The Sales Spectrum Diving into the Supermarket Transaction

1st Duaa Khan
Ghulam Ishaq Khan Institute
2021143

***Abstract*—This report delves into the multifaceted aspects of consumer purchasing patterns within a supermarket setting, utilizing a robust dataset reflecting various transactional details. The primary objective is to unearth subtle correlations, distinct customer segments, and predictive indicators that could empower more tailored marketing strategies and enhanced decision-making processes. By employing advanced data preprocessing techniques, the study ensures a refined dataset, facilitating the application of machine learning models and statistical analysis.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Retail analytics in supermarkets focuses on using data to understand and predict customer behavior, optimize operations, and increase sales. This field integrates statistical and machine learning techniques to analyze sales data and customer demographics, playing a crucial role in shaping decisions from product placement to promotional strategies.

Analyzing supermarket sales data is very important in today's highly competitive market. It helps firms find out what sales trends are there, makes customers happier and raises operational efficiency. As big data and analytics develop, the supermarkets are getting closer to understanding human behavior more deeply. This definitely brings a lot of benefits for consumers in terms of service but also directly contributes to profitability.

II. Methodology

A. Dataset

The dataset in discussion is a comprehensive collection of supermarket sales data, encompassing various transactional details that reflect customer purchasing behavior. It includes a range of variables such as Invoice ID, branch, city, customer type, gender, product line, unit price, quantity, tax, total sales, date, time, payment method, cogs, gross margin percentage, gross income, and customer ratings. Each entry represents a unique transaction, providing insights into what items are bought, by whom, and the financial aspects of the purchase. This rich dataset allows for an in-depth analysis of sales trends and customer preferences, making it a valuable resource for understanding market dynamics and enhancing decision-making in the retail sector.

B. Detailed Methodology

The project begins with meticulous data cleaning, addressing missing values, anomalies, and standardizing variables. This ensures a solid foundation for analysis. Visual explorations follow, using histograms, box plots, and heatmaps to understand the data's distribution and identify initial patterns or outliers. This exploratory phase is critical for setting the analytical direction and confirming data integrity.

Subsequently, the project delves into advanced techniques. Association Rule Mining and FP-Growth algorithms uncover prevalent purchasing combinations, while Sequential Pattern Mining reveals temporal purchasing trends. Clustering techniques like K-means and DBSCAN segment customers into meaningful groups, facilitating targeted strategies. These methods collectively provide a detailed understanding of customer preferences and behaviors.

Finally, the project employs Classification and Regression models to predict customer segments and spending patterns, respectively. It leverages algorithms suited to the data's characteristics, ensuring accurate predictions. Alongside, outlier detection techniques are applied to single out and examine unusual data points, maintaining the overall quality and reliability of the study's insights. This comprehensive approach allows for robust, actionable conclusions, guiding strategic business decisions.

C. Data Cleaning and Preparation

In the initial phase of the analysis, the study emphasized rigorous data cleaning to ensure the integrity and quality of the supermarket sales dataset. This process involved identifying and handling missing values, either through removal or imputation, to prevent gaps in data from skewing the analysis. However, there were no missing values in our dataset. However, I parsed the Date column into additional Year, Month, Date and YearMonth columns. Outliers were detected and addressed to avoid undue influence on the results. I also performed Transaction Encoding. This process converted the transactional data, which listed categories purchased in each transaction, into a format that Apriori and FP-Growth algorithms can interpret. Each unique item or category across all transactions was encoded as a separate feature, and transactions were then represented as binary vectors indicating the presence or absence of each item. Categorical variables underwent encoding, transforming them into a numerical format suitable for algorithmic analysis, while ensuring that the ordinal nature

