MARKET BASKET ANALYSIS

INTRODUCTION

The objective 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.

TOOLS USED

PYTHON

CONTENTS

EDA

MARKET BASKET ANALYSIS

ASSOCIATION RULES

RECCOMENDATIONS

DATA EXPLORATION PROCESS

The dataset consists of 20641 observations from the store.

The data file consists of 3 columns that are date, Order_id and Product.

There are 2 object data types and 1 integer data type.

	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

```
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
```

DATA EXPLORATION PROCESS

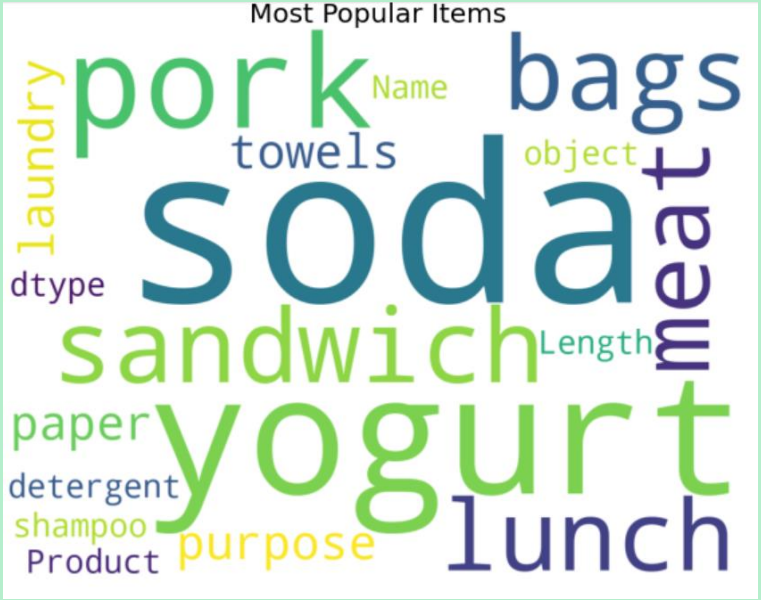
There are 4730 duplicate values and there are no missing values in the dataset.

```
4730
```

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

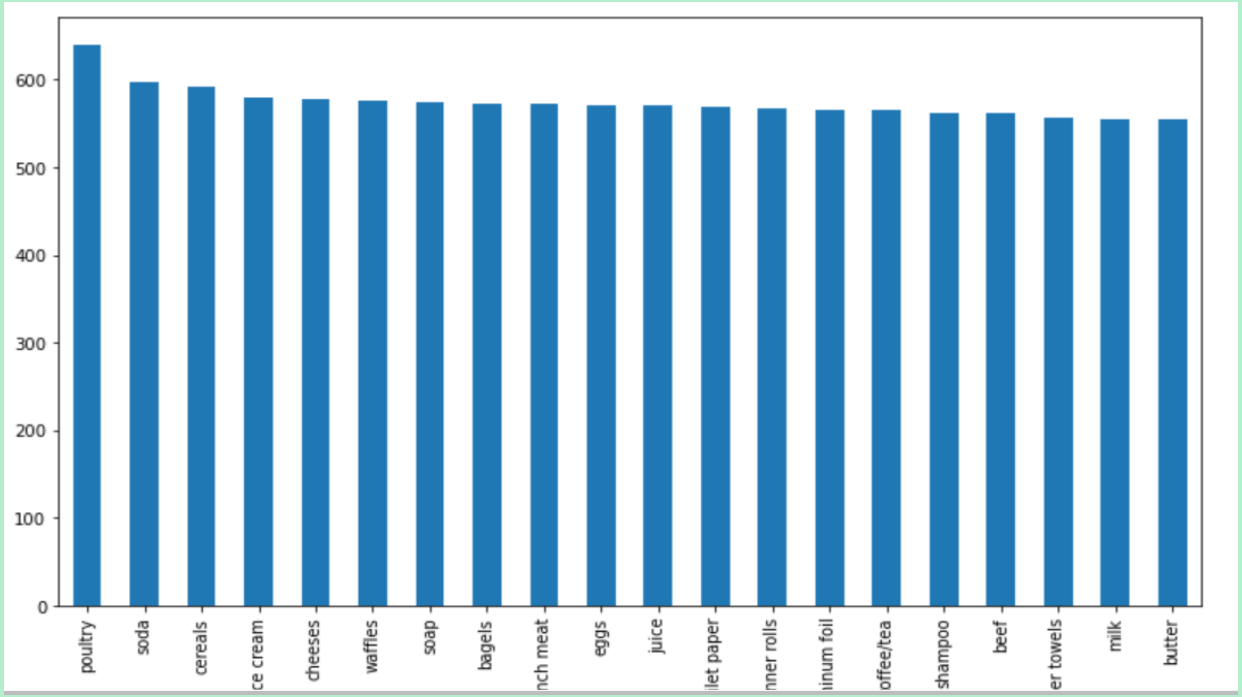
	count	mean	std	min	25%	50%	75%	max
Order_id	20641.0	575.986289	328.557078	1.0	292.0	581.0	862.0	1139.0

