MARKET BASKET ANALYSIS

Great Learning
POWER AHEAD

By,
Abhishek K Hiremath

Agenda

1. **Executive Summary**
   - Problem Statement
2. **About data**
   - Data Characteristics
   - Exploratory data analysis
   - Summary & Insights
3. **About Market Basket**
   - What is Market Basket?
   - Approach
   - KNIME Workflow
4. **Associations**
   - Association Rule
   - Market Basket Recommendations
5. **Overall Recommendations**

Executive Summary:

- **Data:** from 01-01-2018 to 26-02-2020
- **Objective:** project involves conducting a thorough analysis of Point of Sale (POS) Data for providing recommendations through which a grocery store can increase its revenue by popular combo offers & discounts for customers.
- **Dataset:** We have received the 2 years and 2 months data of a Grocery store. Consisting 20641 entries with 3 variable details regarding the demography of the transaction and item information.
- **Missing values :** None
- **Duplicate values:** 4730
- The exploratory analysis and insights provide a clear understanding of the data and highlight the key trends and patterns in sales.
- **Market Basket Analysis** using association rules was performed to identify the relationships between the products purchased by the customers.
- This analysis helped to identify the products that are frequently purchased together, which can be used to create lucrative offers for the customers.
- **Yearly Comparisons** (2018 & 2019):
    - In **2019,** a **decrease of 26** in the total number of **orders** compared to **2018**.
    - A **decrease of 155 units** in the total number of **products sold**.

Problem Statement

The Company's Data Challenge

A grocery store has entrusted us with their transactional data, seeking a solution to enhance their revenue-generation strategies. They are grappling with optimizing their customer offerings and need data-driven insights to overcome this hurdle.

Objective:

Our objective is to :

- **Analyse data**: Analyze POS data to identify common item combinations in customer orders.
- **Recommend**: Develop data-driven strategies for **popular combo offers and discounts**.
- **Increase Revenue**: Use insights to boost the grocery store's revenue through tailored customer incentives.

ABOUT DATA

About Data : Data Characteristics

Sample of dataset

	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
...	...	...	...
20636	2020-02-25	1138	soda
20637	2020-02-25	1138	paper towels
20638	2020-02-26	1139	soda
20639	2020-02-26	1139	laundry detergent
20640	2020-02-26	1139	shampoo

20641 rows × 3 columns

```
RangeIndex: 20641 entries, 0 to 20640
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       20641 non-null  datetime64[ns]
 1   Order_id   20641 non-null  int64
 2   Product    20641 non-null  object
dtypes: datetime64[ns](1), int64(1), object(1)
```

Column	Dictionary
Date	Date of Order/Transaction
Order_id	Order ID
Product	Product purchased

- **Shape of the data**: The dataset contains **20641** rows and 3 columns.
- **Data types** : We have the columns with data type as datetime64(1), int64(1), object(1)
- Key column includes the Product & Order id

Note: The above data is before Exploratory data analysis & data cleaning

About Data : Data Cleaning

Duplicate Values

```
Total duplicate values: 4730
```

Duplicate Value check

	Date	Order_id	Product
4	2018-01-01	1	all- purpose
10	2018-01-01	1	all- purpose
11	2018-01-01	1	dinner rolls
13	2018-01-01	1	all- purpose
18	2018-01-01	1	dinner rolls
...	...	...	...
20632	2020-02-25	1138	sandwich bags
20633	2020-02-25	1138	toilet paper
20634	2020-02-25	1138	soda
20635	2020-02-25	1138	soda
20636	2020-02-25	1138	soda

8613 rows × 3 columns

Missing values

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

Insights

- **No Missing** values in the data
- **4730 duplicates** found, that were removed.
- Total Unique rows: 15911
- No other Irregularity found

Duplicate Values:

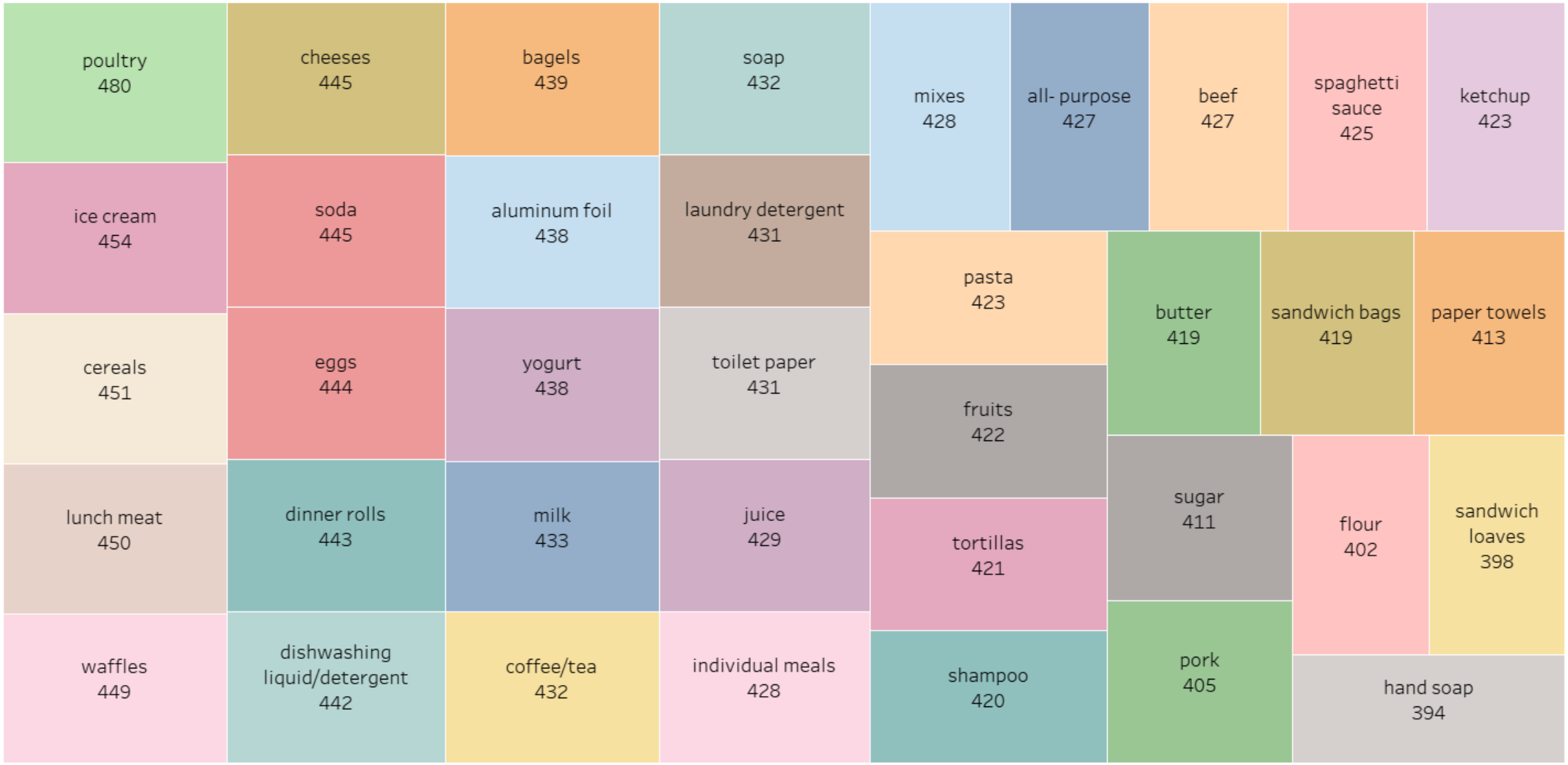
- It is generally a good practice to drop duplicate rows in a dataset as they do not provide any additional information and can skew the results of any analysis performed on the dataset.
- However, in this particular case, dropping duplicate rows may not be appropriate as there is no unique identifier 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.

