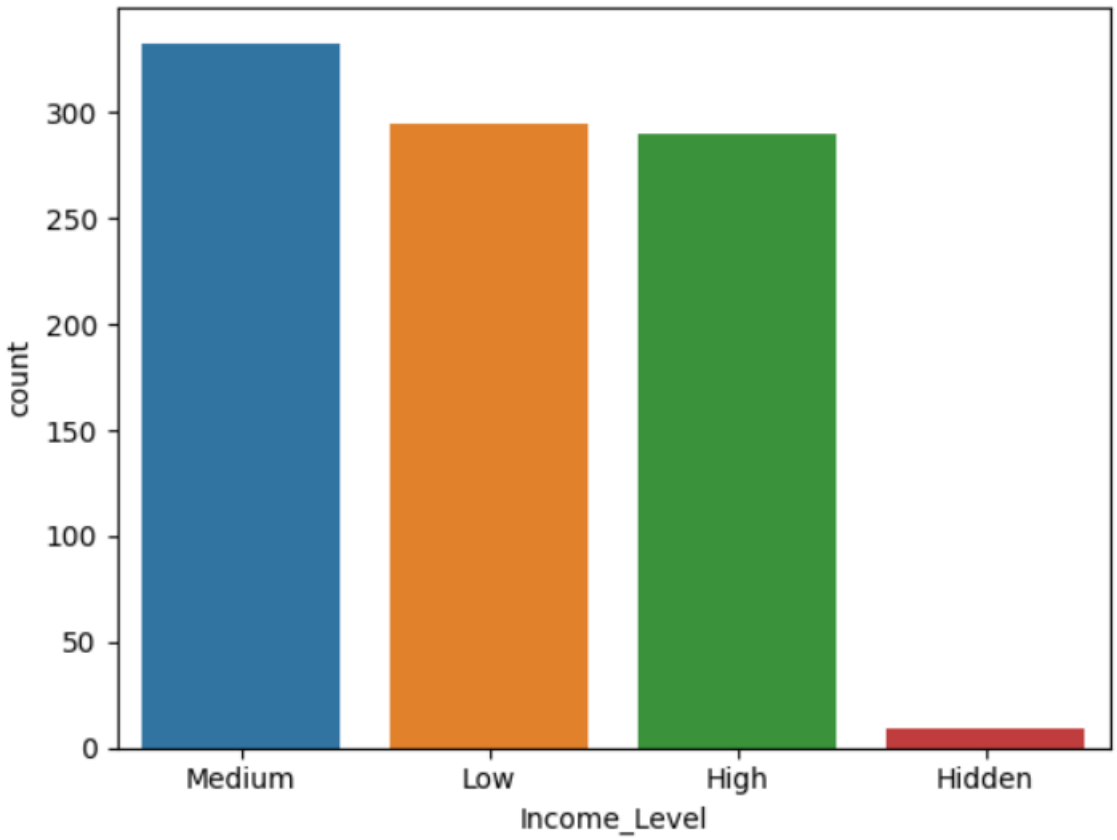
- The age group with the highest frequency (appearing to be around 70 individuals) is the 80-year-old bracket.
- The age group with the lowest frequency (around 20 individuals) is in the 30-year-old bracket.
- There is a general trend where the frequency decreases from the 20-year-old bracket to the 30-year-old bracket and then fluctuates in the subsequent age brackets.
- The age groups of 20s and 80s stand out because they have the highest frequencies compared to the other age groups.
- The graph shows a relatively uniform distribution among the age groups of 40 to 70, with frequencies ranging approximately between 40 and 60 individuals.
- There is no smooth gradient of change; rather, there are peaks and troughs, which suggests that the distribution is not normally distributed.

Histogram of Gender Distribution



Key insights:

- Female Representation: The highest count is for females, suggesting that more females than males or individuals of other genders are associated with the dataset. This could indicate that females are either the primary customers, employees, or the target demographic.
- Other Genders: The category labeled "Other" has a significant representation, which is quite close to the count of females. This suggests that Imtiaz store is inclusive of various gender identities beyond the traditional male and female categories.
- Male Representation: The count for males is also substantial, though less than females and others. It shows that males are also a significant part of the customer base, workforce, or demographic focus of the store.
- Hidden Gender: There's a small representation of individuals whose gender is not disclosed or is hidden. This could be due to privacy choices or the unavailability of data.



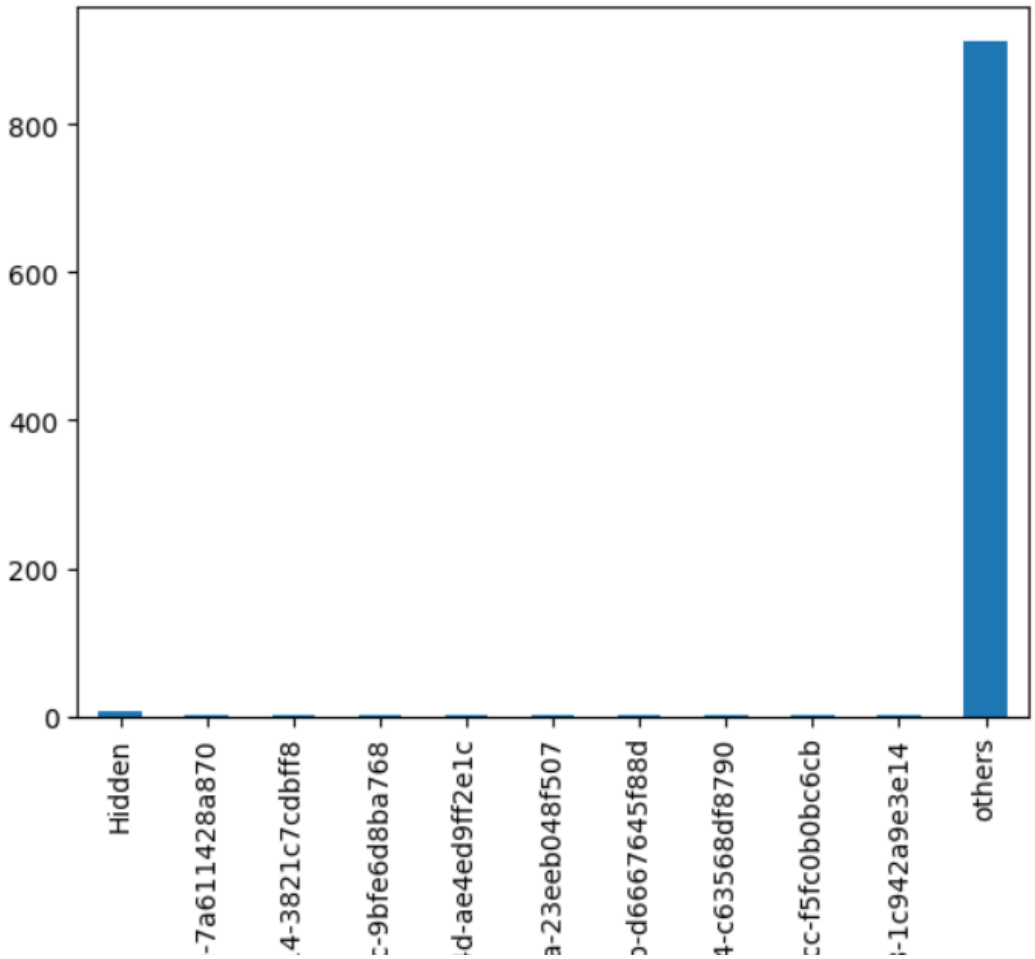
0

Key insights:

Medium Income Predominance: Most individuals fall into the 'Medium' income level category. This suggests that the store's primary customer base, or possibly its employee demographic, has a medium level of income.

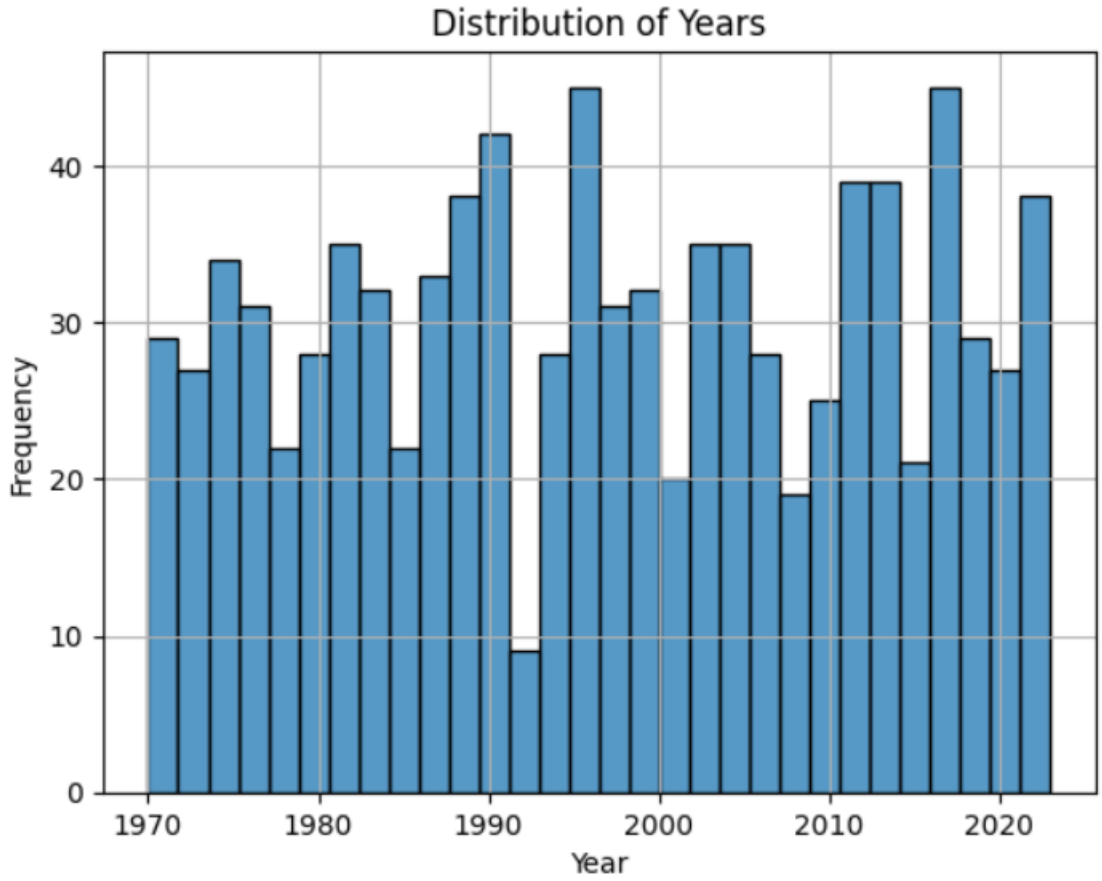
- Low- and High-Income Representation: Both the 'Low' and 'High' income levels have a substantial count, indicating a significant presence of both lower and higher-income individuals. However, the 'High' income category is slightly more represented than the 'Low' income category.

