

Marketing & Retail Analytics. Milestone 2

THILAK RAJ | Batch: 18 FEB 2022

Content of the Presentation:

- **Problem Statement.**
- **Data Analysis**
 - Data Info, Shape and Summary of the Data
- **Exploratory Data Analysis and Inferences.**
 - Charts
 - Summary and recommendations.
- **Market Basket Analysis.**
 - KNIME Work Flow
 - Rule parameters
 - Output and recommendations

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)

Agenda:

- Executive Summary of the data
- Exploratory Data Analysis
- Market Basket Analysis
- Associations Identified
- Recommendation

Executive Summary of the data:

We have received the 26 months of data of a grocery store consisting 20641 entries with 3 variable details.

```
# Statistical Summary  
df.describe().T
```

	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

Data Analysis:

```
<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
```

```
# nulls in data set  
df.isnull().sum()
```

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

```
# duplicate in data set  
df.duplicated().sum()
```

```
4730
```

Data Info:

- Dataset is having 20641 variables.
- 1 is Categorical variables, 1 is numerical variable and one is date and time field variable.
- Found no missing values.
- Found 4730 duplicated data.
- Total number of observations are 20641.

Duplicate Values:

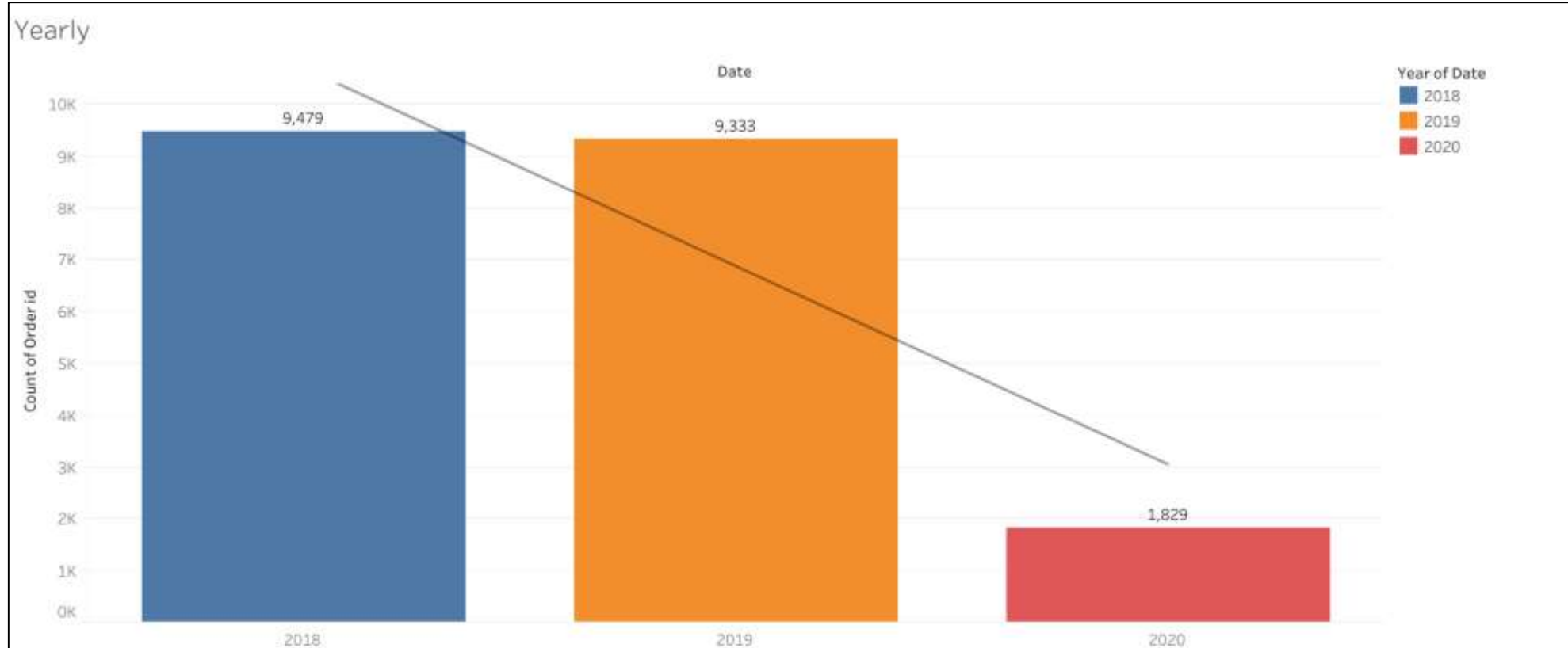
- It is generally a good practice to drop duplicate rows in a dataset as they do not provide any additional information and can skew the results of any analysis performed on the dataset.
- However, in this particular case, dropping duplicate rows may not be appropriate as there is no unique identifier for each row.
- Each row consists of a Date, a customer ID, and a Product purchased, but the same product can be purchased by multiple customers on the same date.
- Therefore, we drop duplicate rows, it may inadvertently remove valid information from the dataset.
- So duplicate values are not removed from the dataset.

Data Assumption:

- The data represents a list of items purchased at a grocery store on various dates.
- Each entry in the data represents a single item purchased.
- The first column in the data represents the date the item was purchased.
- The second column represents the customer who made the purchase.
- The third column represents the item purchased.
- The same item can be purchased by multiple customers on different dates.
- There is no information provided about the quantity or price of each item.
- We have not dropped the duplicated values.

Exploratory Data Analysis

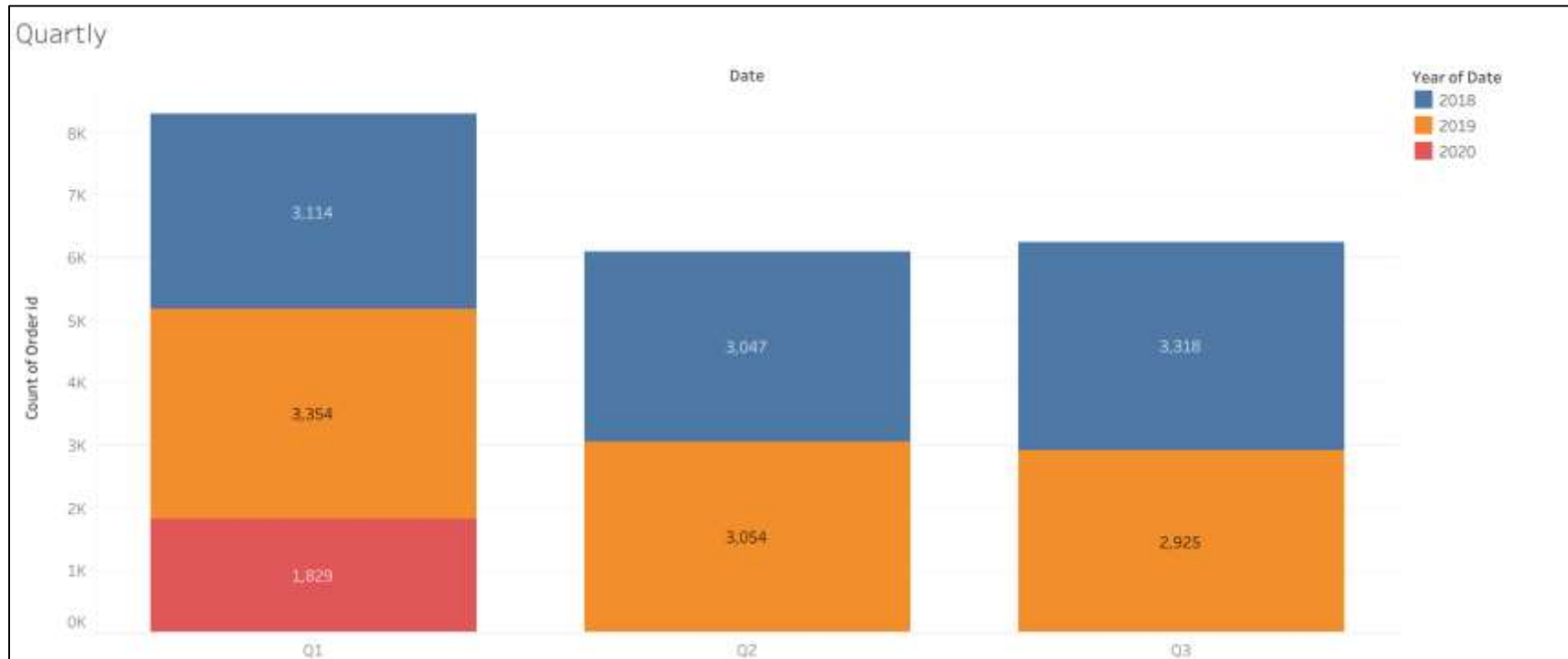
Yearly count of products sold:



- As we have data till 26 Feb 2020 therefore the count of the products sold in 2020 is low.

