Report of Q4
Recommendation and Business Analysis

50015627 JIANG Zhuoyang

1. Question Description:

I was offered 1M retail transaction records from a specific city covering the period from 2009 to 2011. Assuming the role of a Business Analyst(BA), my objective is to compose a report based on this dataset to provide business insights for a retail supermarket. This report encompass the following two components:

- Utilize visual methods to create a minimum of 10 pictures, delivering no fewer than 10 business insights to the supermarket owner.
- Utilize association rule analysis to offer the supermarket owner no fewer than 10 sales (recommendation) suggestions.

2. Load Data and Do Preprocessing

Load data and do Data Preprocessing as below:

- Handle non-finite values in 'Quantity' column before conversion
- Convert 'Quantity' column to integer type after handling non-finite values
- Fill missing values in 'Description' column with 'Unknown'
- Fill missing values in 'Customer ID' column using forward fill method
- Convert 'Date' column to datetime type
- Remove duplicate records and outliers

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 7 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   Transaction_id  1000000 non-null  object
 1   Product_id      1000000 non-null  object
 2   Description     995669 non-null   object
 3   Quantity        1000000 non-null  float64
 4   Date            1000000 non-null  object
 5   Price           1000000 non-null  float64
 6   Customer ID     774502 non-null   float64
dtypes: float64(3), object(4)
memory usage: 56.0+ MB
```

Figure.1 Load data and its information before data preprocessing

```
<class 'pandas.core.frame.DataFrame'>
Index: 942779 entries, 0 to 999999
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Transaction_id  942779 non-null  object
 1   Product_id      942779 non-null  object
 2   Description     942779 non-null  object
 3   Quantity        942779 non-null  Int64
 4   Date            942779 non-null  datetime64[ns]
 5   Price           942779 non-null  float64
 6   Customer ID     942779 non-null  float64
dtypes: Int64(1), datetime64[ns](1), float64(2), object(3)
memory usage: 58.4+ MB
```

Figure.2 Data information after data preprocessing

3. EDA:

(1) EDA to Get Basic Information:

Firstly I obtained the time span of the data and presented a bar chart showing the count of unique Customer IDs, Transaction IDs, and Product IDs. Printing out the date range in the data, these visualizations help us understand the fundamental characteristics and distribution of the data.

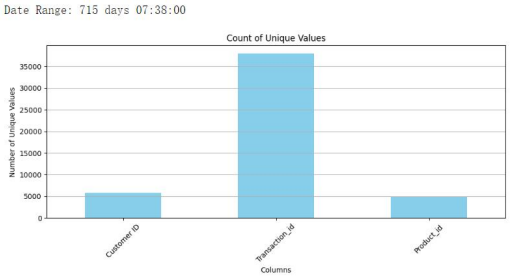


Figure.3 Basic Information

(2) Sales Trend (Exploration about Time):

① Monthly Sales Trend:

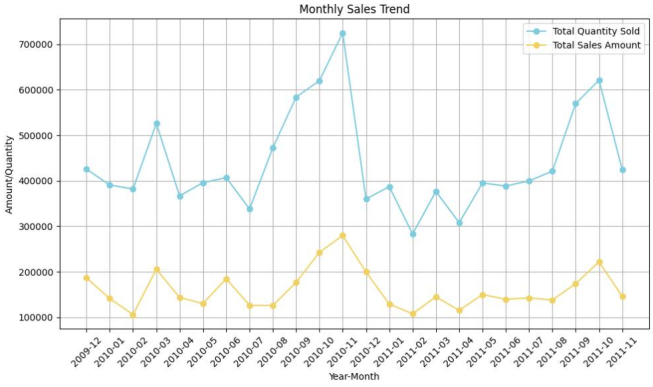


Figure.4 Monthly Sales Trend

Insight 1: Overall, late autumn to early winter marks a peak shopping period each year. Adequate restocking plans need to be established beforehand to prevent severe shortages. There's also a notable surge in demand between February and March annually, requiring proactive preparation.