Assumptions:

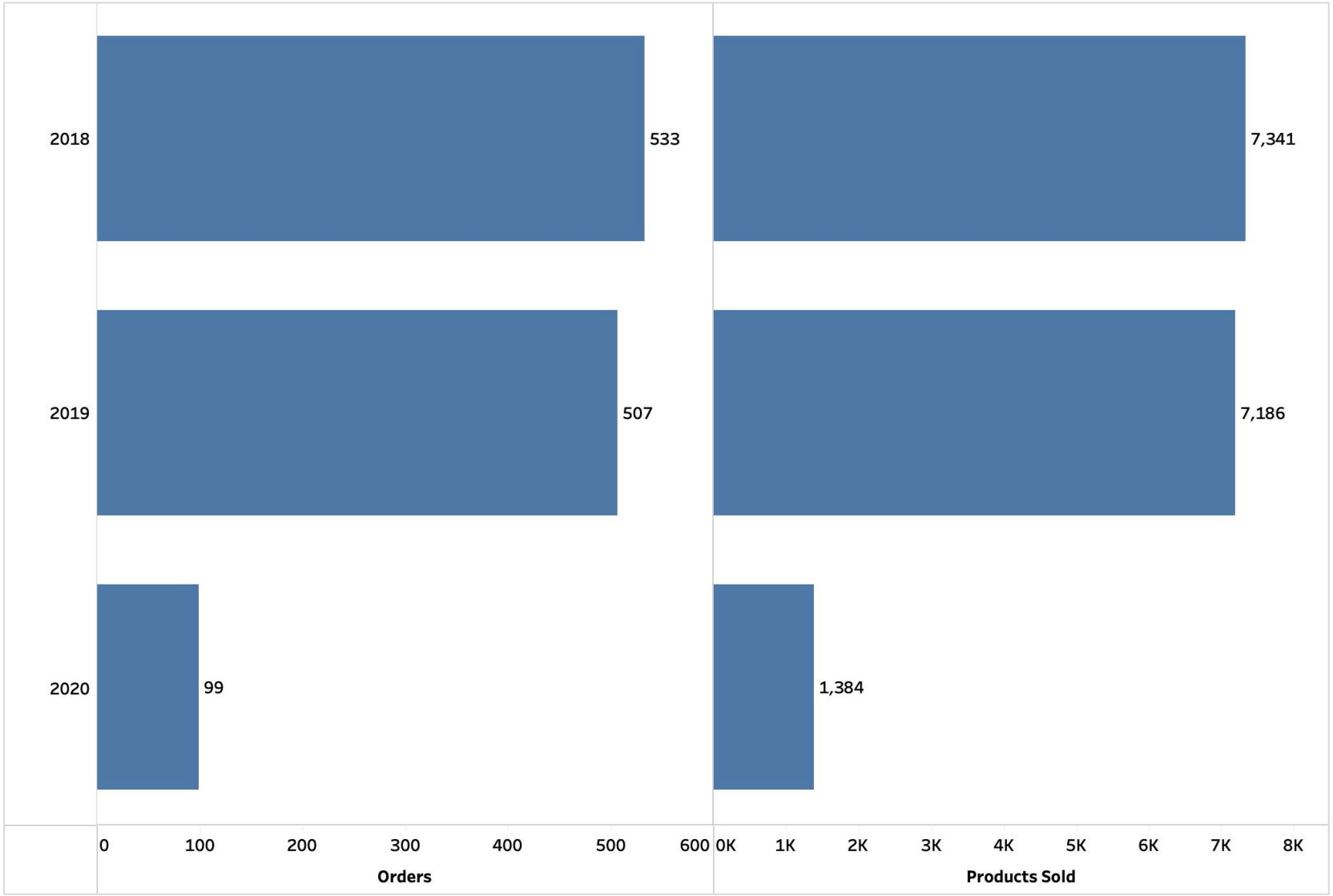
- The data represents a list of items purchased at a grocery store on various dates.
- Each entry in the data represents a single item purchased.
- The first column in the data represents the date the item was purchased.
- The second column represents the customer who made the purchase.
- The third column 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