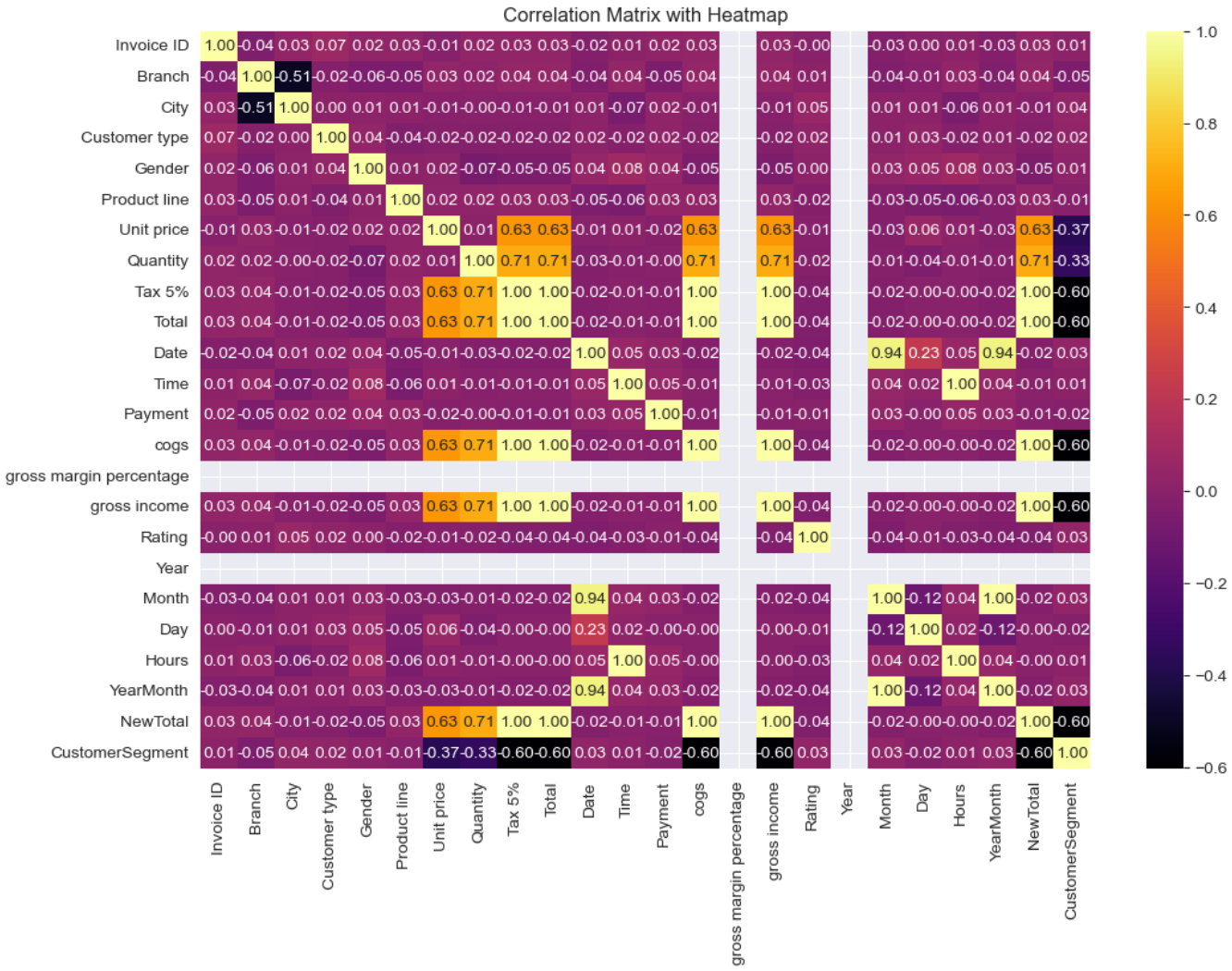

Fig. 1. Correlation Matrix

of certain variables was maintained. Numerical features were standardized, mitigating the risk of scale discrepancies impacting the models. This comprehensive cleaning process laid a solid foundation for the subsequent exploratory and predictive analyses, ensuring that the insights derived from the data were accurate and reliable. Each step was meticulously documented and executed, reflecting best practices in data cleaning and preparation. The train-test split is a critical step in model evaluation, ensuring that a data mining model generalizes well to new, unseen data. By dividing the dataset into separate training and testing sets, it helps in detecting over-fitting, enables robust hyper-parameter tuning, and allows for the objective comparison of different models. This process ensures the model's performance is reliable and credible, reflecting its true predictive power when deployed in real-world scenarios.

D. Visualisations

I performed various advanced visualisations that provided meaningful insights into our dataset. These effective visualization aided in identifying trends, outliers, and correlations.

E. Association Rule Mining

This segment of my analysis focuses on implementing the Apriori algorithm for association rule mining, a key method in uncovering relationships between items in transaction data. I generated frequent itemsets with a minimum support threshold of 0.03, ensuring that only item combinations that appear relatively frequently in the dataset are considered. Following this, association rules are derived from these itemsets based on confidence and lift metrics, with a focus on rules that have a minimum confidence threshold of 0.1. Performance metrics, including execution time and memory usage, are captured to evaluate the efficiency of the Apriori algorithm. The resulting rules are then explored and sorted based on lift, confidence, and support, providing different perspectives on the strength and prevalence of item associations. The results were visualised through many visuals. Such as, a scatter plot visualizes the relationship between support and confidence for

Invoice ID	Branch	City	Customer	Gender	Product lin	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross marg	gross inco	Rating
750-67-84	A	Yangon	Member	Female	Health anc	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
226-31-30	C	Naypyitaw	Normal	Female	Electronic	15.28	5	3.82	80.22	3/8/2019	10:29	Cash	76.4	4.761905	3.82	9.6
631-41-31	A	Yangon	Normal	Male	Home and	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit carc	324.31	4.761905	16.2155	7.4
123-19-11	A	Yangon	Member	Male	Health anc	58.22	8	23.288	489.048	########	20:33	Ewallet	465.76	4.761905	23.288	8.4
373-73-79	A	Yangon	Normal	Male	Sports and	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
699-14-30	C	Naypyitaw	Normal	Male	Electronic	85.39	7	29.8865	627.6165	########	18:30	Ewallet	597.73	4.761905	29.8865	4.1
355-53-59	A	Yangon	Member	Female	Electronic	68.84	6	20.652	433.692	########	14:36	Ewallet	413.04	4.761905	20.652	5.8
315-22-56	C	Naypyitaw	Normal	Female	Home and	73.56	10	36.78	772.38	########	11:38	Ewallet	735.6	4.761905	36.78	8
665-32-91	A	Yangon	Member	Female	Health anc	36.26	2	3.626	76.146	########	17:15	Credit carc	72.52	4.761905	3.626	7.2
692-92-55	B	Mandalay	Member	Female	Food and b	54.84	3	8.226	172.746	########	13:27	Credit carc	164.52	4.761905	8.226	5.9
351-62-08	B	Mandalay	Member	Female	Fashion ac	14.48	4	2.896	60.816	2/6/2019	18:07	Ewallet	57.92	4.761905	2.896	4.5
529-56-39	B	Mandalay	Member	Male	Electronic	25.51	4	5.102	107.142	3/9/2019	17:03	Cash	102.04	4.761905	5.102	6.8
365-64-05	A	Yangon	Normal	Female	Electronic	46.95	5	11.7375	246.4875	########	10:25	Ewallet	234.75	4.761905	11.7375	7.1
252-56-26	A	Yangon	Normal	Male	Food and b	43.19	10	21.595	453.495	2/7/2019	16:48	Ewallet	431.9	4.761905	21.595	8.2
829-34-39	A	Yangon	Normal	Female	Health anc	71.38	10	35.69	749.49	########	19:21	Cash	713.8	4.761905	35.69	5.7
299-46-18	B	Mandalay	Member	Female	Sports and	93.72	6	28.116	590.436	########	16:19	Cash	562.32	4.761905	28.116	4.5
656-95-93	A	Yangon	Member	Female	Health anc	68.93	7	24.1255	506.6355	########	11:03	Credit carc	482.51	4.761905	24.1255	4.6
765-26-69	A	Yangon	Normal	Male	Sports and	72.61	6	21.783	457.443	1/1/2019	10:39	Credit carc	435.66	4.761905	21.783	6.9
329-62-15	A	Yangon	Normal	Male	Food and b	54.67	3	8.2005	172.2105	########	18:00	Credit carc	164.01	4.761905	8.2005	8.6
319-50-33	B	Mandalay	Normal	Female	Home and	40.3	2	4.03	84.63	########	15:30	Ewallet	80.6	4.761905	4.03	4.4
300-71-46	C	Naypyitaw	Member	Male	Electronic	86.04	5	21.51	451.71	########	11:24	Ewallet	430.2	4.761905	21.51	4.8
371-85-57	B	Mandalay	Normal	Male	Health anc	87.98	3	13.197	277.137	3/5/2019	10:40	Ewallet	263.94	4.761905	13.197	5.1
273-16-66	B	Mandalay	Normal	Male	Home and	33.2	2	3.32	69.72	########	12:20	Credit carc	66.4	4.761905	3.32	4.4
636-48-82	A	Yangon	Normal	Male	Electronic	34.56	5	8.64	181.44	########	11:15	Ewallet	172.8	4.761905	8.64	9.9

