MRA PROJECT - 2

Harikrishnan M
PGP-DSBA

AGENDA

Problem Statement

Exploratory Data Analysis

Use of Market Basket Analysis - Association Rules

Associations Identified

Inferences

Problem Statement

A grocery store shared the transactional data with you. Your job 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 DATA ANALYSIS

DATA DICTIONARY

Sno	Column	Description
1	Date	
2	Order_id	
3	Product	

Exploratory Data Analysis

INFO

1. Dataset is having 3 variables out of which 1 is date field, 1 is numerical and remaining one is categorical.
2. There is no missing values and duplicate values found.
3. Total Number of observations are 20641

```
<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  datetime64[ns]
 1   Order_id  20641 non-null  int64
 2   Product   20641 non-null  object
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 483.9+ KB
```

Data Shape

```
(20641, 3)
```

Data Summary

	count	mean	std	min	25%	50%	75%	max
Order_id	20641.0	575.986289	328.557078	1.0	292.0	581.0	862.0	1139.0

Exploratory Data Analysis

Top 5 Records from Dataset

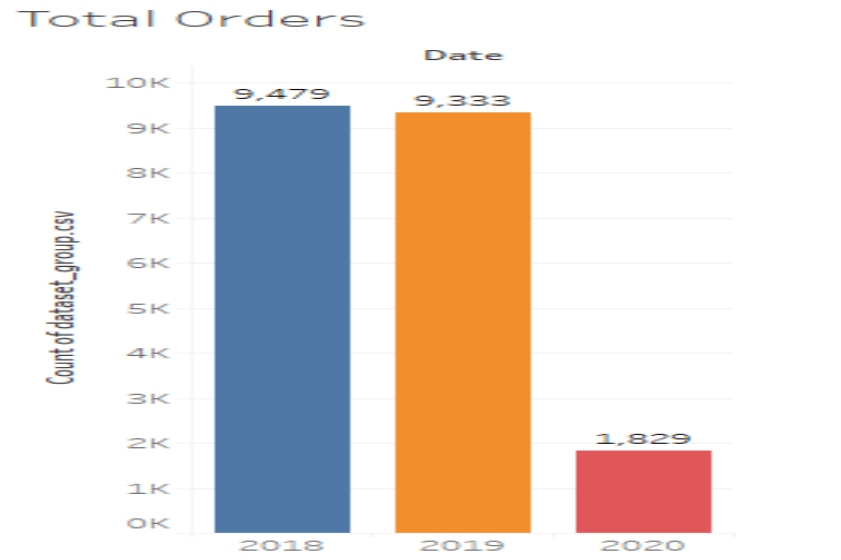
	Date	Order_id	Product
0	01-01-2018	1	yogurt
1	01-01-2018	1	pork
2	01-01-2018	1	sandwich bags
3	01-01-2018	1	lunch meat
4	01-01-2018	1	all- purpose

Last 5 Records from Dataset

	Date	Order_id	Product
20636	2020-02-25	1138	soda
20637	2020-02-25	1138	paper towels
20638	2020-02-26	1139	soda
20639	2020-02-26	1139	laundry detergent
20640	2020-02-26	1139	shampoo

Total number of orders Vs Total number of Products

YEAR	Num_of_Orders	Num_of_Products
2018	533	9479
2019	507	9333
2020	99	1829

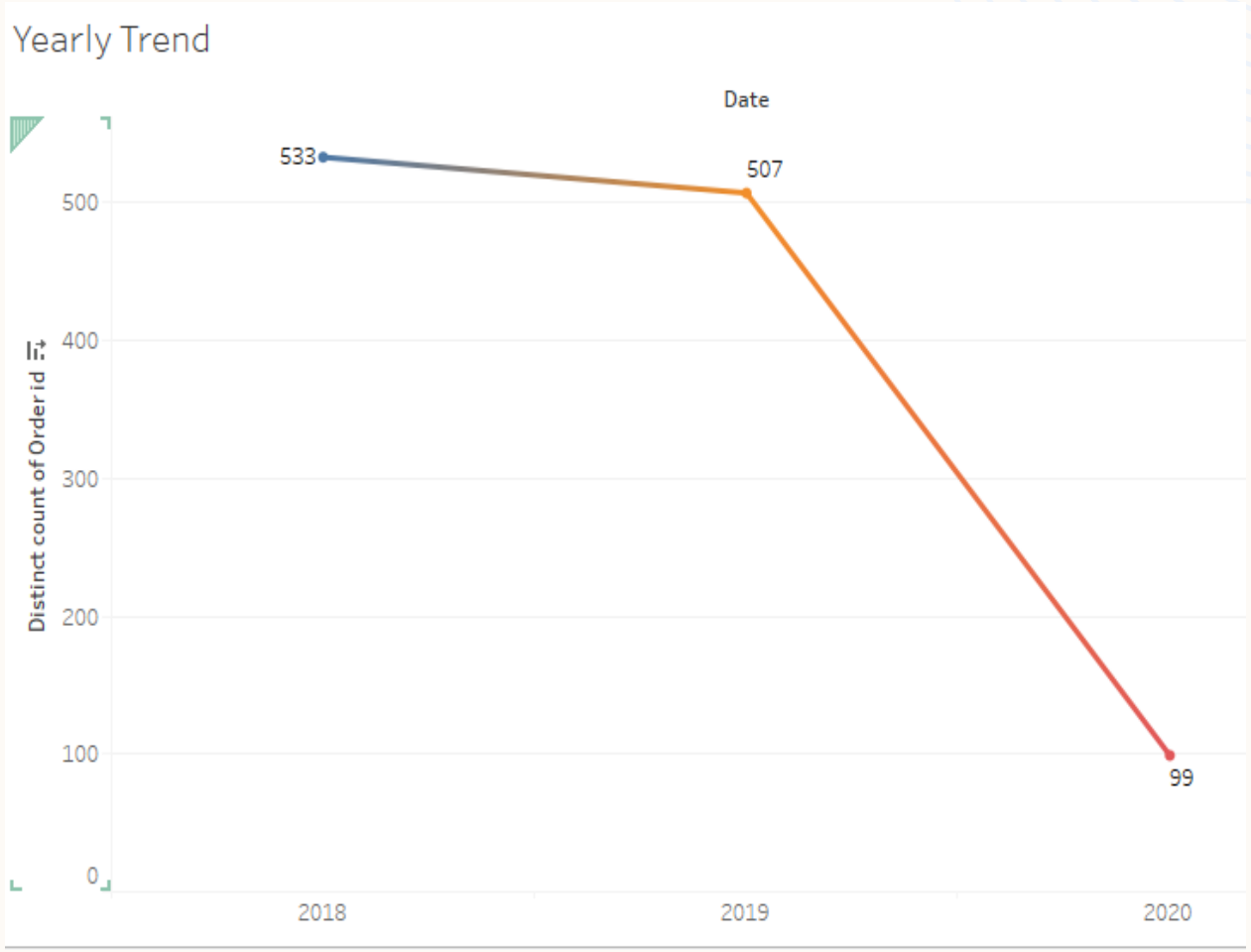


Yearly Trend

1. Kindly note that the transactions are full years of 2018, 2019 and only first 2 months of 2020.

2. In 2018, the number of the orders are highest (533) amongst the consolidated data and 2020 is the lowest orders (99) but having only months of data.