Frequency of Products Sold



Yearly Orders & Products Sold

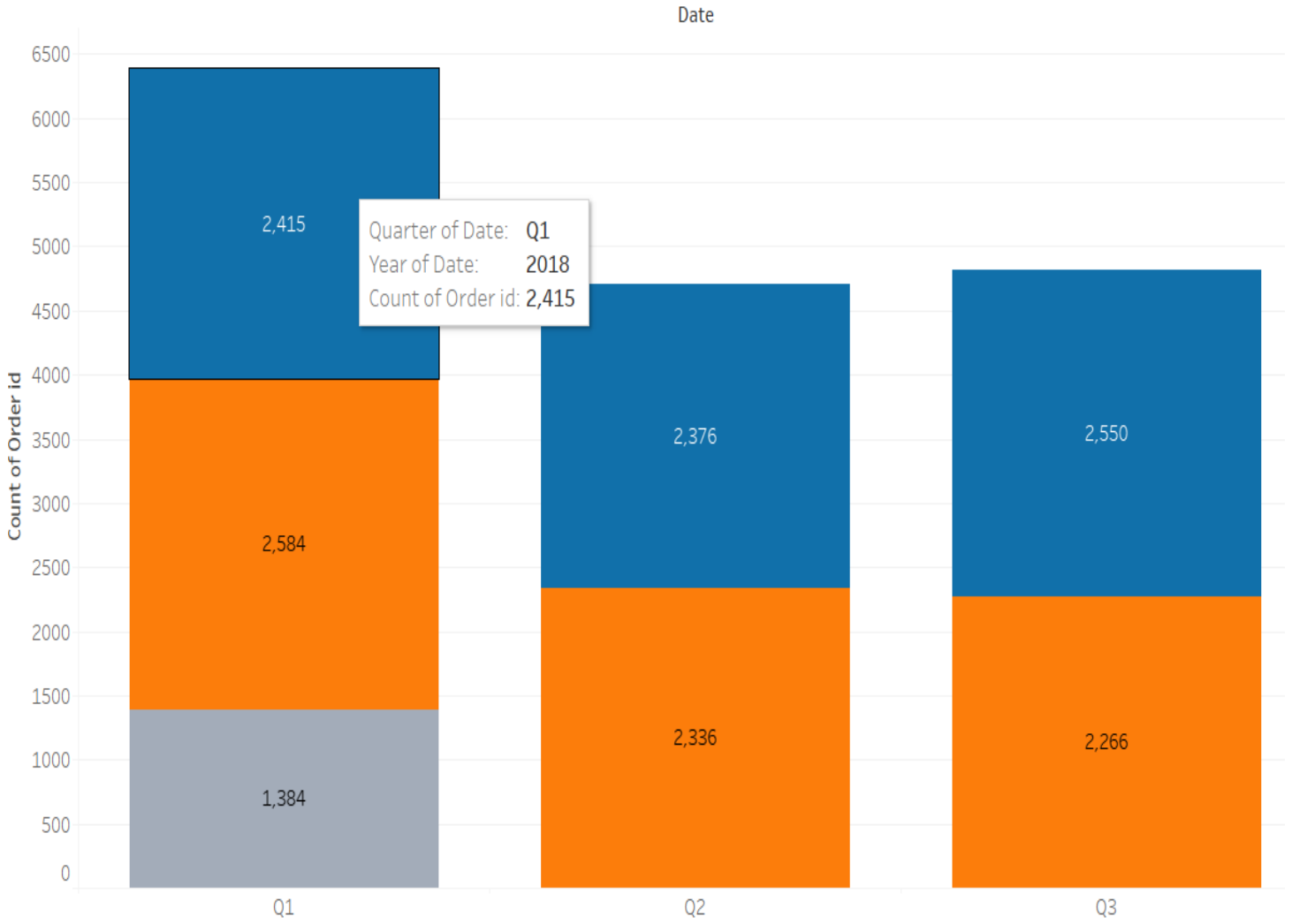


Inference:

- **Decrease** in the **Total orders and count of products sold** in **2019 compared to 2018**
- The **Orders decreased** to **507** in **2019** as compared to **533 in 2018**
- Similarly the **Product sold** count also **reduced to 7186 in 2019 from 7341 in 2018.**

*2020 has not been considered as it has only 2 months of data

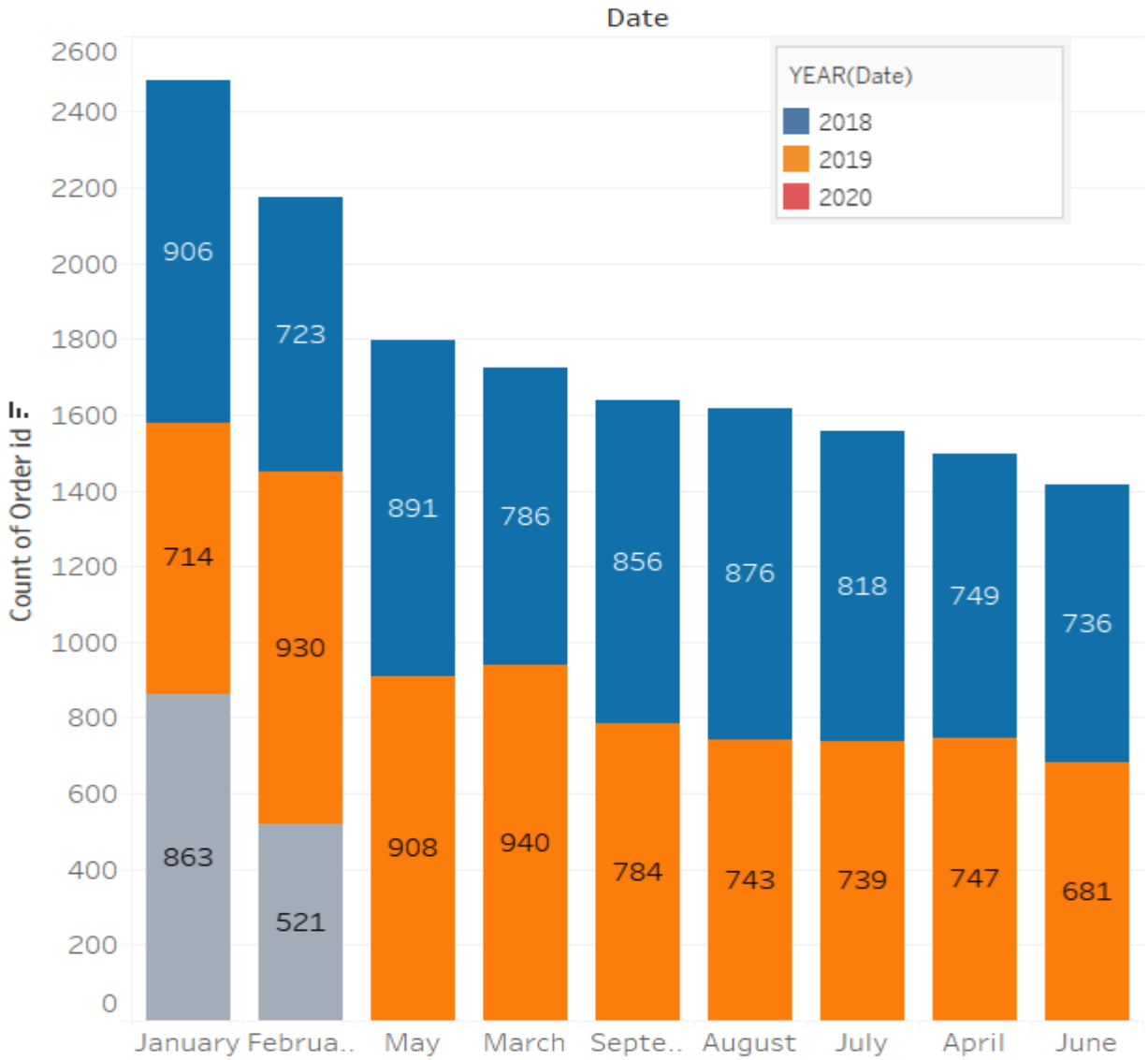
Quarterly Count of Products Sold



Inference:

