MARKET BASKET ANALYSIS

Aaditya Desai
DSBA – Aug 20
aadidesai9@gmail.com

AGENDA [TABLE OF CONTENT]

- Problem Statement
- Data Introduction
- About the Data
- EDA, Timely Trends and Inferences
- Executive Summary
- Market Basket Analysis
- Applications of Market Basket Analysis
- MBA workflow diagram using KNIME and thresholds
- Support, Confidence and Lift with Example
- Association Rules and its Interpretation
- Recommendations and Combo offers suggested

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 items in the customer orders. The Store doesn't have any combo offers. Can you suggest the best combos & offers?

The columns in the dataset are – *DATE, ORDER_ID, PRODUCT*

Following is a snapshot of how data looks like : -

Date	Order_id	Product
01-01-18	1	yogurt
01-01-18	1	pork
01-01-18	1	sandwich bags
01-01-18	1	lunch meat
01-01-18	1	all- purpose
01-01-18	1	flour
01-01-18	1	soda
01-01-18	1	butter
01-01-18	1	beef
01-01-18	1	aluminum foil
01-01-18	1	all- purpose
01-01-18	1	dinner rolls
01-01-18	1	shampoo
01-01-18	1	all- purpose
01-01-18	1	mixes
01-01-18	1	soap
01-01-18	1	laundry detergent
01-01-18	1	ice cream

ABOUT DATA

The data is the given dataset is based on the information collected in the past 3 years. It has 3 columns and has more than 20,000 line items about bill date, order id and products purchased in that order.

The data has no missing values but still needs some computation to make it meaningful and derive some insights. So, we need to do some analysis , EDA, visualizations and perform Market Basket Analysis to derive the insights regarding what recommendations should the system suggest when a customer buys an item or combination of items. This will help the Grocery Store increase its sales by offering combo and special discounts.

EDA

INFORMATION REGARDING DATA USING INFO(),SHAPE, ISNULL().SUM(),DUPLICATED(), UNIQUE()

We have split the date into year, month and day for further analysis.

```
Number of rows- 20641
Number of columns- 6
```

Shape shows no. of rows and 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
 3   day       20641 non-null  object
 4   month     20641 non-null  object
 5   year      20641 non-null  object
dtypes: int64(1), object(5)
memory usage: 967.7+ KB
```

DataTypes of all columns are object except order_id

```
Date        0
Order_id    0
Product     0
day         0
month       0
year        0
dtype: int64
```

There are no Null Values

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

```
4730
```

```
1  df.drop_duplicates(inplace=True)
```

```
1  df.shape
```

```
(15911, 6)
```

Finding duplicates (4730 duplicates found)
And dropping them

DESCRIPTION OF CATEGORICAL DATA-

NUMBER OF UNIQUE VALUE, DIFFERENT VALUES AND ITS OCCURRENCES USING UNIQUE() , VALUE_COUNTS() AND SORTING BY SORT_VALUES()