Fig. 2. Dataset

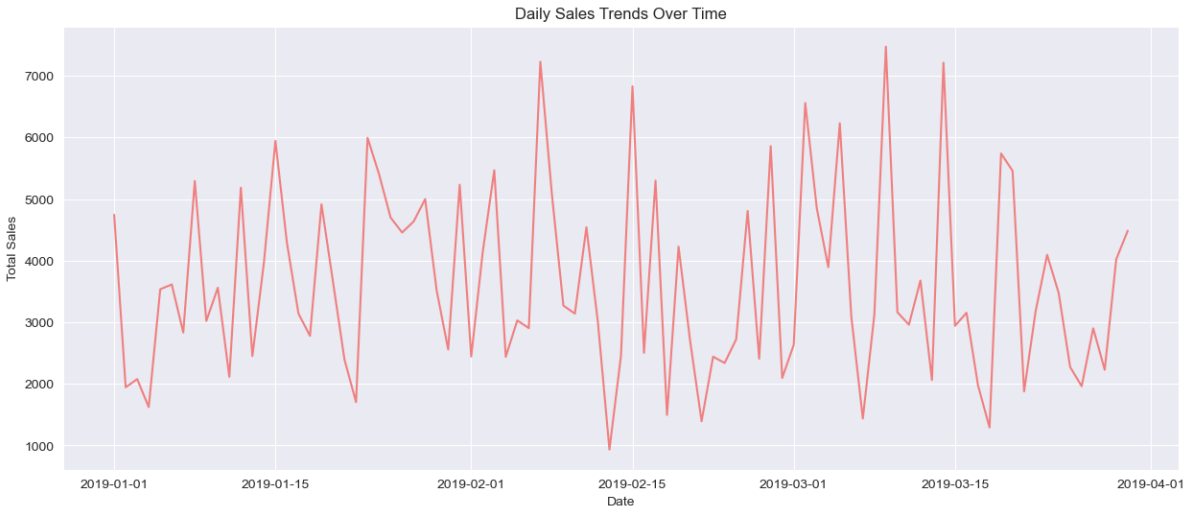

Fig. 3. Daily Sales Trends Over Time

each rule, offering a visual representation of rule robustness and prevalence. Overall, the application of the Apriori algorithm and subsequent rule analysis provides valuable insights into the purchasing patterns within the dataset, aiding in strategic decision-making for business operations and customer relationship management. The visualizations complement the numerical analysis, offering an intuitive understanding of the data and findings.

F. FP Growth Algorithm

Similarly, FP-Growth algorithm was applied to the transaction data to find frequent itemsets, with a minimum support threshold of 0.03. This approach efficiently discovers the commonly occurring product combinations in the dataset. Subsequently, association rules were generated from the frequent itemsets with a focus on those with a minimum confidence level of 0.1, aiming to understand the strength of implications among products. Visualization: The top 30 most frequent products were visualized using a bar chart, highlighting their relative support values.

G. Comparative Analysis

Table I reflects the performance metrics of two frequent itemset mining algorithms, Apriori and FP-Growth, applied to a dataset for identifying frequent patterns. The comparison focuses on execution time and memory usage: Apriori Algorithm: Completed in approximately 0.006 seconds, using 10,989 bytes of memory. FP-Growth Algorithm: Completed virtually instantaneously (reported as 0.0 seconds), with the same memory usage as Apriori at 10,989 bytes. This brief per-

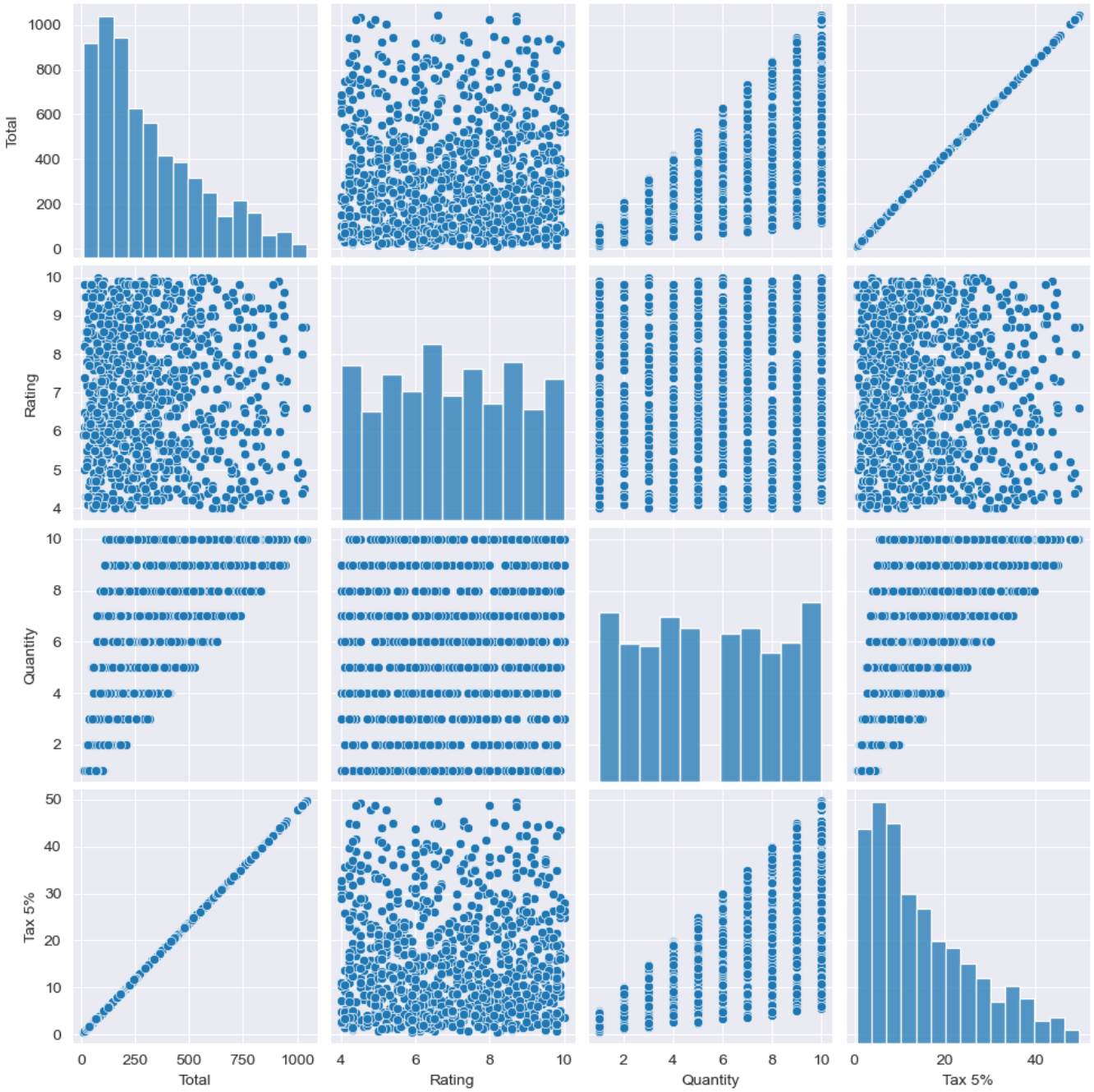

Fig. 4. Numerical Bivariate Analysis

formance comparison suggests that while both algorithms used the same amount of memory, FP-Growth was notably faster in this instance. The immediate execution time of FP-Growth indicates its efficiency and suitability for datasets where quick performance is crucial. This efficiency, particularly in time complexity, is a known characteristic of FP-Growth due to its more concise data structure and recursive divide-and-conquer approach, compared to the pair-wise comparison approach of Apriori. Such results advocate for FP-Growth's application in frequent itemset mining, especially in larger datasets or when time is a critical factor.

