Market Retail Analysis – Grocery POS Coded Project

Part B

Business Report

DSBA – Course

Created by – Rishabh Gupta

Foreword

Business Context:

In the highly competitive grocery retail industry, understanding customer buying patterns is crucial for enhancing sales, increasing customer satisfaction, and improving profitability. By identifying frequently purchased item combinations, grocery stores can craft effective marketing strategies, optimize inventory management, and tailor promotions to meet customer needs. Leveraging Point of Sale (POS) data can unlock valuable insights that drive customer-centric offerings, such as combo packs, discounts, and targeted promotions, which can increase basket size and improve customer retention. This analysis aligns with business goals by maximizing revenue, reducing operational costs, and boosting customer loyalty.

Objective:

As a business analyst, the goal is to analyze the POS transactional data to identify frequently purchased item combinations. Using association rule mining or similar techniques, the aim is to uncover patterns that will help the store create targeted combo offers and discounts, ultimately driving revenue growth by increasing customer purchases and average basket size.

Contents

Sr. No	Topics	Pages
1	Objective	5
2	Data Overview	6
3	Statistical summary of data	7
4	Data Preprocessing	6
5	Exploratory Data Analysis	8
6	Market Based Analysis (MBA)	18
7	Insights and Recommendations	25

List of Tables

Sr. No	Name of Tables	Pages
1	Top 5 rows	6
2	Basic info of dataset	6
3	Statistical summary	7
4	Association rules of metrics	
5	Association final output table	

List of Figures

Sr. No	Name of Figures	Pages
1	Distribution of transactions overtime	7
2	Top 20 frequent selling products	8
3	Weekly trends	9
4	Monthly trends	10
5	Quarterly trends	11
6	Yearly Trends	12
7	Heatmap analysis	13
8	Boxplot for product per orders	14
9	Scatterplot – Support vs confidence	15
10	Andrew plot for distribution	16
11	MBA – Knime Workflow Diagram	20
12	Scatterplot – Life vs confidence (KNIME)	21
13	Bar plot – Lift, Confidence and Support (KNIME)	22
14	Line plot 0 Lift vs Confidence (KNIME)	23

Objective

As a business analyst, the goal is to analyze the POS transactional data to identify frequently purchased item combinations. Using association rule mining or similar techniques, the aim is to uncover patterns that will help the store create targeted combo offers and discounts, ultimately driving revenue growth by increasing customer purchases and average basket size.

For this assignment, we will analyze data do EDA using python and MBA using Knime.

Data Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name –

1. *dataset_group.xlsx*

Part – B

Data Dictionary –

The dataset consists of transactional data from a grocery store, where each row represents a product purchased in a specific order. The columns in the dataset are as follows:

- **Date:** The date when the transaction took place.
- **Order_id:** A unique identifier for each customer order.
- **Product:** The individual item purchased in the transaction.

Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 20641 number of rows with 3 columns.

	Date	Order_id	Product
0	01-01-2018	1	yogurt
1	01-01-2018	1	pork
2	01-01-2018	1	sandwich bags
3	01-01-2018	1	lunch meat
4	01-01-2018	1	all- purpose