② Sales Unit Price Trend:

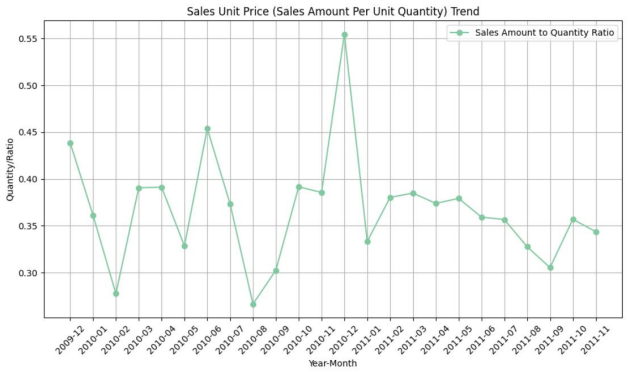


Figure.5 Sales Unit Price (Sales Amount Per Unit Quantity) Trend

Insight 2: In June and December of 2010, there were notably high sales-to-volume ratios, indicating potential value in certain product sales strategies during these months.

(3) Exploration about Pricing Strategy and Stock Selection Strategy:

① Price Distribution:

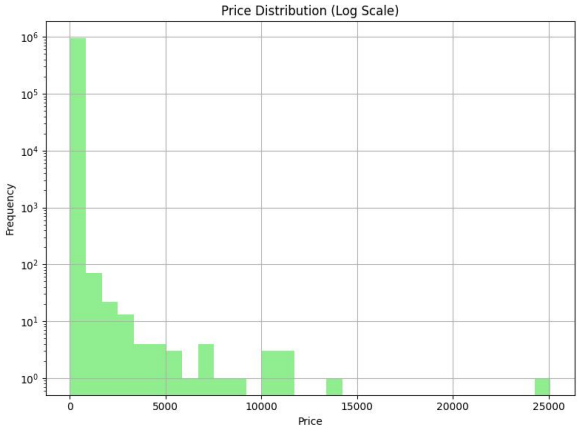


Figure.6 Price Distribution (Log Scale)

Insight 3: In June and December of 2010, there were notably high sales-to-volume ratios, suggesting that there might be valuable sales strategies for certain products during these two months.

② Sales Amount vs. Price:

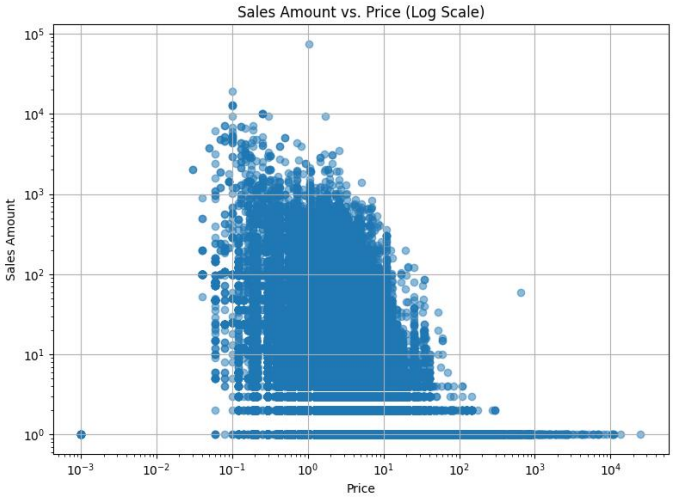
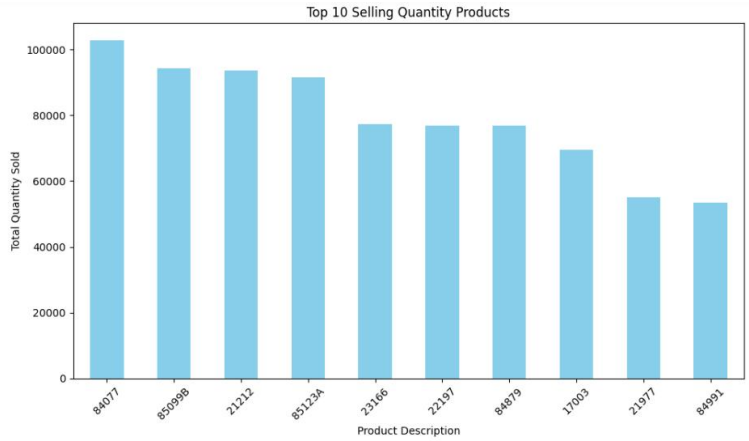


Figure.7 Sales Amount vs. Price (Log Scale)

Insight 4: Given the need for exploration within a wide range of values, constructing a scatter plot using a logarithmic coordinate system can better illustrate the relationship between price and sales. Overall, we can observe that there are more products with lower prices and higher sales volumes, while there are fewer products with both high prices and high sales volumes. As the price increases, the quantity of products with high sales decreases.