3. 2019 has the order count of 507 which is slightly lower than 2018



Exploratory Data Analysis

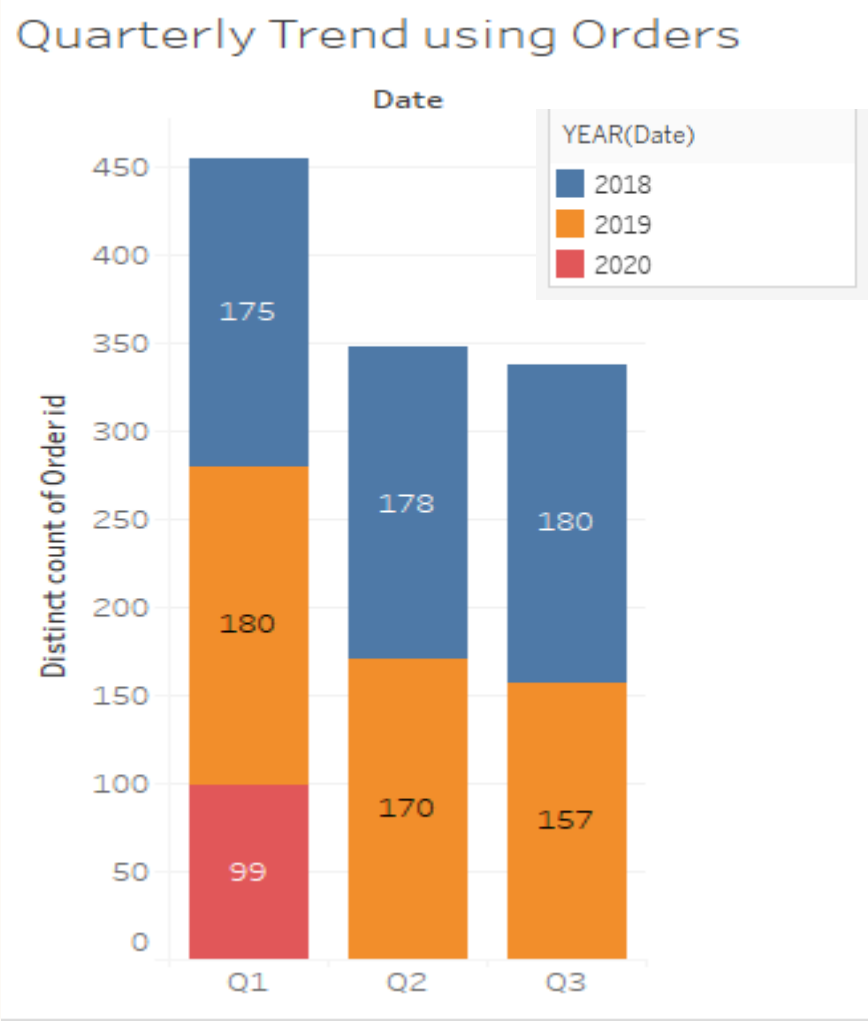
Quarterly Trend

1. There seems to dip in Q2 every year from the high of Q1.

2. In 2018, 3rd Quarter the number of the orders are highest (180) amongst the consolidated data.

3. 2020, 1st Quarter has the order count of 99 since it has only 2 months of data

