

# MRA Project

## Part-B

Shweta Tripathi

07-01-2024



## **Problem Statement:**

A grocery store shared the transactional data with you. Your job is to conduct a thorough analysis of Point of Sale (POS) data, identify the most commonly occurring sets of items in the customer orders, and provide recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.

## **Dataset:**

Grocery Store Data: [dataset\\_group.csv](#)

Top few records to get a feel of the data structure are shown 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

	Date	Order_id	Product
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

# Exploratory data analysis (EDA)

The dataset contains 20641 rows and 3 columns

```
df.shape
```

```
(20641, 3)
```

The info gives the information of the dataset such as Column name, Non-null count, Dtype. The dataset contains 3 columns in which 1 int and 2 object datatype

```
<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 five point summary of the dataset can be obtained from the describe function.

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

- Null values should be checked before performing calculations. Checking the null values by using isnull function. We have observed that there are no null values present in the dataset.

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

- On checking the duplicate values we have observed that there are 4730 duplicate values. Thus before calculation we will remove the duplicate values.


```
df.duplicated().sum()
```

```
4730
```

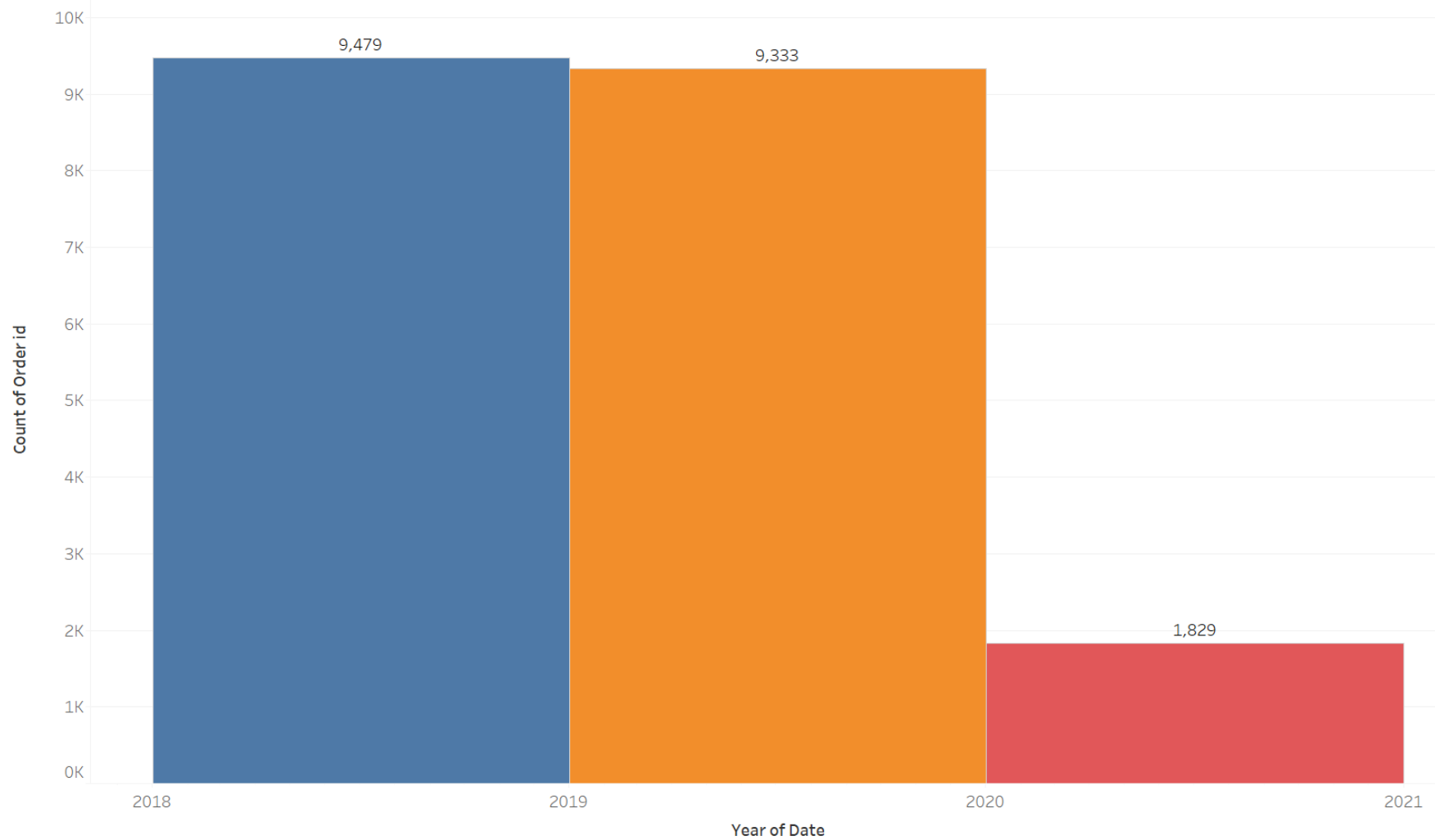
Drop function is used to remove the duplicate values .


The number of duplicate values are 0



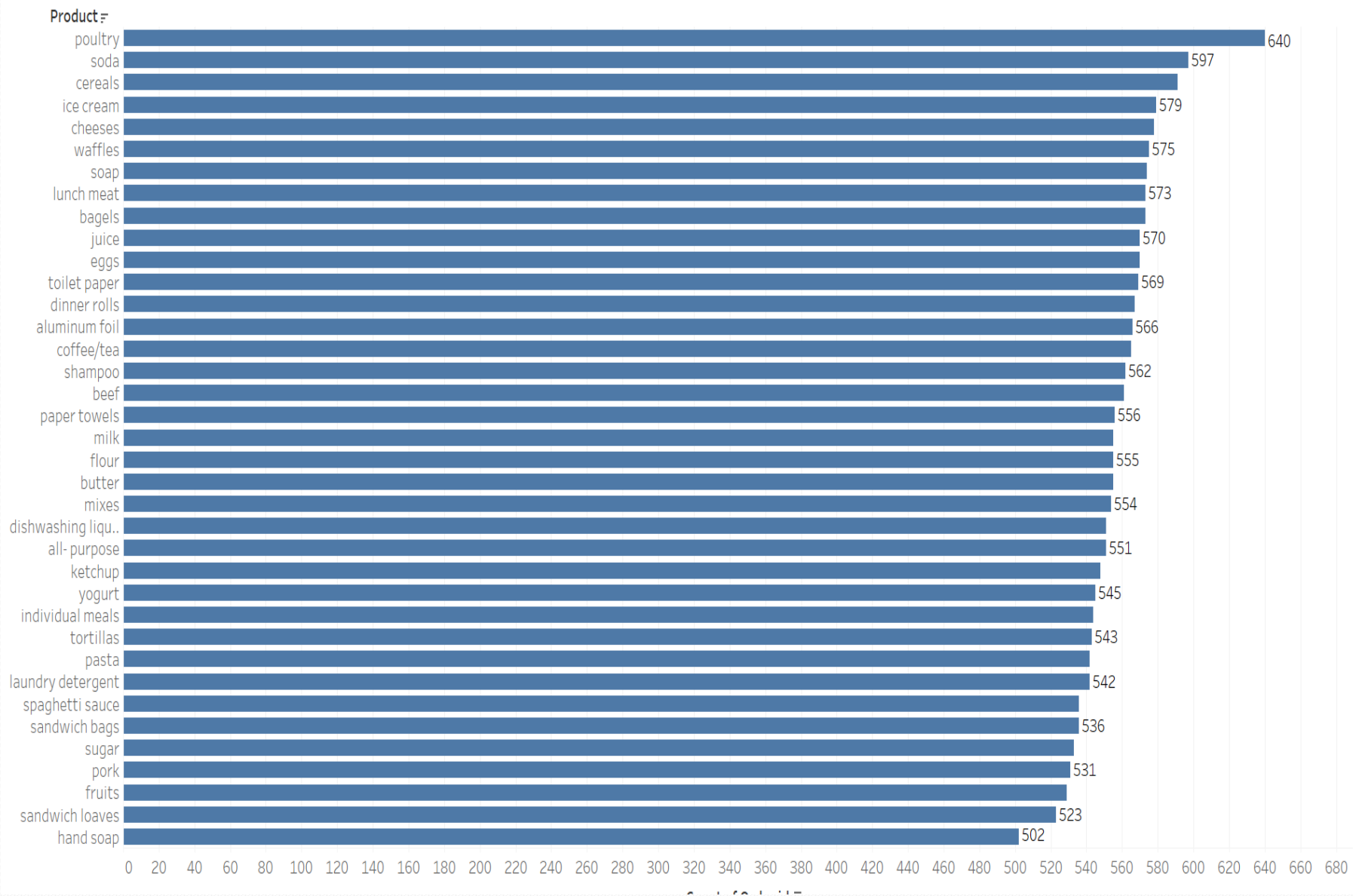
- 
- Order-id is plotted against Year in Tableau.
  - The blue color shows the sale from 2018-2019 , orange color shows the sale from 2019-2020 where as red color shows the sale from 2020-2021.
  - The sale from 2018-2019 is 9479\$, from 2019-2020 it is 9333\$ where as from 2020-2021 is 1829\$.
