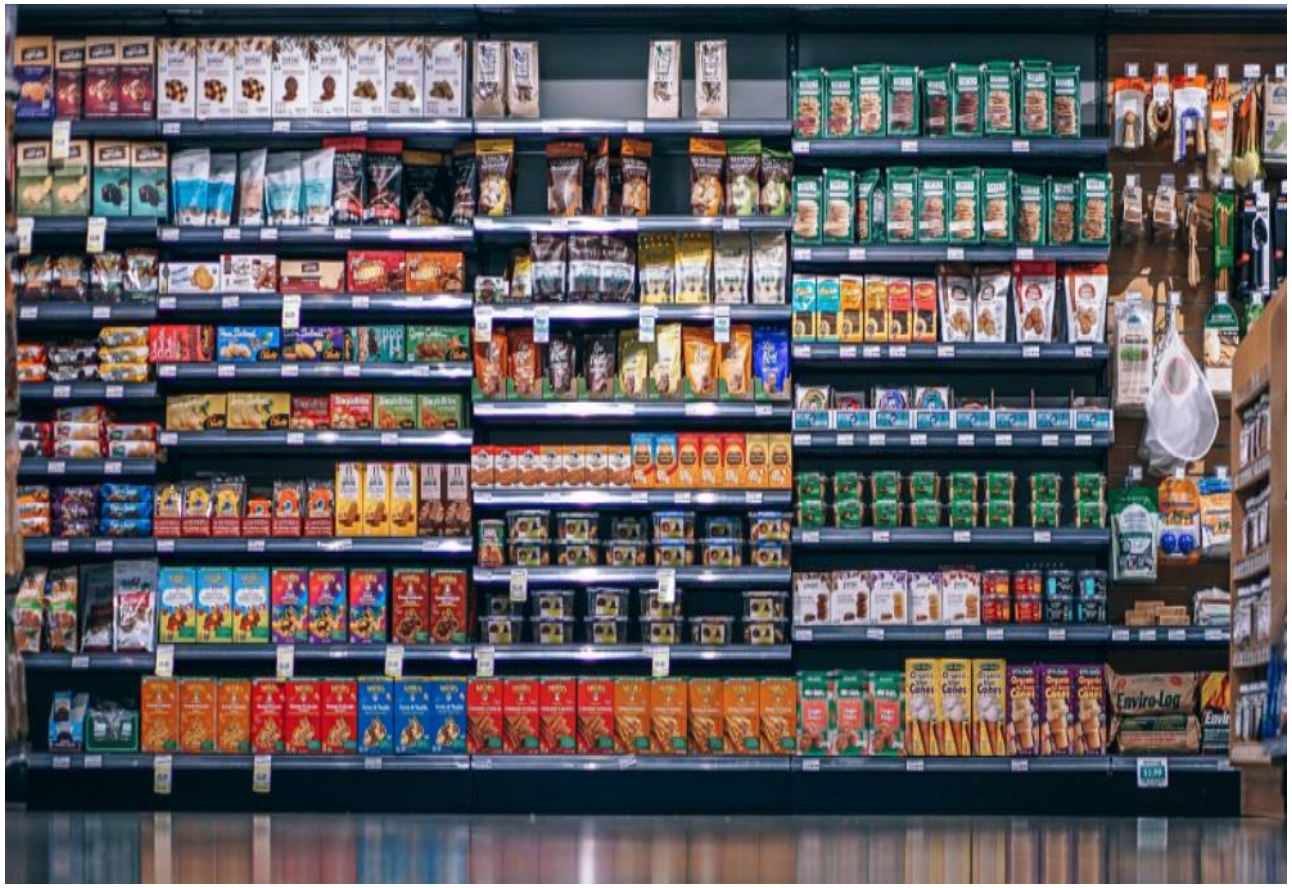
MRA PROJECT

MILESTONE -2

PINAK PANI GOGOI (DSBA AUG'20)

AGENDA

- ❑ Problem Statement
- ❑ About the dataset
- ❑ EDA
- ❑ EDA summary
- ❑ MBA – Association Rule
- ❑ MBA - Summary
- ❑ Conclusion



Problem statement

- A Grocery Store shared the transactional data with you. Your job is to identify the most popular combos that can be suggested to the Grocery Store chain after a thorough analysis of the most commonly occurring sets of items in the customer orders. The Store doesn't have any combo offers. Can you suggest the best combos & offers?

- Grocery Store Data: dataset_group.csv

- The project involves conducting a thorough analysis of Point of Sale (POS) Data for providing recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.

About the Dataset

The dataset has order id and product sold along with the date of transaction. A small glimpse of the dataset as below,