4. We notice that there are no transactions for Q4



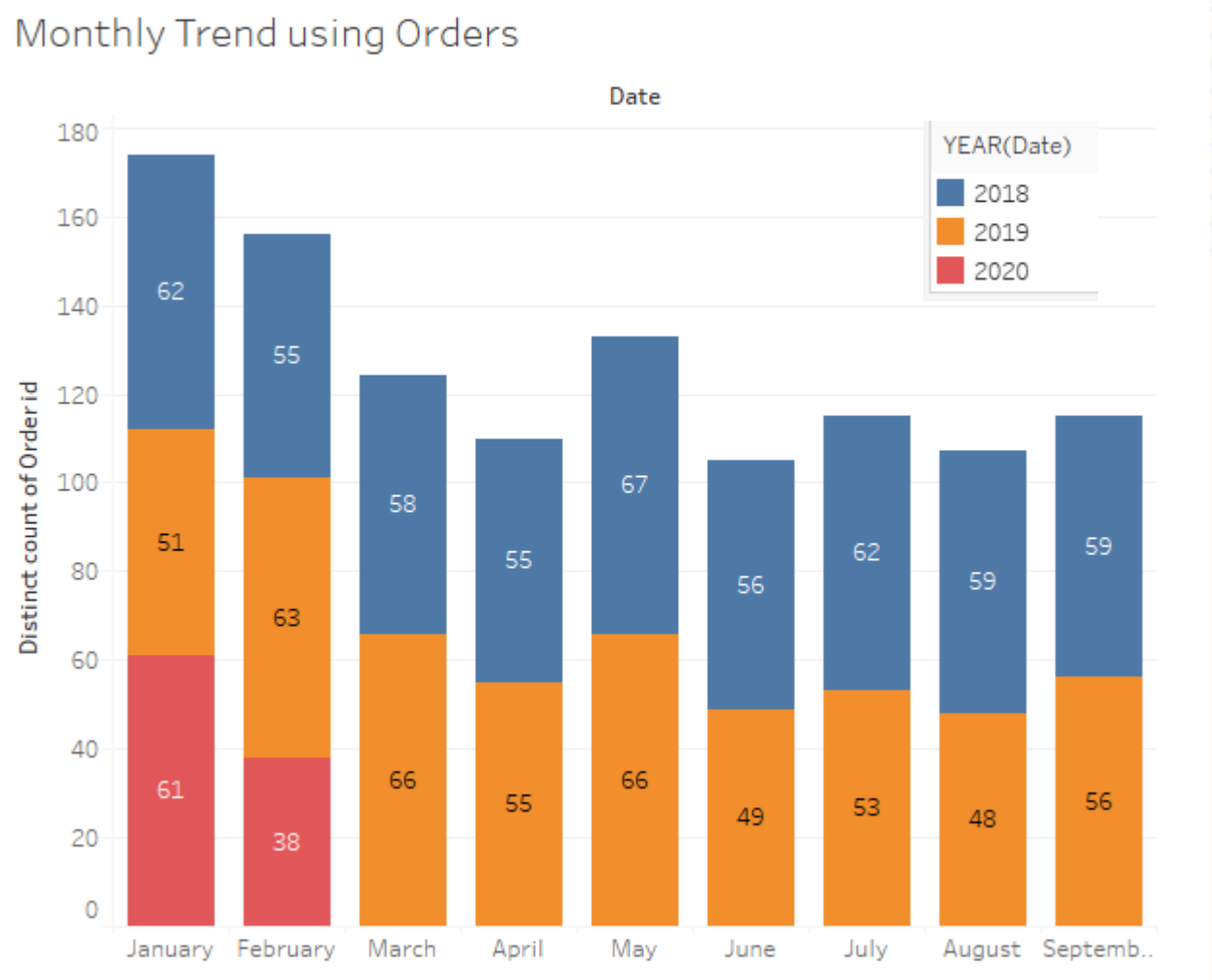
Exploratory Data Analysis

Monthly Trend

1. We notice that there are no transactions for October, November and December

2. January and February having more orders when compared to remaining months.

3. June has least number of orders

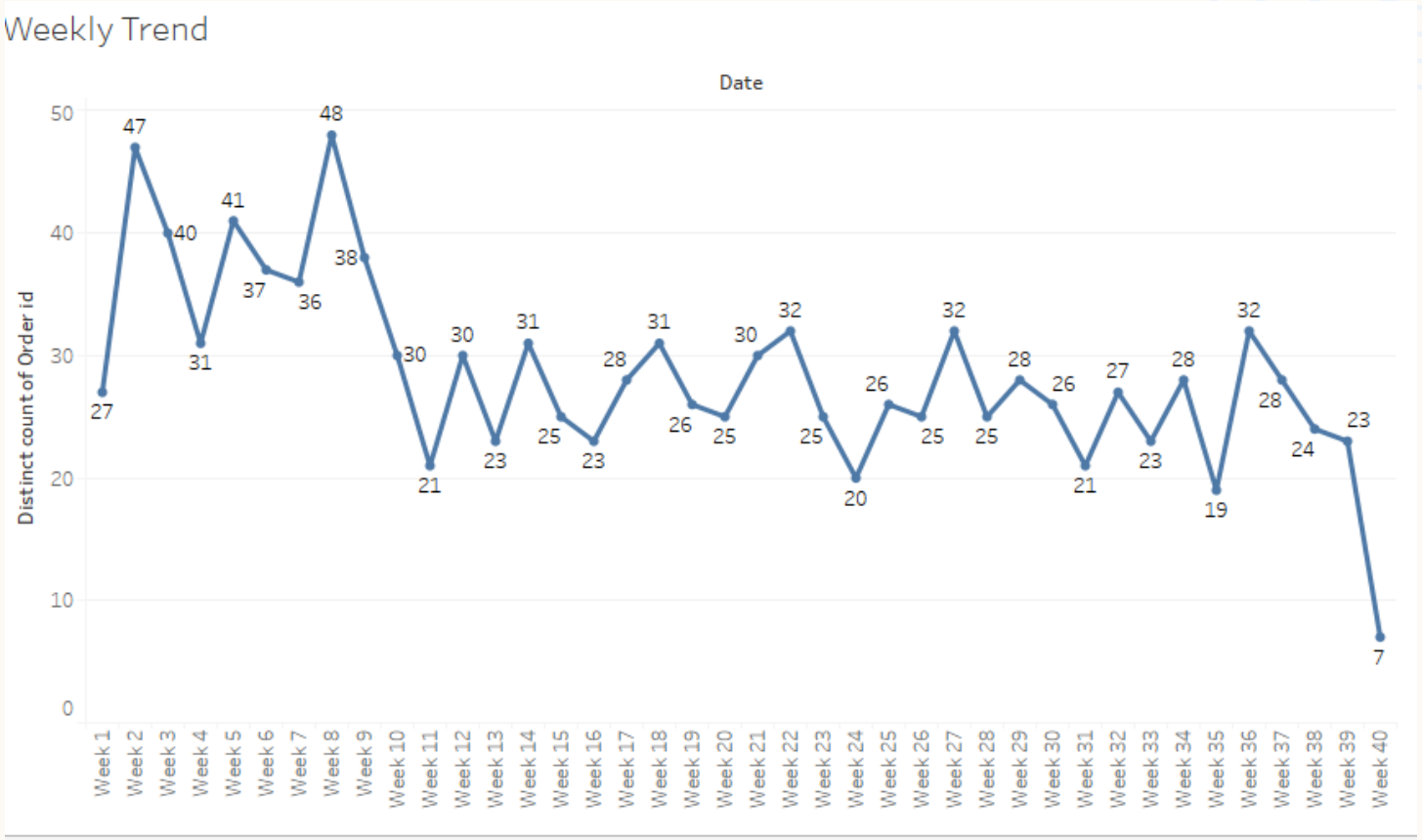


Weekly Trend

1. We notice that there are no transactions for October, November and December

2. Week 2 and Week 8 having more orders

3. Week 40 is having low order counts

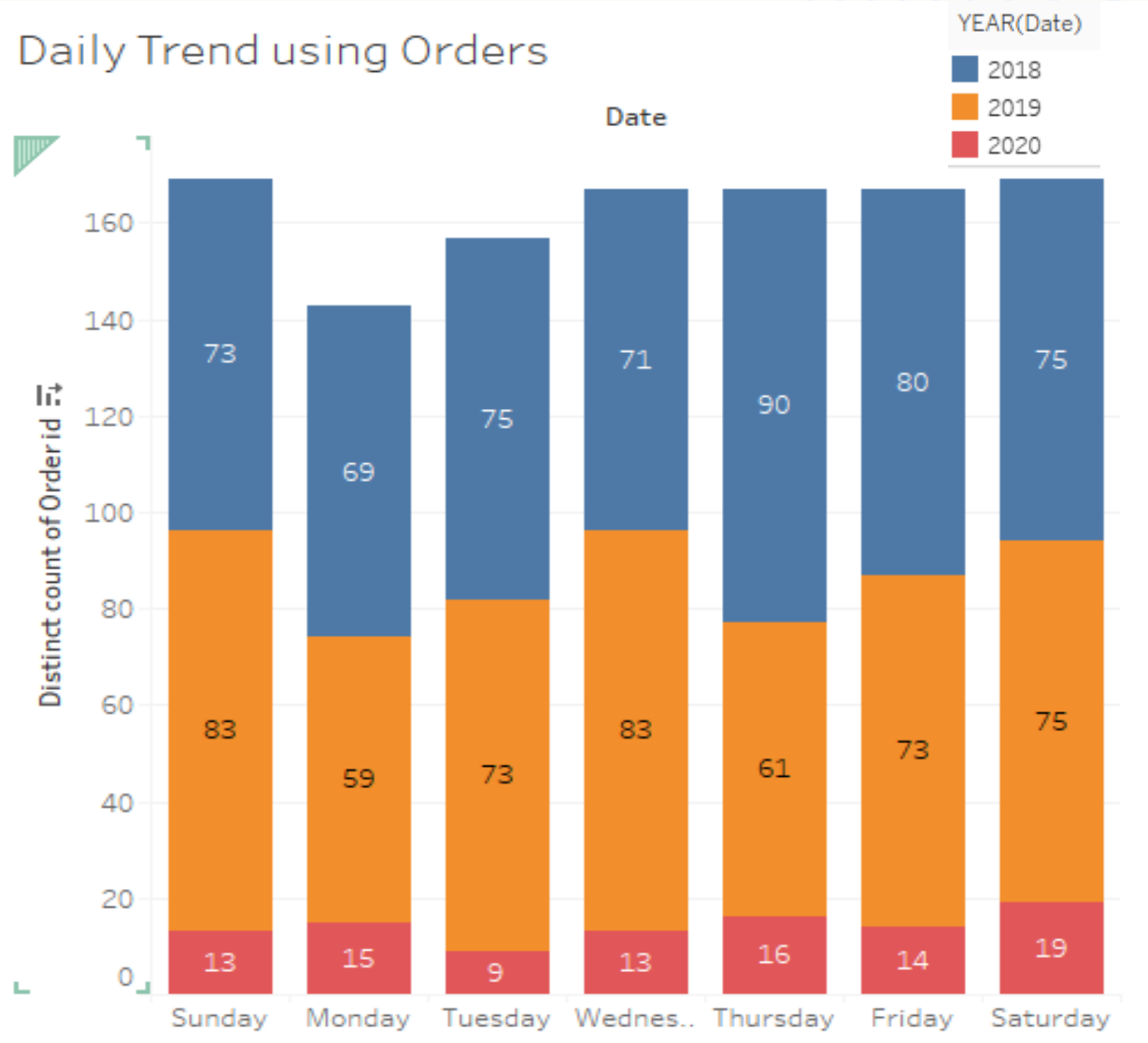


