Marketing and Retail Analytics Project

GROCERY COMBOS

Agenda and Executive Summary of the data

Contents of the presentation

- Problem statement
- Data Analysis
  - Info, Shape, Data type, Null value count, Duplicates, Summary Stats, Unique counts
- Exploratory Analysis
  - Exploratory Data Analysis and Inferences
  - Weekly, Monthly, Quarterly, Yearly Trends in Sales
  - Summary of the inferences from the above analysis
- Market Basket Analysis
  - What is Market Basket Analysis?
  - Association rules and their relevance
  - Threshold values
  - Association rules output
  - KNIME Workflow
  - Threshold values taken
- Associations Identified
  - Associations identified
  - A suggestion of possible combos
  - Discount offers and combos based on Associations
  - Recommendations

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 menu items in the customer orders. The Store doesn't have any combo meals. Can you suggest the best combo meals?

Data: dataset_group.csv

The data is provided from January to September for 2 years (2018 and 2019) and January to February for 1 year (2020).

Data Analysis – Info

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

The dataset has 3 columns and 20641 records as seen in the information of the dataset on the left.

There are 20641 non-null records in all the columns meaning there are no missing records based on the initial analysis that was done.

20641 records means that there were 20641 transactions done.

Data Analysis – Shape

```
(20641, 3)
```

```
The no. of rows:  20641
The no. of columns:  3
```

Shape of the dataset gives the total number of rows and columns in the dataset.

The shape of the data is (20641, 3) meaning the dataset has 20641 rows and 3 columns as shown on the left.

Data Analysis – Data type

```
Date        object
Order_id     int64
Product     object
dtype: object
```

The dataset has 3 variables out of which there are:

- 2 categorical variables and
- 1 numerical variable

Data Analysis – Null Value count

There are no null values or missing values in the dataset.

```
Date        0
Order_id    0
Product     0
dtype: int64
```

Data Analysis – Duplicates

```
The no. of duplicated rows is:  4730
```

There are 4730 duplicate entries.

However, we will not remove them because it is possible that the same consumer purchased numerous copies of the same product using the same order ID.

Data Analysis – Summary stats

	Order_id
count	20641.000000
mean	575.986289
std	328.557078
min	1.000000
25%	292.000000
50%	581.000000
75%	862.000000
max	1139.000000

- The descriptive statistics of the numerical column Order_id is shown in the table.
- There are 20641 records. There are no missing values.
- By looking at the table we can see that the minimum Order ID is 1 and the maximum Order ID is 1139. This means that the number of unique orders is 1139.

Data Analysis – Unique counts

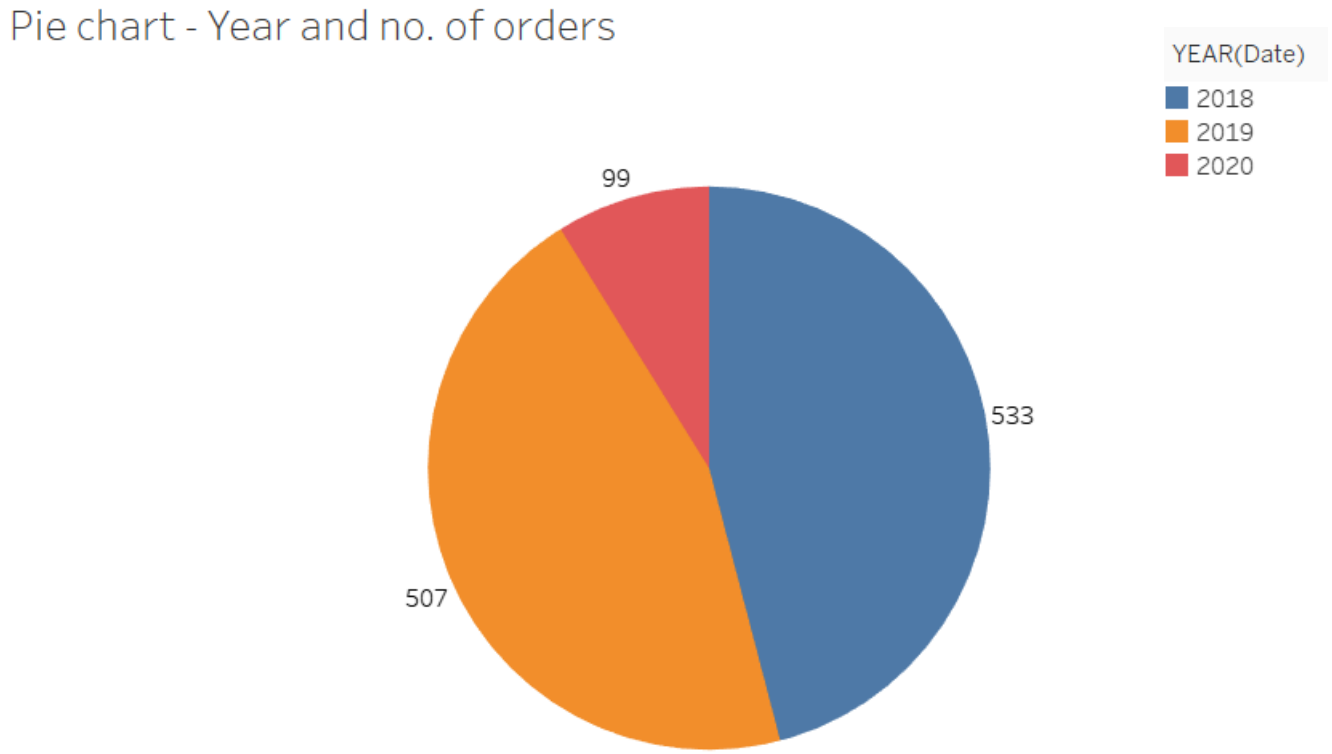
```
The total number of unique dates are 603

The total number of unique order id are 1139

The total number of unique products are 37

array(['yogurt', 'pork', 'sandwich bags', 'lunch meat', 'all- purpose',
       'flour', 'soda', 'butter', 'beef', 'aluminum foil', 'dinner rolls',
       'shampoo', 'mixes', 'soap', 'laundry detergent', 'ice cream',
       'toilet paper', 'hand soap', 'waffles', 'cheeses', 'milk',
       'dishwashing liquid/detergent', 'individual meals', 'cereals',
       'tortillas', 'spaghetti sauce', 'ketchup', 'sandwich loaves',
       'poultry', 'bagels', 'eggs', 'juice', 'pasta', 'paper towels',
       'coffee/tea', 'fruits', 'sugar'], dtype=object)
```

