PROJECT REPORT ON MARKET BASKET ANALYSIS

GROCERY STORE

Presented By: -

SHUBAHM KUMAR

Content

- Information about Data
- Exploratory Data Analysis
- Preliminary Inferences from Data
- Market Basket Analysis
- Association Identified
- Inferences – Recommendation - Suggestions

Exploratory Data Analysis (EDA)

Dataset Samples and Basic Information

First 5 rows of Dataset			
Row No.	**Date**	**Order_id**	**Product**
0	1/1/2018	1	yogurt
1	1/1/2018	1	pork
2	1/1/2018	1	sandwich bags
3	1/1/2018	1	lunch meat
4	1/1/2018	1	all- purpose

Bottom 5 rows of Dataset			
Row No.	**Date**	**Order_id**	**Product**
20636	2/25/2020	1138	soda
20637	2/25/2020	1138	paper towels
20638	2/26/2020	1139	soda
20639	2/26/2020	1139	laundry detergent
20640	2/26/2020	1139	shampoo

Information about the dataset

```
<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  20642 non-null  int64
  2  Product   20643 non-null  object
dtypes: datetime64[ns](1), int64(1), object(1)
```

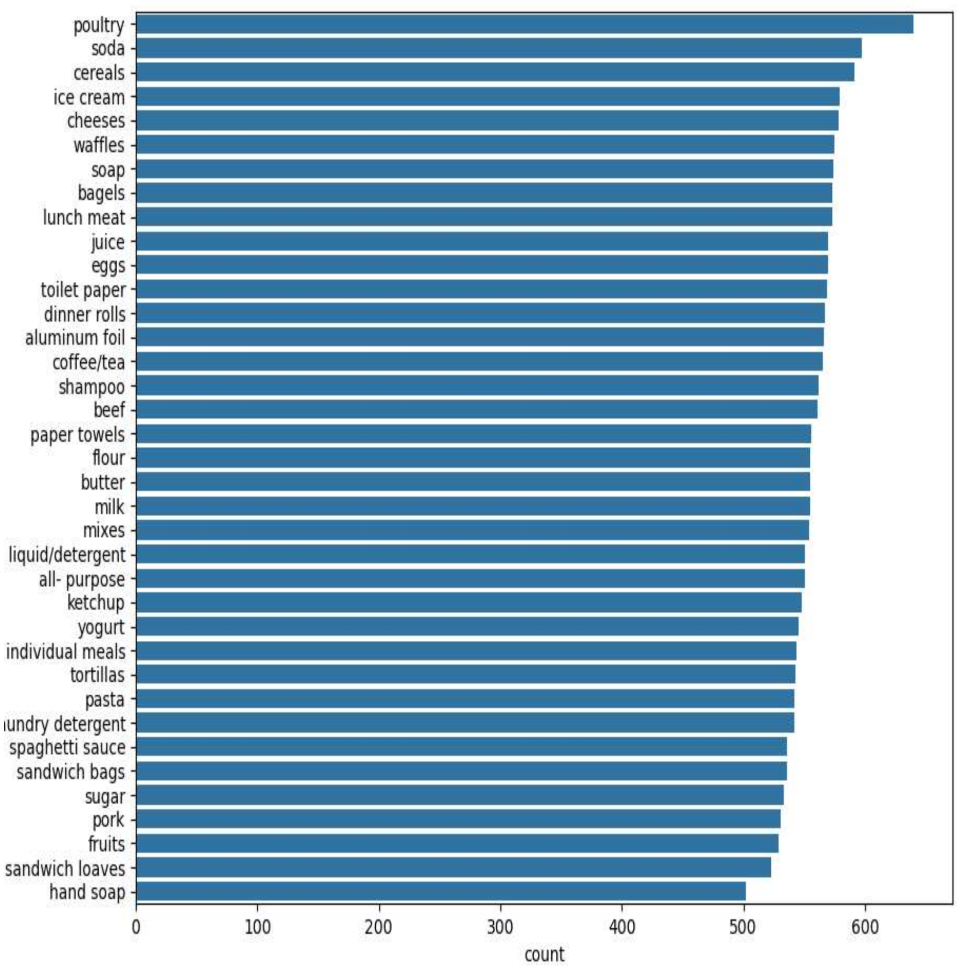
Descriptive Stats of the Numeric Data								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Order_id	20641	575.9863	328.5571	1	292	581	862	1139

Descriptive Stats of the Categorical Data				
	count	**Unique**	**Top**	**freq**
Product	20641	37	Poultry	640

Data Description

- This data set is about the products purchased at the Grocery store by dates across year from 2018 to 2020
- Dataset has 3 columns and 20641 rows
- There are no null entries
- There are 3 datatype in a dataset: -
  - OrderId - Integer
  - Product - Object
  - Date - Datetime
- There are 37 unique Products
- There are 1139 unique OrderId