Transactions per Day of the Week

1. People are more interested to order on Sunday.

2. Monday has the lowest transactions

3. All other days have almost same transactions per day



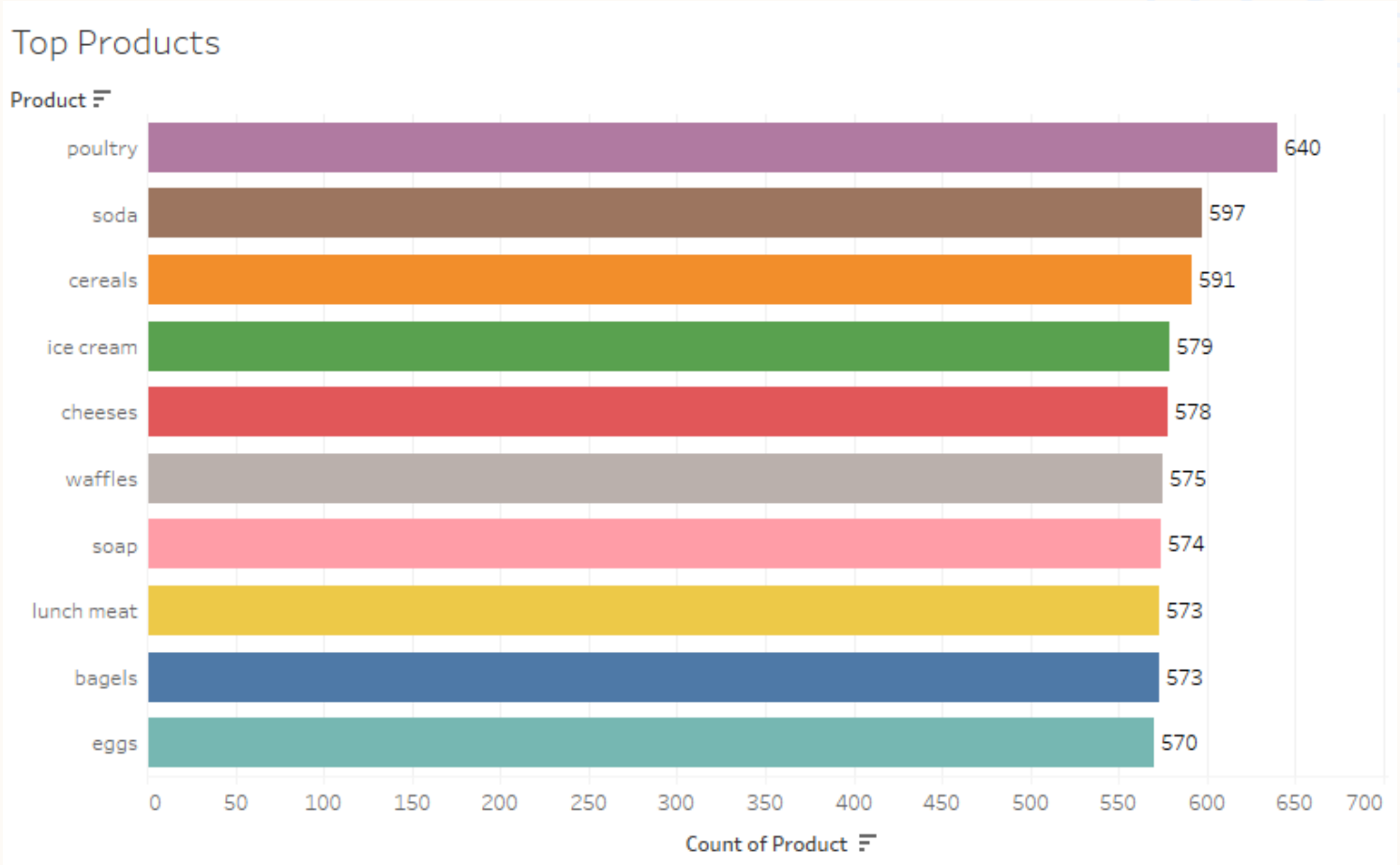
Exploratory Data Analysis

Most ordered Products – Top 10

1. Poultry is the most ordered (640) product amount the products.

2. Soda (597) and Cereals (591) are placed 2nd and 3rd in all the products.

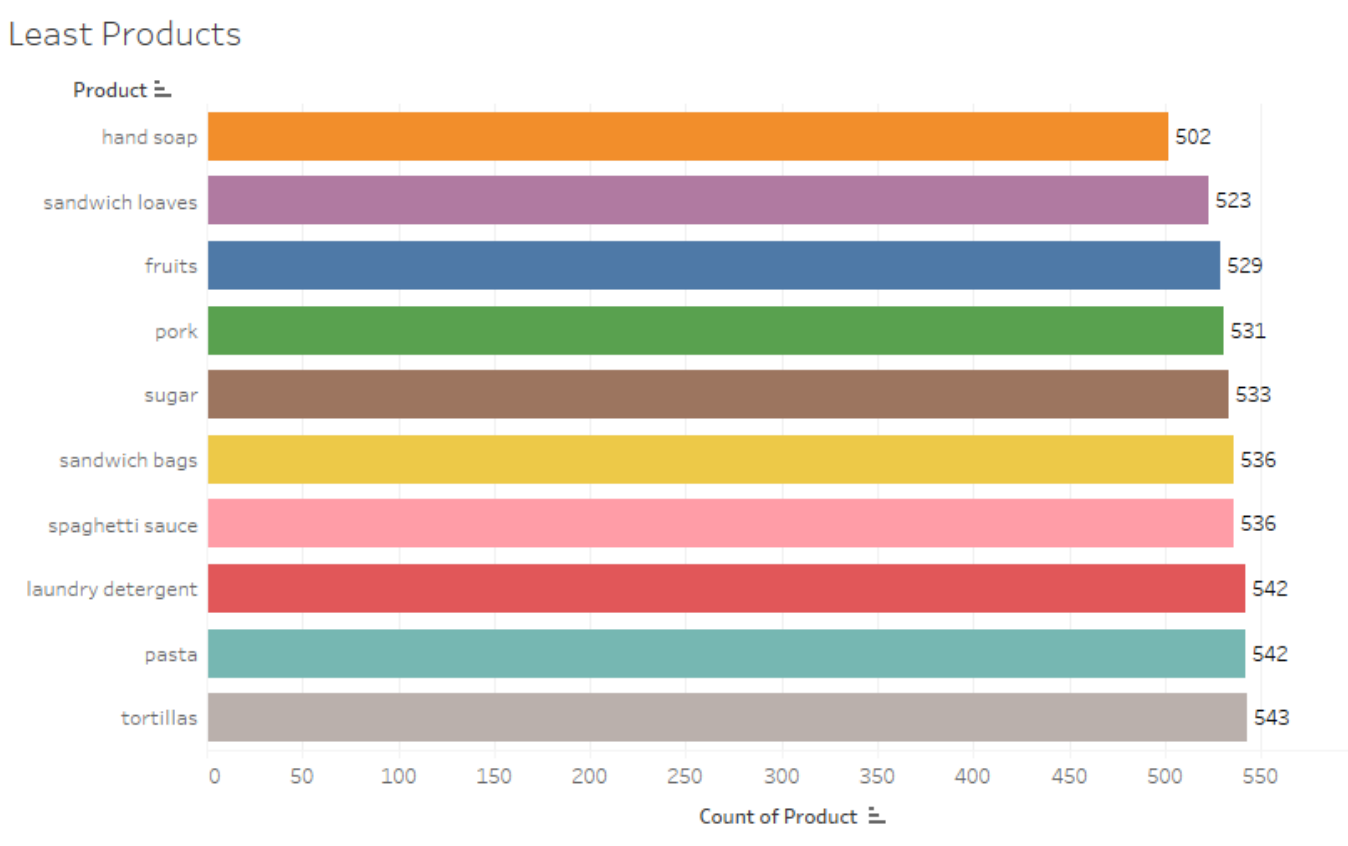
3. Ice cream, cheeses, waffles, soap and lunch, bagles and eggs are ordered as same ratio.



