Data-Science

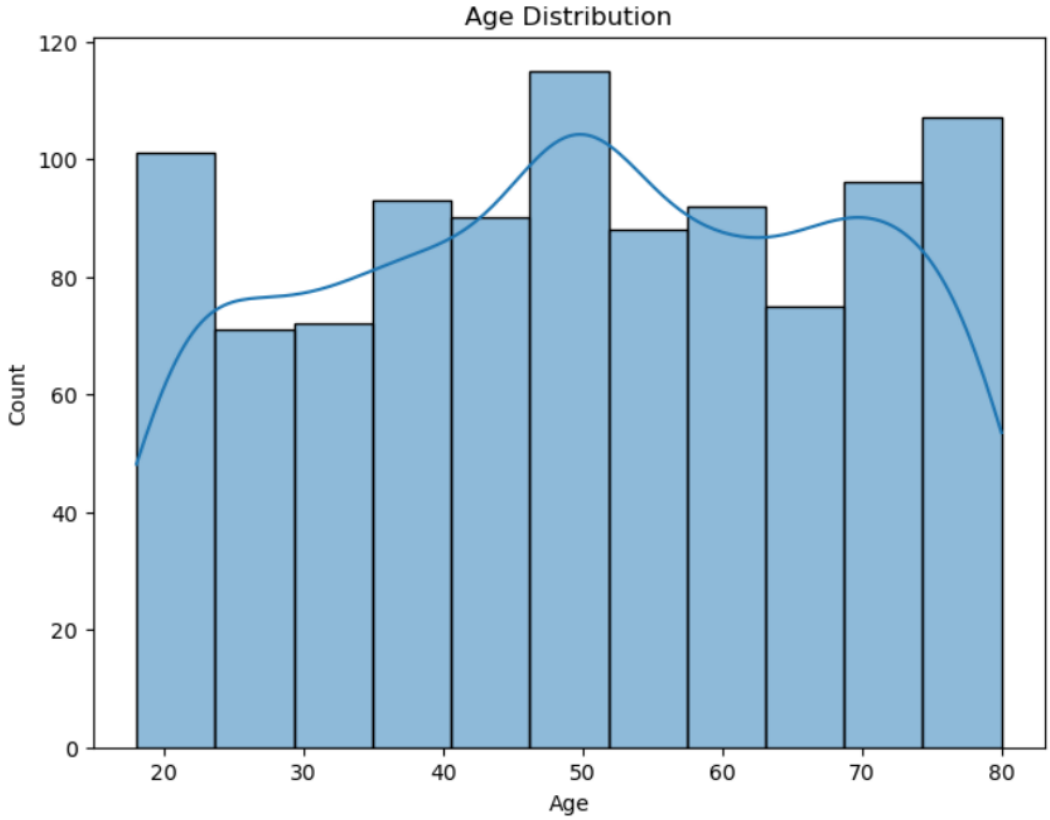
PROJECT

Muhammad Subhan Attique – Abdul Basit Ahsan

Module 2:

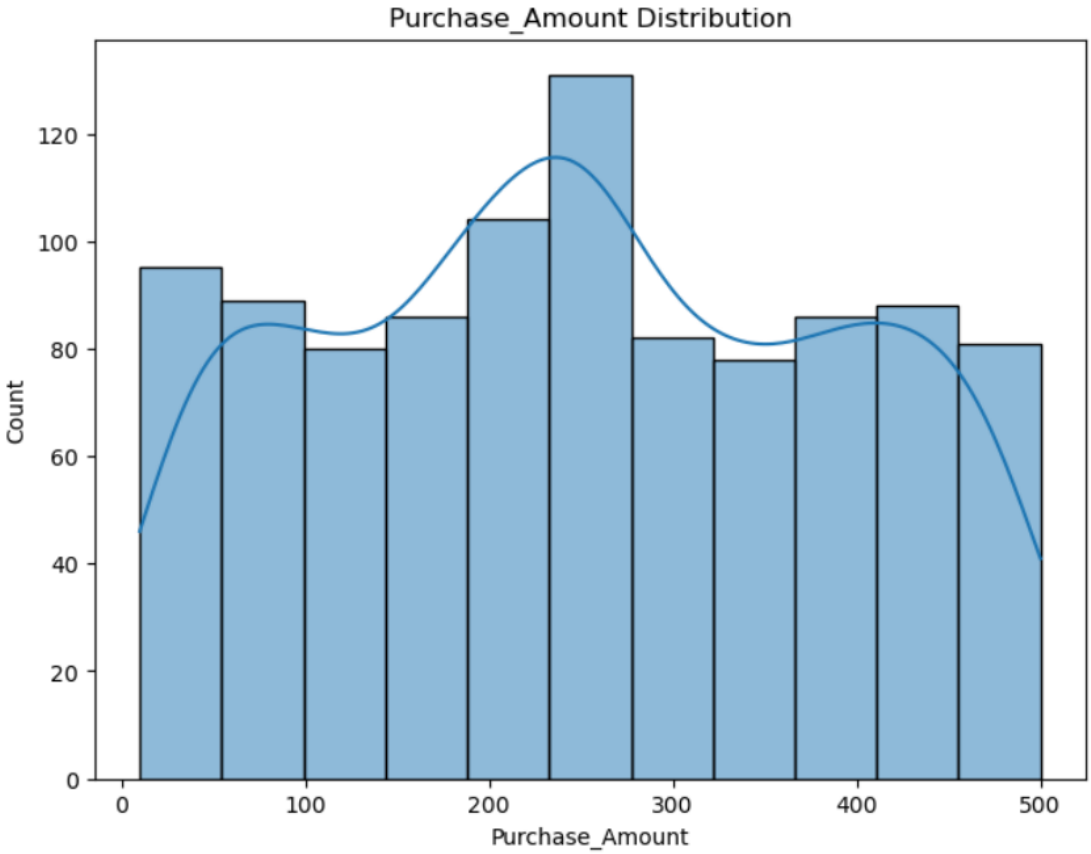
Uni-variate Analysis:

Univariate analysis refers to the statistical analysis technique used to describe and analyze individual variables in isolation. It involves the examination of a single variable at a time to understand its characteristics, distribution, central tendency, dispersion, and other relevant properties. This analysis helps in summarizing and exploring the features of a single variable, usually through descriptive statistics, visualizations such as histograms, box plots, and summary tables.



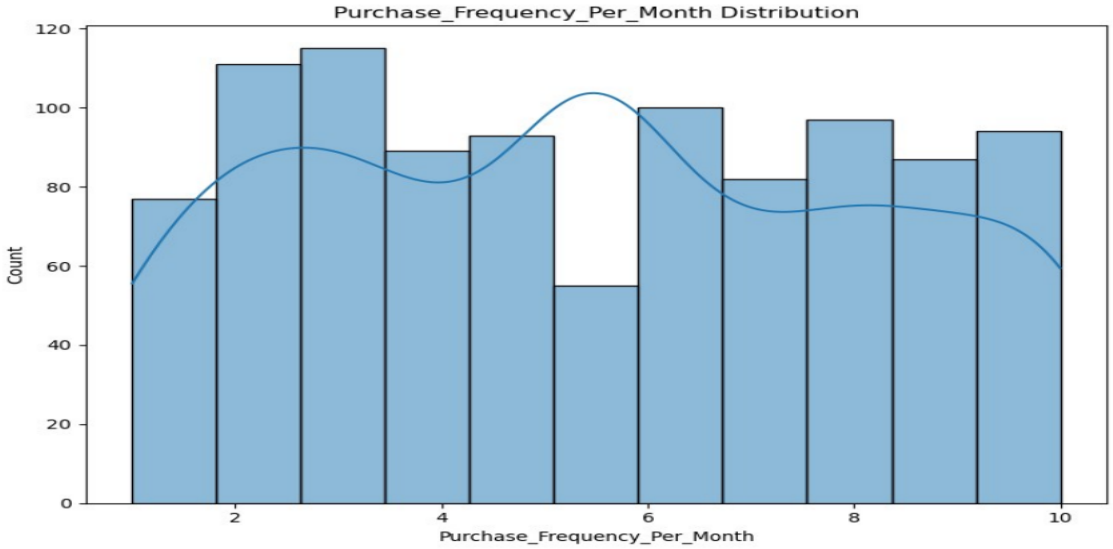
This is the **Histogram** of Age distribution of customers present in the provided dataset, according to this histogram:

- Most of the customers present in the dataset are of age 50 as its bar is higher than all bars.
- Least of the customers present in the dataset are of age 25 as its bar is lower than all bars.
- KDE of the histogram is drawn in order to determine if data is skewed or not and it is not.