- There were 20641 transactions done.
- However, looking at the data, there are
  - 603 unique dates
  - 1139 unique order ids
  - 37 unique products (The 37 unique products are shown on the left)

Exploratory Data Analysis and Inferences

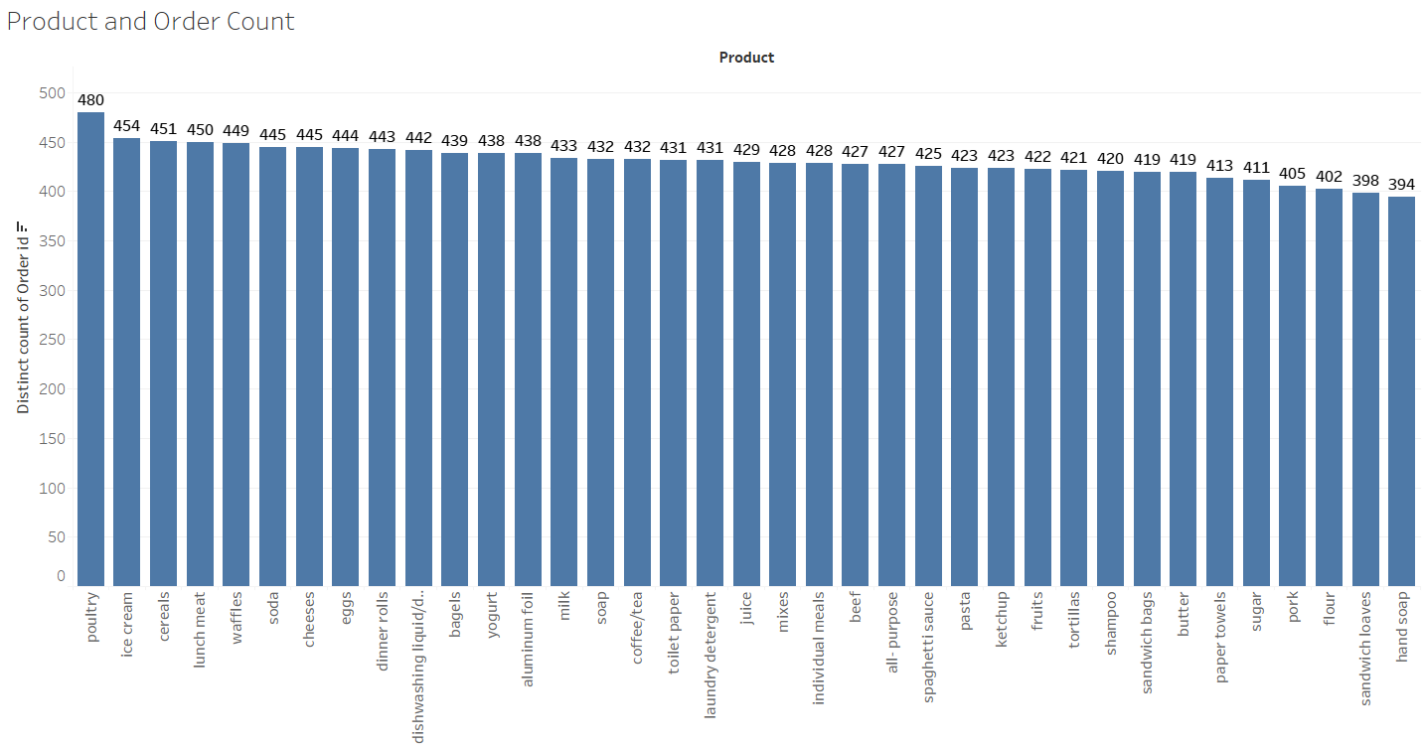
Pie chart of Year and No. of Orders

- 2018 has the highest number of orders (533) followed by 2019 (507 orders).
- 2020 has the least number of orders since the data is for 2 months (January and February) only.