- Minimal Hidden Income Data: A very small number of individuals have their income level categorized as 'Hidden'. This could indicate that most people are open about their income level or that the data collection was able to capture this information effectively.

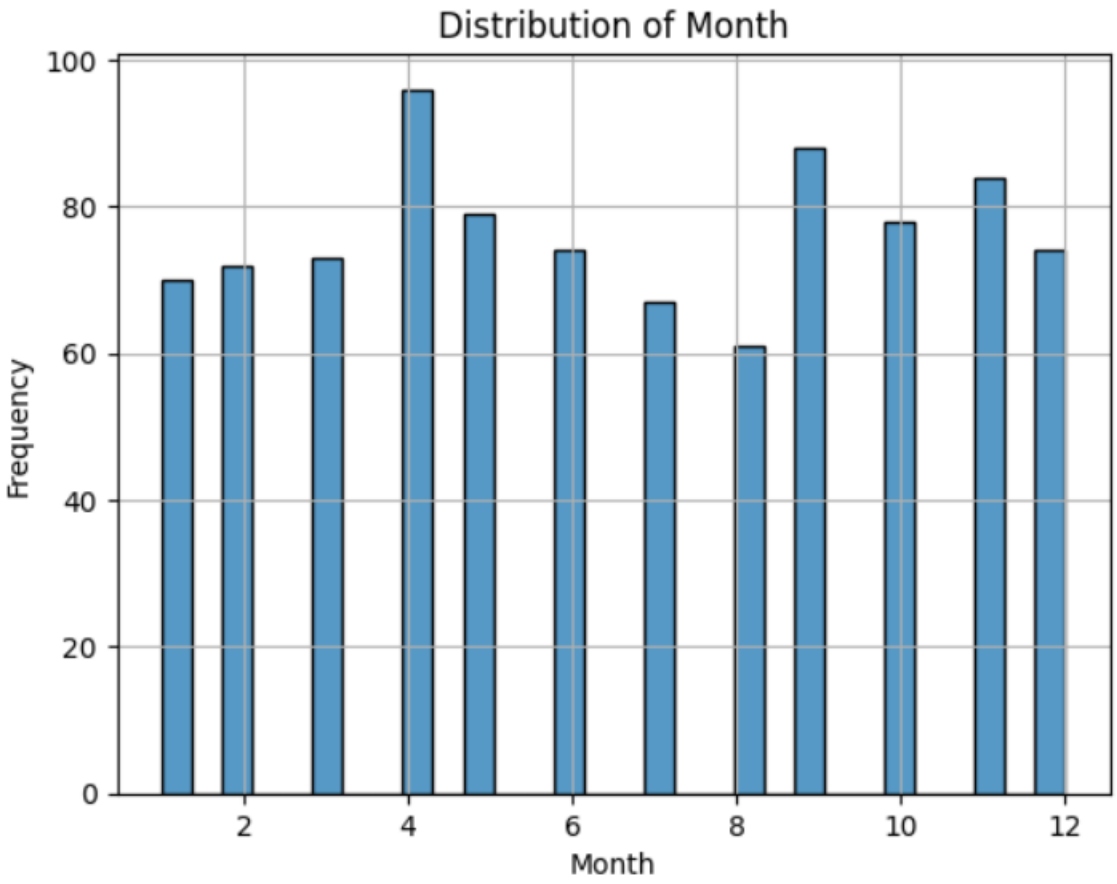


The bar chart appears to represent the distribution of a certain category among a population, with each bar labeled with what seems to be an ID code or a specific categorization tag, except for the last bar labeled "Others".

Distribution of Years:



- Fluctuating Frequencies: There are fluctuations in frequency throughout the years, with no clear trend of increase or decrease over time. This indicates variability in the data year over year.
- Peaks in the 1980s and 2000s: There are noticeable peaks in frequency during certain years in the 1980s and the early 2000s. This could reflect periods of higher activity, interest, or occurrences of the measured variable.
- Dips in the 1990s and 2010s: There are dips in frequency during parts of the 1990s and the 2010s. These could correspond to periods of lower activity or interest in the variable being measured.
- Recent Increase: There appears to be an increase in frequency in the years leading up to 2020, which might suggest a recent resurgence or increase in the variable being measured.
- Data Range: The data spans over 50 years, which indicates a long-term study or collection of the variable in question.



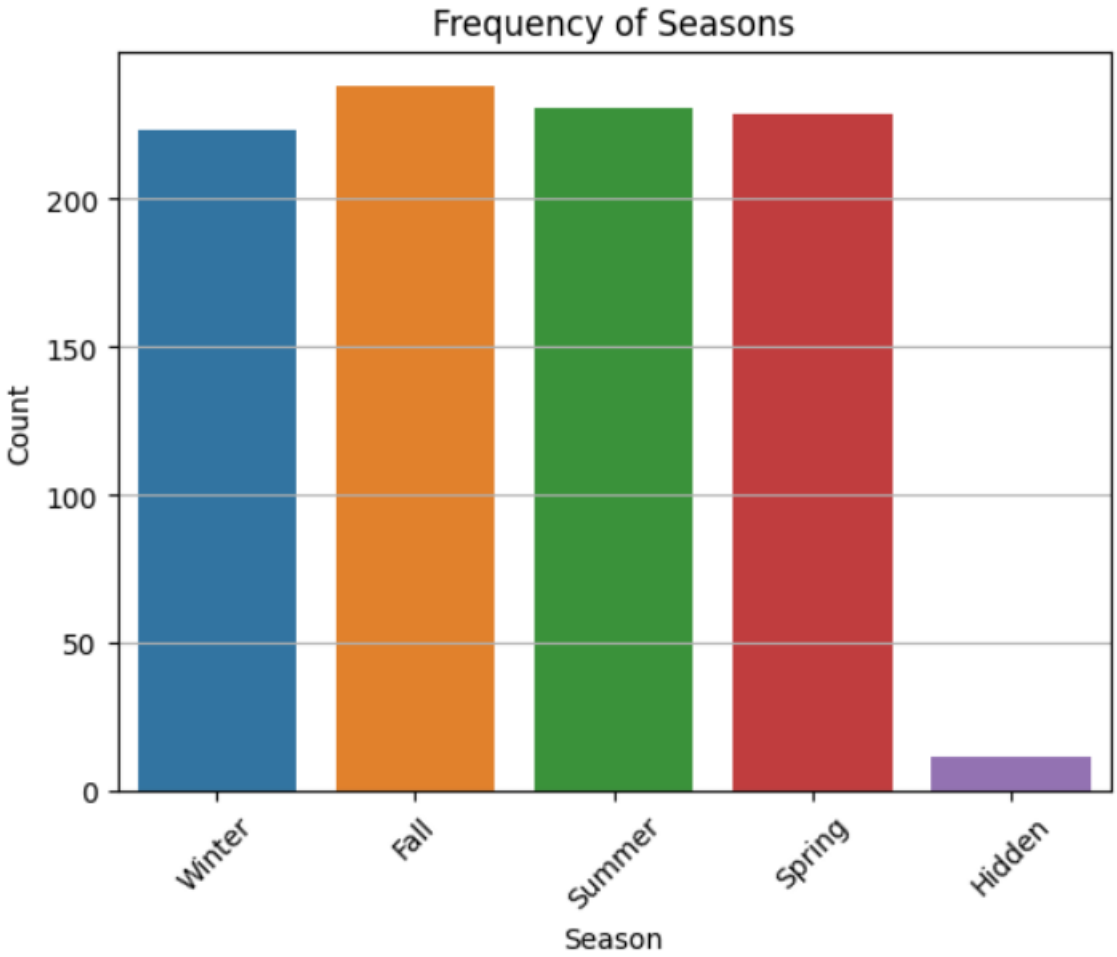
Key insights:

Higher Frequencies in Certain Months: There are peaks in frequency during March, July, and December. This could suggest seasonal trends or specific events that cause an increase in the variable being measured during these months.

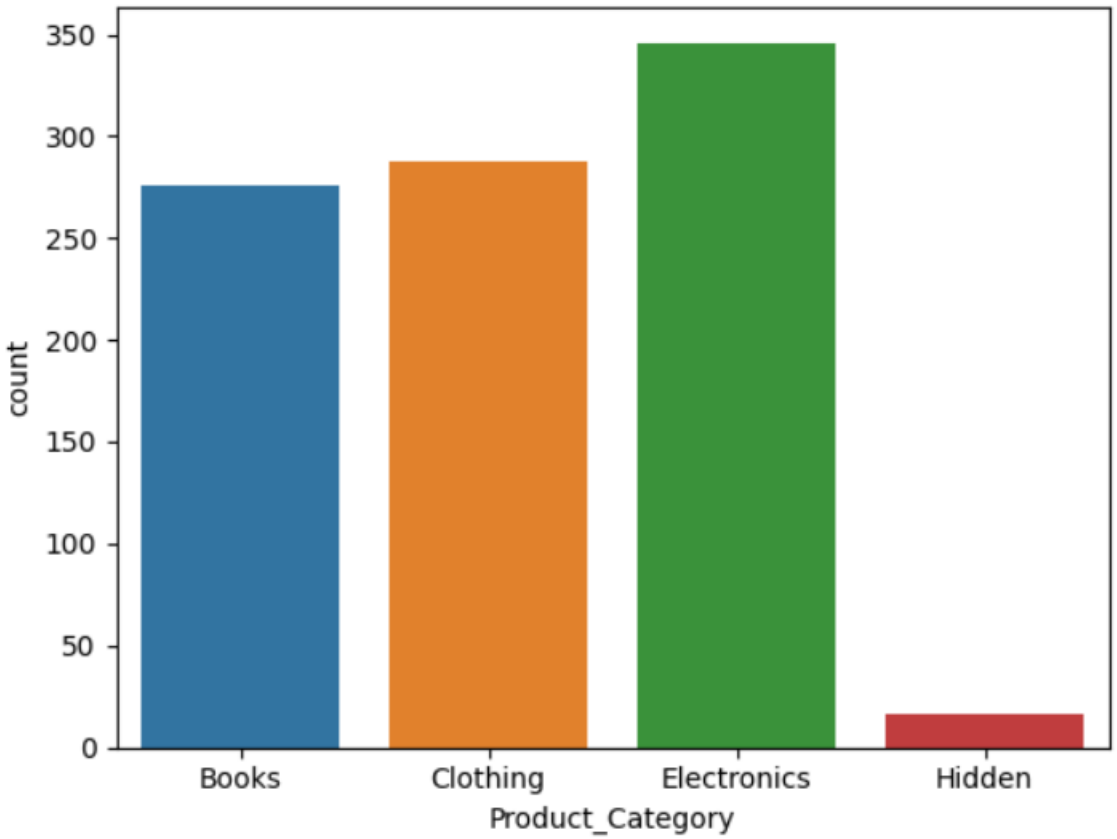
Lower Frequencies in Other Months: There are notably lower frequencies in May and November. These months could be considered off-peak periods for the variable in question.