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

The following steps were followed while analysing the data,

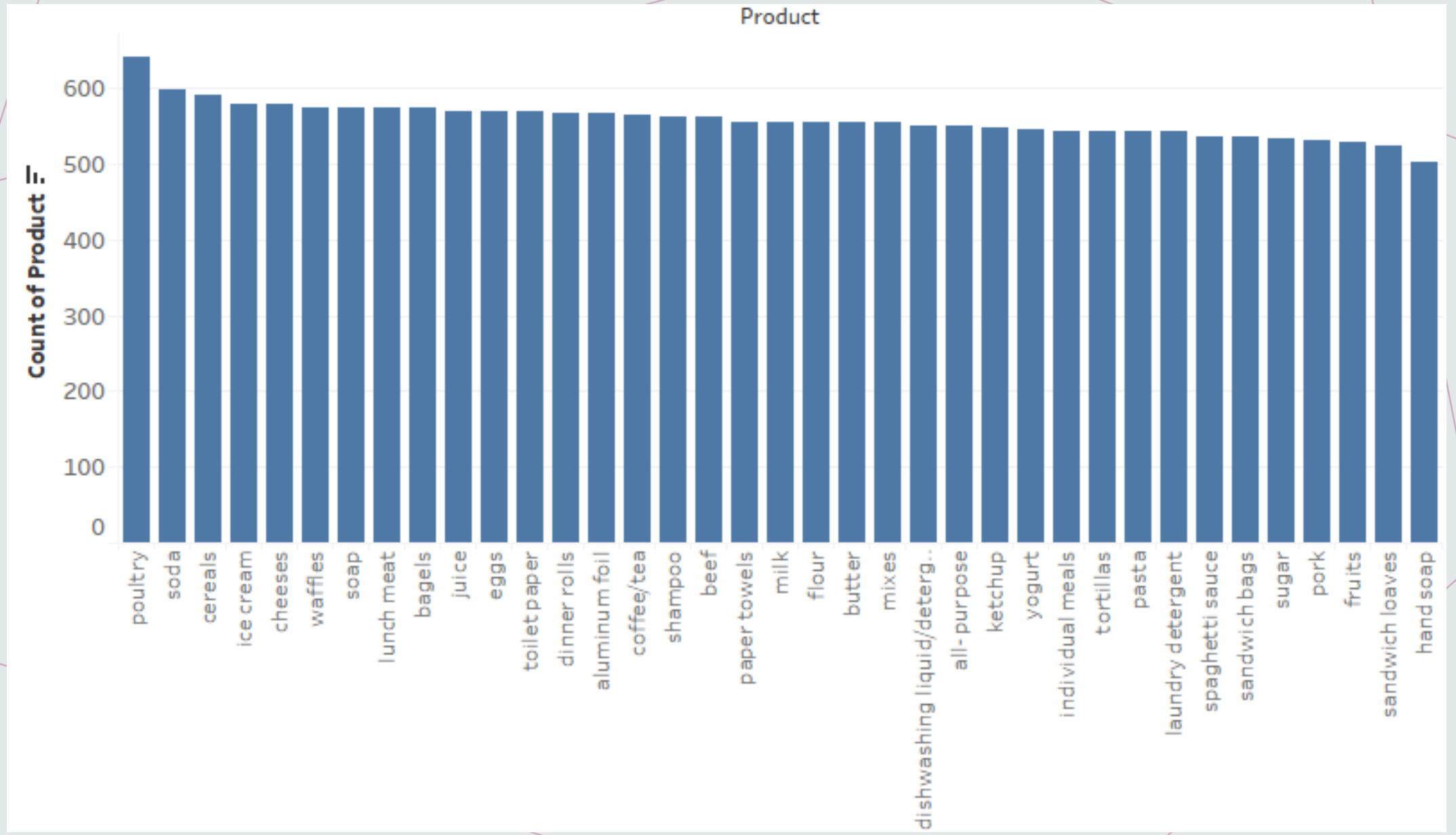
1. The raw dataset was first checked using Python 3 with Pandas and Numpy libraries. The basic stats of the dataset was derived using the above two libraries.
2. Tableau Public was used for data visualization and inferences for EDA.
3. Finally KNIME and Excel was used for MBA analysis.

All about the dataset

1) The dataset has 20641 rows and 3 columns
2) The dataset has 1 numerical, 1 datetime and 1 object variables
3) There is no null value in the dataset
4) There are 4730 duplicates in the dataset
5) The granularity of these duplicates is not available at present. These may be similar items of different brands but are of same item category. So, these duplicates are not omitted from the dataset.
6) If we consider the data as timeseries, we can see that Oct to Dec data is not present in the dataset.

A detailed EDA on the dataset has been performed which is shown in subsequent slides. Tableau public has been used for data visualization.