```
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)
```

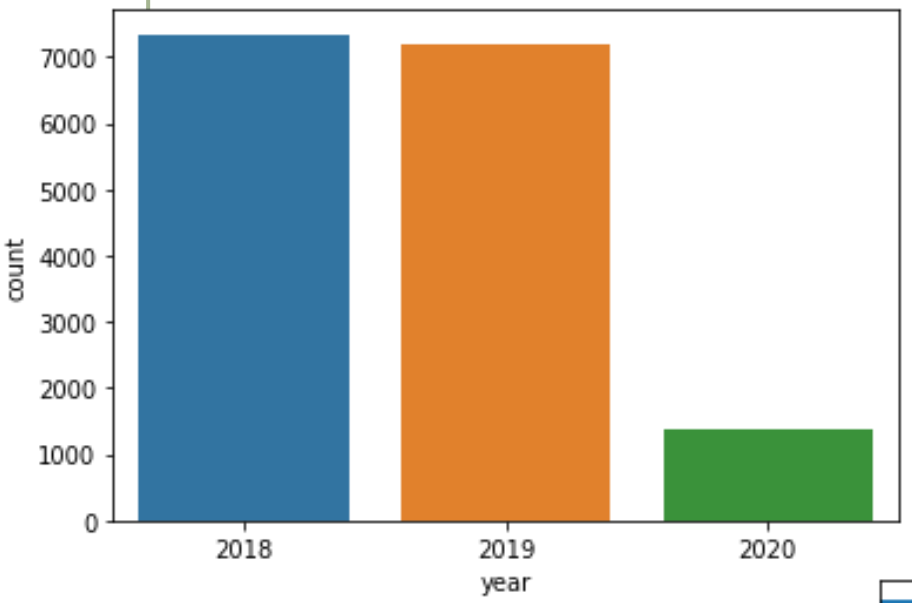
Unique Products

There are 37 such products

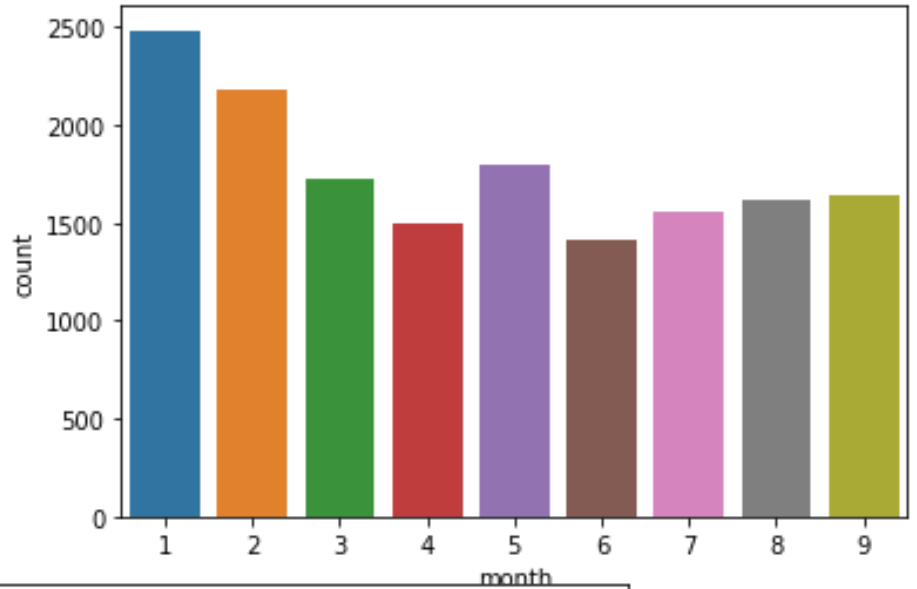
```
poultry                         640
soda                            597
cereals                         591
ice cream                       579
cheeses                         578
waffles                         575
soap                            574
bagels                          573
lunch meat                      573
juice                           570
eggs                            570
toilet paper                    569
dinner rolls                    567
aluminum foil                   566
coffee/tea                      565
shampoo                         562
beef                            561
paper towels                    556
milk                            555
flour                           555
butter                          555
mixes                           554
all- purpose                    551
dishwashing liquid/detergent    551
ketchup                         548
yogurt                          545
individual meals                544
tortillas                       543
pasta                           542
laundry detergent               542
sandwich bags                   536
spaghetti sauce                 536
sugar                           533
pork                            531
fruits                          529
sandwich loaves                 523
hand soap                       502
Name: Product, dtype: int64
```

Unique Products and their counts in all orders.

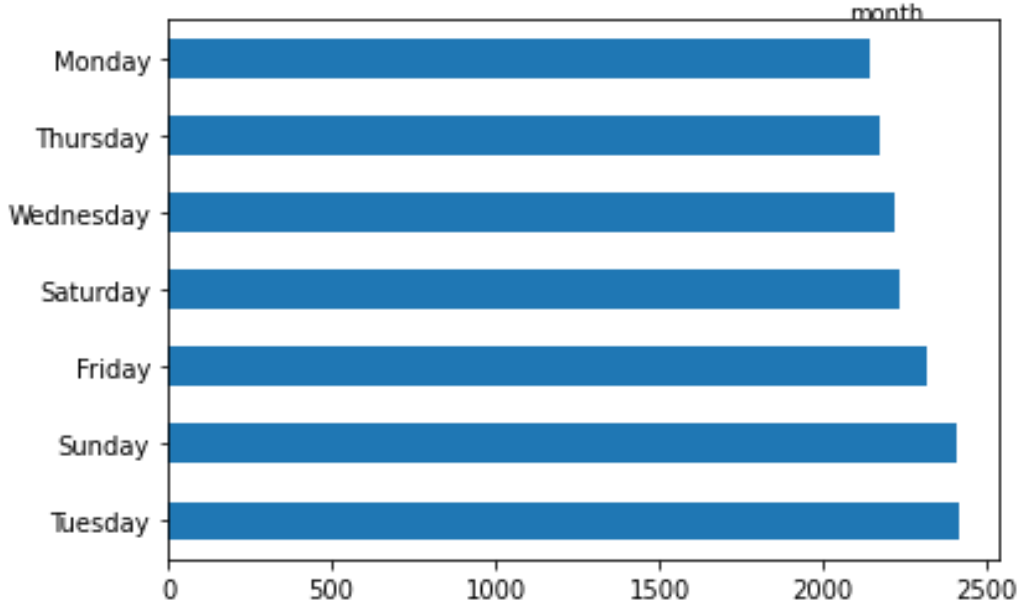
EDA BY DATES (YEAR, MONTH AND DAY OF WEEK)



Data in 2018 and 2019 is similar, data in 2020 is less since we have data till Feb only

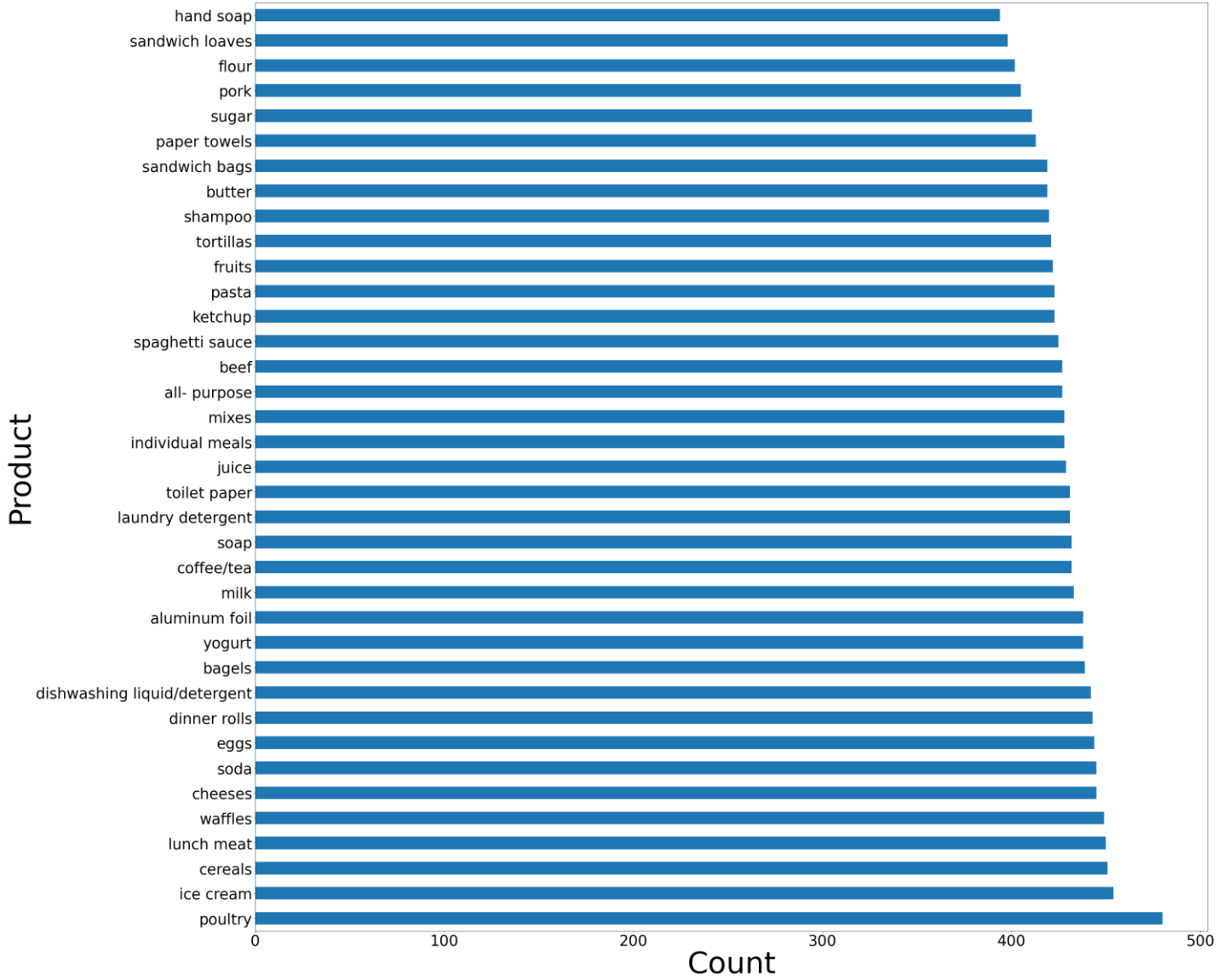


We can see maximum sales in Jan as customers may restock the supplies after the winter break where store wasn't functional



Sales are more on weekends and Tuesdays, it is possible that there might be some discounts or offers on Tuesdays

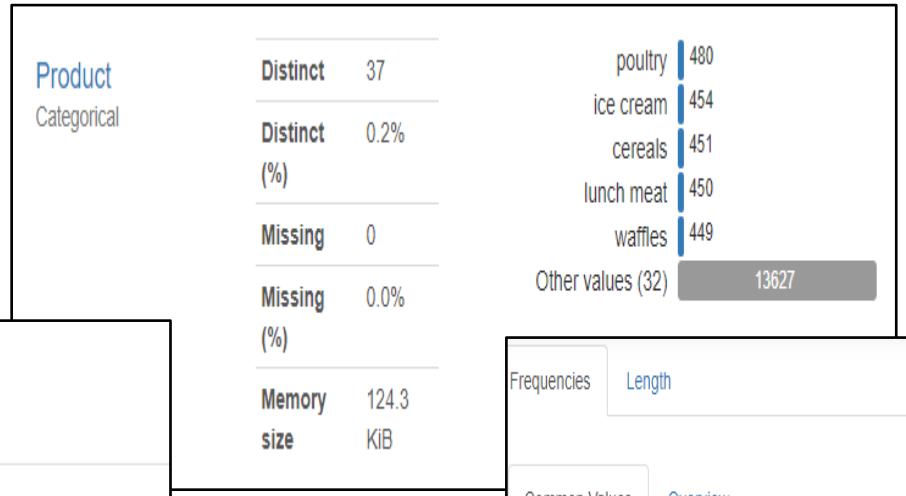
EDA BY PRODUCT AND ITS COUNT

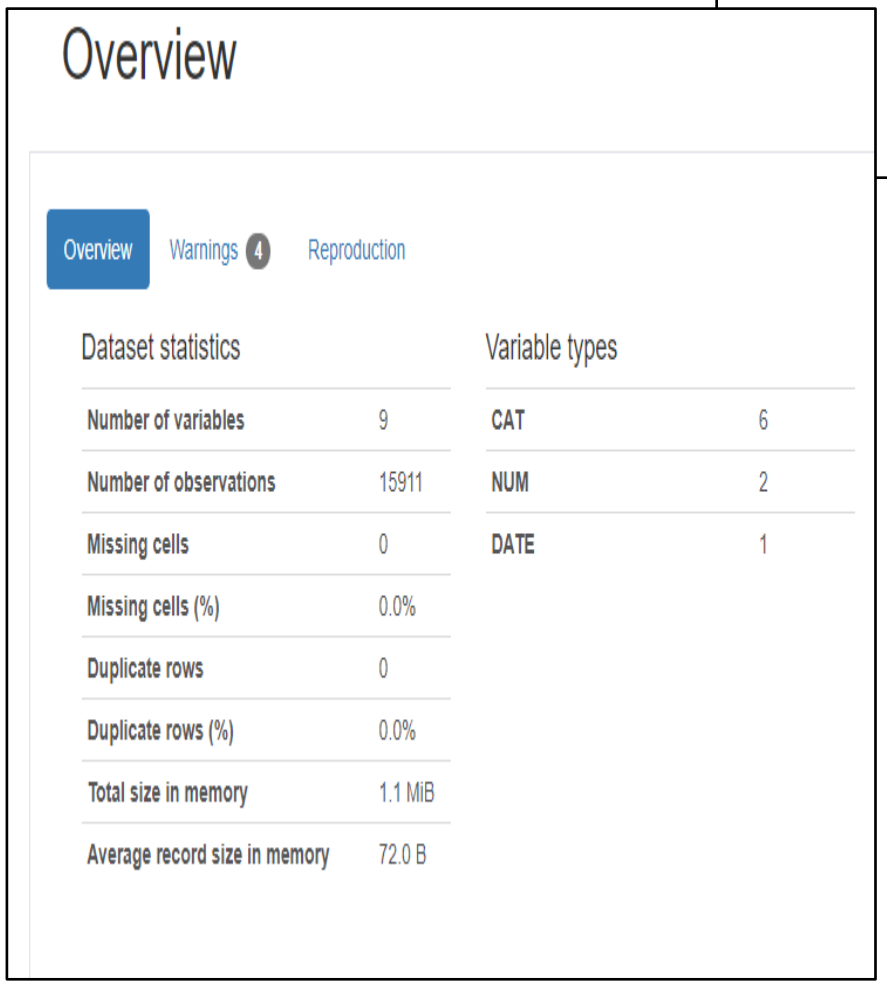


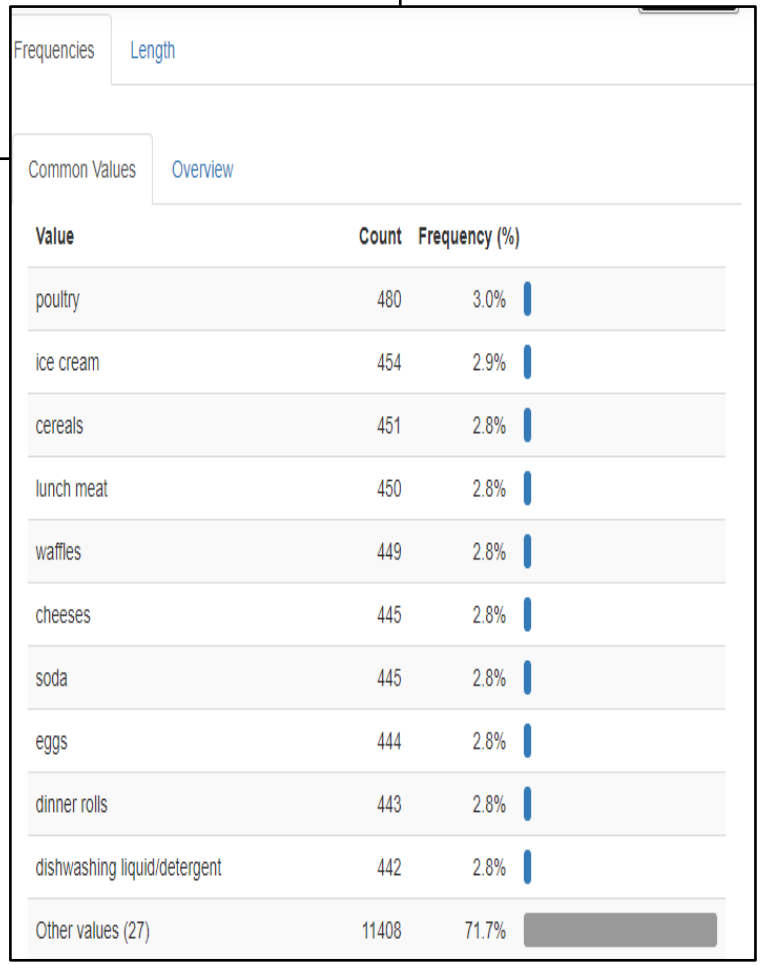
PANDAS PROFILING

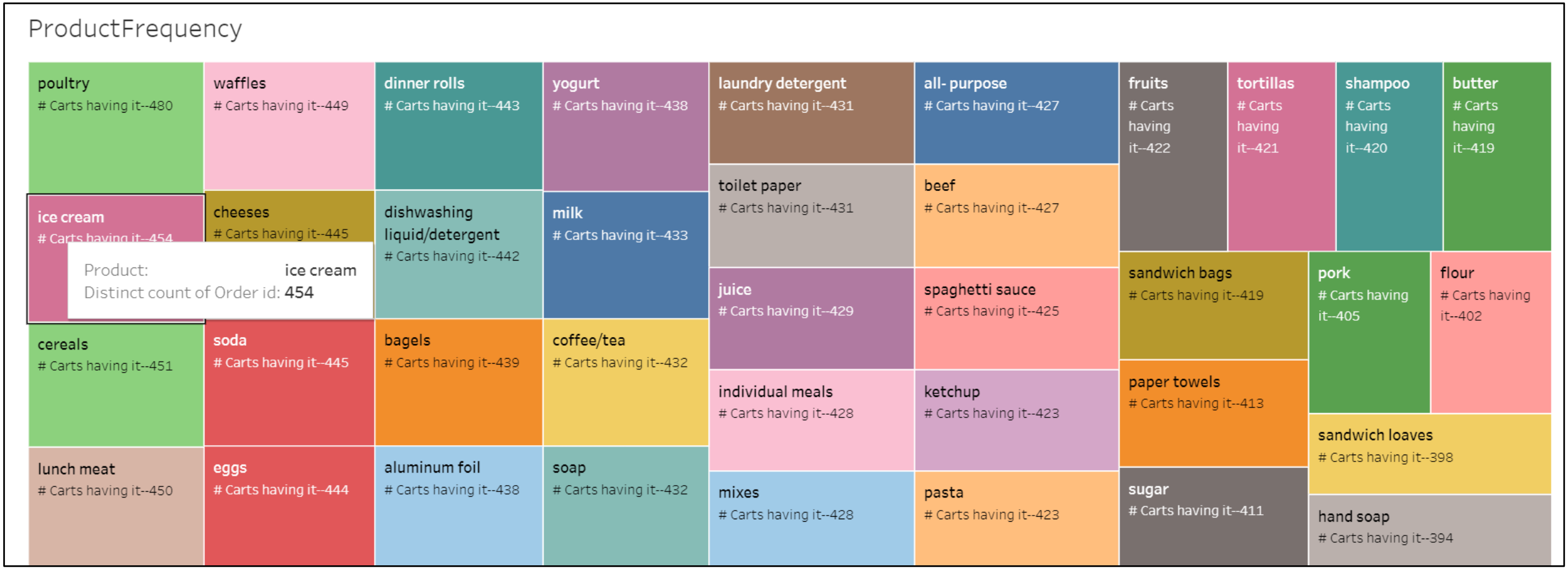
We use pandas profiling for overall analysis of the data, its distribution, eda and its visualization according to variables using profile_report()

Example – Product variables in profile report : -