③ Top 10 Selling Quantity Products:

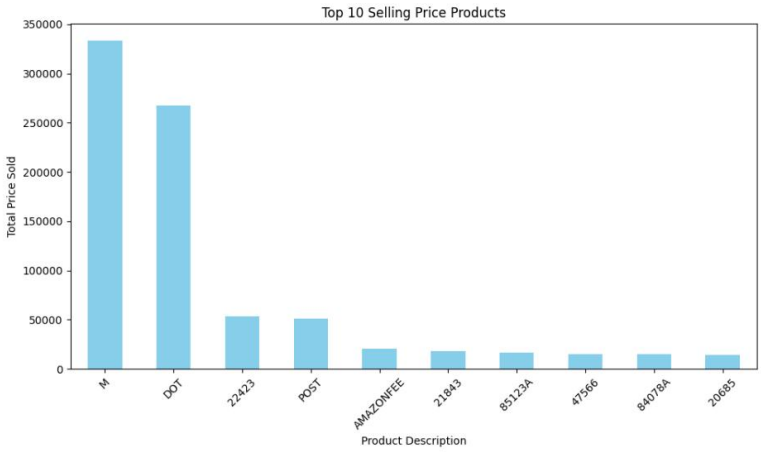


Product_id	Description	Price
17003	BROCADE RING PURSE	0.21
21212	PACK OF 72 RETRO SPOT CAKE CASES	0.55
21977	PACK OF 60 PINK PAISLEY CAKE CASES	0.55
22197	POPCORN HOLDER , SMALL	0.85
23166	MEDIUM CERAMIC TOP STORAGE JAR	1.04
84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	0.21
84879	ASSORTED COLOUR BIRD ORNAMENT	1.69
84991	60 TEATIME FAIRY CAKE CASES	0.55
85099B	JUMBO BAG RED WHITE SPOTTY	1.95
85123A	WHITE HANGING HEART T-LIGHT HOLDER	2.95

Figure.8 Top 10 Selling Quantity Products

Insight 5: Upon analyzing the top ten best-selling products, it's evident that most of these items are compact, affordable, decorative, or intended for everyday use. They are suitable as gifts, decorations, or for daily use. It would be advisable to maintain or even increase their stock quantities.

④ Top 10 Selling Price Products:



Product_id	Description	Price
20685	RED SPOTTY COIR DOORMAT	6.75
21843	RETRO SPOT CAKE STAND	21.7
22423	REGENCY CAKESTAND 3 TIER	12.75
47566	PARTY BUNTING	4.65
84078A	SET/4 WHITE RETRO STORAGE CUBES	34.95
85123A	WHITE HANGING HEART T-LIGHT HOLDER	2.95
AMAZONFEE	AMAZON FEE	6706.71
DOT	DOTCOM POSTAGE	647.19
M	Manual	4.0
POST	POSTAGE	18.0

Figure.9 Top 10 Selling Price Products

Insight 6: After analyzing the top ten highest revenue-generating products in contrast with the top ten best-selling products, it was observed that only the product with ID 85123A (WHITE HANGING HEART T-LIGHT HOLDER) is both the highest in

sales volume and revenue. Other items, to a greater or lesser extent, achieve higher revenue due to their uniqueness or higher unit price. They can be categorized into three main types:

- 1. Special Fees: AMAZON FEE, DOTCOM POSTAGE, POSTAGE
- 2. Special Items: Manual
- 3. Items Similar to the Highest Selling Products: Compact, affordable, d ecorative, or intended for everyday use

This further underscores the commercial value of small, affordable, decorati ve, or utilitarian items."