Algorithm	Time (seconds)	Memory (bytes)
Apriori	0.005930662155151367	10989
FP-Growth	0.0	10989

TABLE I
PERFORMANCE METRICS OF APRIORI AND FP-GROWTH ALGORITHMS

```
Rules sorted by Lift:
                 antecedents                consequents  antecedent support  \
2       (Food and beverages)  (Electronic accessories)            0.292490
3   (Electronic accessories)      (Food and beverages)            0.290514
17     (Fashion accessories)       (Sports and travel)            0.308300
16       (Sports and travel)     (Fashion accessories)            0.290514
11     (Fashion accessories)      (Food and beverages)            0.308300

    consequent support   support  confidence      lift  leverage  conviction  \
2             0.290514  0.073123    0.250000  0.860544  -0.01185    0.945982
3             0.292490  0.073123    0.251701  0.860544  -0.01185    0.945490
17            0.290514  0.077075    0.250000  0.860544  -0.01249    0.945982
16            0.308300  0.077075    0.265306  0.860544  -0.01249    0.941480
11            0.292490  0.077075    0.250000  0.854730  -0.01310    0.943347

    zhangs_metric
2       -0.186364
3       -0.185941
17      -0.189815
16      -0.185941
11      -0.197248
```

Fig. 5. Rules Sorted by List

H. Tracking Patterns and Customer Behavior Analysis

This analysis used frequent itemset mining to break down the dataset by month and determine changes in buying behavior over time. As a comparison, I selected two consecutive periods–the month of January 2019 (2019-01) and 2019 (2019-02) Using the FP-Growth algorithm, frequent itemsets for each period were found in an attempt to detect any new or disappearing patterns of purchases by customers. The results yielded several actual itemsets, showing products often bought together in both months. Some of the combinations remained stable across both periods: ”Health and beauty, Home and lifestyle,” for example; or food and beverages with electronic accessories. This indicates that customer demands in these areas remain constant. Nevertheless, my analysis did not show any new itemsets emerging or old ones disappearing between that period and this one with those blank ”New Itemsets” and ”Disappeared Itmesets” sections. What’s more, the section ”Changed Support for Existing Itemsets” showed no change in support of any itemset. Thus there was little significant shift in backing from these same items between January and February 2019. Does this continuity in item sets and their support during the two months indicate that there were indeed no changes to customer buying habits over these few weeks, or is it merely a result of such a short period being analyzed? More in-depth analysis with supplementary information, perhaps over longer periods or different time frames would reveal more details about the nature of customer shopping habits and market conditions.

I. Sequential Pattern Mining

I applied sequential pattern mining using the PrefixSpan algorithm to discover frequent sequences of items in customer transactions. The minimum support threshold was set to 2, indicating that only sequences occurring at least twice in the dataset were considered frequent. The algorithm identified several frequent sequences of items along with their support counts. These sequences represent patterns of items frequently purchased together by customers. For example, one of the identified sequences was ”Fashion accessories -¿ Food and beverages,” with a support count indicating how many times this sequence occurred in the dataset. To visualize and better understand these sequential patterns, I created a dynamic bar chart. The chart displays the top sequential patterns sorted by their support counts, with the support values shown on the x-axis and the sequences on the y-axis. Overall, sequential pattern mining allows businesses to gain insights into the purchasing behavior of customers, identifying common patterns that can inform marketing strategies, product placement, and recommendations to enhance the customer shopping experience.

J. Clustering

1. K-Means Clustering: K-Means is a popular centroid-based clustering algorithm. It partitions data into ’k’ clusters by minimizing the sum of squared distances between data points and their respective cluster centers. I used it for customer segmentation. The number of clusters ’k’ is a crucial parameter, and the algorithm assigns each data point to the nearest cluster center. In my case, with k=6, K-means gave me 6 clusters.