Exploratory Data Analysis

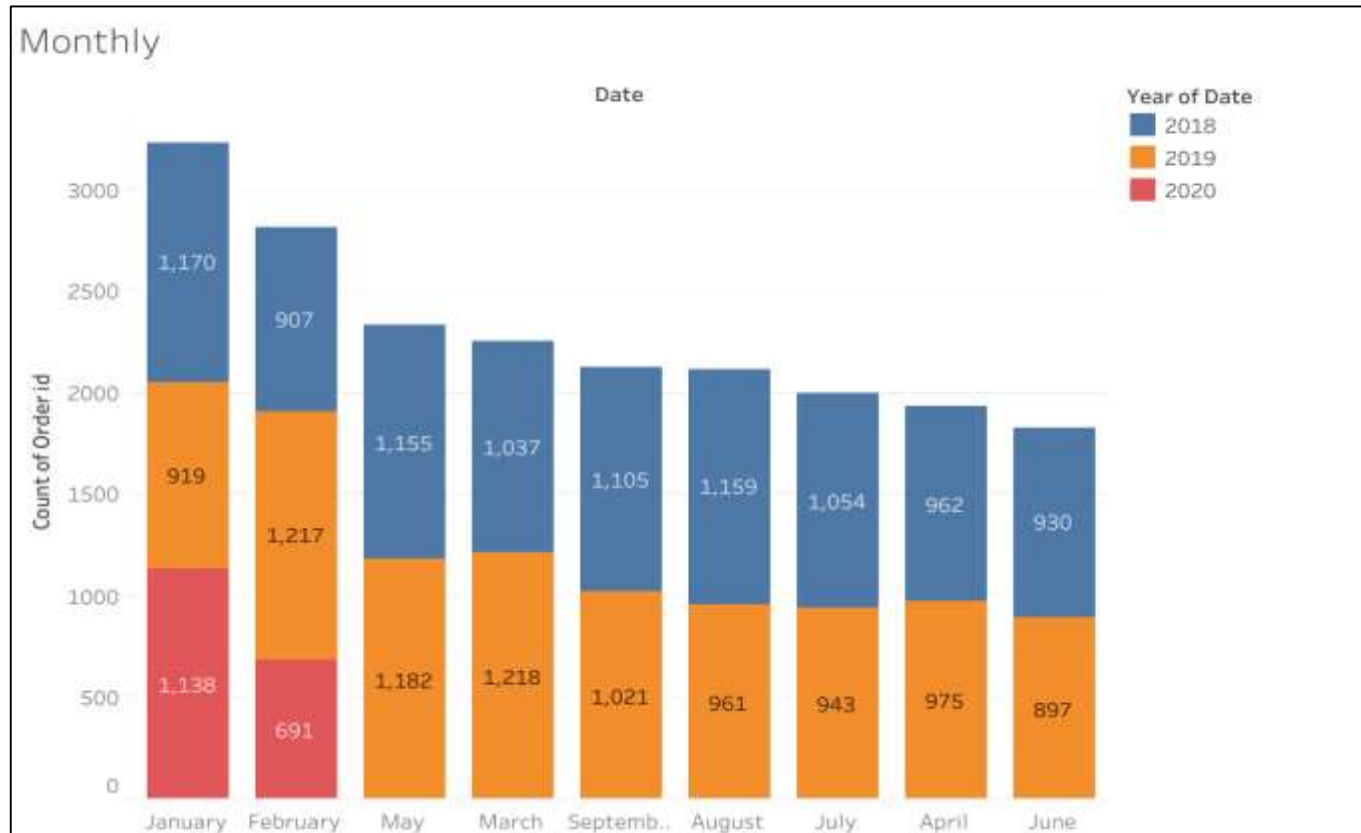
Quarterly Count of Products Sold :



- As we have data till 26th Feb 2020, the count of the products sold in Q1 is more.
- In 2019, Q1 sales was highest
- In 2018, Q3 sales was highest
- Product sold in Q2 is approximately same in the years 2019 and 2018.

Exploratory Data Analysis

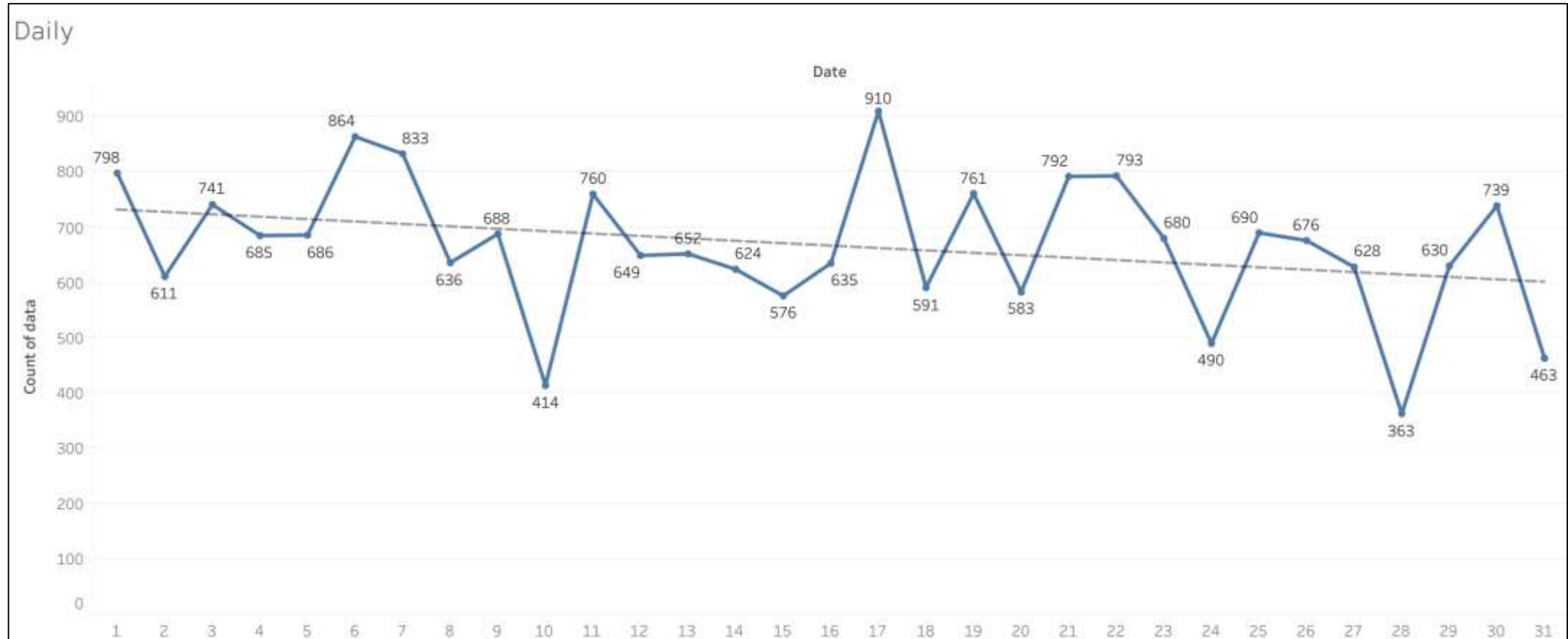
Monthly count of products sold :



- In 2018 most of the products were sold in January and least were sold in February.
- In 2019 most of the products were sold in March and least were sold in January.

