MRA PROJECT REPORT

DSBA

Rohit Nagarahalli

Contents

Table of Tables

Table of Figures

Table of Equations

Part – A

Agenda and Executive Summary

Data Used: Sales_Dala.xlsx

Worksheet 1: Consists of Data for Analysis

Worksheet 2: Consists of Data Dictionary

Goal: To establish RFM and segment customers based on the score.

Expected outcome: Clear understanding of customers who are Gold, Loyal, Verge of Churning and Lost.

Tools Used: KNIME, Tableau, Jupyter-notebook and Excel.

An automobile parts manufacturing company has collected data on transactions for 3 years. The Goal is to find the underlying buying patterns of the customers, provide the company with suitable insights about their customers, and recommend customized marketing strategies for different segments of customers.

About Data

Data Info, Shape, Summary and Assumptions

Column	Dtype
ORDERNUMBER	int64
QUANTITYORDERED	int64
PRICEEACH	float64
ORDERLINENUMBER	int64
SALES	float64
ORDERDATE	int64
DAYS_SINCE_LASTORDER	int64
STATUS	object
PRODUCTLINE	object
MSRP	int64
PRODUCTCODE	object
CUSTOMERNAME	object
PHONE	object
ADDRESSLINE1	object
CITY	object
POSTALCODE	object
COUNTRY	object
CONTACTLASTNAME	object
CONTACTFIRSTNAME	object
DEALSIZE	object

Table 1 Data Info

It was noted that the ORDERDATE column had int64 as the data type which was later converted to converted to proper Date format for analysis.

Data Shape

The data consists of 2747 records and 20 variables.

Summary stats

	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	DAYS_SINCE_LASTORDER	MSRP
count	2747	2747	2747	2747	2747	2747
mean	35.10302148	101.09895	6.491081179	3553.048	1757.085912	100.6917
std	9.762135424	42.04255	4.230543549	1838.954	819.2805763	40.1148
min	6	26.88	1	482.13	42	33
25%	27	68.745	3	2204.35	1077	68
50%	35	95.55	6	3184.8	1761	99
75%	43	127.1	9	4503.095	2436.5	124
max	97	252.87	18	14082.8	3562	214

Table 2 Summary Statistics

Assumptions about the Data

- A total of 2747 records and 20 variables were observed. For analysis of RFM 3 important parameters are required.
- R representing Recency, F representing Frequency and M representing Monetary.
- For Monetary the variable is straight forward and readily available. Sales according to the data summary is the multiplied value of quantity ordered to price of each item.
- Recency too was readily available in the name of DAYS_SINCE_LAST_ORDER. Important to note that the assumption was such that lesser the value most recently the visit.
- Frequency was not a readily available value, instead was extracted from the data by grouping by the data with respect to ORDERNUMBER and identifying the count.
- Only such customers were selected for analysis who had not cancelled their orders.
- It was assumed that the customers whose status was anything but cancelled will be proceeded for RFM analysis

Exploratory Data Analysis

Univariate

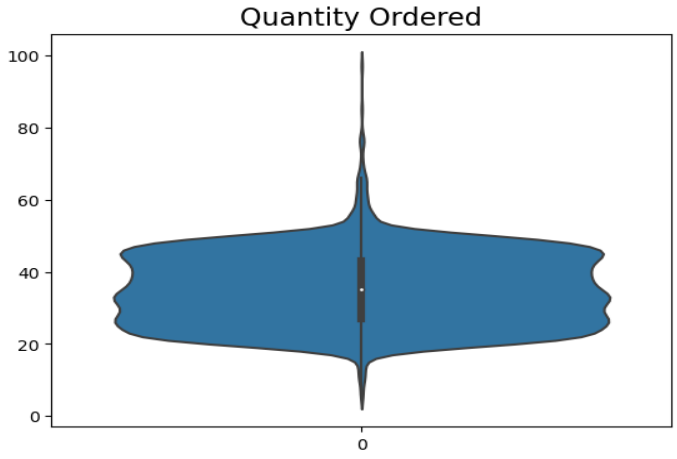

Figure 1 Quantity Ordered

Number of items ordered in each order majority of the time falls between 25 - 50. The "QUANTITYORDERED" variable appears to have slightly skewed towards right indicating presence of high ordered items in each order. The minimum number of orders placed in each order was observed to be 6 and maximum being 97.

Figure 2 Price Each

The variable "PRICEEACH" appears to have skewed towards its right indicating bulk of values with respect to PRICEEACH falling on the lower range and few outlying values where price of each items are higher.

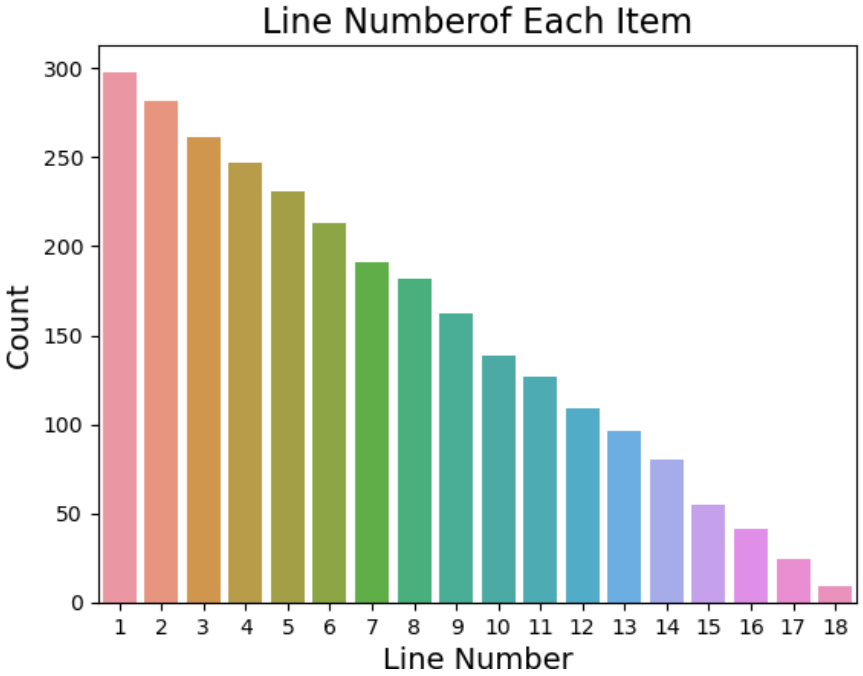

Figure 3 Line Number

The line number of each item within an order where line number 1 has the highest count and 18 being the lowest.

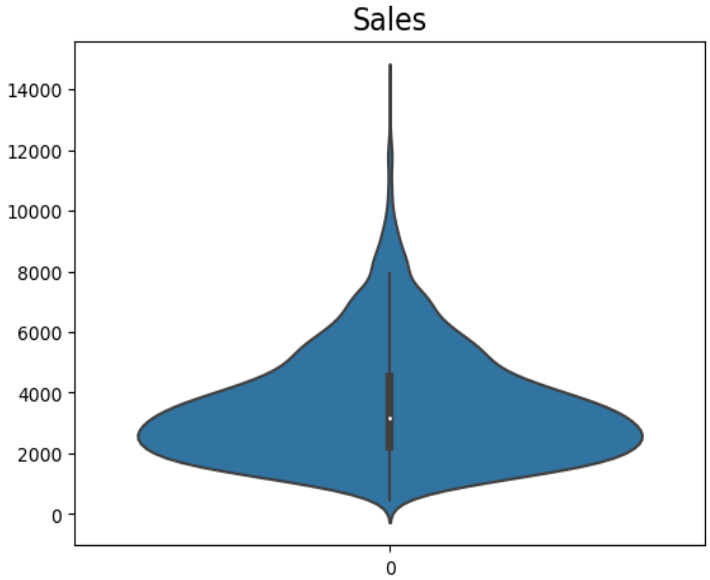

Figure 4 Sales

The Sales variable too has its majority of the data lying in the lower side and few on the positively skewed region. The skewness is elongated to its right and few higher sales have been observed. This was expected owing to the right skewness observed in "QUANTITYORDERED" and "PRICEEACH" since the "SALES" variable is nothing but a calculation of these two.

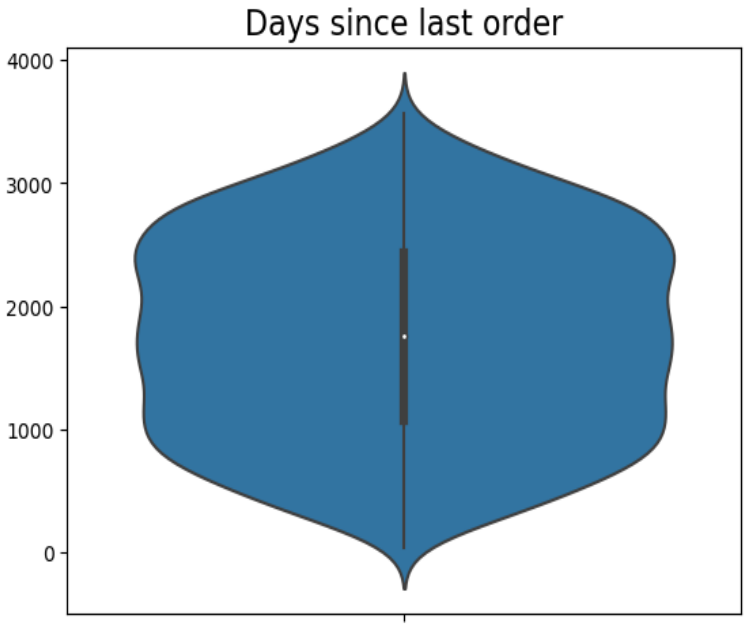

Figure 5 Days since Last Order

This is the metric that goes on to denote Recency in the RFM analysis. It is important to note that there are customers who have visited around 3500 days back and as low as 42 days. Any decisions such as whether or not the customers would come back again who have taken long to visit again will be analysed by segmenting the customers in the later part of analysis.