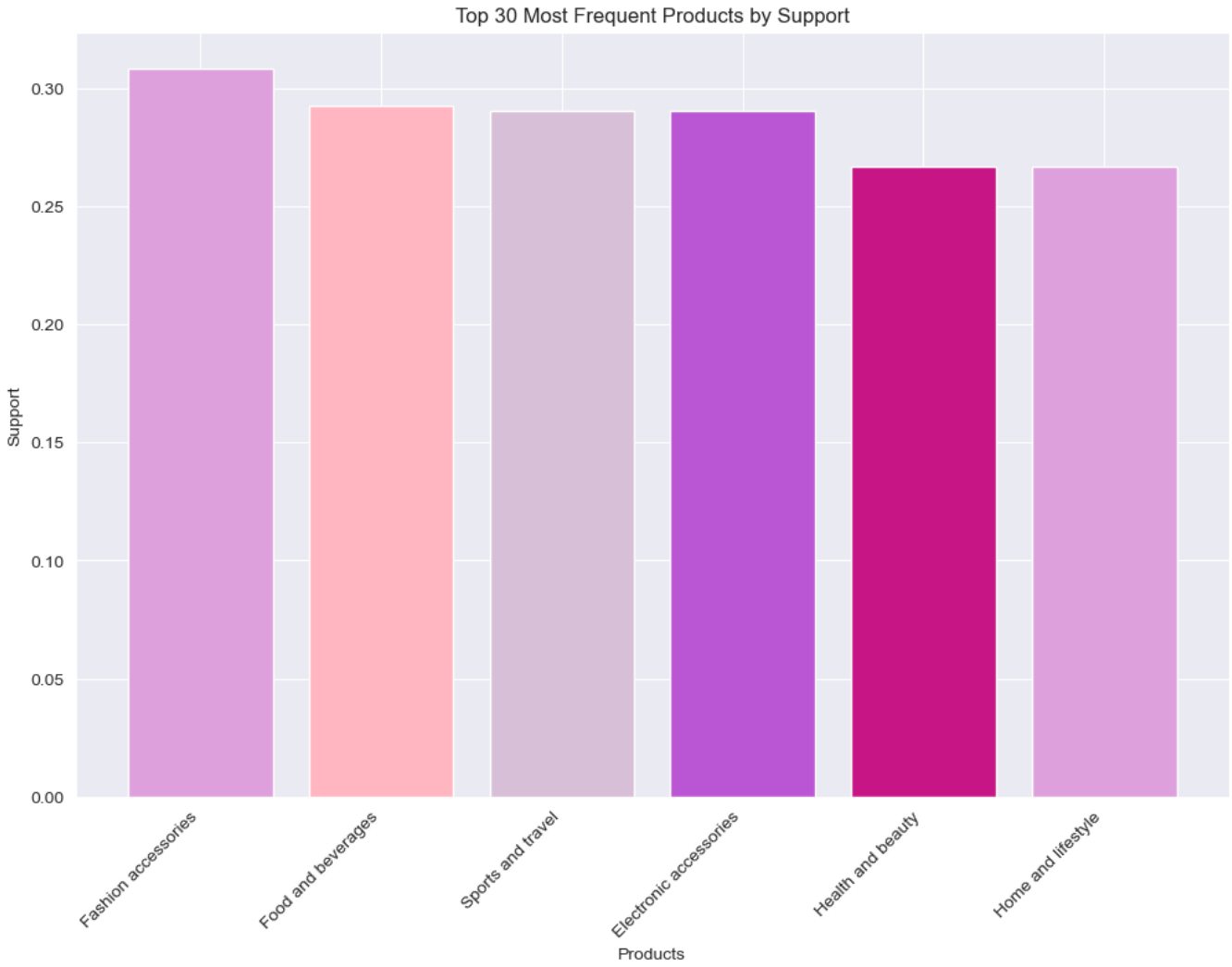

Fig. 6. Top 30 Most Frequent Products by Support

2. Hierarchical Agglomerative Clustering (Hierarchical Clustering): Hierarchical clustering builds a hierarchy of clusters. It starts with individual data points as clusters and recursively merges them into larger clusters based on proximity. It is used for visualizing cluster structures and dendrogram generation. I got 4 clusters since my cut off point was selected in that manner. 3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN groups together data points that are close to each other based on a density criterion. It can discover clusters of arbitrary shapes and identifies noise points. It is useful for anomaly detection and identifying clusters with varying densities. The 'eps' parameter defines the maximum distance between two samples for one to be considered as in the neighborhood of the other. 'min_samples' sets the number of samples required to form a dense region. With eps=0.4, min_samples=3, DBSCAN gave me 5 clusters.

4. Mean Shift Clustering: Mean Shift is a non-parametric clustering algorithm. It shifts data points towards the mode (peak) of the density function to find clusters. It is suitable for applications where the number of clusters is not known in advance, such as image segmentation. In my case, it only gave one cluster when I did not set any bandwidth. However, by iterating my model on different bandwidths I saw which was the best-fit and at the end I selected bandwidth=1.0 since it gave me the optimum result of 6 clusters that matched the rest of my models.

5. Gaussian Mixture Model (GMM): GMM is a probabilistic model that assumes data points are generated from a mixture of several Gaussian distributions. It estimates the parameters of these distributions to find clusters. The number of components (clusters) needs to be specified. This model gave me 6 clusters.

All 5 models gave me clusters in the range of 4-6, the mode being 6. Each of these clustering algorithms has its strengths and weaknesses, making them suitable for different types of data and applications. The choice of which algorithm to use depends on the specific problem and the nature of the data being analyzed.

It's important to note that DBSCAN and MeanShift were

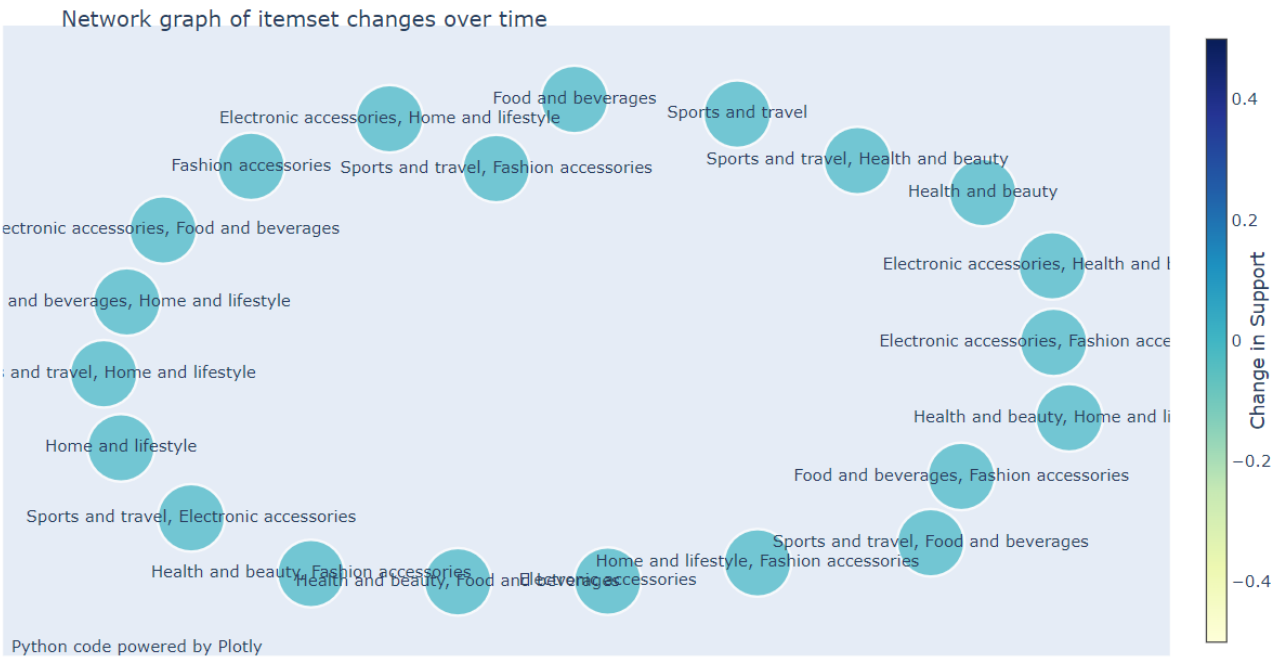

Fig. 7. Network graph of itemset changes over time

skipped because they either resulted in only one cluster or noise points, which may not provide meaningful clustering in this context.

K. Classification

Firstly, I used quantile-based binning to create three segments (Low, Medium, High) as a preprocessing step that discretizes the continuous NewTotal variable into classes suitable for classification. Next, my choice of Random Forest as a classification algorithm is suitable for this customer segmentation task due to its robustness and ability to handle mixed data types. Moreover, it is less prone to overfitting compared to individual decision trees The application of classification allows businesses to assign customers to segments based on their spending behavior, which can inform marketing and targeting strategies. Overall, the use of Random Forest Classifier in this context aligns with the goal of segmenting customers based on their spending behavior, providing valuable insights for business decision-making.