Most Popular Items



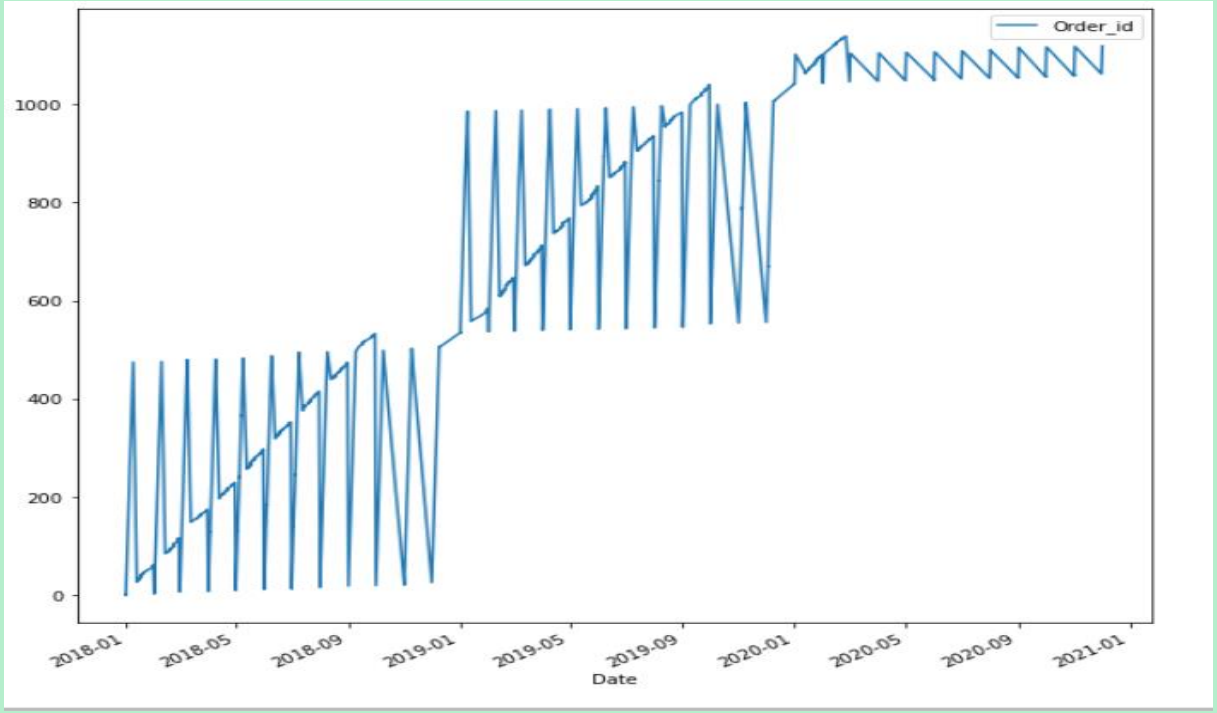
Yogurt, Soda, pork etc; are the products that have been purchased most frequenty.