  - Thus, it is clear from the graph that the sale is decreases as in 2021 lowest sale is observed.

orderid/year

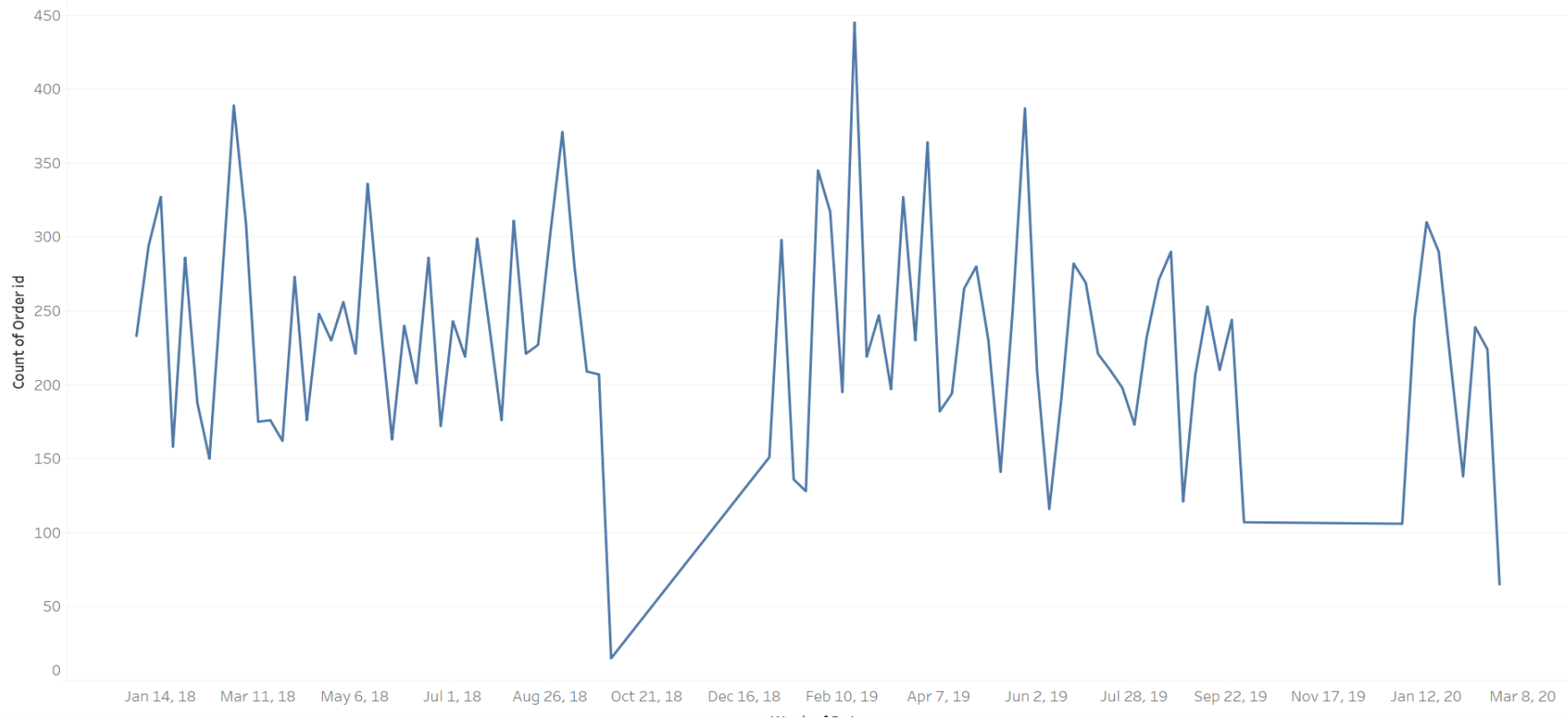


- 
- Tableau is used to make a graph of count of order-id and products that are plotted against each other.
  - It is observed from the graph that the maximum order-id is placed for poltry, soda, cereals where as minimum order-id is placed for handsoaps and sanwich loaves.
  - The maximum order count for poltry is 640 where as minimum order count for hansoaps is 502

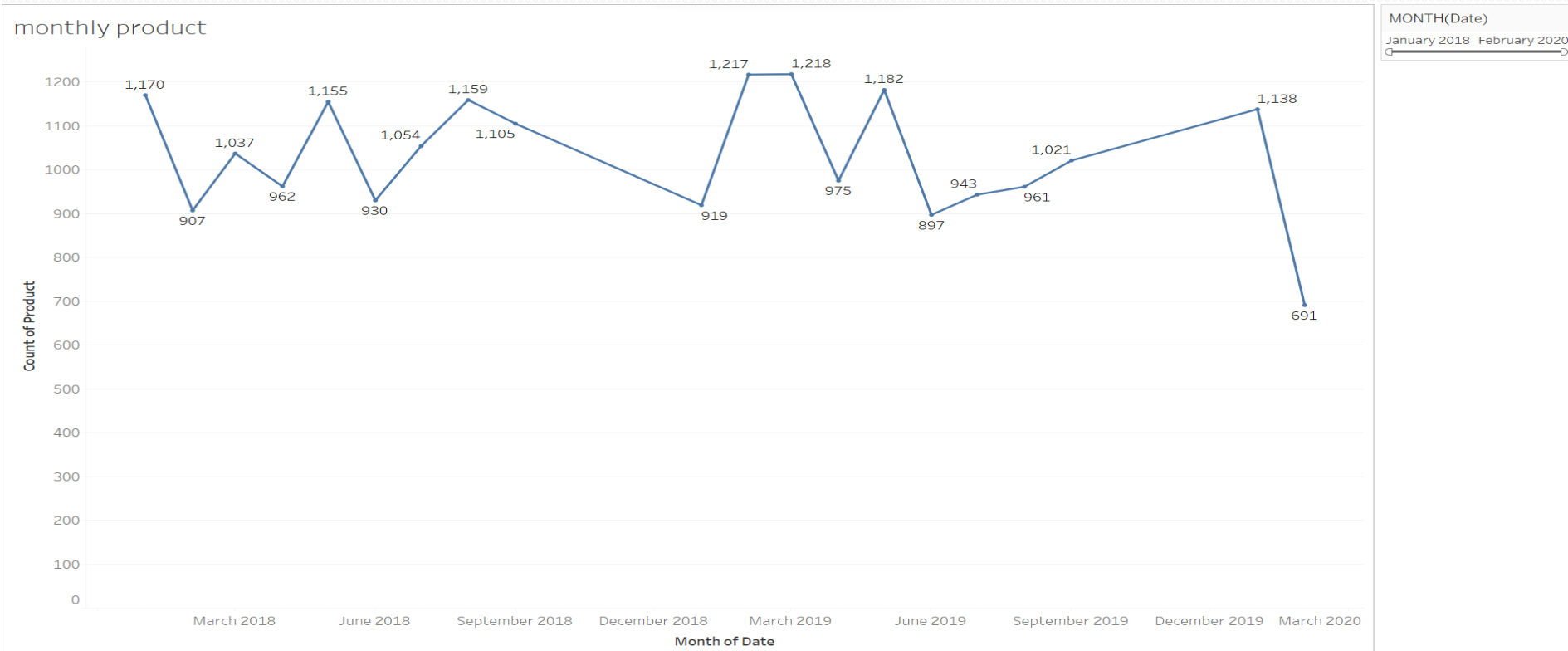
orderid/product



- The weekly order count is plotted in below graph.
- It is observed that only in second week of feb2020 sale was increase other than that the sale is continuously decreasing in 2020



- The monthly product count is shown in below line graph.
- It is observed that product count decreases from 1138 to 691

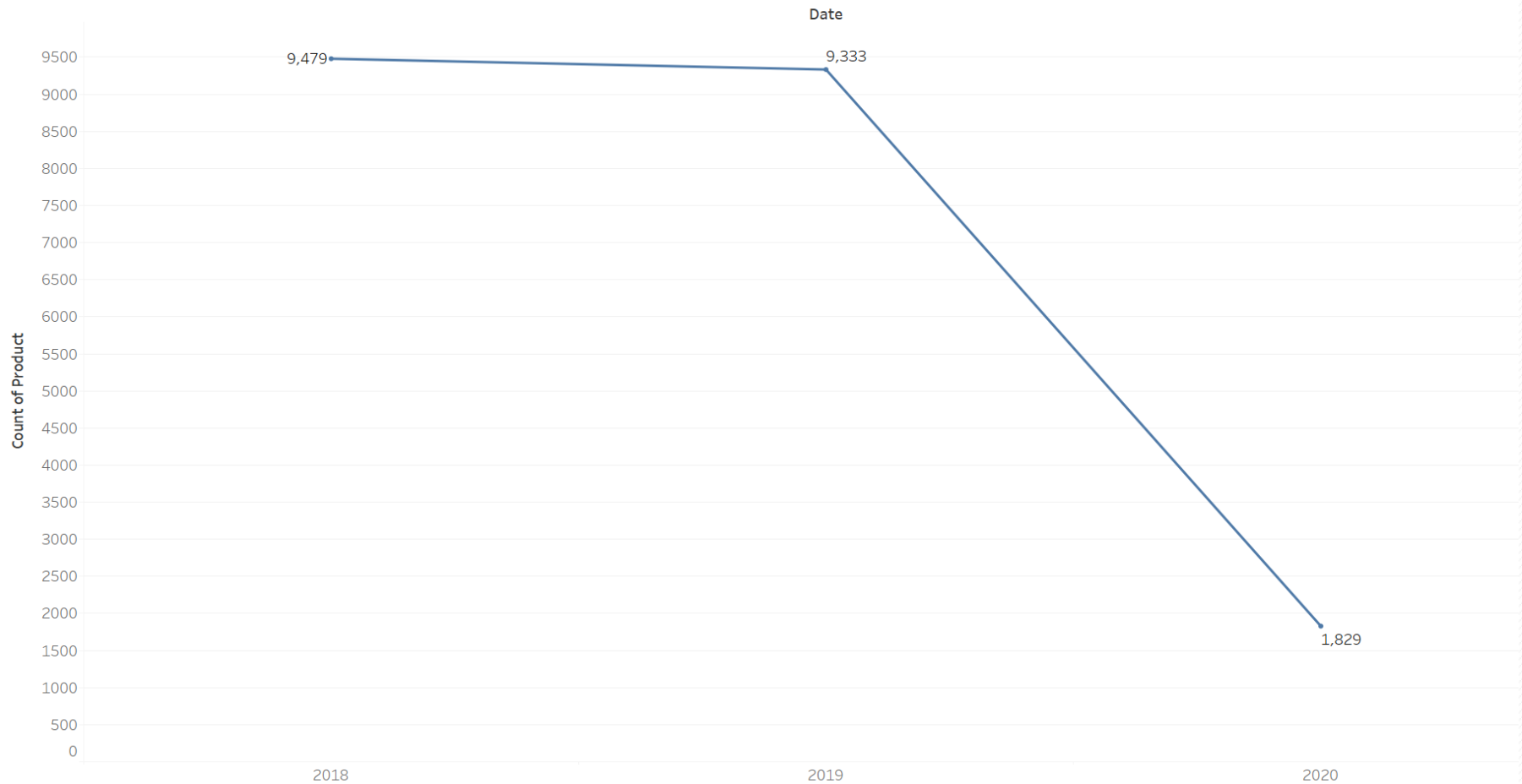


- The quarterly product count is shown below which shows.
- It shows that from 2019 Q3 to 2020Q1 there is huge loss in product count



- The yearly decrease in product count is shown in the below line graph.

yearly product



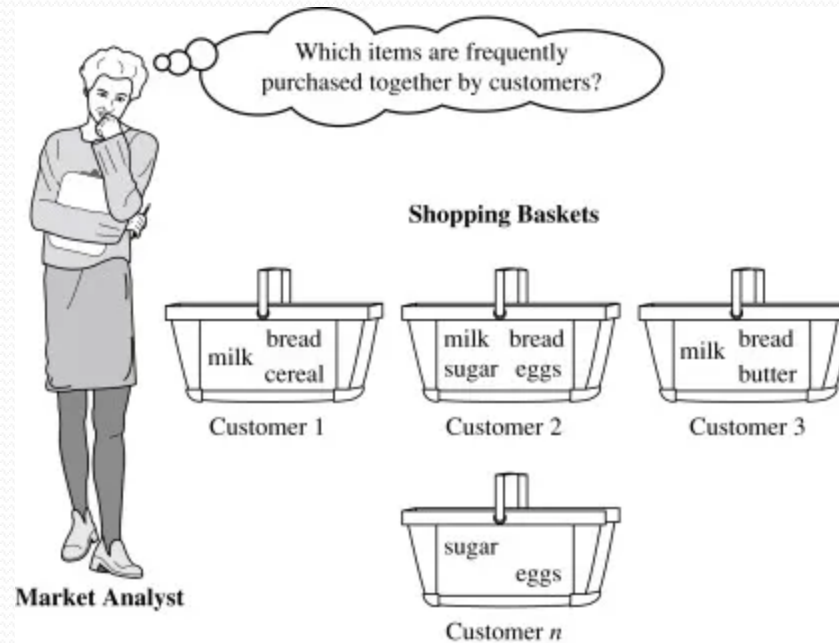


## INFERENCES

- Since sale is decreasing in 2020 from 2<sup>nd</sup> Quarter onwards thus more offers should be provided on combos and update customers about the offers
- Sales can also increase by providing different options to customers such as buy using buy online and pickup in store , in house delivery , reducing wait time at counters etc.
- Sale can also increase by providing discounts at various seasons, weekend etc

# Market Basket Analysis

- Market Basket Analysis is a technique used in Data Mining to increase the sales by understanding customer purchase pattern.



The three important terms which are used in Market Basket Analyses.

$$\begin{aligned} &\textbf{Support} \\ &= (A+B)/\textbf{Total} \end{aligned}$$