EDA

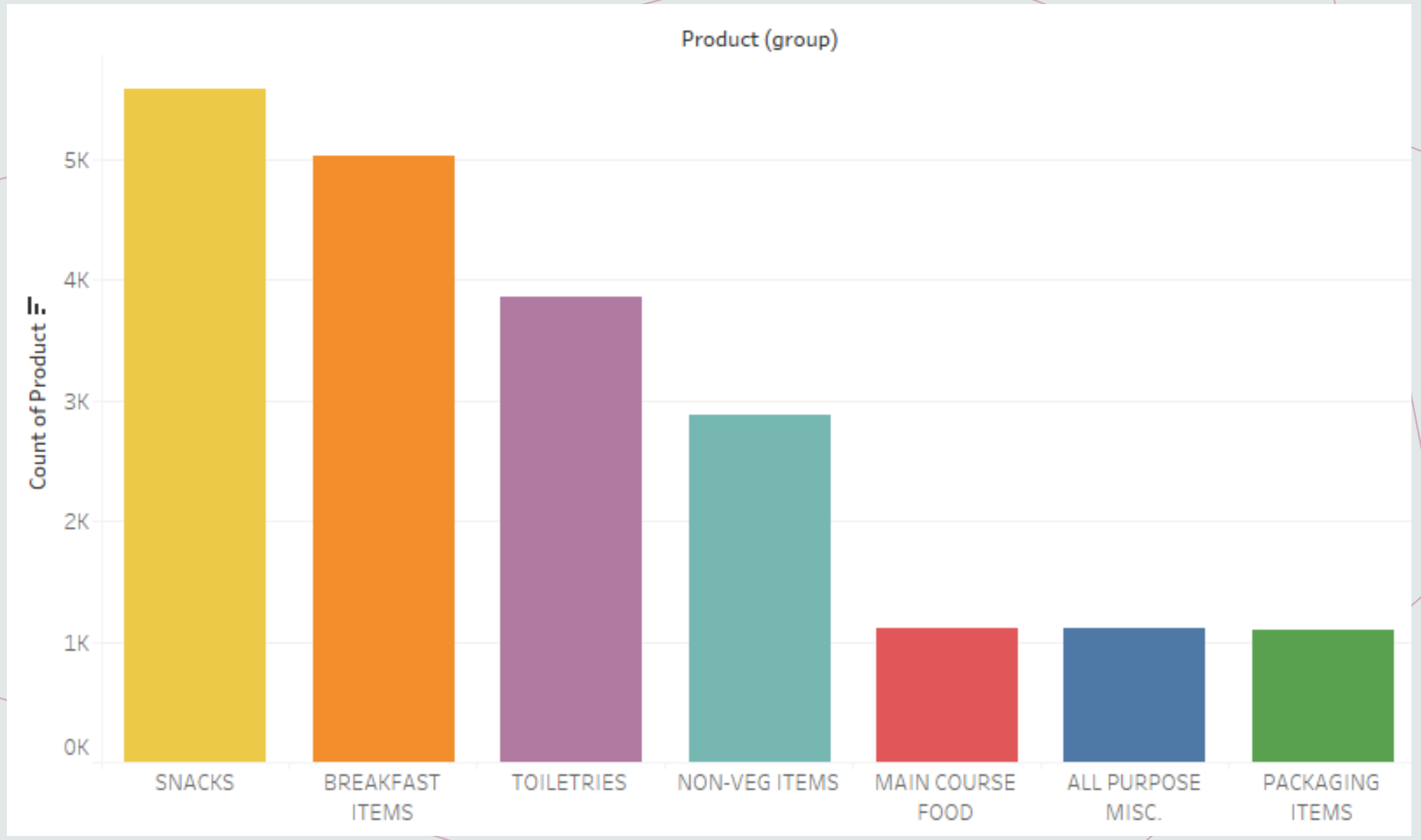

Product
Count of Product
600
500
400
300
200
100
0
poultry
soda
cereals
ice cream
cheeses
waffles
soap
lunch meat
bagels
juice
eggs
toilet paper
dinner rolls
aluminum foil
coffee/tea
shampoo
beef
paper towels
milk
flour
butter
mixes
dishwashing liquid/deterg..
all-purpose
ketchup
yogurt
individual meals
tortillas
pasta
laundry detergent
spaghetti sauce
sandwich bags
sugar
pork
fruits
sandwich loaves
hand soap

EDA

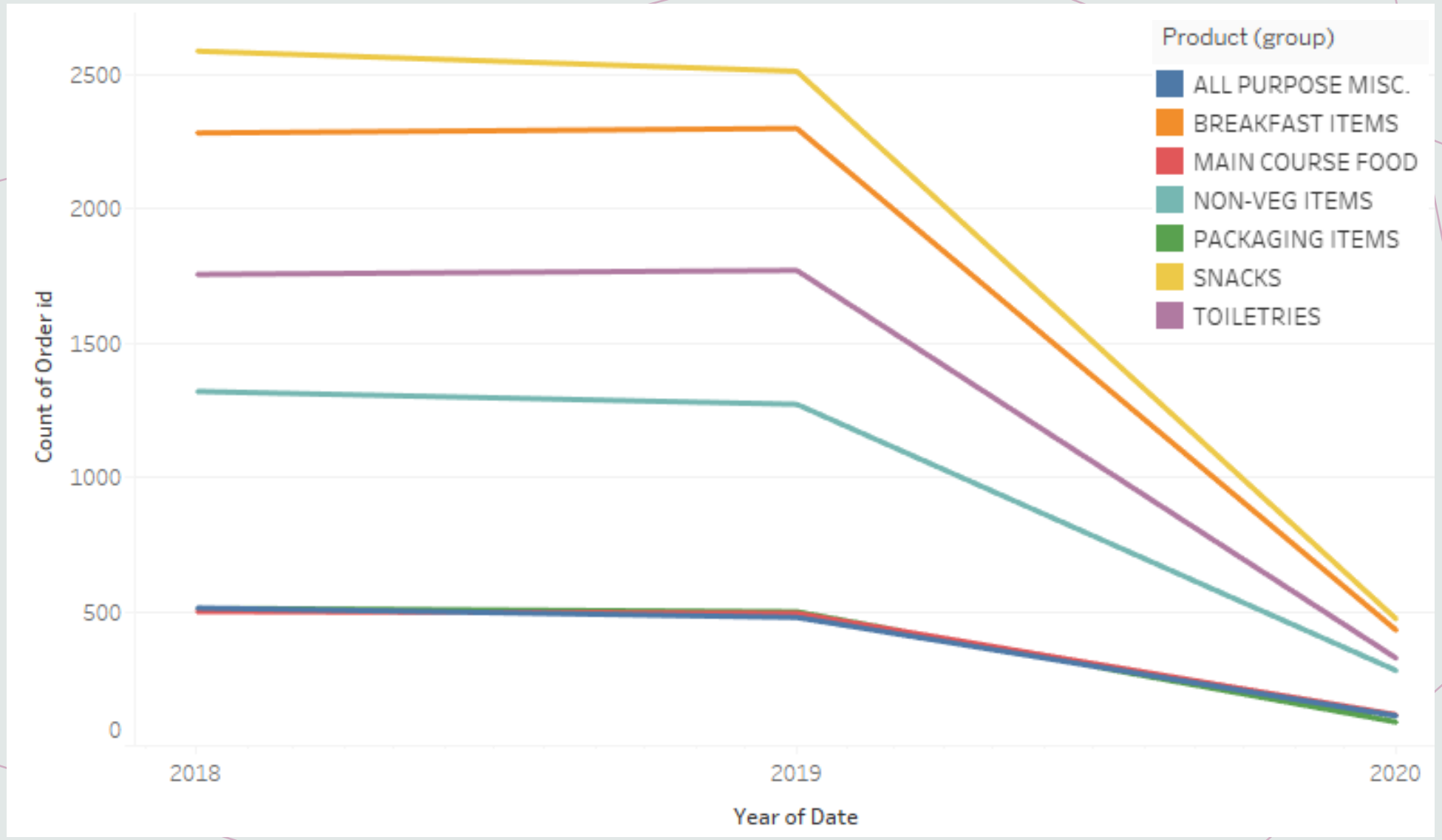
There are around 37 products in the dataset. All those are categorized into main 7 **categories for clarity.** The following are the categories,

