MRA Project - Milestone 2

Name: G. Umamakeshwari
PGPDSBA - April' B21

Agenda

Objective of this project is to analyse the association rules to suggest the best combo and offers for the Grocery Store chain using Market Basket Analysis.

Project to analyse Point of Sale (POS) Data for providing recommendations to increase the revenue of Grocery Store

- ## Executive Summary of the data

The data has 1 date time, 1 int, and 1 Object data types variables. There is no missing values present in the data set. Here 20641 records available without any missing values with 3 variables. Here 4730 duplicate records founded. After duplicate removal we have found 15911 records.

This dataset related to Grocery Store, and they have different products like yogurt, pork, sandwich bags, lunch meat, etc.

The data maintained each transactions entry as order Id and for each order number contained product name with order date.

We noticed that one order number has many different entries with different products.

Contents of the presentation

- Problem Statement.
- About Data
  - Sample data
  - Shape of data
  - Info of data
  - Data Summary
- Exploratory Analysis and Inferences
  - Univariate Analysis – Numerical variable
  - Univariate Analysis – Categorical Variable
  - Bivariate Analysis
  - Time series & Trends in Sales
- Use of Market Basket Analysis (Association Rules)
  - Tool used
  - Write Something about the association rules and their relevance in this case
  - Workflow image of KNIME
  - Write about threshold values of Support and Confidence
- Associations Identified
  - The associations in a tabular manner
    - Output table for association rules
  - Explain about support, confidence, & lift values that are calculated
    - Output table based on lift value
- A suggestion of Possible Combos with Lucrative Offers
  - Recommendations
  - Make discount offers or combos (or buy two get one free) based on the associations and your experience

- # 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?

Data Dictionary:	
Date	Order date
Order_id	Order ID
Product	Order item

- **About Data** (Info, Shape, Summary Stats, your assumptions about data)

Sample data

	Date	Order_id	Product
0	2018-01-01	1	yogurt
1	2018-01-01	1	pork
2	2018-01-01	1	sandwich bags
3	2018-01-01	1	lunch meat
4	2018-01-01	1	all- purpose

Shape of data

- There are 3 variables available regarding the orders of the product with 20641 records.

Info of data

- The data has 1 date time, 1 int, and 1 Object data types variables. There is no missing values present in the data set.

```
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)
```

Data Summary

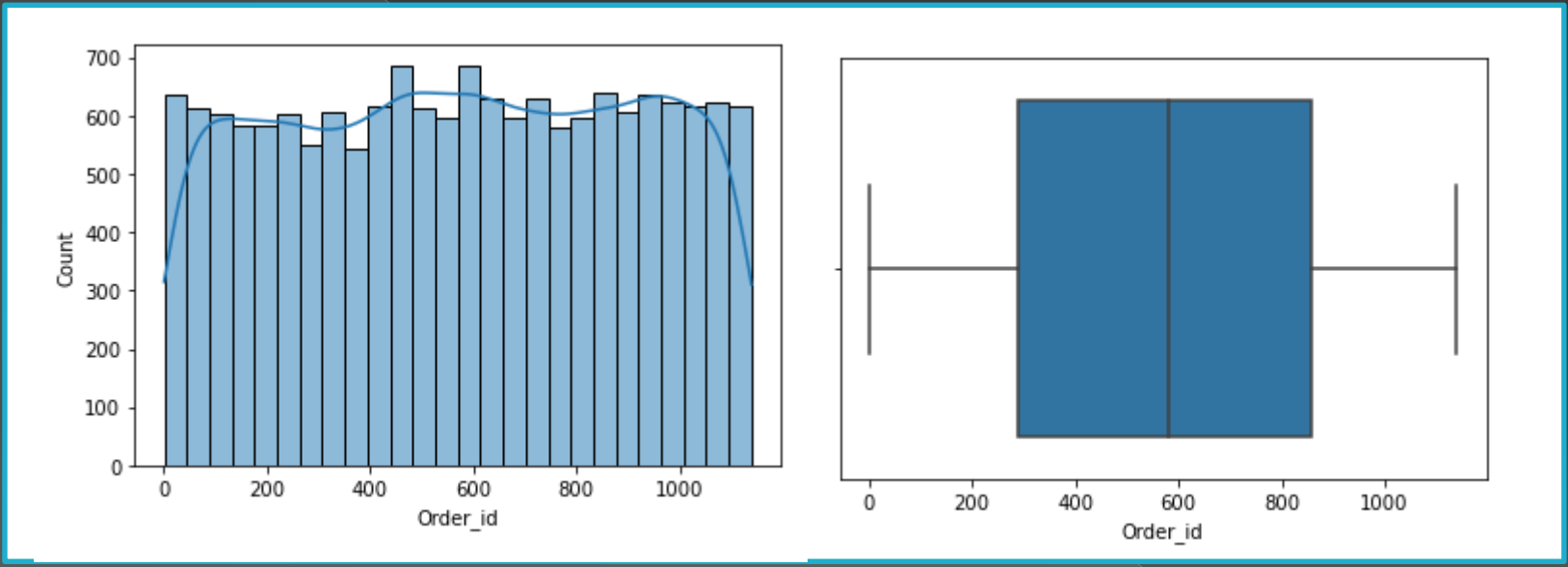
	Date	Order_id	Product
count	20641	20641.000000	20641
unique	603	NaN	37
top	2019-02-08	NaN	poultry
freq	183	NaN	640
mean	NaN	575.986289	NaN
std	NaN	328.557078	NaN
min	NaN	1.000000	NaN
25%	NaN	292.000000	NaN
50%	NaN	581.000000	NaN
75%	NaN	862.000000	NaN
max	NaN	1139.000000	NaN