General Trend: Apart from the peaks and troughs, there is a relatively consistent frequency across the other months, suggesting a stable trend outside of the peak and off-peak periods.

End of Year Increase: There is a notable increase in frequency in December which might be associated with year-end activities such as holidays or sales events.

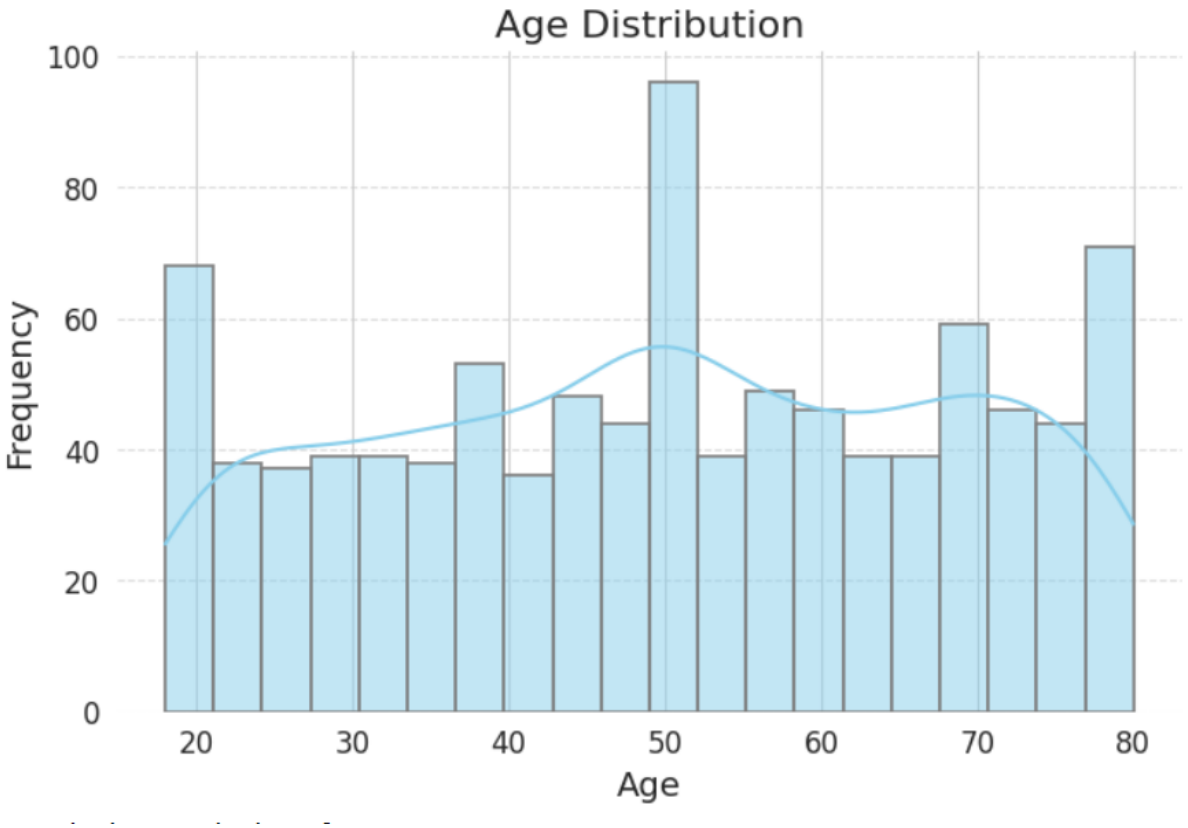


- Balanced Seasonal Distribution: The frequency counts for Winter, Fall, Summer, and Spring are balanced, with each season showing a substantial count. This suggests that the variable measured is relatively stable across different seasons.
- Slight Variations: There are slight variations between the seasons, but none of the seasons drastically outnumbers the others. This could imply that the variable is not heavily influenced by seasonal changes.
- Minimal Hidden Data: The 'Hidden' category has a very low count, indicating that there is little missing or unclassified data in terms of seasonal distribution.



- Electronics as the Leading Category: Electronics have the highest count, indicating that they are either the most stocked or the best-selling category in the store.
- Books and Clothing with Comparable Counts: The Books and Clothing categories have similar counts, which are substantial but less than Electronics. This suggests that these categories also perform well and are significant to the store's inventory or sales.
- Minimal Hidden Data: The 'Hidden' category has a very low count, suggesting that the data for most products is well categorized and there are very few items that are not classified under a specific category.

Module # 2:



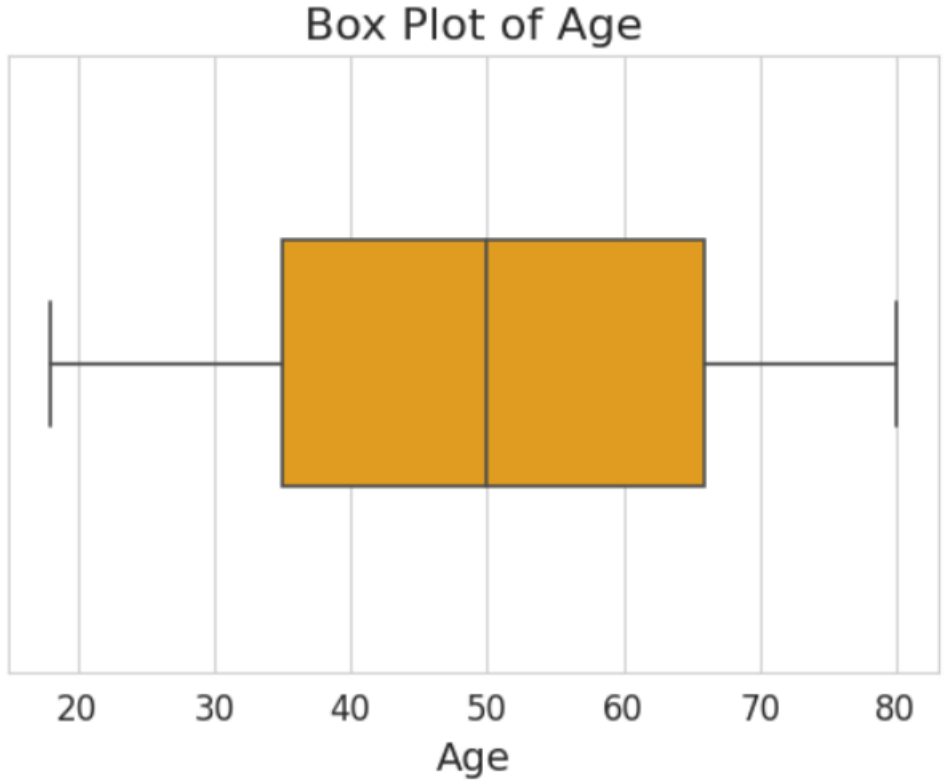
Major Peaks: There are major peaks in the age groups around the early 50s and the early 70s, suggesting that these are the most common ages within the group.

Mid-Age Trough: There is a noticeable trough in the mid-age range, particularly around the late 50s, indicating fewer individuals in this age group.

Consistent Presence of Young Adults: Individuals in their late 20s to early 40s are consistently present, with the frequency gradually increasing from the 20s to the 40s.

Lower Frequency in Youth and Seniors: The lower frequency at the beginning (ages 20s) and the end (ages 80s) of the age spectrum suggests that these age groups are the least common within the population.

Overall Trend: The line graph, which likely represents a moving average, shows the general trend is a gradual increase from the 20s to the 50s, a sharp drop in the late 50s, followed by a peak in the early 70s, and then a decline towards the 80s.

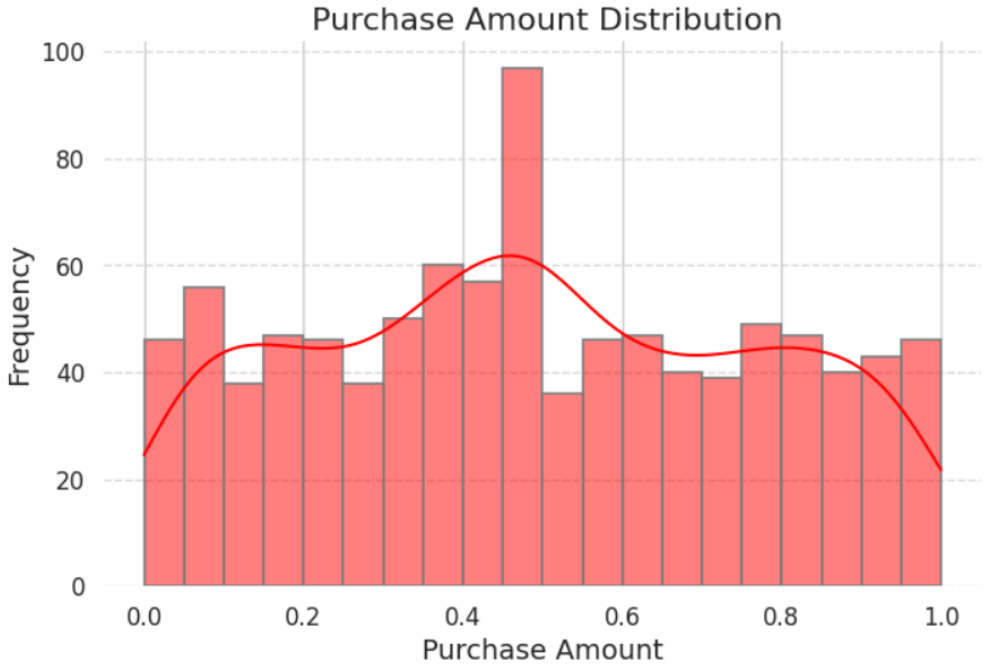


- Median Age: The line within the box represents the median age, which appears to be around the early to mid-50s. This is the age where half of the population is younger and the other half is older.

- Interquartile Range (IQR): The box represents the middle 50% of the data, known as the interquartile range. It spans from what appears to be the mid-30s to the mid-70s. This indicates that half of the population's ages fall within this range.

- Spread of Ages: The 'whiskers' of the box plot (the lines extending from the box) seem to go from approximately the mid-20s to the early 80s. These represent the range of the data excluding outliers.

- Potential Outliers: There are no dots outside of the whiskers which would indicate outliers, suggesting that all ages fall within a predictable range without extreme values.