1. ALL PURPOSE MISC – All purpose , mixes
2. BREAKFAST ITEMS - Butter, cereals, cheeses , Coffee/Tea, fruits, milk, sandwich loaves, waffles, yogurt
3. MAIN COURSE FOOD - Dinner rolls, individual meals
4. NON-VEG ITEMS - Beef, eggs, lunch meat, pork, poultry
5. PACKAGING ITEMS – Aluminium foil, sandwich bags
6. SNACKS - Bagels, flour, ice-cream, juice, ketchup, pasta, soda, spaghetti sauce, sugar, tortillas
7. TOILETRIES - dishwashing liquid/detergent, hand soap, laundry detergent , paper towels, Shampoo, soap, toilet paper

EDA

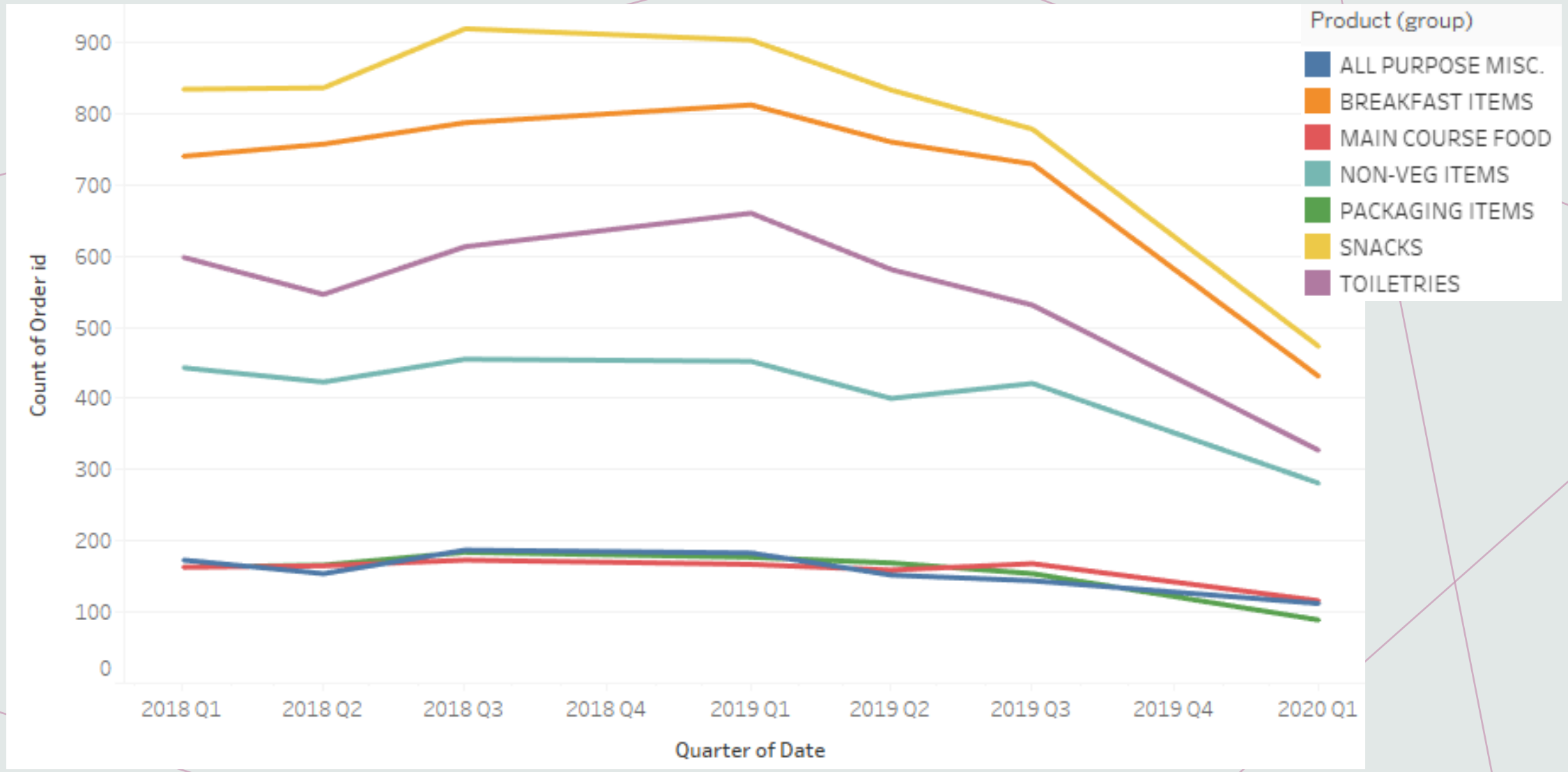

Product (group)
Count of Product
5K
4K
3K
2K
1K
0K
SNACKS
BREAKFAST ITEMS
TOILETRIES
NON-VEG ITEMS
MAIN COURSE FOOD
ALL PURPOSE MISC.
PACKAGING ITEMS