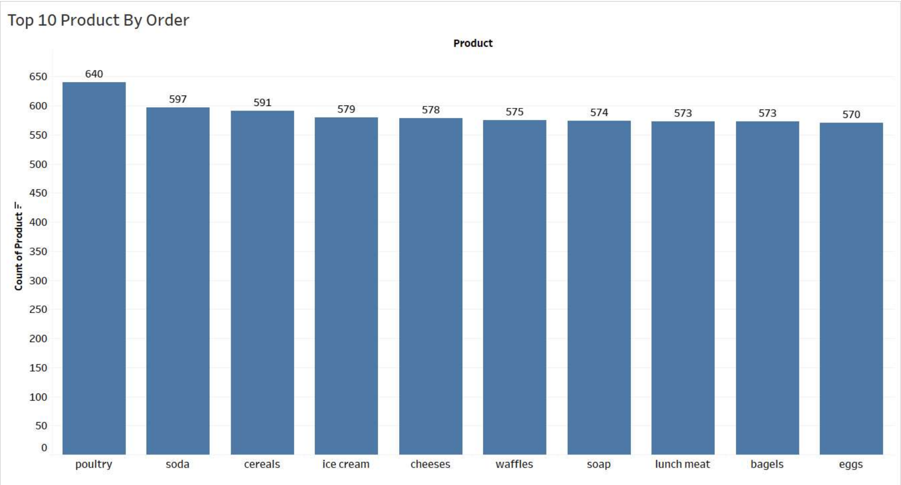
Visualization of the Dataset

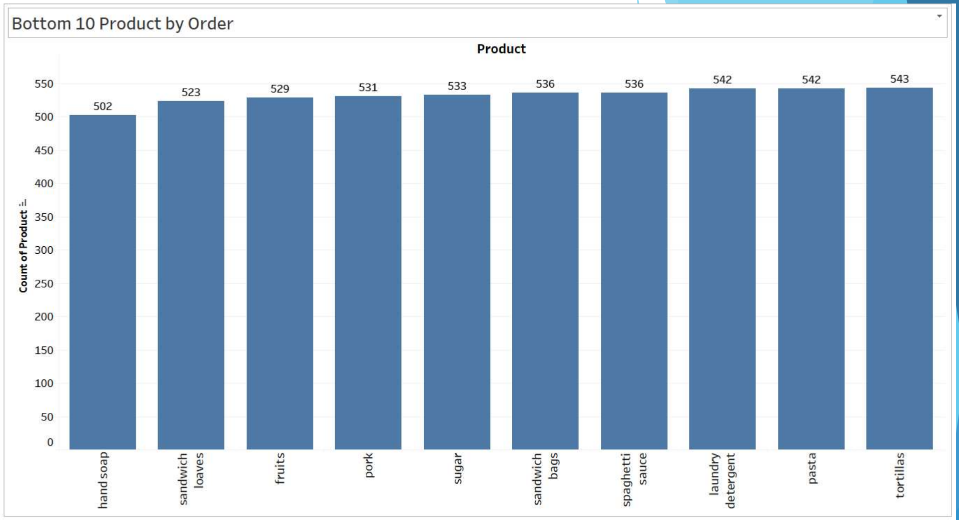


Inferences

- Poultry Products has highest count = 640
- Hand Soap has least count = 502
- There are total 37 products
- None of the products has count less than 500
- Only Poultry Products count is more than 600
- 36 Products are in range of 500 to 600
- Mean Product count = 543

Top & Bottom 10 Products

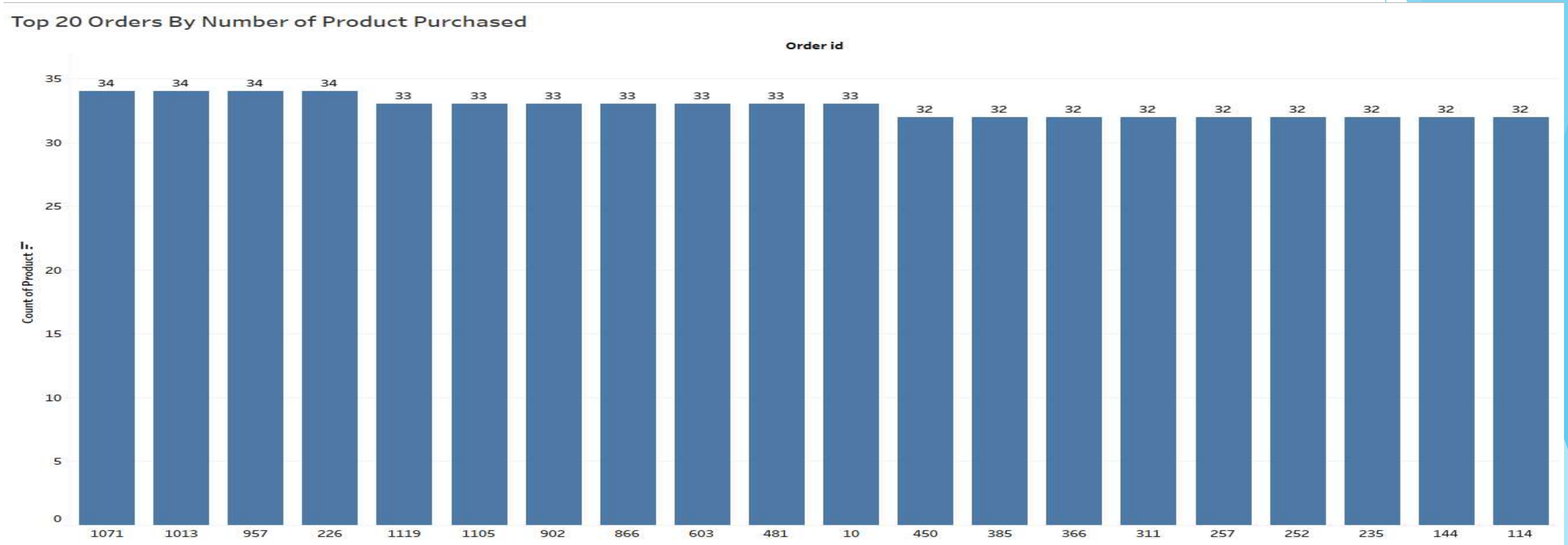




INFERENCES

- Top 10 Products are Poultry, Soda, Cereals, Ice cream, Cheese, Waffles, Soap, Lunch Meat, Bagels and Eggs
- Bottom 10 Products are Hand Soap, Sandwich Loaves, Fruits, Pork, Sugar, Sandwich Bags, Spaghetti Sauce, Laundry Detergent, Pasta and Tortillas