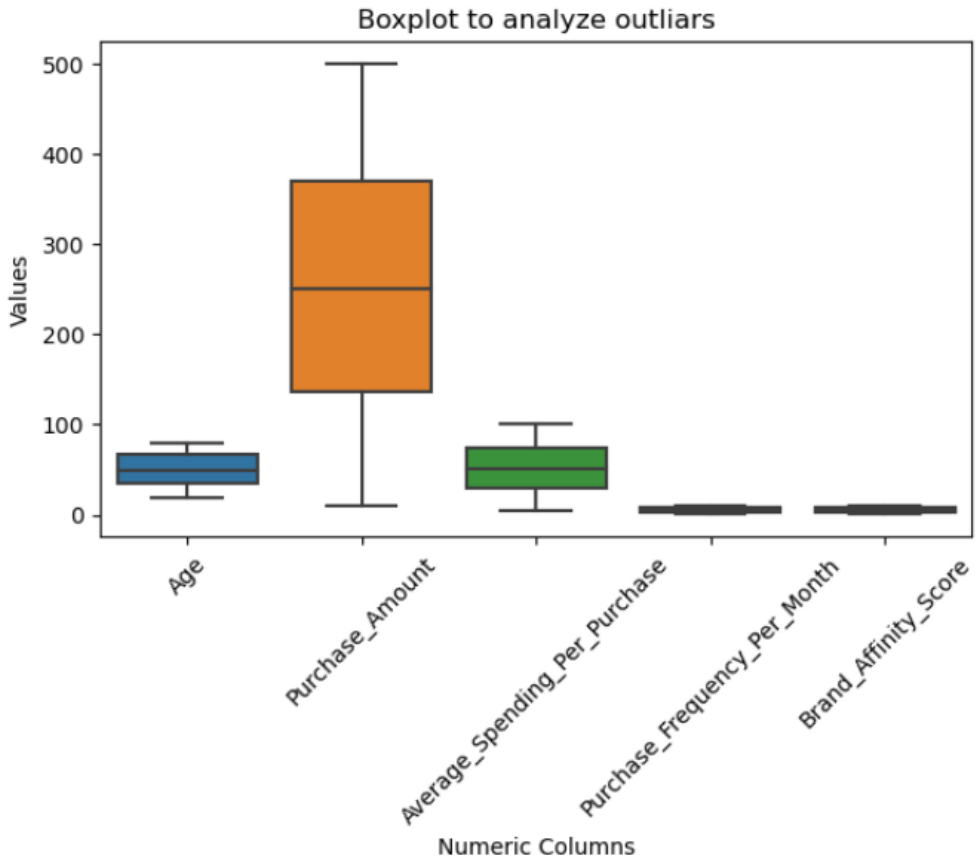
This is the Histogram of purchase amount distribution among the customers present in the dataset, according to this histogram:

- Most frequent age group of customers who make the most purchases are in the 30-40 age group.
- Least frequent age group of customers who make the least purchases are in the 20-29 and 60-79 age.
- The number of customers making purchases gradually decreases as age increases.



This is the Histogram of purchase frequency per month that describes most sales made in a month and according to this histogram:

- In months 2-3 there are most of the purchases made other than any of the month.
- In the 5th month, there are the least of the purchases made.
- Purchases made in months differ and vary between all of the months.

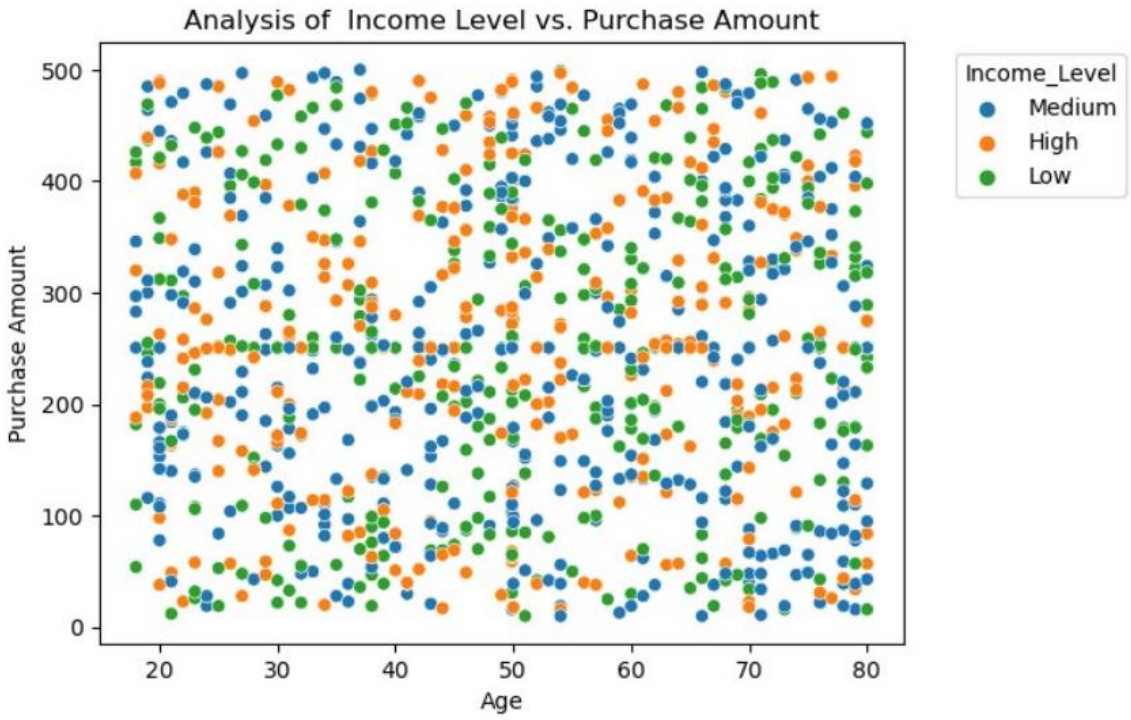


These are all the boxplots of numeric columns present in the dataset, boxplots are primarily made to see the range and spread of data and to analyze outliers. According to these plots:

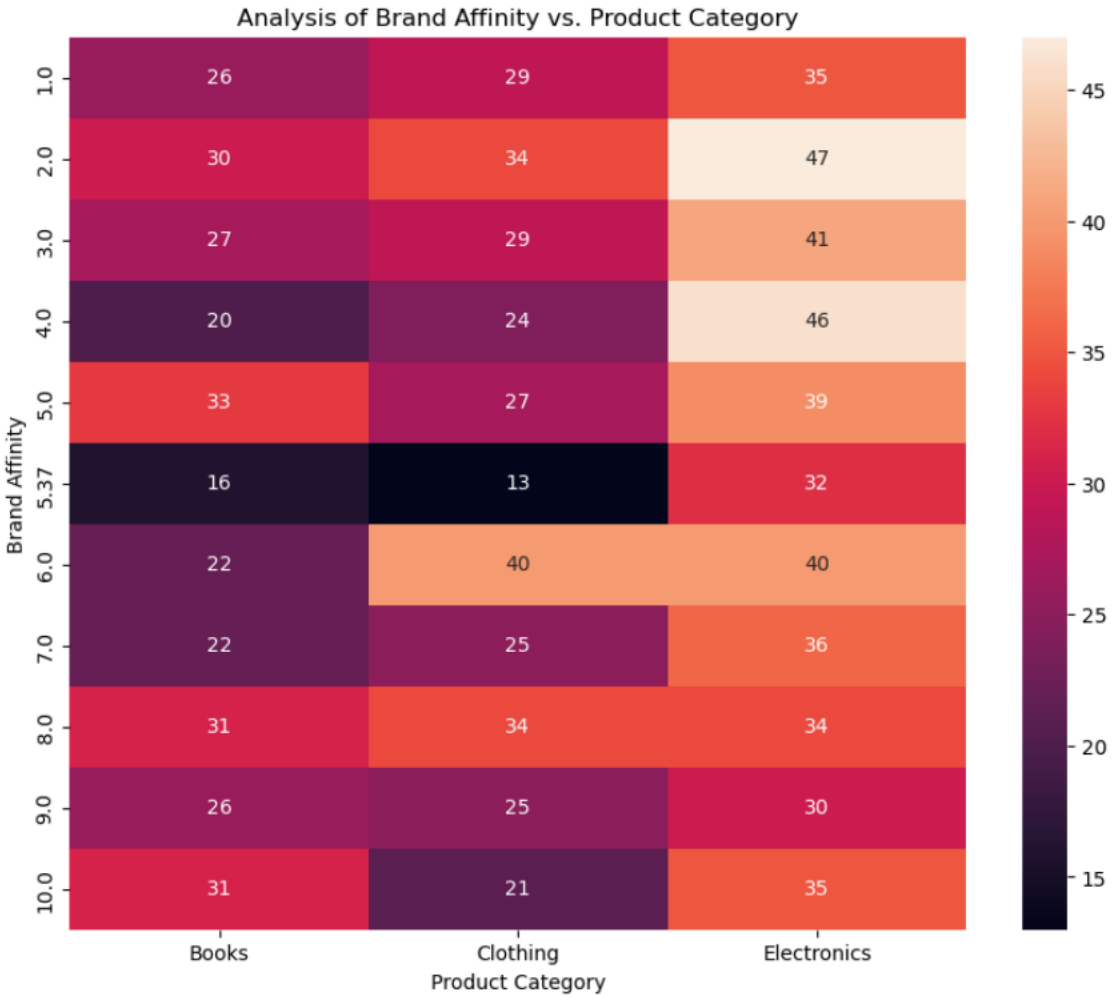
- Data of purchase amount is in the most range as its whiskers extend the most.
- Data of age and average spending per purchase is also spread but not more than purchase amount.
- Data of purchase frequency per month and brand affinity score is least spread among all columns.
- No outliers exist in any of the columns.