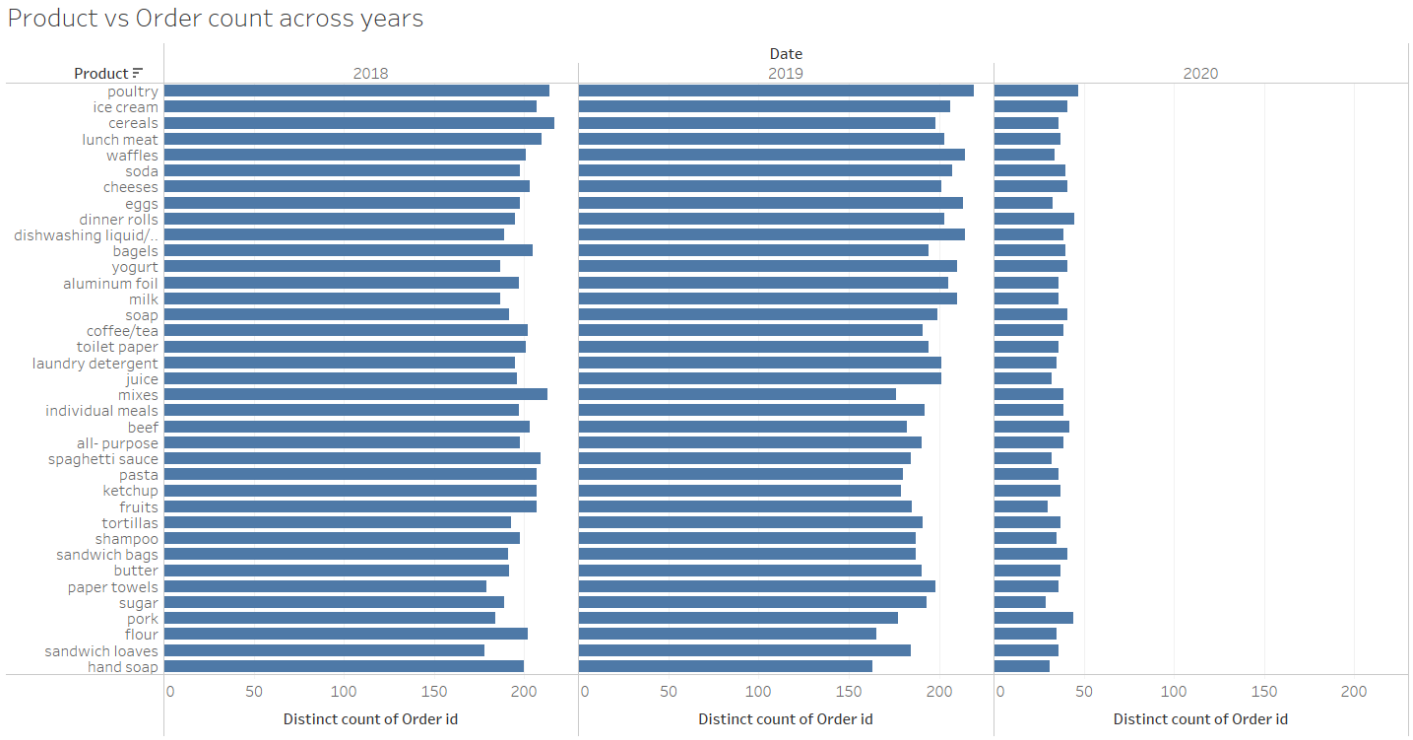
Product and No. of Orders

- Poultry is the most ordered product. It has 480 orders.
- Poultry is followed by soda which has 454 orders.
- The product with the least number of orders is hand shop.



Product vs Order count across years

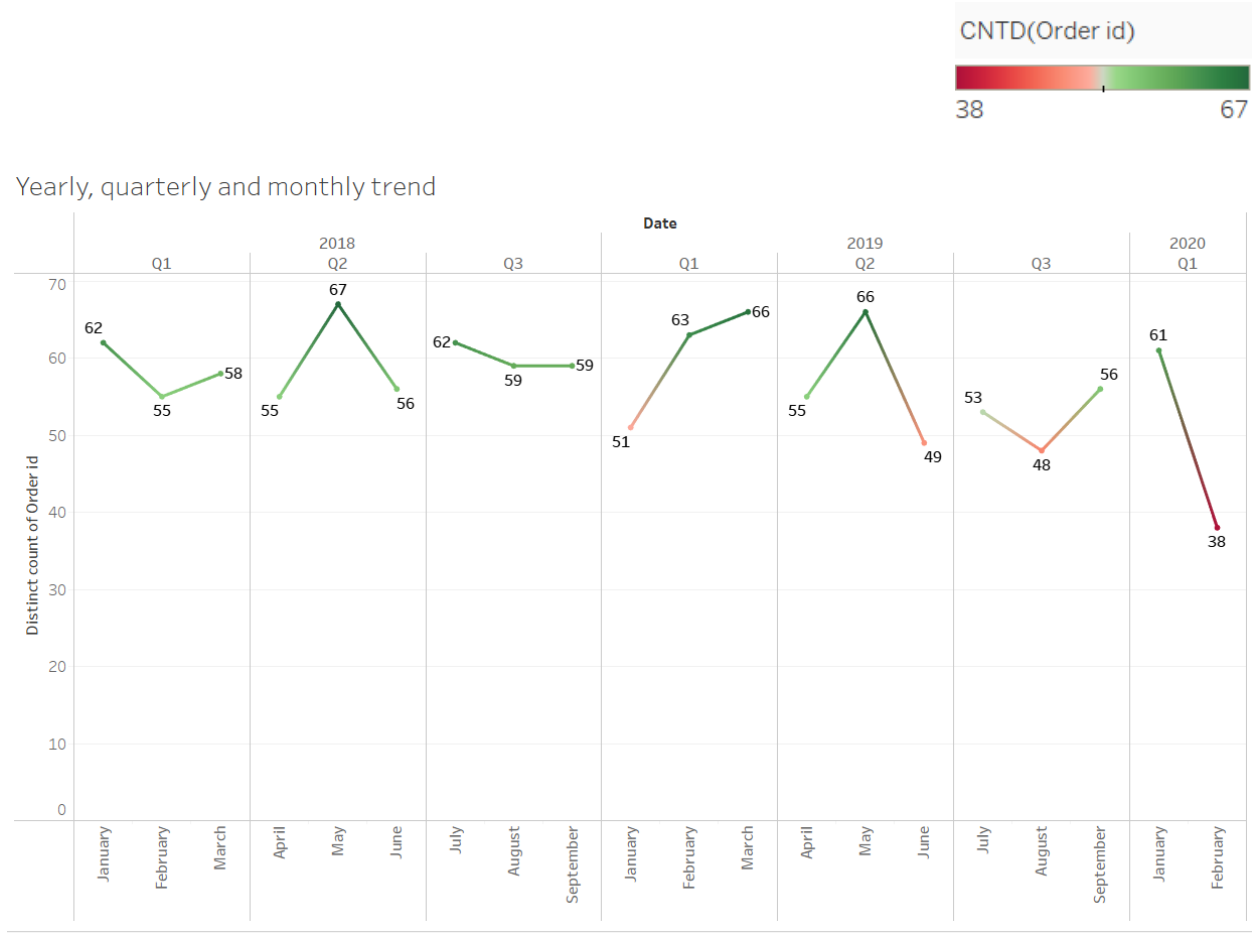
- In 2018, cereal is the most sold product is cereal and sandwich loaves is the least sold product.
- In 2019, the most sold product is poultry and the least sold product is hand soap.
- In 2020, poultry is the most sold product and sugar is the least sold product.