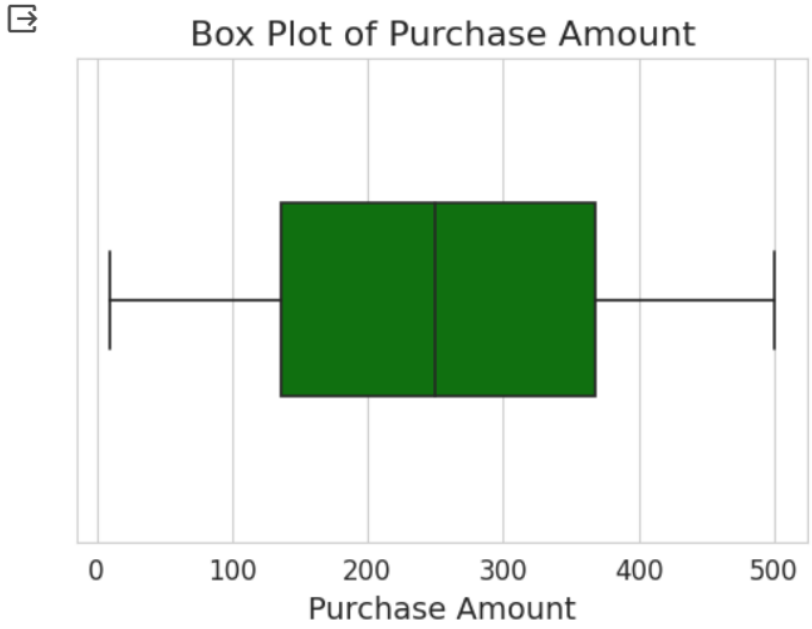
- Common Purchase Amounts: The most frequent purchase amount is around the 0.6 mark. This suggests that the majority of purchases fall into this normalized value, which could represent a particular price range if these are normalized figures.

- Bimodal Distribution: The distribution appears to be bimodal, with two peaks: one lower peak around the 0.4 mark and the main peak at 0.6. This indicates two common purchase amount ranges.

- Variability in Purchase Amounts: There is considerable variability in purchase amounts, as indicated by the spread of bars across the range.

- Lower Frequency at Extremes: The frequency of purchases is lower at the extreme ends of the scale, particularly close to 0 and 1.

- Overall Trend: The line graph shows an overall trend with increases and decreases, mirroring the bimodal nature of the histogram.



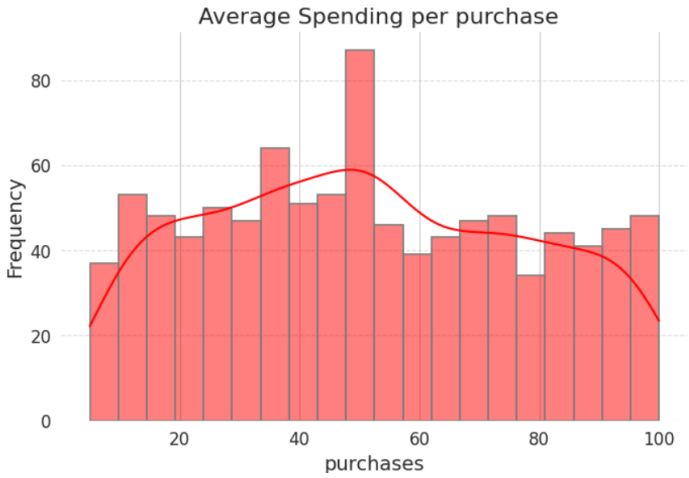
- Median Purchase Amount: The median purchase amount, indicated by the line within the box, is around 200 units. This is the central value of the dataset.

- Interquartile Range (IQR): The box itself shows the interquartile range, extending from approximately 150 to 250 units, indicating where the middle 50% of the data lies.

- Range of Purchase Amounts: The 'whiskers' of the box plot extend from the lowest to the highest purchase amounts within 1.5 IQRs from the lower and upper quartiles, which appear to range from close to 100 units to nearly 400 units.

- No Outliers Present: There are no points beyond the whiskers, indicating that there are no outliers in this dataset, or that all purchase amounts are within an expected range.

[25]

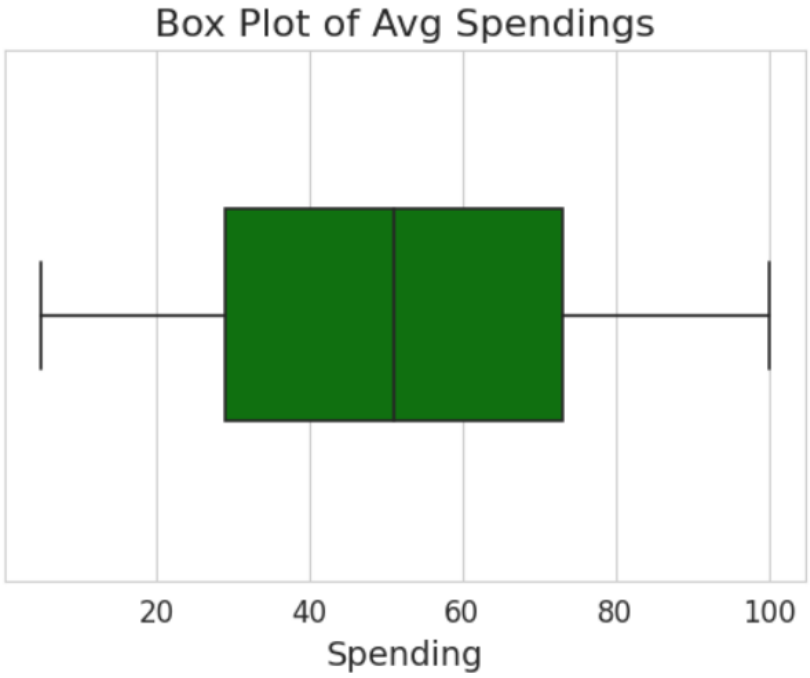


- Distribution Shape: The distribution appears to be roughly normal with a slight right-skew, indicated by the tail extending towards higher purchase numbers.

- Most Common Purchase Range: The peak of the histogram and the line graph is around the 40-50 purchases range, suggesting this is the most common purchase amount range.

- Range of Purchase Amounts: The distribution of purchases is spread out from less than 20 purchases to over 100, with the frequency gradually declining as the purchase amount increases.

- Highest Frequency Purchase Amount: There is a particularly high frequency at one point, significantly higher than the surrounding values, which could represent a mode in the data.



- Median Spending: The median value, represented by the line in the middle of the box, is around 50 units of spending. This is the middle value of the dataset, with half of the spendings being lower and half being higher.

- Interquartile Range (IQR): The box shows the interquartile range, from approximately 35 to 65 units of spending, which contains the middle 50% of the data. This indicates that half of the average spending amounts are within this range.

- Range of Spendings: The 'whiskers' extend to the minimum and maximum values within 1.5 times the IQR from the lower and upper quartiles. They seem to range from about 20 to just over 80 units of spending.

- Potential Outliers: The box plot does not show any points beyond the whiskers, which would indicate outliers. Thus, all spendings are within an expected range without extreme deviations.

- Distribution Across Income Levels: The purchase amounts are distributed across all income levels. However, it's noticeable that the spread is fairly similar across Medium, Low, and High income levels, indicating that purchase amounts are not solely dependent on the income level.

- High Income Level Purchases: Individuals in the High income level have purchase amounts that span the entire range, including the highest purchase amounts observed in the data. However, it is not exclusive to high-income individuals, as similar purchase amounts are seen in Medium and Low income levels as well.

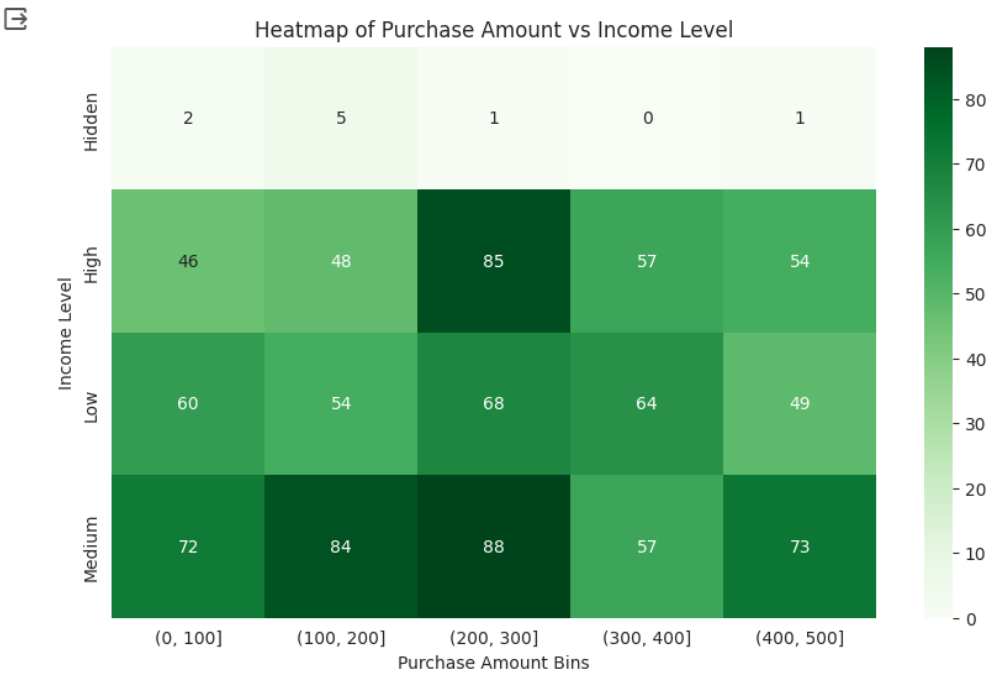
- Consistency in Medium and Low Income Levels: Medium and Low income levels show a high concentration of data points across a consistent range, suggesting that individuals from these income levels tend to spend within a certain band, without many extremes.

- Consistent Purchase Frequencies Across Income Levels: All three income levels (Medium, Low, and High) have a similar range of purchase frequencies per month, generally between 2 and 10 times.

- High Income Level Distribution: High-income level individuals have a few more instances of higher frequency purchases, with some points going up to 10 times per month.

- No Clear Correlation: There's no clear correlation that higher income levels result in a higher frequency of purchases. All income groups shop with similar frequencies.

- Hidden Income Level Data: The 'Hidden' category has fewer data points, suggesting less information available for these individuals. The frequency for this group also spans a range similar to the other income levels.'