TOP 20 PRODUCTS



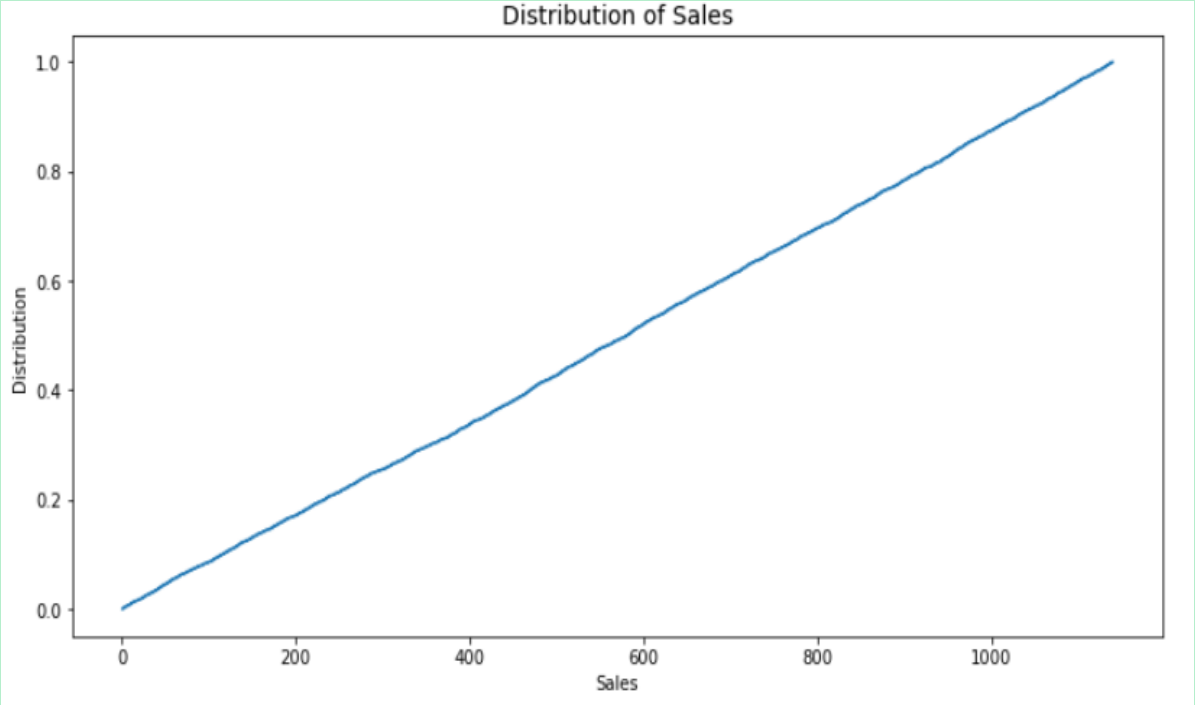
From this bar plot, we can clearly observe that the top 3 products of the store are poultry, soda and cerelary.