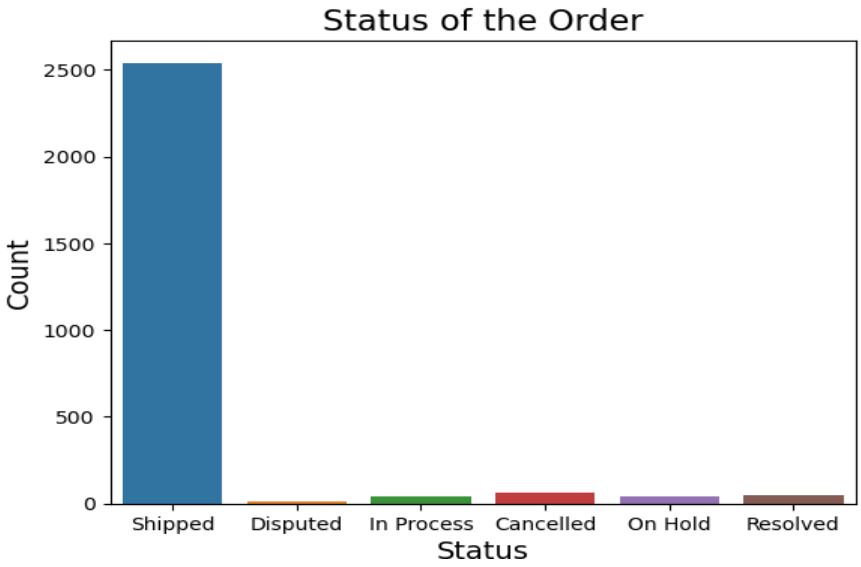

Figure 6 Status of the Order

Almost 95% of the orders have been shipped. Rest other categories of which "In Process", "On Hold" and "Resolved" would be shipped. However, there are very few disputed statuses and let us assume that the disputes shall be resolved by the team and proceed for shipping. The only records that cannot be taken further for analysis is the ones that are "Cancelled". Let us assume that cancelled orders refund back the money to customers and cannot be included in the further analysis.

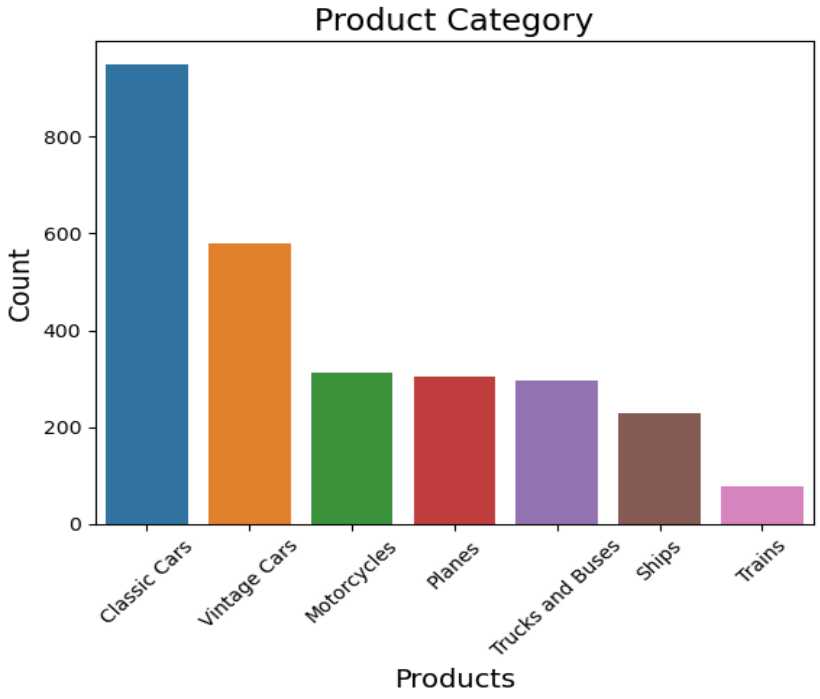

Figure 7 Products

Majority of the items belong to the Classic Cars Product line followed by Vintage Cars and the least being Trains.

The Manufacturer suggested selling price of each item. This attribute can be evaluated alongside the PRICEEACH attribute. Let us assume that the suggested selling price is the maximum price tag that can be allotted to the item and no items should cross the MSRP and could be less than MSRP and wouldn't harm the seller.

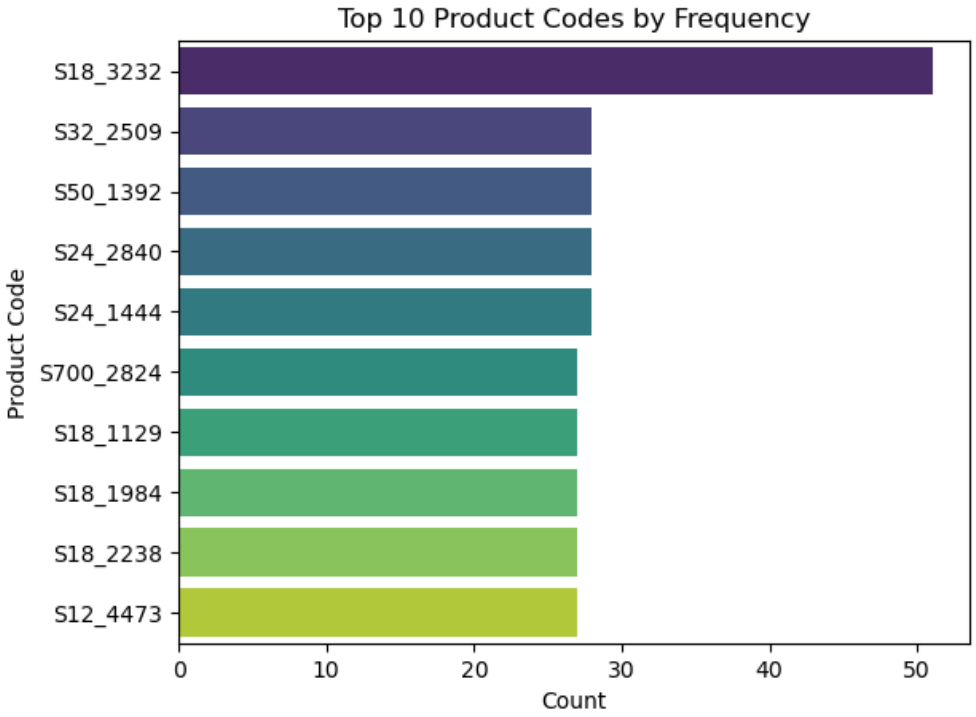

Figure 8 Top 10 Product Codes by Frequency

S18_3232 was observed to be the highest Product code that was preferred. Having enough of such highest count of products in stock more often could possibly address the demand for such products. Analysing the details of product and its need could potentially help promoting the product further and there by further increasing the sale.

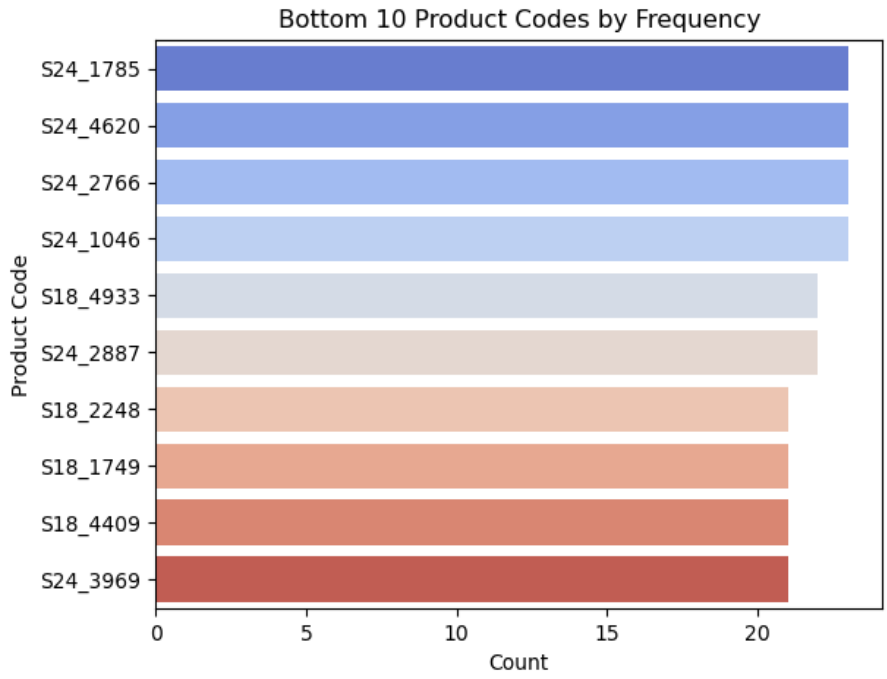

Figure 9 Bottom 10 Product

The top 10 and the bottom 10 product category count. Product code S18_3232 has the highest count indicating having sold the highest product. The bottom 4 have the same count and are the least sold products.

