10/30/2022

MRA – Milestone 2 Project

Shailesh Pande

Contents

Index for Figures and Tables

➢ Problem Statement:

A Grocery Store transactional data is available. We need 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 menu items in the customer orders. The Store doesn't have any combo meals. Suggest the best combo meals?

➢ Exploratory Analysis

- Exploratory Analysis of data & an executive summary of top findings, supported by graphs.

❖ Read the data (top 5 and bottom 5 rows)

Table 1: The original Data - top 5 and bottom 5 rows

	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

20641 rows × 3 columns

❖ Dimensions of the data

```
The number of rows (observations) is 20641
The number of columns (variables) is 3
```

❖ Datatype after converting Date to Date type and Order Id to string

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

❖ Check for duplicate rows

```
Number of Duplicated Rows are  4730
```

❖ Dimensions of the data after dropping the duplicate rows

```
The number of rows (observations) is 15911
The number of columns (variables) is 3
```

❖ Interpretation of Statistical Description of the features

Table 2 : Statistical Summary of the Data

	count	unique	top	freq	first	last
Date	15911	603	2019-02-08 00:00:00	138	2018-01-01	2020-02-26
Order_id	15911	1139	311	26	NaT	NaT
Product	15911	37	poultry	480	NaT	NaT

- The date range is from **1st Jan 2018 to 26th Feb 2020**
- There are **1139 unique Order Ids** (Marketing Baskets) ranging from 1 to 1139.
- There are **37 unique Products** , with 'poultry' Being present in 480 baskets out of a total of 1139 Marketing Baskets

- Noticeable trends across months/years/quarters/days etc.