TABLE 1 - TOP 5 ROWS OF DATASET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20641 entries, 0 to 20640
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Date      20641 non-null  object
 1   Order_id  20641 non-null  int64 
 2   Product   20641 non-null  object
dtypes: int64(1), object(2)
memory usage: 483.9+ KB
```

TABLE 2 - BASIC INFO. OF DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

• We can observe that there is 1 numerical and 2 object columns.

Data Pre-processing-

Missing value treatment and Analysis-

- On analysis, we can observe there are no missing values.

Statistical Summary –

Using Describe () function, we can analyses the summary statistics of the dataset –

	Order_id
count	20641.000000
mean	575.986289
std	328.557078
min	1.000000
25%	292.000000
50%	581.000000
75%	862.000000
max	1139.000000

TABLE 3 - STATISTICAL SUMMARY OF DATASET

Observations-

- There are a total of 20,641 order records.
- The average Order ID is around 576.
- The Order IDs range from a minimum of 1 to a maximum of 1139.
- The median Order ID (50th percentile) is 581, which is quite close to the mean.
- The interquartile range (difference between 75th and 25th percentile) is 570 (862 - 292), showing the spread of the central 50% of the data.
- The standard deviation is approximately 328.56.

Exploratory Data Analysis

Lets have a look at Distribution –

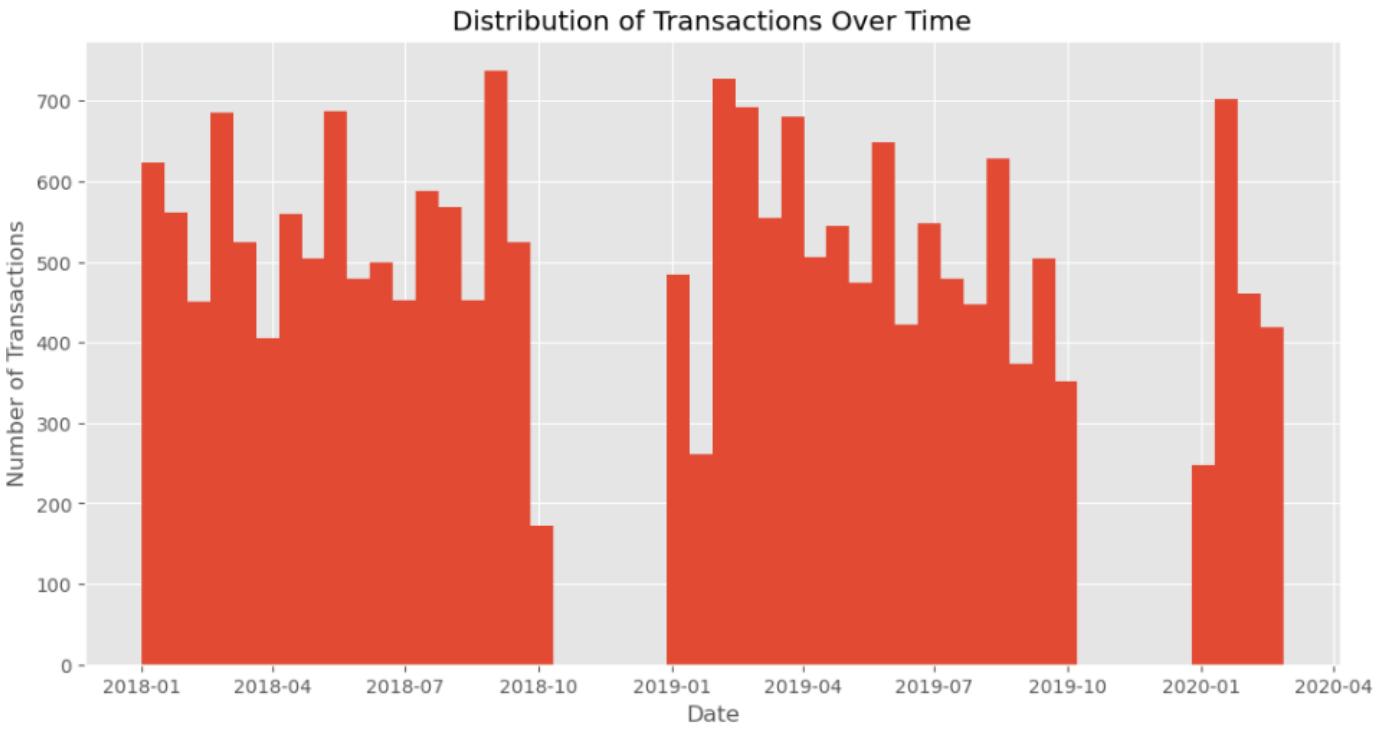

FIGURE 1 - DISTRIBUTION OF TRANSACTIONS OVERTIME

Observations –

- The histogram displays the distribution of grocery store transactions over time, spanning from early 2018 to early 2020.
- We observe cyclical patterns, suggesting regular fluctuations in transaction volume. There appear to be periods of higher transaction activity followed by dips.
- A noticeable drop in transactions occurs around late 2018, followed by a recovery in early 2019. Another significant decline is visible towards the end of 2019.
- The period around early 2020 shows a very low number of transactions, potentially indicating an external event impacting store activity.
- Overall, the transaction volume seems to exhibit some level of consistency within certain periods, punctuated by these notable drops. Further investigation could explore the reasons behind these fluctuations.

Now let’s review top 20 selling products –

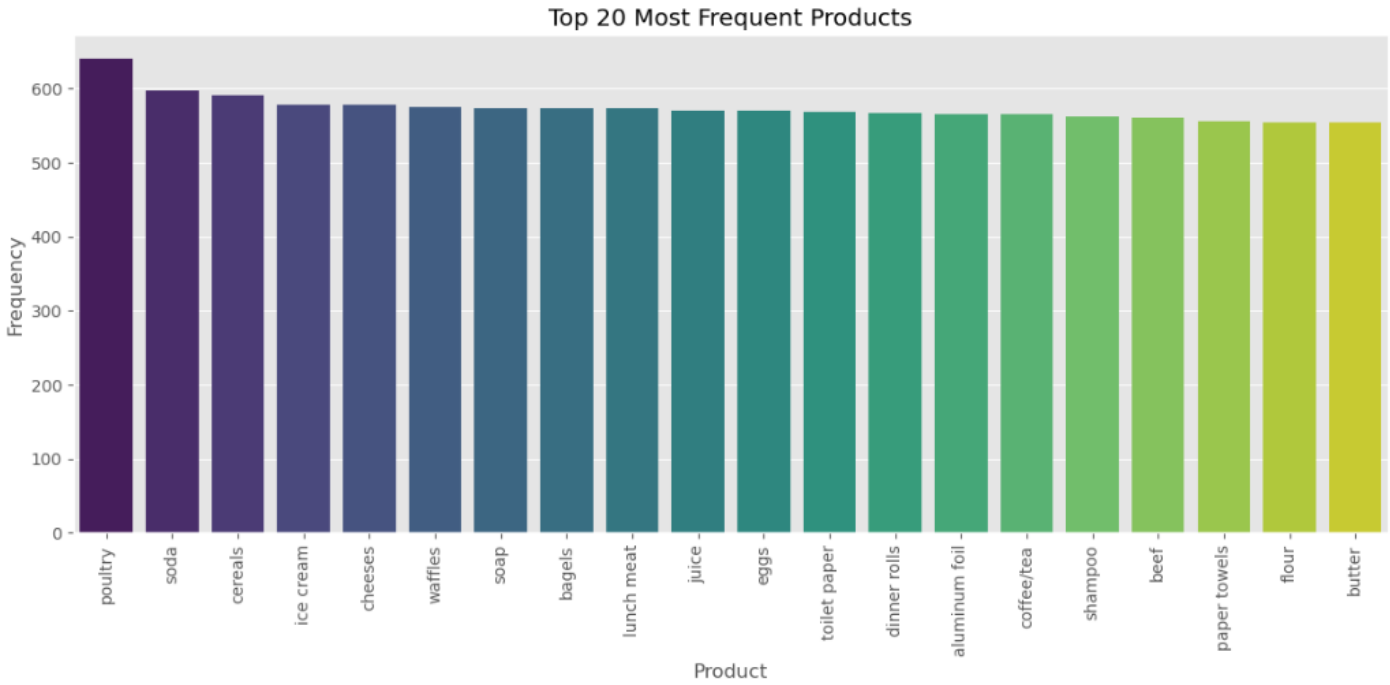

FIGURE 2 - TOP 20 FREQUENT SELLING PRODUCTS

Observations -

- The bar chart showcases the top 20 most frequently purchased products.
- "Poultry" stands out as the most frequent item, with a significantly higher count than others.

- Following poultry, "soda," "cereals," and "ice cream" are also among the top-selling products, with similar high frequencies.
- A gradual decline in frequency is observed as we move towards the less frequent products in the top 20.

- The products towards the end of the list, such as "flour" and "butter," have a noticeably lower purchase frequency compared to the top few.

- This distribution highlights the core popular items driving sales volume, which can inform inventory management and promotional strategies.

Now lets review weekly sales trend–

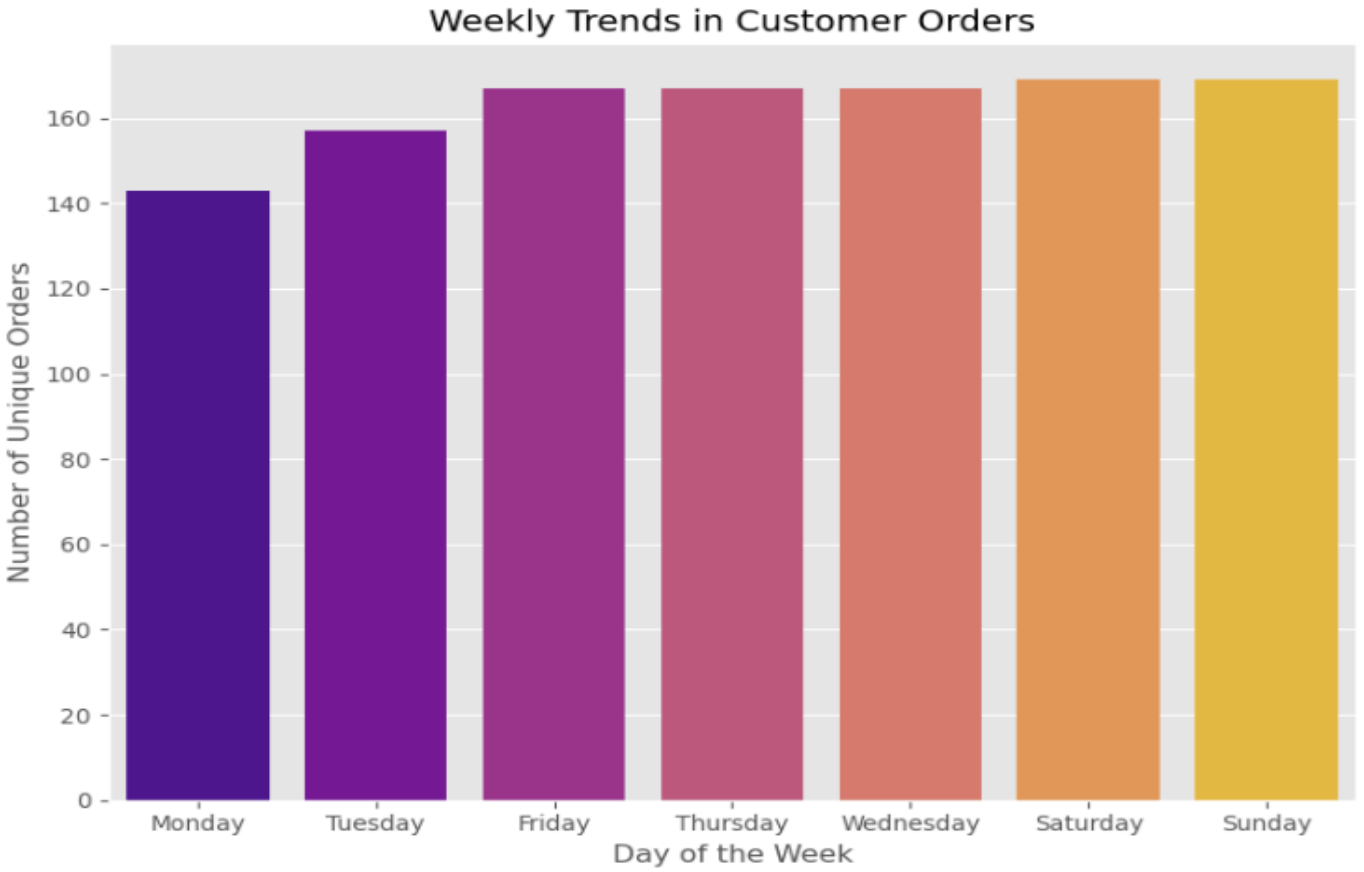

FIGURE 3 - WEEKLY TREND IN CUSTOMER ORDERS

Observations -

- The number of unique customer orders varies across the days of the week.