EDA – Timeseries Yearly



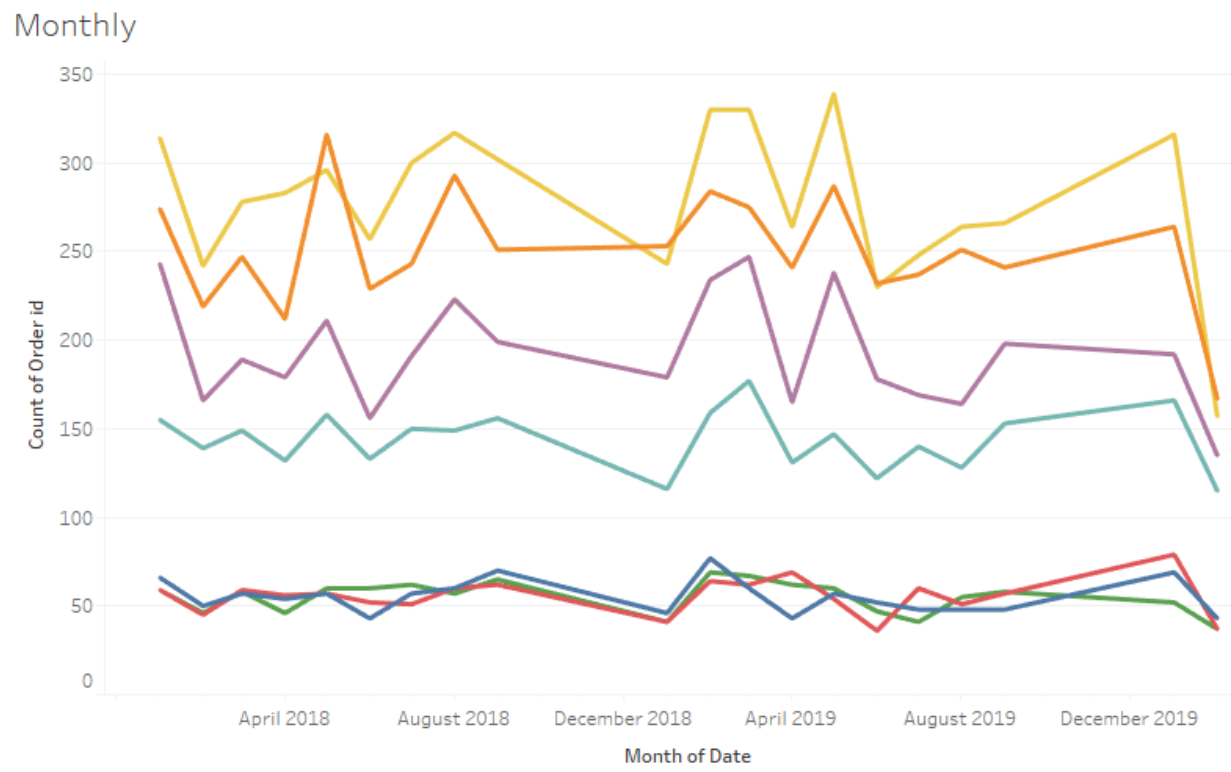
The order count is decreasing from 2019 for all product categories