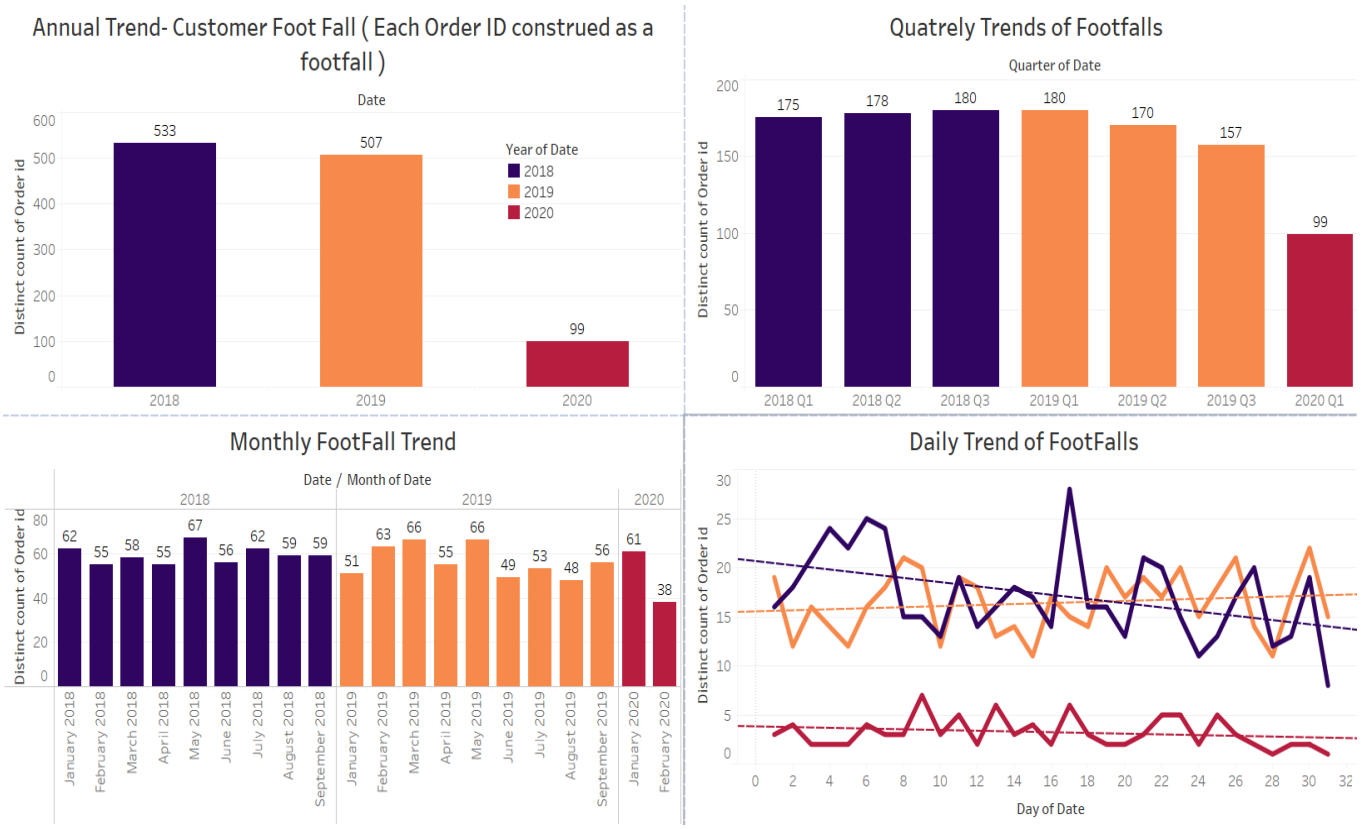

Figure 1: Annual, Quarter, Monthly and Daily trends of Footfalls

- **We note that the Q4 data is not available for years 2018 and 2019**

- Q1 data for 2020 includes data only for the months of Jan and Feb.

- Quarterly Trend above shows a decline in footfalls since Q2 of 2019.

- Monthly Trend indicates decline from June 2019 onwards.

- Feb 2020 is a huge decline

- *Product Popularity , measured by number of order_ids which include the product*

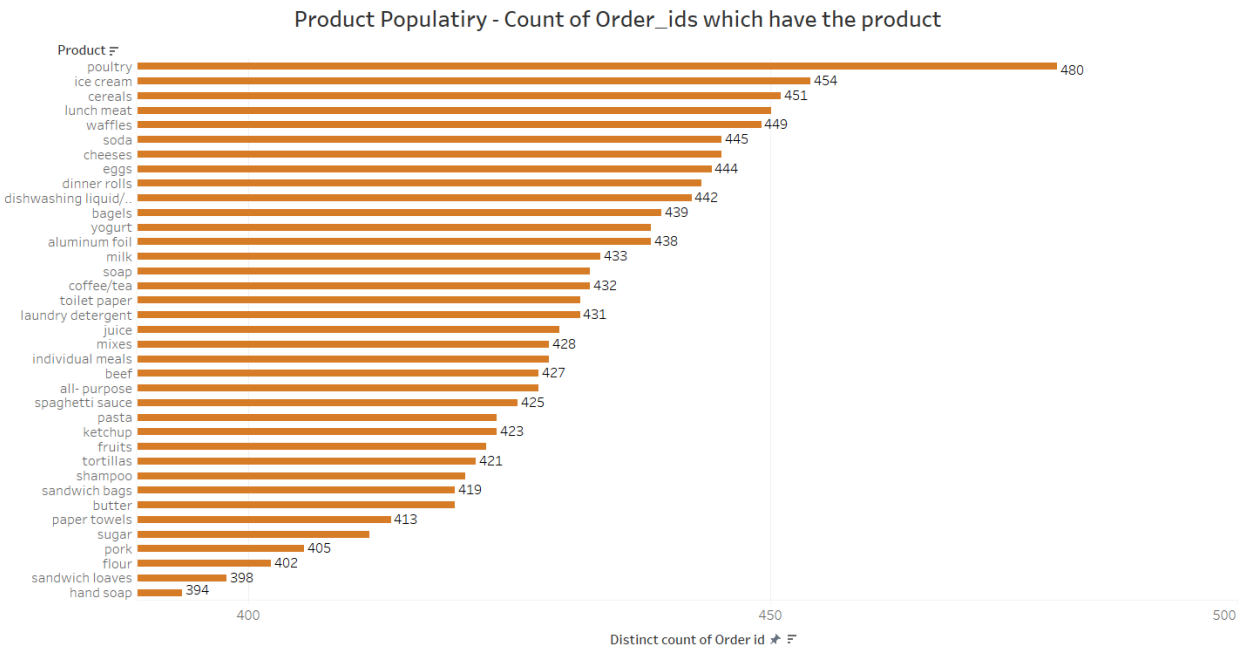

Figure 2: Popularity of Products

- ✓ Poultry as a product is present in 480 order_ids out of a total of 1139 unique order_ids, that is Poultry is present in 480/1139 = 42 % of all transactions
- ✓ Hand Soap , is the least transacted product, being present in 394/1139 = 35 % of all transactions