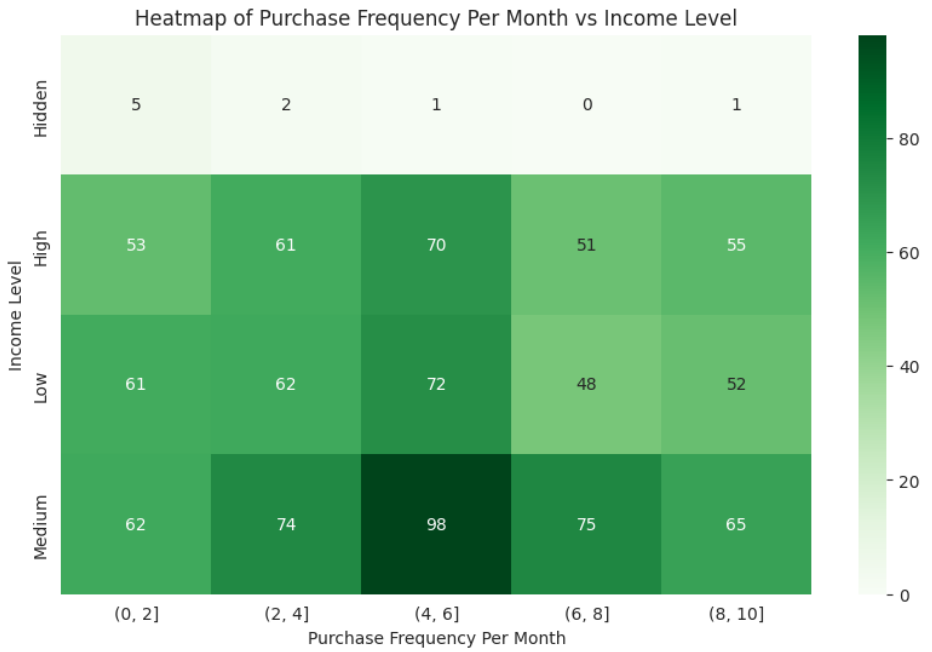
- Medium Income Level: This group has the highest frequency of purchase amounts across all bins. Particularly, the (200, 300] bin is the most common, followed by the (300, 400] and (100, 200] bins.

- Low Income Level: This group also shows a high frequency of purchases, especially in the (200, 300] bin, which is the most frequent, followed by the (0, 100] and (100, 200] bins.

- High Income Level: The highest frequency for this group is in the (200, 300] bin, indicating this is a common purchase amount range for high-income individuals. However, there is also significant activity in the higher purchase amount bins [(300, 400] and (400, 500]].

- Hidden Income Level: Very few purchases have an undisclosed income level. The data points are sparse and spread out across the bins, making it difficult to discern any pattern.

- Overall Trends: The (200, 300] purchase amount bin seems to be the most common across all income levels, suggesting that this price range is popular among customers of Imtiaz store, regardless of their income.



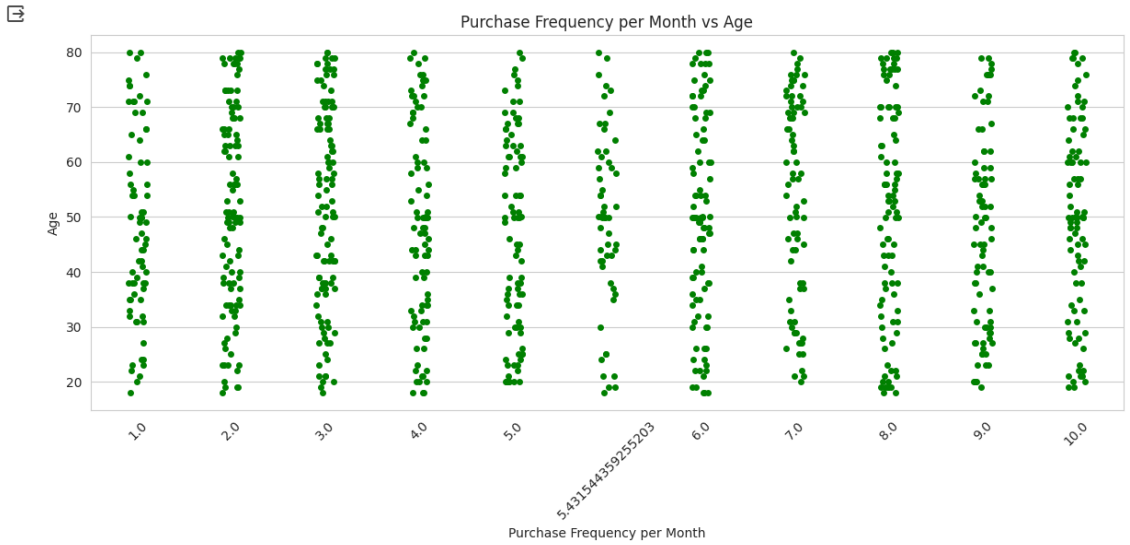
- Medium Income Level: The most frequent purchase interval for this group is within the (4, 6] range, with the frequency being the highest at 98 occurrences. This suggests that individuals in the medium income bracket tend to make purchases 4 to 6 times per month more often than other frequencies.

- Low Income Level: For this group, the (4, 6] purchase frequency per month is also the most common, with 72 occurrences. However, the distribution is fairly even across the (2, 4] and (6, 8] ranges as well.

- High Income Level: The (4, 6] frequency range is also prevalent among high-income individuals, with the highest occurrence being 70. This group also shows a relatively even distribution across the (2, 4] and (6, 8] ranges, with a notable amount in the (8, 10] range.

- Hidden Income Level: There are very few data points for this group, making it difficult to determine any significant patterns.

- General Observation: Across all income levels, the (4, 6] purchase frequency per month is the most common, which could indicate a general shopping habit or store promotion cycle that encourages this frequency of purchases.



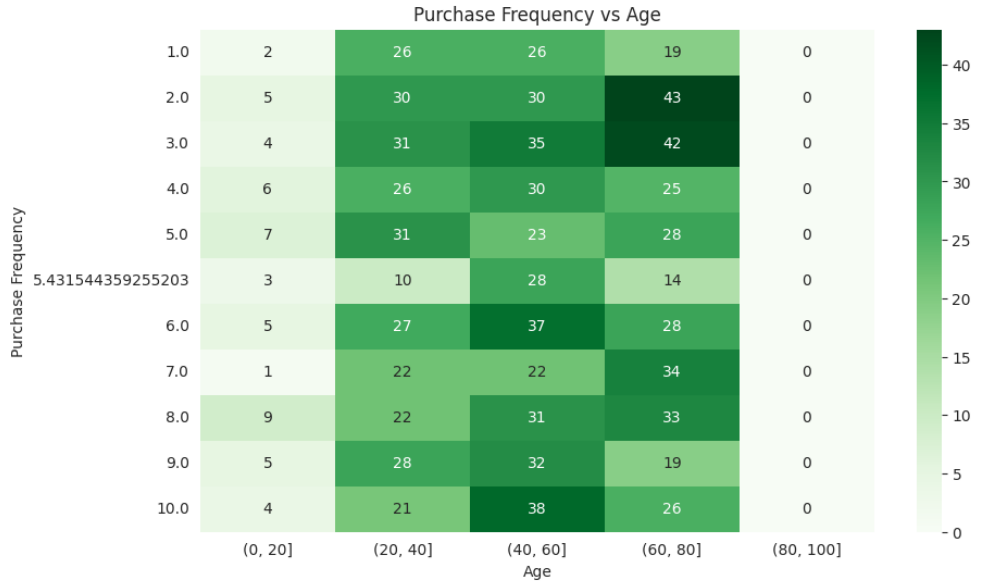
Age Distribution: The ages are spread across the range from the 20s to the 70s. There doesn't appear to be any age group that is particularly more active in purchasing frequency.

Purchase Frequency: The purchase frequency per month ranges from 1 to 10 times. There is a consistent presence of all purchase frequencies across all ages.

No Clear Trend: There is no clear trend that indicates that a particular age group tends to shop more or less frequently than others. Each frequency, from 1 to 10 times per month, is represented across the age spectrum.

Data Density: There is a dense clustering of data points at each frequency level, which suggests a large number of individuals have similar purchase frequencies regardless of age.

Outliers: There don't appear to be any significant outliers; the purchase behavior is consistent across different ages.



Age Groups Most Active: The (40, 60] and (60, 80] age bins appear most active across various purchase frequencies, indicating these age groups shop more frequently per month compared to the younger and older age groups.

Most Frequent Purchase Frequency: Within the (40, 60] age group, a purchase frequency of 6 times per month is the most common, with 37 occurrences. Similarly, in the (60, 80] age group, the most frequent purchase frequency is 3 times per month, with 42 occurrences.

Least Active Age Groups: The youngest (0, 20] and oldest (80, 100] age groups have very low frequencies of purchases per month, which could indicate lower engagement with the store or less need to purchase frequently.

Consistent Shopping Habits: Purchase frequencies of 3 times per month are consistently high across all active age groups, which might suggest a store promotion cycle or customer habit that encourages this frequency of shopping.

No Data for Some Frequencies in Older Age Group: For the oldest age group (80, 100], there is no data for purchase frequencies above 3 times per month, which might suggest less frequent shopping as age increases.

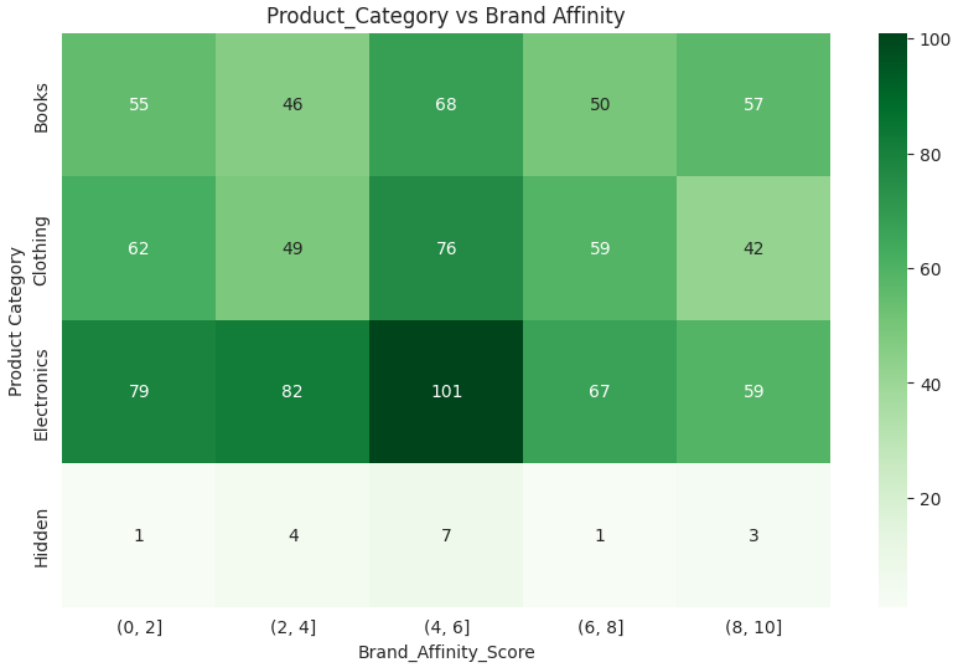
- Brand Affinity Distribution: Brand affinity scores are distributed across a scale from 2 to 10 for Books, Clothing, and Electronics categories, with no scores below 2 or above 10.

- Books and Clothing Categories: Both Books and Clothing categories have a dense distribution of brand affinity scores, spreading across the scale but with many points clustered in the mid to high range (around 4 to 8).

- Electronics Category: The Electronics category shows a wider spread of brand affinity scores, including several points at the highest score of 10. This suggests that customers might have a strong brand loyalty or preference when it comes to electronics.