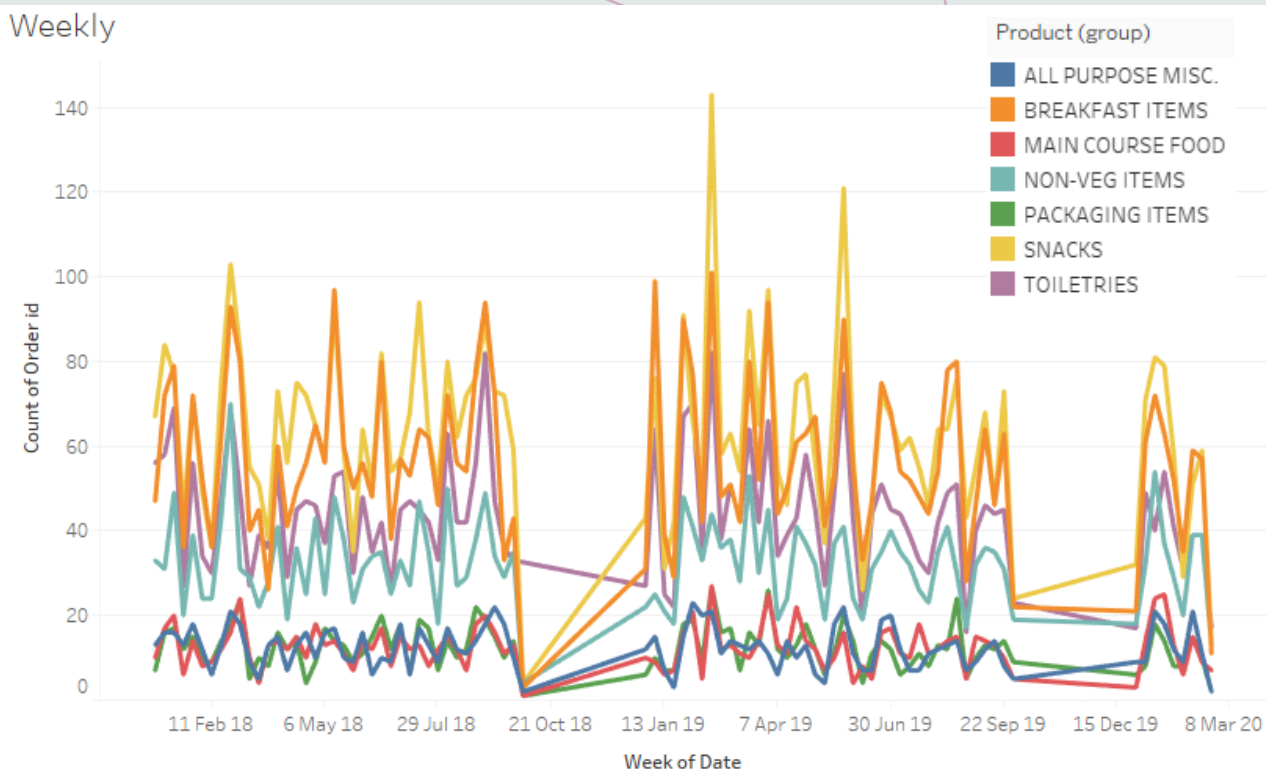
EDA – Timeseries Quarterly



The order count is decreasing from Q1 2019 for all product categories

EDA – Timeseries Monthly/ weekly





- There is no particular trend order counts in monthly and weekly data.
- There is no data from October to December in 2018 as well as 2019. So, basically there is seasonality of missing data, which may be due to annual closure of store or similar reason.

EDA – Summary

1) The dataset has 20641 rows and 3 columns
2) The dataset has 1 numerical, 1 datetime and 1 object variables
3) There is no null value in the dataset
4) There are 4730 duplicates in the dataset
5) There are around 37 products in the dataset. All those are categorized into main 7 categories for clarity
6) Snack, breakfast items and toiletries are the most ordered categories
7) The order count is decreasing from 2019 for all product categories
8) There is no particular trend in order counts in monthly and weekly basis
9) There is no data from October to December in 2018 as well as 2019. So, basically there is seasonality of missing data, which may be due to annual closure of store or similar reason.

MBA – Association rule

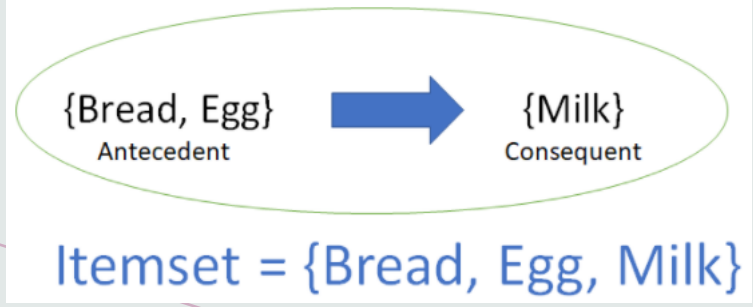
Association Rules is one of the very important concepts of machine learning being used in market basket analysis. It is relevant in the following scenarios.

1) In a store, all vegetables are placed in the same aisle, all dairy items are placed together and cosmetics form another set of such groups.
2) 80% of people who buys book online may also buy online Music.
3) 50% of people who buys health insurance may buy term insurance.

Basically association rule opens up the avenue to the marketing department to acquire propspective customers for the products.

" Association Rules do not extract an individual's preference, rather find relationships between set of elements of every distinct transaction. "

Rule consists of an **antecedent** and a **consequent**, both of which are a list of items.



Some Information written here is taken from Google.com

MBA – Association rule

Various metrics are in place to help us understand the strength of association between antecedent and consequent.

1. Support: This measure gives an idea of how frequent an *itemset* is in all the transactions.

Support of the product = (Number of transactions includes that product) / (Total number of transactions)

2. Confidence: This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents. Confidence can be interpreted as the likelihood of purchasing both the products A and B.

Confidence (A=>B) = (Number of transactions includes both A and B) / (Number of transactions includes only product A)

3. Lift: Lift is the rise in probability of having {Y} on the cart with the knowledge of {X} being present over the probability of having {Y} on the cart without any knowledge about presence of {X}.

In cases where {X} actually leads to {Y} on the cart, value of lift will be

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

Some Information written here is taken from Google.com

MBA – Support/Confidence threshold

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.

1. A minimum support threshold is applied to find all frequent item sets in a database.
2. A minimum confidence constraint is applied to these frequent item sets in order to form rules.

MBA – KNIME Results

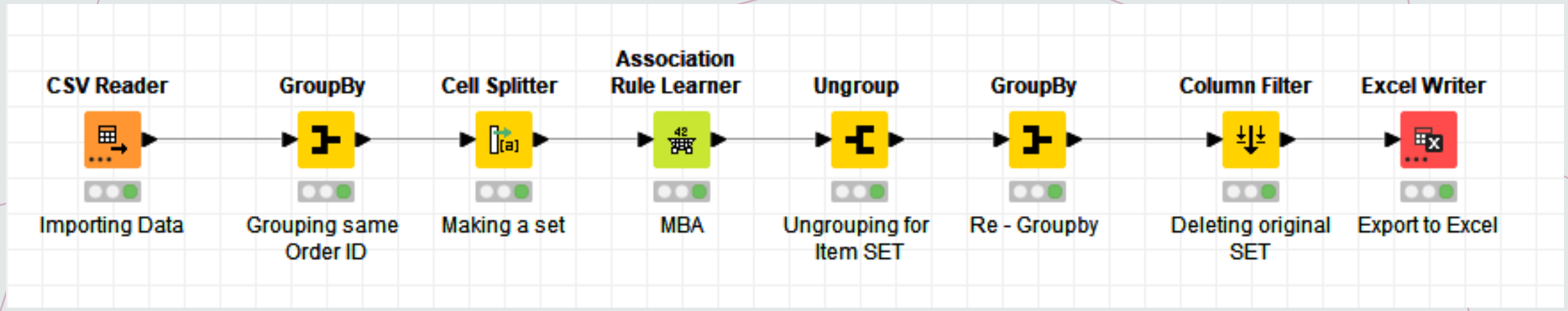
In the present dataset, to arrive at the rules, various support and confidence threshold were checked.