➢ Use of Market Basket Analysis (Association Rules)

- Association rules and their relevance

- ❖ Association rule identifies frequent if-then associations.
- ❖ Support is an indication of how frequently the items appear in the data or Support refers to how often a given rule appears in the database being mined.

$$\text{Support}(X,Y) = \text{Frequency}(X,Y) / \text{Total Transactions}$$

- ❖ Confidence indicates the number of times the if-then statements are found true.

$$\text{Confidence}(X,Y) = \text{Support}(X, Y) / (\text{Support}(X))$$

- A third metric, called *lift*, can be used to compare confidence with expected confidence, or how many times an if-then statement is expected to be found true.

Lift = Support (X,Y) / { Support (X) * Support (Y)} = Confidence (X,Y) / Support (Y)

Confidence (X,Y

- If the lift value is a negative value, then there is a negative correlation between datapoints. If the value is positive, there is a positive correlation, and if the ratio equals 1, then there is no correlation.

- The example given below illustrates the concept with numbers

- Example 1: Lift is < 1 , negative correlation example

Table 3: Association rules Example Matrix , Lift < 1

Per Day Details for a super Store	Number	Calcutated %	Inference
Total Transactions	10000		
Total Transactions in which Product X is present	600		
Total Transactions in which Product Y is present	400		
Total Transactions in which Products X and Y are present	**20**		
Support (X)	600/10000	6.00%	Product X is present in 6 % of all transactions
Support (Y)	400/10000	4.00%	Product Y is present in 4 % of all transactions
Support (X, Y) - X and Y are present together in a basket	20/10000	0.20%	Products X and Y are together present in 0.2 % of all transactions
Confidence (X → Y) , Y will be present in the basket if X is already present	0.2% / 6%	3.33%	There is a 3.3% chance that Y will be present in a Marketing Basket , if X is already present in the basket
Lift (X → Y)	3.33% / 4%	0.83	There is a 4 % chance that Y is present in a Basket, however this chance is reduced to 3.3 % , if X is present in the basket. So the presence of X is reducing the chance of Y being present in the basket by 3.3/4 = 0.833 times. This is a case of **Negative Correlation**

❖ Example 2: Lift is > 1 , positive correlation example

Table 4: Association Rules, Example Matrix, Lift > 1

Per Day Details for a super Store	Number	Calcutated %	Inference
Total Transactions	10000		
Total Transactions in which Product X is present	600		
Total Transactions in which Product Y is present	400		
Total Transactions in which Products X and Y are present	**200**		
Support (X)	600/10000	6.00%	Product X is present in 6 % of all transactions
Support (Y)	400/10000	4.00%	Product Y is present in 4 % of all transactions
Support (X, Y) - X and Y are present together in a basket	200/10000	2.00%	Products X and Y are together present in 2 % of all transactions
Confidence (X → Y) , Y will be present in the basket if X is already present	2% / 6%	33.33%	There is a 33% chance that Y will be present in a Marketing Basket , if X is already present in the basket
Lift (X → Y)	33.33% / 4%	8.33	There is only a 4 % chance that Y is present in a Basket, however this chance is lifted upto 33 % , if X is present in the basket. So the presence of X lifts the chance of Y being present in the basket by by 33/4 = 8.33 times. This is a case of **Positive Correlation**

❖ Example 3 : Lift = 1 , case of no correlation

Table 5: Association Rule , Example Matrix , Lift = 1

Per Day Details for a super Store	Number	Calcutated %	Inference
Total Transactions	10000		
Total Transactions in which Product X is present	600		
Total Transactions in which Product Y is present	400		
Total Transactions in which Products X and Y are present	**24**		
Support (X)	600/10000	6.00%	Product X is present in 6 % of all transactions
Support (Y)	400/10000	4.00%	Product Y is present in 4 % of all transactions
Support (X, Y) - X and Y are present together in a basket	24/10000	0.24%	Products X and Y are together present in 0.24% of all transactions
Confidence (X → Y) , Y will be present in the basket if X is already present	0.24% / 6%	4.00%	There is a 4 % chance that Y will be present in a Marketing Basket , if X is already present in the basket
Lift (X → Y)	4% / 4%	1.00	Chance of Y being present in the basket remains unchanged at 4 % , irrespective whether X is present in the basket or not . This is a case of No Correlation.