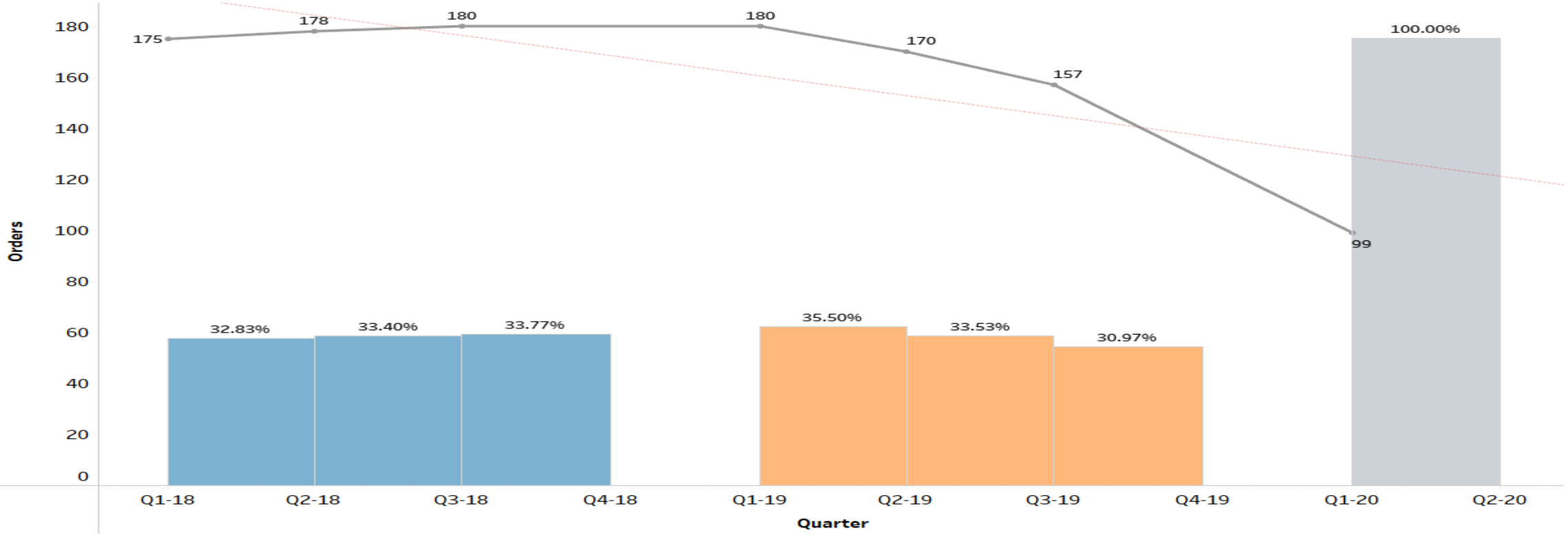
- As we have data till 26 Feb. 2020 that's why the count of products sold in Q1 is Hight.
- In 2019 Q1 sales was highest
- In 2018 Q3 sales was highest
- Count of product sold in Q2 is approx. same in 2019 and 2018.

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.

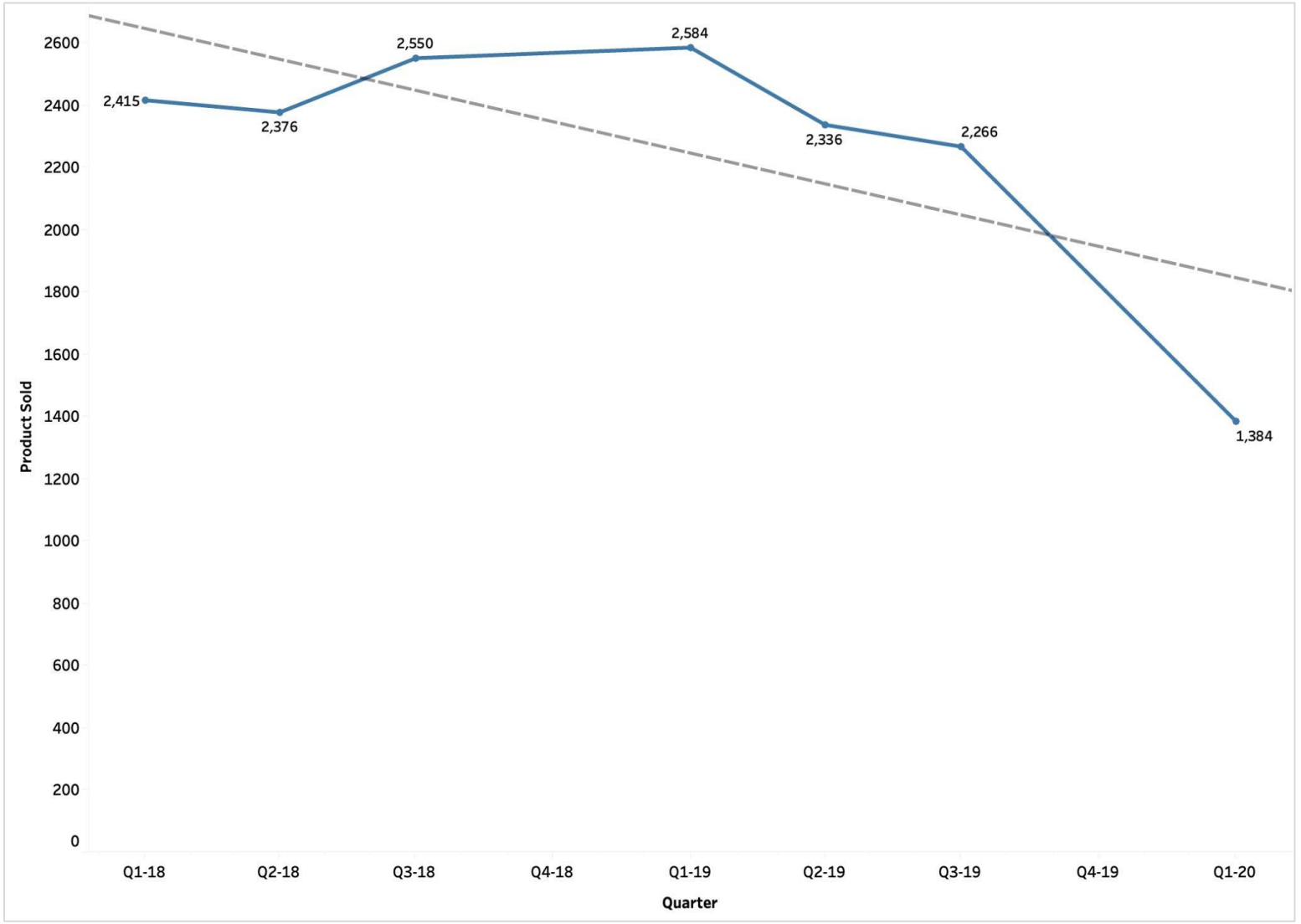
Orders Over Time: Quarterly Orders Placed