	Date	Order_id	Product
count	15911	15911.000000	15911
unique	603	NaN	37
top	2019-02-08	NaN	poultry
freq	138	NaN	480
mean	NaN	574.150462	NaN
std	NaN	328.537425	NaN
min	NaN	1.000000	NaN
25%	NaN	289.500000	NaN
50%	NaN	579.000000	NaN
75%	NaN	859.000000	NaN
max	NaN	1139.000000	NaN

- Here Number of duplicate rows are 4730.
- The data has 1 date time, 1 int, and 1 Object data types variables. There is no missing values present in the data set. Here 20641 records available without any missing values with 3 variables. Here 4730 duplicate records founded. After duplicate removal we have found 15911 records.
- This dataset related to Grocery Store, and they have different products like yogurt, pork, sandwich bags, lunch meat, etc.
- The data maintained each transactions entry as order Id and for each order number contained product name with order date.
- We noticed that one order number has many different entries with different products

Exploratory Analysis and Inferences

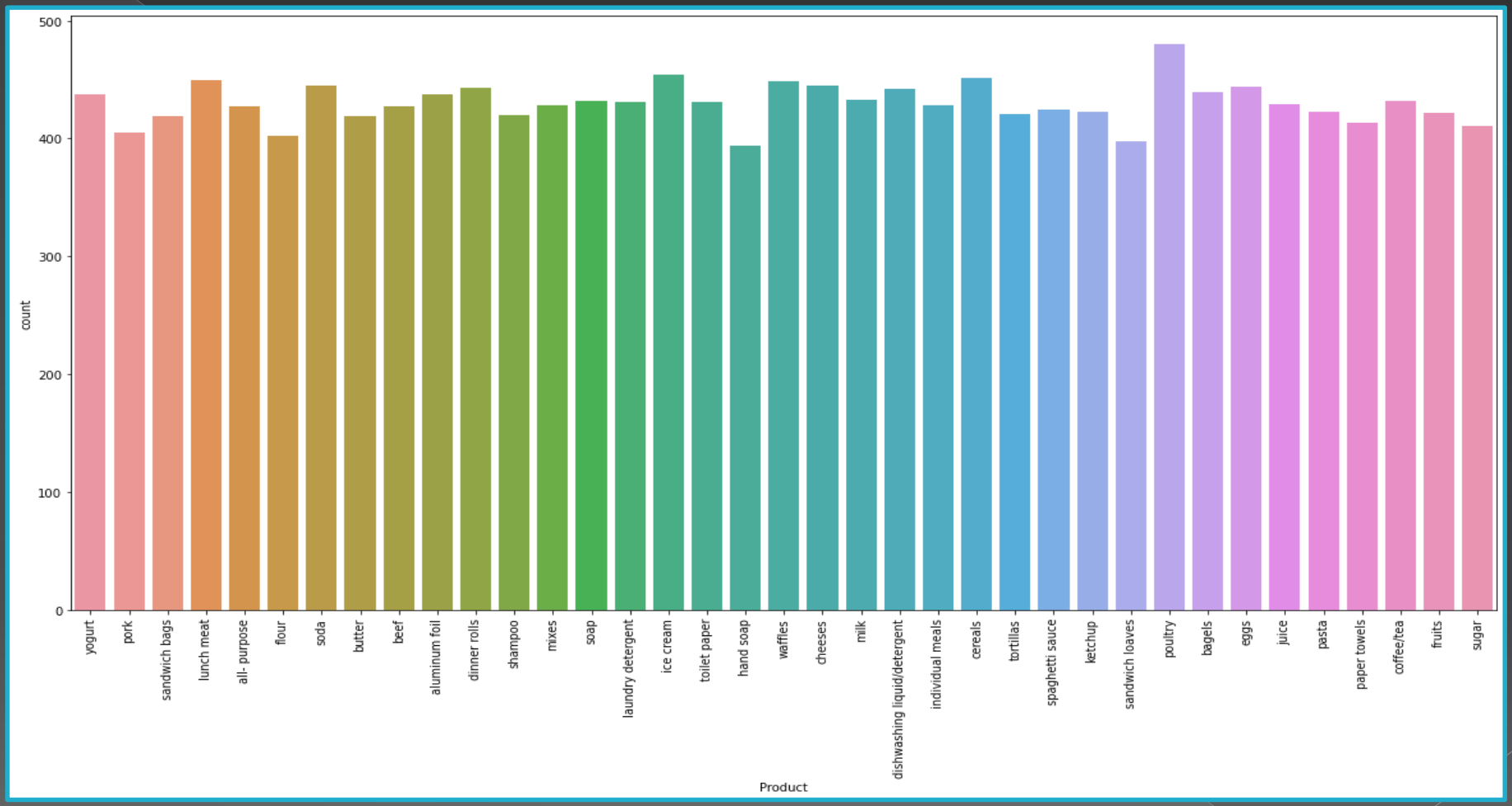
Univariate Analysis – Numerical variable



Inference:

- There is no outliers in order id. Order id have different data pattern.