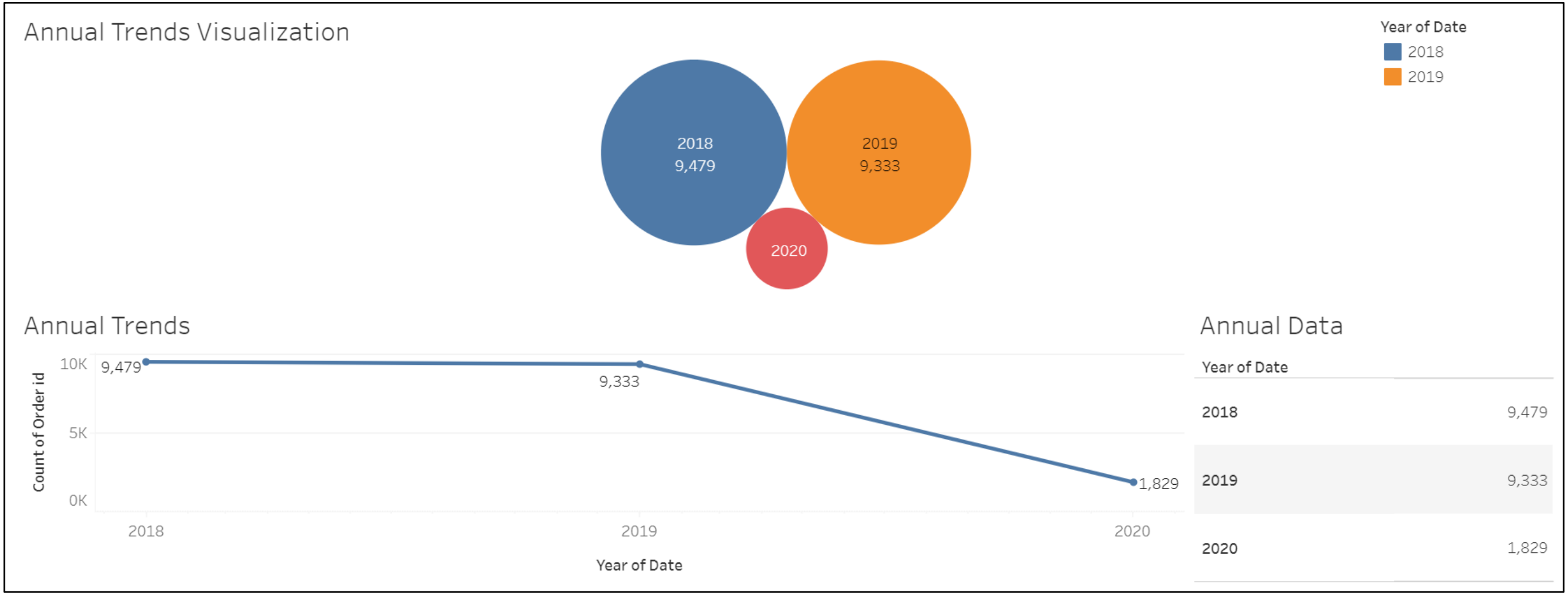


VISUALIZATION TO SEE DEMAND ACROSS DIFFERENT PRODUCTS

This is done using Tableau

ANNUAL TRENDS

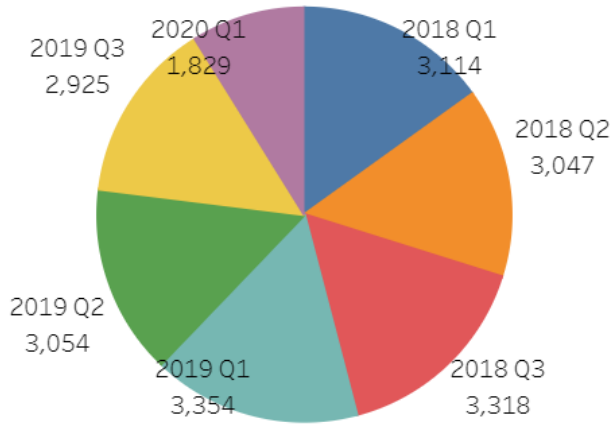
The trends seem normal as per the graphs, not much difference in 2018 and 2019

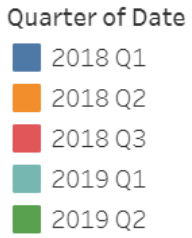


QUARTERLY TRENDS

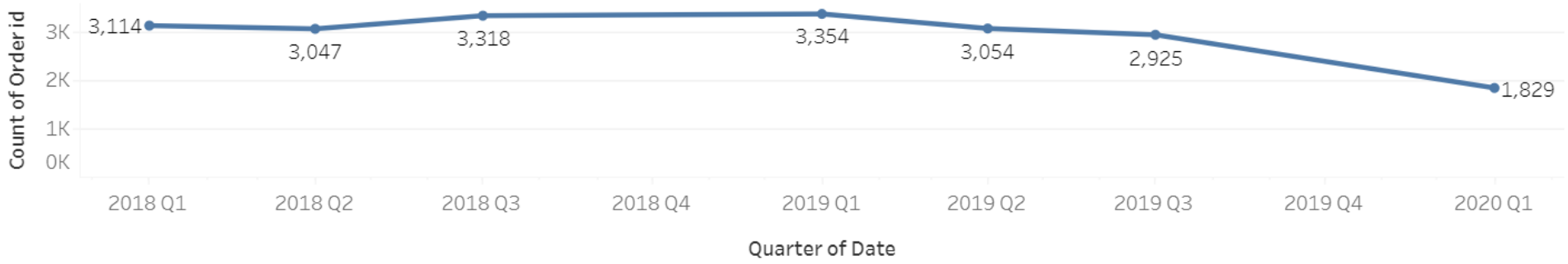
We can see that sales in Q1 are high and that trend continues in both 2018 and 2019, we have data till Feb in 2020, so it would be bias to look at it.

Quarterly Trends Visualization





Quarterly Trends



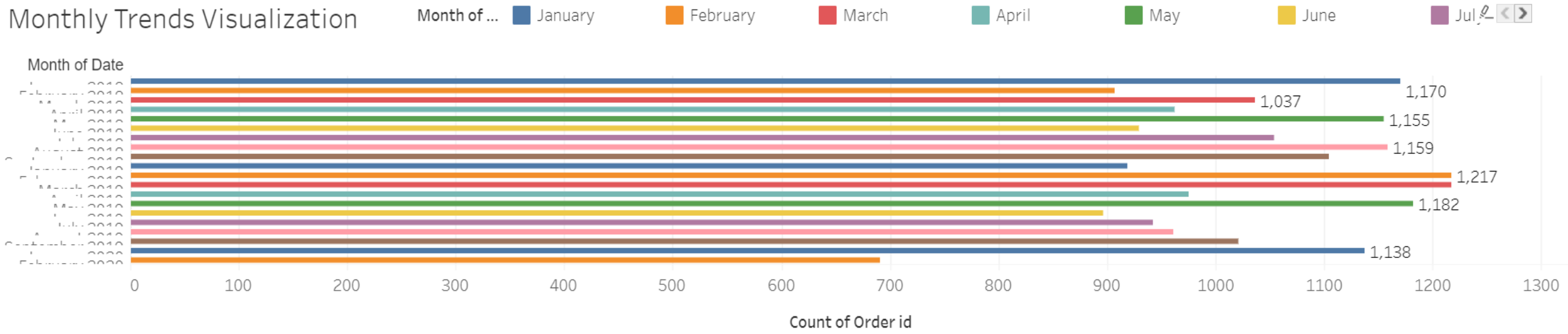
Quarterly Data

Quarter of ..	
2018 Q1	3,114
2018 Q2	3,047
2018 Q3	3,318
2019 Q1	3,354
2019 Q2	3,054
2019 Q3	2,925
2020 Q1	1,829

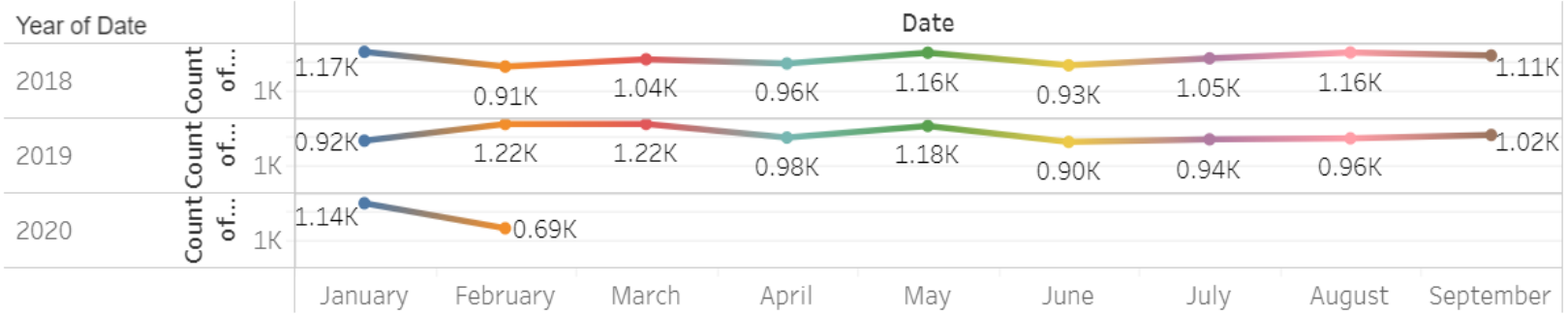
MONTHLY TRENDS

We can see Jan in 2018,2020 Feb and Mar in 2019 and May in both the years contributing to major sales, sales has slightly dropped in Q3 of 2019 and again picked up in Q1 of 2020

Monthly Trends Visualization



Monthly Trends



Monthly Data

Month of Date	Date 2018	2019	2020
January	1,170	919	1,138
February	907	1,217	691
March	1,037	1,218	
April	962	975	
May	1,155	1,182	
June	930	897	
July	1,054	943	
August	1,159	961	
September	1,105	1,021	

WEEKLY TRENDS

Sales on weekends and Tuesdays show a high trends compared to other weekdays

EDA INFERENCES

For more clearer visualization, we can refer - https://public.tableau.com/views/Aadi_MarketBasketAnalysis/DaywiseProducts?:language=en-US&:display_count=n&:origin=viz_share_link

From the EDA, we can see that there are 37 different products that are being sold by the grocery store.

The sales in 2018 and 2019 were similar, but we can see that the data is present only from Jan to Sept, so it is possible that the grocery shop won't be functioning in the winter season.

The poultry products are sold the most and major sales happen in Jan when the store re-opens i.e customers replenish their stock again.

We can see that the sales have decreased since Q4 of 2019, it is possible that the offers may not be exclusive or there might be change in demands.

In weekly trends, we can notice that most sales or orders are placed on the weekends. There are no drastic differences, but weekend sales are visibly higher.

EXECUTIVE SUMMARY

- We can see that poultry products have maximum sales followed by ice cream whereas bread loaves and hand soaps have the least demand.