Bi-variate Analysis:

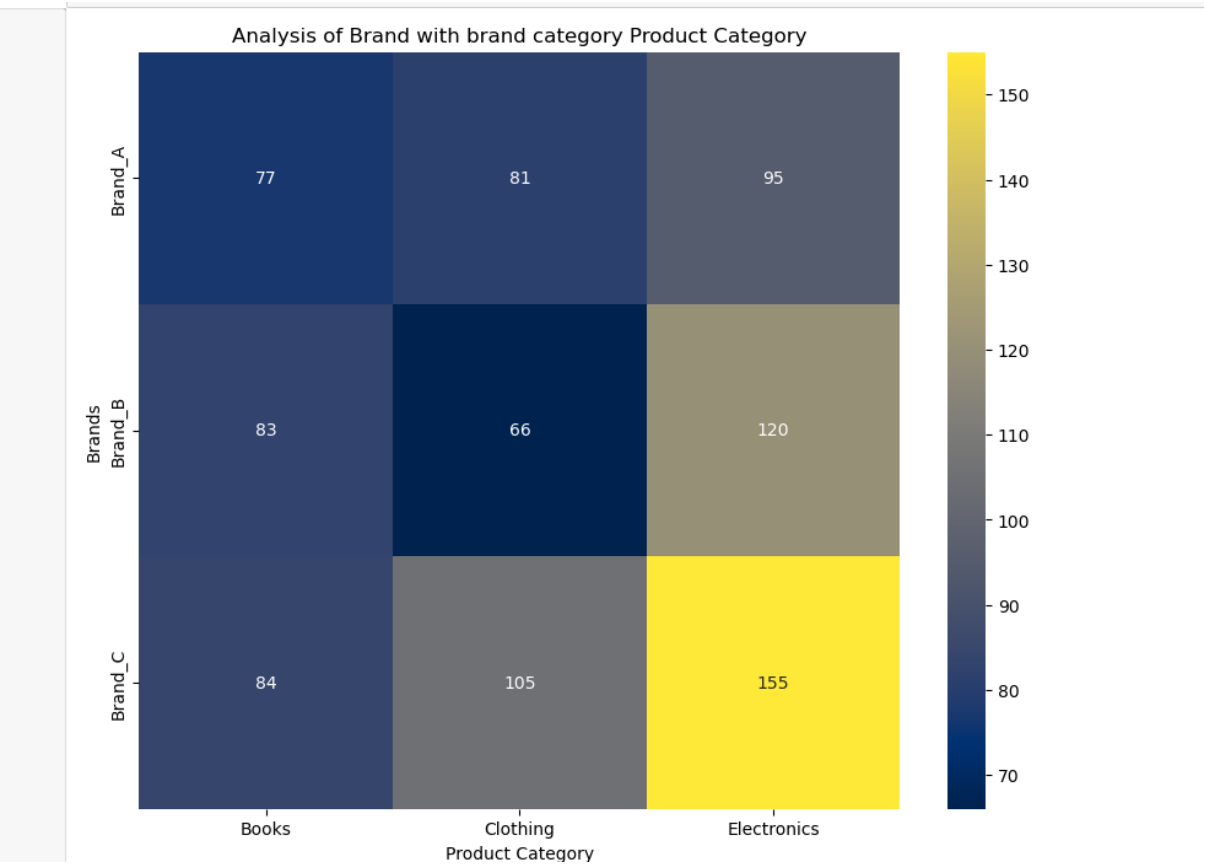
Bivariate analysis is a statistical method that involves the simultaneous examination and analysis of two variables to determine relationships, associations, or correlations between them. It explores how changes or variations in one variable are related to changes or variations in another variable.



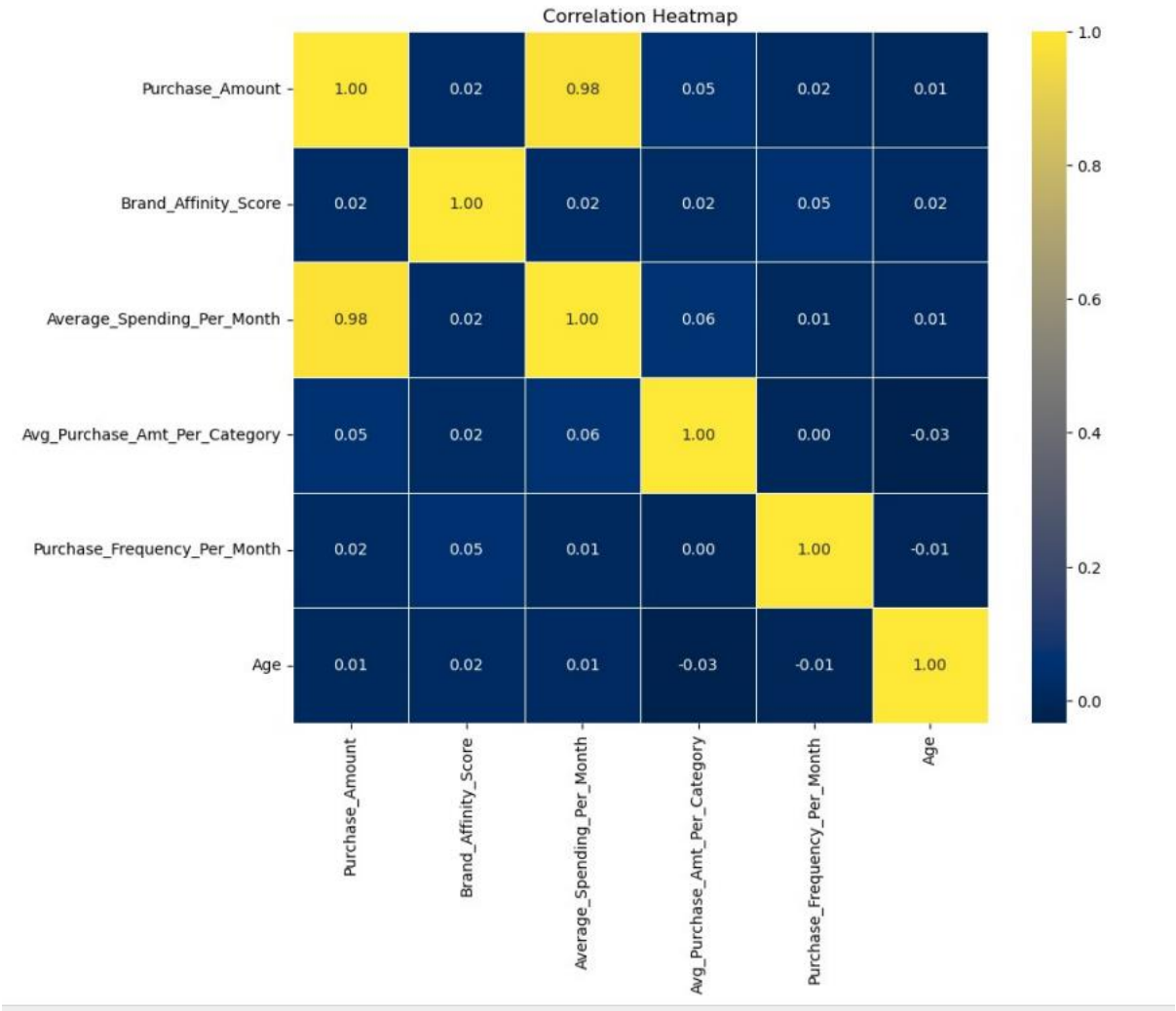
This is the scatter plot of analysis of income level and purchase amount, and it shows that customers of all ages and all income levels make purchases of all amounts. People with low-income levels are most scattered in region of purchase amount 0-250 and a slight of customers are above this purchase amount. Customers with high income scatter all along the plot in all of the purchase amounts and same for the customers with medium income.



This heatmap shows the qualitative analysis of different brands category over brand affinity score the color scale shows the different shades and different colors to represent the different values daker shade represent the higher greater value like clothing have the highest brand affinity which is at 5.37 . Similarly the light color shows the lower value for same category like Electronics has the lowest value of brand affinity at 2.0 which light color of scale 47. This comparison shows the dept of the categorical data which have higher values and which have low values at the dataset. Books data shows that they either have large nor low affinity score and also average brand affinity score.