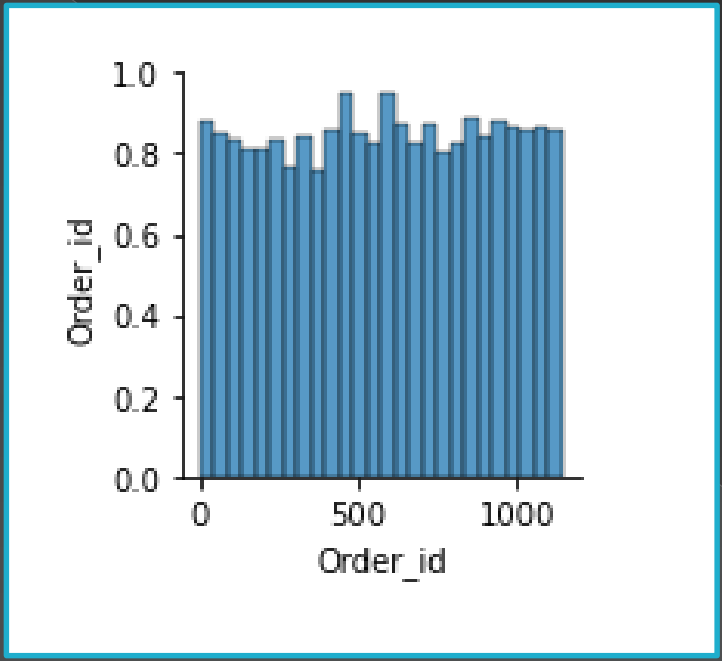
Univariate Analysis – Categorical Variable



Inference:

- As per the plot 'Poultry' and 'Ice cream' are available in highest number of orders. More or less all products are available between 390 to 480 orders in total of 1139 orders.

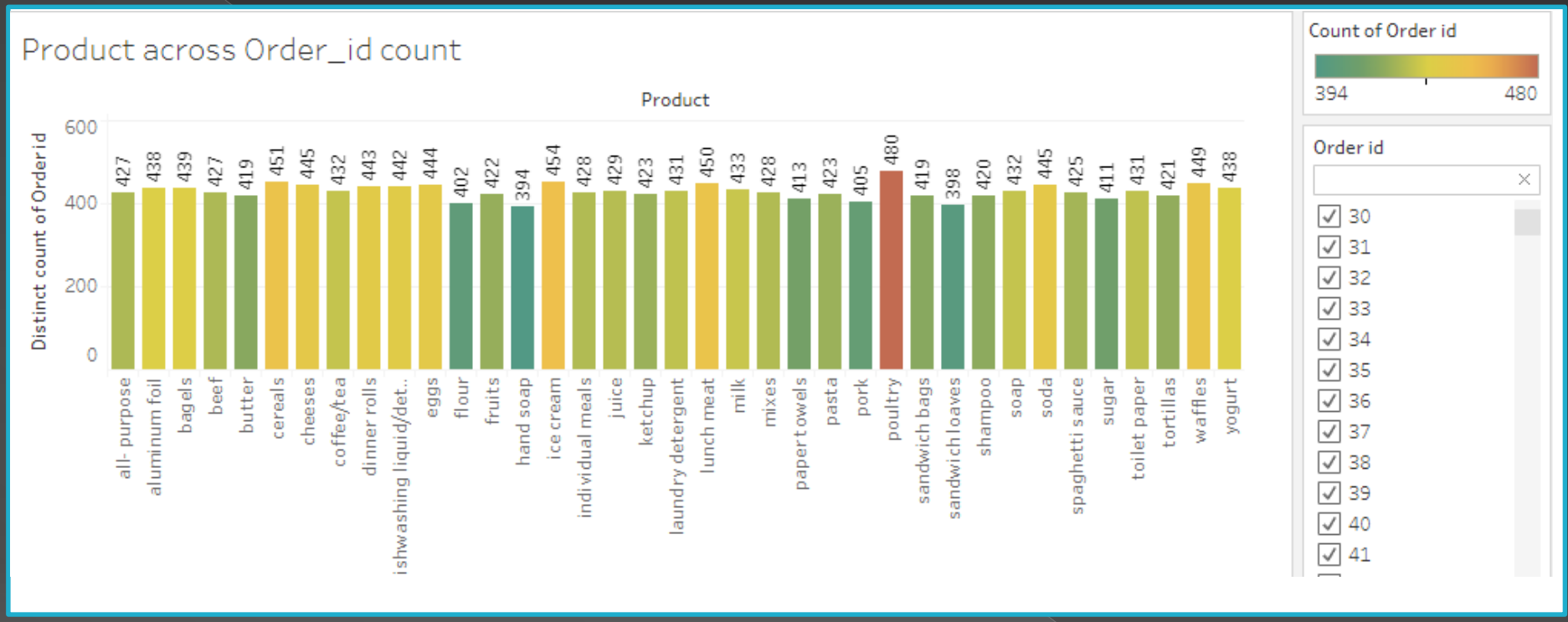
Bivariate Analysis



Inference:

- As per the plot order id only numerical variable. So, no need to do bivariate and multivariate analysis for this dataset.