- Monday exhibits the lowest number of unique orders compared to other days.
- There's a noticeable increase in order volume from Monday to Tuesday.
- The highest number of unique orders is observed on Saturday and Sunday, indicating the peak shopping days.
- Friday, Thursday, and Wednesday show similar, relatively high levels of customer order activity, slightly lower than the weekend.
- This suggests a clear weekly pattern with lower activity at the start of the week and a significant surge towards the weekend.

Now let’s observe monthly trends –

FIGURE 4 - MONTHLY TRENDS IN CUSTOMER ORDERS

- January shows the highest volume of orders, significantly exceeding most other months.
- There's a noticeable dip in order numbers starting from March and reaching the lowest point in April.
- A recovery in order volume is observed in May, followed by fluctuations in the subsequent months.
- June and August record relatively lower order counts compared to the months around them.
- September and July show similar moderate levels of customer order activity.
- The data for October, November, and December appears to be missing from this visualization.
- The significant peak in January could be attributed to post-holiday restocking or specific January-related promotions.
- The low in April might correlate with a specific time of year with fewer shopping needs or potentially external factors not immediately apparent from the data itself.

Now observe quarterly trends –

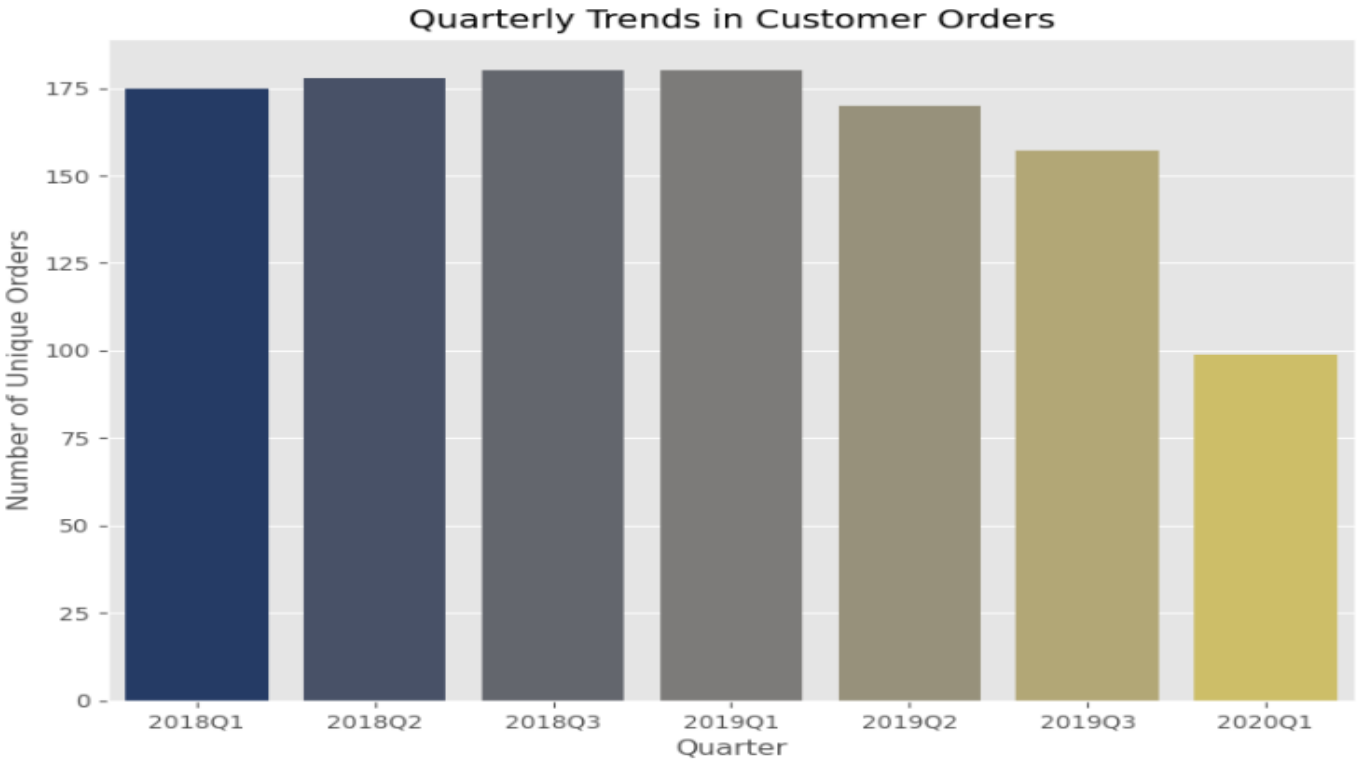

FIGURE 5 - QUARTERLY TRENDS IN CUSTOMER SERVICE

- The bar chart displays the quarterly trends in the number of unique customer orders, spanning from the first quarter of 2018 to the first quarter of 2020.
- The number of unique orders shows a generally stable and relatively high level in the first three quarters of 2018.
- A slight increase is observed in the second and third quarters of 2018 compared to the first.
- The trend continues with similarly high order volumes in the first quarter of 2019.
- A noticeable dip in the number of unique orders occurs in the second quarter of 2019.
- This downward trend continues into the third quarter of 2019, showing a further decrease.
- The first quarter of 2020 exhibits the lowest number of unique orders in the entire period visualized.
- This significant drop in early 2020 might indicate the impact of external factors or a shift in customer behaviour.
- Overall, the quarterly trend suggests a period of stability followed by a decline in customer order volume in the latter part of the observed timeframe.

Now lets analyse Yearly trends –

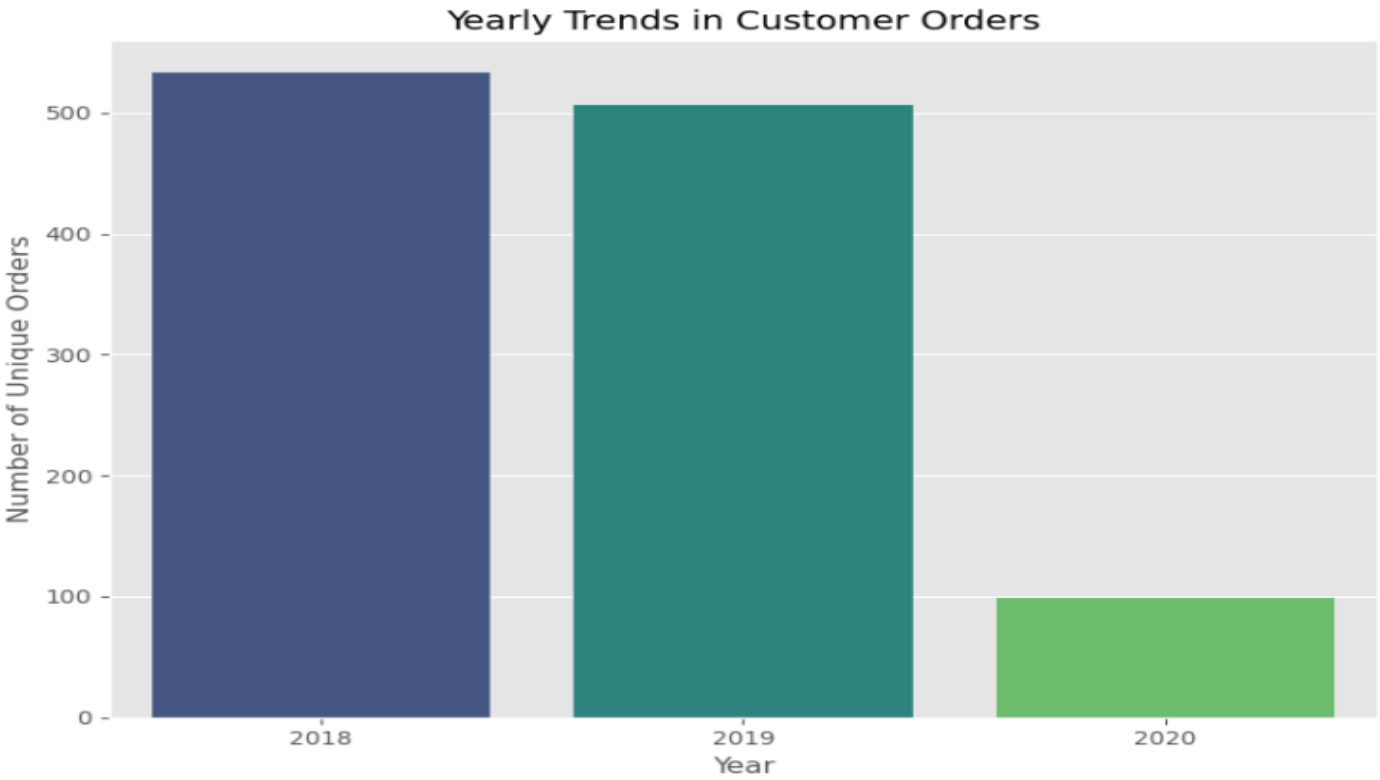

FIGURE 6 - YEARLY TRENDS FOR ORDERS

- The bar chart presents the yearly trends in the number of unique customer orders from 2018 to 2020.
- The year 2018 recorded the highest number of unique customer orders among the three years shown.
- In 2019, there was a noticeable decrease in the total number of unique customer orders compared to the previous year.
- The year 2020 shows a significantly lower number of unique customer orders than both 2018 and 2019.
- The sharp decline in 2020 suggests a substantial change in customer purchasing behavior or external factors impacting the store's operations.