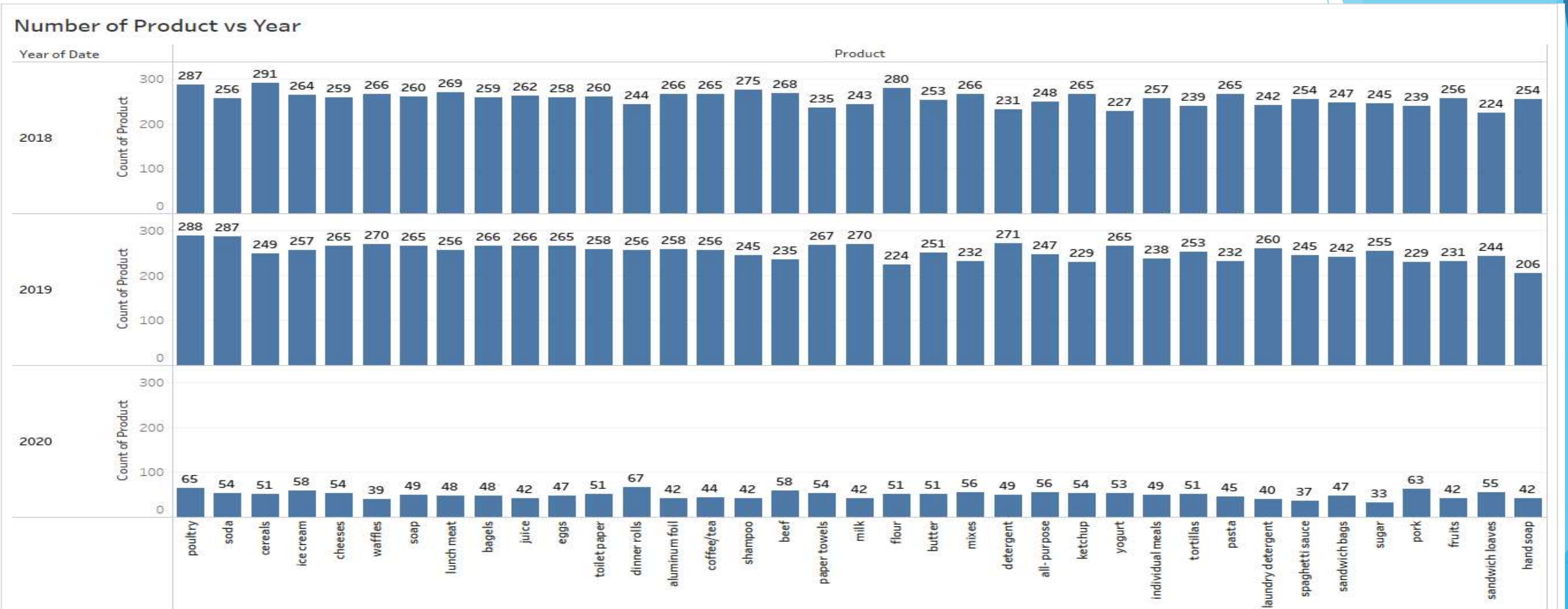
Top 10 Orders By Product Purchased



INFERNCES

- OrderId 1071,1013, 957 & 226 has equal number of product 34 which is also the highest number of Product in a particular OrderId
- Next Highest Number of products Ordered in particular OrderID is 33 followed by 32.

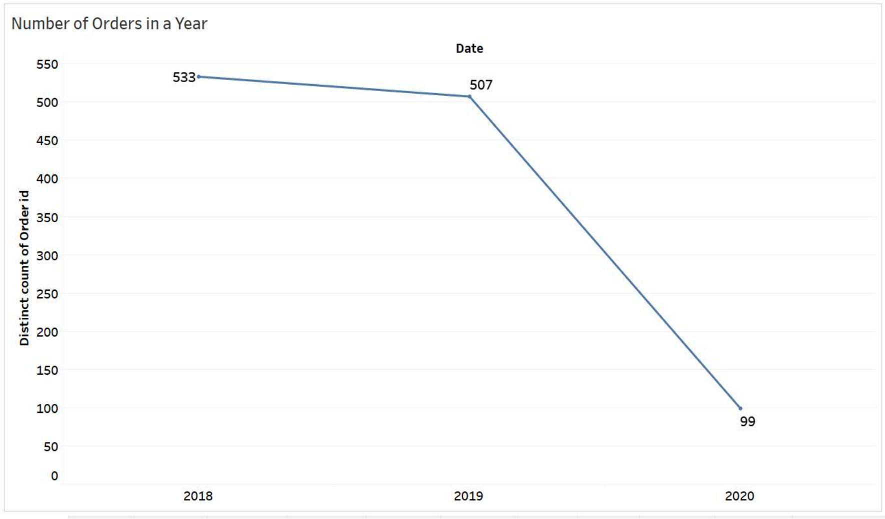
Top Bought Product Every Year



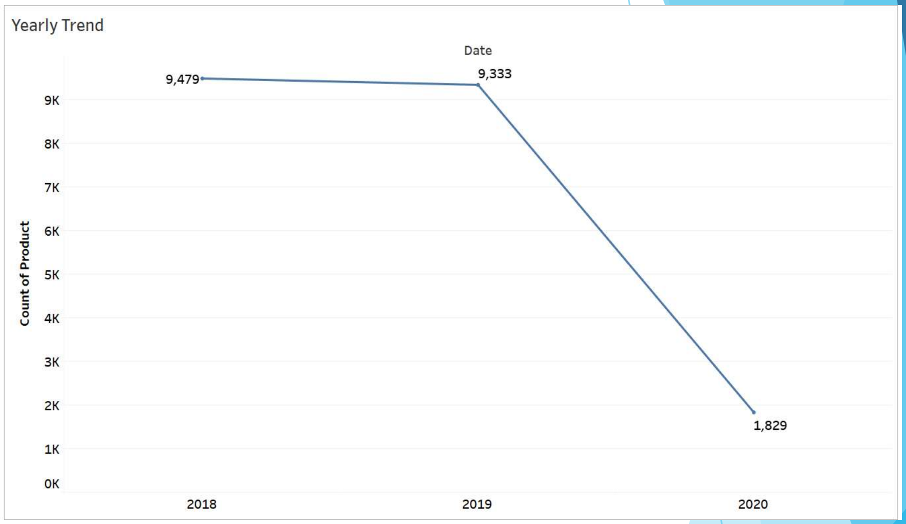
- In 2018, Cereal is bought for most number of times.
- In 2019, Poultry is bought for most number of times.
- In 2020, Dinner Rolls is bought for most number of times.