Time series & Trends in Sales

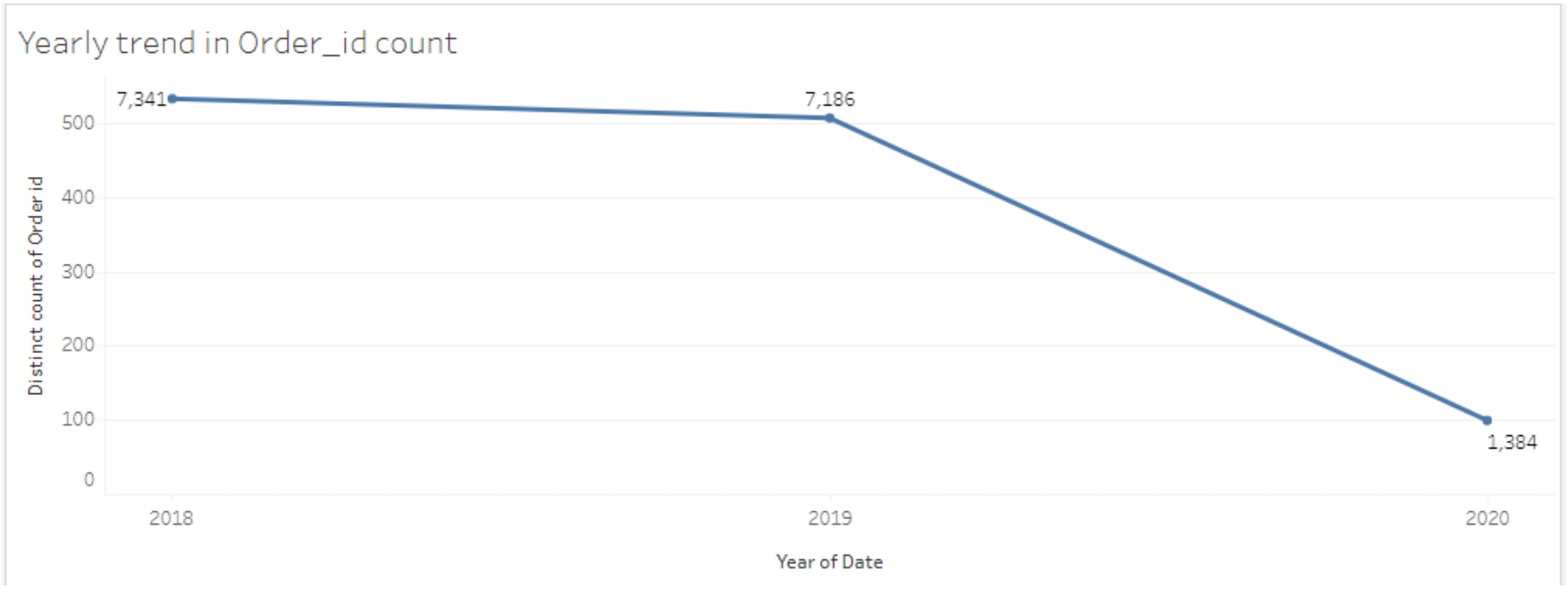


Inference:

- As per the plot ‘Poultry’ and ‘Ice cream’ are available in highest number of orders. More or less all products are available between 390 to 480 orders in total of 1139 orders.
- Here ‘hand soap’, ‘sandwich’ and ‘flour’ are low in order count compared with others.