- The annual sale trends don't differ much, in quarterly trends we see that Q4 has no sales, so we assume that the store won't function during winter. Jan, Feb, Mar, May have maximum orders placed, assuming customers to restock goods after winter and buy newer products when season changes.
- Weekends and Tuesdays have the best sales.
- The data has only 3 columns and to analyse the data, we have grouped them on order_id.

MARKET BASKET ANALYSIS

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations 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 between the items that people buy.

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.

Benefits of Market Basket Analysis-

- Market basket analysis can increase sales and customer satisfaction. Using data to determine that products are often purchased together, retailers can optimize product placement, offer special deals and create new product bundles to encourage further sales of these combinations.
- These improvements can generate additional sales for the retailer, while making the shopping experience more productive and valuable for customers. By using market basket analysis, customers may feel a stronger sentiment or brand loyalty toward the company.

PRACTICAL APPLICATIONS OF MARKET BASKET ANALYSIS

Ecommerce- A list of potentially interesting products for Amazon. Amazon informs the customer that people who bought the item being purchased by them, also reviewed or bought another list of items. A list of applications of Market Basket Analysis in various industries is listed below:

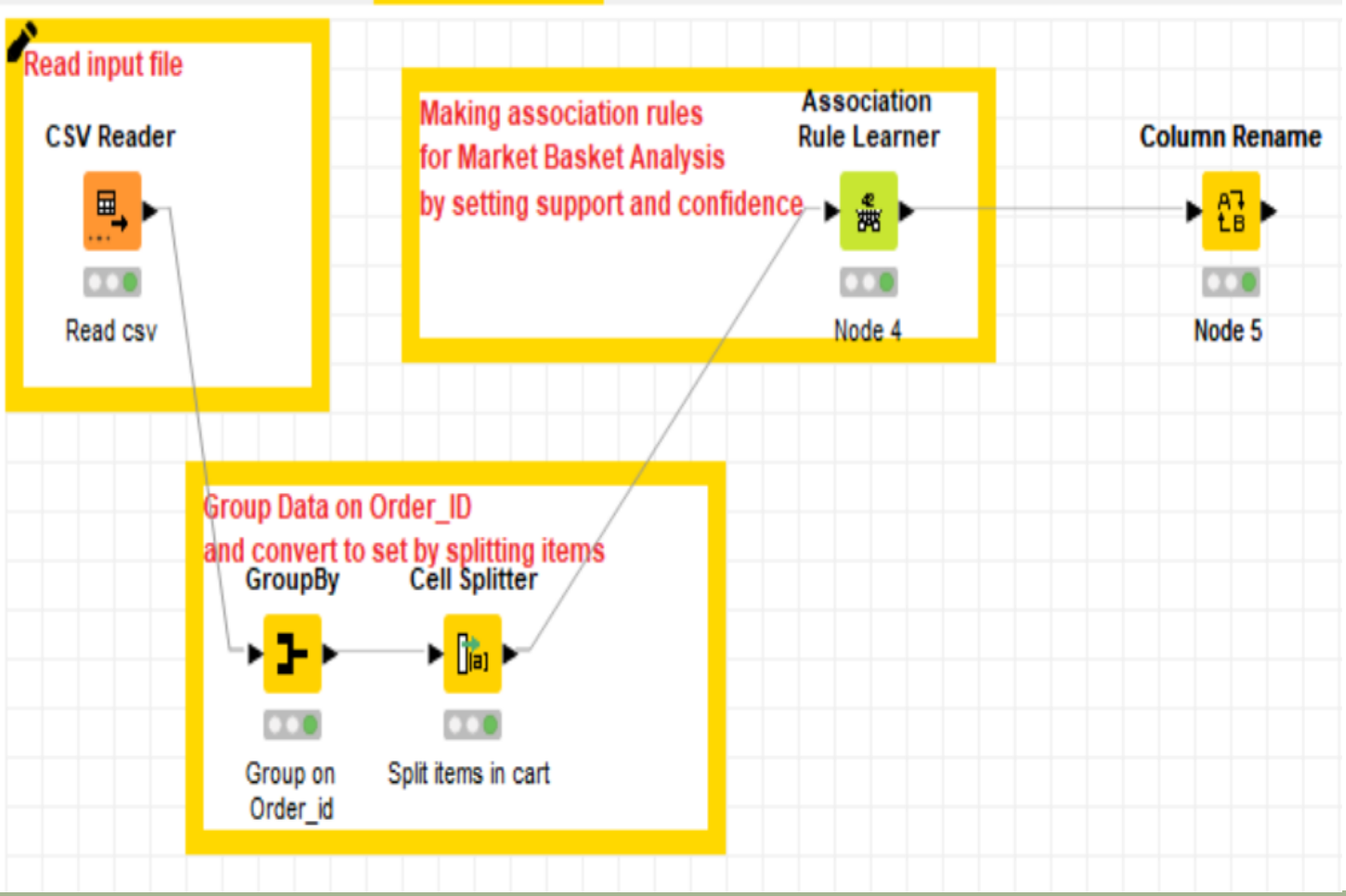
Retail- In Retail, Market Basket Analysis can help determine what items are purchased together, purchased sequentially, and purchased by season. This can assist retailers to determine product placement and promotion optimization (for instance, combining product incentives). Does it make sense to sell soda and chips or soda and crackers?

Telecommunications- In Telecommunications, where high churn rates continue to be a growing concern, Market Basket Analysis can be used to determine what services are being utilized and what packages customers are purchasing. They can use that knowledge to direct marketing efforts at customers who are more likely to follow the same path. For instance, Telecommunications these days is also offering TV and Internet. Creating bundles for purchases can be determined from an analysis of what customers purchase, thereby giving the company an idea of how to price the bundles. This analysis might also lead to determining the capacity requirements.