- This downward trend from 2018 to 2020 indicates a potential challenge or shift in the business over this period.
- Further investigation into the events of 2020 would be crucial to understand the reasons behind this significant reduction in order volume.
- The data highlights a clear need to analyze the factors contributing to the declining trend in customer orders year over year.

Heatmap -

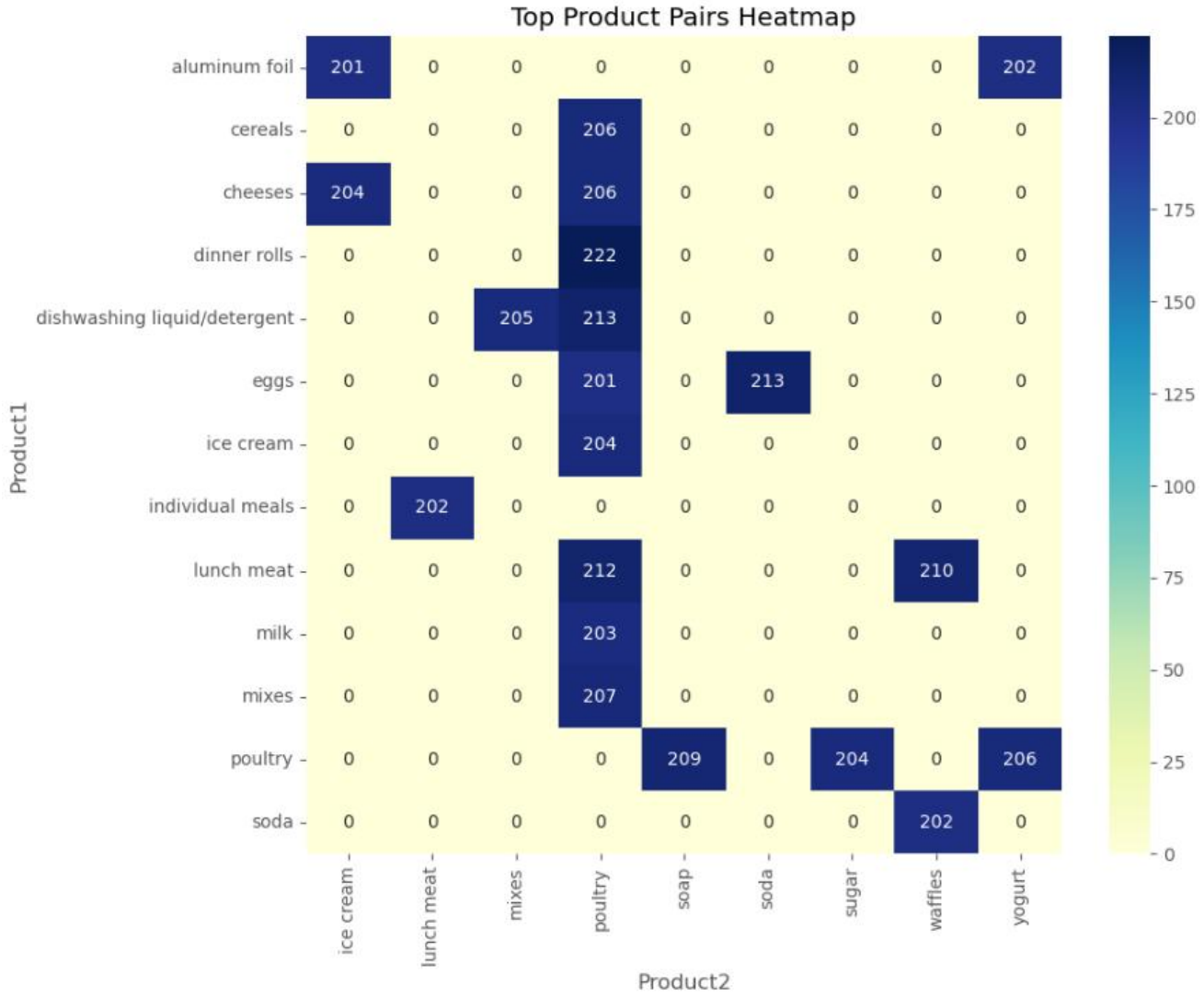

FIGURE 7 - HEATMAP FOR PRODUCTS

- The heatmap displays the frequency of the top 20 most common product pairs purchased together.
- Darker blue cells indicate product pairs with a higher co-occurrence count within the same order.
- Several pairs show strong co-purchase patterns, such as (cereals, poultry) and (cheeses, poultry), both occurring 206 times.
- The absence of dark cells in certain intersections suggests that those specific product pairs from the top list are not frequently bought together.
- This visualization provides direct insights into which products are often part of the same shopping basket, valuable for marketing and merchandising decisions.

Now’s lets a boxplot –

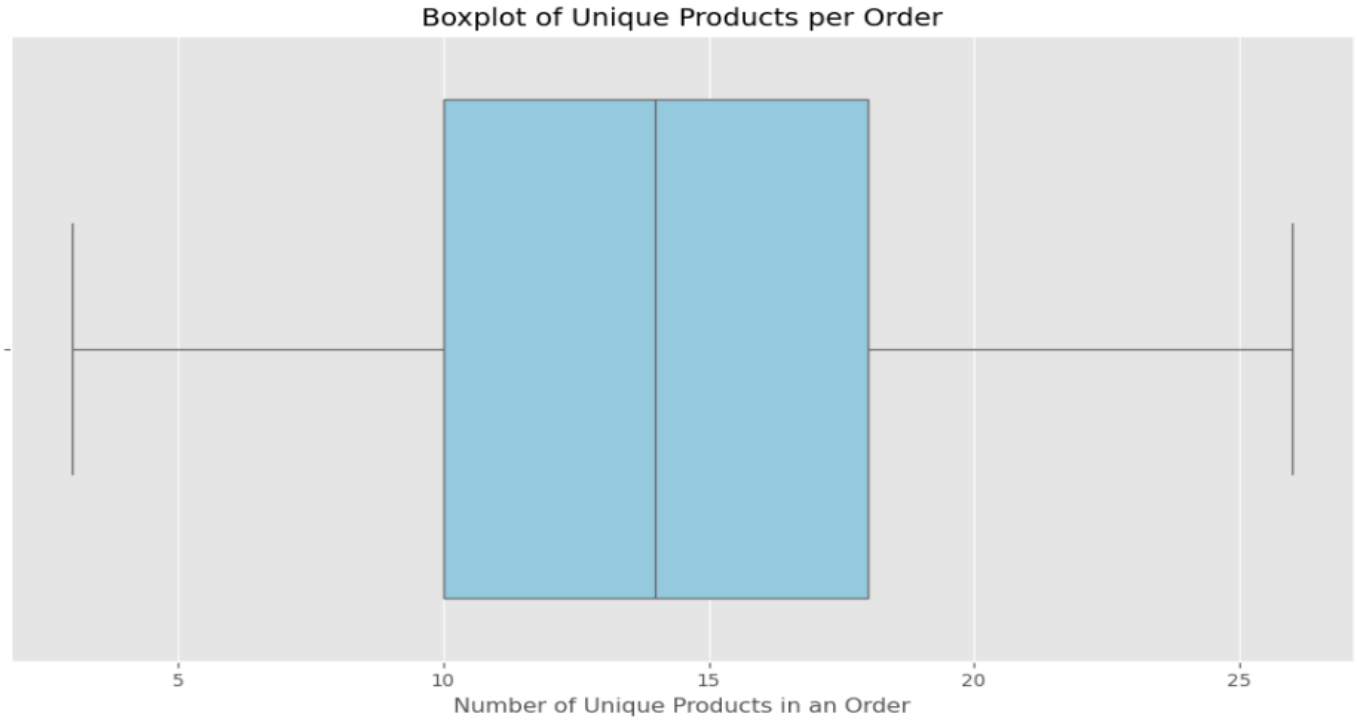

FIGURE 8 - BOXPLOT FOR UNIQUE PRODUCTS PER ORDER

- The median order contains approximately 13 unique products.
- 50% of orders include between 10 and 16 unique products (the IQR).
- The number of unique products per order typically ranges from about 4 to 26.
- Orders with fewer than 4 or more than 26 unique products could be considered outliers.
- The distribution within the IQR appears relatively symmetrical around the median.
- A typical grocery basket contains a variety of around 10 to 16 distinct items.
- This provides a baseline for understanding average purchase diversity and identifying unusual shopping patterns.

Now lets observe Combo opportunities with high support & confidence -

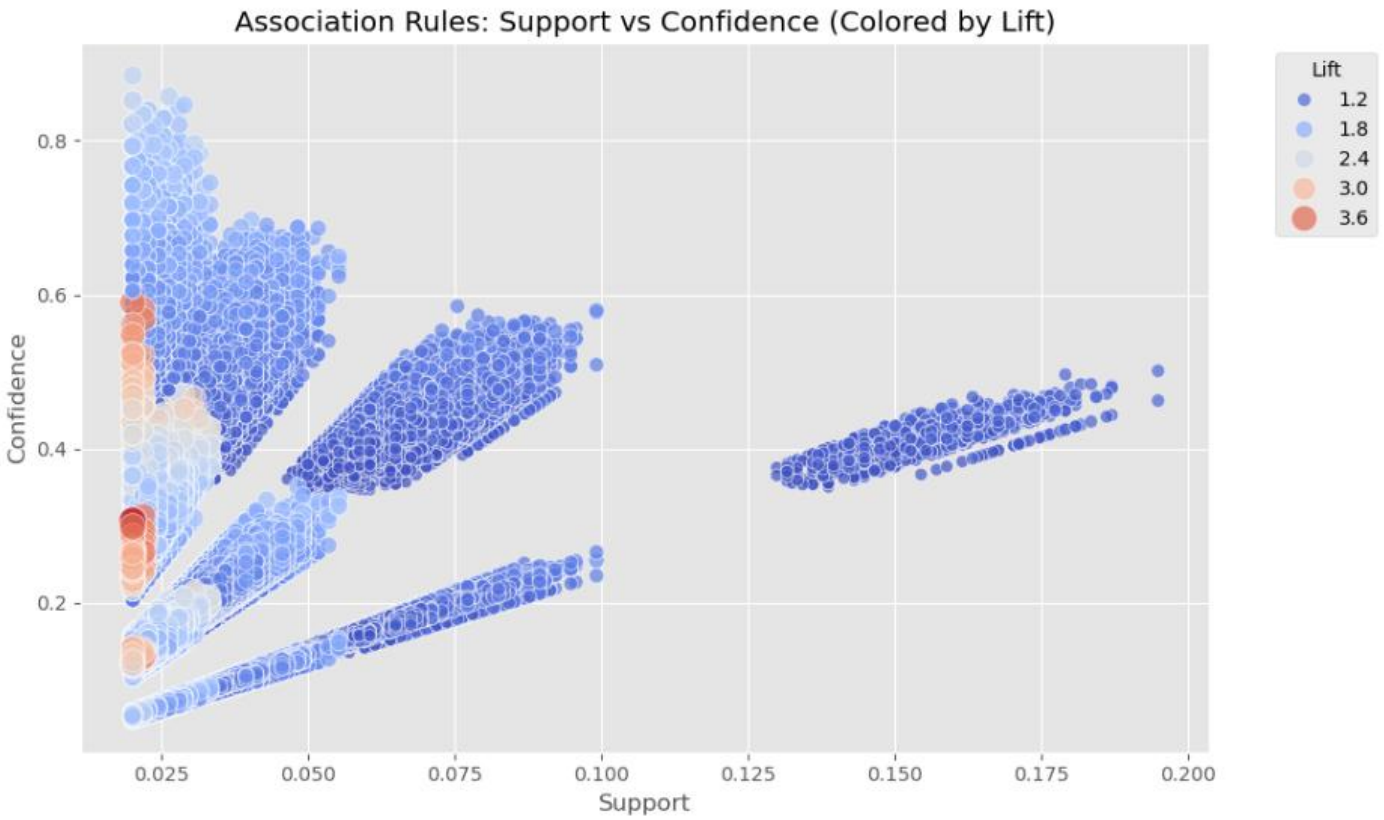

FIGURE 9 - SUPPORT VS CONFIDENCE

- Each point on the scatter plot represents an association rule derived from the transactional data.
- The x-axis indicates the support of the rule, representing how frequently the itemset in the rule appears in the transactions.
- The y-axis indicates the confidence of the rule, showing the probability of the consequent occurring given the antecedent.
- The color of each point represents the lift of the rule, a measure of how much more likely the consequent is to be purchased when the antecedent is also purchased compared to their independent probabilities. Lighter blue indicates lower lift, and darker red indicates higher lift.
- There appears to be a dense cluster of rules with relatively low support (below 0.05) but varying levels of confidence and lift.
- Some rules exhibit higher confidence (above 0.6) but generally have lower support (around or below 0.03), suggesting strong associations for less frequent itemsets.

- A separate cluster of rules shows moderate support (around 0.12 to 0.20) and moderate confidence (around 0.35 to 0.45), with generally lower lift values.
- Rules with the highest lift (darker red) tend to be concentrated in the lower support range, indicating strong associations between less frequently occurring itemsets.
- There are fewer rules with both high support and very high confidence and lift simultaneously in this visualization.