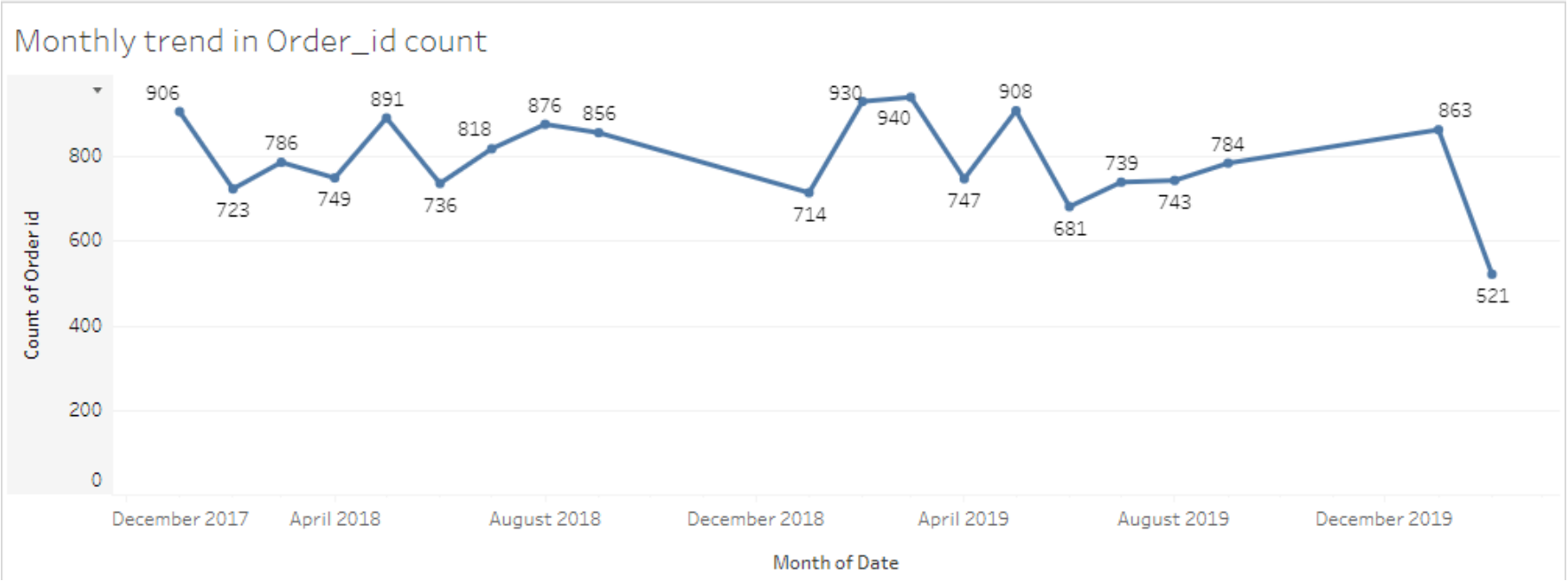
Time series & Trends in Sales

- Daily, Monthly, Quarterly, Yearly Trends in Sales



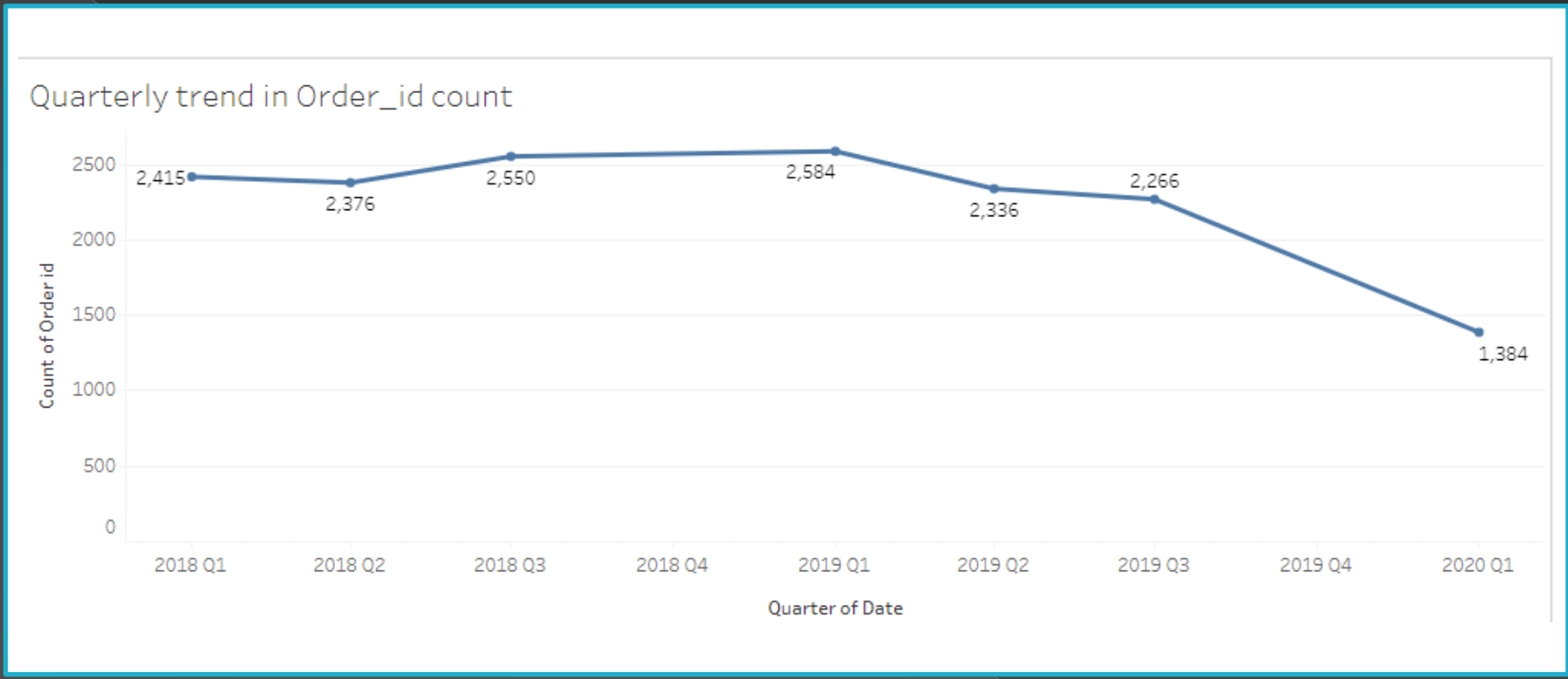
Inference:

- As per Yearly orders, we only have 3 years data. Here 2018 to 2019 sales has no trend but 2019 to 2020 orders has decreased.