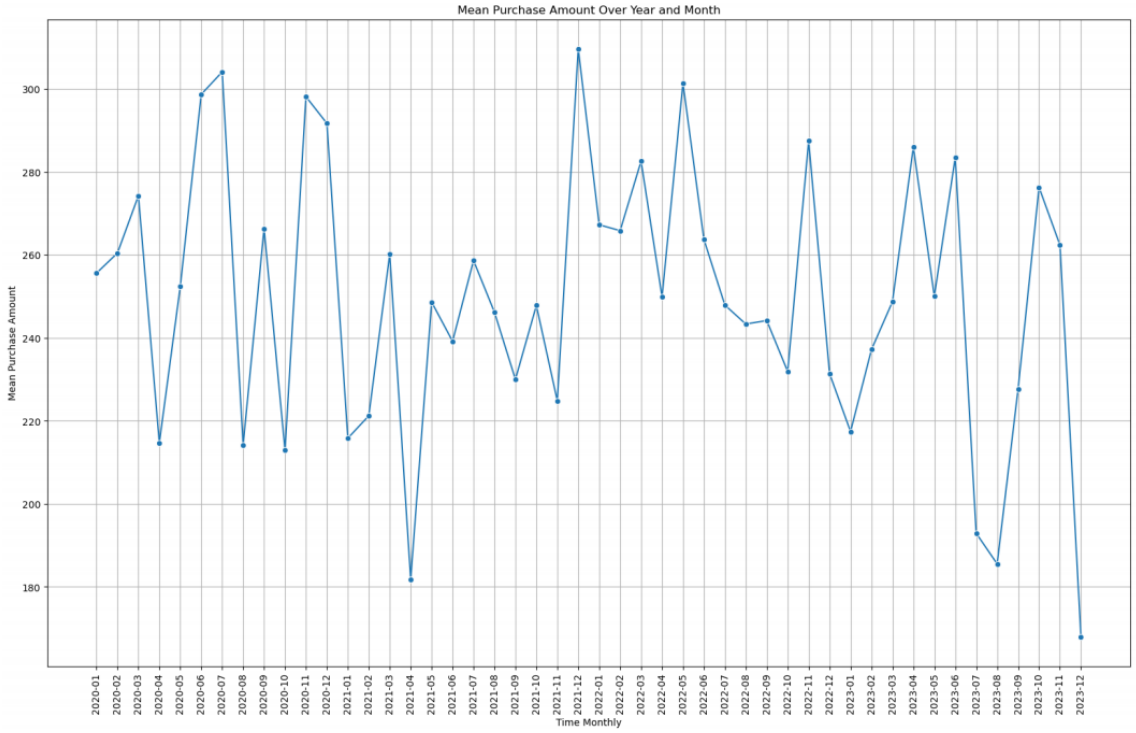
This heatmap show the overall the comparison of the brands over the brand affinity score and give the idea which brand category have the highest score and which brand category have the lowest score and different levels. The color depth of the color shows high and low value for same data . Brand C have the lowest value at the electronics which is equal to the 155. Brand B have the highest value for at clothing.



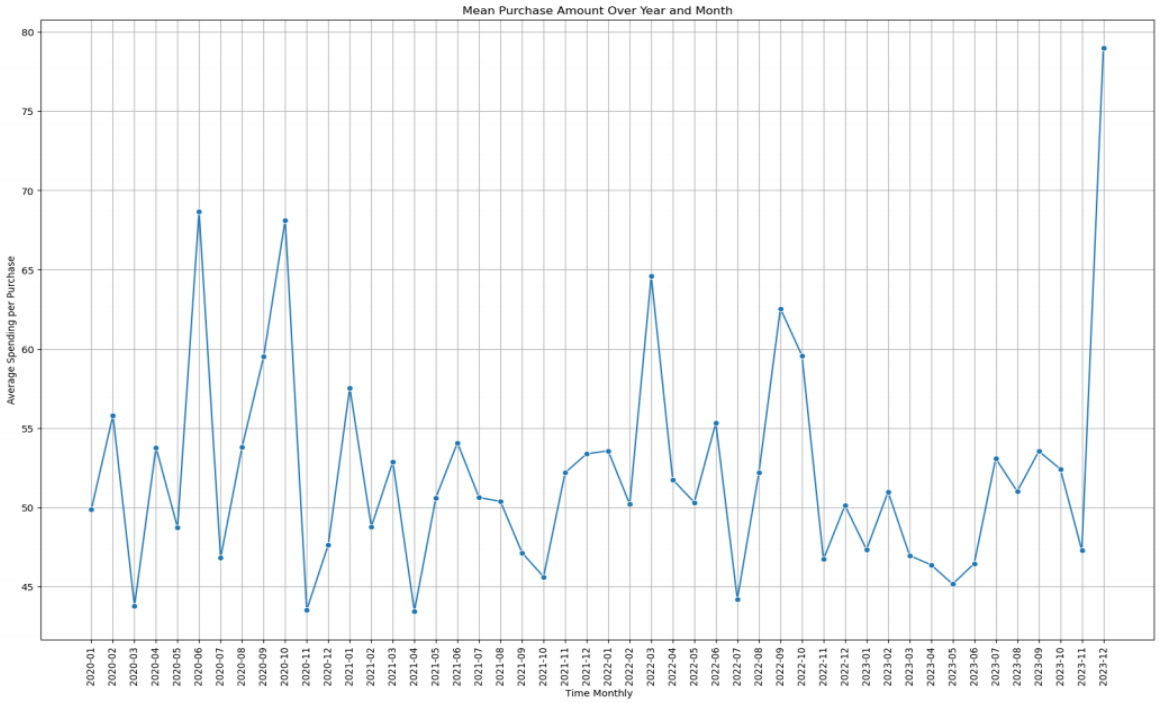
This heatmap indicate the overall correlation of each data. Darker shades or stronger colors on the heatmap indicate a stronger correlation between the corresponding variables. Positive correlations (values closer to 1) are often represented by shades of blue. Negative correlations (values closer to -1) are often represented by shades of red or brown. Heatmaps are particularly useful for visualizing complex datasets with numerous variables. They provide a quick overview of relationships without the need to analyze numerical correlation coefficients directly. After identifying correlated variables, further analysis, such as regression analysis or clustering, can be performed to gain a deeper understanding of the relationships

Temporal Analysis:

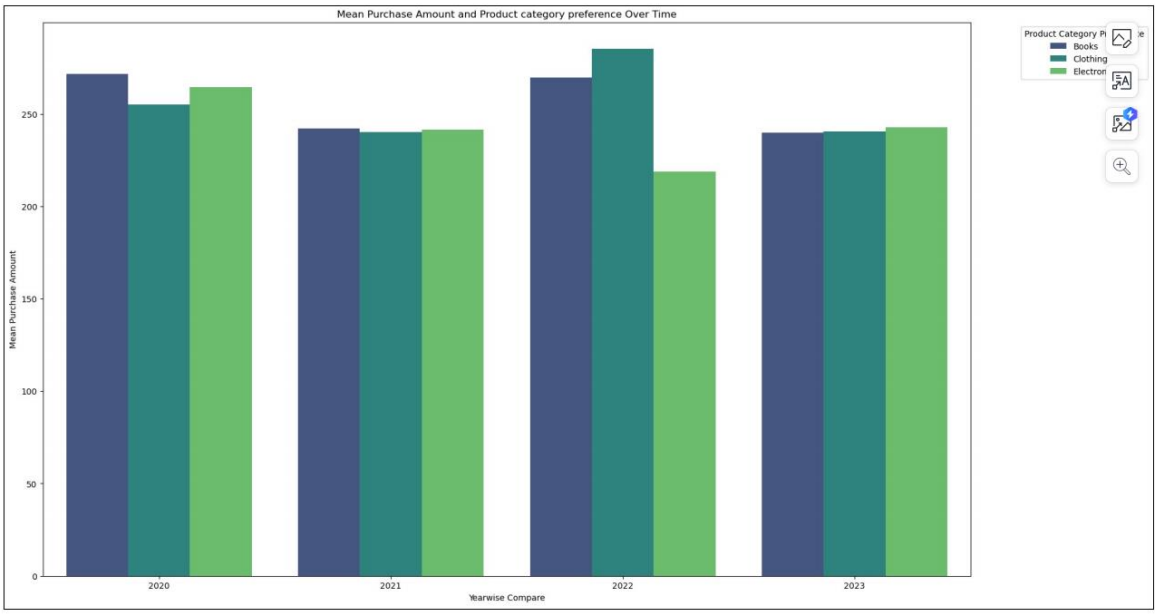
Temporal analysis is a method of examining data over a specific period or timeframe to identify patterns, trends, or variations that occur with respect to time. It involves analysing how data changes or evolves over time intervals, such as seconds, minutes, hours, days, months, or years. Temporal analysis can reveal seasonal patterns, trends, cyclic behaviour, and other time-related characteristics within datasets.



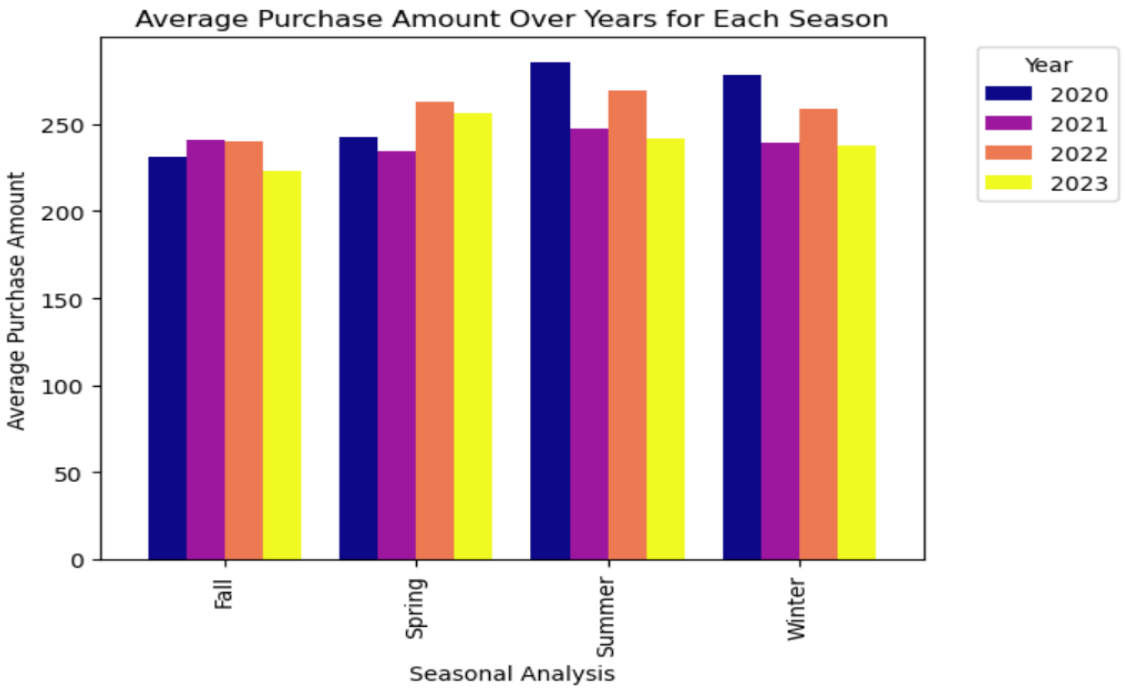
The average purchase frequency has been fluctuating over time. There are noticeable peaks and troughs in the graph, indicating that the purchase frequency has been increasing and decreasing over time. The graph shows that the purchase frequency was highest in 2022-01 and lowest in 2023-12. The purchase frequency has been increasing and decreasing since 2022-07 and is currently at a moderate level.