Exploratory Data Analysis

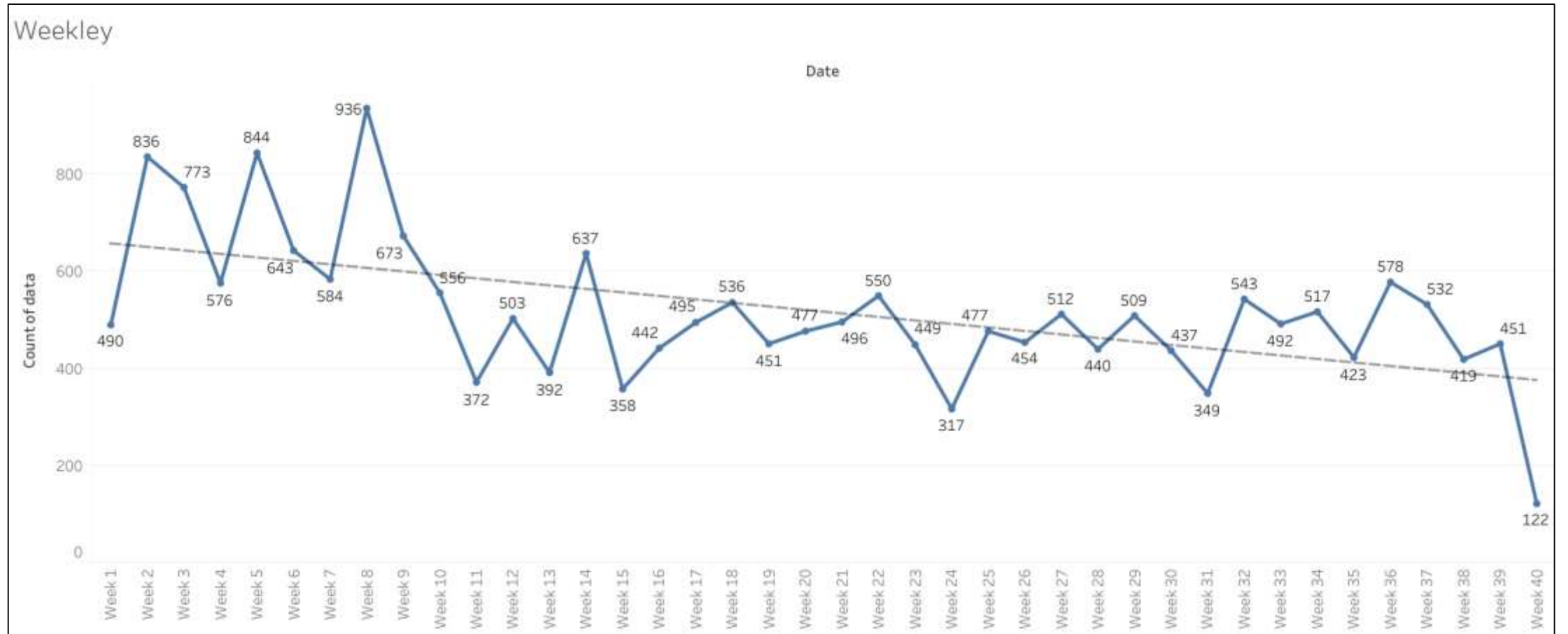
Day of month VS the count of products sold:



- We could see the decrement in the count of sales across the days of the months. The highest sale is on 17th

Exploratory Data Analysis

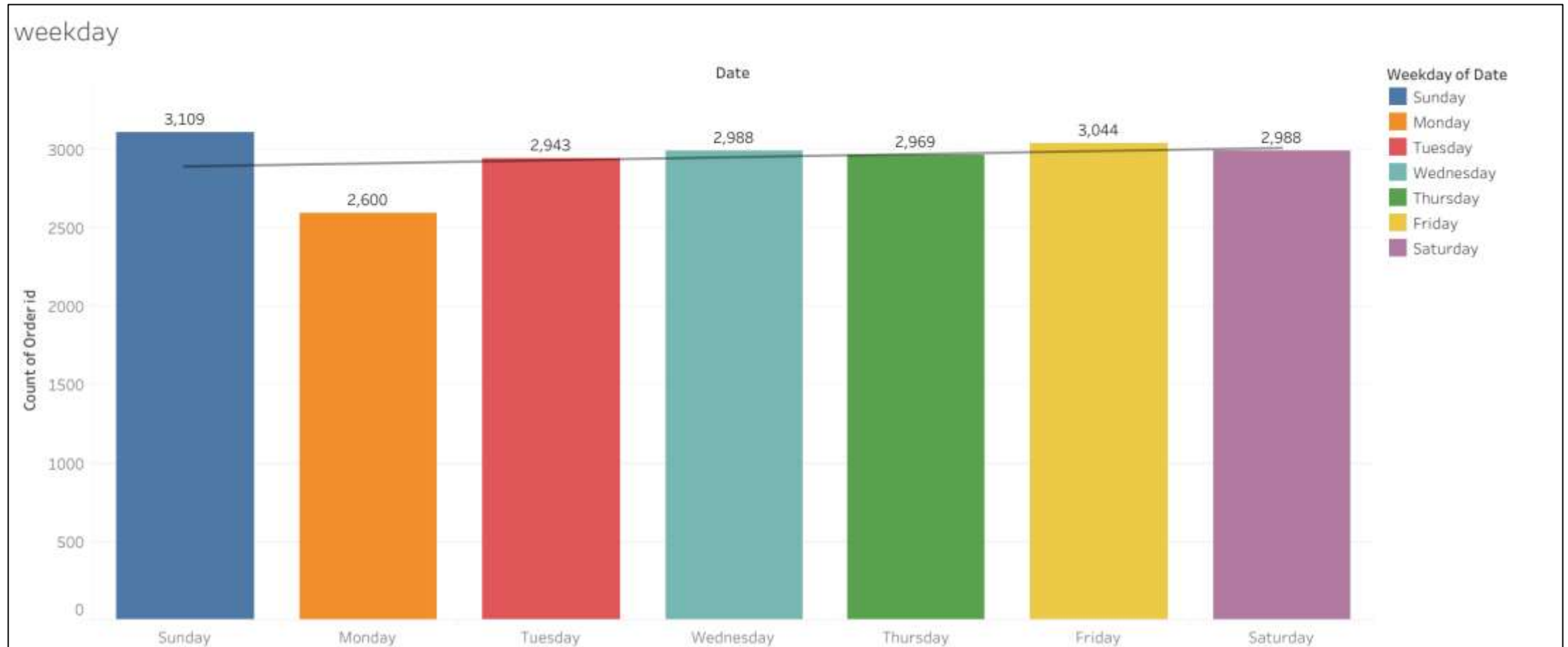
Weekly count of products sold:



- We could see the decrement in the count of sales across weeks which is not a good sign.

Exploratory Data Analysis

Weekday count of products sold:



- We can see the least items sold on Mondays and most items sold on Sundays and Fridays.