Yearly Trend

Number of Orders in a Year



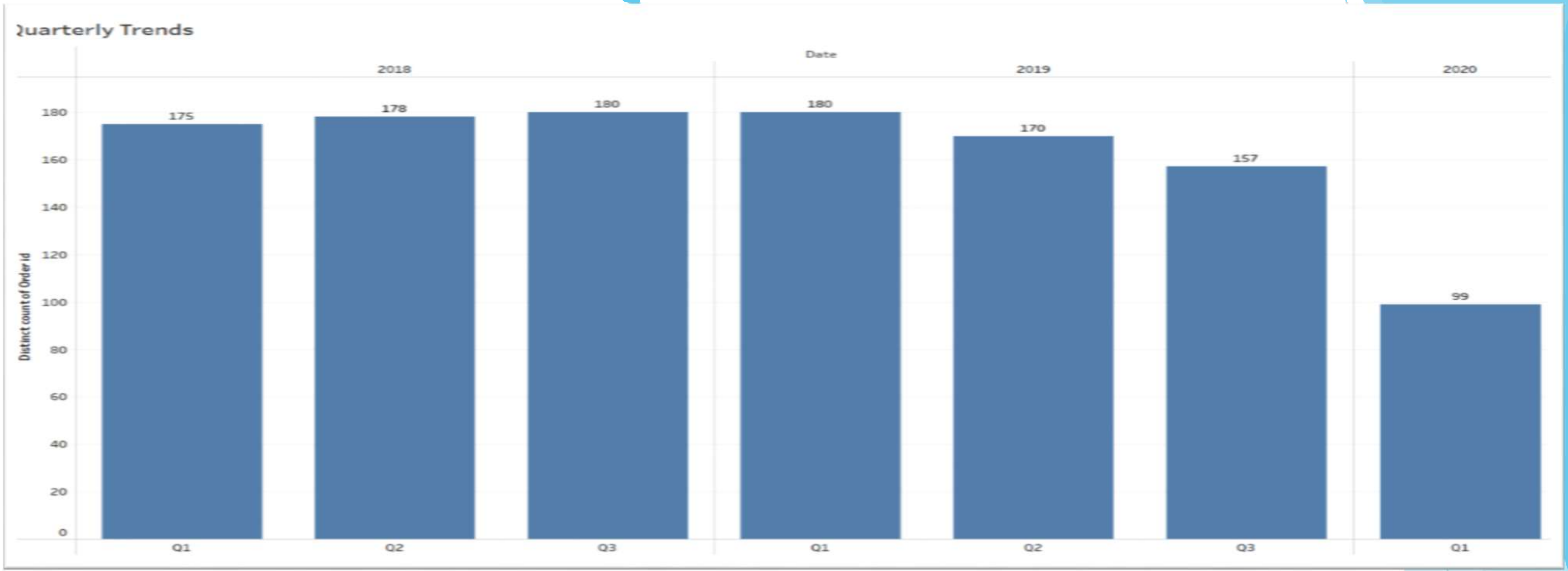
Number of Products sold in a Year



Inferences

- 2018 has highest number of order placed or product purchased.
- Order & Products Purchased has dropped slightly from 2018 to 2019 and sharply from 2019 to 2020.
- Sharp drop in Order or Product Purchased from 2019 to 2020 is due to availability of data. As for 2020 data is available up to month of February Only

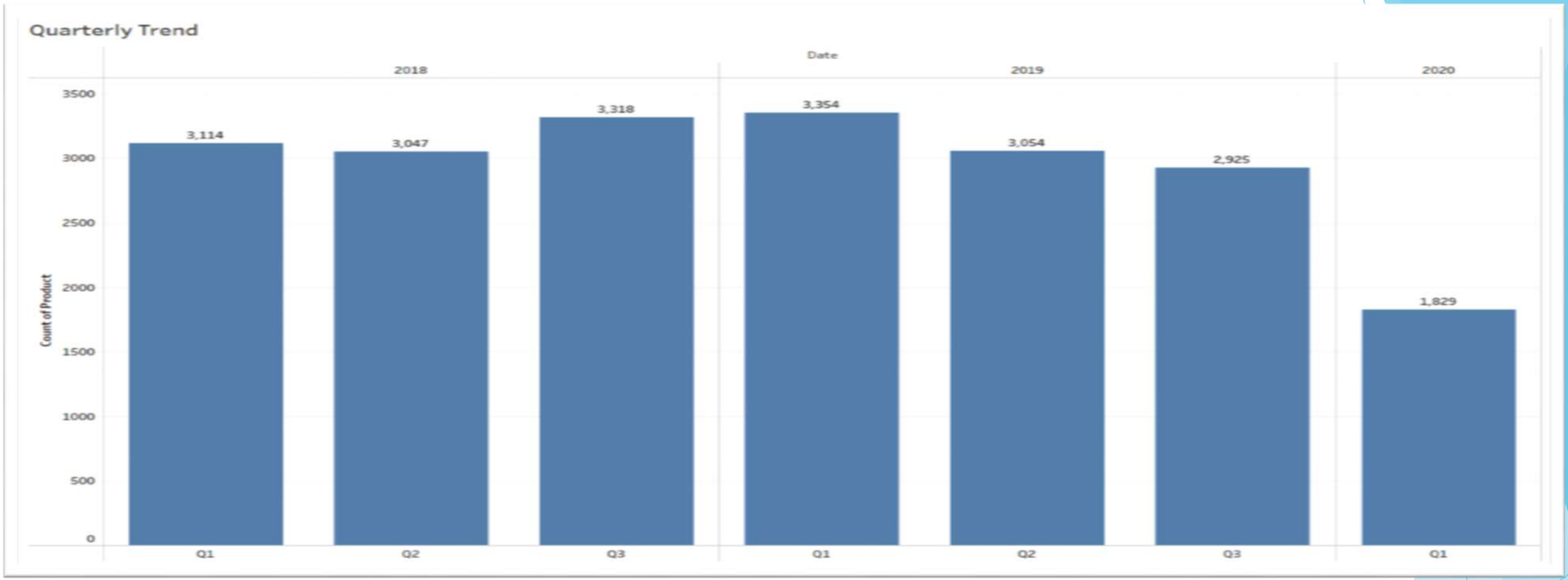
Order Placed in each Quarter



Inferences

- In 2018, Orders increases slightly from 1st Quarter to 3rd Quarter, increasing trend
- In 2019, Orders decreases slightly from 1st Quarter to 2nd Quarter decreasing trend
- In 2020, we only have data for first two months so we cannot capture any Trend