Weekly, Monthly, Quarterly, Yearly Trends in Sales

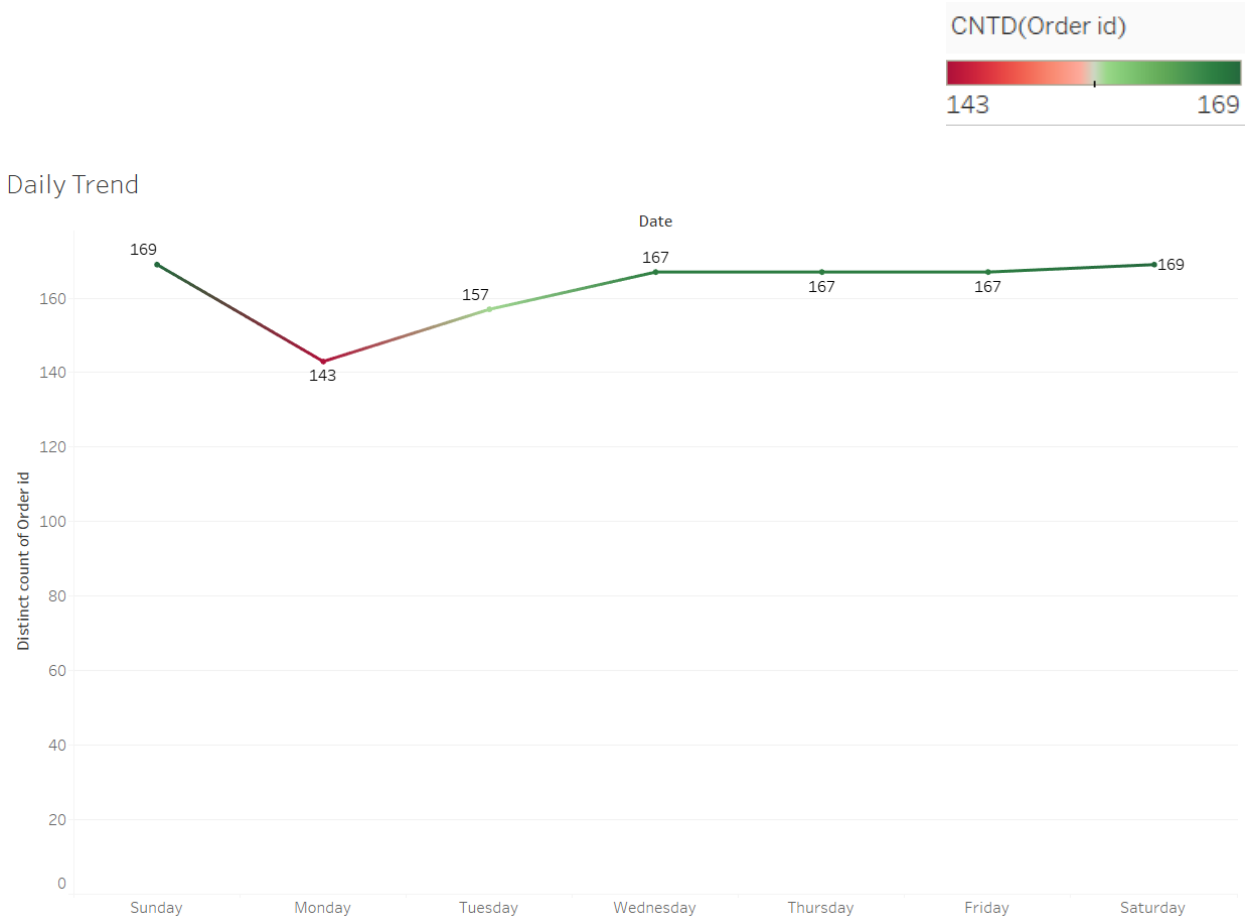
Trend in Count of Orders

- In 2018, May has the highest number of orders.
- In 2019, March and May have the highest number of orders compared to the other months.
- In 2020, highest number of orders are in January.