Now have a look at Andrews plot for Orders -

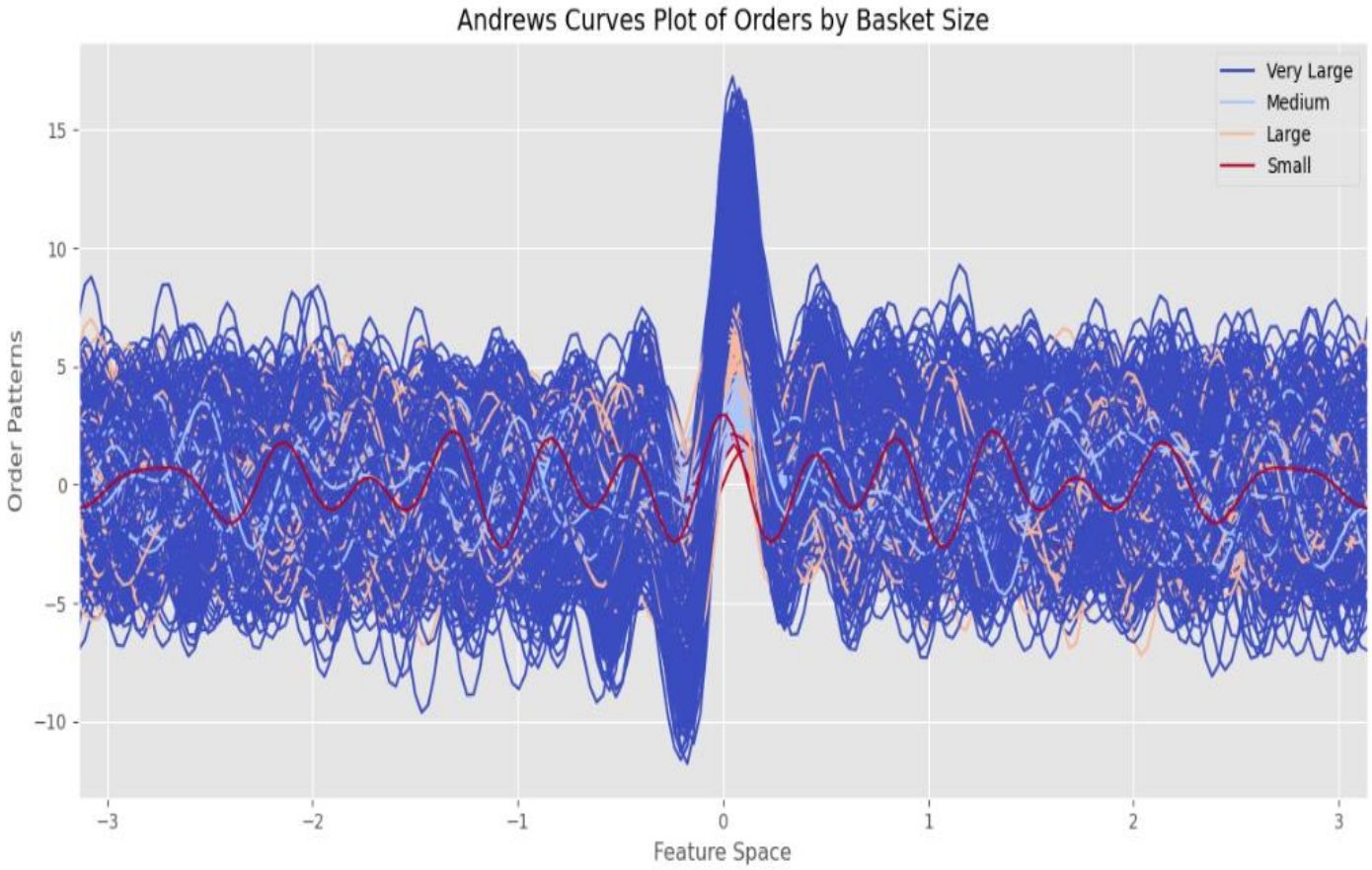

FIGURE 10 - ANDREWS PLOT FOR ORDERS W.R.T BASKET SIZE

- The Andrews Curves plot visualizes customer orders based on their basket size, where each curve represents a single order. The x-axis represents a transformed feature space, and the y-axis shows the resulting curve values

- Orders are categorized by basket size (Very Large, Medium, Large, Small), indicated by different coloured lines. The predominantly blue curves suggest a large number of orders fall into the 'Very Large' basket size category.
- 'Small' basket size orders, represented by red curves, exhibit a distinct, lower amplitude wave pattern compared to the larger baskets. This indicates a different underlying product composition for smaller orders.
- 'Medium' and 'Large' basket size orders (light orange and light blue respectively) show more variability and tend to have higher amplitude curves, particularly around the central feature space.
- The clustering of curves within each color group suggests that orders of similar basket sizes have some commonality in their product composition patterns.
- The sharp peak around the feature space value of 0 indicates a feature that strongly differentiates the order patterns, particularly for larger basket sizes.
- This plot allows for the visual identification of patterns and potential groupings of customer orders based on the characteristics of the products they purchase, linked to their overall basket size.

Market Based Analysis (MBA)

It is a technique used to identify associations between different items that are frequently purchased together. It's often used in retail to understand customer buying behavior, optimize product placement, create targeted promotions, and build recommendation systems.

Lets dive into –

Our dataset contains Point of Sale (POS) transaction data with the following structure:

- **Date**: Transaction date (e.g., "01-01-2018")
- **Order_id**: Unique identifier for each customer order
- **Product**: Product purchased in that order

We want to:

- Identify frequently purchased product combinations
- Use these patterns to build combo offers/discounts
- Increase basket size, revenue, and customer retention
-

Steps -

1. Read the dataset using CSV reader
2. Then use Group by node to sort it by 'Order'
3. Then we use Cell spliter to convert it to Basket format to get ouput as Collection cell.
4. Then we will use Association Rule Learner (from KNIME Labs). It will give us-
   - **Support**: Fraction of orders that contain both items
   - **Confidence**: Likelihood of buying item B given A
   - **Lift**: Measure of how much more likely B is given A vs randomly

And finally we can visualize the results using plots.

At step 4, we get results –

Table "default" - Rows: 1332 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.128	0.346	0.973	pork	<---	[fruits]
rule1	0.128	0.36	0.973	fruits	<---	[pork]
rule2	0.128	0.342	0.929	butter	<---	[beef]
rule3	0.128	0.348	0.929	beef	<---	[butter]
rule4	0.129	0.369	0.976	laundry det...	<---	[sandwich lo...
rule5	0.129	0.341	0.976	sandwich lo...	<---	[laundry det...
rule6	0.13	0.376	1.056	pork	<---	[hand soap]
rule7	0.13	0.365	1.056	hand soap	<---	[pork]
rule8	0.132	0.377	1.06	pork	<---	[sandwich lo...