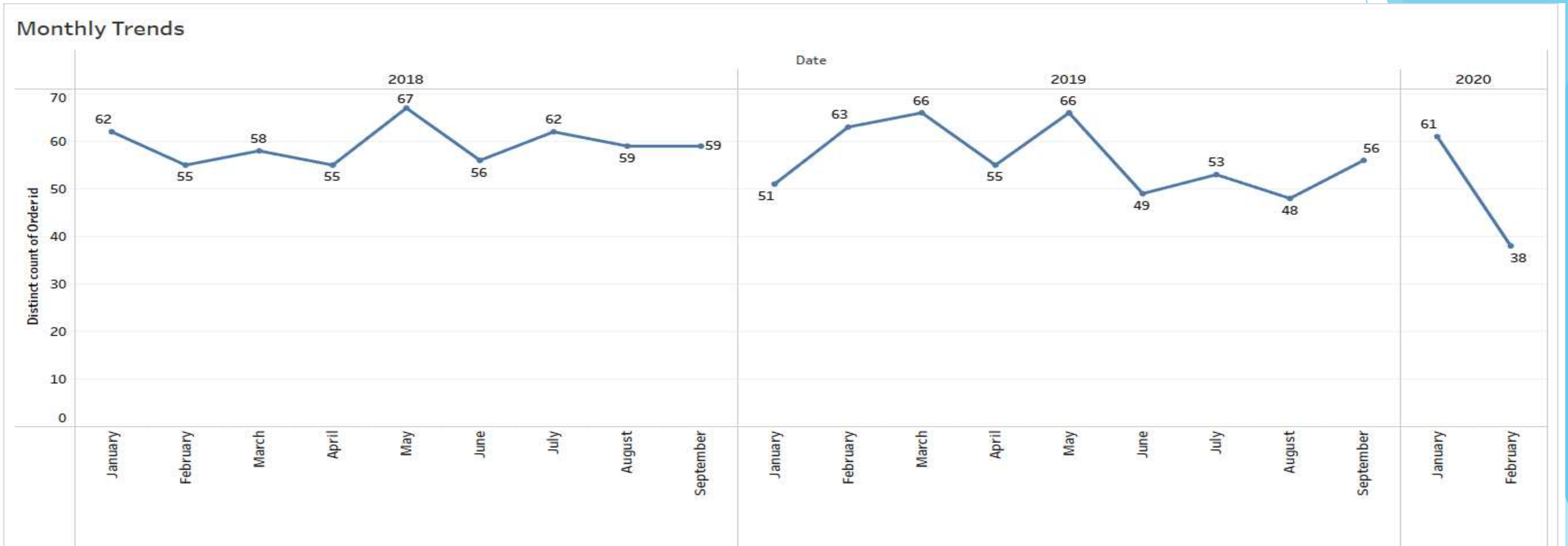
Product Purchased in each Quarterly



Inferences

- In 2018, Orders drops slightly from Q1 to Q2 and increases from Q2 to Q3
- In 2019, Orders decreases slightly from Q1 to Q3, downwards trend
- In 2020, we only have data for first two months so we cannot capture any Trend

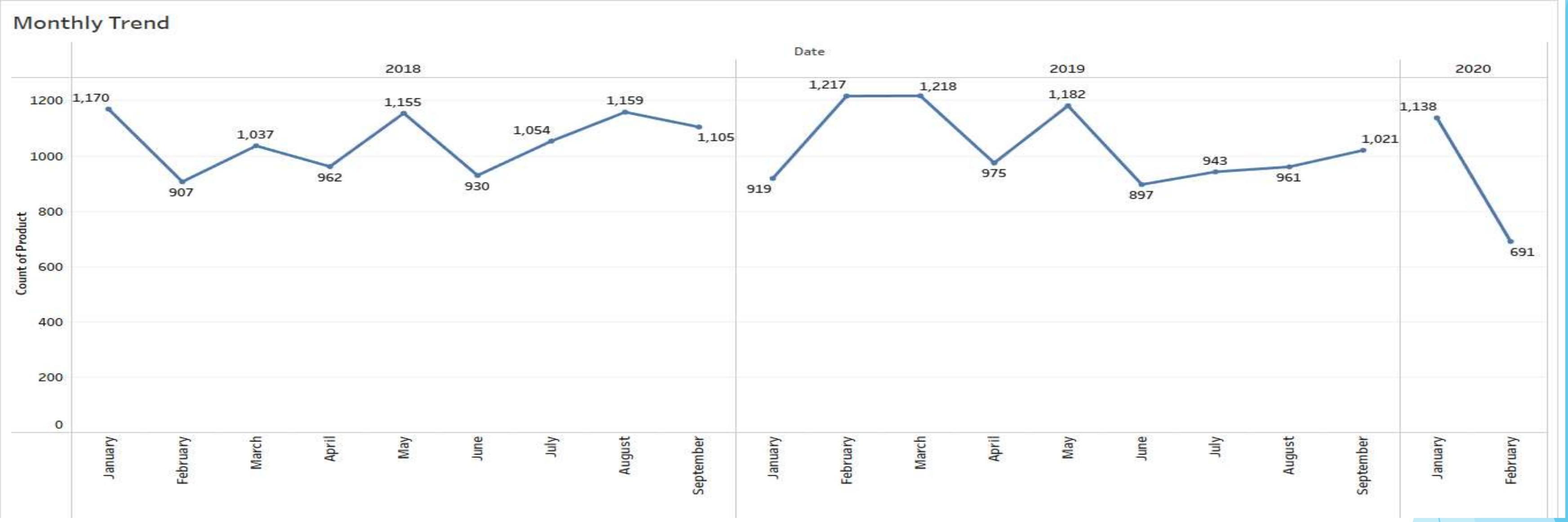
Order Placed in each Month



Inferences

- In 2018 maximum orders came in the month of May. The Orders increases and drops after every consecutive month
- In 2019, Maximum order came in month of March & May. The Orders have increased from January to March and then it drop and increases after every consecutive months