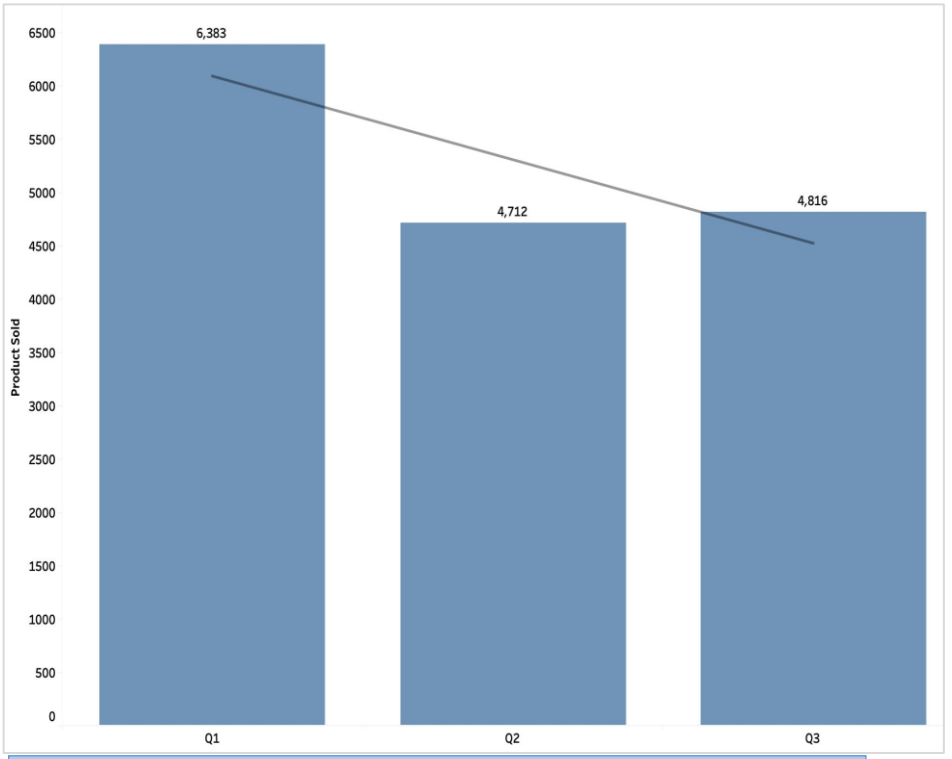


Inference:

- The Order Trend shows **decrease in Sales** over time
- In **2018**, the **quarterly trend** was **slightly increasing,** however, in **2019, the sales decreased from Q1 to Q3**
- **No data found** for **quarter 4** in both **2018 & 2019**. This is critical & business need to check immediately for reason
- **For 2020, only Jan & Feb** month data available

Orders Over Time: Quarterly Products Sold



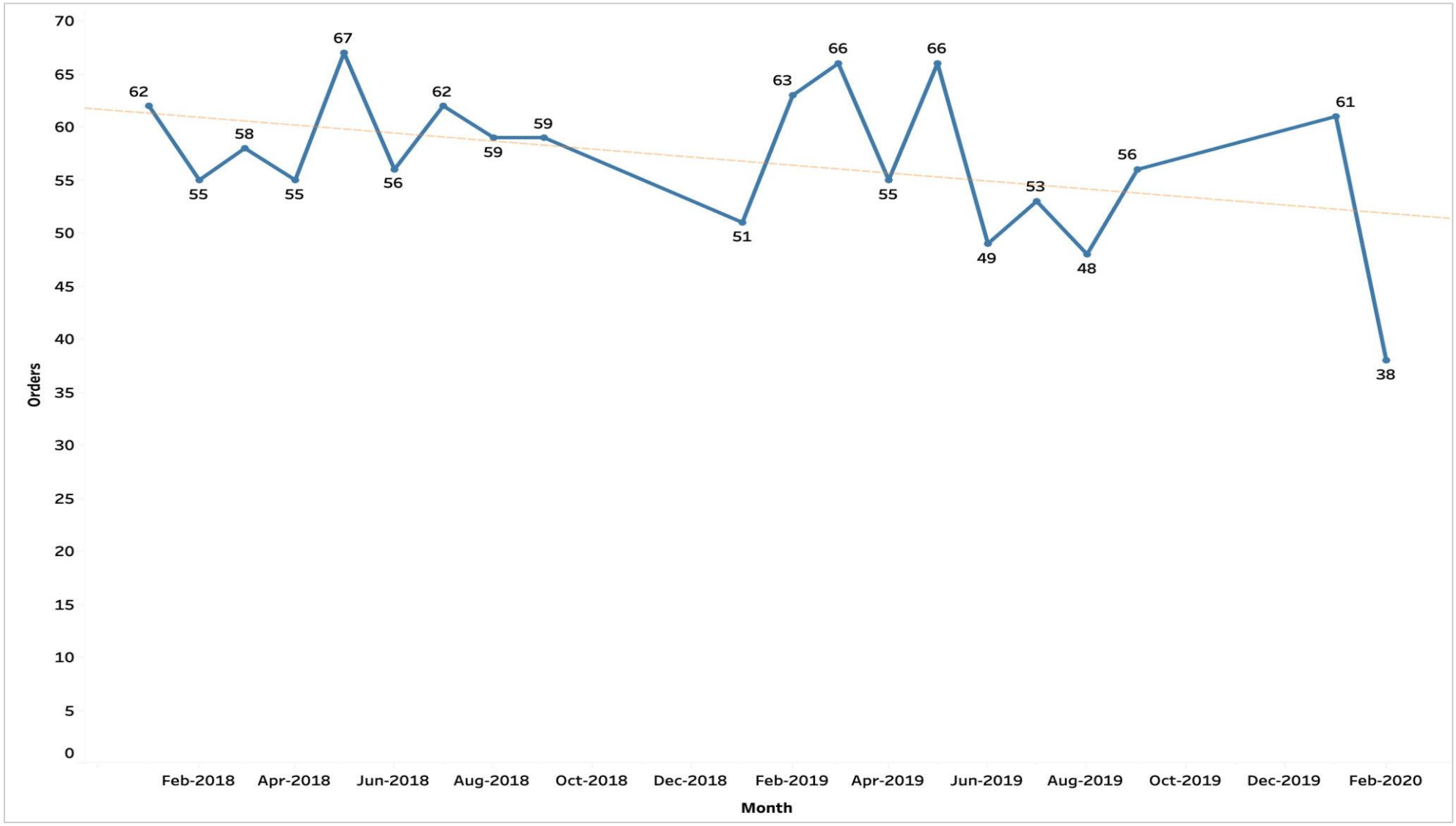


Inference:

- The **overall Order Trend** is in **decreasing** Quarter on Quarter
- On the entire data, **Quarter 1** has the **highest sale** & **Quarter 2 has the lowest**
- **No data found** for **Q4** in both **2018 & 2019**