L. Regression

The linear regression model was employed to predict customer spending, and the model demonstrated exceptional performance. The mean squared error (MSE) was found to be extremely low, with a value of approximately 2.54×10^{-31}, indicating that the model's predictions were in very close proximity to the actual values. Furthermore, the coefficient of determination (R^2) exhibited a perfect score of 1.0, signifying that the model accounted for all the variance in the data, effectively capturing the relationship between the input features and customer spending. This remarkable performance suggests that the linear regression model is highly effective in predicting customer spending based on the provided dataset. Data preprocessing for regression analysis is a crucial step to ensure that the dataset is suitable for modeling and that the regression results are reliable. This process involves selecting relevant features, scaling numeric variables, encoding categorical variables, splitting the data into training and testing sets, and conducting feature engineering as needed. The model then produces accurate predictions, and provides meaningful insights into the relationships between independent and dependent variables. Careful preprocessing enhances the overall quality of regression models and their ability to make sound predictions.

M. Outlier Detection and Statistical Validation

The z-score-based outlier detection method and the one-way ANOVA (Analysis of Variance) statistical test were chosen to identify patterns and anomalies in the dataset. The z-score method allowed me to detect outliers by measuring how many standard deviations a data point is away from the mean. In this case, a threshold of 0.5 standard deviations was set, and data points exceeding this threshold were considered outliers. This method helps identify individual data points that deviate significantly from the average, highlighting potential anomalies in the data. On the other hand, the one-way ANOVA test is used to assess whether there are statistically significant differences in the means of multiple groups. In my context, it helps determine whether there are significant differences in the "Total" variable across different "Product line" groups. The F-

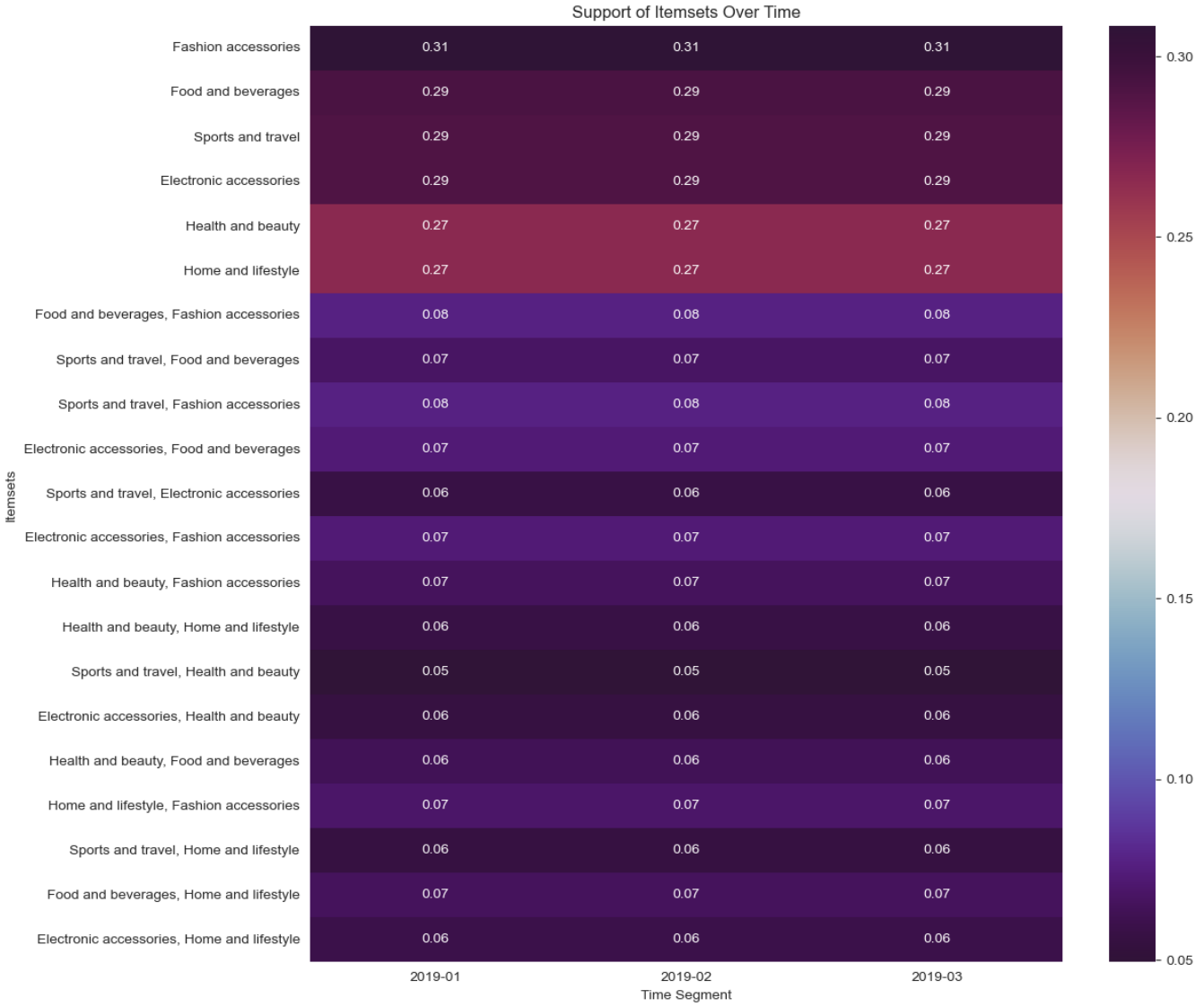

Fig. 8. Support of Itemsets Over Time

Algorithm	Number of Clusters	Silhouette Score	Cluster Sizes
K-Means	6	0.3444	{5: 212, 2: 204, 0: 198, 3: 147, 1: 124, 4: 115}
Hierarchical	4	0.3922	{1: 532, 3: 169, 4: 160, 2: 139}
Gaussian Mixture	6	0.3396	{1: 254, 5: 185, 2: 175, 4: 136, 0: 131, 3: 119}

TABLE II
COMPARISON OF CLUSTERING RESULTS

statistic and p-value obtained from this test provide insights into whether the variations between groups are statistically significant. A low p-value indicates that at least one group's mean is significantly different from the others, highlighting potential patterns or anomalies in customer spending behavior based on product lines.