Least ordered Products – Top 10

1. Hand soap is the least ordered (502) product amount the products.

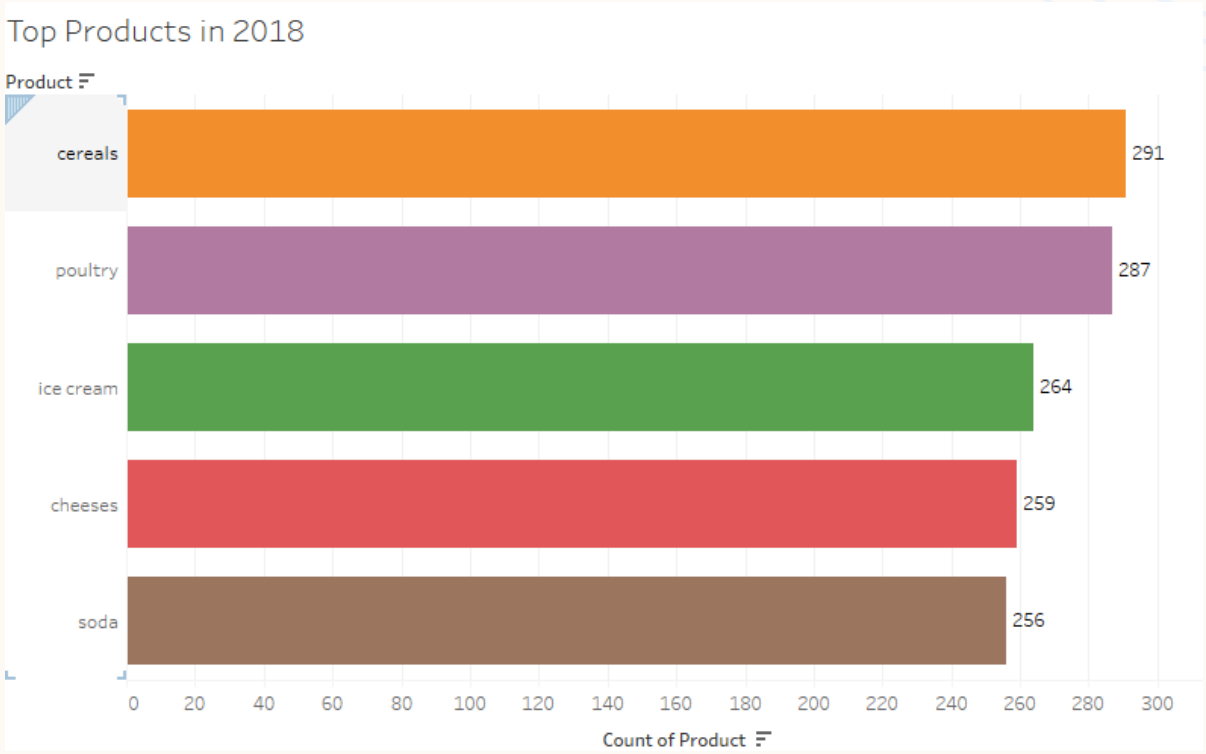
2. Sandwich leaves (523) and Fruits (529) are placed 2nd and 3rd in all the products.



Most ordered Products – 2018

1. Cereals is the most ordered (291) product amount the products.

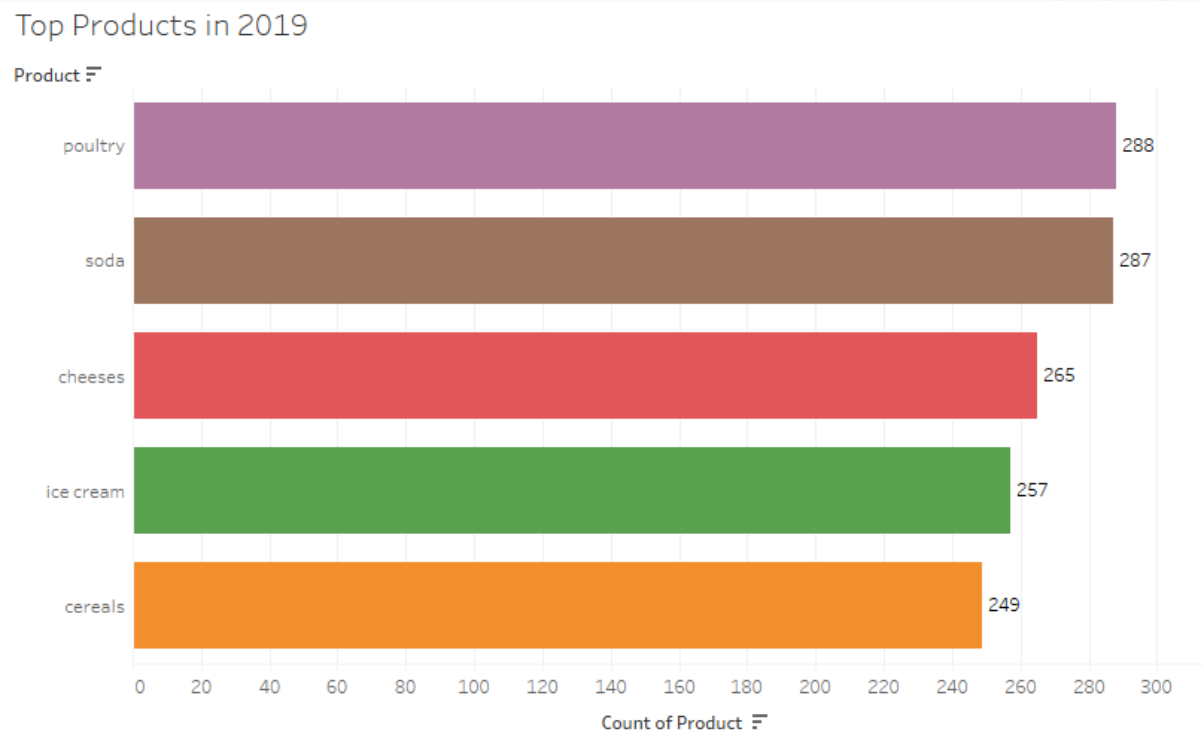
2. Poultry (597) and Flour (591) are placed 2nd and 3rd in all the products.