○ KNIME workflow Image

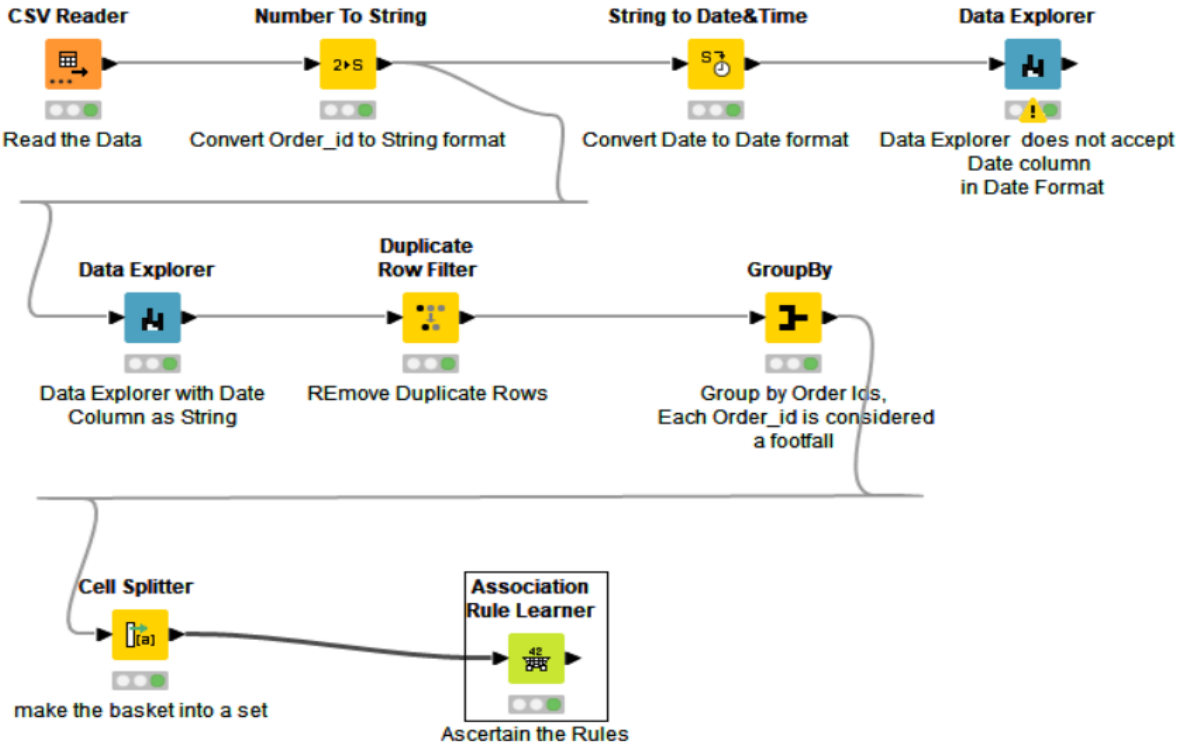

Figure 3: KNIME work flow for ascertaining Association Rules

○ Threshold values of Support and Confidence

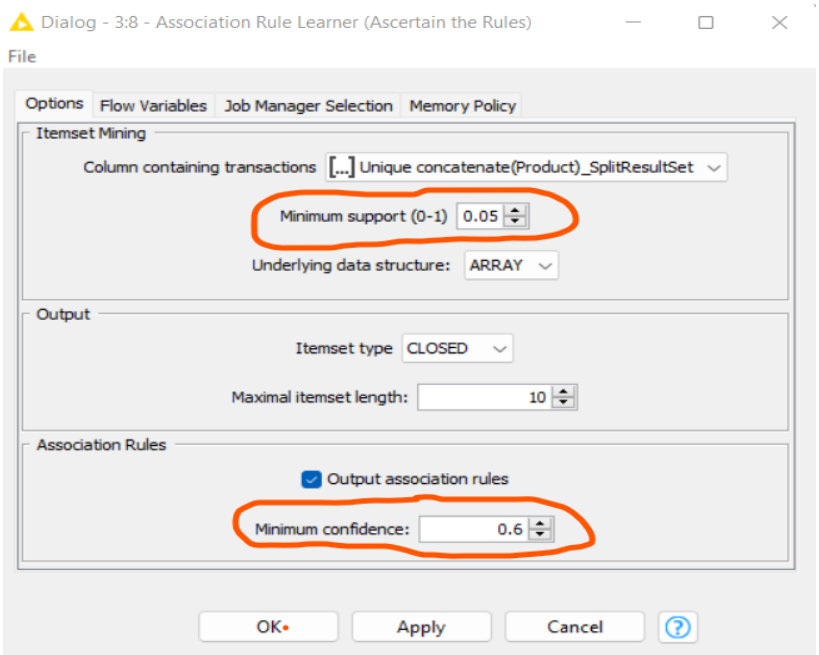

Figure 4: Threshold Values of Support and Confidence

- Minimum Support is taken as 0.05 , that is 5 %
- Minimum Confidence is taken as 0.6 , that is 60 %

We are going to identify those rules that have combination of products , known as 'antecedents' present in at least 5 % of all transactions. The probability of the 'consequent' product being present in the same basket , in which 'antecedent ' product or products are already present , should be 60 % or more.

➢ Associations Identified

○ Associations in tabular manner in descending order of Lift value

Table 6: Association Rules

Rule	Support	Confidence	Lift	Consequent	implies	Items
1	0.055	0.649	1.791	paper towels	<---	[eggs, ice cream, pasta]
2	0.055	0.643	1.731	pasta	<---	[paper towels, eggs, ice cream]
3	0.051	0.674	1.726	cheeses	<---	[bagels, cereals, sandwich bags]
4	0.050	0.640	1.700	juice	<---	[yogurt, toilet paper, aluminum foil]
5	0.051	0.630	1.678	mixes	<---	[yogurt, poultry, aluminum foil]
6	0.051	0.611	1.660	sandwich bags	<---	[cheeses, bagels, cereals]
7	0.054	0.642	1.651	dinner rolls	<---	[poultry, spaghetti sauce, laundry detergent]
8	0.052	0.641	1.649	dinner rolls	<---	[poultry, spaghetti sauce, ice cream]
9	0.050	0.620	1.645	juice	<---	[yogurt, poultry, aluminum foil]