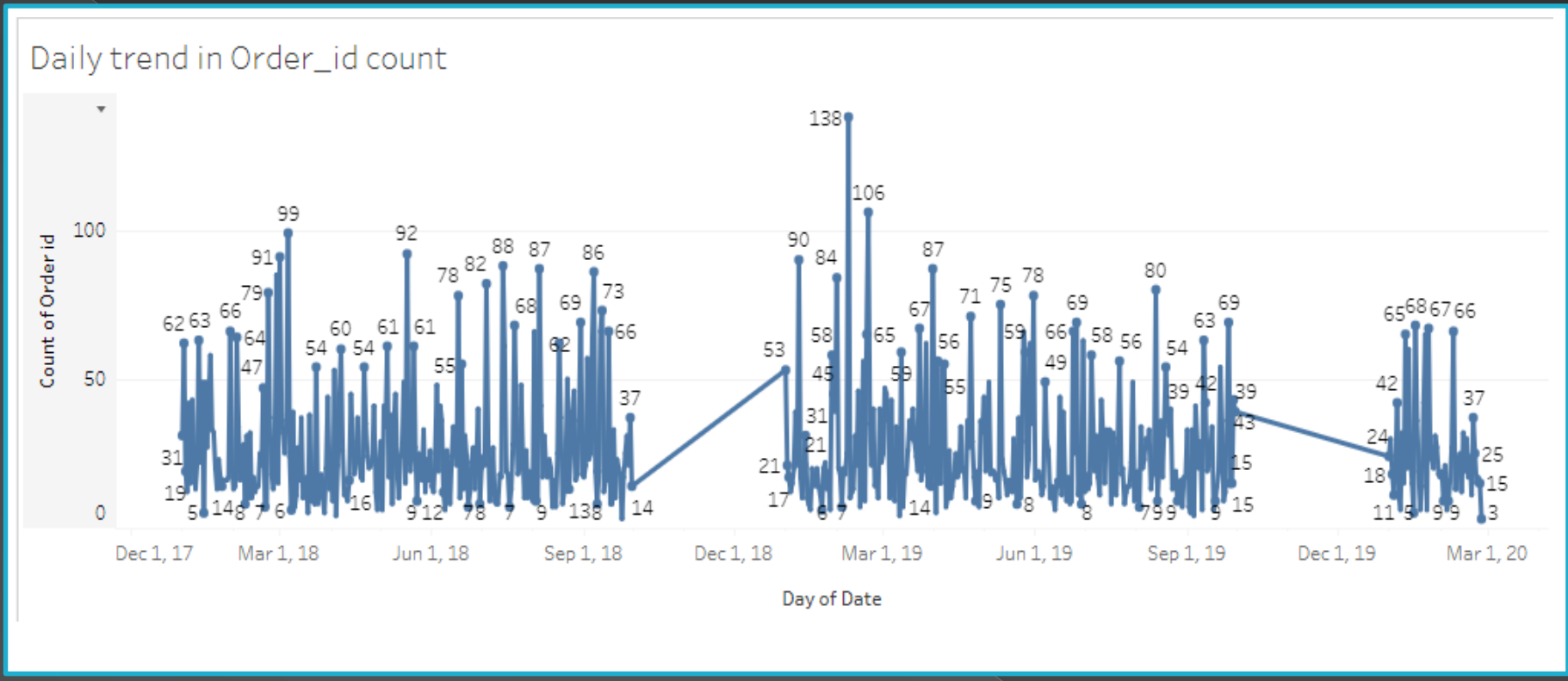
Inference:

- As per monthly orders, we are not found any trend and seasonality. However compared with starting month, ending month have nearly 45% low orders.



Inference:

- As per quarterly orders, we only have 9 quarter data. Here 2018 to 2019 sales has no trend but 2019 to 2020 orders has decreased.



Inference:

- As per daily orders, we only have 3 years data. Here 2018 to 2019 sales has no trend but 2019 to 2020 orders has decreased. Additionally we can see that from Q4 2018 to Q1 2019 orders has gradually increased on the other hand from Q4 2019 to Q1 2020 orders has gradually decreased compared with other quarters.

Use of Market Basket Analysis (Association Rules)

- **Tool used:**

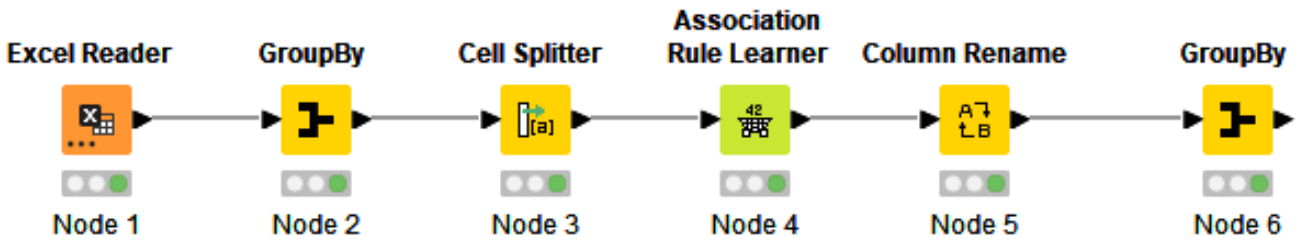
  I have used Python and Tableau for EDA and KNIME Tool for Market Basket Analysis.