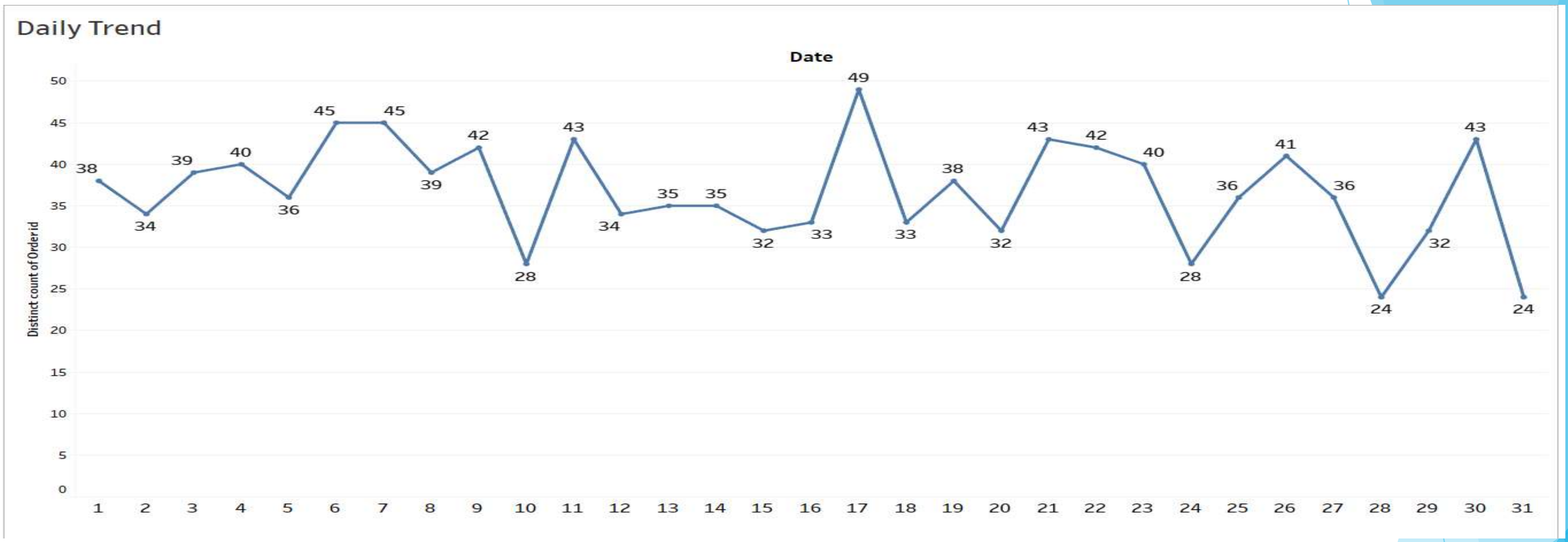
Product Purchased in each Month



Inferences

- In 2018 maximum orders came in the month of January. The Orders increases and drops after every consecutive month
- In 2019, Maximum order came in month of May. We cannot find any particular trend in 2019

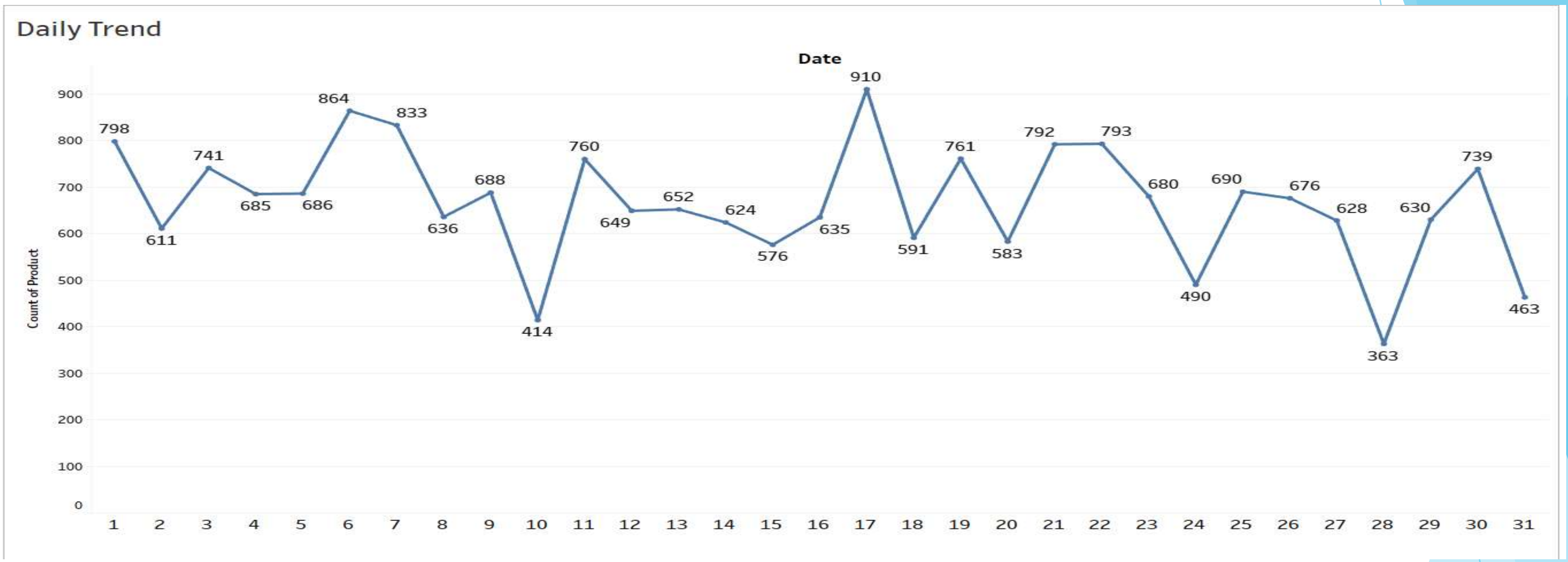
Orders placed in a day



Inferences

- Maximum Orders were placed on 17th in a month i.e., 49 orders
- Least Orders were placed on 28th in a month i.e., 24 orders
- There is no trend observed in Orders Placed on daily basis.

Products Purchased in a day



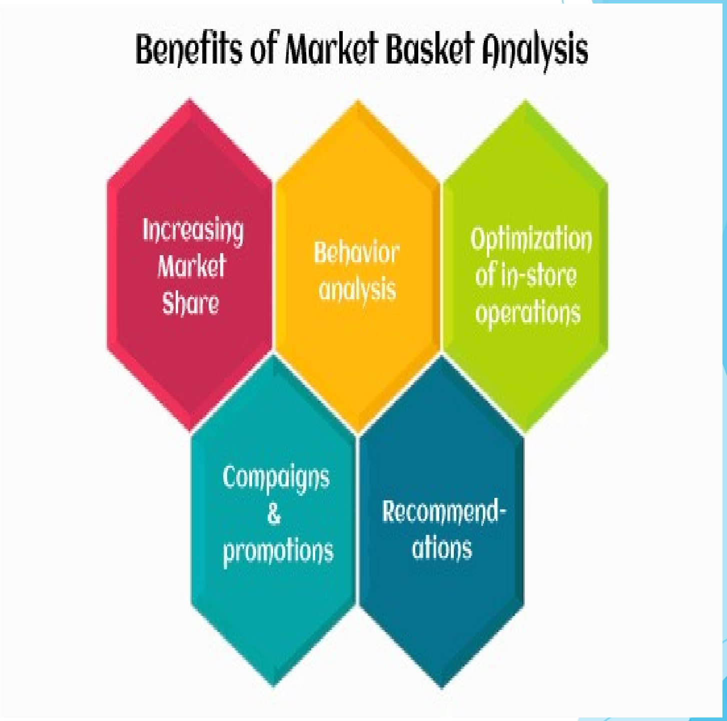
Inferences

- Maximum Orders were placed on 17th of every Month i.e., 910 Products.
- Least Orders were placed on 28th of every Month i.e., 363 Products.
- There is no trend observed in product purchased on daily basis.