As for the least selling product in bottom four which are in the same count, observing their selling price and the sub-products that come under such Product code further giving offers on the selling price or probably clubbing the products with top category and providing a combo offer could help resolve the issue if exists.

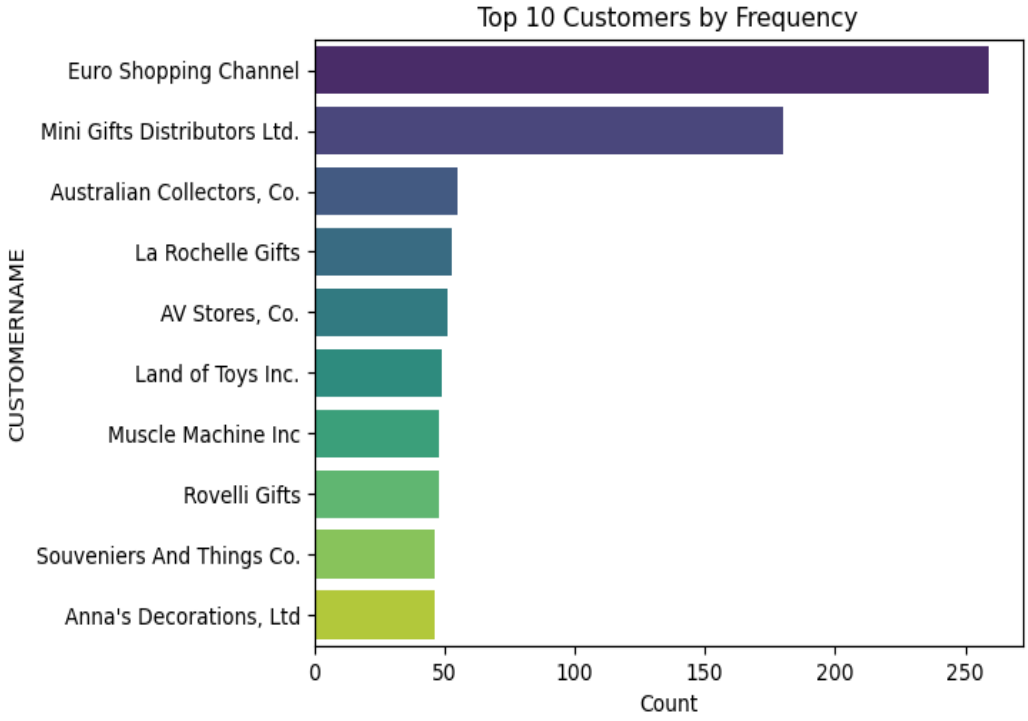

Figure 10 Top 10 Customers

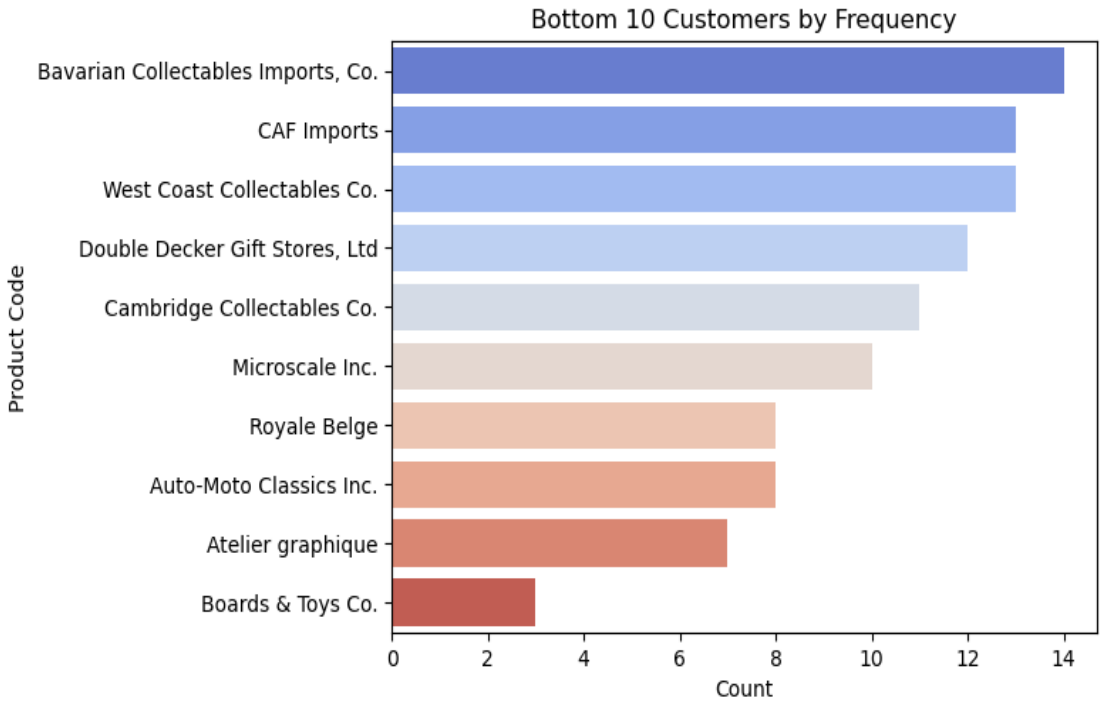

Figure 11 Bottom 10 Customers

The Top and Bottom 10 name of customer who placed the order. Euro shopping channel has placed the highest order and Boards & Toys Co. has placed the least orders. It is highly likely that top 10 customers would place orders in the future, if they don't then special attentions needs to be given to check what went wrong. The Bottom 10 customers who placed the orders might be the customers who recently got to know the company or might be the customers who did not return owing to dissatisfaction after visiting few times.

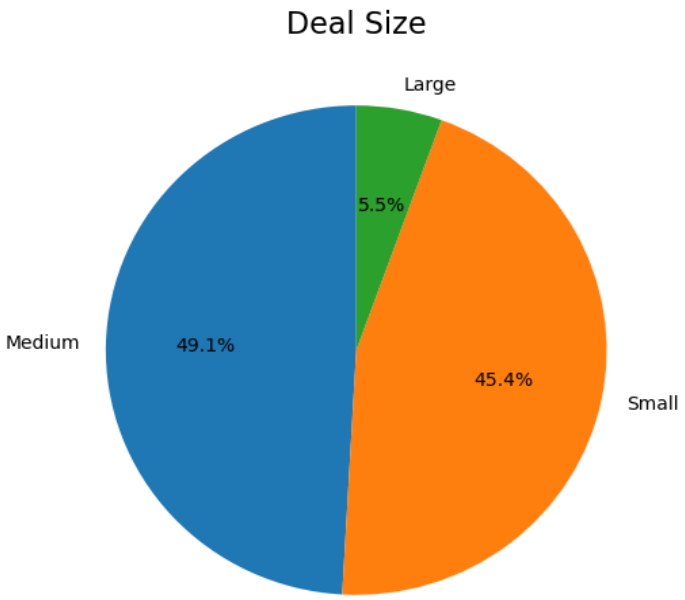

Figure 12 Deal Size

Medium followed by small size of the deals have highest categories of orders placed (approx. 95%) and large accounting only up to 5.5% of the total category.

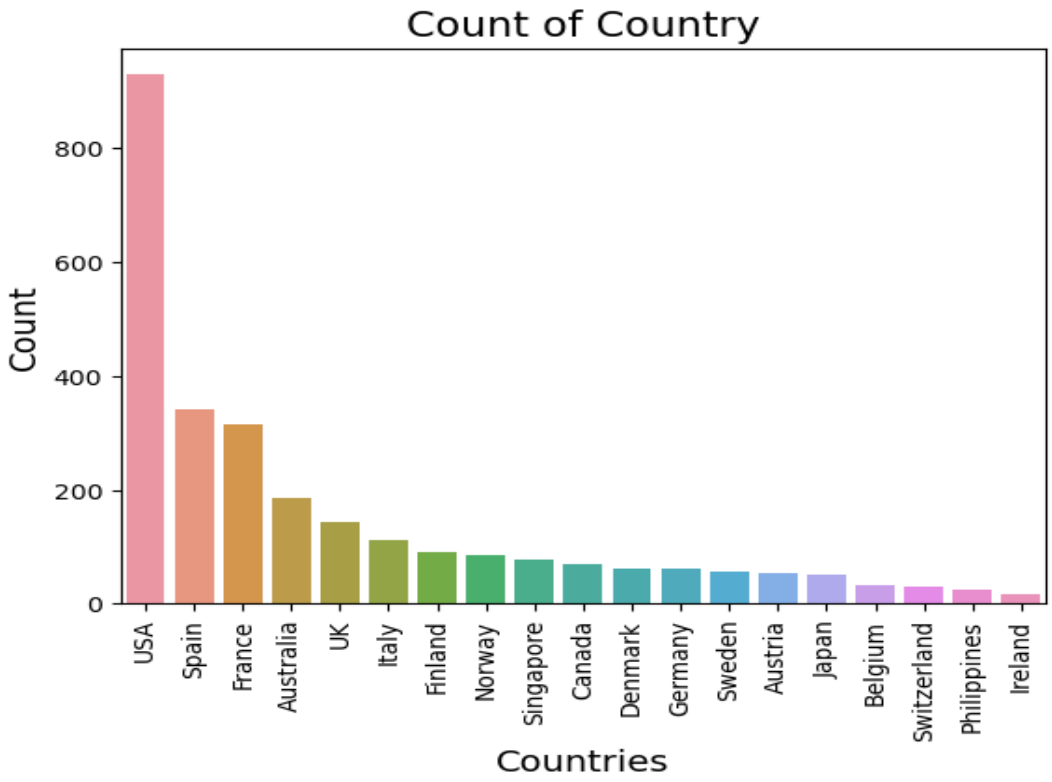

Figure 13 Count of Countries

Of the orders placed USA has the highest count. If there exists a customer who is likely to churn and is from USA a special focus on him/her is needed as he is likely to continue if given attention by providing discounts or other parameters. European countries such as Belgium, Switzerland and Ireland are on the lower side with lesser customers also Philippines. Expansion in such countries could see a growth in the near future by studying combination of patterns.