Exploratory Data Analysis

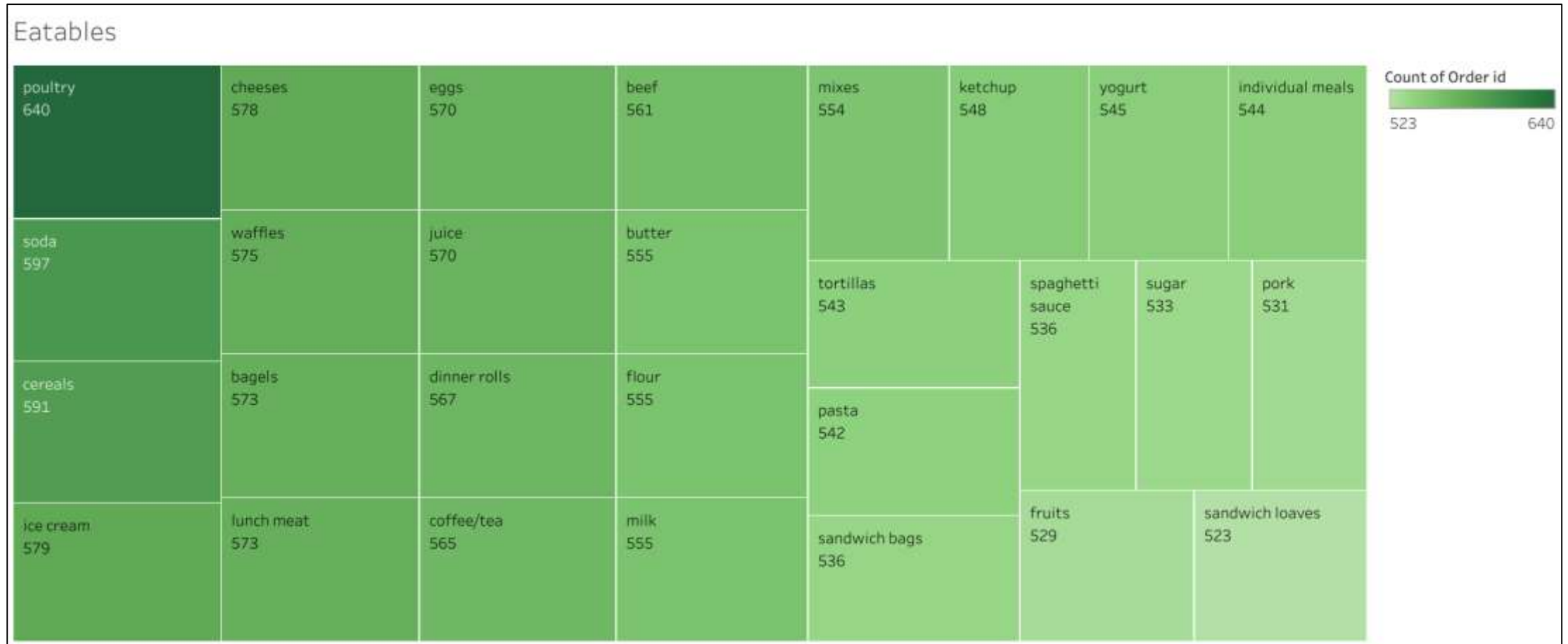
Count of products sold:



- We can see the minimum count is 500 for all the products. Poultry is the only one product with 600+ sales.

Exploratory Data Analysis

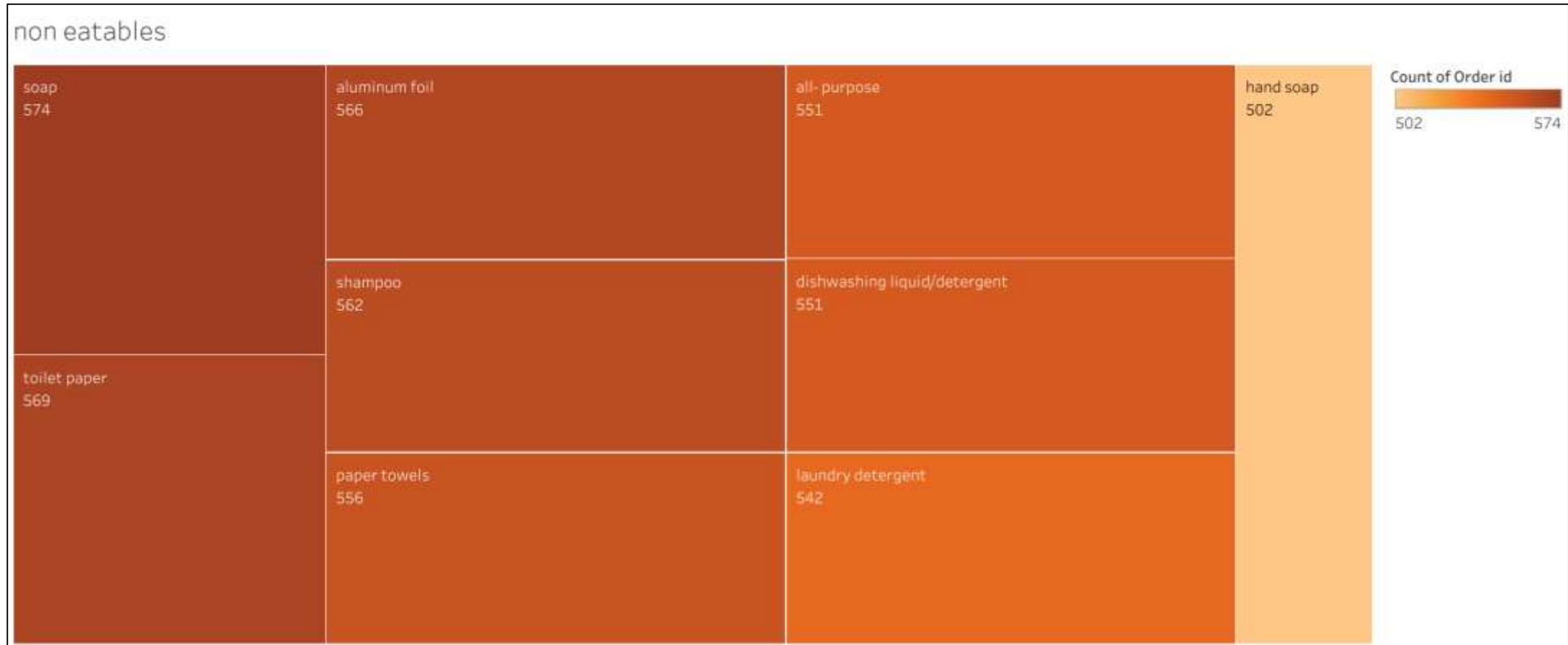
Eatable products sale count heatmap:



- There are total 28 products in this category. poultry , Soda, cereals are the highest and pork, fruits, sandwich loaves are the lowest sold products