MARKET BASKET ANALYSIS (MBA)

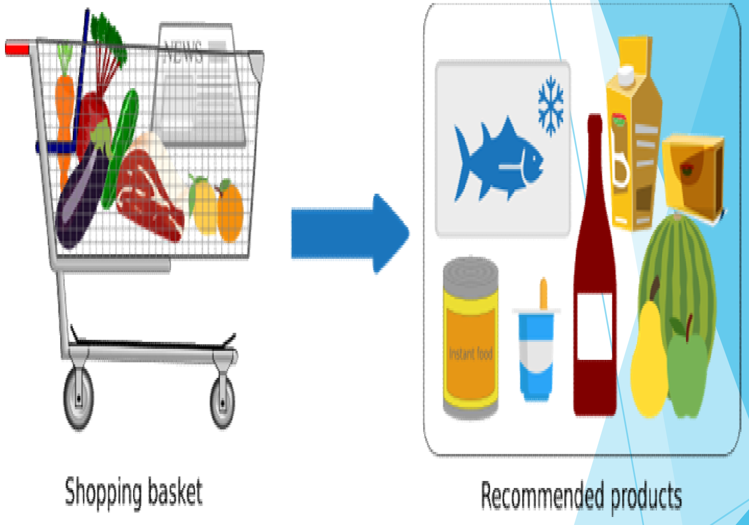
Market Basket Analysis (MBA)

- Market basket analysis in data mining is to analyze the combination of products which been bought together
- In simple terms Basically, Market basket analysis in data mining is to analyze the combination of products which been bought together.
- This concept identifies the pattern of frequent purchase items by customers.
- Market basket analysis mainly works with the ASSOCIATION RULE



Association Rules in Market Basket Analysis

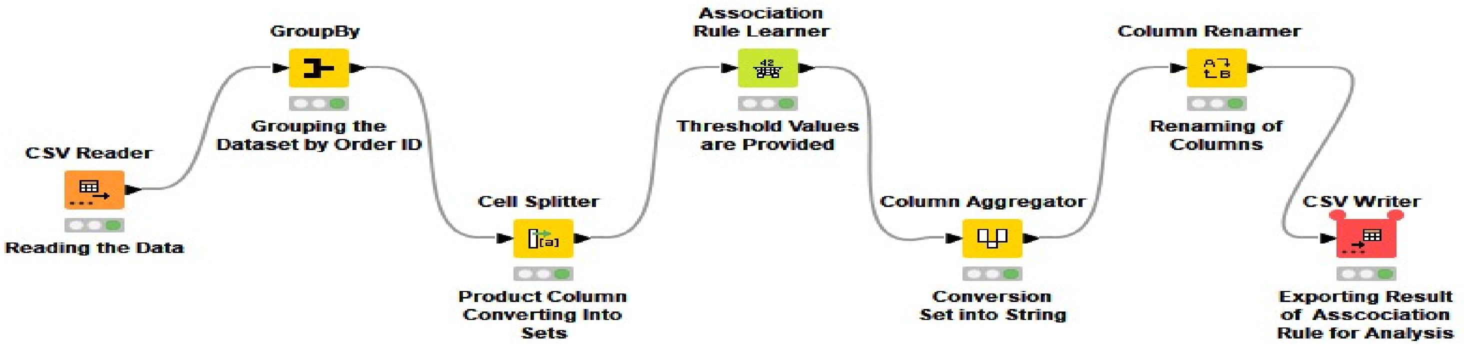
- Association Rule is classified as Unsupervised Learning
- Association Rule in MBA is used to find pattern in transaction data.
- The main concept is based on IF – THEN structure. If A is purchased then B is likely to be bought
- It simply helps in predicting the likelihood of the products being purchased together.
- In this problem, using MBA and Association rules will help finding the best combo and recommendation to increase sales of the grocery store



Keys Points in Association Rule using MBA

- **Support:** - Percentage of transaction or Purchase containing both A and B
  - Support = (A + B) / Total Purchase
- **Confidence:** - Percentage of customer who bought A also bought B.
  - Confidence = combine transactions/individual transactions
  - Confidence = (A + B) / A
- **Lift:** - Lift is calculated for knowing the ratio for the sales
  - Lift = Confidence Percent / Support Percent
  - Lift = [(A + B)/A] / [B / Total Purchase]

KNIME WORKFLOW (MBA using Association Rule)


CSV Reader
Reading the Data
GroupBy
Grouping the Dataset by Order ID
Cell Splitter
Product Column Converting Into Sets
Association Rule Learner
Threshold Values are Provided
Column Aggregator
Conversion Set into String
Column Renamer
Renaming of Columns
CSV Writer
Exporting Result of Asscociation Rule for Analysis

- Support Threshold Value: - 0.05
- Confidence Threshold Value: - 0.50
- Maximum Item set Length:- 3

Association Identified

First 20 Rules Identified

row ID	Support	Confidence	Lift	Recommended item	Recommended with	Items In Basket
rule0	0.064969271	0.506849315	1.202711187	poultry	<----------------------------	[fruits, pork]
rule1	0.064969271	0.503401361	1.327254976	soap	<----------------------------	[sandwich loaves, laundry detergent]
rule2	0.065847234	0.5	1.297266515	bagels	<----------------------------	[pork, sugar]
rule3	0.065847234	0.5	1.416666667	flour	<----------------------------	[dishwashing liquid/detergent, sandwich loaves]
rule4	0.065847234	0.5	1.330607477	mixes	<----------------------------	[butter, hand soap]
rule5	0.065847234	0.510204082	1.357762731	individual meals	<----------------------------	[sandwich loaves, laundry detergent]
rule6	0.065847234	0.5	1.268374165	waffles	<----------------------------	[dishwashing liquid/detergent, sandwich loaves]
rule7	0.066725198	0.5	1.279775281	soda	<----------------------------	[pasta, pork]
rule8	0.066725198	0.5	1.186458333	poultry	<----------------------------	[pasta, pork]
rule9	0.066725198	0.520547945	1.350578837	bagels	<----------------------------	[fruits, pork]
rule10	0.066725198	0.506666667	1.33896365	laundry detergent	<----------------------------	[sandwich bags, sugar]
rule11	0.066725198	0.506666667	1.202277778	poultry	<----------------------------	[sandwich bags, sugar]
rule12	0.066725198	0.503311258	1.336297257	juice	<----------------------------	[spaghetti sauce, flour]
rule13	0.066725198	0.506666667	1.33896365	laundry detergent	<----------------------------	[butter, hand soap]
rule14	0.066725198	0.506666667	1.296838951	cheeses	<----------------------------	[dishwashing liquid/detergent, sandwich loaves]
rule15	0.067603161	0.503267974	1.27382716	lunch meat	<----------------------------	[shampoo, tortillas]
rule16	0.067603161	0.5	1.279775281	soda	<----------------------------	[flour, beef]
rule17	0.067603161	0.503267974	1.293955355	dinner rolls	<----------------------------	[shampoo, tortillas]
rule18	0.067603161	0.503267974	1.339304258	mixes	<----------------------------	[shampoo, tortillas]
rule19	0.067603161	0.503267974	1.34875817	spaghetti sauce	<----------------------------	[sandwich loaves, milk]
rule20	0.067603161	0.513333333	1.366090343	individual meals	<----------------------------	[dishwashing liquid/detergent, sandwich loaves]