In the context of customer behavior analysis, two hypotheses were tested using statistical methods:

1. Tukey's HSD Test: The null hypothesis (H_0) for this test is that there is no significant difference in total spending among different product line groups, while the alternative hypothesis (H_1) is that there are significant differences. To test this, Tukey's HSD test was conducted to compare the means of total spending between different product line groups. The results provide insights into whether customers from different product line groups exhibit significantly different spending behaviors. The implications of this test are to identify which product line groups may contribute more or less to overall sales.

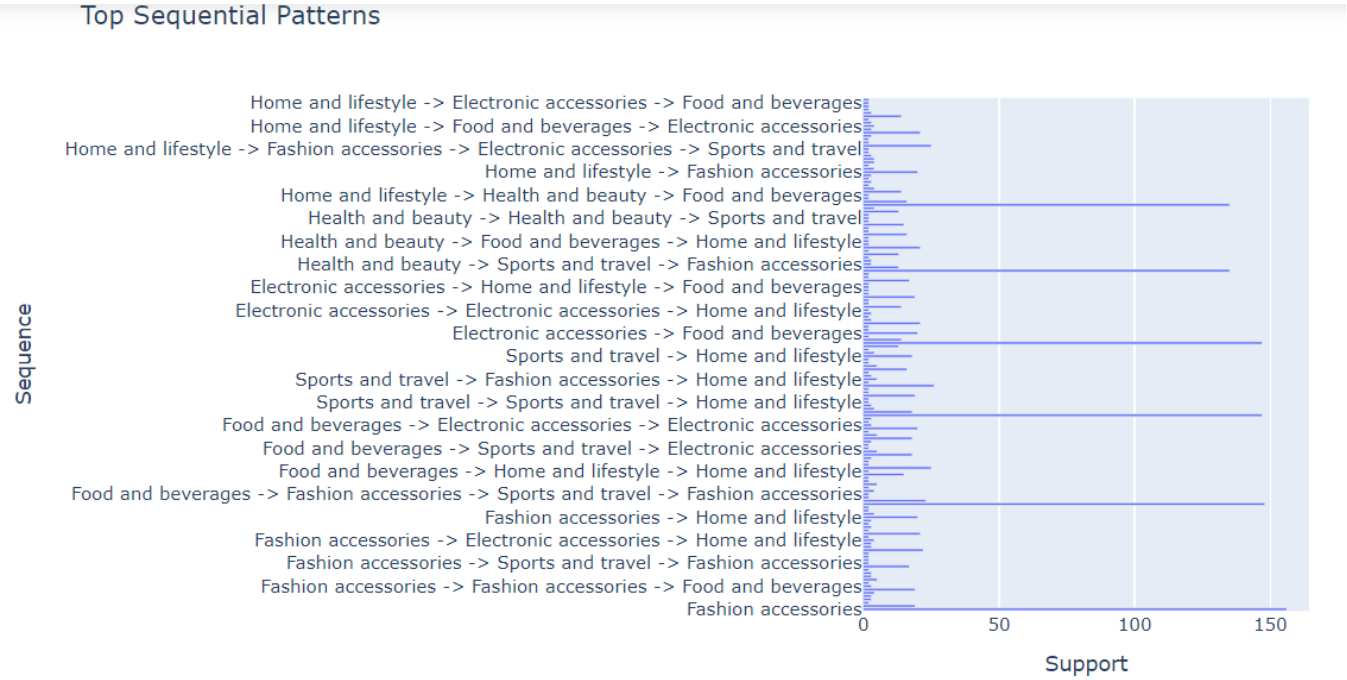

Fig. 9. Top Sequential Patterns

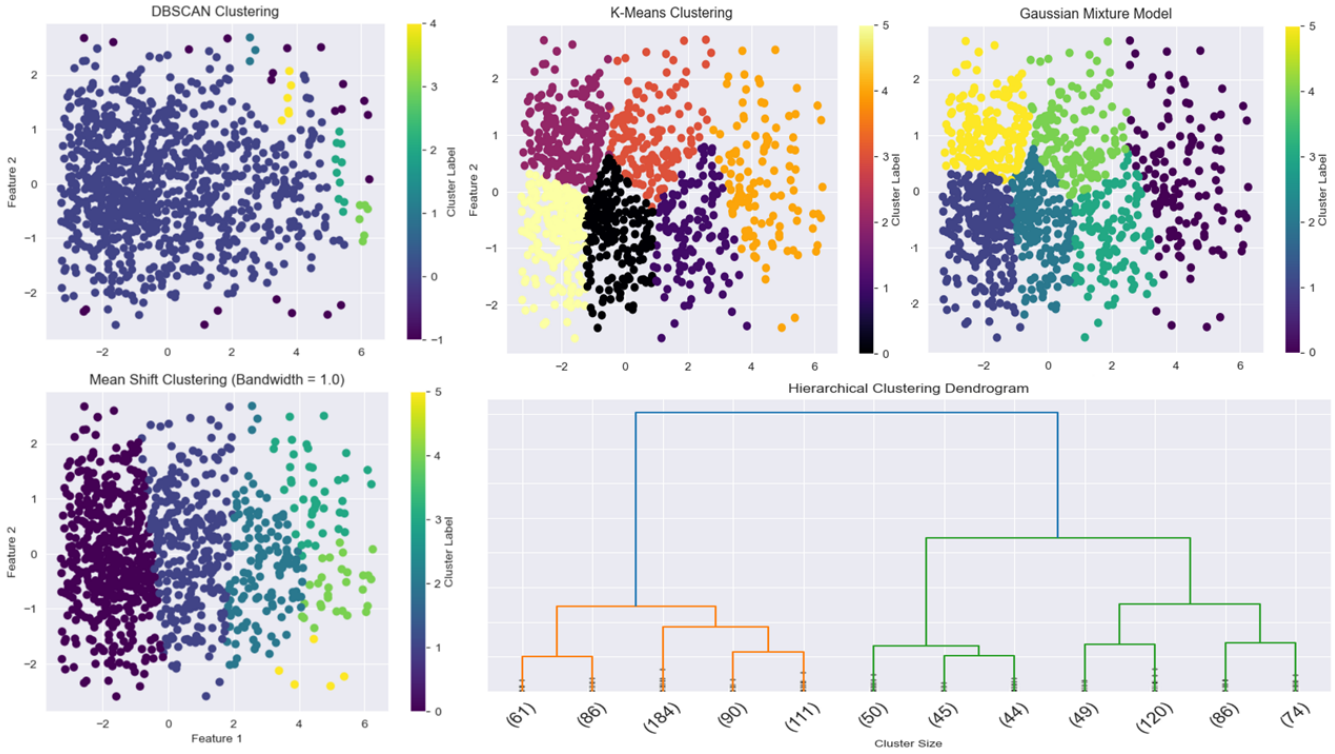

Fig. 10. Clustering Results

2. Independent Samples t-test: The null hypothesis (H_0) for this test is that there is no significant difference in total spending between members and non-members, while the alternative hypothesis (H_1) is that there are significant differences. This test was conducted to compare the means of total spending between two customer types: members and non-members. The