Banks. In Financial (banking for instance), Market Basket Analysis can be used to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities.

Insurance. In Insurance, Market Basket Analysis can be used to build profiles to detect medical insurance claim fraud. By building profiles of claims, you are able to then use the profiles to determine if more than 1 claim belongs to a particular claimee within a specified period of time.

Medical. In Healthcare or Medical, Market Basket Analysis can be used for comorbid conditions and symptom analysis, with which a profile of illness can be better identified. It can also be used to reveal biologically relevant associations between different genes or between environmental effects and gene expression.



- Read the CSV file.
- We group the data on order_id and concatenate the products to make it a single entry.
- No we split the items in cart as different entries.
- Now we use Association Rule Learner by setting the values for support as 0.05 and confidence as 0.5 to derive the association rules.
- We initially start by setting the support value as 0.1 and confidence as 0.8 and keep on iterating by reducing the values until we get the desired number of rules.
- Now we rename the columns to readable and interpretable formats.
- Now we read the association rules by sorting them on Lift and then suggest recommendations and combo offers.

MARKET BASKET ANALYSIS - USING KNIME (AND THRESHOLDS)

SUPPORT, CONFIDENCE & LIFT

- **SUPPORT** is the number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
- **CONFIDENCE** of the rule is the ratio of the number of transactions that include all items in {Y} as well as the number of transactions that include all items in {X} to the number of transactions that include all items in {X}.
- **LIFT** or lift ratio is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.
  - Lift = 1; implies no relationship between X and Y (i.e., X and Y occur together only by chance)
  - Lift > 1; implies that there is a positive relationship between X and Y (i.e., X and Y occur together more often than random)
  - Lift < 1; implies that there is a negative relationship between X and Y (i.e., X and Y occur together less often than random)

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

$$Rule: X \Rightarrow Y \rightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

EXAMPLE