10	0.052	0.686	1.628	poultry	<---	[dinner rolls, spaghetti sauce, ice cream]
11	0.052	0.634	1.627	eggs	<---	[paper towels, dinner rolls, pasta]
12	0.052	0.602	1.621	pasta	<---	[paper towels, eggs, dinner rolls]
13	0.051	0.630	1.621	dinner rolls	<---	[poultry, spaghetti sauce, cereals]
14	0.055	0.630	1.616	eggs	<---	[paper towels, ice cream, pasta]
15	0.050	0.613	1.616	coffee/tea	<---	[yogurt, cheeses, cereals]
16	0.052	0.628	1.614	dinner rolls	<---	[poultry, spaghetti sauce, juice]
17	0.052	0.628	1.610	eggs	<---	[dinner rolls, poultry, soda]
18	0.051	0.604	1.589	milk	<---	[poultry, laundry detergent, cereals]
19	0.055	0.624	1.565	ice cream	<---	[paper towels, eggs, pasta]
20	0.051	0.617	1.558	cereals	<---	[cheeses, bagels, sandwich bags]
21	0.054	0.656	1.556	poultry	<---	[dinner rolls, spaghetti sauce, laundry detergent]
22	0.051	0.637	1.512	poultry	<---	[dinner rolls, spaghetti sauce, cereals]
23	0.052	0.602	1.429	poultry	<---	[dinner rolls, spaghetti sauce, juice]
24	0.050	0.600	1.424	poultry	<---	[dishwashing liquid/detergent, laundry detergent, mixes]

- 24 rules are identified with aforementioned Minimum Support level of 5% and Minimum Confidence of 60 %.
- We remove those rules that have Consequent Products which are already in the top ten list, as they do not need any additional support, such as ' poultry ' , ' ice cream ' etc.
- The rearranged list of rules with top sellers as consequent highlighted