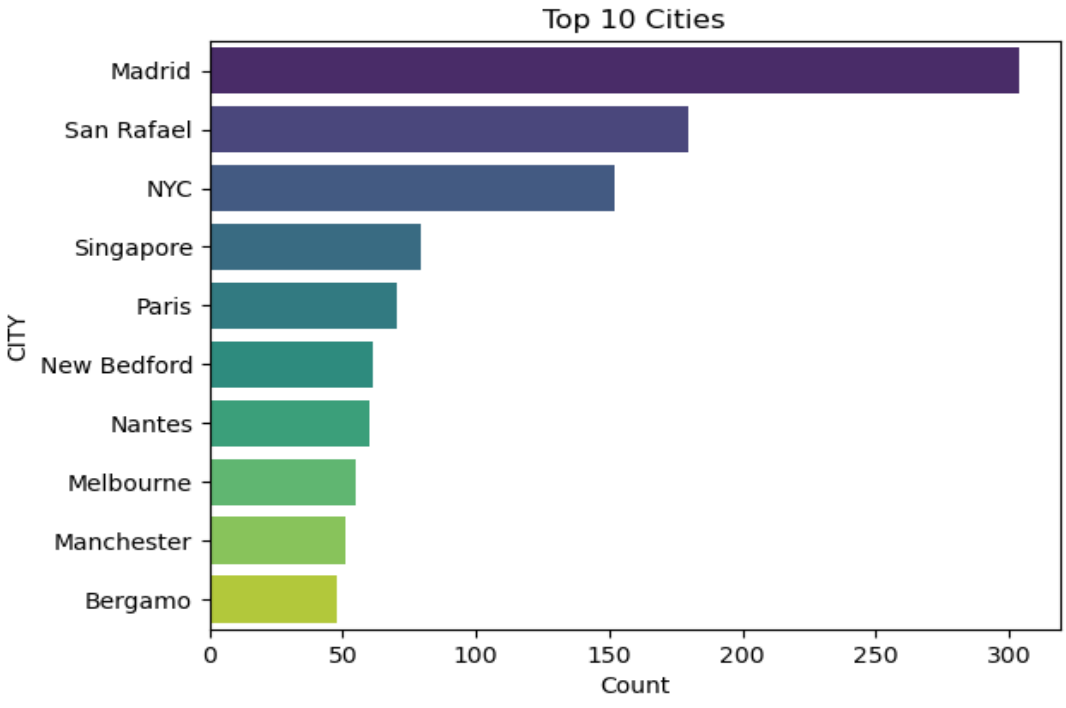

Figure 14 Top 10 Cities Count

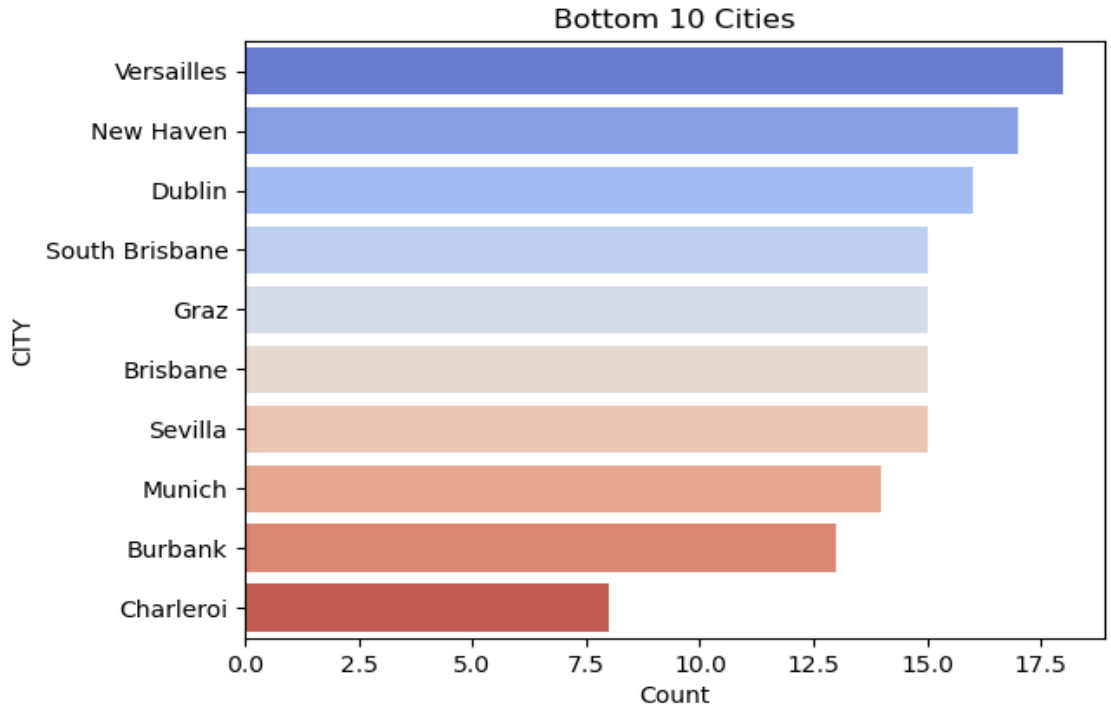

Figure 15 Bottom 10 Cities Count

Munich city has seen the highest orders in terms of cities. Although USA as a country has highest orders it is important to note that it has a greater number of cities. Charleroi a city in Belgium has seen the least number of orders.

Note: Inferences for the Univariate Analysis have been mentioned below the respective figures

Bivariate and Multivariate

Tableau reference link:

https://public.tableau.com/app/profile/rohit6191/viz/KNIME_data_analysis/Month_Year_Sales_Pattern?publish=yes

Inferences

- The yearly, monthly and weekly sales was observed. Although the data describes that the extracted data was for 3 years, the third-year data exists only up to 6 months (half-yearly).

- When analyzed monthly, November proved to be the highest sales for both the year 2018 and 2019. However, there was a drastic decline in the month of December after observing a rapid growth of sales in the month of November.

- It was also observed before hitting the peak at November, the sales were seen going above the average sales in the month of October. In summary, October and November were observed to be the peak months. The lowest sales observed was in January 2018 at 85,132 assuming this to be the start of the retail company it is obvious to seen lesser sales. However, the next lowest sale was observed on June 2020 at 93,497.

- The classic cars product line was observed to have contributed to the highest of sales followed by the Vintage cars and the least being the Trains.

- The product line Sales when observed over deal size it was noted that majority of the sales were by Medium sized deal size followed by the small deal size. However, for classic cars the second largest deal size was large and not small. Apart from classic cars rest categories placed a small amount of large deal size whereas Ships did not place any large deal size as per the data available.

- Analysis on Cancelled orders suggest that only 4 customers have repeatedly cancelled orders. It was observed that they belong to the Spain, Sweden, UK and USA. The customers need to be confronted as to know why they cancelled the orders and a fee has to be charged every time they cancel the order from now on.

- The customers whose status was disputed needs to be verified why were the orders disputed and needs to resolve the issue soon so that they don't go on to cancel the order. Among the orders disputed Euro Shopping Channel had disputes 6 times (Highest) also had cancelled the orders 16 times (Highest). A thorough investigation on this customer needs to be conducted.

- Although the highest ordered among the product line was Classic cars the highest cancelled was that of Ships. Although the Sales of ships was not phenomenal but attracted higher cancels. The company needs to look into this aspect and come forward with a solution to counter such cancellations.

- USA appears to have made the highest sales among all the countries followed by Spain and France. Ireland was the least among the lot.

- Although USA was observed to have made the highest sales among all other countries it is important to know that USA has the highest number of states and cities compared to other countries. But what surprised the analysis is that Madrid city has made the highest sales even more than that of combined sales of San Rafael and NYC (Top 2 sales in USA).

- A pareto chart was plotted to analyze the percentage of sales each country has contributed. It was observed that USA, Spain and France combined contributed to around 58% of the sales meaning rest all (16 countries) combined contributed to only 42% of the total sales.

- Since USA, Spain and France were observed to be the top 3 countries to have contributed higher towards Sales. A city wise pareto chart was observed for these 3 countries and was noted Madrid, San Rafael, NYC and Paris combined had contributed to the 45% of the total sales of the top 3 countries. Special focus to customers from these geographies should be given.

Customer Segmentation using RFM

What is RFM?