rule9	0.132	0.37	1.06	sandwich lo...	<---	[pork]
rule10	0.132	0.365	1.026	pork	<---	[sugar]
rule11	0.132	0.37	1.026	sugar	<---	[pork]
rule12	0.132	0.365	0.992	sandwich bags	<---	[sugar]
rule13	0.132	0.358	0.992	sugar	<---	[sandwich b...
rule14	0.132	0.381	1.035	butter	<---	[hand soap]
rule15	0.132	0.358	1.035	hand soap	<---	[butter]
rule16	0.132	0.342	0.988	hand soap	<---	[bagels]
rule17	0.132	0.381	0.988	bagels	<---	[hand soap]
rule18	0.132	0.377	0.971	dishwashing...	<---	[sandwich lo...
rule19	0.132	0.339	0.971	sandwich lo...	<---	[dishwashin...
rule20	0.133	0.383	1.086	flour	<---	[hand soap]
rule21	0.133	0.376	1.086	hand soap	<---	[flour]
rule22	0.133	0.355	1.007	flour	<---	[spaghetti s...
rule23	0.133	0.376	1.007	spaghetti sa...	<---	[flour]

TABLE 4 - ASSOCIATION RULES WITH METRICS

Explanation of Metrics

- **Support**: % of total transactions that include both A & B

  → E.g., 3% of transactions have both yogurt & pork

- **Confidence**: % of transactions with A that also have B

  → If yogurt is bought, 45% chance pork is also bought

- **Lift**: How much more likely is B given A vs overall

  → Lift > 1 means a positive association (strong rule)

Finally, our output table is –

Association Rules Output Table

Show 10 entries Search:

RowID	Support	Confidence	Lift	Consequent	implies	Items
rule2	0.12818261633011413	0.34192037470725994	0.9294685126290431	butter	<---	[beef]
rule3	0.12818261633011413	0.3484486873508353	0.9294685126290431	beef	<---	[butter]
rule203	0.1431079894644425	0.36302895322939865	0.9549422118436144	milk	<---	[waffles]
rule202	0.1431079894644425	0.37644341801385683	0.9549422118436145	waffles	<---	[milk]
rule153	0.14223002633889376	0.36	0.9693617021276596	pasta	<---	[lunch meat]
rule152	0.14223002633889376	0.3829787234042553	0.9693617021276596	lunch meat	<---	[pasta]
rule19	0.13169446883230904	0.33936651583710403	0.9712021646695013	sandwich loaves	<---	[dishwashing liquid/detergent]
rule18	0.13169446883230904	0.37688442211055273	0.9712021646695013	dishwashing liquid/detergent	<---	[sandwich loaves]
rule0	0.12818261633011413	0.3459715639810427	0.9729916330232287	pork	<---	[fruits]
rule1	0.12818261633011413	0.36049382716049383	0.9729916330232287	fruits	<---	[pork]

Showing 1 to 10 of 1,332 entries Previous 1 2 3 4 5 … 134 Next

TABLE 5 - ASSOCIATION OUTPUT TABLE

KNIME Workflow-

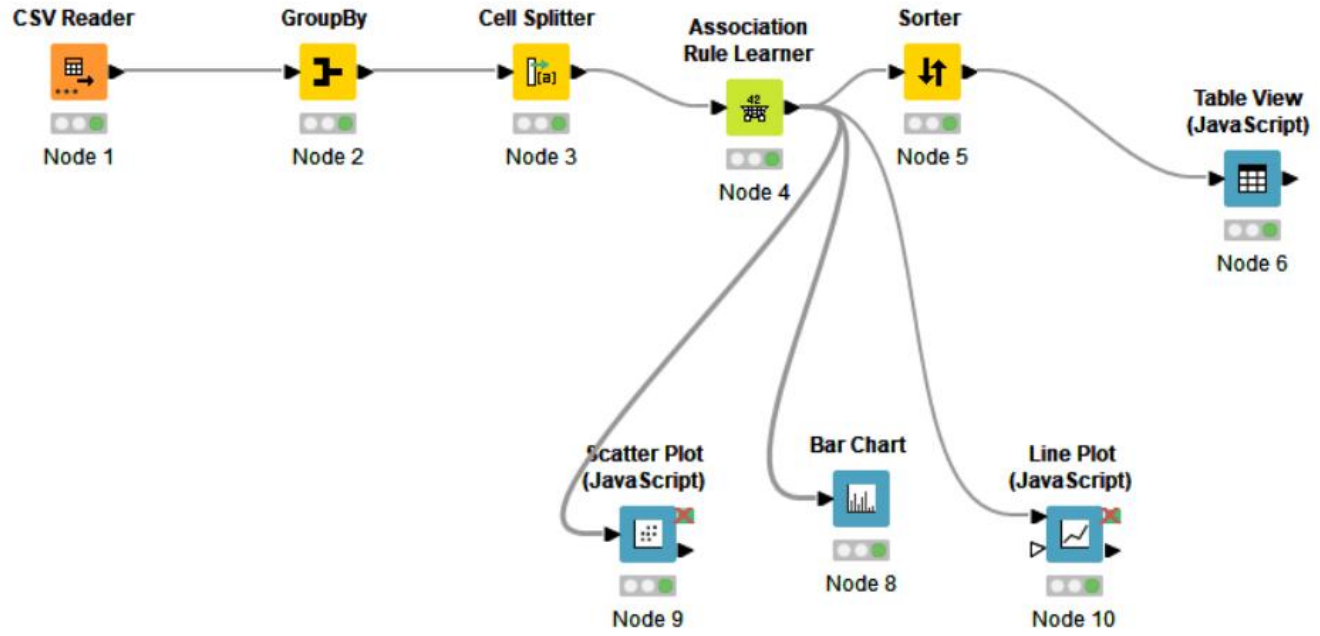

FIGURE 11 - MBA - KNIME WORKFLOW DIAGRAM

Now lets different plots for MBA –

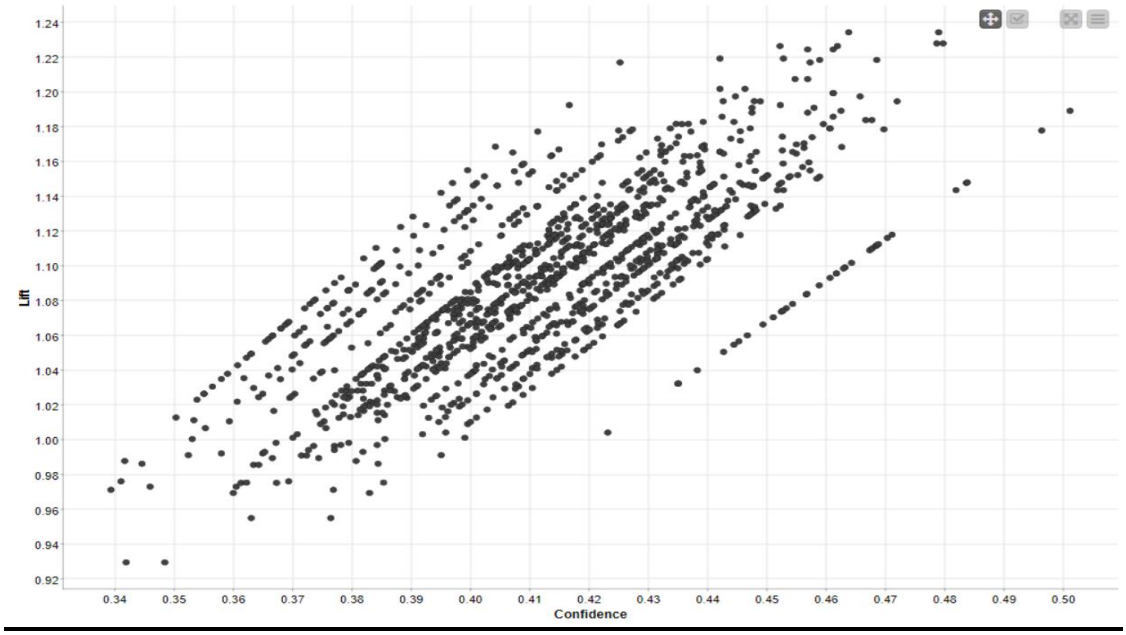

FIGURE 12 - SCATTER PLOT - LIFT VS CONFIDENCE

- The scatter plot shows a strong positive linear relationship between lift and confidence.
- Most rules cluster between confidence values of 0.37 to 0.45 and lift between 1.02 to 1.14.
- A few outlier rules have high confidence (>0.47) and lift above 1.2, indicating valuable associations.
- Rules with lift < 1.0 suggest no added association value — these should be ignored for offers.
- The tight linear bands suggest consistent rule quality and confidence distribution across products.
- Lift values are generally moderate, meaning products are somewhat dependent but not extremely so.
- These insights support identifying targeted combos with moderate reliability but scalable potential.

Also,

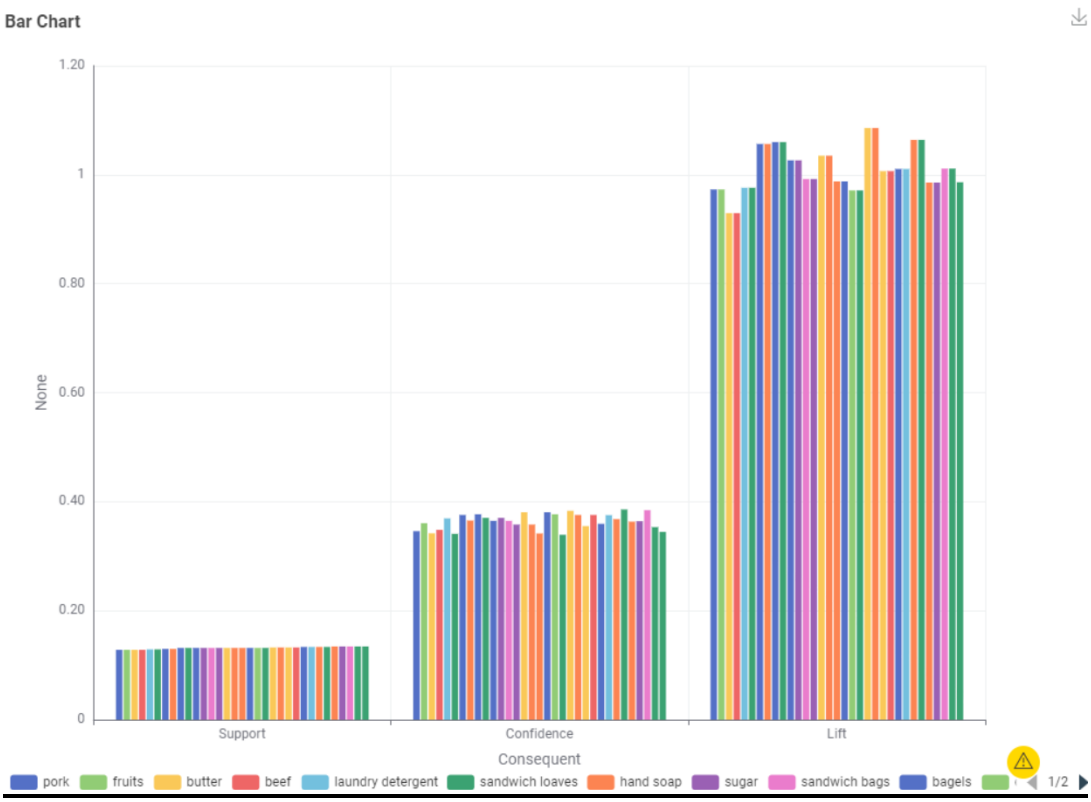