$$\begin{aligned} &\textbf{Confidence} \\ &= (A+B)/A \end{aligned}$$

$$\begin{aligned} &\textbf{Lift} \\ &= ((A+B)/A) / (B/\textbf{Total}) \end{aligned}$$

The KNIME tool is used to do the Market Basket Analysis

The data is imported in the KNIME using the CSV node.

Rows: 20641 | Columns: 3

<input type="checkbox"/>	#	RowID	Date String	<input type="checkbox"/>	Order_id Number (integer)	<input type="checkbox"/>	Product String	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	1	Row0	01-01-2018		1		yogurt		
<input type="checkbox"/>	2	Row1	01-01-2018		1		pork		
<input type="checkbox"/>	3	Row2	01-01-2018		1		sandwich bags		
<input type="checkbox"/>	4	Row3	01-01-2018		1		lunch meat		
<input type="checkbox"/>	5	Row4	01-01-2018		1		all- purpose		
<input type="checkbox"/>	6	Row5	01-01-2018		1		flour		
<input type="checkbox"/>	7	Row6	01-01-2018		1		soda		
<input type="checkbox"/>	8	Row7	01-01-2018		1		butter		
<input type="checkbox"/>	9	Row8	01-01-2018		1		beef		
<input type="checkbox"/>	10	Row9	01-01-2018		1		aluminum foil		
<input type="checkbox"/>	11	Row...	01-01-2018		1		all- purpose		
<input type="checkbox"/>	12	Row...	01-01-2018		1		dinner rolls		
<input type="checkbox"/>	13	Row...	01-01-2018		1		shampoo		
<input type="checkbox"/>	14	Row...	01-01-2018		1		all- purpose		
<input type="checkbox"/>	15	Row...	01-01-2018		1		mixes		
<input type="checkbox"/>	16	Row...	01-01-2018		1		soap		
<input type="checkbox"/>	17	Row...	01-01-2018		1		laundry detergent		
<input type="checkbox"/>	18	Row...	01-01-2018		1		ice cream		
<input type="checkbox"/>	19	Row...	01-01-2018		1		dinner rolls		
<input type="checkbox"/>	20	Row...	01-01-2018		2		toilet paper		
<input type="checkbox"/>	21	Row...	01-01-2018		2		shampoo		
<input type="checkbox"/>	22	Row...	01-01-2018		2		hand soap		
<input type="checkbox"/>	23	Row...	01-01-2018		2		waffles		
<input type="checkbox"/>	24	Row...	01-01-2018		2		cheeses		