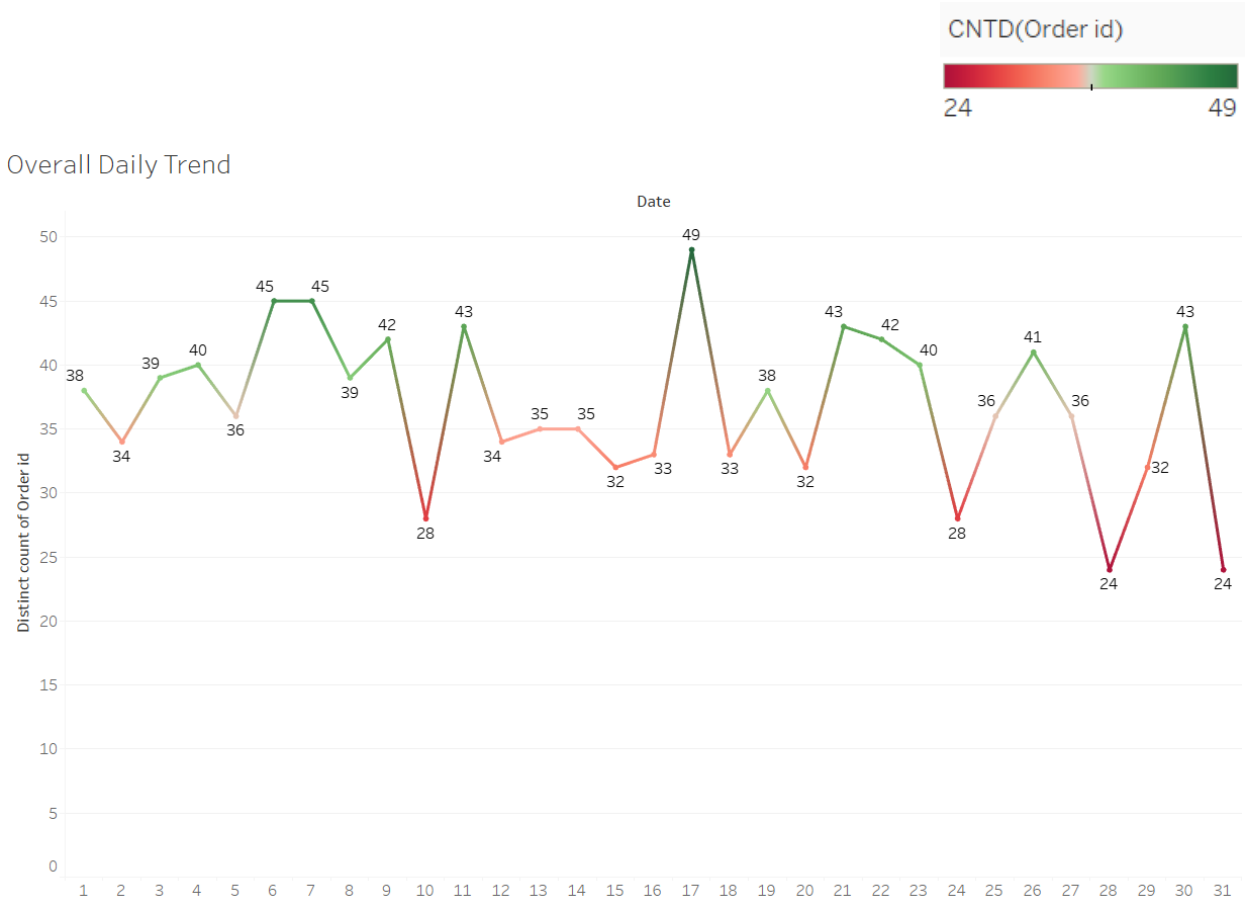
Daily Trend in Count of Orders

- Monday has the lowest count of orders compared to all other days. This maybe because people stock up the groceries during the weekend.
- Saturday and Sunday have the highest count of orders followed by Wednesday and Friday. This maybe because people work from Monday to Friday and do their shopping leisurely during their days off.



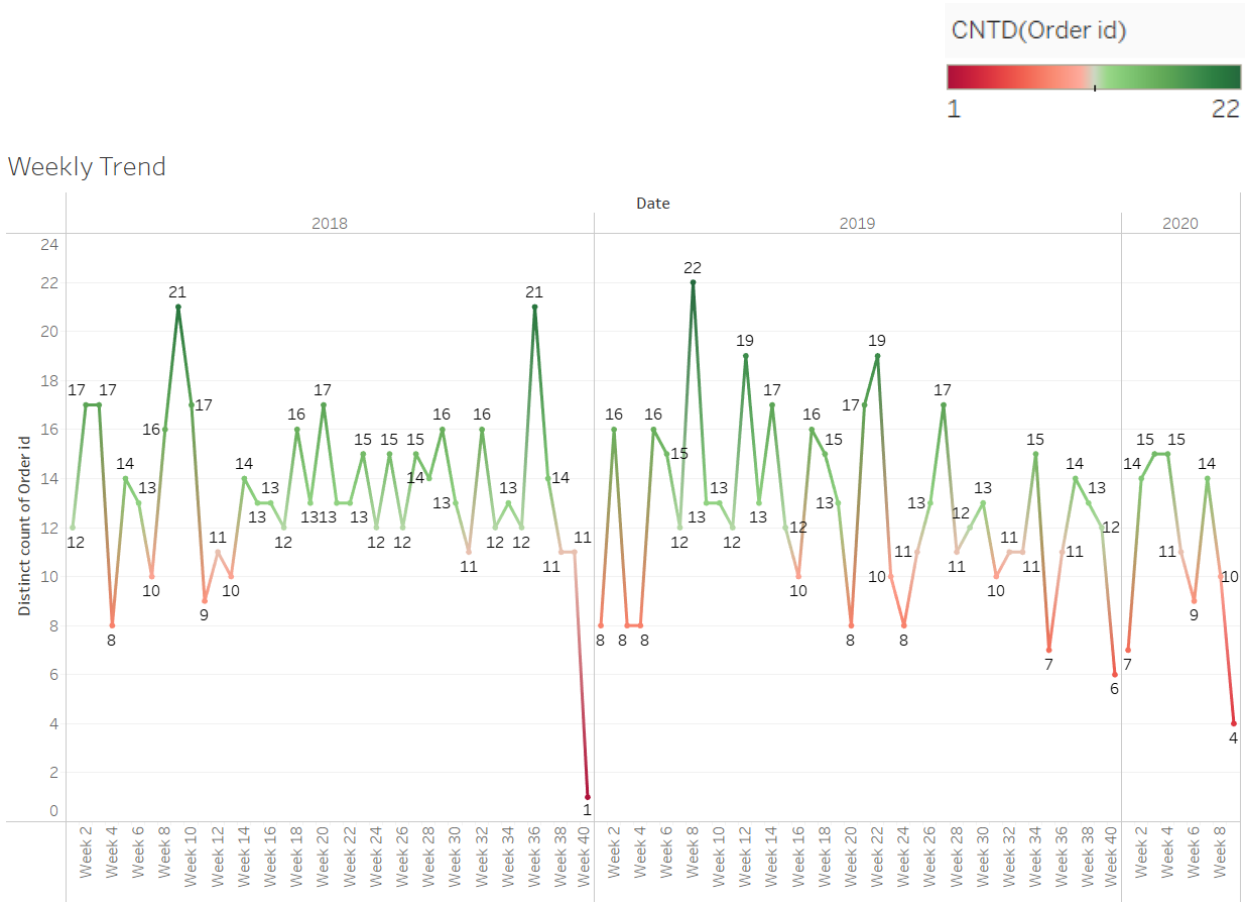
Overall Daily Trend in Count of Orders

- From the trend seen in the graph, we can notice that the highest number of orders are made in the mid of the month.
- The number of orders are low in the starting and end of the month.