The average spending per purchase has been seen fluctuating throughout 2020-2023, there are some noticeable peaks and troughs throughout the graph but average spending per purchase can be seen highest in 2023-12 and lowest in 2021-04 but it has been mostly increasing since 2021-10 and sometimes decreasing.



This bar plot shows the average purchase amount and product category preferences over the years and it can be seen that books and clothing were at peak in terms of purchase amount in the year 2022 and electronics were most down at that time and in 2023 all of three product categories were even but electronics was a bit higher.



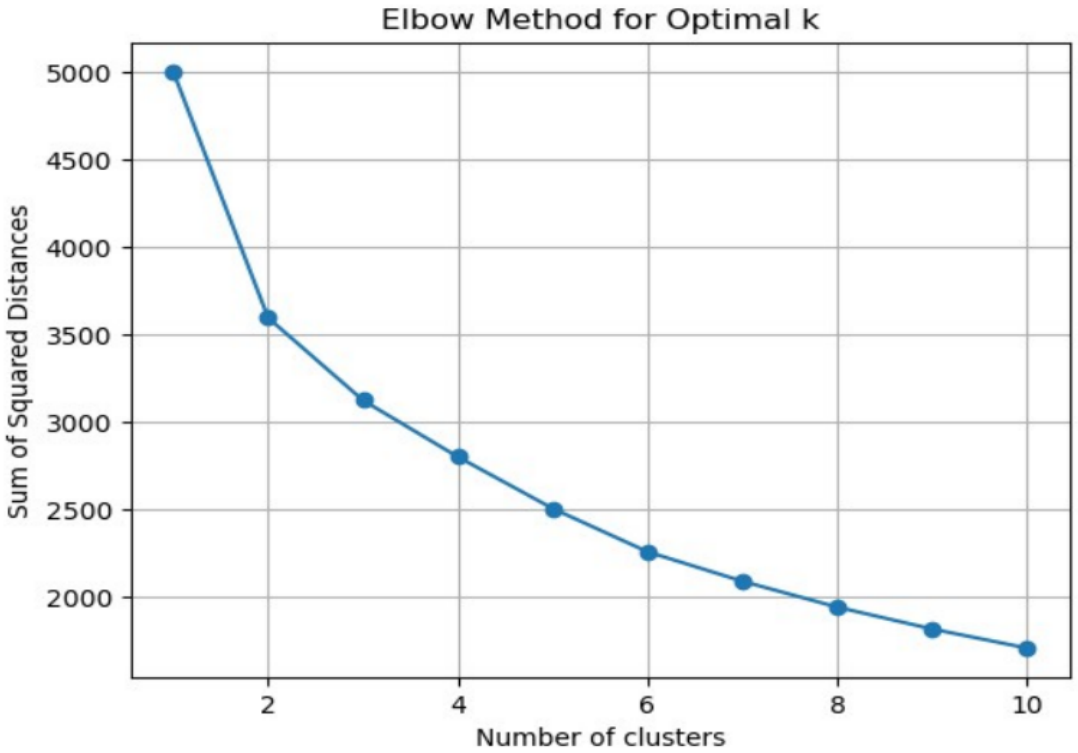
This bar plot shows the seasonal analysis of purchase amount over the years in all seasons and this can be seen that in summer 2020, there were the most purchases made and lowest purchases were made in 2023 fall.

Module-3

K-Means Clustering:

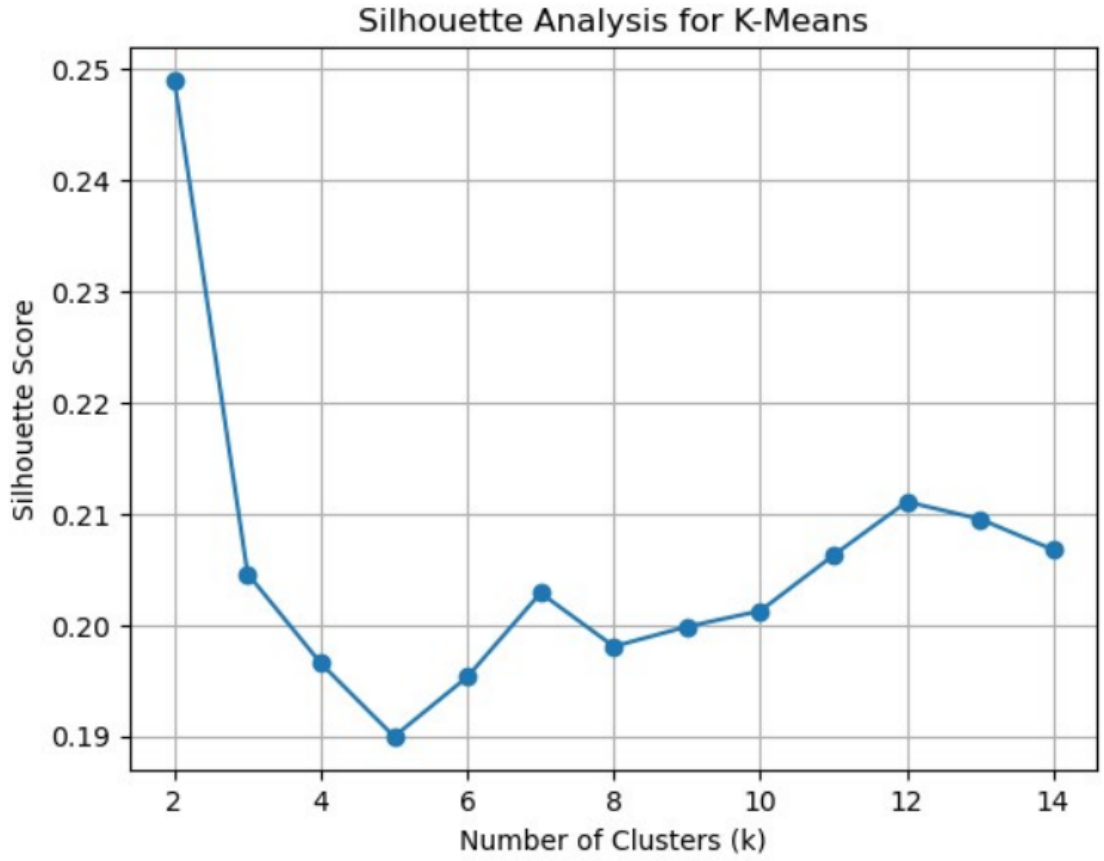
K-Means Clustering is an unsupervised machine learning algorithm that groups similar data points into clusters. The algorithm iteratively assigns points to the cluster with the nearest centroid and updates centroids based on the new assignments. The user defines the number of clusters ('k'). Efficient and scalable, K-Means is widely used in customer segmentation, image compression, and other applications. It requires selecting an appropriate 'k' value and can be sensitive to initial centroid placement. Despite its simplicity, K-Means offers valuable insights into underlying data patterns, aiding in various analyses and decision-making processes.