- Group By node is used to convert the data in the form of set as shown below :-

Rows: 1139 | Columns: 2

<input type="checkbox"/>	#	RowID	Order_id <i>Number (integer)</i>	Set(Product) <i>Set</i>
<input type="checkbox"/>	1	Row0	1	[yogurt,pork,sandwich bags,...]
<input type="checkbox"/>	2	Row1	2	[toilet paper,shampoo,hand soap,...]
<input type="checkbox"/>	3	Row2	3	[soda,pork,soap,...]
<input type="checkbox"/>	4	Row3	4	[cereals,juice,lunch meat,...]
<input type="checkbox"/>	5	Row4	5	[sandwich loaves,pasta,tortillas,...]
<input type="checkbox"/>	6	Row5	6	[laundry detergent,toilet paper,eggs,...]
<input type="checkbox"/>	7	Row6	7	[individual meals,paper towels,tortillas,...]
<input type="checkbox"/>	8	Row7	8	[ice cream,juice,paper towels,...]
<input type="checkbox"/>	9	Row8	9	[juice,poultry,coffee/tea,...]
<input type="checkbox"/>	10	Row9	10	[ketchup,coffee/tea,toilet paper,...]
<input type="checkbox"/>	11	Row...	11	[sandwich loaves,ice cream,soda,...]
<input type="checkbox"/>	12	Row...	12	[pork,tortillas,shampoo,...]
<input type="checkbox"/>	13	Row...	13	[sugar,fruits,all- purpose,...]
<input type="checkbox"/>	14	Row...	14	[fruits,dinner rolls,individual meals,...]
<input type="checkbox"/>	15	Row...	15	[individual meals,ice cream,cereals,...]
<input type="checkbox"/>	16	Row...	16	[sugar,sandwich bags,flour,...]
<input type="checkbox"/>	17	Row...	17	[milk,hand soap,pasta,...]
<input type="checkbox"/>	18	Row...	18	[sandwich bags,toilet paper,bagels,...]
<input type="checkbox"/>	19	Row...	19	[individual meals,laundry detergent,coffee/tea,...]
<input type="checkbox"/>	20	Row...	20	[shampoo,dishwashing liquid/detergent,yogurt,...]
<input type="checkbox"/>	21	Row...	21	[waffles,fruits,all- purpose,...]
<input type="checkbox"/>	22	Row...	22	[cheeses,cereals,sugar,...]
<input type="checkbox"/>	23	Row...	23	[aluminum foil,bagels,shampoo,...]
<input type="checkbox"/>	24	Row...	24	[fruits,all- purpose,pasta,...]

The Association Rule Learner node is used to get the Support, Confidence, Lift, Consequent as shown in table below