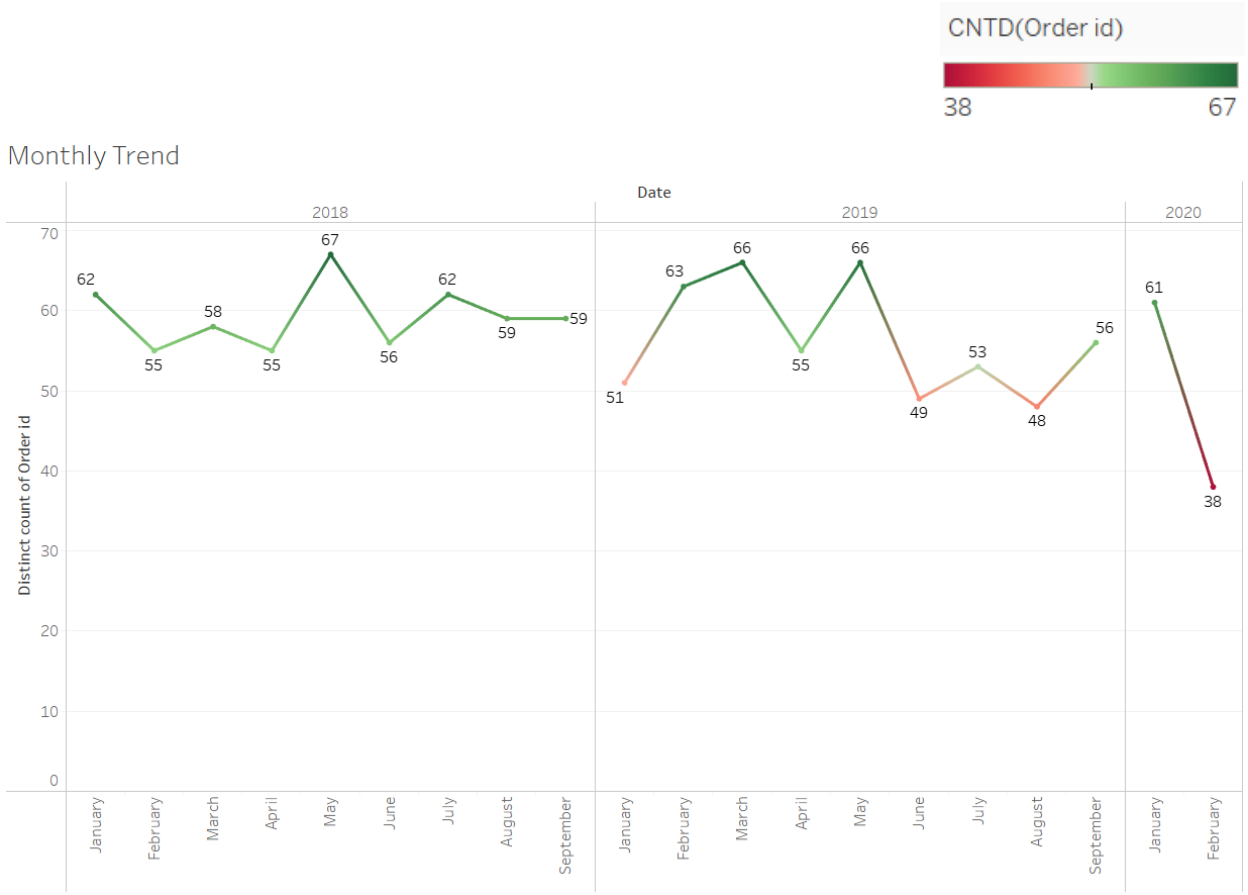
Weekly Trend in Count of Orders

- Week 9 and 36 has the highest count of orders compared to all other weeks in 2018.
- In 2019, week 8 has the highest number of orders.



Monthly Trend in Count of Orders

- In 2018, there are fluctuations in the number of orders from January to April after which there is a sudden increase in May followed by fluctuations again.
- In 2019, the number of orders peak in March and May.
- In 2020, the number of orders is highest in January.
- We also notice that there is no trend or seasonality in the data.



Quarterly Trend in Count of Orders

- In 2018, the highest number of orders were in the 3rd quarter.
- In 2019, the opposite was observed. The highest number of orders were in the 1st quarter.
- There is a downward or decreasing trend from the 1st to 3rd quarter in 2019.
- In 2020, we only have data till February. The number of orders in Q1 is 99.

Yearly Trend in Count of Orders

- 2018 has the highest number of orders.

Summary of the inferences from the above analysis

- The data is provided from January to September for 2 years (2018 and 2019) and January to February for 1 year (2020).
- Because full year data is not available, the three years cannot be accounted for and compared. It will not be an apple to apple comparison.
- However, when comparing these three years, 2018 appears to be the best.
- The year 2020 has the lowest data, which could explain why the number of orders are low this year.
- Because the year 2020 does not have a whole quarter, it is possible that the number of orders is the lowest in contrast to the other years' quarters.
- The daily trend may be related to a person's work day, as weekday orders are lower than weekend orders. Saturday and Sunday has the highest amount of orders.
- The dip in the number of orders at the end of the month could be due to employee pay cycles, as pay day is normally in the first week of the year.
- There appears to be a pick-up in the number of orders around the middle of the month, as we witness an increase in orders numbers.

Market Basket Analysis

What is Market Basket Analysis?

- Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.
- It is a prediction of what a customer would buy based on what the customer has bought in the past. Based on this the company can entice the customer with coupons, offers etc., It is based on conditional probability.
- It is one of the key techniques used by large retailers to uncover associations between and also the strength of association between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships and patterns of co-occurrence between the items that people buy. (A co-occurrence is when two or more things take place together).

Association Rules and their relevance

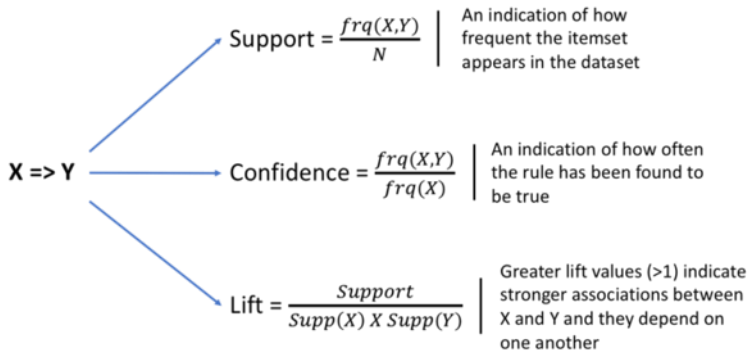
- Association Rules are ‘if-then’ statements that help to show probability of relationships between data items. For example, if an item A is purchased then item B is likely to be purchased.
- These rules are derived from the frequencies of co-occurrence in the observations (frequency is the proportion of baskets that contain the items of interest).
- Market Basket Analysis takes data at transactional level which lists all items bought by a customer in a single purchase. The technique determines relationships of what products were purchased with which other products. These relationships are then used to build profiles containing If-then rules of the items purchased. The rules can be written as: If {A} then {B}.
- Association rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.
- These rules can be used in pricing strategies, placement or arrangement of products in a store and various types of cross-selling strategies.
- The association rule has three measures that express the degree of confidence in the rule – support, confidence and lift.