Great Learning
POWER AHEAD

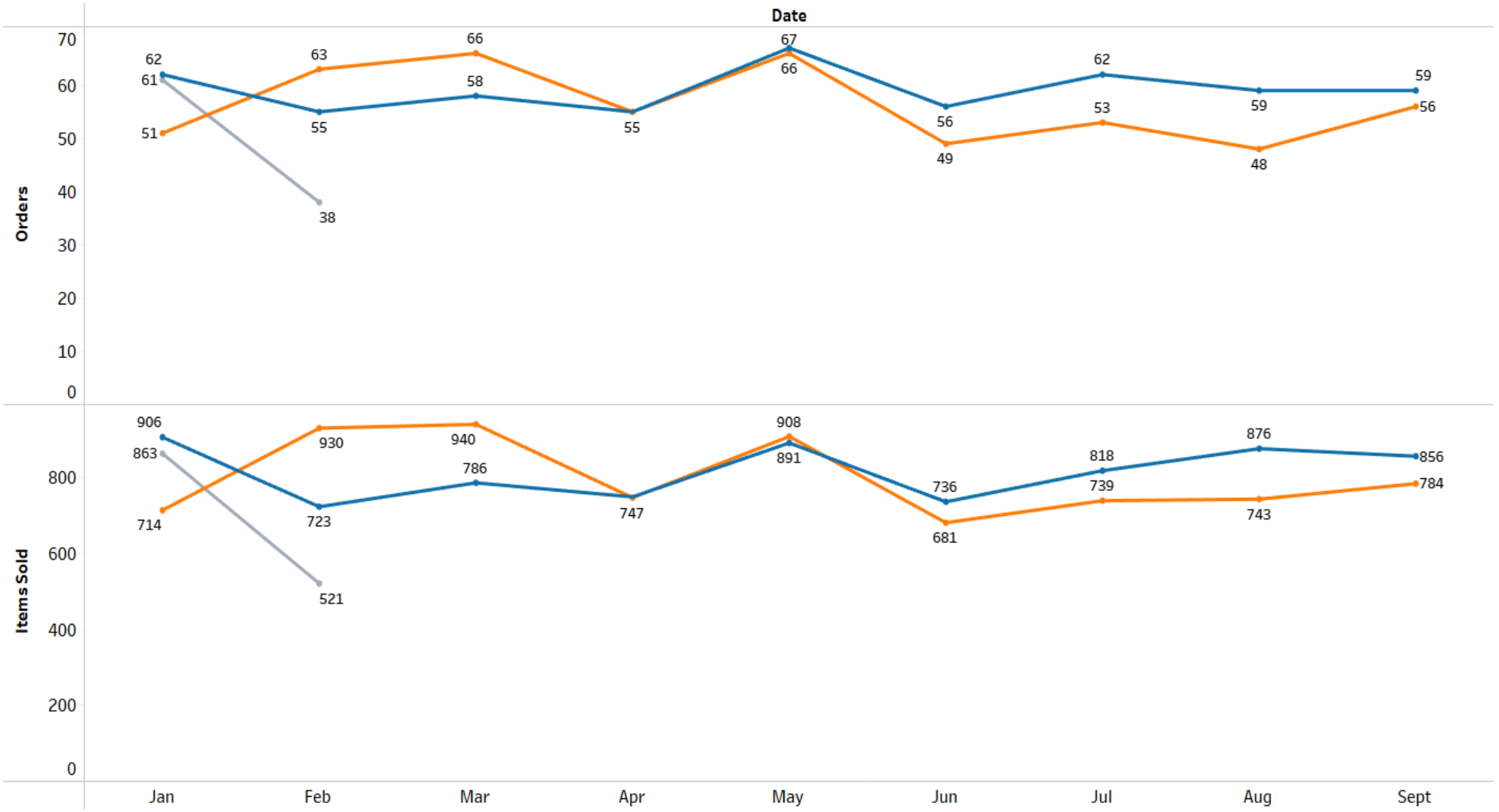
Orders Over Time: Monthly Order Trend



Inference:

- **Highest sales** observed in **May 2018 (67 Orders)**.
- This is followed by **Mar 2019 & May 2019 (66 Orders each**