⑤ Bottom 10 Selling Price Products:

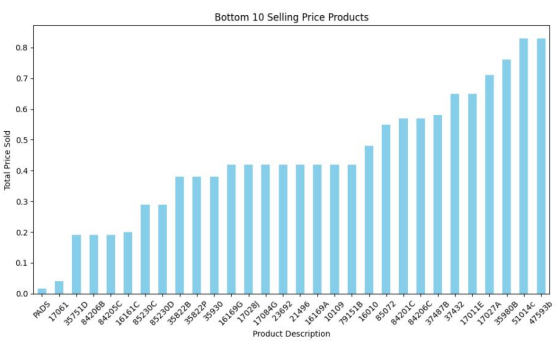


Figure.10 Bottom 10 Selling Price Products

Insight 7: The ten lowest revenue-generating products were analyzed, and as expected, these items have relatively low unit prices and very low sales volume. They yield limited profits and can be reasonably concluded as non-essential items, providing low returns. Therefore, it might be advisable to consider reducing their purchase quantities.

(4) Exploration about the Customers:

① Distribution of Transactions per Customer:

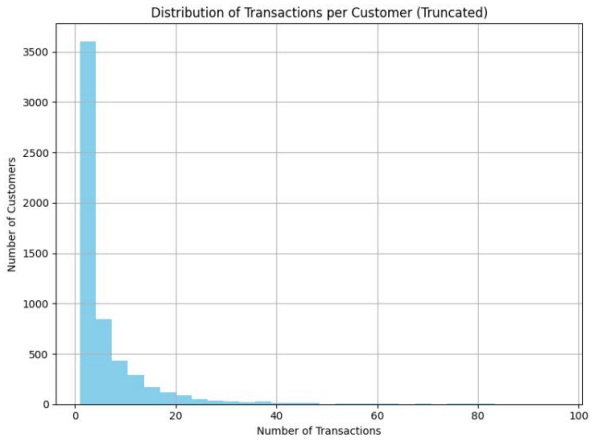


Figure.11 Distribution of Transactions per Customer (Truncated)

Insight 8: The analysis of customer transaction volume revealed a pronounced long-tail effect. This phenomenon indicates that only a small portion of customers engage in a large number of transactions, while the majority have relatively few transactions. To address this, we truncated the data, focusing solely on the transaction behavior of the majority of customers. Users with transactions exceeding 20 were considered loyal customers. Analyzing their extensive transaction characteristics enabled personalized recommendations tailored to their preferences.

② **Log Purchase Frequency Distribution per Customer:**

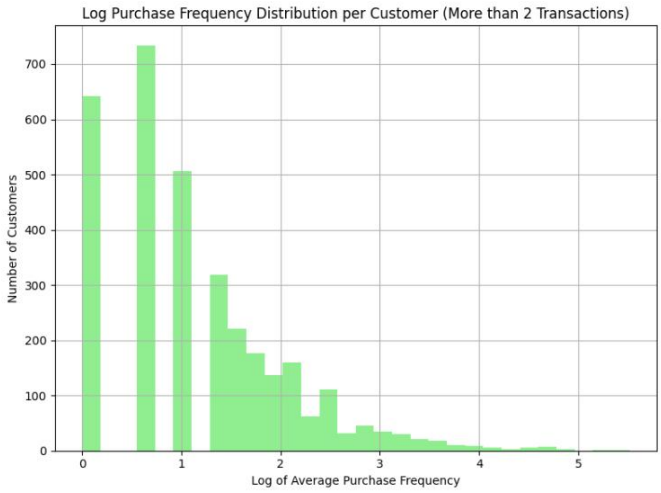


Figure.12 Log Purchase Frequency Distribution per Customer (More than 2 Transactions)

Insight 9: By calculating the average transaction interval between purchase dates for each customer (in days), we estimated their purchasing frequency. To mitigate the impact of the long-tail effect on visualization, we utilized logarithmic scaling. Our analysis indicates that customers with multiple transaction records may predominantly consist of loyal customers due to their lower purchase intervals and higher frequency.