Table 7: Association Rules with top selling products as Consequent highlighted

Rule	Suppo	Confiden	Lift	Consequent	implie	Items
1	0.050	0.613	1.616	coffee/tea	<---	[yogurt, cheeses, cereals]
2	0.050	0.640	1.700	juice	<---	[yogurt, toilet paper, aluminum foil]
3	0.050	0.620	1.645	juice	<---	[yogurt, poultry, aluminum foil]
4	0.051	0.604	1.589	milk	<---	[poultry, laundry detergent, cereals]
5	0.051	0.630	1.678	mixes	<---	[yogurt, poultry, aluminum foil]
6	0.055	0.649	1.791	paper towels	<---	[eggs, ice cream, pasta]
7	0.055	0.643	1.731	pasta	<---	[paper towels, eggs, ice cream]
8	0.052	0.602	1.621	pasta	<---	[paper towels, eggs, dinner rolls]
9	0.051	0.611	1.660	sandwich bags	<---	[cheeses, bagels, cereals]
10	0.051	0.617	1.558	cereals	<---	[cheeses, bagels, sandwich bags]
11	0.051	0.674	1.726	cheeses	<---	[bagels, cereals, sandwich bags]
12	0.054	0.642	1.651	dinner rolls	<---	[poultry, spaghetti sauce, laundry detergent]
13	0.052	0.641	1.649	dinner rolls	<---	[poultry, spaghetti sauce, ice cream]
14	0.051	0.630	1.621	dinner rolls	<---	[poultry, spaghetti sauce, cereals]
15	0.052	0.628	1.614	dinner rolls	<---	[poultry, spaghetti sauce, juice]
16	0.052	0.634	1.627	eggs	<---	[paper towels, dinner rolls, pasta]
17	0.055	0.630	1.616	eggs	<---	[paper towels, ice cream, pasta]
18	0.052	0.628	1.610	eggs	<---	[dinner rolls, poultry, soda]
19	0.055	0.624	1.565	ice cream	<---	[paper towels, eggs, pasta]
20	0.052	0.686	1.628	poultry	<---	[dinner rolls, spaghetti sauce, ice cream]
21	0.054	0.656	1.556	poultry	<---	[dinner rolls, spaghetti sauce, laundry detergent]
22	0.051	0.637	1.512	poultry	<---	[dinner rolls, spaghetti sauce, cereals]
23	0.052	0.602	1.429	poultry	<---	[dinner rolls, spaghetti sauce, juice]
24	0.050	0.600	1.424	poultry	<---	[dishwashing liquid/detergent, laundry detergent, mixes]

- Rules 10 – 24 are not considered for reasons stated above, consequent products (highlighted in yellow above) are already top sellers.

○ Support, Confidence, & Lift - explanation on calculated values

- Support values are all above 5 %, which means that the combination of products as listed in the ‘ Antecedents ‘ column occur in a basket more than 5 % times in all transactions.

- The Confidence values indicate the probability of the consequent product being in the basket if the antecedent products have already been put in the basket. All values for the selected rules are above 60 %

- Lift values for all Rules are more than 1 and range from 1.56 – 1.79. This shows that there is a positive correlation between the consequent product and the antecedent product combinations.

○ Suggested Possible Combos with Lucrative Offers

Table 8: Final Set of Rules to be used for Recommendations

Rule	Support	Confidence	Lift	Consequent	implies	Items (antecedents)	Combo's	Push Purchase Product ,which are in Bottom 10
6	0.055	0.649	1.791	paper towels	<---	[eggs, ice cream, pasta]	1	butter
7	0.055	0.643	1.731	pasta	<---	[paper towels, eggs, ice cream]		
8	0.052	0.602	1.621	pasta	<---	[paper towels, eggs, dinner rolls]		
2	0.050	0.640	1.700	juice	<---	[yogurt, toilet paper, aluminum foil]	2	sugar
5	0.051	0.630	1.678	mixes	<---	[yogurt, poultry, aluminum foil]		
3	0.050	0.620	1.645	juice	<---	[yogurt, poultry, aluminum foil]		
1	0.050	0.613	1.616	coffee/tea	<---	[yogurt, cheeses, cereals]	3	sandwich loaves
9	0.051	0.611	1.660	sandwich bags	<---	[cheeses, bagels, cereals]		
4	0.051	0.604	1.589	milk	<---	[poultry, laundry detergent, cereals]	4	hand soap

- Rules 1 -9 are studied deeper and rearranged as above so that the combination of the antecedent products and the consequent products form a common bundle.