Rows: 20272 | Columns: 6

<input type="checkbox"/>	#	RowID	Support Number (double)	Confidence Number (double)	Lift Number (double)	Consequent String	implies String	Items Set
<input type="checkbox"/>	1	rule0	0.05	0.576	1.497	yogurt	<---	[toilet paper,juice,al...
<input type="checkbox"/>	2	rule1	0.05	0.594	1.544	aluminum foil	<---	[yogurt,toilet paper,j...
<input type="checkbox"/>	3	rule2	0.05	0.538	1.421	toilet paper	<---	[yogurt,juice,alumin...
<input type="checkbox"/>	4	rule3	0.05	0.64	1.7	juice	<---	[yogurt,toilet paper,...
<input type="checkbox"/>	5	rule4	0.05	0.523	1.36	yogurt	<---	[poultry,juice,alumin...
<input type="checkbox"/>	6	rule5	0.05	0.564	1.468	aluminum foil	<---	[yogurt,poultry,juice]
<input type="checkbox"/>	7	rule6	0.05	0.538	1.276	poultry	<---	[yogurt,juice,alumin...
<input type="checkbox"/>	8	rule7	0.05	0.62	1.645	juice	<---	[yogurt,poultry,alum...
<input type="checkbox"/>	9	rule8	0.05	0.588	1.528	yogurt	<---	[cheeses,cereals,co...
<input type="checkbox"/>	10	rule9	0.05	0.594	1.52	cheeses	<---	[yogurt,cereals,coff...
<input type="checkbox"/>	11	rule10	0.05	0.582	1.469	cereals	<---	[yogurt,cheeses,cof...
<input type="checkbox"/>	12	rule11	0.05	0.613	1.616	coffee/tea	<---	[yogurt,cheeses,cer...
<input type="checkbox"/>	13	rule12	0.05	0.588	1.564	mixes	<---	[dishwashing liquid...
<input type="checkbox"/>	14	rule13	0.05	0.528	1.395	laundry detergent	<---	[dishwashing liquid...
<input type="checkbox"/>	15	rule14	0.05	0.576	1.484	dishwashing liquid/...	<---	[poultry,laundry det...
<input type="checkbox"/>	16	rule15	0.05	0.6	1.424	poultry	<---	[dishwashing liquid...
<input type="checkbox"/>	17	rule16	0.051	0.563	1.464	yogurt	<---	[poultry,mixes,alumi...
<input type="checkbox"/>	18	rule17	0.051	0.569	1.479	aluminum foil	<---	[yogurt,poultry,mixes]
<input type="checkbox"/>	19	rule18	0.051	0.63	1.678	mixes	<---	[yogurt,poultry,alum...