- Hidden Category: There are fewer data points in the Hidden category, and the brand affinity scores are lower and less dense compared to the other categories. This may indicate less brand loyalty or insufficient data to determine a pattern.

- High Scores for Electronics: The presence of several high scores (10) in the Electronics category could suggest that customers are particularly loyal to certain brands within this category, which might be due to the nature of electronics purchases where brand reputation and trust are significant factors.



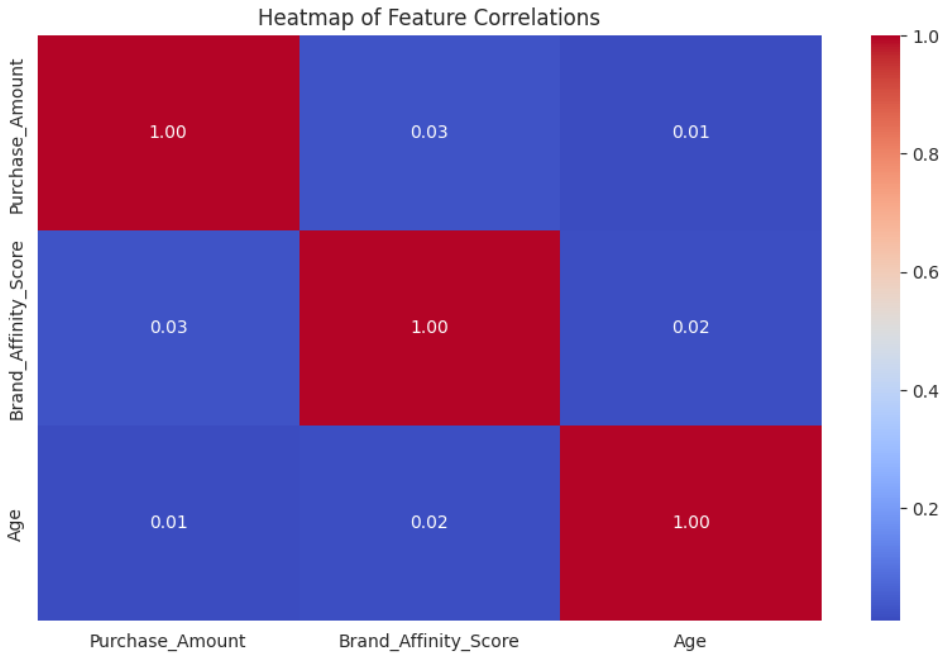
- Electronics Category: The highest concentration of brand affinity scores is in the (4, 6] range, with the Electronics category showing the most significant number at 101. This suggests strong brand loyalty among customers purchasing electronics.

- Books and Clothing Categories: Both categories show a relatively even distribution of brand affinity scores, but the (4, 6] range is still the most frequent. This indicates moderate brand loyalty for these product types.

- High Scores in Electronics: The Electronics category also has a high number of scores in the (6, 8] and (8, 10] ranges, indicating that customers have a higher likelihood of strong brand preference or loyalty in this category.

- Lower Scores in Hidden Category: The Hidden category, which likely represents data where the product category is not disclosed, has lower overall counts, but still shows some brand affinity in the (4, 6] range.

- Overall Brand Affinity: Across all visible product categories, the majority of brand affinity scores are concentrated in the middle ranges, with fewer at the extreme low or high ends.

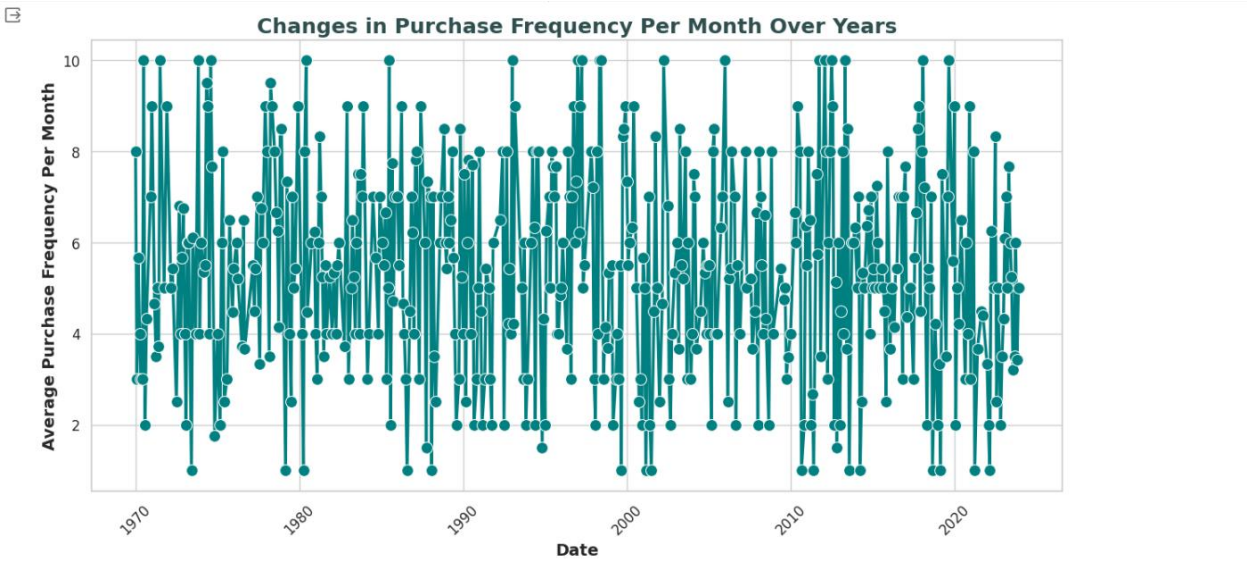


- Low Correlation Values: All off-diagonal values, which represent the correlations between different variables, are very low, close to zero. This suggests there is no linear relationship between these variables.

- Purchase Amount and Brand Affinity Score: The correlation between purchase amount and brand affinity score is 0.03, indicating almost no linear relationship.

- Purchase Amount and Age: The correlation between purchase amount and age is 0.01, also indicating no significant linear relationship.

- Brand Affinity Score and Age: The correlation between brand affinity score and age is 0.02, which is again very low and suggests no meaningful linear relationship.



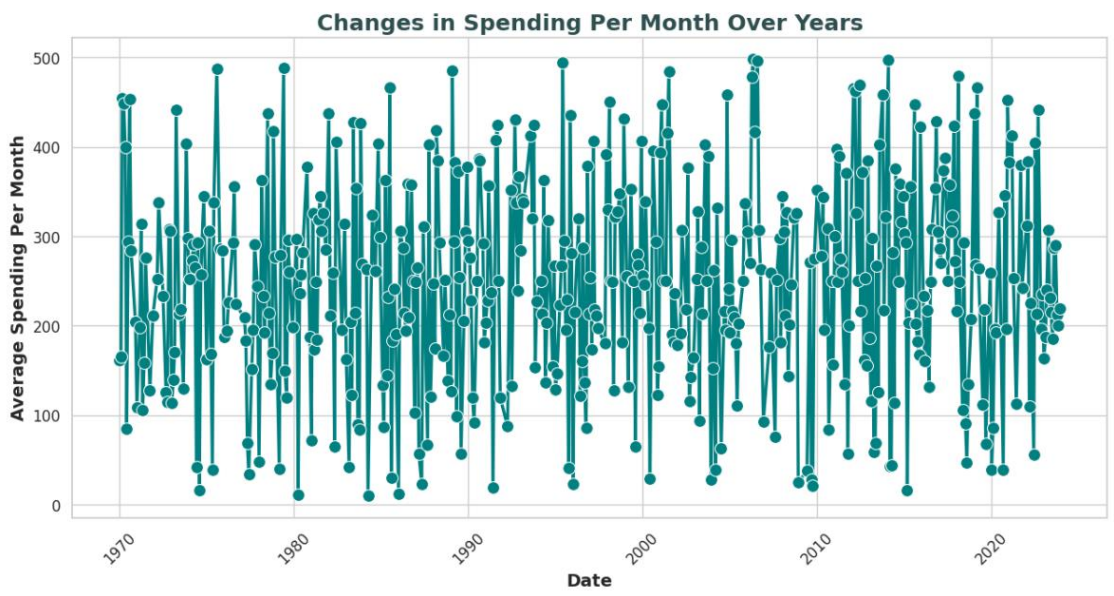
- Variance in Purchase Frequency: There is considerable variance in purchase frequency throughout the years, with some months showing an average frequency as low as 2 and others as high as 10.

- No Clear Trend: There is no discernible long-term upward or downward trend in the average purchase frequency per month over the years. The data points fluctuate significantly from year to year.

- Consistent Range: Despite the fluctuations, the purchase frequencies mostly stay within a range of approximately 2 to 10, suggesting that while individual months may vary, the overall range of frequencies doesn't change drastically.

- Yearly Patterns: Without clear labeling for each data point, it's challenging to ascertain any specific yearly patterns or to identify any outliers that might indicate significant changes in customer behavior.

- Data Density: There are a lot of data points clustered around the middle frequency numbers (4-6), which might indicate that these are the most common purchase frequencies for customers.

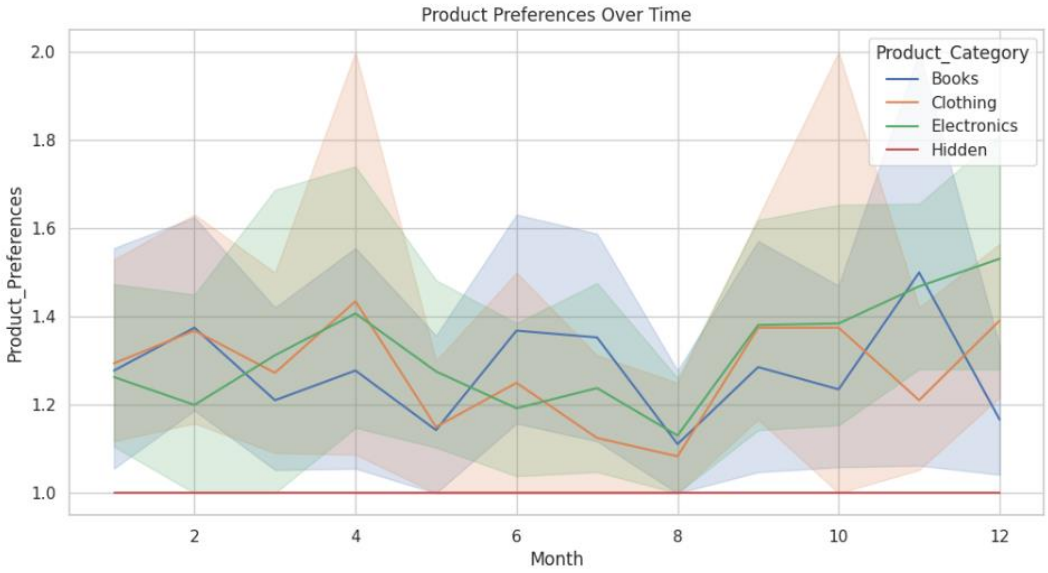


- Fluctuations in Spending: There are significant fluctuations in average spending per month throughout the years. This could be indicative of varying economic conditions, changes in store pricing, or shifts in customer purchasing power.