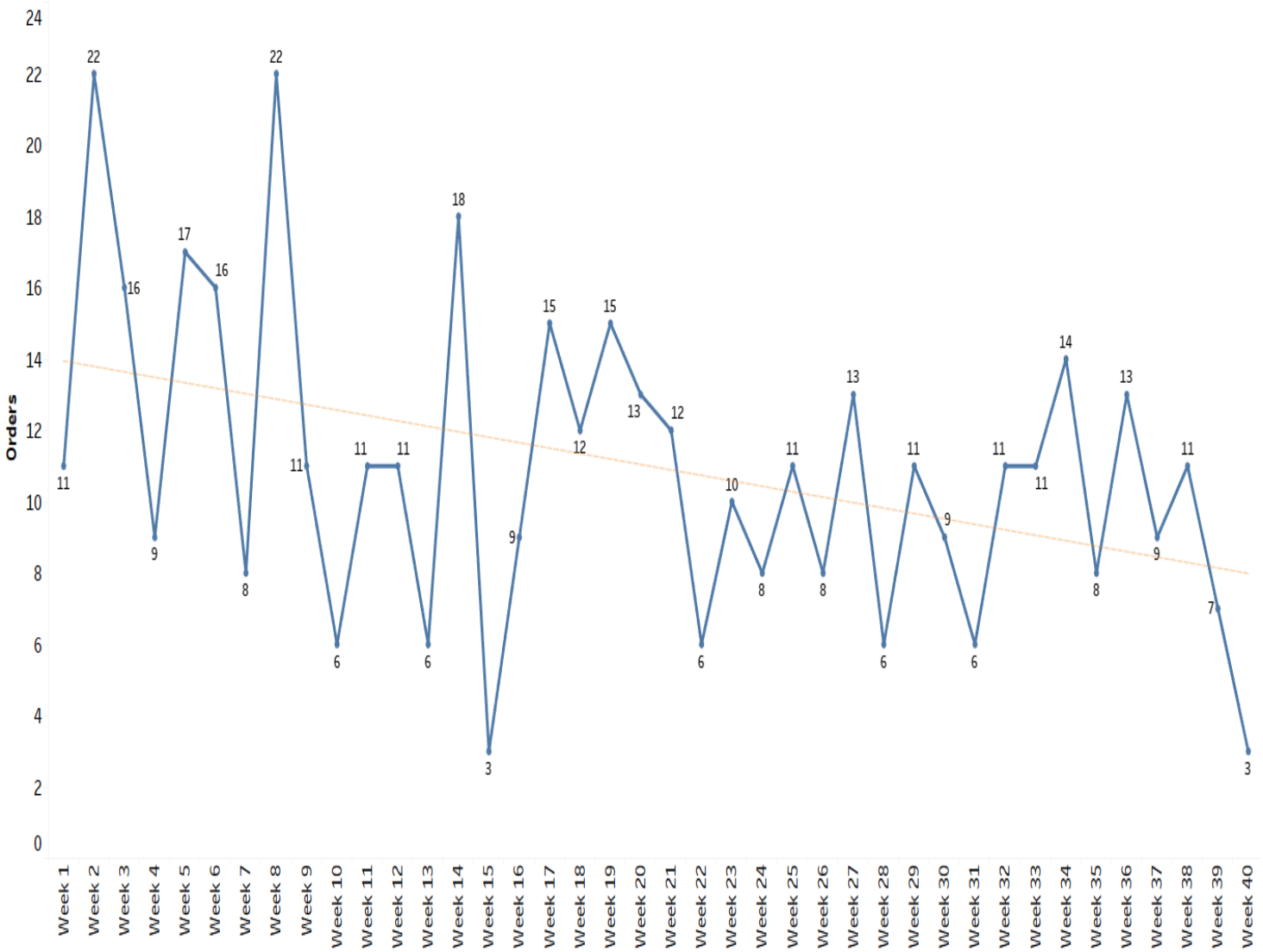
Orders Over Time: YOY Comparison

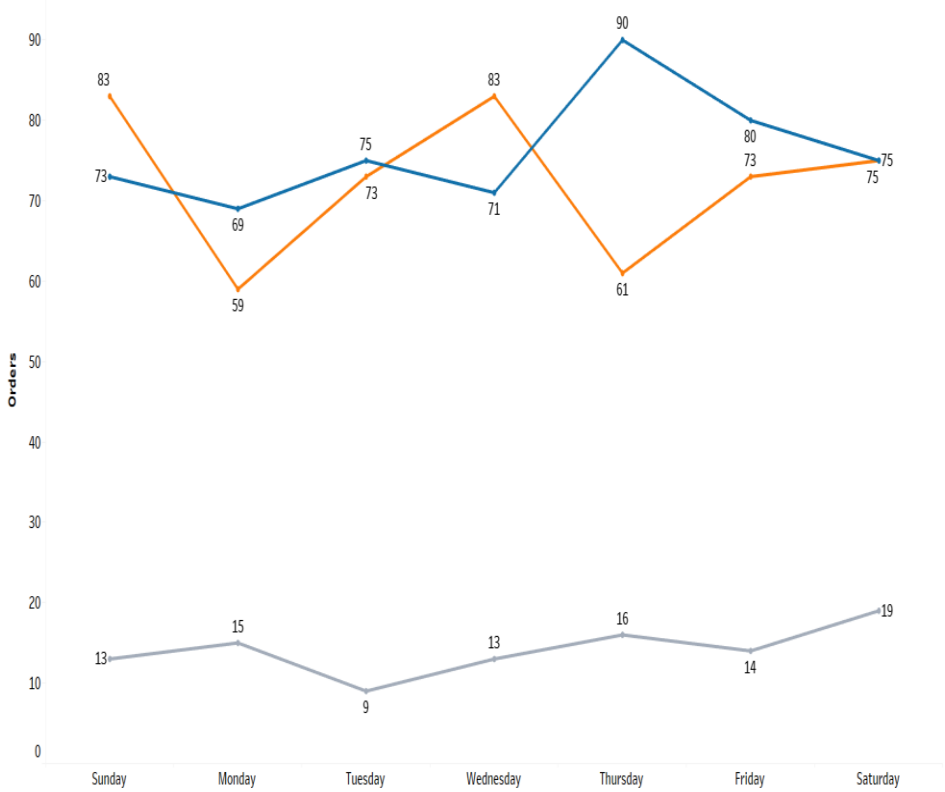


Inference:

- Orders were highest in **May month** for both **2018 & 2019**
- Item sold hd a different peak than order count for both years.
- **Apr & June** shows the **dip** for orders.
- For 2020, only 2 months data available, However, Feb'20 shows huge dip

Orders Over Time: WOW Order Trend

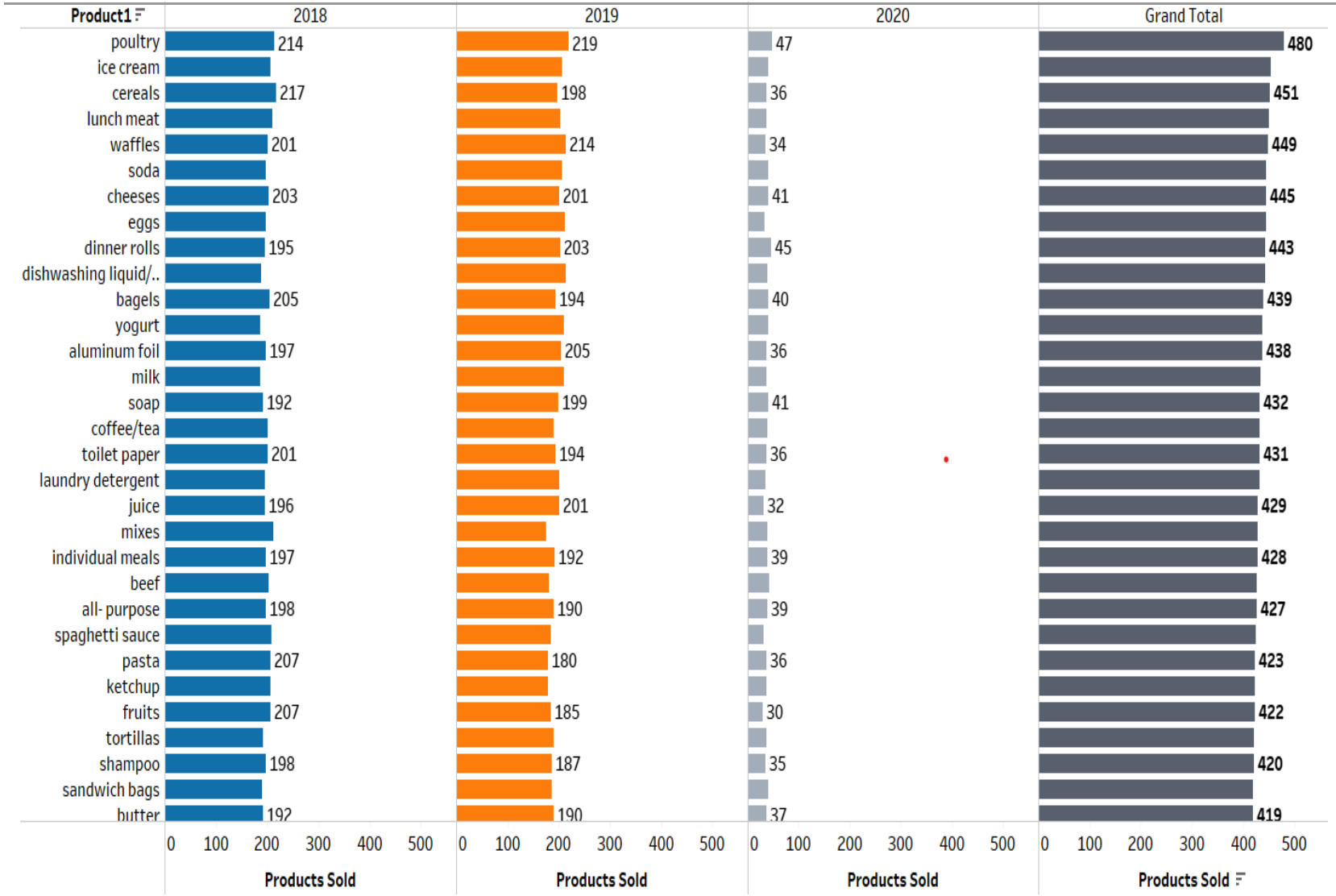




Inference:

- The Week Wise Trend is **decreasing**
- **Highest sales** are in **Week 8.**
- No particular pattern found in basis Weekday
- For 2020, only 2 months data available