<input type="checkbox"/>	20	rule19	0.051	0.586	1.39	poultry	<---	[yogurt,mixes,alumi...
<input type="checkbox"/>	21	rule20	0.051	0.611	1.66	sandwich bags	<---	[cheeses,bagels,cer...
<input type="checkbox"/>	22	rule21	0.051	0.674	1.726	cheeses	<---	[bagels,cereals,san...

- The recommended items according to basket is obtained from the KNIME tool as shown below :

Rows: 20272 | Columns: 5

<input type="checkbox"/>	#	RowID	Rec_Item String ↑	Support Number (double)	Confidence Number (double)	Lift Number (double)	Items_basket String
<input type="checkbox"/>	1	Row0	all- purpose	0.07	0.444	1.186	aluminum foil, ketchup
<input type="checkbox"/>	2	Row1	all- purpose	0.073	0.464	1.237	aluminum foil, soap
<input type="checkbox"/>	3	Row2	all- purpose	0.157	0.409	1.09	aluminum foil
<input type="checkbox"/>	4	Row3	all- purpose	0.07	0.417	1.111	bagels, aluminum foil
<input type="checkbox"/>	5	Row4	all- purpose	0.062	0.41	1.095	bagels, coffee/tea
<input type="checkbox"/>	6	Row5	all- purpose	0.067	0.411	1.096	bagels, juice
<input type="checkbox"/>	7	Row6	all- purpose	0.063	0.414	1.104	bagels, laundry detergent
<input type="checkbox"/>	8	Row7	all- purpose	0.07	0.444	1.186	bagels, lunch meat
<input type="checkbox"/>	9	Row8	all- purpose	0.063	0.416	1.11	bagels, pasta
<input type="checkbox"/>	10	Row9	all- purpose	0.065	0.418	1.115	bagels, pork
<input type="checkbox"/>	11	Row...	all- purpose	0.06	0.41	1.093	bagels, soap
<input type="checkbox"/>	12	Row...	all- purpose	0.062	0.413	1.101	bagels, sugar
<input type="checkbox"/>	13	Row...	all- purpose	0.067	0.409	1.09	bagels, toilet paper
<input type="checkbox"/>	14	Row...	all- purpose	0.066	0.405	1.081	bagels, tortillas
<input type="checkbox"/>	15	Row...	all- purpose	0.065	0.407	1.085	bagels, waffles
<input type="checkbox"/>	16	Row...	all- purpose	0.067	0.418	1.114	beef, aluminum foil
<input type="checkbox"/>	17	Row...	all- purpose	0.058	0.41	1.093	beef, coffee/tea
<input type="checkbox"/>	18	Row...	all- purpose	0.066	0.424	1.13	beef, juice