For example, minimum support of 0.1 and minimum confidence of 0.5 yielded only 1 rule. So, as a thumb rule, minimum threshold confidence is kept at 0.75 for analysis.

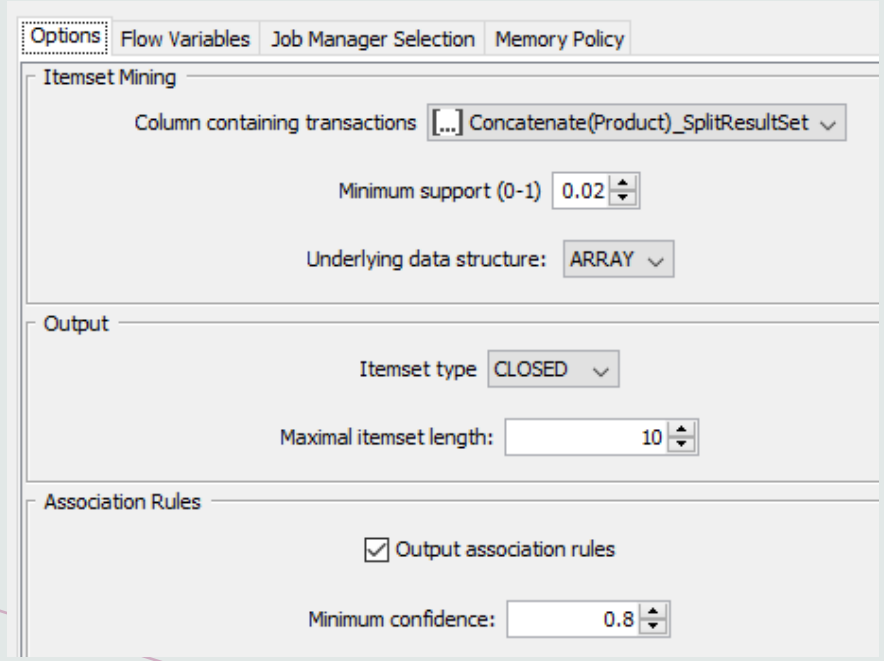
Minimum support of 0.03 and minimum confidence of 0.75 yielded only 4 rules.
Minimum support of 0.02 and minimum confidence of 0.8 yielded 39 rules.

So, more the rules, more is the flexibility to devise strategies. Lift value decides the preferred ones.

MBA – KNIME Workflow



Minimum support of 0.02 and minimum confidence of 0.8 yielded 39 rules.



MBA – Associations rules

RowID	Support	Confidence	Lift	Consequent	impli	Items (#1)
Row14	0.020	0.852	2.349	paper towels	<---	eggs, dinner rolls, ice cream, pasta, lunch meat
Row37	0.020	0.852	2.267	mixes	<---	yogurt, dishwashing liquid/detergent, all- purpose, hand soap
Row15	0.020	0.821	2.265	paper towels	<---	eggs, dinner rolls, poultry, ice cream, pasta
Row35	0.023	0.839	2.258	ketchup	<---	tortillas, coffee/tea, juice, soap
Row21	0.022	0.833	2.244	pasta	<---	paper towels, dishwashing liquid/detergent, eggs, dinner rolls, ice cream
Row22	0.020	0.885	2.219	ice cream	<---	paper towels, eggs, dinner rolls, pasta, lunch meat
Row36	0.021	0.828	2.218	spaghetti sauce	<---	waffles, laundry detergent, mixes, soap
Row27	0.021	0.828	2.208	beef	<---	poultry, fruits, hand soap, sugar
Row19	0.026	0.857	2.194	cheeses	<---	paper towels, cereals, sandwich bags, sugar
Row29	0.020	0.821	2.191	beef	<---	shampoo, fruits, lunch meat, pork
Row34	0.023	0.813	2.188	ketchup	<---	toilet paper, mixes, coffee/tea, soap
Row2	0.020	0.852	2.180	soda	<---	bagels, pasta, individual meals, pork
Row16	0.020	0.821	2.161	milk	<---	eggs, poultry, beef, sandwich bags
Row31	0.024	0.818	2.157	soap	<---	spaghetti sauce, all- purpose, sandwich bags, ketchup
Row13	0.021	0.828	2.152	yogurt	<---	dishwashing liquid/detergent, eggs, juice, sandwich bags
Row28	0.021	0.828	2.147	bagels	<---	sandwich loaves, fruits, toilet paper, juice
Row38	0.023	0.813	2.142	coffee/tea	<---	yogurt, ice cream, tortillas, cereals
Row20	0.022	0.833	2.138	eggs	<---	paper towels, dishwashing liquid/detergent, dinner rolls, ice cream, pasta
Row5	0.020	0.821	2.136	yogurt	<---	cheeses, all- purpose, tortillas, coffee/tea
Row18	0.022	0.833	2.133	soda	<---	ice cream, waffles, milk, pork
Row33	0.025	0.829	2.130	dinner rolls	<---	spaghetti sauce, poultry, waffles, laundry detergent
Row0	0.025	0.800	2.109	soap	<---	all- purpose, flour, soda, ketchup
Row4	0.023	0.813	2.080	cheeses	<---	butter, spaghetti sauce, ice cream, lunch meat
Row24	0.020	0.821	2.079	lunch meat	<---	paper towels, milk, individual meals, coffee/tea
Row23	0.024	0.818	2.076	waffles	<---	paper towels, laundry detergent, soda, sugar
Row25	0.025	0.800	2.062	dishwashing liquid/detergen	<---	paper towels, spaghetti sauce, milk, laundry detergent
Row1	0.025	0.800	2.048	soda	<---	all- purpose, waffles, laundry detergent, juice
Row26	0.025	0.806	2.021	ice cream	<---	paper towels, yogurt, pasta, lunch meat
Row9	0.029	0.846	2.008	poultry	<---	dinner rolls, spaghetti sauce, hand soap, sugar
Row6	0.028	0.842	1.998	poultry	<---	dinner rolls, spaghetti sauce, beef, sugar
Row8	0.023	0.839	1.990	poultry	<---	dinner rolls, spaghetti sauce, hand soap, soap
Row30	0.025	0.829	1.966	poultry	<---	shampoo, hand soap, juice, sugar
Row17	0.021	0.828	1.964	poultry	<---	eggs, tortillas, coffee/tea, sugar
Row12	0.028	0.821	1.947	poultry	<---	dinner rolls, spaghetti sauce, sandwich loaves, soap
Row32	0.023	0.813	1.928	poultry	<---	spaghetti sauce, laundry detergent, mixes, sugar
Row3	0.022	0.806	1.914	poultry	<---	butter, cheeses, sandwich loaves, laundry detergent
Row10	0.025	0.806	1.912	poultry	<---	dinner rolls, spaghetti sauce, ice cream, beef