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

$$\text{Support} = \frac{(A + B)}{\text{Total}}$$

$$\text{Support for Basket 1} = \frac{(\text{Milk} + \text{Cheese})}{\text{Total}} = \frac{6}{9} = .6666667$$

$$\text{Confidence} = \frac{(A + B)}{A}$$

$$\text{Confidence for Basket 1} = \frac{(\text{Milk} + \text{Cheese})}{\text{Milk}} = \frac{6}{6} = 1.000$$

$$\text{Lift} = \left(\frac{\left(\frac{(A + B)}{A}\right)}{\left(\frac{B}{\text{Total}}\right)}\right)$$

$$\text{Lift for Basket 1} = \left(\frac{\left(\frac{(\text{Milk} + \text{Cheese})}{\text{Milk}}\right)}{\left(\frac{\text{Cheese}}{\text{Total}}\right)}\right) = \left(\frac{\left(\frac{6}{6}\right)}{\left(\frac{7}{9}\right)}\right) = \left(\frac{1}{.7777778}\right) = 1.2857$$

ASSOCIATION RULES

These are the top18 association rules which were derived when we set the thresholds as minimum support to 0.05 and minimum confidence as 0.6 sort them on lift values as it is the strongest measure to show association.

Setting these thresholds, we derive 24 rules.

# Rules	Support	Confidence	Lift	Recommended Items		Items in Cart
Rule #24	0.05531	0.64948	1.79119	paper towels	<---	[eggs, ice cream, pasta]
Rule #23	0.05531	0.64286	1.73100	pasta	<---	[paper towels, eggs, ice cream]
Rule #7	0.05092	0.67442	1.72621	cheeses	<---	[bagels, cereals, sandwich bags]
Rule #1	0.05004	0.64045	1.70040	juice	<---	[yogurt, toilet paper, aluminum foil]
Rule #5	0.05092	0.63043	1.67772	mixes	<---	[yogurt, poultry, aluminum foil]
Rule #6	0.05092	0.61053	1.65964	sandwich bags	<---	[cheeses, bagels, cereals]
Rule #19	0.05356	0.64211	1.65092	dinner rolls	<---	[spaghetti sauce, poultry, laundry detergent]
Rule #13	0.05180	0.64130	1.64886	dinner rolls	<---	[spaghetti sauce, poultry, ice cream]
Rule #2	0.05004	0.61957	1.64495	juice	<---	[yogurt, poultry, aluminum foil]
Rule #14	0.05180	0.68605	1.62793	poultry	<---	[dinner rolls, spaghetti sauce, ice cream]
Rule #17	0.05180	0.63441	1.62746	eggs	<---	[paper towels, dinner rolls, pasta]
Rule #18	0.05180	0.60204	1.62110	pasta	<---	[paper towels, eggs, dinner rolls]
Rule #9	0.05092	0.63043	1.62091	dinner rolls	<---	[spaghetti sauce, poultry, cereals]
Rule #22	0.05531	0.63000	1.61615	eggs	<---	[paper towels, ice cream, pasta]
Rule #3	0.05004	0.61290	1.61596	coffee/tea	<---	[yogurt, cheeses, cereals]
Rule #15	0.05180	0.62766	1.61378	dinner rolls	<---	[spaghetti sauce, poultry, juice]
Rule #12	0.05180	0.62766	1.61014	eggs	<---	[dinner rolls, poultry, soda]