<input type="checkbox"/>	19	Row...	all- purpose	0.062	0.415	1.108	beef, pork
<input type="checkbox"/>	20	Row...	all- purpose	0.064	0.41	1.094	beef, sandwich bags
<input type="checkbox"/>	21	Row...	all- purpose	0.072	0.451	1.202	beef, soap
<input type="checkbox"/>	22	Row...	all- purpose	0.062	0.401	1.07	beef, soda
<input type="checkbox"/>	23	Row...	all- purpose	0.066	0.419	1.118	beef, toilet paper
<input type="checkbox"/>	24	Row...	all- purpose	0.064	0.422	1.126	butter. bagels

# INFERENCES

- The customer will buy dinner rolls, spaghetti, sauce , bagels, cereals then he will buy poultry because its probability is approx. 70%

Rec_Item String	Support Number (double)	Confidence ↓ Number (double)	Lift Number (double)	Items_basket String
poultry	0.052	0.686	1.628	dinner rolls, spaghetti sauce,...
cheeses	0.051	0.674	1.726	bagels, cereals, sandwich ba...
poultry	0.054	0.656	1.556	dinner rolls, spaghetti sauce,...
paper towels	0.055	0.649	1.791	eggs, ice cream, pasta
pasta	0.055	0.643	1.731	paper towels, eggs, ice cream
dinner rolls	0.054	0.642	1.651	spaghetti sauce, poultry, laun...

- A lift greater than 1 indicated that the chances of buying item increases where less than 1 indicates that there are less chances of buying the item. Hence, paper towel, pasta, cheeses, juices are more with egg, ice-cream, cereal etc

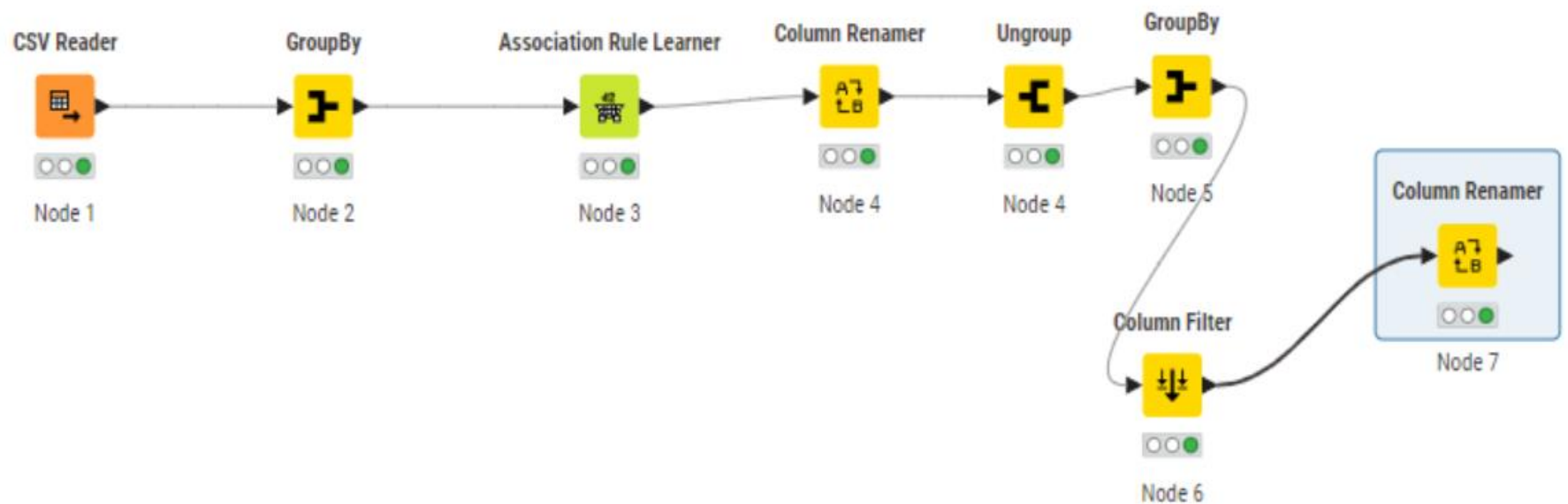


Rows: 20272 | Columns: 5

<input type="checkbox"/>	#	RowID	Rec_Item String	Support Number (double)	Confidence Number (double)	Lift ↓ Number (double)	Items_basket String
<input type="checkbox"/>	12265	Row...	paper towels	0.055	0.649	1.791	eggs, ice cream, pasta
<input type="checkbox"/>	13006	Row...	pasta	0.055	0.643	1.731	paper towels, eggs, ice cream
<input type="checkbox"/>	3315	Row...	cheeses	0.051	0.674	1.726	bagels, cereals, sandwich ba...
<input type="checkbox"/>	9164	Row...	juice	0.05	0.64	1.7	yogurt, toilet paper, aluminu...
<input type="checkbox"/>	12060	Row...	mixes	0.051	0.63	1.678	yogurt, poultry, aluminum foil