Exploratory Data Analysis

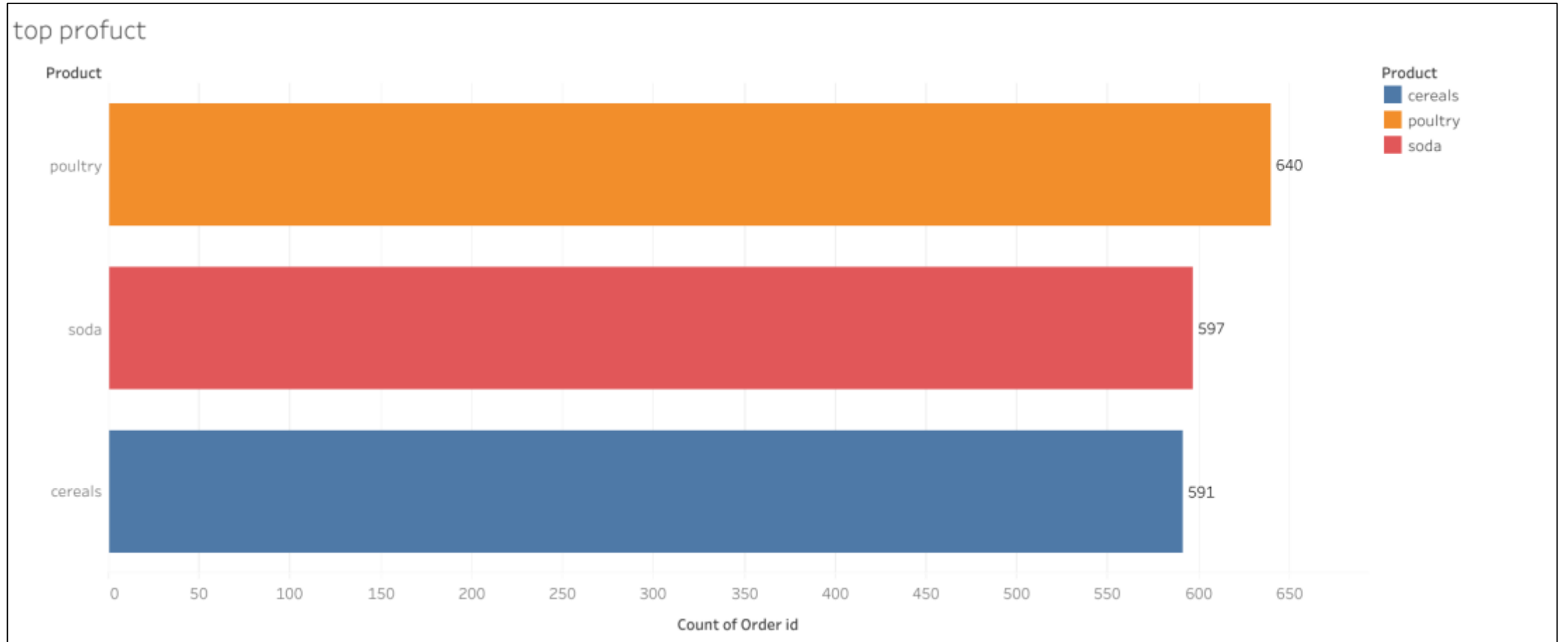
Non-eatable products sale count heatmap:



- There are total 9 products in this category. Soap and Toilet Paper are the highest and Hand Soap is the lowest sold products

Exploratory Data Analysis

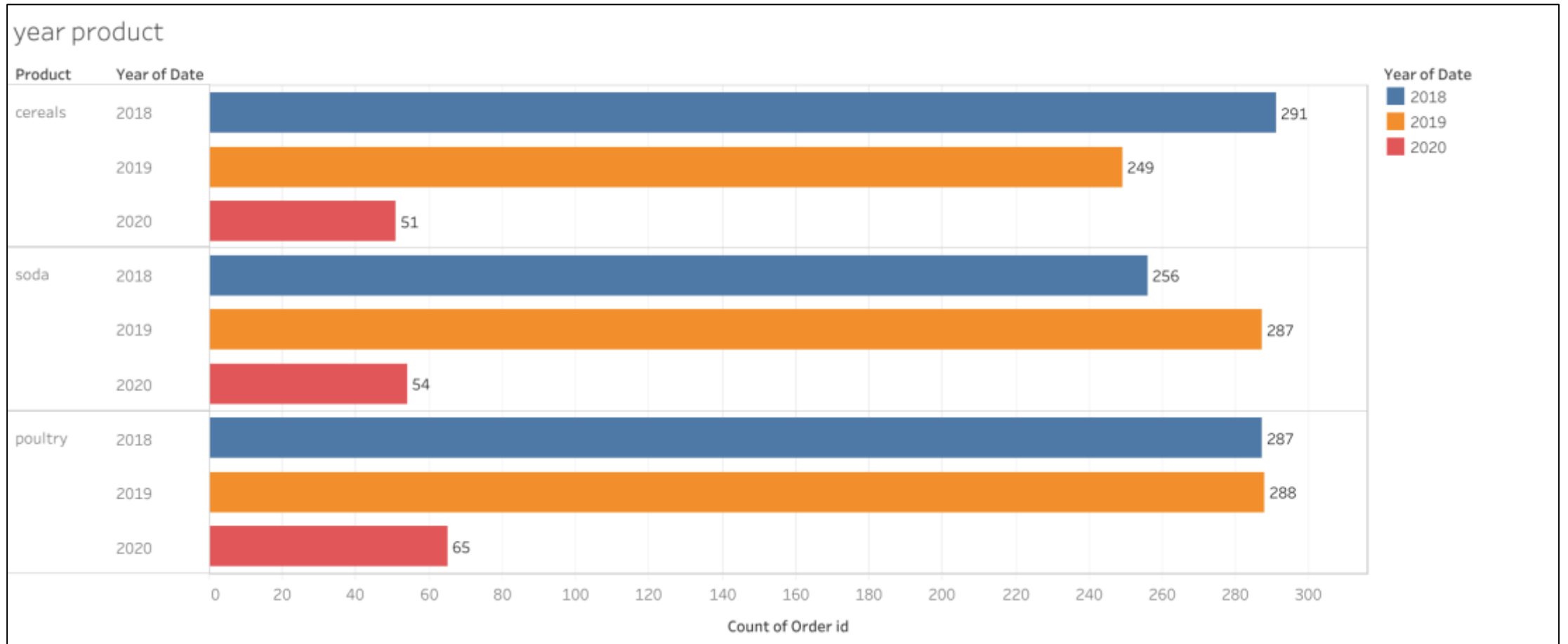
Top 3 products over the years:



- Poultry, Soda and Cereals are the top 3 products over the years. We will explore more about it.

Exploratory Data Analysis

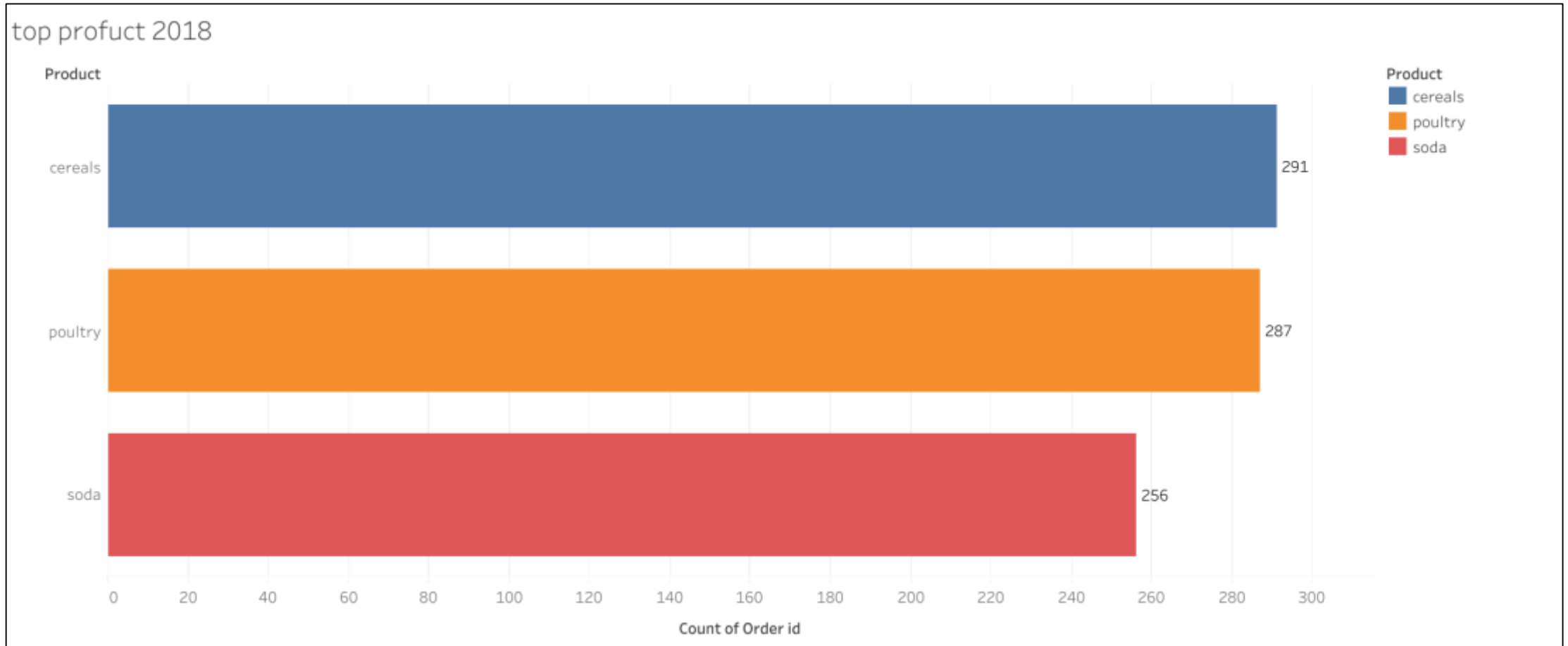
Top 3 products over the years:



- Above is the sale count of top 3 products over the years. Cereal sale count decreased and Soda sale count is increased through the years but the poultry count is almost same if we ignore the year 2020 (Not complete data of 2020).