- **Write Something about the association rules and their relevance in this case**

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. Association rules are working based on "if-then" statements, that help to show the probability of relationships between data items, within large data sets in various types of databases. The Association rule is very useful in analysing the retail basket or transaction data.

Here, The Point of Sale(POS) data is given by grocery store. The database consists of a large number of transaction records for individual order id with product. Based on the Association Rules, we can find the Support, Confidence and Lift values to find out the best combination of sales.

- **Workflow image of KNIME**



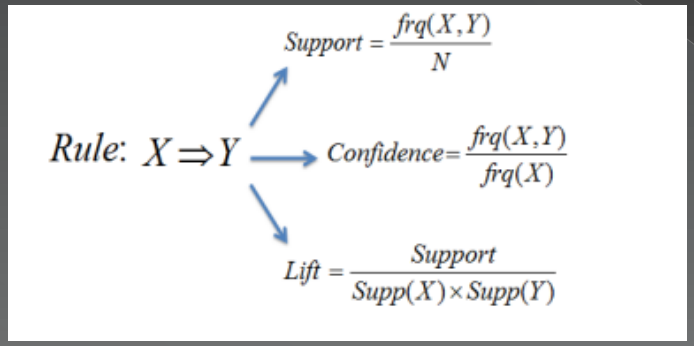
Write about threshold values of Support and Confidence

Threshold values of Support and Confidence is states that the value of the proposed support has the ability to generate a rule when viewed from the quality as support produces at a lift ratio value of > 1.

Support: This says how popular an product set is, as measured by the proportion of items x and y by total no.of items in the dataset appears.

Confidence: This is calculated using probability of customer who bought both x and y with probability of y.

Lift : Increase in the sale of x when you sell y.



Associations Identified

- **The associations in a tabular manner**

We can observe here that there is 1,139 unique order ids from out dataset with 15911, All-Purpose have no value here. So after filtering the data with out All-Purpose have 15484.

This will help us to classify the products for our further Market Basket Analysis.

Here we can set up minimum threshold value for Support and Confidence. So that form that value to maximum of 1 for support and confidence will be calculated as per association rules for each product.

Based on the support and confidence value, we can calculate lift value of recommended combo items for each product. Finally we will filter the combo products for each product by using maximum value of lift value.

- **Output table for association rules**

File Edit Hilite Navigation View

Table "default" - Rows: 787 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	D Support	D Confide...	D Lift	S Rec_Item	S implies	[...] Items
rule0	0.15	0.408	1.148	pork	<---	[sandwich ba...
rule1	0.15	0.422	1.148	sandwich bags	<---	[pork]
rule2	0.15	0.4	1.126	pork	<---	[beef]
rule3	0.15	0.422	1.126	beef	<---	[pork]
rule4	0.15	0.422	1.081	cheeses	<---	[pork]
rule5	0.15	0.422	1.066	cereals	<---	[pork]
rule6	0.15	0.414	1.126	sandwich bags	<---	[paper towels]
rule7	0.15	0.408	1.126	paper towels	<---	[sandwich ba...
rule8	0.15	0.425	1.132	mixes	<---	[flour]
rule9	0.15	0.407	1.042	soda	<---	[shampoo]
rule10	0.15	0.405	1.037	soda	<---	[fruits]
rule11	0.15	0.408	1.031	cereals	<---	[butter]
rule12	0.15	0.402	1.094	butter	<---	[spaghetti sa...
rule13	0.15	0.408	1.094	spaghetti sa...	<---	[butter]
rule14	0.15	0.43	1.117	aluminum foil	<---	[sandwich loa...
rule15	0.15	0.407	1.021	ice cream	<---	[shampoo]
rule16	0.15	0.43	1.165	shampoo	<---	[sandwich loa...
rule17	0.15	0.407	1.165	sandwich lo...	<---	[shampoo]
rule18	0.15	0.43	1.151	spaghetti sa...	<---	[sandwich loa...
rule19	0.15	0.402	1.151	sandwich lo...	<---	[spaghetti sa...
rule20	0.15	0.405	1.118	paper towels	<---	[fruits]
rule21	0.15	0.414	1.118	fruits	<---	[paper towels]
rule22	0.151	0.425	1.089	eggs	<---	[pork]
rule23	0.151	0.411	1.067	aluminum foil	<---	[sandwich ba...
rule24	0.151	0.428	1.128	soap	<---	[flour]
rule25	0.151	0.41	1.08	soap	<---	[shampoo]
rule26	0.151	0.407	1.103	shampoo	<---	[ketchup]
rule27	0.151	0.41	1.103	ketchup	<---	[shampoo]
rule28	0.151	0.41	1.051	eggs	<---	[shampoo]
rule29	0.151	0.407	1.103	shampoo	<---	[pasta]
rule30	0.151	0.41	1.103	pasta	<---	[shampoo]

- # Explain about support, confidence, & lift values that are calculated

▪ We can see that 787 rows finalized as per association rules. In that for same product contains a different rules with different recommended product for combo sales.