RFM stands for Recency, Frequency, and Monetary Value. It's a method used in marketing to analyze customer behavior and segment them into different groups. By understanding these three factors, businesses can target their marketing campaigns more effectively.

- Recency: This refers to how recently a customer has made a purchase. Customers who have purchased recently are generally considered more valuable as they are more likely to buy again soon.
- Frequency: This refers to how often a customer makes purchases. Customers who purchase frequently are also valuable as they are loyal to the brand.
- Monetary Value: This refers to how much money a customer spends on average per purchase. Customers who spend more are considered more valuable.

By combining these three factors, businesses can create a score for each customer. This score can then be used to segment customers into different groups which is mentioned below.

Parameters and Assumptions Made

- For Recency parameter was readily available in the name of DAYS_SINCE_LASTORDER. However, a separate parameter too could have been created since order date too was available.

- For Monetary, parameter in the name of sales was available which was calculated by multiplying the quantity ordered by the price of each item. Since, this was readily available no calculations had to be implemented.

- Frequency was not readily available. For finding out frequency the data was extracted into the KNIME and was grouped by the order ID since the data describes that it is the unique identification number assigned to each order.

- Manual aggregation was performed in which customer name was assigned count since every unique order id placing multiple orders had the same customer name. Count of order date too could have done the job since it was observed multiple orders placed by unique order ID too had same order date. However, in our case customer name was preferred.

- A four-bin approach was preferred in the segmentation. It was binned in such a way that the desired outcome explains that the best customers had maximum frequency (visits), maximum monetary (amount spent on purchase) and a minimum Recency (Days take to revisit).

- Customers who had Cancelled the orders were exempted from the RFM analysis. This was made an assumption since the customers who cancelled got the money refunded.

KNIME Workflow Image

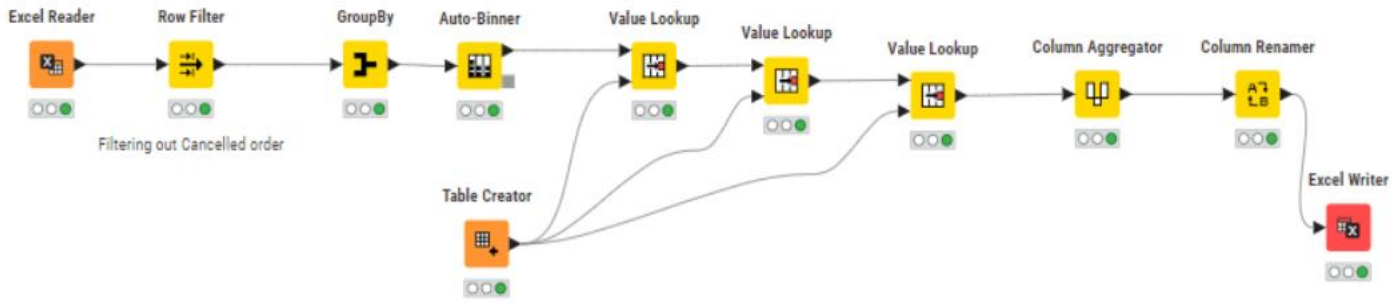

Figure 16 Customer Segmentation using RFM

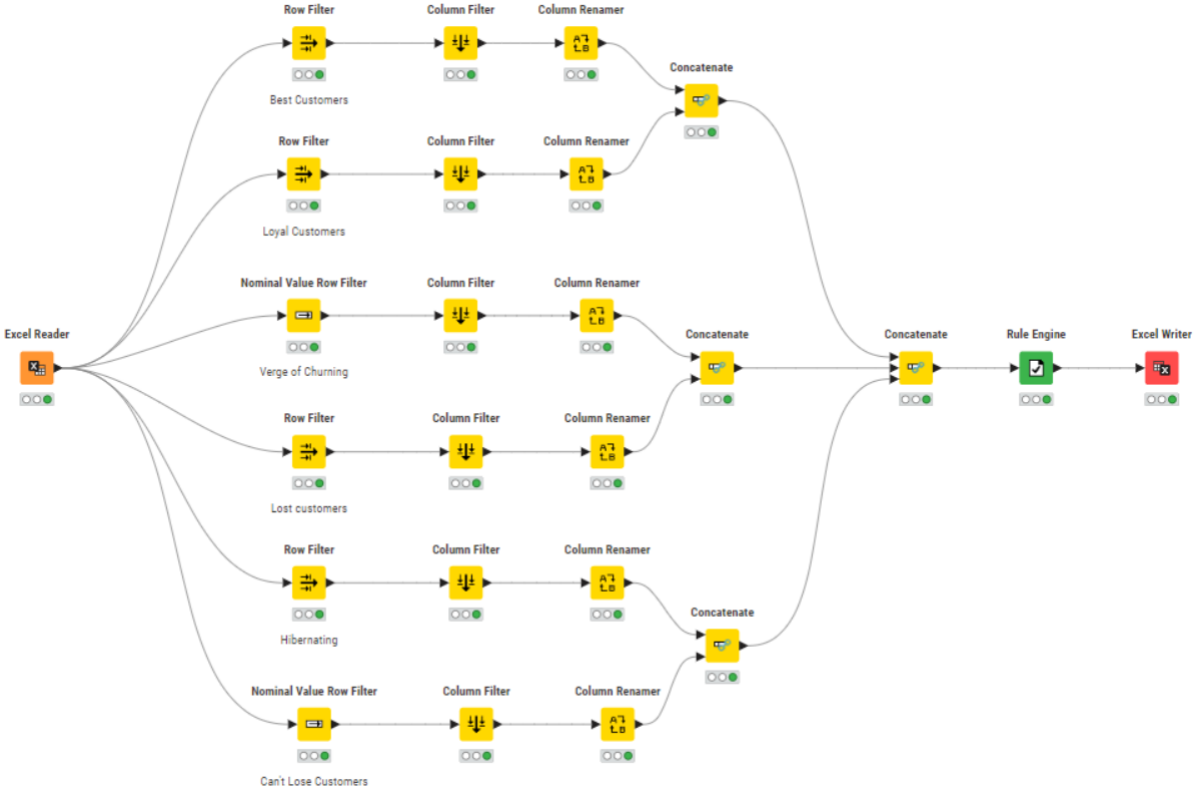

Figure 17 Segment Analysis

Figure 16 represents the KNIME workflow of how the data was imported into the tool and the cancelled customers where filtered, bins created for R, F and M. Further the R, F and were aggregated to RFM and renamed to RFM Score and the segmented data was exported into a new excel.

Figure 17 reads the segmented excel that was created in Figure 16. This was made to further bifurcate the segments as Best/Core customers, Lost customers, Loyal customers, Verge of churning customers, Hibernating and Cannot lose them customers.

Results Output head

Figure 16 output table head:

Note: Only relevant columns such as ORDERNUMBER, DAYS_SINCE_LASTORDER (RECENCY), CUSTOMERNAME(Frequency), SALES(Monetary), RFM Score, contact First and Last Name and Country

ORDERNUMBER	SALES	FREQUENCY	Recency	FIRSTNAME	LASTNAME	COUNTRY	RFM_Score
10100	12133.25	4	1429	Valarie	Young	USA	111
10101	11432.34	4	1573	Roland	Keitel	Germany	111
10102	6864.05	2	1327	Michael	Frick	USA	111
10103	54702	16	878	Jonas	Bergulfsen	Norway	244
10104	44621.96	13	1102	Diego	Freyre	Spain	133
10105	58871.11	15	939	Jytte	Petersen	Denmark	244
10106	56181.32	18	1361	Giovanni	Rovelli	Italy	144
10107	25783.76	8	828	Kwai	Yu	USA	222
10108	55245.02	16	971	Arnold	Cruz	Philippines	244
10109	27398.82	6	1241	Rosa	Hernandez	USA	122

Table 3 Output table

Identified Segments and Inferences

Segments Identified

Best Customers: Bought recently, buy often and spend the most.

Actions taken (If necessary) to revive: Reward them. They can become evangelists and early adopters of new products.

ORDERNUMBER	SALES	Frequency	COUNTRY	FIRSTNAME	LASTNAME	Recency	RFM_Score	Customer Analytics
10273	47760.48	15	Belgium	Catherine	Dewey	426	444	Best Customers
10275	56002.9	18	France	Janine	Labrune	326	444	Best Customers
10280	56078.26	17	Italy	Paolo	Accorti	328	444	Best Customers
10287	67281.01	17	Switzerland	Michael	Holz	440	444	Best Customers
10304	59074.9	17	France	Daniel	Tonini	275	444	Best Customers
10306	57827.61	17	UK	Victoria	Ashworth	421	444	Best Customers
10308	46873.04	16	USA	Steve	Frick	295	444	Best Customers
10310	68943.4	17	Germany	Henriette	Pfalzheim	395	444	Best Customers
10312	63075.08	17	USA	Valarie	Nelson	266	444	Best Customers
10314	60273.94	15	Denmark	Palle	Ibsen	414	444	Best Customers

Table 4 Best/Core Customers

Customers on the Verge of Churning: Some time since they've purchased. Need to bring them back.

Actions taken (If necessary) to revive: Send personalized email or other messages to reconnect. Provide good offers and share valuable resources.