Rule #11	0.05092	0.60417	1.58925	milk	<---	[poultry, laundry detergent, cereals]

INTERPRETING THE ASSOCIATION RULES

Let us look at the first rule**(Rule #24)** -

Support: Items in Basket#24 {eggs, ice cream, pasta} are present in almost 5.5% of total orders.

Confidence: The probability that someone will buy paper towels given that Basket#24 items {eggs, ice cream, pasta} were already purchased is 64.9%

Lift: It is given as **confidence** ({eggs, ice cream, pasta ⇒ paper towels})/**support** ({paper towels}) =1.79 i.e >1 which means that there is a positive relationship in the association i.e strong association.

Let us look at the first rule**(Rule #23)** -

Support: Items in Basket#23 {eggs, ice cream, paper towels} are present in almost 5.5% of total orders.

Confidence: The probability that someone will buy pasta given that Basket#23 items {eggs, ice cream, paper towels} were already purchased is 64.2%

Lift: It is given as **confidence** ({eggs, ice cream, paper towels ⇒ pasta})/**support** ({pasta}) =1.73 i.e >1 which means that there is a positive relationship in the association i.e strong association.

ASSOCIATION RULES

These are some of the association rules which were derived when we set the thresholds as minimum support to 0.05 and minimum confidence as 0.55 and 0.5 respectively
sort them on lift values as it is the strongest measure to show association.

Setting these thresholds, we derive 84 rules and 1247 rules respectively.

Confidence = 0.55

# Rules	Support	Confidence	Lift	Recommended Items		Items in Cart
Rule #29	0.05180	0.56731	1.45861	dinner rolls	<---	[eggs, poultry, soda]
Rule #28	0.05180	0.56731	1.45205	soda	<---	[eggs, dinner rolls, poultry]
Rule #31	0.05180	0.56731	1.45205	soda	<---	[eggs, dinner rolls, pasta]
Rule #69	0.08692	0.56250	1.44625	dinner rolls	<---	[poultry, hand soap]
Rule #26	0.05092	0.56863	1.43607	cereals	<---	[poultry, milk, laundry detergent]
Rule #68	0.08692	0.55932	1.43484	eggs	<---	[beef, soda]
Rule #61	0.08341	0.55556	1.43162	dishwashing liquid/detergent	<---	[mixes, soda]
Rule #39	0.05180	0.60204	1.42859	poultry	<---	[dinner rolls, spaghetti sauce, juice]
Rule #12	0.05004	0.60000	1.42375	poultry	<---	[dishwashing liquid/detergent, laundry detergent, mixes]
Rule #71	0.08780	0.55556	1.42197	cheeses	<---	[cereals, sandwich bags]
Rule #81	0.09482	0.55385	1.42079	eggs	<---	[dinner rolls, soda]
Rule #66	0.08604	0.55367	1.42034	eggs	<---	[paper towels, dinner rolls]
Rule #51	0.07287	0.55333	1.41628	soda	<---	[sandwich bags, sugar]