Exploratory Data Analysis

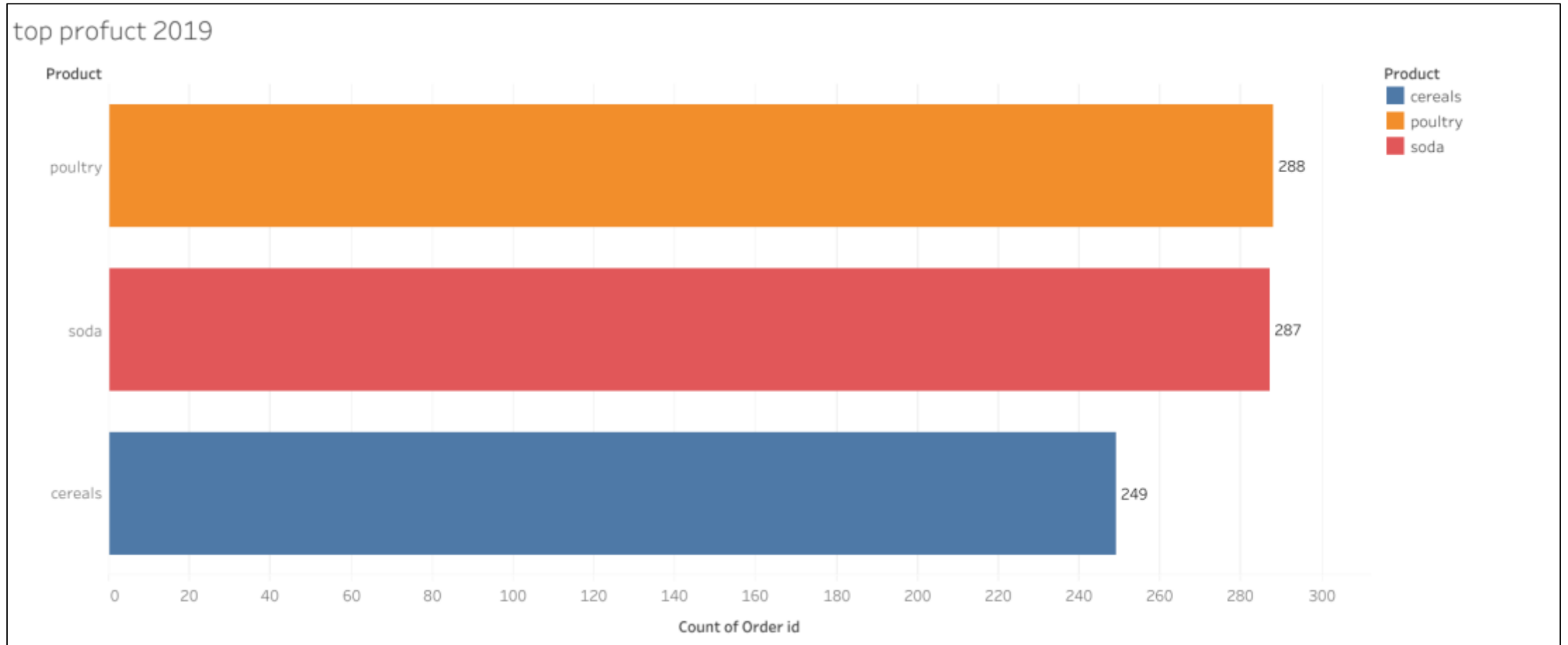
Top 3 products of 2018:



- Cereals, Poultry and Soda are the top 3 products of 2018.

Exploratory Data Analysis

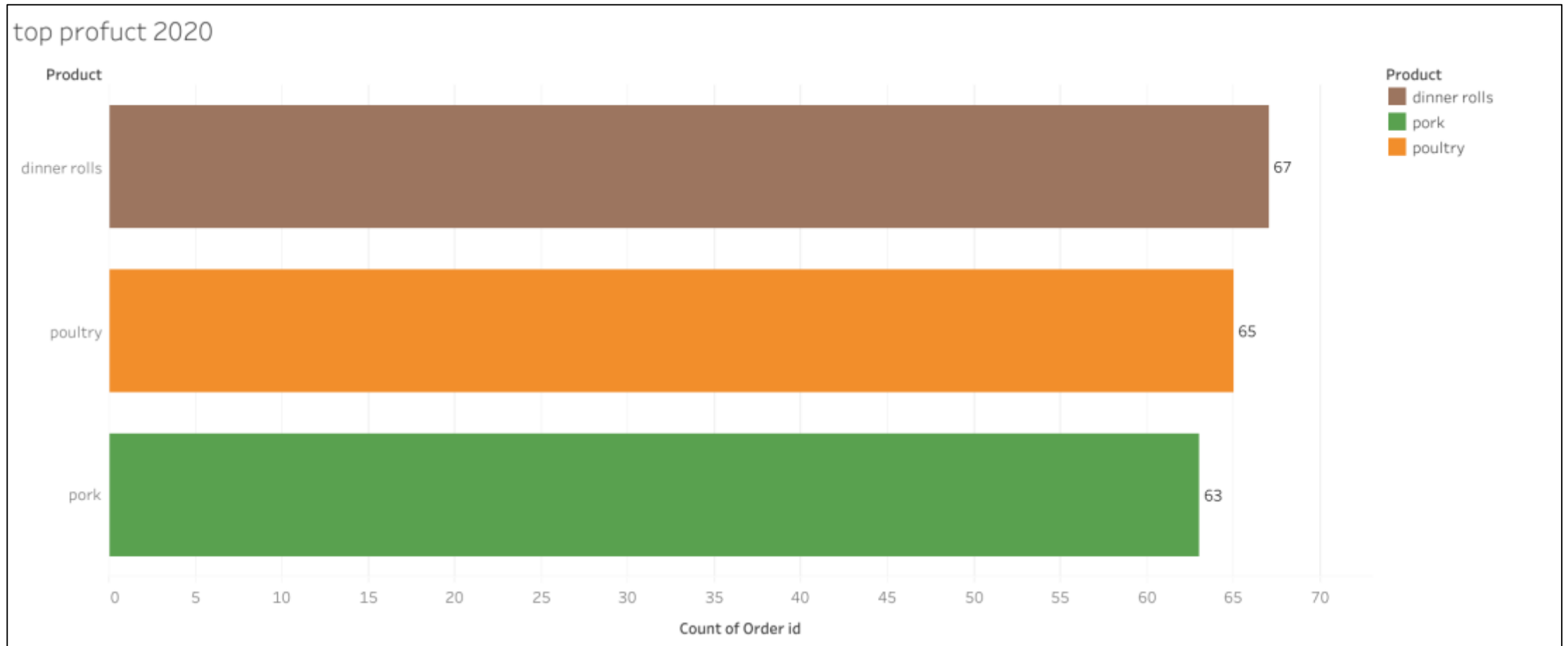
Top 3 products of 2019:



- Poultry, Soda and Cereals are the top 3 products of 2019. Top products are same as of 2018 but the positions are interchanged.