ORDERNUMBER	SALES	Frequency	COUNTRY	FIRSTNAME	LASTNAME	Recency	RFM_Score	Customer Analytics
10107	25783.76	8	USA	Kwai	Yu	828	222	Verge of Churning
10129	32376.29	9	UK	Ann	Brown	820	222	Verge of Churning
10141	31569.43	9	Finland	Kalle	Suominen	1022	222	Verge of Churning
10196	42498.76	8	USA	Leslie	Murphy	735	223	Verge of Churning
10226	25872.22	7	USA	Valarie	Thompson	972	222	Verge of Churning
10244	28327.64	9	Spain	Diego	Freyre	863	222	Verge of Churning
10297	18971.96	7	Ireland	Dean	Cassidy	1021	222	Verge of Churning
10300	27257.79	8	Germany	Roland	Keitel	771	222	Verge of Churning
10327	24078.61	8	Denmark	Jytte	Petersen	748	222	Verge of Churning
10344	20136.86	7	France	Laurence	Lebihan	757	222	Verge of Churning
10353	29343.35	9	USA	Dan	Lewis	947	222	Verge of Churning

Table 5 Verge of Churning customers Table

Lost Churning: Last purchase was long back and very low number of orders. May be lost.

ORDERNUMBER	SALES	Frequency	COUNTRY	FIRSTNAME	LASTNAME	Recency	RFM_Score	Customer Analytics
10100	12133.25	4	USA	Valarie	Young	1429	111	Lost Customers
10101	11432.34	4	Germany	Roland	Keitel	1573	111	Lost Customers
10102	6864.05	2	USA	Michael	Frick	1327	111	Lost Customers
10116	1711.26	1	Belgium	Pascale	Cartrain	2963	111	Lost Customers
10118	4219.2	1	Spain	Eduardo	Saavedra	3389	111	Lost Customers
10123	16560.3	4	France	Carine	Schmitt	1245	111	Lost Customers
10125	9738.18	2	Australia	Peter	Ferguson	1197	111	Lost Customers
10130	7277.35	2	USA	Leslie	Taylor	1615	111	Lost Customers
10137	15146.32	4	France	Paul	Henriot	1195	111	Lost Customers
10144	1637.2	1	Belgium	Pascale	Cartrain	2893	111	Lost Customers

Table 6 Lost Customers

Loyal Customers: Buy on a regular basis. Responsive to promotions. Customers who have come recently and purchase above average and are above average frequent visitors too come under this bracket.

Actions taken (If necessary) to revive: Up-sell higher value products. Engage them. Ask for reviews.

ORDERNUMBER	SALES	Frequency	COUNTRY	FIRSTNAME	LASTNAME	Recency	RFM_Score	Customer Analytics
10263	44130.52	11	USA	Julie	King	350	433	Loyal Customers
10270	41297.14	11	Australia	Adrian	Huxley	356	433	Loyal Customers
10281	44781.35	14	USA	Kyung	Yu	425	433	Loyal Customers
10283	43332.35	14	Canada	Elizabeth	Lincoln	397	433	Loyal Customers
10299	42744.06	11	Finland	Matti	Karttunen	259	433	Loyal Customers
10313	36576.71	11	Canada	Yoshi	Tannamuri	364	433	Loyal Customers
10341	45001.11	10	Austria	Georg	Pipps	208	433	Loyal Customers
10357	42152.11	10	USA	Valarie	Nelson	219	433	Loyal Customers
10361	37905.15	14	Australia	Adrian	Huxley	186	433	Loyal Customers
10371	36124.27	12	USA	Valarie	Nelson	326	433	Loyal Customers
10391	35160.25	10	Australia	Anna	O'Hara	131	433	Loyal Customers
10425	43784.69	13	France	Janine	Labrune	150	433	Loyal Customers

Table 7 Loyal Customers

Can't lose Customers: Customers who Used to purchase frequently but haven't returned for a long time.

Actions taken (If necessary) to revive: Win them back. Talk to them. Make them special offers. Make them feel valuable.

ORDERNUMBER	SALES	Frequency	COUNTRY	FIRSTNAME	LASTNAME	Recency	RFM_Score	Customer Analytics
10104	44621.96	13	Spain	Diego	Freyre	1102	133	Cannot lose Customers
10106	56181.32	18	Italy	Giovanni	Rovelli	1361	144	Cannot lose Customers
10110	51017.92	16	UK	Victoria	Ashworth	1307	144	Cannot lose Customers
10124	33847.62	13	USA	Sue	King	1296	133	Cannot lose Customers
10148	47191.76	14	Australia	Anna	O'Hara	1059	134	Cannot lose Customers
10149	34100.03	11	USA	Sue	Taylor	1085	133	Cannot lose Customers
10177	34311.35	11	Spain	Jesus	Fernandez	1525	133	Cannot lose Customers
10328	41696.69	14	Italy	Giovanni	Rovelli	1387	133	Cannot lose Customers

Table 8 Cannot lose Customers

Inferences

- We can identify customers with high RFM scores, indicating frequent purchases, recent activity, and high spending. These are likely your most valuable customers with the highest CLV.

- Customers with low recency scores (haven't purchased recently) are at risk of churning. We can identify these customers and develop win-back campaigns with special offers or loyalty programs.

- RFM helps us segment customers into distinct groups based on their behavior. This allows us to tailor marketing campaigns to each segment. For instance, loyal high-spending customers might receive exclusive discounts, while new customers might get welcome offers.

- Analyze customer response to specific marketing campaigns segmented by RFM. This helps in understanding which campaigns resonate with different customer groups and optimize future efforts.

- By analyzing purchase frequency of specific products across customer segments, one can identify which products are popular with different customer groups. This can inform product development and marketing strategies.

Overall, RFM analysis provides valuable insights into customer behavior, allowing us to:

- Improve customer retention: By identifying at-risk customers and implementing targeted win-back campaigns.

- Boost customer engagement: By tailoring marketing messages and promotions to each customer segment based on their RFM profile.

- Increase customer lifetime value: By focusing efforts on high-value customers and encouraging repeat purchases

Part – B

Problem Statement

A grocery store shared the transactional data. The goal is to conduct a thorough analysis of Point of Sale (POS) data, identify the most commonly occurring sets of items in the customer orders, and provide recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.

Exploratory Analysis

The data consists of three variable Date, Order ID and Product. There are 20641 records of data. The Date was object data type which was converted to date format for further analysis.