Association Identified

- With Support as 0.05, confidence as 0.5 we are able to identify 1187 rules
- **For Support**: - Higher the support value, item is more likely to be ordered
- **For Confidence**: - Higher the confidence value, better is the chances of product combos to succeed
- **Lift**: -
  - Lift = 1: - There is no correlation within itemset
  - Lift > 1: - There is strong correlation within itemset
  - Lift < 1:- There is Negative Correlation within itemset
- All the rules has lift value greater than 1 which means there is a positive correlation within item set

Top 5 Rules with Highest Support					
row ID	Support	Confidence	Lift	Recommended item	Items In Basket
rule1186	0.194907814	0.501128668	1.189136569	poultry	[dinner rolls]
rule1183	0.099209833	0.579487179	1.489923019	dinner rolls	[spaghetti sauce, poultry]
rule1184	0.099209833	0.509009009	1.364144144	spaghetti sauce	[dinner rolls, poultry]
rule1185	0.099209833	0.576530612	1.368059099	poultry	[dinner rolls, spaghetti sauce]
rule1180	0.095697981	0.542288557	1.410197869	aluminum foil	[poultry, juice]

Top 5 Rules with Highest Confidence					
row ID	Support	Confidence	Lift	Recommended item	Items In Basket
rule287	0.075504829	0.585034014	1.388236961	poultry	[sandwich loaves, laundry detergent]
rule1183	0.099209833	0.579487179	1.489923019	dinner rolls	[spaghetti sauce, poultry]
rule1185	0.099209833	0.576530612	1.368059099	poultry	[dinner rolls, spaghetti sauce]
rule556	0.079016681	0.573248408	1.360270701	poultry	[mixes, sugar]
rule1047	0.086918349	0.565714286	1.342392857	poultry	[lunch meat, mixes]

Top 5 Rules with Highest Lift					
row ID	Support	Confidence	Lift	Recommended item	Items In Basket
rule810	0.082528534	0.562874251	1.497929375	individual meals	[sandwich loaves, lunch meat]
rule1183	0.099209833	0.579487179	1.489923019	dinner rolls	[spaghetti sauce, poultry]
rule465	0.078138718	0.559748428	1.486138599	juice	[shampoo, spaghetti sauce]
rule849	0.082528534	0.513661202	1.470000275	sandwich loaves	[cheeses, ketchup]
rule1030	0.086040386	0.547486034	1.46726257	spaghetti sauce	[dinner rolls, juice]

Inferences - Recommendation - Suggestions

INFERNCES

- Poultry is most recommended item i.e., 225 times which is almost 20% of total recommendation
- Second most highest recommendation is of Soda i.e., 68 times which is almost 6% of total recommended item.
- Flour, Sandwich Loaves and Sugar are least recommended item
- Butter, Hand Soap & Pork have no recommendation.

Recommended Item	No. of times Recommended
poultry	225
soda	68
lunch meat	67
yogurt	65
cheeses	64
eggs	62
waffles	54
ice cream	52
dinner rolls	51
dishwashing liquid/detergent	46
cereals	45
juice	44
aluminum foil	42
mixes	37
soap	34
bagels	25
ketchup	24

Recommended Item	No. of times Recommended
spaghetti sauce	20
milk	18
paper towels	17
individual meals	16
beef	15
laundry detergent	14
sandwich bags	14
shampoo	13
fruits	11
coffee/tea	10
pasta	10
toilet paper	9
all- purpose	6
tortillas	3
flour	2
sandwich loaves	2
sugar	2

RECOMMENDATIONS

- Poultry is the most recommended item. Hence it could be suggested as combo with less recommended eatable items such as sugar, sandwich loaves , flour.
- 5% or 10% discounts can be given items such as soda, lunch meat, yogurt or cheese in order to increase the sales
- We can make Combos of daily usage and less recommended items such as shampoo, Fruits, coffee/tea.
- Combos such as [sugar, milk, eggs] or [Tortilla, Dinner rolls, Cereals, Meat] or more such combos can be prepared as these products are used together in a household
- Items with high support, confidence and lift should be given offers like “Combo pack”, “Buy 1 Get 1” or “Discounts” to increase sales
- Some Reward Points system can be introduced in order to attract the new or retain the old customers

Combos or Discount Offers Based on Association

- Combos of items with which recommended items will go well as they have high lifts: -
  - [Spaghetti Sauce, Poultry] -------> Dinner Rolls
  - [Sandwich Loaves, Lunch Meats] -------> Individual Meats
  - [Cheeses, Ketchup] -------> Sandwich Loaves
  - [Dinner Roll, Juice] -------> Spaghetti Sauce
- Daily use Products Combos which can be useful: -
  - Pasta and Cheese
  - Waffles and Ice Cream
  - Milk and Sugar
  - Sandwich Loaves and Butter
  - Lunch Meat and Soda or Beef and Soda
  - Fruits and Yogurt
- Discounts on Purchase of Butter, Hand Soap and Pork can be provided as they don't go well with recommendation.
- Discounts such as buy 2 Soda and get 1 Soda free can be useful as people tends buy more than 1 soda mostly.
- Discounts of 5% to 10% on Dishwashing Liquid/Detergent or Laundry Detergent can be given.