Exploratory Data Analysis

Top 3 products of 2019:



- Dinner role, Poultry and Pork are the top 3 products of 2020 with the limited data.

Summary of Exploratory data analysis:

- Cereals, Poultry, and Soda are the top 3 products sold in 2018 and 2019.
- Dinner rolls, Poultry, and Pork are the top 3 products sold in January and February 2020.
- Poultry, Cereals, and Soda are the highly sold products over the years.
- The highest sold products are Soap and Toilet Paper, and the lowest sold product is Hand Soap in non eatables products.
- The highest sold products in the eatables category are Poultry, Soda, and Cereals, and the least sold products are Pork, Fruits, and Sandwich Loaves.
- Most of the products were sold on Sundays, and the least were sold on Mondays.
- In 2018, the most products were sold in January, and the least were sold in February. In 2019, the most products were sold in March, and the least were sold in January.
- The sales were the highest in Q1 2019 and Q3 2018.
- The count of products sold in Q2 is approximately the same in 2019 and 2018.
- The count of products sold in 2020 is low, possibly due to the data being only until 26th February.

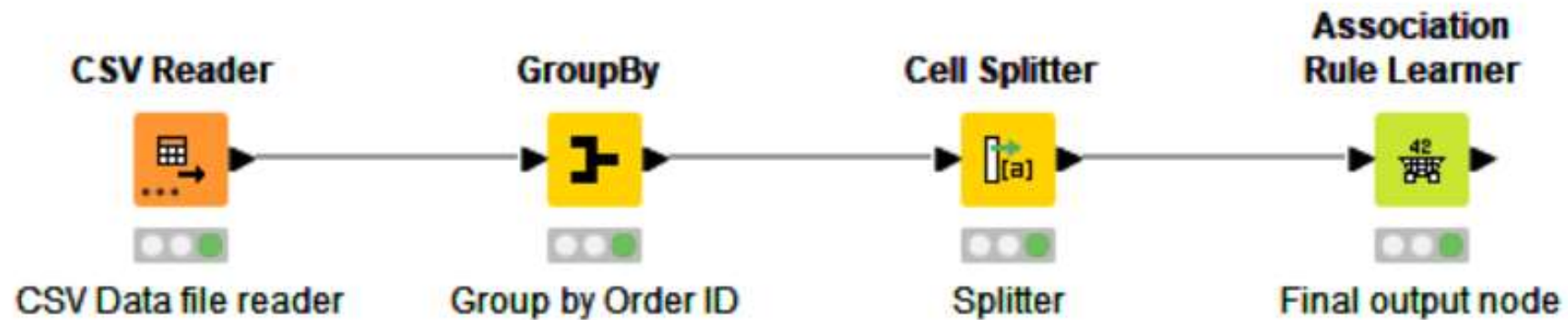
Recommendations:

- Focus on promoting and stocking up on poultry, soda, and cereals as they are consistently top-selling products.
- Consider increasing the stock of soap and toilet paper as they are the highest sold non-eatable products.
- Evaluate the reasons behind the low sales of hand soap and take measures to increase its sales.
- Schedule promotions and offers on Sundays to maximize sales on the day with the highest sales.
- Plan marketing campaigns and discounts during February to increase sales during the historically low-sales month.
- Plan marketing campaigns and discounts during January and March to increase sales during the historically high-sales months.
- Aim to replicate the sales patterns of Q1 2019 and Q3 2018.
- Keep the stock of products sold in Q2 consistent with the previous years to maintain sales levels.
- Keep in mind the limited data for 2020 while making sales and marketing decisions.

MARKET BASKET ANALYSIS:

- **DEFINITION:** Market Basket Analysis is a statistical technique that analyzes customer purchase patterns to identify associations between different products. It helps businesses understand which products are frequently purchased together and how customers' buying habits affect sales.
- **DATA:** To conduct market basket analysis, businesses need transactional data that includes details such as customer ID, product ID, and transaction date. This data is then used to create a matrix that represents the relationships between different products.
- **ASSOCIATION RULES:** Association rules are used to identify the strength of the relationship between different products. These rules are expressed in terms of support, confidence, and lift. Support refers to the frequency of co-occurrence of items in a transaction, while confidence measures the probability that if a customer buys one item, they will also buy another. Lift measures the degree of correlation between two items.
- **APPLICATIONS:** Market Basket Analysis is used in a variety of industries, including retail, e-commerce, and marketing. Retailers use this technique to optimize product placement and promotions. E-commerce companies use it to personalize product recommendations, and marketers use it to develop targeted advertising campaigns.
- **BENEFITS:** Market Basket Analysis helps businesses increase revenue by identifying cross-selling opportunities and developing targeted promotions. It also helps improve customer satisfaction by providing personalized recommendations and improving the overall shopping experience.

Market Basket Analysis KNIME Flow:



Input and output data:

Input to the KNIME Flow:

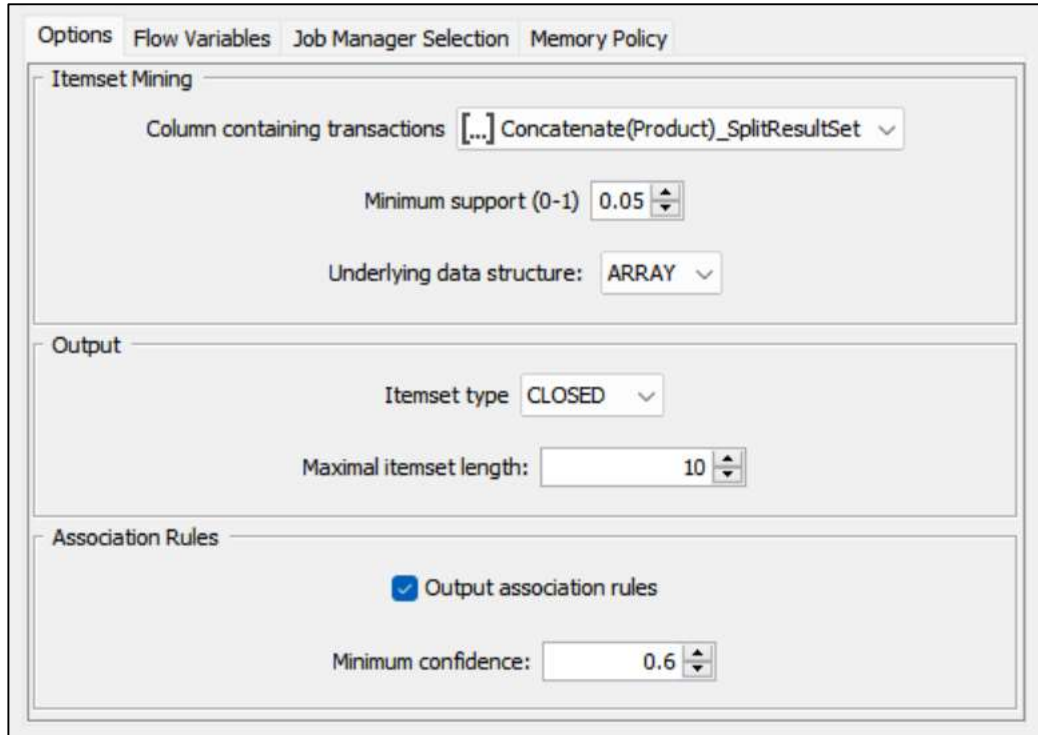
Table "default" - Rows: 20641 Spec - Columns: 3 Properties Flow Variables				
Row ID	[S] Date	[I] Order_id	[S] Product	
Row0	2018-01-01	1	yogurt	
Row1	2018-01-01	1	pork	
Row2	2018-01-01	1	sandwich bags	
Row3	2018-01-01	1	lunch meat	
Row4	2018-01-01	1	all- purpose	
Row5	2018-01-01	1	flour	
Row6	2018-01-01	1	soda	
Row7	2018-01-01	1	butter	
Row8	2018-01-01	1	beef	