- Rules 6,7,8 suggest combos of following products (orange highlighted above)

Combo 1 (Dinner items)

- ***eggs, ice cream, pasta, paper towels, dinner rolls***

- Rules 2,5,3 suggest combos of the following products (blue highlighted)

Combo 2 (Mixed)

- ***yogourt , toilet paper, aluminium foil, poultry, juices , mixes***

- Rules 1,9 suggest combos of following products (grey highlighted above)

Combo 3 (Breakfast Items)

- ***cheeses, bagels, cereals, sandwich bags, yogurt, coffee/tea***

- Rule 9 suggests

- ***Combo 4 (Food , Laundry)***
- ***milk, poultry, laundry detergent, cereals***

➢ Write recommendations

- Make discount offers or combos based on the associations and your experience

❖ We again have a look at the top ten selling products and the bottom ten products

- Top 10 Selling Products based on presence in Order Ids

➢

Count of Order_id / Footfalls	2018	2019	2020
poultry	214	219	47
ice cream	207	206	41
cereals	217	198	36
lunch meat	210	203	37
waffles	201	214	34
soda	198	207	40
cheeses	203	201	41
eggs	198	213	33
dinner rolls	195	203	45
dishwashing liquid/detergent	189	214	39

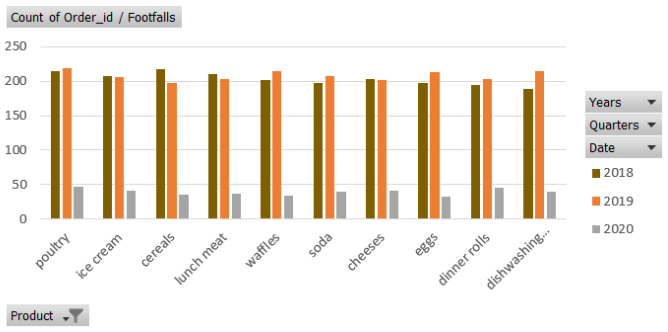

Figure 5: Top 10 Selling Products

- Bottom 10 Selling Products based on presence in Order Ids

Count of Order_id / Footfalls	2018	2019	2020
tortillas	193	191	37
shampoo	198	187	35
sandwich bags	191	187	41
butter	192	190	37
paper towels	179	198	36
sugar	189	193	29
pork	184	177	44
flour	202	165	35
sandwich loaves	178	184	36
hand soap	200	163	31

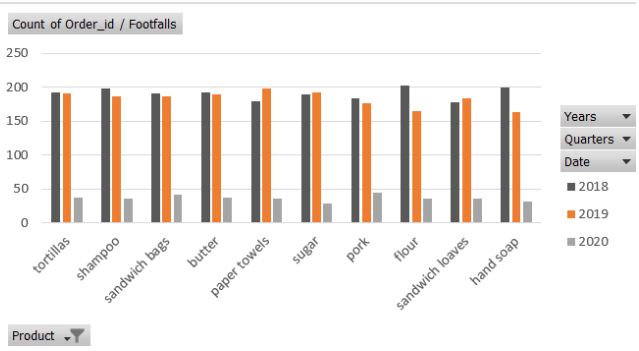

Figure 6: Last 10 selling Products

❖ We will use the Association Rules to push the sales of some of the products that are in the bottom 10 list

- **Recommendation 1 (considering Combo 1)**

- ❖ For purchase of **minimum 2 units** of minimum pack size each of
    - Eggs (antecedent)
    - Dinner Rolls (antecedent)
    - Pasta (antecedent)
    - **Butter (push sale product , in the bottom 10)**

  1 set of free Paper Towel (the consequent product, of the above listed antecedent products)

- **Recommendation 2 (considering Combo 2)**

  - ❖ For purchase of **minimum 2 units** each of

    - Yogurt (antecedent)
    - Aluminum Foil (antecedent)
    - Toilet Paper (antecedent)
    - **Sugar (push sale product , in the bottom 10)**

    1 unit of free Juice (the consequent product, of the above listed antecedent products)

- **Recommendation 3 (considering Combo 3)**

    - For purchase of **minimum 2 units** each of

        - Cheese (antecedent)
        - Cereals(antecedent)
        - Bagels (antecedent)
        - **Sandwich Loaves (push sale product , in the bottom 10)**

          1 unit of free Sandwich Bags (the consequent product, of the above listed antecedent products)

- **Recommendation 4 (considering Combo 4)**

    - For purchase of **minimum 2 units** each of

        - Poultry (antecedent)
        - Laundry Detergent (antecedent)
        - Cereals (antecedent)
        - **Hand Soap (push sale product , in the bottom 10)**

          1 unit of free Milk (the consequent product, of the above listed antecedent products)

END