TRENDS ACROSS THE YEARS

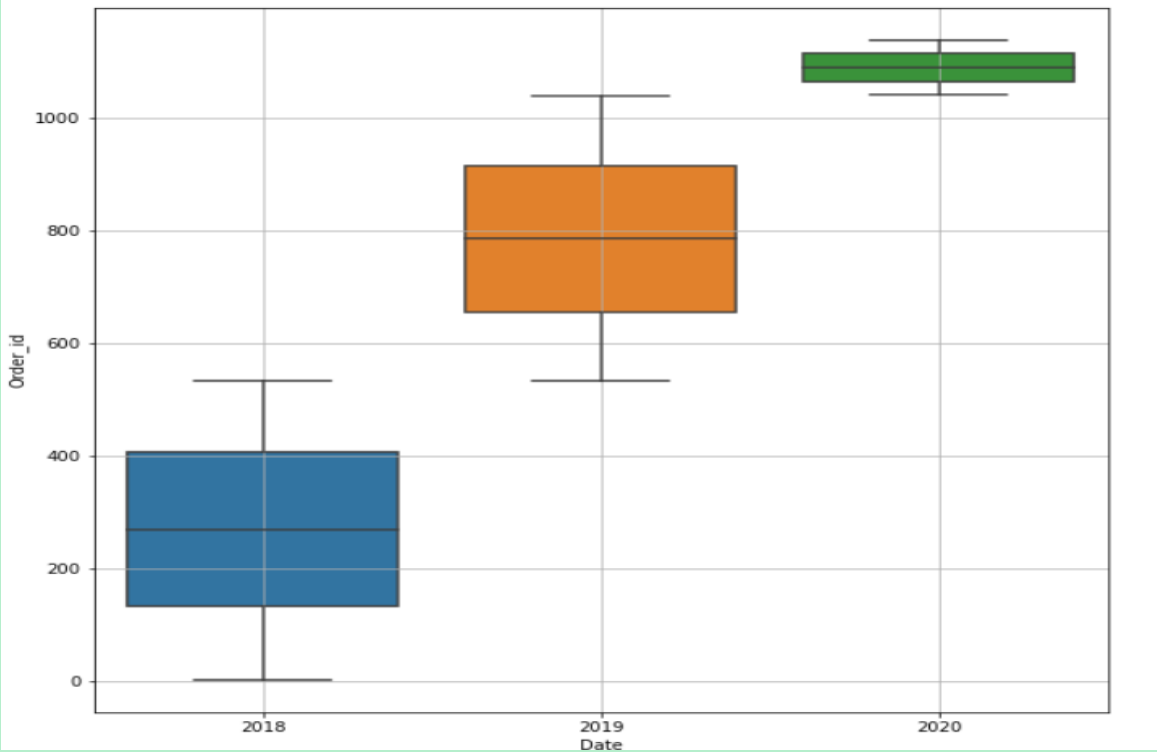


The sales keep increasing steadily yearly.