Rule #80	0.09219	0.55263	1.41449	soda	<---	[eggs, soap]
Rule #56	0.08077	0.55090	1.41005	cheeses	<---	[shampoo, sandwich bags]
Rule #16	0.05092	0.58586	1.39019	poultry	<---	[yogurt, mixes, aluminum foil]
Rule #27	0.05092	0.58586	1.39019	poultry	<---	[milk, laundry detergent, cereals]

Confidence = 0.5

# Rules	Support	Confidence	Lift	Recommended Items		Items in Cart
Rule #69	0.06673	0.50000	1.18646	poultry	<---	[pasta, pork]
Rule #105	0.06936	0.50000	1.18646	poultry	<---	[hand soap, toilet paper]
Rule #106	0.06936	0.50000	1.18646	poultry	<---	[fruits, toilet paper]
Rule #107	0.06936	0.50000	1.18646	poultry	<---	[cheeses, hand soap]
Rule #108	0.06936	0.50000	1.18646	poultry	<---	[hand soap, ketchup]
Rule #110	0.06936	0.50000	1.18646	poultry	<---	[sugar, ketchup]
Rule #232	0.07375	0.50000	1.18646	poultry	<---	[tortillas, pork]
Rule #255	0.07375	0.50000	1.18646	poultry	<---	[sandwich loaves, mixes]
Rule #269	0.07463	0.50000	1.18646	poultry	<---	[lunch meat, pork]
Rule #407	0.07638	0.50000	1.18646	poultry	<---	[shampoo, coffee/tea]
Rule #485	0.07726	0.50000	1.18646	poultry	<---	[tortillas, coffee/tea]
Rule #661	0.07989	0.50000	1.18646	poultry	<---	[all- purpose, pasta]
Rule #707	0.07989	0.50000	1.18646	poultry	<---	[waffles, sugar]
Rule #892	0.08253	0.50000	1.18646	poultry	<---	[mixes, ketchup]
Rule #907	0.08253	0.50000	1.18646	poultry	<---	[spaghetti sauce, cheeses]
Rule #961	0.08341	0.50000	1.18646	poultry	<---	[ice cream, soap]

RECOMMENDATIONS AND OFFERS SUGGESTED

We can recommend keeping the paper towels near the stands where breads, eggs and other edible items are placed.

We can also place the things used in making sandwiches separately, like bread loaves, sandwich bags, beefs and vegetables differently, so that customer might roam around the store and eventually end up buying more products.

We also recommend the grocery store to launch an online application that serves their customers even during winters , so that the sales continue to increase rather than having no sales for 3 months in a row

We can also offer combo discounts like Buying pasta + ice cream + paper rolls will offer 50% off on buying eggs.

We see that hand soaps are not selling as good as other products, so we can offer discounts on hand soaps like 10-20% based on the cart value.

We can offer full meal offers like {bread loaves + sandwich bags + aluminium foils + butter + kechup + mixes + cheese} for 10% off on cart order + paper rolls FREE