③ **Transaction of Customers with Transaction Intervals Between 3 and 30 Days:**

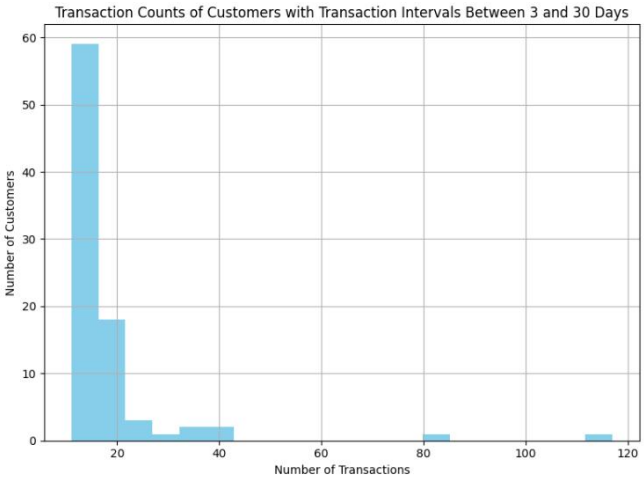


Figure.13 Transaction of Customers with Transaction Intervals Between 3 and 30 Days

Insight 10: However, upon further analysis of the specific transaction amounts for users with transaction intervals between 3 and 30, it appears that their transaction counts are not particularly high. Hence, it's essential to consider transaction amounts in conjunction with these findings.

④ **Purchase Frequency vs Transaction Counts for Selected Core Customers:**

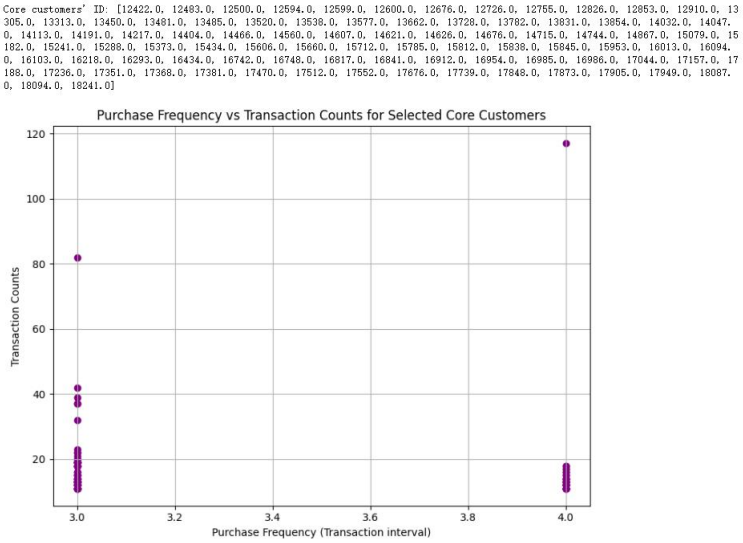


Figure.14 Purchase Frequency vs Transaction Counts for Selected Core Customers

Insight 11: I further identified users with transaction counts exceeding 10 and transaction intervals between 5 to 30 days, considering them as core users. Observing their transaction behavior affirmed this as a reasonable conclusion.

4. Association Rule Analysis

(1) Algorithm:

There are two association mining algorithms that I have considered, one is the FP growth algorithm, and the other is the Apriori algorithm.

The FP growth algorithm is an association rule mining algorithm based on the FP tree. It represents the dataset by constructing an FP tree and utilizes the properties of the FP tree to mine frequent itemsets. Compared to the Apriori algorithm, the FP growth algorithm has higher efficiency, especially in handling large-scale datasets. It only requires two scans of the dataset, making it more efficient in terms of computational overhead.

So I ultimately chose the FP growth algorithm.