Output of splitter to the Association rule learner:

Table "default" - Rows: 1139 Spec - Columns: 3 Properties Flow Variables				
Row ID	[I] Order_id	[S] Concatenate(Product)	[...] Concatenate(Product)_SplitResult...	
Row0	1	yogurt, pork, sandwich bags,...	[yogurt,pork,sandwich bags,...]	
Row1	2	toilet paper, shampoo, hand ...	[toilet paper,shampoo,hand soap,...]	
Row2	3	soda, pork, soap, ice cream, ...	[soda,pork,soap,...]	
Row3	4	cereals, juice, lunch meat, so...	[cereals,juice,lunch meat,...]	
Row4	5	sandwich loaves, pasta, tortil...	[sandwich loaves,pasta,tortillas,...]	
Row5	6	laundry detergent, toilet pap...	[laundry detergent,toilet paper,eggs,...]	
Row6	7	individual meals, paper towel...	[individual meals,paper towels,tortillas,...]	
Row7	8	ice cream, juice, paper towel...	[ice cream,juice,paper towels,...]	
Row8	9	juice, poultry, coffee/tea, co...	[juice,poultry,coffee/tea,...]	

Association rule parameters:

- Support of Minimum: 0.05
- Maximum Item Set Length : 10
- Minimum Confidence Level:0.6



The screenshot shows a software interface for configuring association rule parameters. It features four tabs at the top: 'Options', 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Options' tab is active, displaying three sections: 'Itemset Mining', 'Output', and 'Association Rules'. In the 'Itemset Mining' section, the 'Column containing transactions' is set to 'Concatenate(Product)_SplitResultSet', 'Minimum support (0-1)' is 0.05, and 'Underlying data structure' is 'ARRAY'. The 'Output' section shows 'Itemset type' as 'CLOSED' and 'Maximal itemset length' as 10. The 'Association Rules' section has 'Output association rules' checked and 'Minimum confidence' set to 0.6.

Options Flow Variables Job Manager Selection Memory Policy

Itemset Mining

Column containing transactions [...] Concatenate(Product)_SplitResultSet ▾

Minimum support (0-1) 0.05 ▴ ▾

Underlying data structure: ARRAY ▾

Output

Itemset type CLOSED ▾

Maximal itemset length: 10 ▴ ▾

Association Rules

☒ Output association rules

Minimum confidence: 0.6 ▴ ▾

Market Basket Analysis, Support, Confidence, and Lift values:

- In market basket analysis, support, confidence, and lift values are used to measure the strength of association between items in a transaction dataset.
- **Support:** It is the probability of observing the items together in a transaction. It is calculated as the number of transactions that contain both items divided by the total number of transactions. It measures how frequent the item set occurs in the dataset. High support indicates that the item set is popular and should be considered for promotion or placement together.
- **Confidence:** It is the conditional probability that a transaction containing one item also contains another item. It is calculated as the number of transactions containing both items divided by the number of transactions containing the first item. It measures the strength of the association between two items. High confidence indicates that the items are likely to be bought together, and can be used to recommend or suggest items to customers.
- **Lift:** It is the measure of how much more often two items occur together than expected if they were independent of each other. It is calculated as the support of the item set divided by the product of the individual supports of the items. A lift value of 1 indicates that the items are independent, while a value greater than 1 indicates a positive association between the items. A lift value less than 1 indicates a negative association between the items. High lift indicates that the items have a strong association and can be used for cross-selling or bundling.

Association Rules:

- Association rules are a technique used to find relationships or associations between items in a large dataset. These rules are based on the concept of frequent item sets, which are sets of items that appear together frequently in a transactional dataset.
- 24 rules have been found with the dataset and set parameters.

Row ID	D Support	D Confide...	D Lift	S Conseq...	S implies	[...] Items
rule0	0.05	0.64	1.7	juice	<---	[yogurt,toilet paper,aluminum foil]
rule1	0.05	0.62	1.645	juice	<---	[yogurt,poultry,aluminum foil]
rule2	0.05	0.613	1.616	coffee/tea	<---	[yogurt,cheeses,cereals]
rule3	0.05	0.6	1.424	poultry	<---	[dishwashing liquid/detergent,laundry detergent,mixes]
rule4	0.051	0.63	1.678	mixes	<---	[yogurt,poultry,aluminum foil]
rule5	0.051	0.611	1.66	sandwich bags	<---	[cheeses,bagels,cereals]
rule6	0.051	0.674	1.726	cheeses	<---	[bagels,cereals,sandwich bags]
rule7	0.051	0.617	1.558	cereals	<---	[cheeses,bagels,sandwich bags]
rule8	0.051	0.63	1.621	dinner rolls	<---	[spaghetti sauce,poultry,cereals]
rule9	0.051	0.637	1.512	poultry	<---	[dinner rolls,spaghetti sauce,cereals]
rule10	0.051	0.604	1.589	milk	<---	[poultry,laundry detergent,cereals]
rule11	0.052	0.628	1.61	eggs	<---	[dinner rolls,poultry,soda]
rule12	0.052	0.641	1.649	dinner rolls	<---	[spaghetti sauce,poultry,ice cream]
rule13	0.052	0.686	1.628	poultry	<---	[dinner rolls,spaghetti sauce,ice cream]
rule14	0.052	0.628	1.614	dinner rolls	<---	[spaghetti sauce,poultry,juice]
rule15	0.052	0.602	1.429	poultry	<---	[dinner rolls,spaghetti sauce,juice]
rule16	0.052	0.634	1.627	eggs	<---	[paper towels,dinner rolls,pasta]
rule17	0.052	0.602	1.621	pasta	<---	[paper towels,eggs,dinner rolls]
rule18	0.054	0.642	1.651	dinner rolls	<---	[spaghetti sauce,poultry,laundry detergent]
rule19	0.054	0.656	1.556	poultry	<---	[dinner rolls,spaghetti sauce,laundry detergent]
rule20	0.055	0.624	1.565	ice cream	<---	[paper towels,eggs,pasta]
rule21	0.055	0.63	1.616	eggs	<---	[paper towels,ice cream,pasta]
rule22	0.055	0.643	1.731	pasta	<---	[paper towels,eggs,ice cream]
rule23	0.055	0.649	1.791	paper towels	<---	[eggs,ice cream,pasta]

Recommendations by MBA:

- Create a combo deal where customers can purchase cereals, bagels, and sandwich bags together at a discounted price.
- Offer a "*Buy Two Get One Free*" promotion on yogurt, poultry, and aluminium foil to encourage customers to purchase more items at once.
- Offer a discount on mixes when purchased with yogurt, poultry, or aluminium foil.
- Provide a discount on dinner rolls when purchased with spaghetti sauce or poultry.
- Create a "Paper Products Bundle" offer that includes paper towels, toilet paper, and/or tissues at a discounted price.
- These discount offers and combos can help increase sales by providing customers with more value for their money and encouraging them to purchase more items. It is important to promote these offers through in-store signage, advertisements, and social media to ensure customers are aware of the deals available.

Summary:

- The analysis identified the products that are often purchased together by customers, which can help the store to optimize its product placement and promotions.
- Yogurt, poultry, aluminium foil, cheeses, cereals, and dinner rolls are some of the most frequently purchased products.
- Some of the product associations are unexpected, such as poultry with dishwashing liquid/detergent, laundry detergent, and mixes.
- The analysis suggests that offering discounts or combos, such as "buy two get one free," on certain products can encourage customers to purchase more.
- The store can also consider placing complementary products in close proximity to each other to increase the likelihood of customers making additional purchases.
- Overall, the market basket analysis can help the store to better understand customer behavior and preferences, and to make informed decisions about product placement and promotions.

THANK YOU