Most ordered Products – 2019

1. Poultry is the most ordered (288) product amount the products.

2. Sod (287) and Cheeses (265) are placed 2nd and 3rd in all the products.

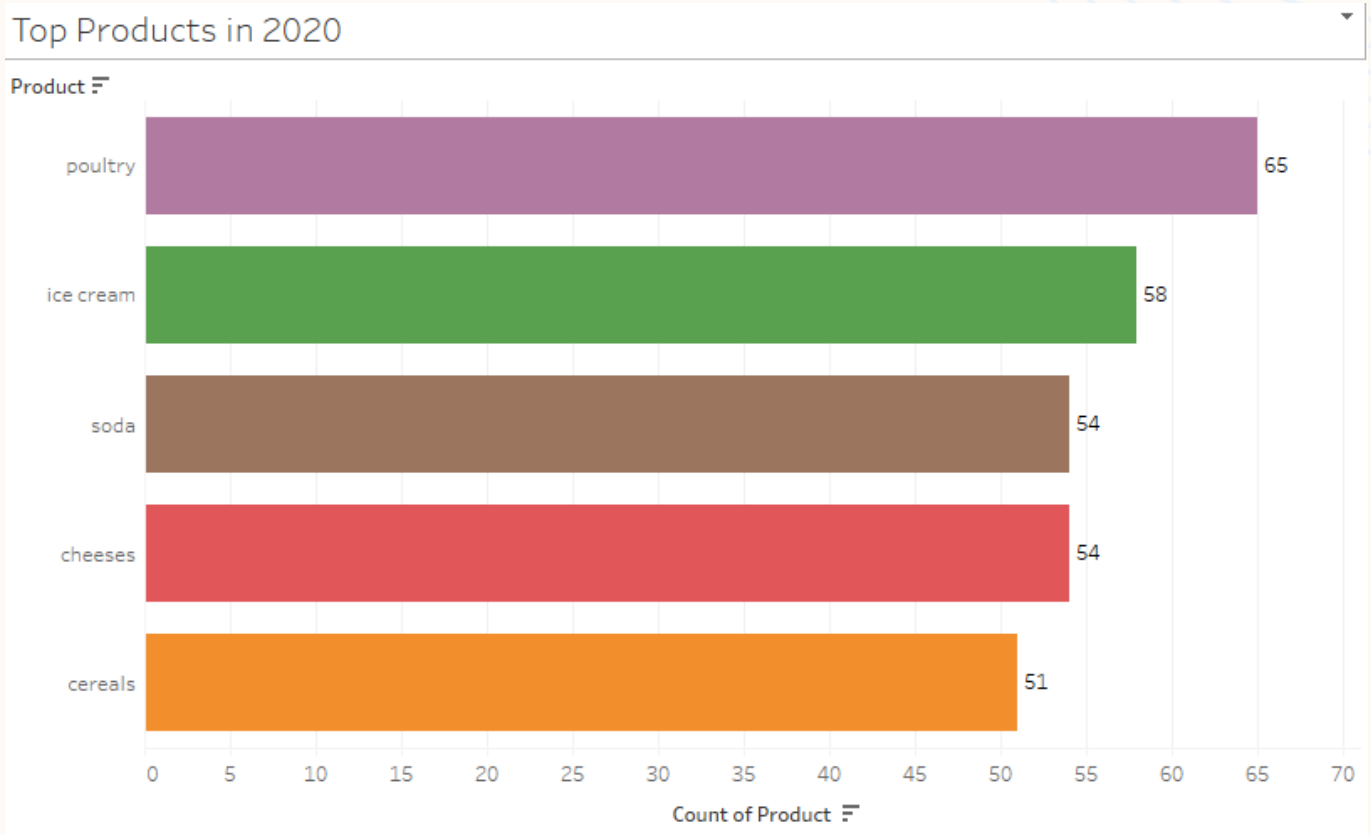


Exploratory Data Analysis

Most ordered Products – 2020

1. Poultry is the most ordered (65) product amount the products.

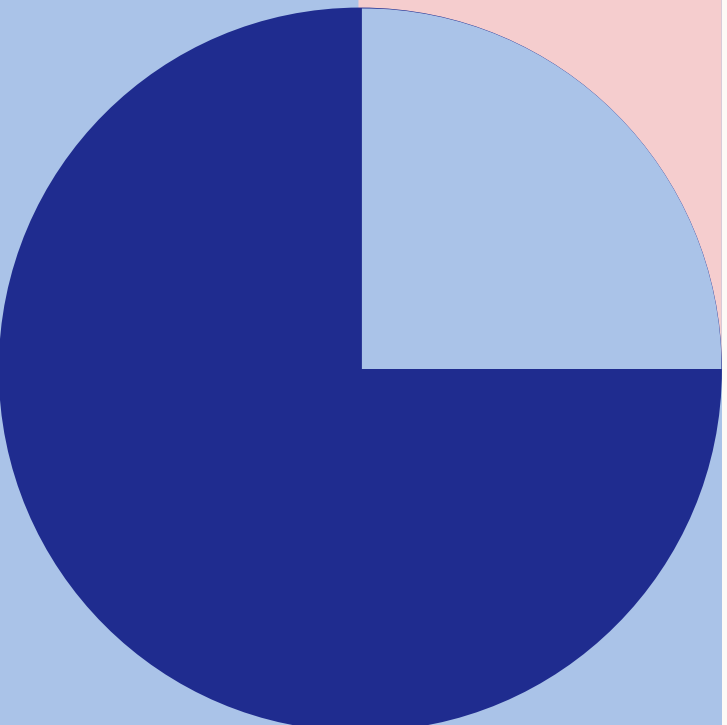
2. Ice cream (58) and Soda (54) are placed 2nd and 3rd in all the products.



MARKET BASKET ANALYSIS

USE OF MBA

- MBA is a strategy adopted by the retailers to gauge customer buying pattern.
- Retailers use analytics methods like market basket analysis (MBA) to comprehend the purchasing patterns of their customers
- It is all about understanding customer's basket behavior
- It investigates general group of items and customers end up buying together.
- It is used to find out which products customers usually buy together or put in the same basket. This purchasing data is used to increase the efficiency of sales and marketing.
- MBA finds relationships between the items in a customer's shopping cart based on various metrics.
- This level of understanding of the customer's shopping behavior is used the retailers in Target strategy and Recommendations systems