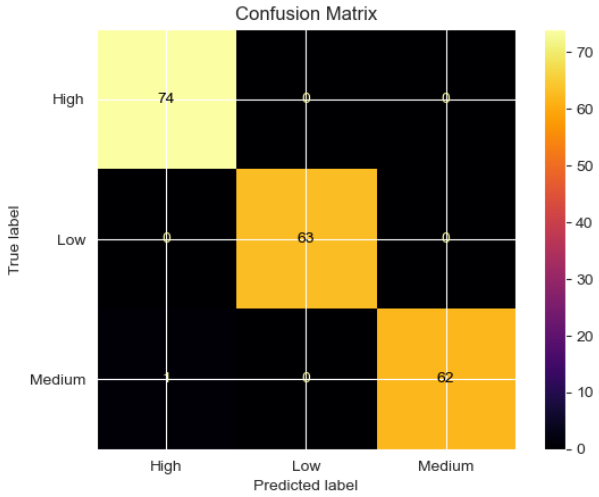

Fig. 11. Confusion Matrix

TABLE III
CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
High	0.99	1.00	0.99	74
Low	1.00	1.00	1.00	63
Medium	1.00	0.98	0.99	63
Accuracy			0.99	200
Macro Avg	1.00	0.99	1.00	200
Weighted Avg	1.00	0.99	0.99	200

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower  upper  reject
-----------------------------------------------------
     0      1  -0.0592  0.994  -0.366 0.2477  False
     0      2   0.0124    1.0 -0.2962 0.3209  False
     0      3   0.0163    1.0 -0.3031 0.3357  False
     0      4   0.0692  0.989  -0.246 0.3844  False
     0      5   0.0506 0.9974 -0.2616 0.3628  False
     1      2   0.0715 0.9852 -0.2335 0.3766  False
     1      3   0.0755 0.9839 -0.2405 0.3915  False
     1      4   0.1284 0.8484 -0.1834 0.4401  False
     1      5   0.1098  0.913  -0.199 0.4185  False
     2      3    0.004    1.0 -0.3137 0.3216  False
     2      4   0.0568 0.9955 -0.2566 0.3702  False
     2      5   0.0382 0.9993 -0.2722 0.3487  False
     3      4   0.0529 0.9973 -0.2712  0.377  False
     3      5   0.0343 0.9996  -0.287 0.3555  False
     4      5  -0.0186    1.0 -0.3356 0.2984  False
-----------------------------------------------------
```

Fig. 12. Tukey's HSD Test

results determine whether there is a significant difference in spending behavior between these two customer types. The implications are to assess whether membership status has a notable impact on customer spending.

In the case of the t-test, the output indicates that we "Fail to reject H_0," which means there is no significant difference in total spending between members and non-members. This suggests that membership status alone may not be a strong predictor of spending behavior. These statistical tests provide

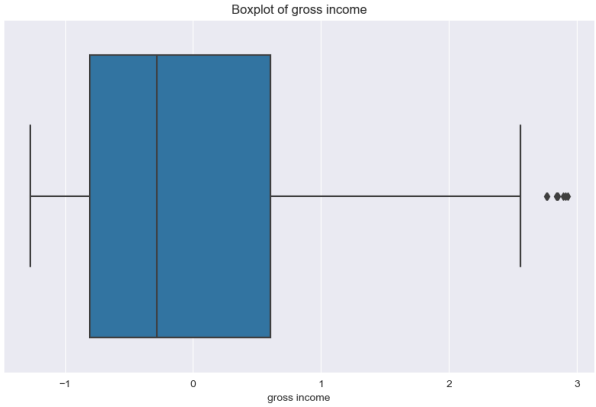

Fig. 13. Box Plot of one feature

valuable insights into customer behavior and help guide marketing and business strategies.

III. CONCLUSION

In this comprehensive analysis of supermarket transaction data, I delved deep into understanding customer behavior and purchasing patterns. The study began with the segmentation of customers based on their purchasing behavior, providing valuable insights into distinct customer groups. Classification models, including Random Forest, were employed to classify customers into segments such as "Low," "Medium," and "High" spenders. This segmentation is pivotal for supermarkets to tailor marketing strategies, personalize customer experiences, and optimize inventory management.

Additionally, I performed regression analysis to predict customer spending, offering a predictive tool for supermarkets to anticipate future sales and understand the factors driving customer expenditures. Moreover, I applied Outlier detection techniques to identify unusual behavior in customer transactions, ensuring data quality and reliability. Clustering methods helped group customers based on their spending patterns, enabling supermarkets to target specific customer segments more effectively. Furthermore, statistical tests, including Tukey's HSD and t-tests, were conducted to evaluate the impact of product lines and membership status on customer spending, providing actionable insights for decision-makers.

In conclusion, this analysis equips supermarkets with data-driven insights to enhance their operations, optimize marketing strategies, and improve customer satisfaction. By understanding customer behavior and the factors influencing spending, supermarkets can stay competitive in a rapidly evolving market and create a shopping experience tailored to the unique needs of their customers.

IV. DISCUSSION

Let's discuss some limitations and future directions of this project.

A. Limitations

While this analysis provides valuable insights into customer behavior and purchasing patterns, it is essential to acknowledge its limitations. Firstly, the dataset used for this study is limited to a specific supermarket, and the findings may not be entirely generalizable to other retail contexts. Additionally, the analysis assumes that the data is representative of the entire customer population, which may not account for potential biases or variations in customer demographics. Moreover, the choice of classification and regression models, as well as the parameter settings, could impact the accuracy of predictions and classifications. Finally, the study primarily focuses on historical data and does not consider external factors such as economic conditions or seasonal trends, which can influence customer behavior. Despite these limitations, this analysis serves as a valuable foundation for understanding and optimizing supermarket operations and customer interactions.

B. Future Directions

Looking ahead, there are several promising avenues for extending this study and further enhancing our understanding of customer behavior in the context of supermarkets. First, incorporating external data sources, such as economic indicators, weather conditions, and local events, could provide a more comprehensive view of the factors influencing customer purchasing decisions. Additionally, conducting surveys or customer feedback analysis could help capture qualitative insights, preferences, and sentiments that complement the quantitative findings. Exploring advanced machine learning techniques like deep learning and natural language processing for sentiment analysis could also yield richer insights from customer reviews and comments. Furthermore, implementing real-time data analytics and personalized marketing strategies based on customer segmentation could enhance customer engagement and loyalty. Lastly, collaboration with other supermarkets or retail chains to compare and benchmark customer behavior patterns could lead to industry-wide best practices and insights.