▪ It has created multiple rules on the basis of threshold limit that we have set for the Association Rule Learner Node.

Here we have set,

Threshold value for Support as 0.15

Threshold value for Confidence 0.4

▪ Consequent column contains recommended products and we can filter the highest lift values for each product for the better recommendations.

For instance, Assume there are 50 customers. 10 of them bought milk, 8 bought bread and 6 bought both of them.

- Support = P(Milk & Butter) = 6/100 = 0.06
- Confidence = support/P(Butter) = 0.06/0.08 = 0.75
- Lift = Confidence/P(Milk) = 0.75/0.10 = 7

Output table based on lift value

Table "default" - Rows: 36 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	[...] Items	D Lift	D Support	D Confide...	S Rec_Item
Row0	[aluminum foil]	1.199	0.163	0.425	flour
Row1	[bagels]	1.174	0.162	0.42	sandwich bags
Row2	[beef]	1.195	0.159	0.423	pork
Row3	[butter]	1.135	0.155	0.422	cereals
Row4	[cereals]	1.168	0.167	0.422	shampoo
Row5	[cheeses]	1.179	0.166	0.424	milk
Row6	[coffee/tea]	1.166	0.159	0.42	flour
Row7	[dinner rolls]	1.197	0.166	0.427	soap
Row8	[dishwashin...	1.234	0.166	0.428	ketchup
Row9	[eggs]	1.228	0.167	0.428	aluminum foil
Row10	[flour]	1.159	0.155	0.438	mixes
Row11	[fruits]	1.195	0.16	0.432	soda
Row12	[hand soap]	1.169	0.153	0.443	mixes
Row13	[ice cream]	1.178	0.168	0.42	toilet paper
Row14	[individual m...	1.217	0.159	0.423	shampoo
Row15	[juice]	1.226	0.162	0.431	dishwashing...
Row16	[ketchup]	1.202	0.158	0.426	shampoo
Row17	[laundry det...	1.173	0.16	0.422	individual m...
Row18	[lunch meat]	1.195	0.166	0.421	eggs
Row19	[milk]	1.172	0.163	0.429	juice
Row20	[mixes]	1.234	0.162	0.43	hand soap
Row21	[paper towels]	1.219	0.158	0.435	sandwich bags
Row22	[pasta]	1.219	0.159	0.429	shampoo
Row23	[pork]	1.167	0.153	0.431	sandwich bags
Row24	[poultry]	1.189	0.178	0.421	beef
Row25	[sandwich b...	1.202	0.157	0.428	pork
Row26	[sandwich lo...	1.217	0.155	0.444	aluminum foil
Row27	[shampoo]	1.226	0.156	0.424	soda
Row28	[soap]	1.192	0.162	0.428	laundry det...
Row29	[soda]	1.228	0.166	0.424	cereals
Row30	[spaghetti s...	1.224	0.161	0.431	butter
Row31	[sugar]	1.178	0.157	0.436	bagels
Row32	[toilet paper]	1.207	0.161	0.424	spaghetti sa...
Row33	[tortillas]	1.182	0.158	0.427	milk
Row34	[waffles]	1.184	0.168	0.427	sandwich bags
Row35	[yogurt]	1.218	0.165	0.428	sandwich bags

As per the highest lift value of each product we can get 36 rows as final output for recommendations.

A suggestion of Possible Combos with Lucrative Offers

- **Recommendations**
  - If we see the result table of the Association Rule Learner best combo product is finalized based on lift value in single bracket order.
  - So generally we recommend the products for combo offers that are listed in consequent feature which has a higher lift value. That means it has the higher probability of being purchased as combo by maximum customer and it will result highest purchase by more customer if we give combo offers.
  - Here we can recommend to arrange products in near by racks with the highest selling and the lowest selling products with combo offers. So, that the lowest selling product may seen by customer and it make them purchase.
  - Also, we can recommend to buy two get one or buy one get one offer for same product of lower selling product or with highest selling product. So it will increase the sales.
  - We can recommend that some discounts based on the purchased value. If may be cash discount in total value or coupon code for next purchase with valid date or surprise free product offer with some lower selling product based on purchased value.

- **Make discount offers or combos (or buy two get one free) based on the associations and your experience**

  - As per the result, **Sandwich bags** is associated with **bagels, paper towels, pork, waffles and yogurt**. So we can give a buy 2 or 3 same or different items in this list to get a free **Sandwich bags.**

  - Same for the items **cereals, individual meals, ketchup or pasta** to get free **shampoo**

  - Combo offer with some discount for aluminium foil with flour, coffee/tea with flour, beef with pork, butter and soda with cereals, tortillas and cheeses with milk, flour and hand soap with mixes and so on.

  - Buy 1 get 1 or buy 2 get 1 offer for milk with juice, lunch meat with eggs, spaghetti sauce with butter, poultry with beef. This have low expire days because of daily consumable food product. So, this offer results us a high sales as well as the prevent form loss of product.