DISTRIBUTION OF SALES

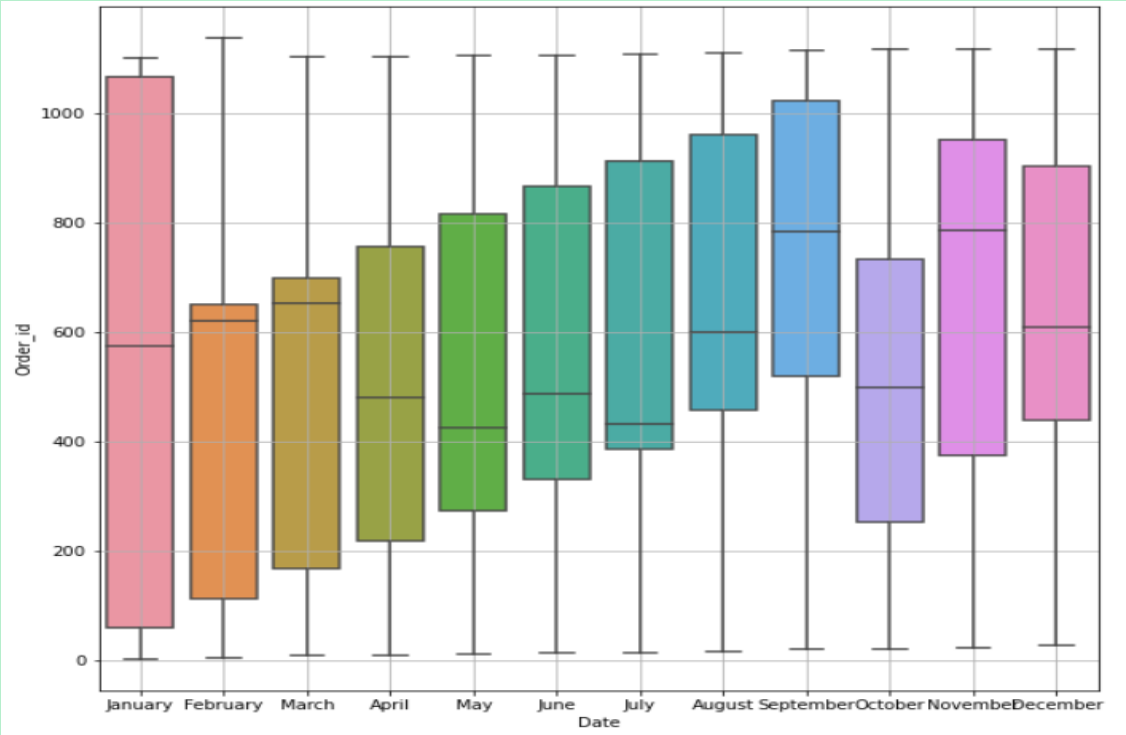


TRENDS ACROSS THE YEARS

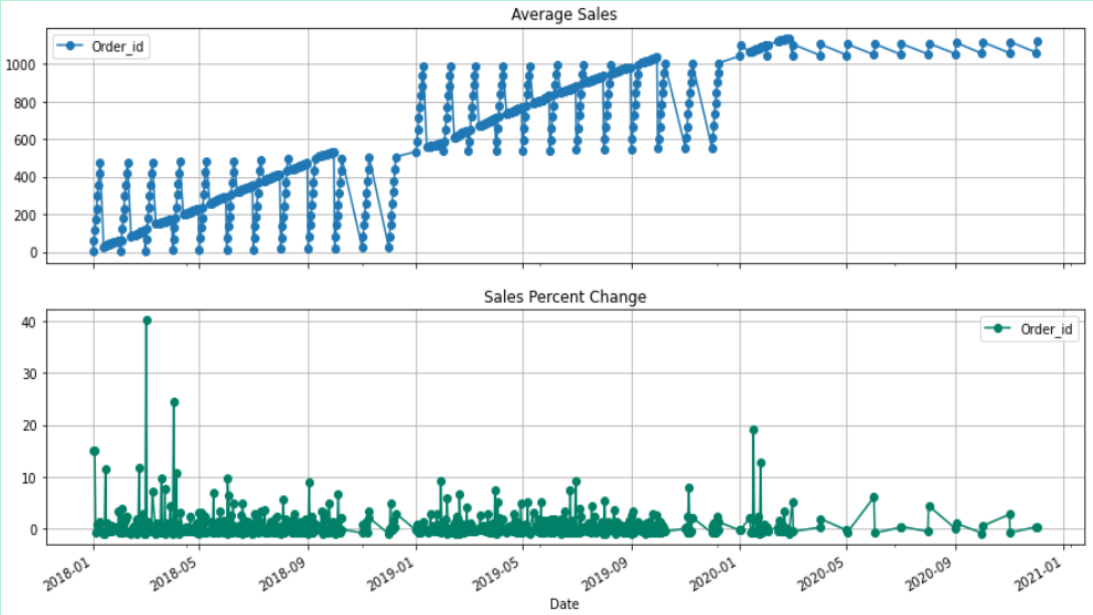


Highest number of orders are made in the year 2020 and the lowest in 2018.