<input type="checkbox"/>	14314	Row...	sandwich bags	0.051	0.611	1.66	cheeses, bagels, cereals
<input type="checkbox"/>	5057	Row...	dinner rolls	0.054	0.642	1.651	spaghetti sauce, poultry, laun...
<input type="checkbox"/>	5055	Row...	dinner rolls	0.052	0.641	1.649	spaghetti sauce, poultry, ice ...
<input type="checkbox"/>	9155	Row...	juice	0.05	0.62	1.645	yogurt, poultry, aluminum foil
<input type="checkbox"/>	13722	Row...	poultry	0.052	0.686	1.628	dinner rolls, spaghetti sauce,...
<input type="checkbox"/>	6136	Row...	eggs	0.052	0.634	1.627	paper towels, dinner rolls, pa...
<input type="checkbox"/>	13005	Row...	pasta	0.052	0.602	1.621	paper towels, eggs, dinner ro...
<input type="checkbox"/>	5054	Row...	dinner rolls	0.051	0.63	1.621	spaghetti sauce, poultry, cer...
<input type="checkbox"/>	6142	Row...	eggs	0.055	0.63	1.616	paper towels, ice cream, pasta
<input type="checkbox"/>	4480	Row...	coffee/tea	0.05	0.613	1.616	yogurt, cheeses, cereals
<input type="checkbox"/>	5056	Row...	dinner rolls	0.052	0.628	1.614	spaghetti sauce, poultry, juice
<input type="checkbox"/>	5911	Row...	eggs	0.052	0.628	1.61	dinner rolls, poultry, soda
<input type="checkbox"/>	17089	Row...	spaghetti sauce	0.054	0.598	1.603	dinner rolls, poultry, laundry ...
<input type="checkbox"/>	11335	Row...	milk	0.051	0.604	1.589	poultry, laundry detergent, ce...

- From the Association Rule Learner table we can get the products which are having high possibility of purchase.
- Discounts should be given on combos.
- Shopkeeper should also put some recommended items on sale to increase the sale
- Offers should also be given on some items to increase the sales

# KNIME MODEL



# Tableau Public

[https://public.tableau.com/app/profile/shweta.tripathi2394/viz/MRAB\\_17046350614370/weekwiseproductdistribution?publish=yes](https://public.tableau.com/app/profile/shweta.tripathi2394/viz/MRAB_17046350614370/weekwiseproductdistribution?publish=yes)