Elbow Method:



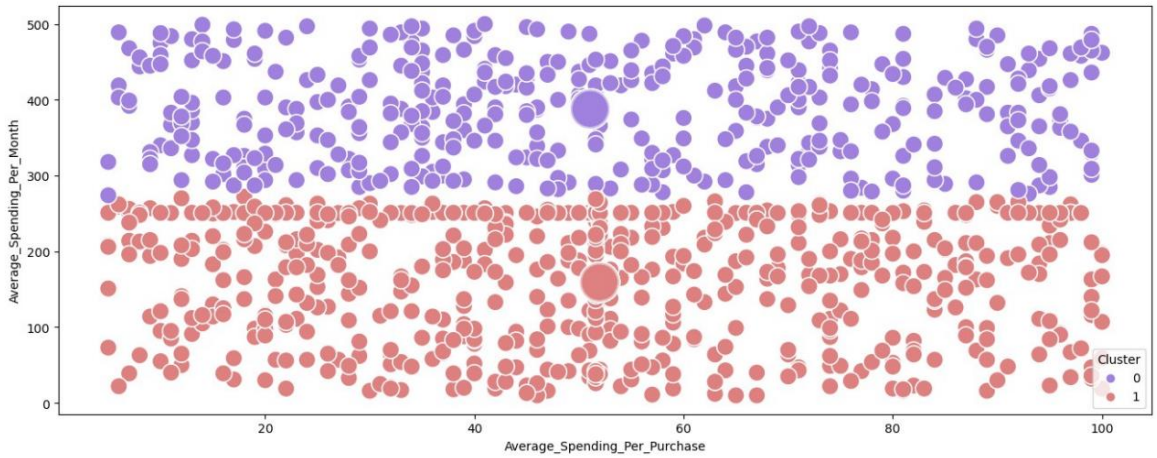
The Elbow Method is a technique to find the optimal number of clusters (k) in K-Means clustering. It involves plotting the within-cluster sum of squares against different k values and looking for an "elbow" point where the rate of decrease in sum of square slows down. This point indicates a balance between underfitting and overfitting, helping to choose the optimal number of clusters for a dataset. This graph represents the bending of the line at k=2. So we create two cluster of the numeric data in the dataset. This graph is created by taking all the numeric data from the dataset like Age, Purchase Amount, Average_Spending_Per_Purchase, Purchase_Frequency_Per_Month. This allows to identify the clustering for this dataset.

Silhouette Analysis:

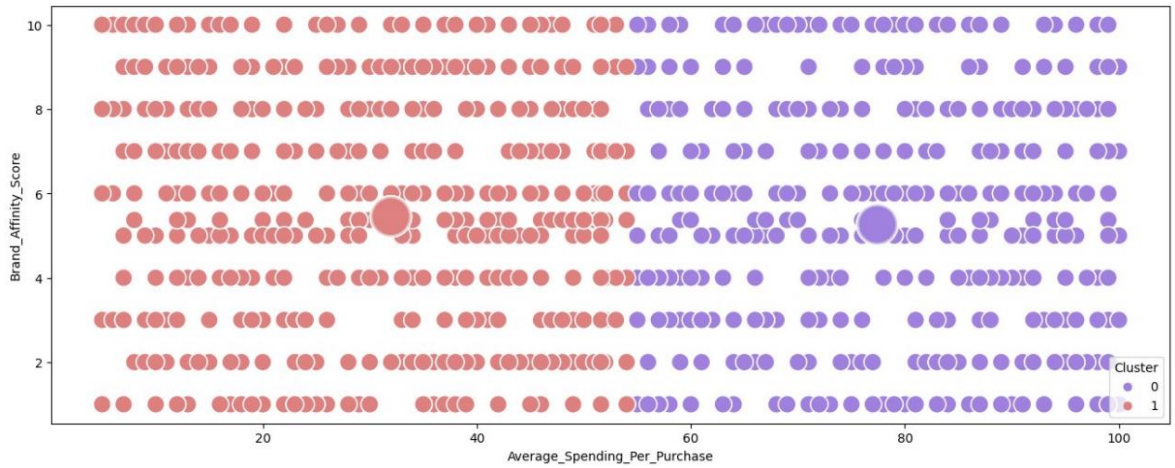


Silhouette analysis is a technique used to determine the optimal number of clusters in a clustering algorithm, such as K-Means. It provides a measure of how well-separated the clusters are, helping to identify the most appropriate number of clusters for a given dataset. The silhouette score ranges from -1 to 1, where higher values indicate better-defined clusters. The Silhouette analysis for this dataset shows that the optimal value of cluster starting with k=2. The highest value or the value closes to 1 give the value of the clustering for dataset. This analysis show the exact number of clustering formed in the dataset. This method give us the best value at k=2

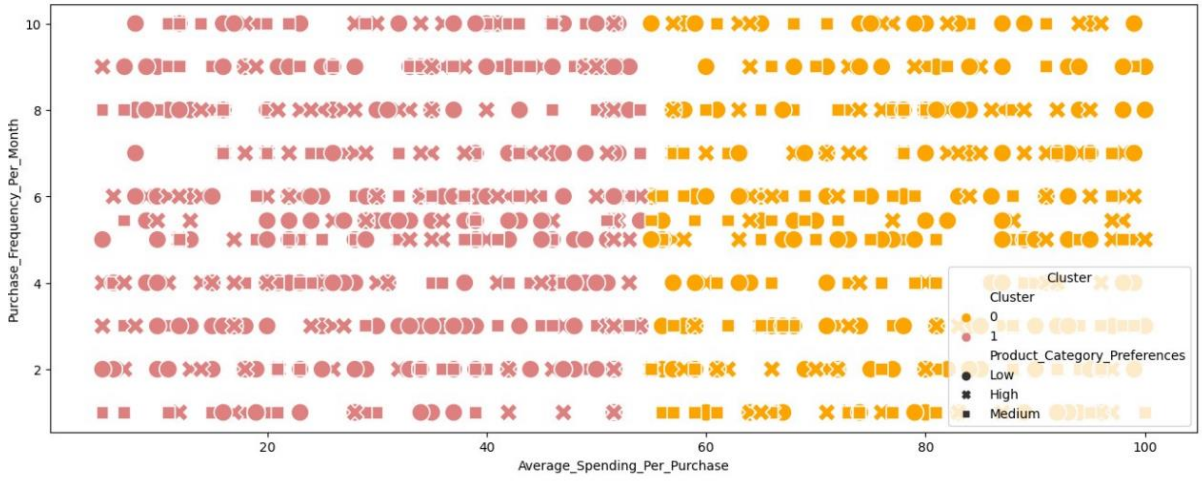
Clustered Points for Age with Purchase amount



Clustered Points for Brand and Average Purchase



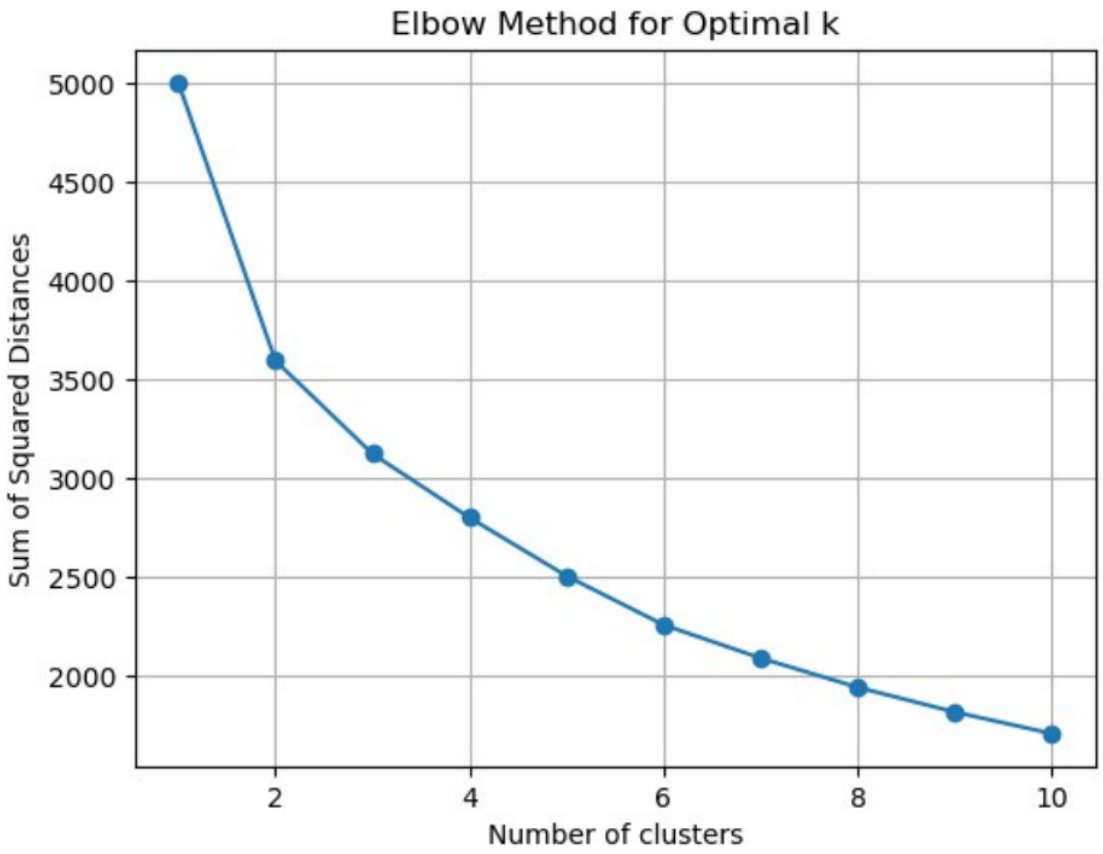
Clustered Data to identify the Purchasing Behavior



K-Means++ Clustering:

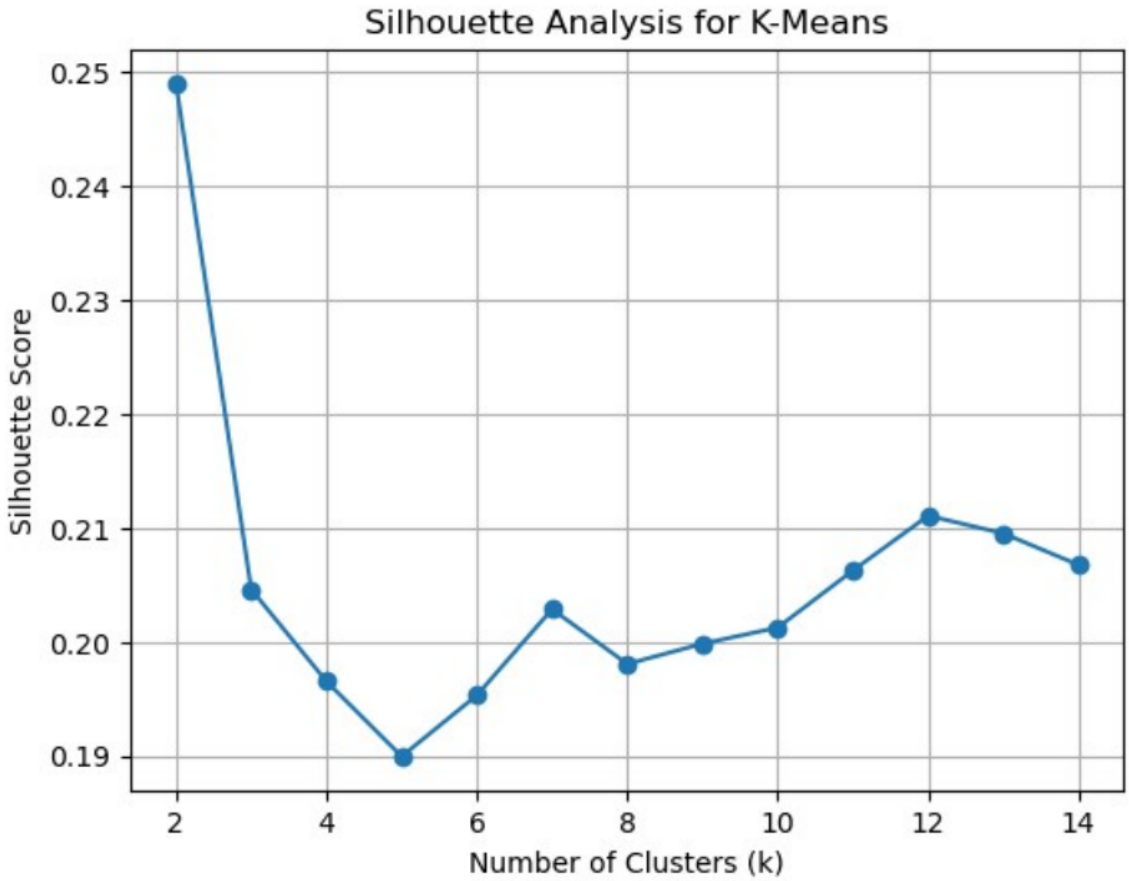
K-Means++ is an enhanced initialization technique designed to improve the performance of the K-Means clustering algorithm. Unlike the random initialization of centroids in traditional K-Means, K-Means++ strategically selects initial centroids. It begins by choosing one centroid randomly and subsequently selects additional centroids with a probability proportional to their squared distance from the nearest existing centroid. This initialization method spreads out the initial centroids more effectively, reducing sensitivity to initial placements and leading to faster convergence. K-Means++ is widely utilized in practice for its ability to enhance the stability and accuracy of clustering results, making it a preferred choice in various data analysis applications.

Elbow Method:



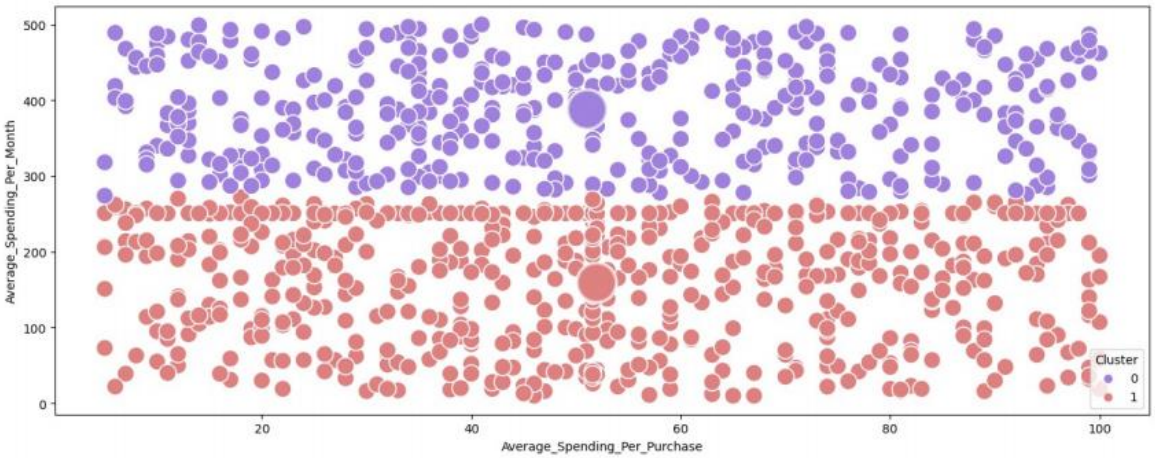
K-Means++ helps in selecting better initial cluster centroids, leading to faster convergence and potentially more accurate results. When applying the Elbow Method with K-Means++, the graph represents the bending of the line at the optimal k value. The bending observed at k=2, it indicates that creating two clusters is optimal for your dataset. The numeric data in the dataset Age, Purchase Amount, Average_Spending_Per_Purchase, and Purchase_Frequency_Per_Month contribute to the clustering process, allowing you to identify distinct groups within your data. the Elbow Method with K-Means++ is a valuable tool for determining the optimal number of clusters by analyzing the rate of decrease in within-cluster sum of squares.

Silhouette Analysis:



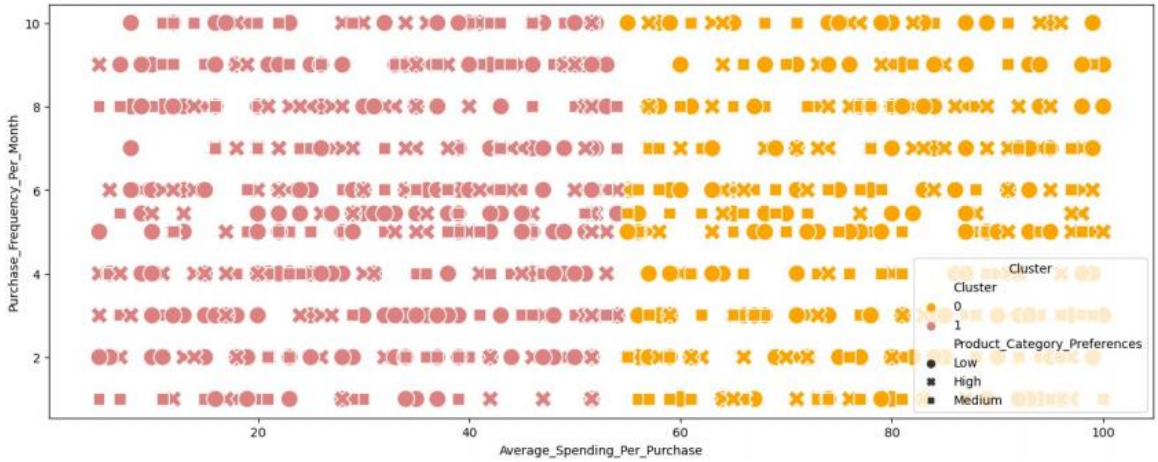
Silhouette analysis is a technique used to determine the optimal number of clusters in a clustering algorithm, such as K-Means. It provides a measure of how well-separated the clusters are, helping to identify the most appropriate number of clusters for a given dataset. The silhouette score ranges from -1 to 1, where higher values indicate better-defined clusters. The Silhouette analysis for this dataset shows that the optimal value of cluster starting with k=2. The highest value or the value closes to 1 give the value of the clustering for dataset. This analysis show the exact number of clustering formed in the dataset. This method give us the best value at k=2

Clustered Points for Age with Purchase amount



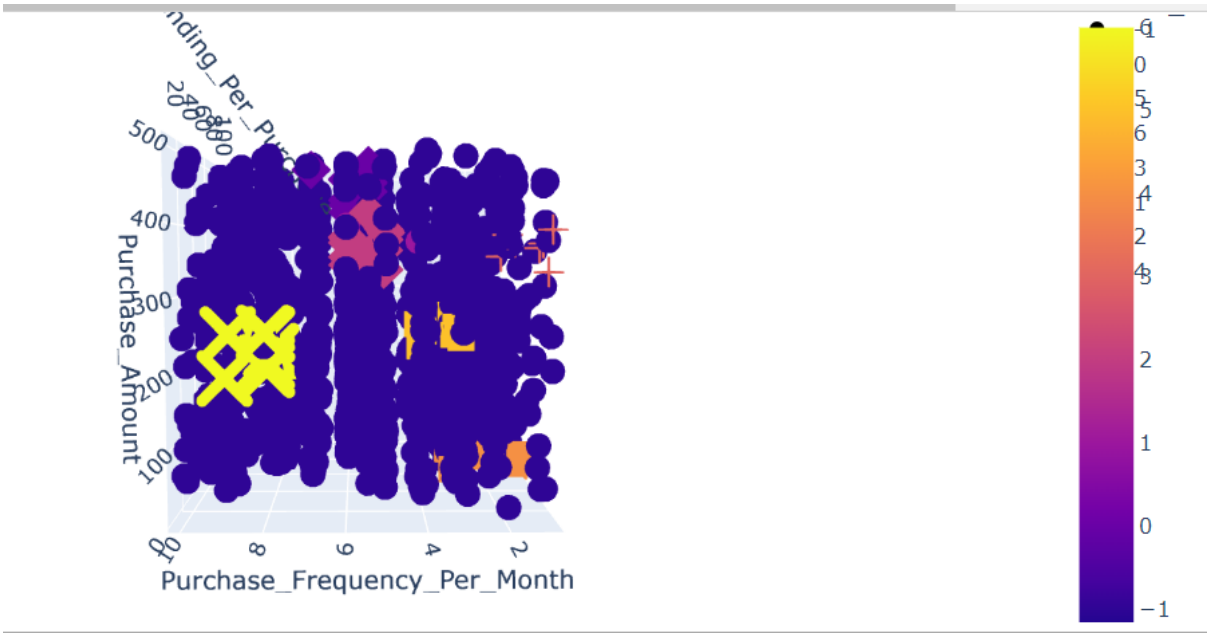
Clustered Points for Brand and Average Purchase