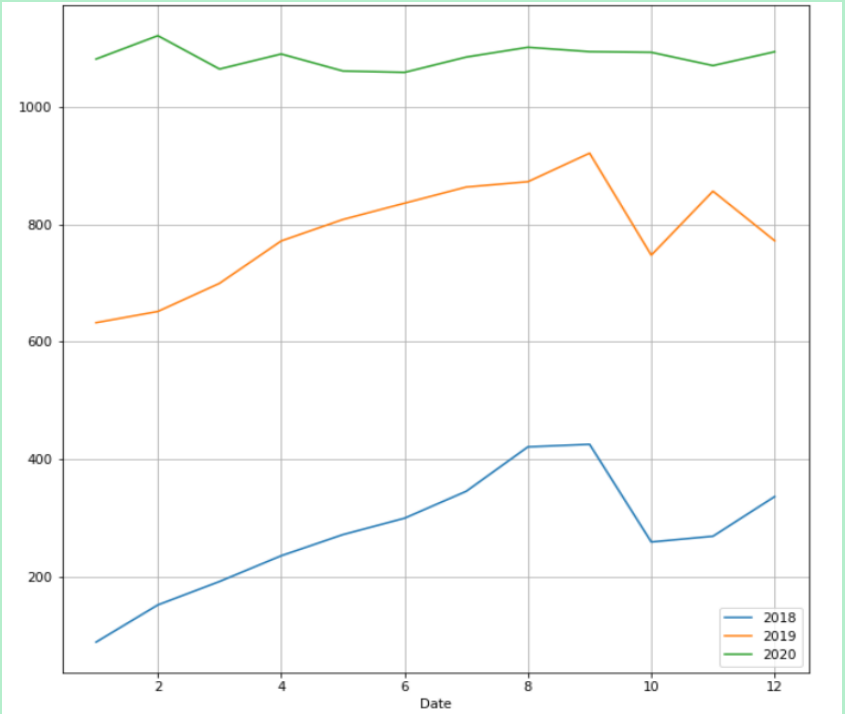
TRENDS ACROSS THE YEARS



From this box plot, we can clearly observe that the the high number of orders are received in the month of January followed by September . Also, there are no outliers present. The lowest number of orders received are in February.


Average Sales
Order_id
1000
800
600
400
200
0
Sales Percent Change
Order_id
40
30
20
10
0
2018-01
2018-05
2018-09
2019-01
2019-05
2019-09
2020-01
2020-05
2020-09
2021-01
Date

Date	2018	2019	2020
Date			
1	89.125609	632.501699	1081.015815
2	152.317186	651.580328	1120.710280
3	192.152731	699.412500	1064.000000
4	235.959402	771.545455	1089.533333
5	271.914336	807.997906	1060.687500
6	299.839923	835.753138	1058.311475
7	345.662681	863.277293	1084.321429
8	421.204698	872.293137	1101.020408
9	425.521368	920.856092	1093.545455
10	259.387850	747.331288	1092.405405
11	269.150000	856.179641	1069.830189
12	336.559259	771.931034	1093.350000



02.

MARKET BASKET ANALYSIS

Market Basket Analysis Process

01 **Construct association rules**

02 **Identify items frequently purchased together**

03 **Construct Recommendations**

WHAT ARE ASSOCIATION RULES?

It is a technique to identify underlying relationships between different items in a dataset to reveal how seemingly unrelated items have associations among them.

Association Rules are generally expressed in an 'if-then' format. We can also use the term 'antecedent' for if and 'consequent' for then.

Ex:{antecedent}→{consequent}
{vegetable}→{fruit}

Association rules are usually written as **i(j)-> i(k)**. This means that there is a strong relationship between the purchase of item i(j) and item i(k). Both these items were purchased together in the same transaction.

Support Threshold Value

The support of an itemset X, *supp(X)* is the proportion of transaction in the database in which the item X appears.
where N refers to the total number of transactions.

$$Support = \frac{Frequency\ (X,Y)}{N}$$

It signifies the popularity of an itemset which means that Higher the support, more frequently the items appears in the data set. A low support value can be helpful to find “hidden” relationships between the items.

Confidence Threshold Value

Confidence of a rule signifies the likelihood of item Y being purchased when item X is purchased. A confidence of .5 in the above example would mean that in 50% of the cases where Baby Gel and Soap were purchased, the purchase also included Cookies and Chips.

$$Confidence\ (A \rightarrow B) = \frac{Support\ (A \cup B)}{Support\ (A)}$$

Higher the confidence value, more reliable the rule is.

03.