- Range of Spending: The average spending per month ranges from close to 0 to just over 500 units. The variations in spending seem to increase slightly over time, with some of the higher spendings occurring in more recent years.

- No Clear Long-term Trend: There does not appear to be a clear long-term trend in average spending. While there are spikes and dips, there is no consistent pattern of increase or decrease.

- High Spending Peaks: There are certain years where the average spending peaks significantly, which might be due to specific events or promotions that drove higher sales.

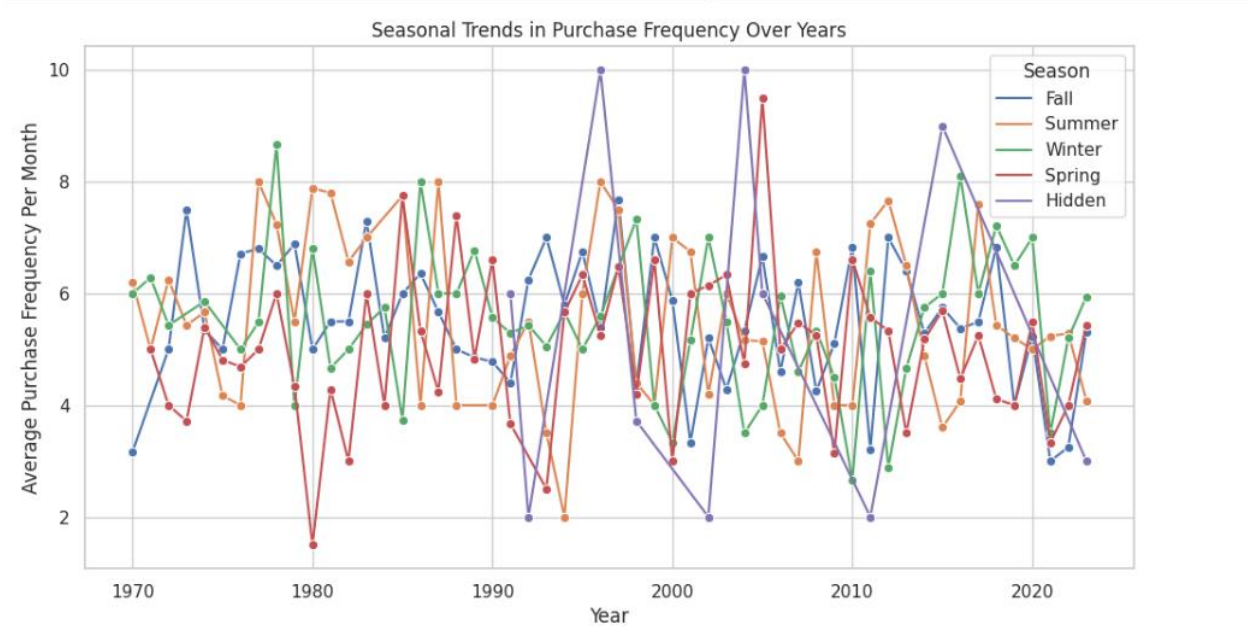


- Monthly Trends: Each product category exhibits some variability in preference throughout the year. All categories show a change in preferences, suggesting seasonal trends or promotional impacts.

- Peak Preferences: There are noticeable peaks in preferences for certain categories during specific months. For instance, the Books category shows a significant peak around the 10th month, which could be indicative of an event like a literary festival or back-to-school season.

- Clothing Category: This category shows a relatively stable preference throughout the year with slight increases around the 2nd and 10th months, which could correspond to seasonal changes like spring and autumn fashion lines.

- Electronics Category: There is a noticeable peak in the Electronics category around the 4th month and another rise towards the end of the year, possibly aligned with new tech releases or holiday shopping.

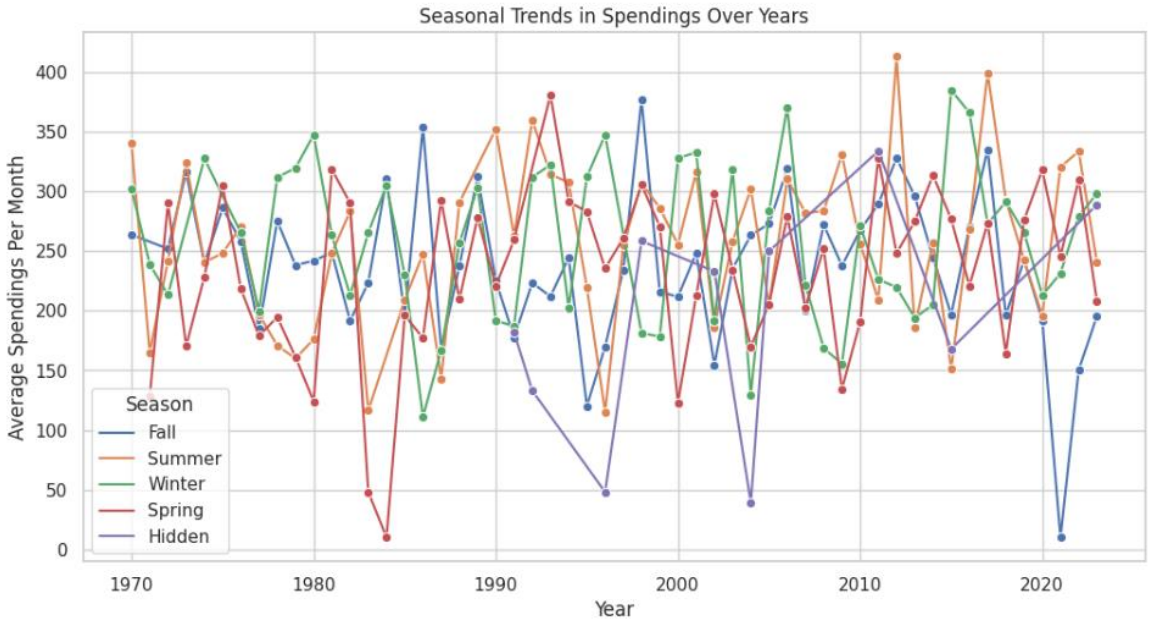


- Seasonal Variability: All seasons show variability in purchase frequency over the years. This indicates that seasonality does affect purchasing behavior to some extent.

- Trend Analysis: There isn't a consistent upward or downward trend across the seasons, suggesting that while purchase frequency may vary by month and year, it doesn't consistently increase or decrease over time.

- Seasonal Peaks and Troughs: Each season has its peaks and troughs. For example, it appears that Winter and Spring have several points where they peak to the highest purchase frequencies.

- Comparison Between Seasons: At different points in time, certain seasons have higher purchase frequencies than others, which could be linked to factors like holidays, store promotions, or weather-related shopping behavior.

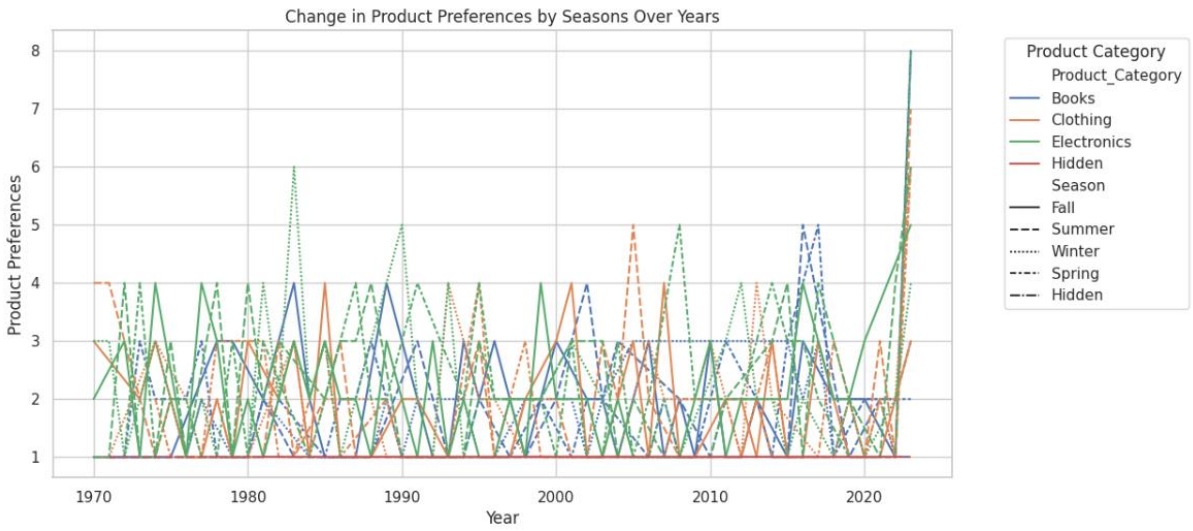


- Seasonal Variations: There are clear seasonal variations in spending, with all seasons exhibiting fluctuations. This suggests that seasonality impacts consumer spending behavior.

- Spending Peaks: There are noticeable peaks in spending during certain seasons across various years. These could correspond to holiday seasons or specific seasonal promotions that encourage higher spending.

- Winter Spending: The Winter season frequently shows higher spending peaks, which might be related to end-of-year holidays.

- Spring and Fall Spending: Spring and Fall also show significant spending, but with less pronounced peaks than Winter.



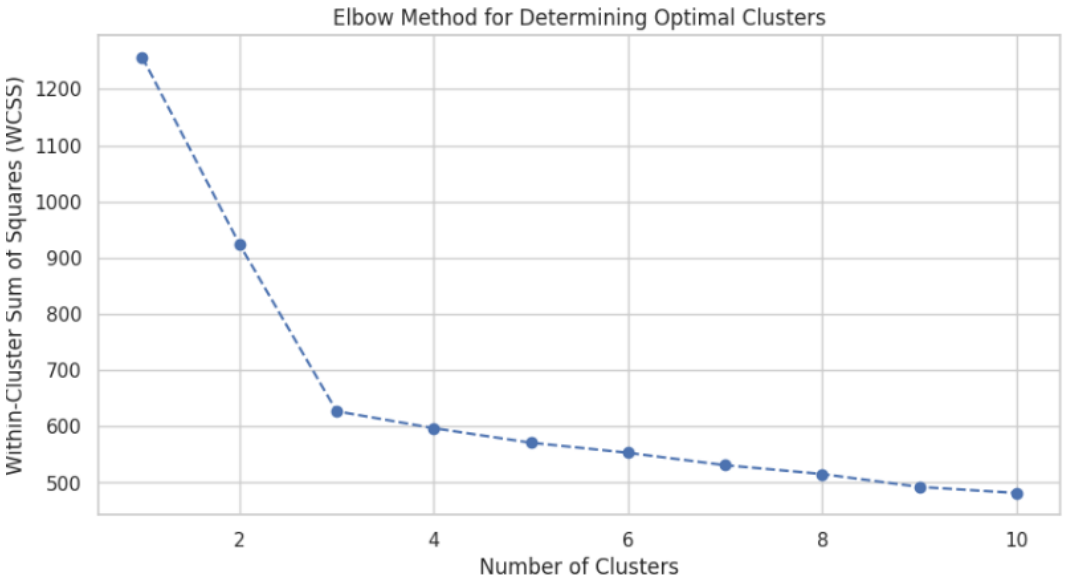
- High Variability: The chart shows high variability in product preferences over the years across all product categories and seasons. This suggests that consumer preferences are not constant and fluctuate significantly over time.