Threshold values

Support: It is the default popularity of an item. In mathematical terms, the support of item A is nothing but the ratio of transactions involving A to the total number of transactions.

Confidence: Likelihood that the customer who bought both A and B. it divides the number of transactions involving both A and B by the number of transactions involving B.

Lift: Increase in the sale of A when you sell B.

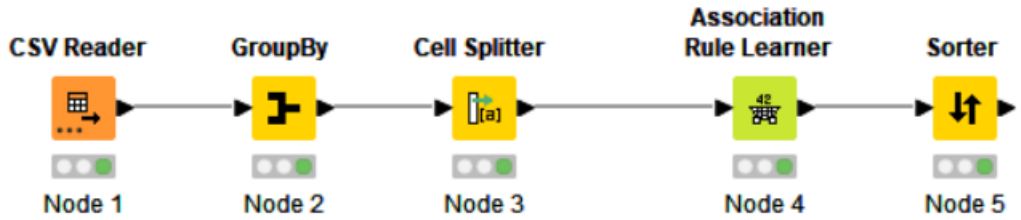


Association Rule Output

Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule0	0.031	0.479	1.247	yogurt	<---	[shampoo,pork,sandwich bags]
rule1	0.031	0.461	1.295	pork	<---	[yogurt,shampoo,sandwich bags]
rule2	0.031	0.449	1.22	sandwich bags	<---	[yogurt,shampoo,pork]
rule3	0.031	0.486	1.318	shampoo	<---	[yogurt,pork,sandwich bags]
rule4	0.031	0.486	1.264	yogurt	<---	[cereals,pork,sandwich bags]
rule5	0.031	0.417	1.172	pork	<---	[yogurt,cereals,sandwich bags]
rule6	0.031	0.455	1.236	sandwich bags	<---	[yogurt,cereals,pork]
rule7	0.031	0.486	1.228	cereals	<---	[yogurt,pork,sandwich bags]
rule8	0.031	0.412	1.071	yogurt	<---	[poultry,pork,sandwich bags]
rule9	0.031	0.407	1.145	pork	<---	[yogurt,poultry,sandwich bags]
rule10	0.031	0.438	1.189	sandwich bags	<---	[yogurt,poultry,pork]
rule11	0.031	0.486	1.154	poultry	<---	[yogurt,pork,sandwich bags]
rule12	0.031	0.427	1.11	yogurt	<---	[tortillas,lunch meat,pork]
rule13	0.031	0.398	1.119	pork	<---	[yogurt,tortillas,lunch meat]
rule14	0.031	0.5	1.266	lunch meat	<---	[yogurt,tortillas,pork]
rule15	0.031	0.507	1.372	tortillas	<---	[yogurt,lunch meat,pork]
rule16	0.031	0.461	1.198	yogurt	<---	[butter,all-purpose,pork]
rule17	0.031	0.473	1.33	pork	<---	[yogurt,butter,all-purpose]
rule18	0.031	0.538	1.436	all-purpose	<---	[yogurt,butter,pork]
rule19	0.031	0.443	1.204	butter	<---	[yogurt,all-purpose,pork]
rule20	0.031	0.493	1.282	yogurt	<---	[all-purpose,individual meals,pork]
rule21	0.031	0.449	1.262	pork	<---	[yogurt,all-purpose,individual meals]
rule22	0.031	0.507	1.353	all-purpose	<---	[yogurt,individual meals,pork]
rule23	0.031	0.443	1.179	individual meals	<---	[yogurt,all-purpose,pork]
rule24	0.031	0.507	1.319	yogurt	<---	[all-purpose,cereals,pork]
rule25	0.031	0.407	1.145	pork	<---	[yogurt,all-purpose,cereals]
rule26	0.031	0.455	1.212	all-purpose	<---	[yogurt,cereals,pork]
rule27	0.031	0.443	1.119	cereals	<---	[yogurt,all-purpose,pork]
rule28	0.031	0.467	1.214	yogurt	<---	[flour,pork,juice]
rule29	0.031	0.443	1.246	pork	<---	[yogurt,flour,juice]
rule30	0.031	0.398	1.127	flour	<---	[yogurt,pork,juice]
rule31	0.031	0.486	1.291	juice	<---	[yogurt,flour,pork]
rule32	0.031	0.467	1.214	yogurt	<---	[mixes,pork,soda]
rule33	0.031	0.422	1.186	pork	<---	[yogurt,mixes,soda]
rule34	0.031	0.461	1.179	soda	<---	[yogurt,mixes,pork]
rule35	0.031	0.449	1.194	mixes	<---	[yogurt,pork,soda]
rule36	0.031	0.493	1.282	yogurt	<---	[toilet paper,pork,soda]
rule37	0.031	0.398	1.119	pork	<---	[yogurt,toilet paper,soda]

KNIME workflow

KNIME is used to perform the MBA Analysis and the workflow diagram is shown on the right.



Threshold values taken

The association rules are actionable. These rules can be used to target customers in the form of marketing, placing of products in the supermarkets (common combinations of products can be placed close to each other) and to make decisions about offering discounts, combo offers etc.,

Due to this, the ‘association rule learner’ is the most important node in Market Basket Analysis. Determining the value of the thresholds is an iterative process. It is a balance between getting more volume of data at the cost of accuracy.

Support and confidence are used to create thresholds.

The values will change depending on the business scenario, the dataset, etc., The value of minimum support and minimum confidence was found by starting the threshold with a value of 0.9 and reducing the values if enough rules isn’t got.

In this case study, the initial value was taken to be 0.9 and the optimum value for **minimum support** was found to be **0.03** i.e., 3% sell of a product from overall transactions and optimum value for **minimum confidence** was found to be **0.3**.