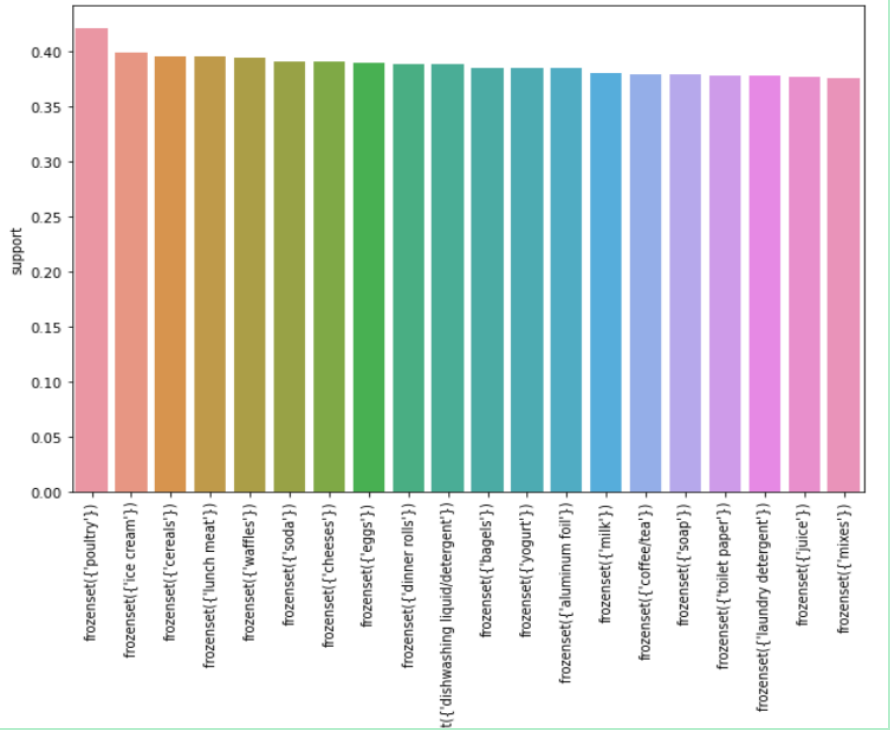
ASSOCIATION RULES

Association Rules

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
21916458	(ketchup, sugar, sandwich bags, all-purpose)	(laundry detergent, soap, flour)	0.023705	0.058824	0.011414	0.481481	8.185185	0.010019	1.815126
21916523	(laundry detergent, soap, flour)	(ketchup, sugar, sandwich bags, all- purpose)	0.058824	0.023705	0.011414	0.194030	8.185185	0.010019	1.211329
21907008	(toilet paper, fruits, all- purpose, coffee/tea)	(butter, beef, cereals)	0.022827	0.058824	0.010536	0.461538	7.846154	0.009193	1.747899
21907073	(butter, beef, cereals)	(toilet paper, fruits, all- purpose, coffee/tea)	0.058824	0.022827	0.010536	0.179104	7.846154	0.009193	1.190374
22026341	(pork, milk, individual meals, ice cream)	(sandwich loaves, shampoo, cereals)	0.020193	0.067603	0.010536	0.521739	7.717674	0.009170	1.949557
22026384	(sandwich loaves, shampoo, cereals)	(pork, milk, individual meals, ice cream)	0.067603	0.020193	0.010536	0.155844	7.717674	0.009170	1.160694
22029873	(ketchup, cheeses, lunch meat, milk)	(pork, soap, coffee/tea)	0.022827	0.065847	0.011414	0.500000	7.593333	0.009910	1.868306
22029908	(pork, soap, coffee/tea)	(ketchup, cheeses, lunch meat, milk)	0.065847	0.022827	0.011414	0.173333	7.593333	0.009910	1.182064
21916456	(ketchup, laundry detergent, sugar, all-purpose)	(flour, soap, sandwich bags)	0.025461	0.059701	0.011414	0.448276	7.508621	0.009893	1.704291
21916525	(flour, soap, sandwich bags)	(ketchup, laundry detergent, sugar, all-purpose)	0.059701	0.025461	0.011414	0.191176	7.508621	0.009893	1.204885
22002793	(pasta, lunch meat, beef, sandwich bags)	(shampoo, fruits, spaghetti sauce)	0.026339	0.053556	0.010536	0.400000	7.468852	0.009125	1.577407
22002808	(shampoo, fruits, spaghetti sauce)	(pasta, lunch meat, beef, sandwich bags)	0.053556	0.026339	0.010536	0.196721	7.468852	0.009125	1.212109
21916490	(flour, laundry detergent, soap, sandwich bags)	(ketchup, sugar, all- purpose)	0.027217	0.057068	0.011414	0.419355	7.348387	0.009860	1.623939
21916491	(ketchup, sugar, all- purpose)	(flour, laundry detergent, soap, sandwich bags)	0.057068	0.027217	0.011414	0.200000	7.348387	0.009860	1.215979
21986699	(pasta, pork, soap)	(soda, ketchup, waffles, bagels)	0.047410	0.030729	0.010536	0.222222	7.231746	0.009079	1.246206
21986646	(soda, ketchup, waffles, bagels)	(pasta, pork, soap)	0.030729	0.047410	0.010536	0.342857	7.231746	0.009079	1.449593
21916481	(flour, sugar, sandwich bags, all- purpose)	(ketchup, laundry detergent, soap)	0.022827	0.069359	0.011414	0.500000	7.208861	0.009830	1.861282