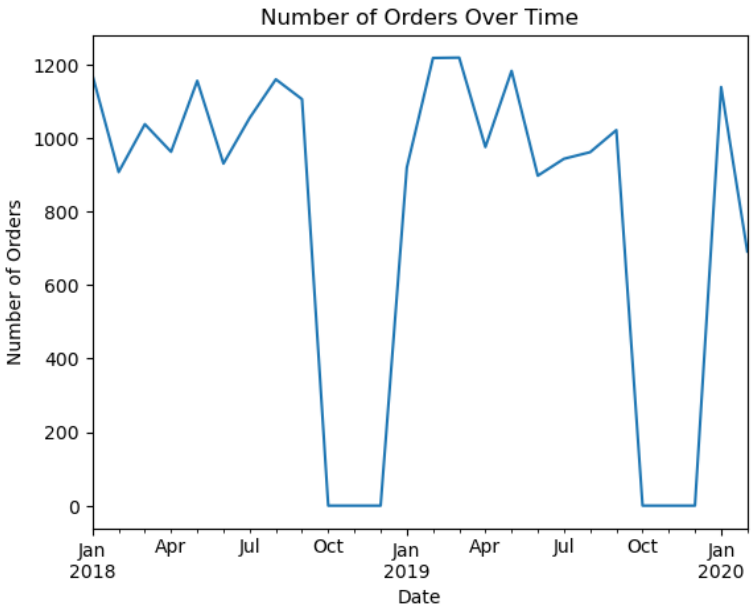

Figure 18 Monthly orders trend

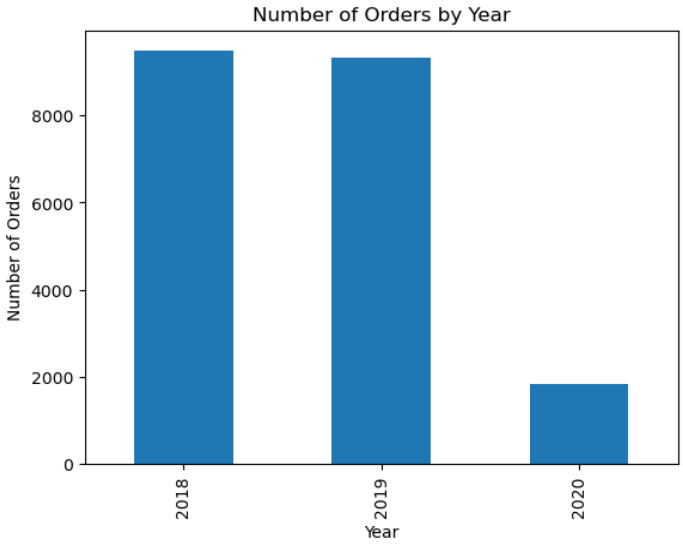

Figure 19 Yearly Orders

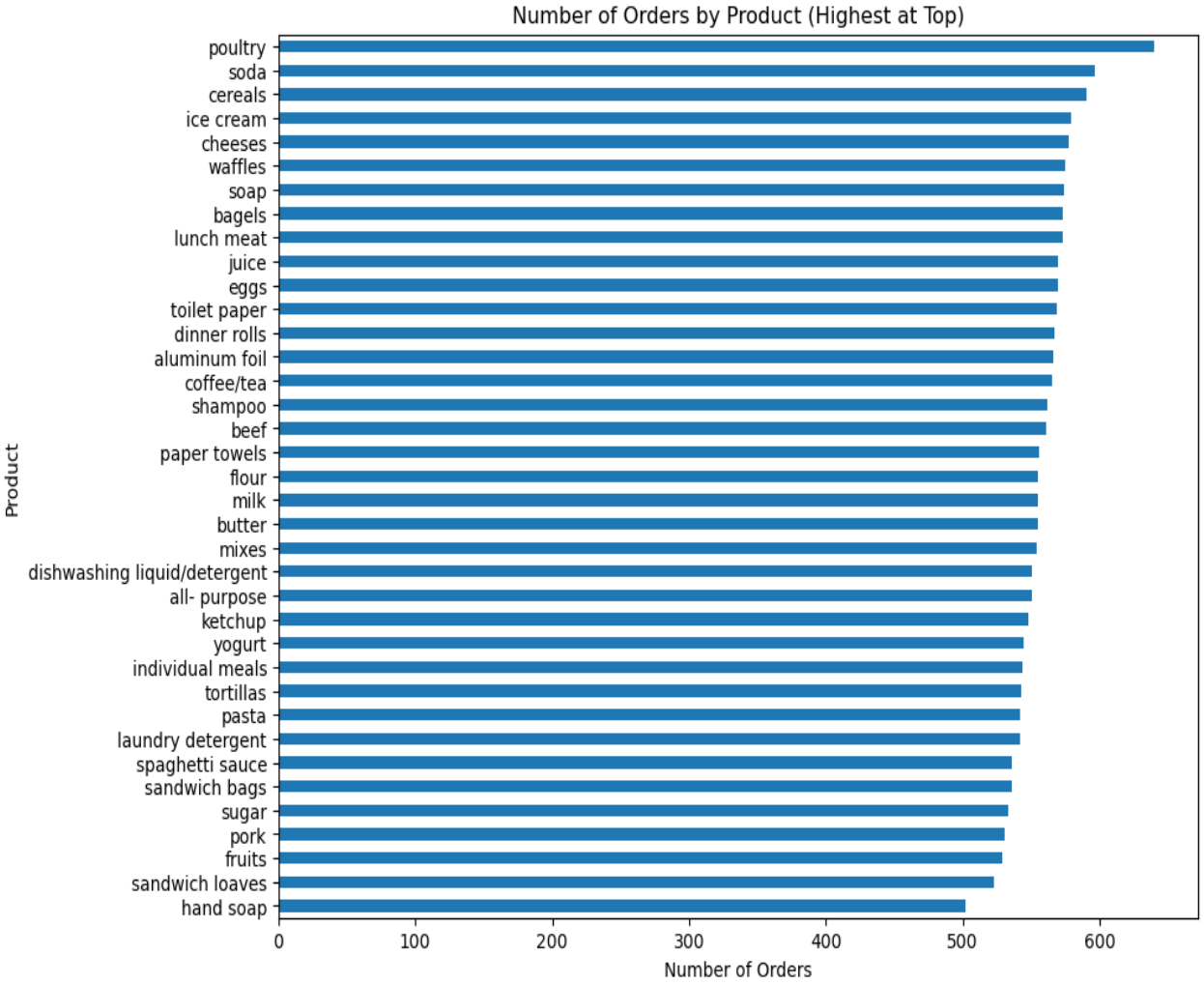

Figure 20 Total Orders per product

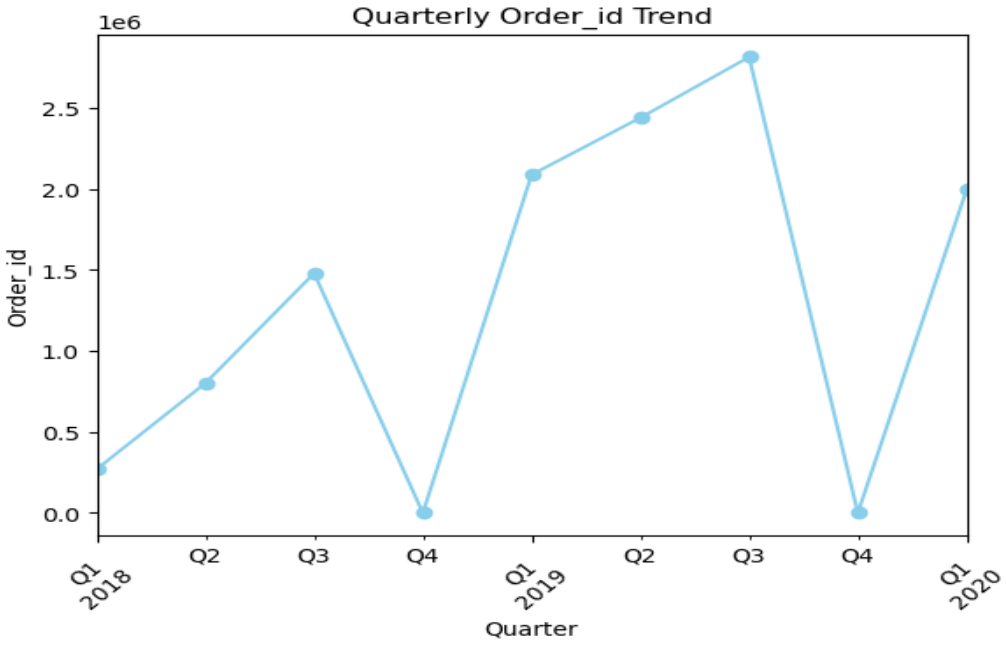

Figure 21 Quarterly order trends

Products that have been seen a decline in 2019 when compared to 2018

1. Beef
2. Cereals
3. flour
4. fruits
5. hand wash
6. ketchup
7. mixes
8. shampoo

Products that have been ordered more in 2019 when compared to 2018

1. Dishwashing liquid/Detergent
2. Laundry detergent
3. Milk
4. Paper towels
5. Soda
6. Yogurt

Note: The proof in the form of plots for Product that have been ordered more and that have seen a decline have not been mentioned in the report as the simple plots in large number could look cluttered. However, the information above is sufficient to convey the analysis. For the plots please refer the jupyter notebook in the format .ipynb

Summary and Inferences

- Although the data seem to describe that it has 3 years of data but has only up-to 2 months of data for the year 2020. Any trends observed will only be limited to that of two years of data which is 2018 and 2019.

- It was observed that for both the years 2018 and 2019, the month of October, November and December the sales were NIL. This could be the company's shutdown owing to many possible reasons such as festivals, maintenance, etc. This is of least concern for any decisions but can be noted as an observation.

- The total orders in the year 2018 and 2019 when analyzed the bar graph indicates no dip in orders instead were almost in the same step. As for the year 2020 it could be difficult to interpret since the data was available only up-to February.

- The number of orders placed product wise were analyzed and poultry was the highest among the lot as shown in the Figure 20. Hand soap was the least ordered product.

- The quarterly orders were analyzed and a clear increasing trend was observed quarter by quarter up-to the 4th quarter. The 4th quarter was NIL in the year 2018 and 2019.

- The 3rd quarter was observed to have seen higher order in both the years. Important to have stock of products since this demands more products owing to increase in orders. The higher orders in the 3rd quarter could also be a result of company taking shutdown in the 4th quarter.

Market Basket Analysis

Association rules are a fundamental technique in data mining and are particularly useful in market basket analysis. They help identify interesting relationships, patterns, and associations among a set of items in large datasets. These rules are used to discover how the presence of one item in a transaction influences the presence of another item.

Concept of Association Rules

1. Itemset: A collection of one or more items. In our example, the grocery store, an itemset might be {yogurt, pork, soda}.

2. Support: This measures how frequently an itemset appears in the dataset. It is calculated as the proportion of transactions that contain the itemset. For an itemset X

$$\text{Support}(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}$$

Equation 1 Support

3. Confidence: This measures how often items in Y appear in transactions that contain X. It is the conditional probability that a transaction containing X also contains Y

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

Equation 2 Confidence

4. Lift: This measures how much more likely Y is to be bought when X is bought compared to the likelihood of buying Y without X. It is the ratio of the observed support to that expected if X and Y were independent:

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) \times \text{Support}(Y)}$$

Equation 3 Lift

A small example of how is Support, Confidence and Lift associated to our case

In our case the Support, confidence and lift play a significant role.

Let us consider an example of products such as yogurt, pork, soda and aluminum foil. The Support identified the probability of yogurt alone being purchased but the confidence is what tells us the association of two products bought together, consider yogurt and pork cooked forms a dish and this leading to often purchase history of yogurt and pork.

When a customer adds yogurt, the confidence is what tells how often he/she would be adding pork in the basket. It works on the basis of conditional probability.

The lift formula written out would look something like: Interpreted as: How much our confidence has increased that pork will be purchased given that yogurt was purchased. Greater lifts indicate stronger associations.