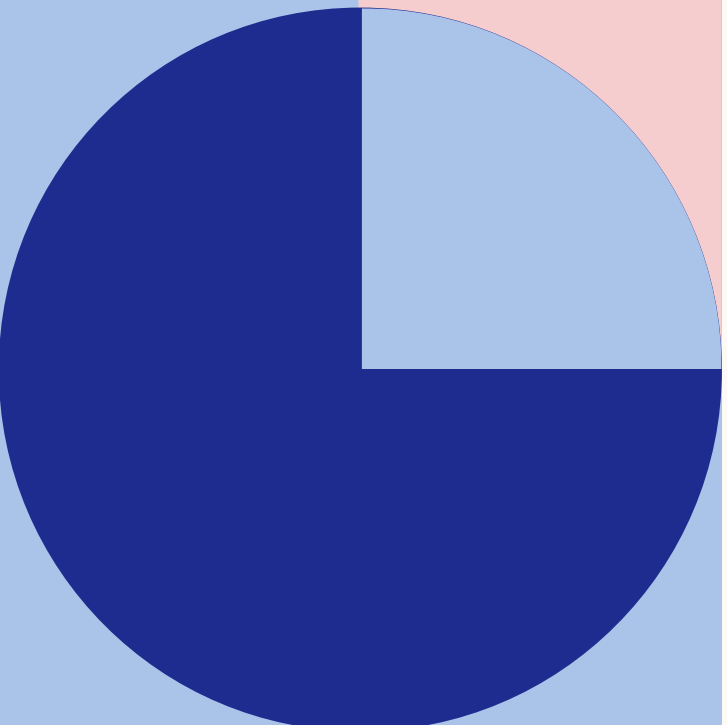


USE OF MBA

- Data from point of sale (PoS) systems that pertain to customers can be used in market basket analysis (MBA). Retailers benefit from its:
- Increasing sales and return on investment
- Boosts consumer engagement
- Increasing client satisfaction
- Aid in improving customer comprehension
- Identifies patterns and behavior of customers
- Improves marketing initiatives and strategies

ASSOCIATION RULES

- Association Rule is the most important Data Mining technique used in Market Basket Analysis
- It tries to associate different items in a shopping cart with some others using some metrics Increasing client satisfaction
- Mainly, it is related to the statement “What goes with What”
- Association Rules give a result like “Set A → Set B”
  IF (items in Set A are bought)
  THEN (items in Set B will be bought) Identifies patterns and behavior of customers
- It is a directional rule, and the inverse does not necessarily hold true
- Improves Here, Set A is called ‘Precedent’ and Set B is called ‘Consequent



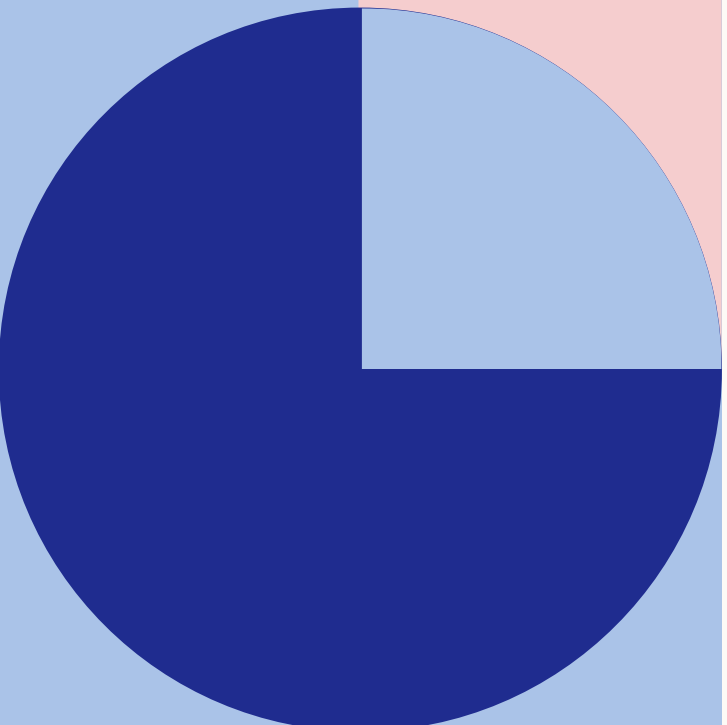
SUPPORT

- Support of A – is the fraction of transactions of A out of the total transactions.
- If item A is bought 10 times out of the total 1000 transactions of the store, then
  Support of A = 100/1000 ➔ 0.1 (10%)
- Similarly, if items A and B are together bought 50 times, then support of A and B = 50/1000 ➔ 0.05 (5%)

$$\text{Support of A} = \frac{\text{Number of Transactions containing A}}{\text{Total Transactions}}$$

CONFIDENCE

- Confidence of (A ➔ B) – is the likelihood of a customer buying item A, will also buy item B.
- This is the Probability of B give that A has been thought.
- Out of the 100 times that A has been bought, If B is bought 40 times along with A, then,
  Confidence (A ➔ B) = 40/100 ➔ 0.4 (40%)



$$\text{Confidence of A} \rightarrow \text{B} = P(B \mid A)$$

$$\frac{\text{Number of Transactions containing A and B}}{\text{Total Transactions containing A}}$$

LIFT

- Lift is the most important metric to consider when choosing an association rule
- Given A is bought, then –

  Lift is the % increase in chance of buying B
- Lift(A → B) < 1 ➔ Presence of A has decreased the chance of buying B
- Lift(A → B) > 1 ➔ Presence of A has increased the chance of buying B
- For example, Lift = 1.57 ➔ Chance of buying B has increased by 57%

$$\text{Lift of A} \rightarrow \text{B} = \frac{\text{Confidence of A} \rightarrow \text{B}}{\text{Support of B}}$$

Association Rules

- We perform MBA using KNIME on the given Grocery Store Dataset.
- We choose multiple thresholds for Min Support and Min Confidence to filter out less frequent and less appropriate rules
- Finally chosen thresholds for

SUPPORT

- Support of minimum = 0.02
- We want to create rules with only those items which appear in at least 3% of transctions.

CONFIDENCE

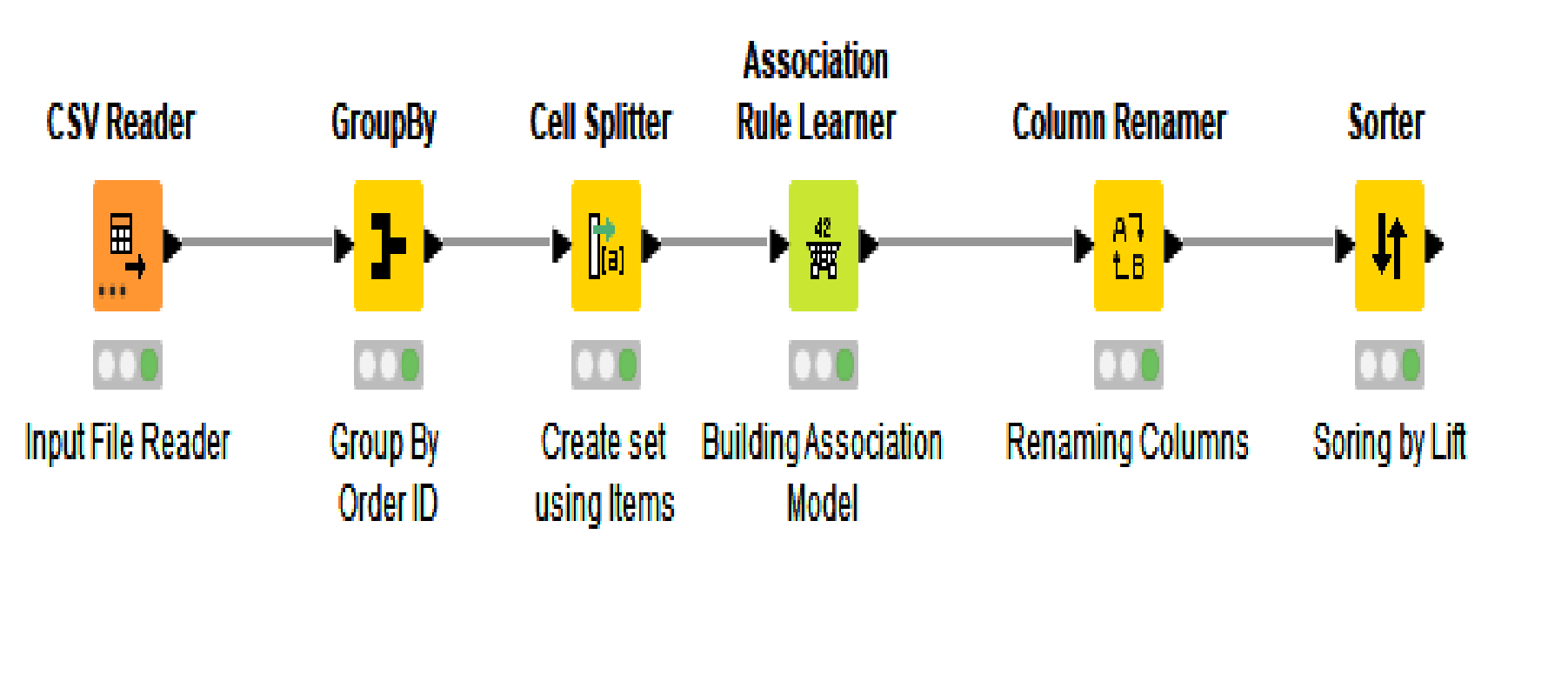
- Pursue Minimum Confidence Level = 0.8
- We want a minimum Confidence 80% in the rule.