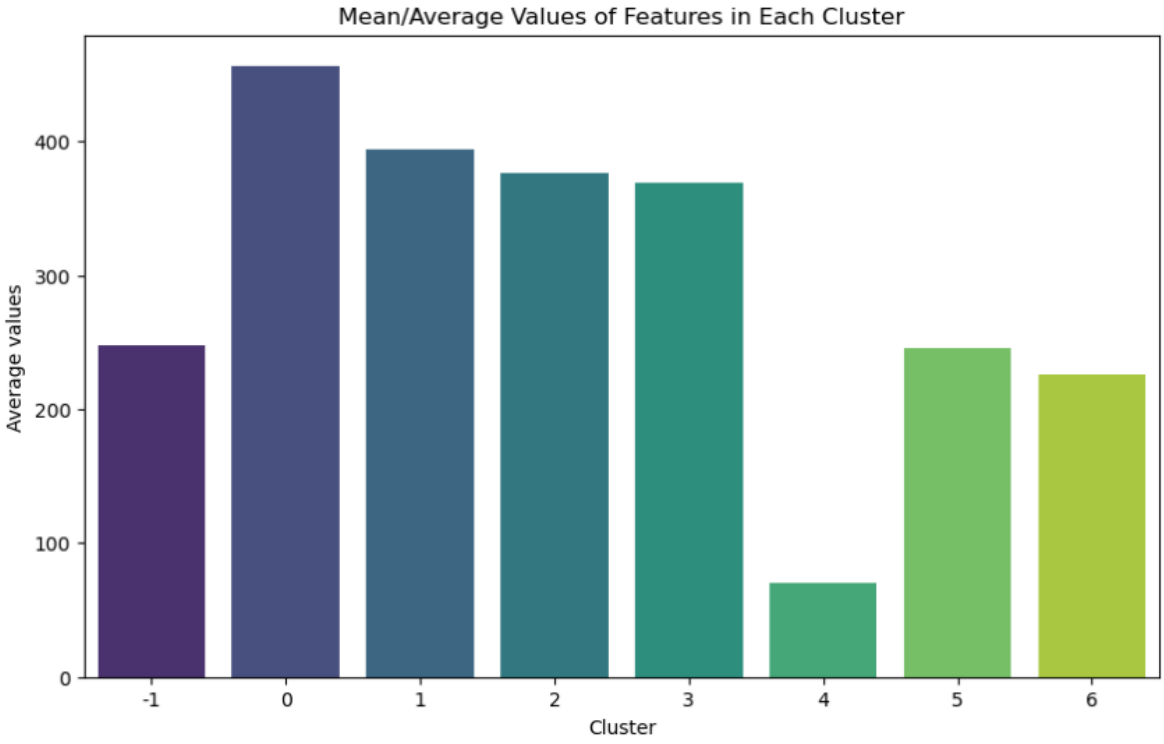
Clustered Data to identify the Purchasing Behavior



DBSCAN Clustering:

Distribution of clusters and core points:





This bar plot shows average values of each feature in each cluster and it indicates that cluster 0 has the maximum number of average values of each feature and cluster 4 has the minimum number of average values of each feature.

Module 4:

Comparison of all techniques:

Silhouette Score for K-Means Clustering: 0.03326499218535305

- Positive score indicates that all of the clusters are well separated.

Calinski-Harabasz Score for K-Means Clustering: 13.683537926826611

- Higher score suggests that cluster are dense, and there are well-separated clusters.

Davies-Bouldin Index for K-Means Clustering: 7.230089065332304

- Lower index indicates better clustering.

Silhouette Score for DBSCAN Clustering: 0.10017518189000402

- High score indicates well-defined, dense clusters.

Calinski-Harabasz Score for DBSCAN Clustering: 8.840340366926815

- Low score suggests fewer compact clusters.

Davies-Bouldin Index for DBSCAN Clustering: 4.061674740363794

- Very low index indicates well-separated clusters.

Silhouette Score for K-Means++ Clustering: 0.03326499218535305

- Same as k means.

Calinski-Harabasz Score for K-Means++ Clustering: 13.683537926826611

- Same as k means.

Davies-Bouldin Index for K-Means++ Clustering: 7.230089065332304

- Same as k means.

Metrics:

Optimal Parameters:

eps: 0.5

min_samples: 1

Best Silhouette Score: 0.10017518189000402

In the context of the silhouette scores:

Silhouette score evaluation is a metric used to assess the goodness of clusters produced by clustering algorithms. The silhouette score measures how similar an object is to its cluster (cohesion) compared to other clusters (separation). The score ranges between -1 to 1, where:

Near 1 or positive value: Indicates that the sample is far away from neighbouring clusters, implying that the clustering configuration is appropriate.

0: Indicates overlapping clusters.

Near -1: Suggests that the sample might have been assigned to the wrong cluster.

The silhouette score of our data is positive means that clusters are appropriate to some extent, there are other metrics for evaluation of clusters in all clustering techniques, such as Calinski-Harabasz score and Davies-Bouldin index.

Advantages/Disadvantages:

K-Means:

Advantages:

- K-Means is relatively straightforward to understand and implement.
- It is computationally faster and more efficient, making it suitable for large datasets.
- Performs well with many variables.

Disadvantages:

- The final clusters can be sensitive to the initial placement of centroids, which may result in different outcomes.

- It requires prior knowledge of the number of clusters (K), which may not always be known.
- Sensitive to outliers, which can significantly affect cluster centroids and boundaries.

DBSCAN:

Advantages:

- DBSCAN can discover arbitrary-shaped clusters and does not require specifying the number of clusters in advance.
- It is robust to outliers and noise in the data.
- Can identify clusters of varying shapes and densities.

Disadvantages:

- Performance can be affected by the choice of epsilon (ε) and minPts parameters.
- Struggles with clusters of varying densities or when the clusters are within each other.
- Struggles with high-dimensional data due to the curse of dimensionality.

K-Means++:

Advantages:

- K-Means++ selects initial centroids that are farther apart, reducing sensitivity to the initial centroids' placement.
- Often converges faster and provides better clustering results compared to random initialization.
- Minimizes the likelihood of converging to suboptimal solutions.

Disadvantages:

- Slightly more computationally intensive due to the additional step of selecting better initial centroids.
- While it reduces sensitivity to initial centroids, it doesn't eliminate the risk of converging to suboptimal solutions.

Conclusion:

1. **Customer Segments within the Electronics Section:**

Clusters Identified:

Based on the clustering analysis, distinct customer segments within the electronics section have been identified using KMeans and DBSCAN.

KMeans and KMeans++ produced similar clusters, while DBSCAN identified noise and a primary cluster, there was also differences in number of clusters made in DBSCAN, Kmeans and Kmeans++.

2. **Key Factors Differentiating Customer Segments:**

Differentiating Factors are:

- Average spending per purchase.
- Age
- Income Level
- Purchase frequency per month.
- Brand affinity score.
- Product category preferences.

3. **Purchasing Behaviour Patterns:**

Cluster Characteristics:

Customers in different clusters exhibit varied purchasing behaviour patterns.

For instance, one cluster may consist of high-spending, infrequent customers, while another may include frequent, lower-spending customers.

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

Recommendations:

- Customize marketing campaigns based on the identified segments.
- Introduce loyalty programs targeting specific customer behaviours. So, they can become your primary and permanent customers.
- Implement personalized product recommendations to enhance user experience.

5. **Potential Applications of Clustering Results:**

Personalized Recommendations:

Leverage clusters for personalized product recommendations. For example, sales.

Targeted Marketing:

Design targeted marketing campaigns tailored to specific customer segments.

Loyalty Programs:

Develop loyalty programs that resonate with the preferences of each cluster. Like make offers that suits a specific cluster.

Further Analysis and Investigations:

Conduct further analysis to gain a deeper understanding of customer behaviour.

Explore additional factors that may optimize the performance of the electronics section.

7. Conclusion:

The clustering analysis provides valuable insights into customer segmentation within the electronics section. Understanding these segments can drive targeted strategies for customer retention and sales growth.