FIGURE 13 - BAR CHART SHOWING SUPPORT, CONFIDENCE, AND LIFT ACROSS CONSEQUENTS

- All consequents show moderate confidence (~0.34–0.37), suggesting consistent reliability in these associations.
- Lift values exceed 1.0 for every product, indicating positive correlations—customers tend to buy these items together.
- Items like hand soap, sandwich loaves, and bagels have slightly higher lift, making them ideal for bundled offers.
- Support values are uniformly low (~0.12), implying these combinations aren't frequent but still statistically strong.
- Laundry detergent and pork also show good lift with balanced confidence—great candidates for household combo packs.
- These patterns highlight potential for "Buy X, get Y at discount" or "Everyday bundle deals" targeting frequent household essentials.

Line plot-

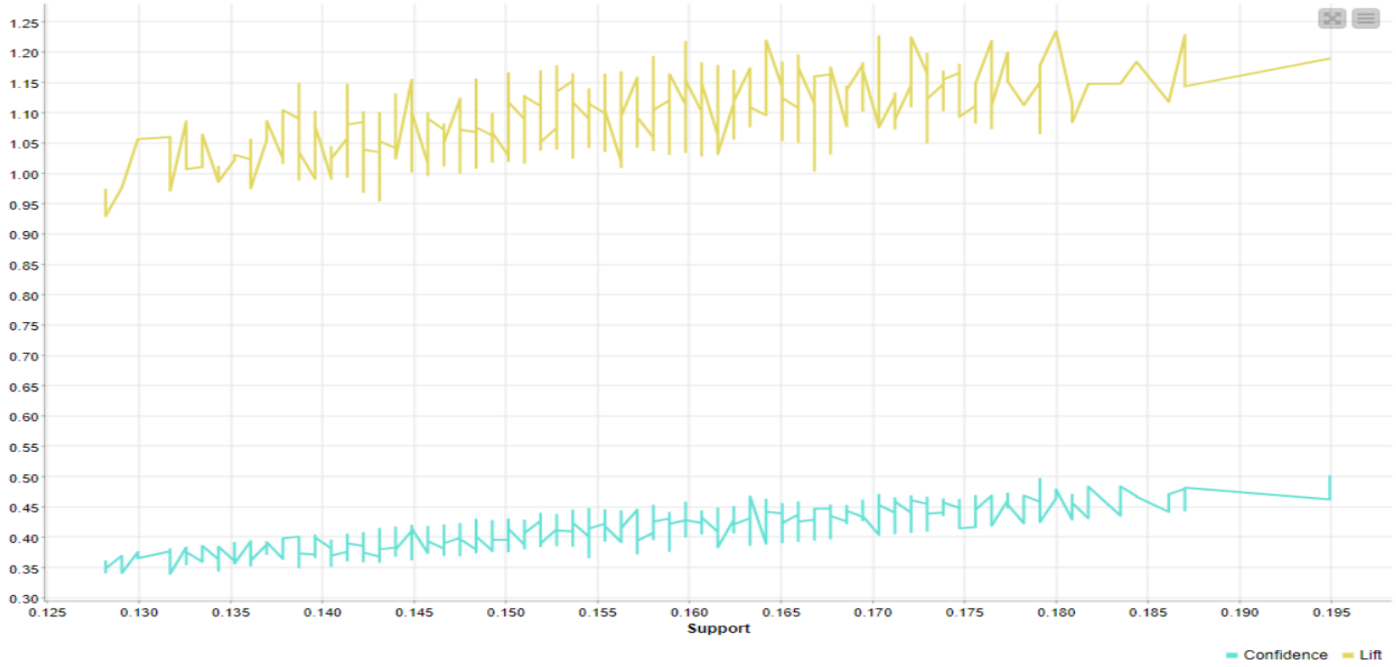

FIGURE 14 - LINEPLOT LIFT VS CONFIDENCE

- **Positive Correlation Trend:** Both confidence (cyan) and lift (yellow) exhibit a general positive trend as the support (x-axis) increases. This suggests that rules based on more frequently occurring itemsets tend to have higher confidence and lift values, indicating stronger and more reliable associations.

- **Higher Lift Values:** Consistently, the lift values are significantly higher than the confidence values for the majority of the plotted rules across the observed support range. This implies that the presence of the antecedent itemset increases the likelihood of the consequent itemset occurring together more substantially than just the overall frequency of the consequent.

- **Scatter and Variability:** Despite the general upward trend, there is considerable scatter in both confidence and lift values for any given level of support. This indicates that for itemsets with similar support, the strength and reliability of the association rules can vary significantly depending on the specific items involved.
- **Denser Clustering at Lower Support:** The plot appears to show a higher density of data points in the lower support range (towards the left side of the x-axis). This suggests that a larger number of association rules are generated from less frequent itemsets compared to those with higher support.

- **Wider Range at Higher Support:** As support increases, the range of both confidence and lift values seems to widen. This could imply that while higher support rules tend to be stronger overall, there's still a considerable variation in the strength of these associations.
- **Support as a Filter:** The visualization implicitly highlights the role of support as a potential initial filter for association rules. Rules with very low support might represent coincidental co-occurrences rather than

meaningful associations, and focusing on rules with higher support could be a starting point for identifying more robust patterns.

- **Potential for Actionable Insights at Higher Lift:** Rules with higher lift values, especially those with reasonably good support and confidence, are often the most actionable for business decisions, such as product placement or targeted promotions, as they indicate strong positive dependencies between items.