Support for itemsets using Apriori

	support	itemsets
0	0.374890	(all- purpose)
1	0.384548	(aluminum foil)
2	0.385426	(bagels)
3	0.374890	(beef)
4	0.367867	(butter)
...	...	...
610567	0.010536	(ketchup, mixes, spaghetti sauce, pork, soap, ...
610568	0.011414	(ketchup, waffles, mixes, spaghetti sauce, soa...
610569	0.010536	(sandwich loaves, laundry detergent, soap, lun...
610570	0.011414	(yogurt, mixes, milk, sandwich bags, lunch mea...
610571	0.010536	(yogurt, tortillas, mixes, milk, sandwich bags...



Association Rules

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	convict
21916458	(ketchup, sugar, sandwich bags, all-purpose)	(laundry detergent, soap, flour)	0.023705	0.058824	0.011414	0.481481	8.185185	0.010019	1.815
21916523	(laundry detergent, soap, flour)	(ketchup, sugar, sandwich bags, all- purpose)	0.058824	0.023705	0.011414	0.194030	8.185185	0.010019	1.211
21907008	(toilet paper, fruits, all- purpose, coffee/tea)	(butter, beef, cereals)	0.022827	0.058824	0.010536	0.461538	7.846154	0.009193	1.747

Running the algorithm with the parameters set as follows: miminum_support = 0.01 . We'll be able to see all the items with 2 or more items appear in the frequent itemset, along with their rules. How the metrics are calculated are mentioned in the previous slides.

The first itemset shows the association: "If Ketchup, Sugar then Laundry Detergent , soap, flour " with suport vlue at 0.011 i.e around 1.1% of all transactions have this combination.

The confidence value is 0.48 i.e 48% confidence that the sales of first item set of antecedents happen whenever consequents are purchased.

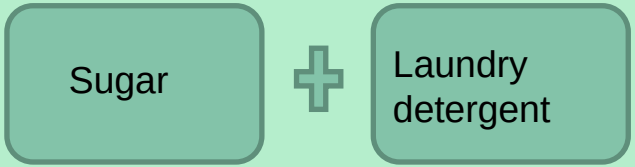
Lift metric is used to evaluate relationship between items. If a lift value is less than 1, then it likely means that there is no relationship between Ketchup and Soap items. But here Lift value is 8.18, which suggests that item ketchup is likely bought when soap/flour items are purchased.

04.

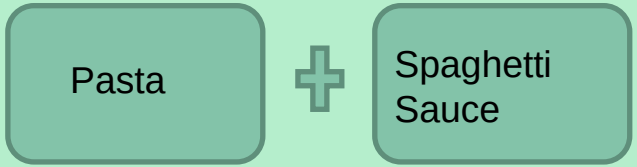
RECCOMENDATIONS

Combo Offers

- HOME DEAL



- ITALIAN OFFER



- BREAKFAST DEAL

- RANDOM OFFER

Combo offers on these items are made as they have high lift value of around 8.

Buy 2 get 1 free offers

RECCOMMENDATIONS

A

Combo Offers. Offering combo will increase the sales in the store.

B

Offering Discount to the low sale items and it can increase the sales of these specific items

C

Buy 2 get 1 free is also a great way to boost the sales. It will increase the revenue among various items.