Association Rule Output sorted by Lift values

Row ID	D Support	D Confidence	D Lift	S Consequent	S implies	[...] Items
rule18260	0.031	0.795	2.194	paper towels	<---	[eggs,ice cream,pasta,...]
rule35220	0.032	0.783	2.158	paper towels	<---	[eggs,ice cream,pasta,...]
rule18261	0.031	0.729	2.066	flour	<---	[dishwashing liquid/detergent,cheeses,waffles,...]
rule51470	0.032	0.74	2.041	paper towels	<---	[eggs,dinner rolls,ice cream,...]
rule35225	0.032	0.72	1.986	paper towels	<---	[eggs,poultry,ice cream,...]
rule18257	0.031	0.778	1.951	ice cream	<---	[paper towels,eggs,pasta,...]
rule18262	0.031	0.761	1.947	soda	<---	[dishwashing liquid/detergent,cheeses,flour,...]
rule66274	0.033	0.717	1.931	pasta	<---	[paper towels,dishwashing liquid/detergent,eggs,...]
rule125779	0.04	0.697	1.922	paper towels	<---	[all- purpose,individual meals,toilet paper]
rule18273	0.031	0.714	1.914	spaghetti sauce	<---	[dinner rolls,poultry,laundry detergent,...]
rule66273	0.033	0.745	1.911	eggs	<---	[paper towels,dishwashing liquid/detergent,ice cream,...]
rule66275	0.033	0.691	1.905	paper towels	<---	[dishwashing liquid/detergent,eggs,ice cream,...]
rule35224	0.032	0.706	1.901	pasta	<---	[paper towels,eggs,poultry,...]
rule35223	0.032	0.72	1.847	eggs	<---	[paper towels,poultry,ice cream,...]
rule51469	0.032	0.685	1.845	pasta	<---	[paper towels,eggs,dinner rolls,...]
rule94971	0.036	0.641	1.833	sandwich loaves	<---	[all- purpose,flour,individual meals]
rule51468	0.032	0.712	1.825	eggs	<---	[paper towels,dinner rolls,ice cream,...]
rule125903	0.04	0.676	1.822	pasta	<---	[hand soap,soda,aluminum foil]
rule126175	0.04	0.676	1.822	ketchup	<---	[butter,aluminum foil,soap]
rule111902	0.038	0.632	1.81	sandwich loaves	<---	[paper towels,flour,individual meals]
rule127779	0.041	0.671	1.808	ketchup	<---	[pork,sandwich bags,soap]
rule67559	0.034	0.629	1.8	sandwich loaves	<---	[yogurt,hand soap,soap]
rule18269	0.031	0.7	1.796	eggs	<---	[dishwashing liquid/detergent,ice cream,pasta,...]
rule133522	0.046	0.65	1.793	paper towels	<---	[ice cream,pasta,lunch meat]
rule135467	0.055	0.649	1.791	paper towels	<---	[eggs,ice cream,pasta]
rule110607	0.038	0.662	1.786	fruits	<---	[all- purpose,beef,lunch meat]
rule120797	0.039	0.688	1.784	bagels	<---	[sandwich loaves,fruits,juice]
rule132469	0.044	0.676	1.781	soap	<---	[sandwich loaves,all- purpose,ketchup]
rule35218	0.032	0.692	1.776	eggs	<---	[paper towels,ice cream,pasta,...]
rule57628	0.033	0.667	1.774	mixes	<---	[all- purpose,hand soap,tortillas]
rule110034	0.038	0.672	1.767	milk	<---	[sandwich loaves,pork,soda]
rule111836	0.038	0.623	1.766	flour	<---	[pasta,mixes,coffee/tea]
rule35219	0.032	0.655	1.762	pasta	<---	[paper towels,eggs,ice cream,...]
rule115095	0.038	0.614	1.758	sandwich loaves	<---	[cheeses,hand soap,ketchup]
rule18264	0.031	0.686	1.757	cheeses	<---	[dishwashing liquid/detergent,flour,waffles,...]
rule124453	0.04	0.652	1.756	ketchup	<---	[tortillas,coffee/tea,juice]
rule132153	0.044	0.617	1.749	flour	<---	[yogurt,pasta,coffee/tea]
rule130627	0.042	0.649	1.747	pasta	<---	[dinner rolls,hand soap,individual meals]

Associations Identified

The association rules output is shown previously. The table shows 158925 records in which each row contains different rules.

These rules were created on the basis of the threshold values of minimum support and minimum confidence that were found and optimized in the ‘Association Rule Learner’ node.

The product with higher lift value is the most recommended product to the customer. Therefore, the table is sorted based on the lift values and shown previously.

The consequent column contains recommended products and the table is sorted according to the lift values in descending order to find the recommendations.

A suggestion of possible combos with lucrative offers

When a consumer purchases eggs, ice cream, pasta and lunch meat there is a probability that the customer will also buy paper towels. Therefore, the company can offer this combo along with paper towels at a confidence of 0.795.

When a customer buys eggs, ice cream, pasta and cereal there is a probability that the company can offer them along with paper towels at a confidence of 0.783

When a customer buys dishwashing liquid/detergent, cheeses, waffles, soda there is a probability that we can offer them along with flour at a confidence of 0.729

When a customer buys eggs, dinner rolls, ice cream, pasta there is a probability that we can offer them along with paper towel at a confidence of 0.74

Discount offers and combos based on Associations

If a consumer purchases eggs, ice cream, pasta and lunch meat we can offer paper towels for free because in comparison to the other products purchased by the customer, the price of paper towels is relatively less. Therefore in the long run, it is a gain for the business.

If a costumers purchases an order above a certain value, few products which high lift can be offered at a discounted rate or free.

If a customer buys dishwashing liquid/detergent, cheeses, waffles, soda there is a probability that the customer will also buy flour. Therefore, flour can be offered at a discounted rate.

Recommendations

The company can offer discount offers or combos (or buy two get one free) based on the associations.

Generally the products that are listed in consequent feature which has a higher lift value is recommended. That means it has the higher probability of being purchased by the customer.

Products that are likely to be purchased together can be placed close to each other in the supermarket. The combos that has been identified can be placed together and brochures can be handed out to advertise these combo offers.

The combinations of products can be advertised across the city.

The combos can be offered at a discounted rate to attract more customers. This will not only increase the overall order Id count but also increase customers for the company. New customers can be attracted which will increase the business for the company in the long run.

Giving rewards for best customers with better offers in these combos could help the company maintain the customer lifetime value with the company.

TABLEAU Link

https://public.tableau.com/app/profile/athulya4213/viz/MRAProjectMilestone2_16535158945070/YearlyTrend?publish=yes

Thank you