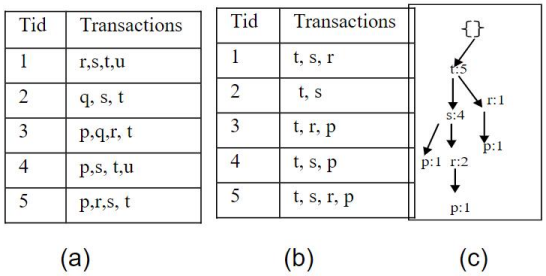


Figure.15 Example: a) Initial Dataset b) Projected Dataset with min-threshold= 50% c) FP-tree

(2) Result of association rule analysis

Sort the rules by the 'lift' column and selecting the top 100 rules. For each rule Checking inclusion in other rules, retain a rule if its itemset is not included in others or it has a higher lift. Output the top 10 rules in the remained rules as below:

	antecedents_description	consequents_description	support	confidence	lift
1742	[POPPY'S PLAYHOUSE LIVINGROOM , POPPY'S PLAYHO...	[POPPY'S PLAYHOUSE BEDROOM]	0.010158	0.869074	50.286906
1640	[WOODEN STAR CHRISTMAS SCANDISPOT]	[WOODEN TREE CHRISTMAS SCANDISPOT]	0.010369	0.635922	47.915420
618	[COFFEE MUG CAT + BIRD DESIGN]	[COFFEE MUG DOG + BALL DESIGN]	0.010079	0.622150	46.599761
36	[SMALL DOLLY MIX DESIGN ORANGE BOWL]	[SMALL MARSHMALLOWS PINK BOWL]	0.011108	0.632132	43.012222
1638	[WOODEN HEART CHRISTMAS SCANDISPOT]	[WOODEN STAR CHRISTMAS SCANDISPOT]	0.011609	0.635838	38.993958
1204	[KEY FOB , BACK DOOR]	[GARAGE KEY FOB]	0.010000	0.583975	37.704711
822	[SET/10 PINK POLKADOT PARTY CANDLES]	[SET/10 BLUE POLKADOT PARTY CANDLES]	0.012612	0.608917	37.525141
1770	[GARDENERS KNEELING PAD CUP OF TEA]	[GARDENERS KNEELING PAD KEEP CALM]	0.011451	0.719735	37.469703
1206	[KEY FOB , SHED]	[GARAGE KEY FOB]	0.011530	0.554569	35.806043
786	[PINK POLKADOT CUP]	[BLUE POLKADOT CUP]	0.013509	0.705234	35.543051

Figure.16 Top ten filtered rules

(3) Sales/Recommendation Suggestions

From this, we can suggest that merchants combine or bundle the pre - and post product sets of these ten rules to sell to customers, thereby generating ten recommendations/sales recommendations:

- **Recommendation 1 by the Rule 1742:** Suggest putting ["poppy's playhouse livingroom ", "poppy's playhouse kitchen"] together with ["poppy's playhouse bedroom "] or bundling them for sale
- **Recommendation 2 by the Rule 1640:** Suggest to put ['wooden star christmas scandispot'] together with ['wooden tree christmas scandispot'] or to bundle them for sale
- **Recommendation 3 by the Rule 618:** Recommend to put ['coffee mug cat + bird design'] together with ['coffee mug dog + ball design'] or to bundle them for sale
- **Recommendation 4 by the Rule 36:** Suggest to put ['small dolly mix design orange bowl'] together with ['small marshmallows pink bowl'] or to bundle them for sale
- **Recommendation 5 by the Rule 1638:** Suggest to put ['wooden heart christmas scandispot'] together with ['wooden star christmas scandispot'] or to bundle them for sale
- **Recommendation 6 by the Rule 1204:** Suggest to put ['key fob , back door '] together with ['garage key fob'] or to bundle them for sale
- **Recommendation 7 by the Rule 822:** Suggest to put ['set/10 pink polkadot party candles'] together with ['set/10 blue polkadot party candles'] or to bundle them for sale
- **Recommendation 8 by the Rule 1770:** Suggest to put ['gardeners kneeling pad cup of tea '] together with ['gardeners kneeling pad keep calm '] or to bundle them for sale
- **Recommendation 9 by the Rule 1206:** Suggest to put ['key fob , shed'] together with ['garage key fob'] or bundling them for sale
- **Recommendation 10 by the Rule 786:** Recommend to put ['pink polkadot cup'] together with ['blue polkadot cup'] or to bundle them for sale