KNIME Workflow

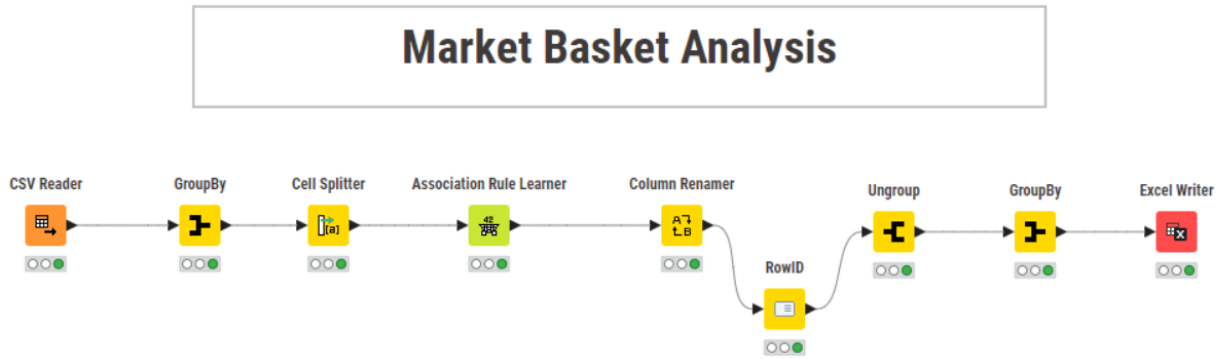

Figure 22 KNIME workflow

Threshold values of Support and Confidence

Support:

- Represents how frequently a set of items appears together in transactions.
- Calculated as the proportion of transactions containing the itemset.
- Higher support thresholds indicate more common purchases.
- We might set a lower support for exploratory analysis to uncover unexpected patterns, then raise it to focus on more frequent ones.

Confidence:

- Indicates the likelihood of finding item B in a transaction, given that item A is already present (written as A -> B).
- Calculated as the support for A and B divided by the support for A alone.
- Higher confidence thresholds suggest a stronger association between the items.

Associations Identified

Note: The data in the table below is sorted in the descending order of lift.

Support	Confidence	Lift	Recommended Item	In Basket Items
0.099209833	0.579487179	1.489923019	dinner rolls	spaghetti sauce, poultry
0.095697981	0.545	1.446981352	juice	poultry, aluminum foil
0.090430202	0.528205128	1.422282839	pasta	dinner rolls, soda
0.094820018	0.553846154	1.420790021	eggs	dinner rolls, soda
0.092186128	0.552631579	1.414488468	soda	eggs, soap
0.091308165	0.525252525	1.41433245	pasta	eggs, dinner rolls
0.095697981	0.542288557	1.410197869	aluminum foil	poultry, juice
0.093064091	0.540816327	1.406369397	yogurt	juice, aluminum foil
0.094820018	0.545454545	1.396118488	soda	eggs, dinner rolls
0.093064091	0.524752475	1.393223938	juice	yogurt, aluminum foil
0.091308165	0.538860104	1.385466497	dinner rolls	eggs, pasta
0.090430202	0.533678756	1.372144703	dinner rolls	pasta, soda
0.093064091	0.527363184	1.371385084	aluminum foil	yogurt, juice
0.090430202	0.515	1.370525701	mixes	poultry, aluminum foil
0.099209833	0.576530612	1.368059099	poultry	dinner rolls, spaghetti sauce
0.099209833	0.509009009	1.364144144	spaghetti sauce	dinner rolls, poultry
0.090430202	0.536458333	1.360859781	waffles	ice cream, soda
0.091308165	0.527918782	1.354278136	eggs	poultry, soda
0.091308165	0.527918782	1.354278136	eggs	dinner rolls, pasta
0.092186128	0.527638191	1.353558332	eggs	soda, soap
0.094820018	0.507042254	1.349348427	mixes	dishwashing liquid/detergent, poultry
0.094820018	0.52173913	1.344481605	dishwashing liquid/detergent	poultry, mixes
0.090430202	0.533678756	1.338898907	ice cream	cheeses, aluminum foil
0.090430202	0.52284264	1.338242172	soda	dinner rolls, pasta
0.090430202	0.52284264	1.338242172	soda	ice cream, waffles
0.091308165	0.562162162	1.333963964	poultry	dinner rolls, lunch meat
0.091308165	0.517412935	1.330323552	dinner rolls	eggs, poultry
0.091308165	0.517412935	1.324344569	soda	eggs, poultry
0.090430202	0.556756757	1.321137387	poultry	dinner rolls, mixes
0.095697981	0.556122449	1.319632228	poultry	juice, aluminum foil
0.090430202	0.504901961	1.312975647	aluminum foil	poultry, ice cream
0.090430202	0.504901961	1.312975647	aluminum foil	ice cream, cheeses
0.090430202	0.512437811	1.311610487	cheeses	ice cream, aluminum foil
0.092186128	0.509708738	1.310515242	dinner rolls	poultry, cereals
0.091308165	0.509803922	1.304868914	cheeses	poultry, ice cream
0.094820018	0.507042254	1.303659428	dinner rolls	eggs, soda
0.090430202	0.515	1.292037445	ice cream	poultry, aluminum foil
0.090430202	0.5	1.288461538	dishwashing liquid/detergent	poultry, cereals

0.090430202	0.50990099	1.279244995	ice cream	waffles, soda
0.092186128	0.538461538	1.277724359	poultry	dinner rolls, cereals
0.091308165	0.504854369	1.266583978	ice cream	poultry, cheeses
0.090430202	0.533678756	1.266375216	poultry	dishwashing liquid/detergent, cereals
0.094820018	0.526829268	1.250121951	poultry	dishwashing liquid/detergent, mixes
0.090430202	0.525510204	1.246991922	poultry	mixes, aluminum foil
0.091308165	0.525252525	1.246380471	poultry	eggs, dinner rolls
0.090430202	0.512437811	1.215972222	poultry	ice cream, aluminum foil
0.091308165	0.509803922	1.209722222	poultry	ice cream, cheeses
0.194907814	0.501128668	1.189136569	poultry	dinner rolls

Table 9 Associations table

Support, Confidence and Lift Calculated

The threshold values for support and confidence in association rule mining are often determined through a trial-and-error process. There is no pre-defined way where one can identify the optimal values. The nature of solving problem only goes by trial-and-error approach.

The threshold value for Support and Confidence from where we could analyse was **0.09 and 0.5** respectively.

It is important to note than any value greater 0.09 for support fetched no results and similarly any value greater than 0.5 of confidence did not fetch any records. Hence, a **support** of **0.09** and a **confidence** of **0.5** was identified as the optimal value for our analysis.

Since Lift values are often a result of calculations post identifying Support and Confidence. The Lift with highest values is how it is sorted and providing combos and lucrative offers could further promote purchasing the recommended products.

Recommendations/Suggestions of possible combos

Top 5 Associations Identified

Support: 0.099210

Confidence: 0.579487

Lift: 1.489923

Recommended Item: Dinner Rolls

In Basket Items: Spaghetti Sauce, Poultry

Association Rule: High lift (1.489923) indicates a strong association between Spaghetti Sauce, Poultry, and Dinner Rolls.

Offer: Buy one bottle of Spaghetti Sauce and one pack of Poultry, get a pack of Dinner Rolls for free.

Rationale: This combo leverages the strong association and encourages customers to buy all three items together.

Support: 0.095697

Confidence: 0.545

Lift: 1.446981

Recommended Item: Juice

In Basket Items: Poultry, Aluminum foil

Association Rule: High lift (1.4469) for poultry and aluminum foil with juice.

Offer: Buy one pack of poultry and one pack of aluminum foil, get a bottle of juice at a 50% discount.

Rationale: This offer encourages customers to purchase these frequently bought together items, promoting cross-category sales.

Support: 0.090430

Confidence: 0.528205

Lift: 1.422282

Recommended Item: Pasta

In Basket Items: Dinner rolls, Soda

Association Rule: High lift (1.4223) between dinner rolls, soda, and pasta.

Offer: Buy two packs of dinner rolls and one soda, get a pack of pasta for free.

Rationale: This offer promotes the sale of dinner rolls and soda while incentivizing the purchase of pasta.

Support: 0.09482

Confidence: 0.55384

Lift: 1.42079

Recommended Item: Eggs

In Basket Items: Dinner rolls, Soda

Association Rule: High lift (1.4208) for dinner rolls and soda with eggs.

Offer: Buy one pack of dinner rolls and one soda, get a dozen eggs at a 30% discount.

Rationale: This combo encourages the purchase of breakfast items together, driving up the overall basket value.

Support: 0.0921861

Confidence: 0.5526315

Lift: 1.414483

Recommended Item: Soda

In Basket Items: Eggs, Soap

Association Rule: High lift (1.4145) for eggs and soap with soda.

Offer: Buy one dozen eggs and one bar of soap, get a can of soda for free.

Rationale: This offer encourages customers to buy household essentials together with a refreshing drink, making the combo attractive for customers looking to stock up on basics.

Implementation Strategy

Prominent In-Store Display:

- Place the items in these combos near each other in the store.
- Use end-cap displays to feature the combos prominently.

Digital Marketing Campaigns:

- Send targeted emails and social media posts highlighting these combos and offers.
- Use in-app notifications to inform customers of these deals.

In-Store Promotions:

- Use signage and posters to advertise these combos and discount offers.
- Train staff to inform customers about these offers and encourage them to take advantage.

Monitor and Adjust:

- Track the sales data to evaluate the effectiveness of these offers.
- Collect customer feedback to understand preferences and adjust the offers accordingly.