- **Need for Further Investigation:** This scatter plot provides a good overview of the generated association rules. However, to gain deeper insights, it would be necessary to examine the specific itemsets corresponding to the points of interest (e.g., high lift, high confidence, high support) and apply further filtering or analysis based on business objectives.

So as per MBA, we can confirm on below points-

Key Insights:

- High-lift rules show strong item affinity — bundle these into combo offers.

- Items like milk, bread, and eggs may appear in high-support rules — these can be part of loss leader promotions.

- Confidence-heavy rules help with recommendation systems (e.g., Amazon-style “People who bought X also bought Y”).

Recommendations:

- Create Combo Offers: Based on top rules by lift/confidence.

- Cross-Selling: Use heatmap insights to place related items together in-store.

- Targeted Promotions: Use frequent itemsets for personalized coupons.

- Inventory Optimization: Stock top-right quadrant items (from scatter plot) more heavily.

- Customer Loyalty: Offer loyalty points when customers buy high-lift combinations.

Inferences and Recommendations

Key Insights from the overall Analysis

1. Association Rule Metrics

Rules with high lift (>3) and moderate to high confidence (>0.5) indicate strong, non-random item relationships.

Majority of high-lift associations occur at lower support levels, suggesting niche but valuable combinations.

2. Top Product Pairs Heatmap

Frequently purchased together pairs include:

- cheese & milk
- dinner rolls & bread
- ice cream & individual meals
- poultry & cheese
- dishwashing liquid & detergent

These indicate common basket combos across groceries and household categories.

3. Boxplot of Basket Size

Median basket contains around 5–6 unique items, with many orders on the lower end.

Suggests opportunities to increase basket size through intelligent combo/discount strategies.

4. Transaction Trend Over Time

Regular seasonal dips (likely holidays or system downtimes).

Peak months (e.g., mid-2018 and early 2019) show when promotional offers might have succeeded.

Business Recommendations

1. Combo Offers Based on Strong Rules

Combo Items	Strategy Suggestion	Why?
Milk + Cheese	"Breakfast Starter Pack"	High lift + high confidence
Dinner Rolls + Bread	"Family Meal Deal"	Frequently paired
Ice Cream + Individual Meals	"Chill & Dine Combo"	Target evening/snack buyers
Poultry + Cheese	"Protein Pack Combo"	Relevant for non-veg customers
Dishwashing Liquid + Detergent	"Home Essentials Bundle"	Utility-driven purchase combo

2. Discount Offers for Volume & Retention

Offer Type	Targeted Items	Expected Impact
Buy 2 Get 1 Free	Ice Cream, Sodas, Dinner Rolls	Boost volume on popular low-margin goods
10% off on Combos	Poultry + Cheese, Milk + Bread	Encourage bundling of high-demand items
Loyalty Discount	Repeat buyers of detergents or cereals	Boost repeat purchase and retention
Bundle Packs for Weekends	Family Meal Combo (bread, rolls, cheese)	Drive weekend grocery traffic

3. Personalized Offers & Merchandising

- Milk/Cheese/Ice Cream could be placed adjacently in aisles or highlighted in weekly flyers.

- Detergents and Dishwashing Liquid: Target with home care bundle ads.

- Use email/SMS campaigns for users who’ve previously bought any item from a rule to recommend the related item(s).

4. Monitor KPIs Going Forward

Track: Lift, Confidence, Avg Basket Size, Redemption Rate on Offers