- Seasonal Patterns: While the data is noisy, there seem to be certain years where seasonal preferences spike, which could correlate with specific events or promotions driving interest in certain product categories during particular seasons.

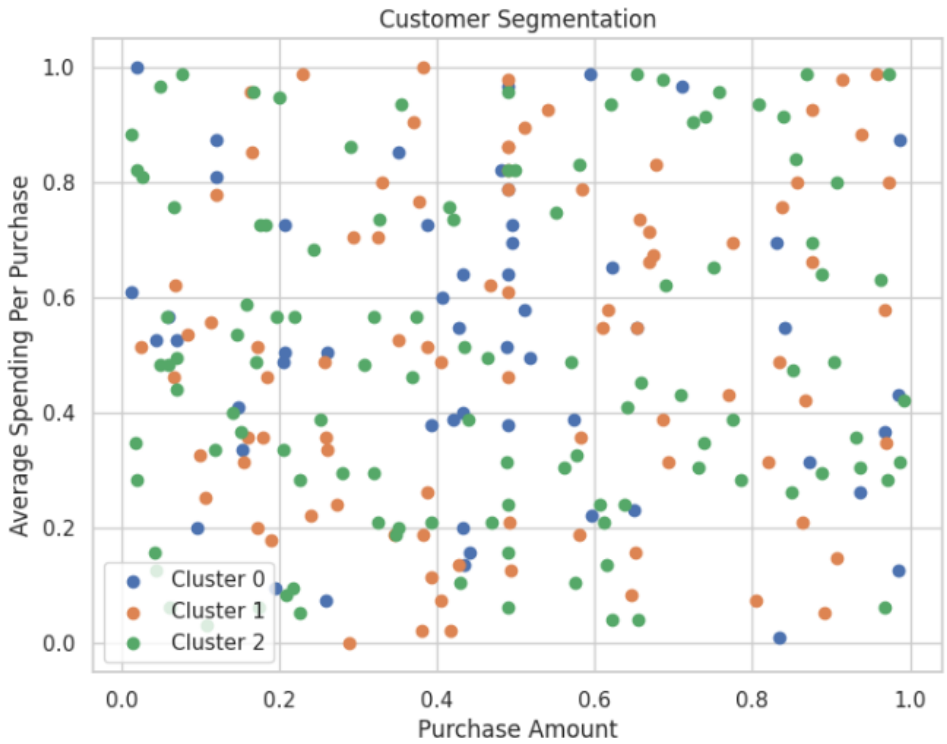
- Product Categories: The preferences for Books, Clothing, and Electronics all show distinct trends over the years, with some seasons showing higher preferences than others.

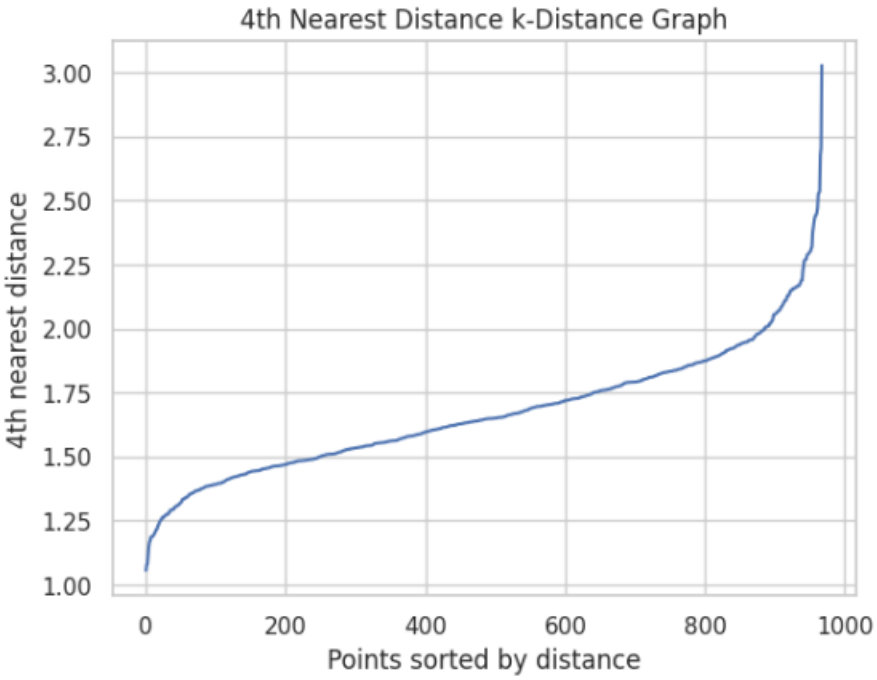
- Significant Changes: There are points in time where certain seasons and categories exhibit sharp increases or decreases in preferences, which could be due to a variety of factors, such as changes in market trends, introduction of new products, or strategic shifts in the store's focus.

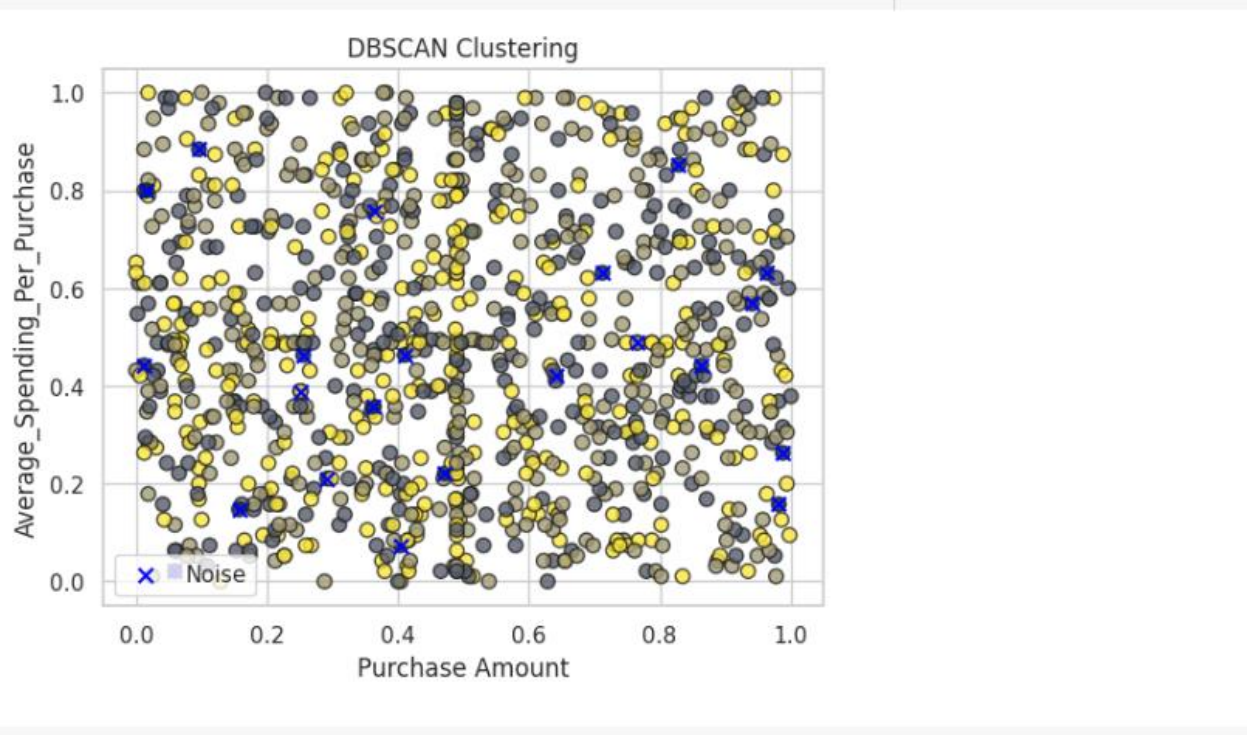
Recent Years: In the most recent years on the chart, there's a noticeable increase in the variability of preferences. This could be due to increased competition, changes in consumer behavior, or a diversification of the store's product offerings.



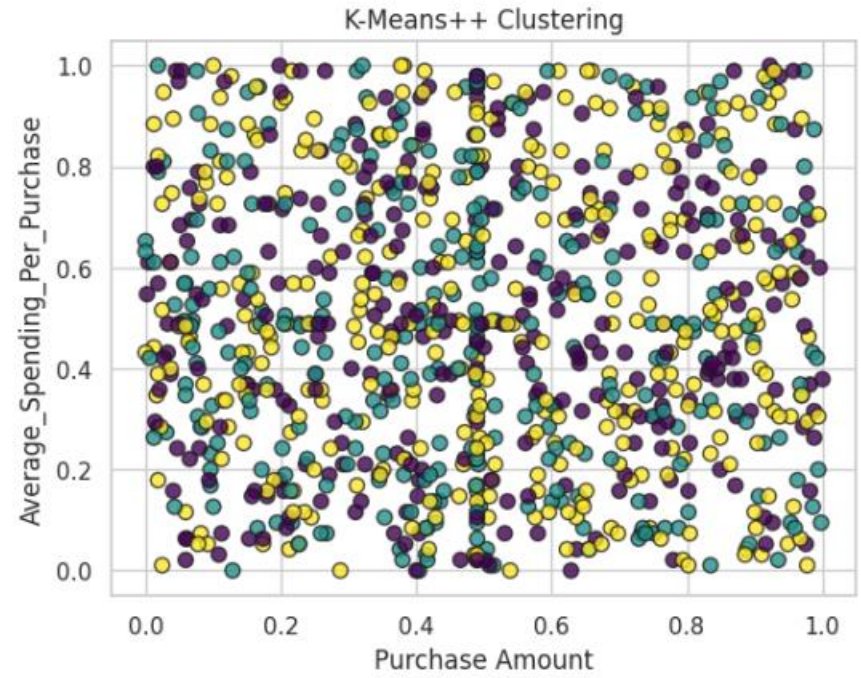
Elbow Method suggests that 3 clusters are best.




4th Nearest Distance k-Distance Graph
4th nearest distance
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0
200
400
600
800
1000
Points sorted by distance



X represents the outliers in the data



Module # 4

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.

2. **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.

3. **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.