Sorting by Lift in descending order

MBA – Associations rules

INTERPRETATION OF SUPPORT, CONFIDENCE and LIFT

(Assuming each order id is equivalent to one customer)

RowID	Support	Confidence	Lift	Consequent	implies	Items (#1)
Row14	0.020	0.852	2.349	paper towels	<---	eggs, dinner rolls, ice cream, pasta, lunch meat
Row37	0.020	0.852	2.267	mixes	<---	yogurt, dishwashing liquid/detergent, all- purpose, hand soap
Row15	0.020	0.821	2.265	paper towels	<---	eggs, dinner rolls, poultry, ice cream, pasta
Row35	0.023	0.839	2.258	ketchup	<---	tortillas, coffee/tea, juice, soap
Row21	0.022	0.833	2.244	pasta	<---	paper towels, dishwashing liquid/detergent, eggs, dinner rolls, ice cream

1 . Rule 1 (row 14) : Support says that 2% of customers purchased eggs, dinner rolls, ice cream, pasta, lunch meat and Paper towels. Confidence is that 85.2% of customers who bought eggs, dinner rolls, ice cream, pasta, lunch meat also bought Paper towels. Lift represents the 2.35 times increase in expectation that someone will buy Paper towels, when we know that they bought eggs, dinner rolls, ice cream, pasta, lunch meat.

2. Rule 2 (row 37) : Support says that 2% of customers purchased yogurt, dishwashing liquid/detergent, all- purpose, hand soap and Mixes. Confidence is that 85.2% of customers who bought yogurt, dishwashing liquid/detergent, all- purpose, hand soap also bought mixes. Lift represents the 2.27 times increase in expectation that someone will buy Mixes, when we know that they bought yogurt, dishwashing liquid/detergent, all- purpose, hand soap.

MBA – Recommendations

Based on the association rules, following are the offers/recommendation for future marketing campaigns,

1. First of all, based on the rules, it is very important that the consequents are placed strategically near to antecedents in the retail outlets. Continuous observation on the sales of these items are required and change the locations if required. Same can be implemented in the Online sales portals where these associations will be flashed as suggestions to the customers.

2. Basic outcome of a MBA is combo offers to boost sales. Some of such offers are discussed in subsequent section.

3. We have observed that most of the associations are valid with item numbers 4 or more. Store can also explore the cross selling with orders less than 3 different items. Association rules with very less support and confidence threshold can yield to such combos. But in our current analysis, such combos were not detected due to high confidence cut off.

4. Store can focus on targeted marketing campaigns by sending out promotional coupons to customers for products related to items they recently purchased.

MBA – COMBOS

Based on the association rules, following are the combo offers that can be displayed in the store,

1. Due to high lift value, a discount of 5% can be offered on a bundled pack purchase of eggs, dinner rolls, ice cream, pasta, lunch meat and Paper towels.

MBA – COMBOS

2. A basket of Eggs, poultry, beef, sandwich bags and Milks can be offered at 10% discount.

MBA – COMBOS

3. In the association rules, it is observed that Poultry is a very common purchase along with most of the items. So, Poultry is a very popular item. So, a combo offer on Poultry and Eggs is suggested as both are most preferred by customers.

MBA – COMBOS

4. Brand play can be a good strategy along with the combo offers. For example, a premium brand item (where store can have a bigger sales margin) can be pushed along with the combos for a little discounted price. For example, Yogurts made from A1 milks along with other associated items. Or, Ketchup made from tomatoes grown organically along with combo items etc…