MAXIMUM ITEM SET LENGTH

- Maximum Item set length = 10

Association Rules

KNIME WORKFLOW



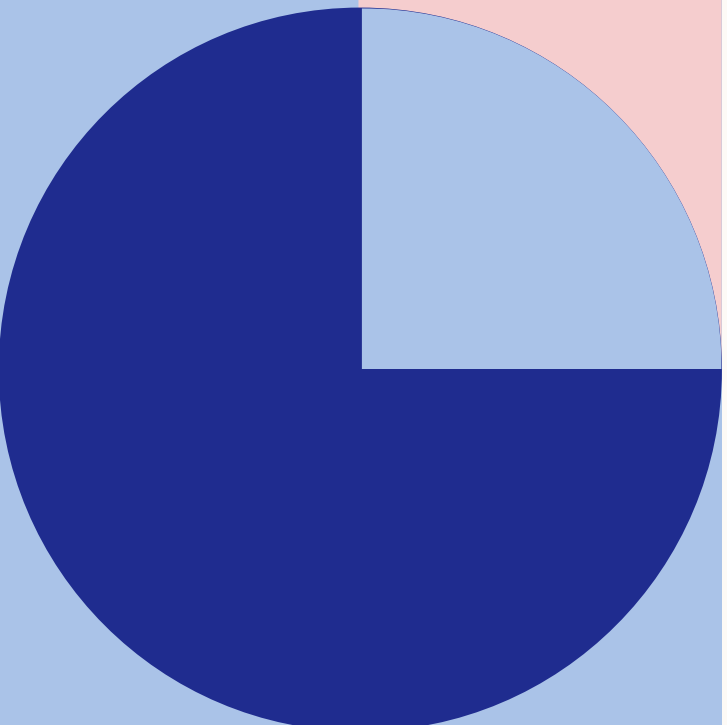
We have observed that, in the given data set we found out 39 association rules

Association Rules

Top 10 Combo suggestion using Lift

Row ID	Support	Confide...	Lift	Recom...	Recom...	Items_list
rule7	0.02	0.852	2.349	paper towels	<---	[eggs,dinner rolls,ice cream,...]
rule0	0.02	0.852	2.267	mixes	<---	[yogurt,dishwashing liquid/detergent,all- purpose,...]
rule8	0.02	0.821	2.265	paper towels	<---	[eggs,dinner rolls,poultry,...]
rule23	0.023	0.839	2.258	ketchup	<---	[tortillas,coffee/tea,juice,...]
rule17	0.022	0.833	2.244	pasta	<---	[paper towels,dishwashing liquid/detergent,eggs,...]
rule6	0.02	0.885	2.219	ice cream	<---	[paper towels,eggs,dinner rolls,...]
rule11	0.021	0.828	2.218	spaghetti sa...	<---	[waffles,laundry detergent,mixes,...]
rule10	0.021	0.828	2.208	beef	<---	[poultry,fruits,hand soap,...]
rule35	0.026	0.857	2.194	cheeses	<---	[paper towels,cereals,sandwich bags,...]
rule2	0.02	0.821	2.191	beef	<---	[shampoo,fruits,lunch meat,...]

SUGGESTIONS & RECOMMENDATIONS



- Rules consist of Precedent(Items_list) and Consequent (Recommended_item) which gives us rules if Items_list are bought, then Recommended_item likely to bought.
- Rule #1 ➔ If [yogurt, dishwashing liquid/detergent, all- purpose, hand soap] is bought then there is 2.267 times likelihood that [mixes] will be bought with 85% confidence.
- Rule #8 ➔ If [eggs, dinner rolls, ice cream, pasta, lunch meat] is bought then there is 2.349 times likelihood that [paper_towels] will be bought with 85% confidence.
- Rule #24 ➔ [tortillas, coffee/tea, juice, soap] is bought then there is 2.258 times likelihood that [ketchup] will be bought with 84% confidence.
- Rule #3 ➔ [bagels, pasta, individual meals, pork] is bought then there is 2.18 times likelihood that [soda] will be bought with 85% confidence.
- Rule #2 ➔ [cheeses, all- purpose, tortillas, coffee/tea] is bought then there is 2.136 times likelihood that [yogurt] will be bought with 85% confidence.
- Rule #13 ➔ [sandwich loaves, fruits, toilet paper, juice] is bought then there is 2.147 times likelihood that [bagels] will be bought with 83% confidence.
- Rule #36 ➔ [paper towels, cereals, sandwich bags, sugar] is bought then there is 2.194 times likelihood that [cheeses] will be bought with 85% confidence.
- Rule #39 ➔ [dinner rolls, spaghetti sauce, hand soap, sugar] is bought then there is 2.008 times likelihood that [poultry] will be bought with 85% confidence..

Suggestion & Recommendations

- Poultry could be suggested s combo offer with most of the food and snacks items such as dinner rolls and spaghetti sauce.
- Also maximize the Poultry sales by cross selling other items with this like – BUY 2 POULTRY GET 20% ON PAPER TOWELS
- Soda could be another item which can be offered in a combo.
- Beef/Pork buyers are seen to have a high likelihood of also buying cleaning products such as soap, hand soap, shampoo and dishwashing liquid.
- We can easily lift-up counter of the top combinations near sales counter to increase the preferring combinations.
- Since Poultry and Soda are most sold items and hand soap and Sandwich loaves are the least sold items. So, a combo offer of these would eventually increase a sale of hand soap and sandwich loaves as well.
- We can offer special discount on the least sold products to increase the sale of least product and increase the frequency of the customers.
- Make an offer of BUY 2 DETERGENTS and GET 1 SOAP FREE.

THANK YOU

Harikrishnan M