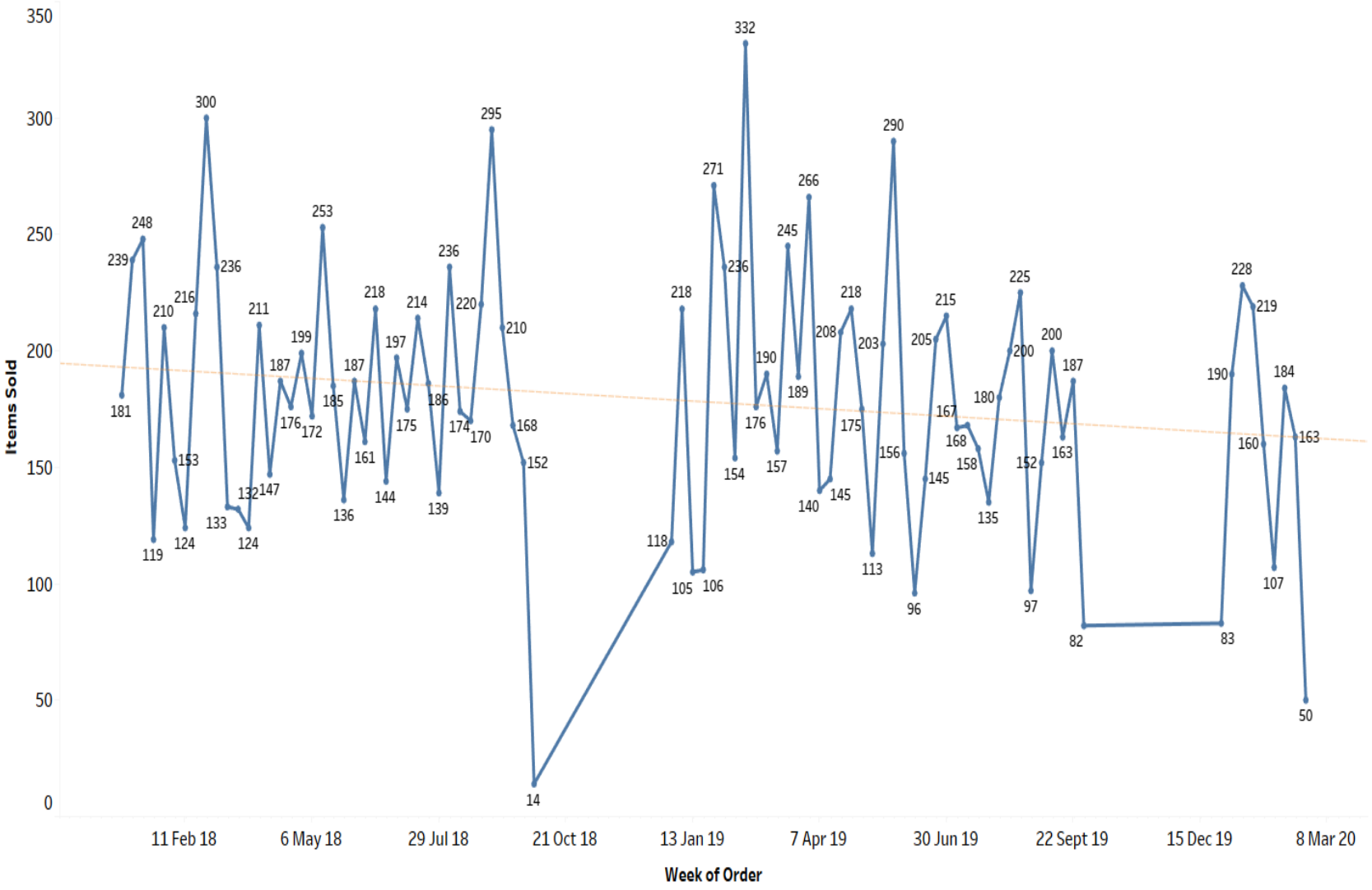
Top 10 Items Sold



Inference:

- **Poultry, ice cream, Cereals** are the most popular products.
- **Hand Soap, Sandwich Loaves, flour, pork** are the **least popular** products.

Weekly Product Demand



Insights

- Product sold is in decreasing trend for most of the items.
- On Analysing the Trend line, coefficient of slope for all products are in negative.
- This suggests decreasing trend, **except Yogurt** which has a slight positive Slope coefficient .
- Hence, we can say that, except Yogurt, all products demand is **decreasing**

Summary of Exploratory Analysis

Sales Overview:

- The data is, starting from January 1, 2018 having **Over a 2-year and 2-month period**. A total of **15,911 products** were sold through **1,139 orders**.
- Unfortunately, there's no data available for the fourth quarter in both years.

Yearly Comparisons:

- In 2019, the total number of orders **decreased by 26** compared to the previous year (2018).
- Additionally, the total number of products sold experienced a **decrease of 155 units in 2019** compared to 2018.

Trend Analysis:

- A **consistent decreasing** trend was observed in the number of orders over the analyzed period.
- There was a **mild decline in the total number of products sold**, although not as pronounced.
- A gradual decline was noted in sales and orders **after May**.

Product Popularity

- Among **37 products**, clear preferences emerged.
- **Poultry, ice cream, and cereals** were the **most popular** choices among customers.
- **Conversely, hand soap, sandwich loaves, flour, and pork** ranked as the **least popular** products during this period.

Weekly Demand Trends:

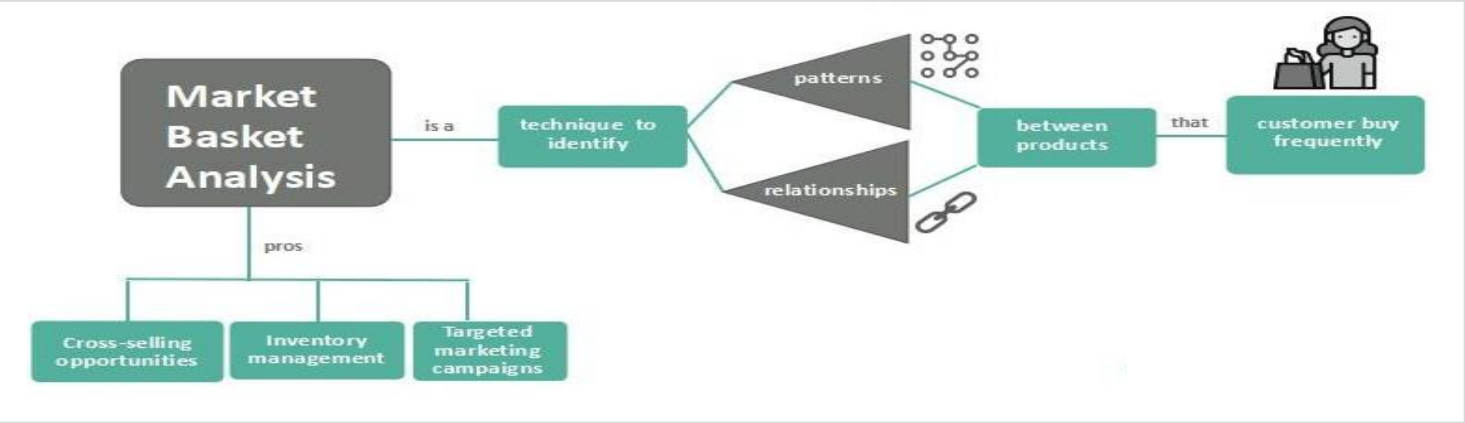
- Weekly demand trends for most products displayed a **consistent decrease** over time.
- One product, yogurt, showed a slight positive increase in demand.

Recommendation:

- 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.

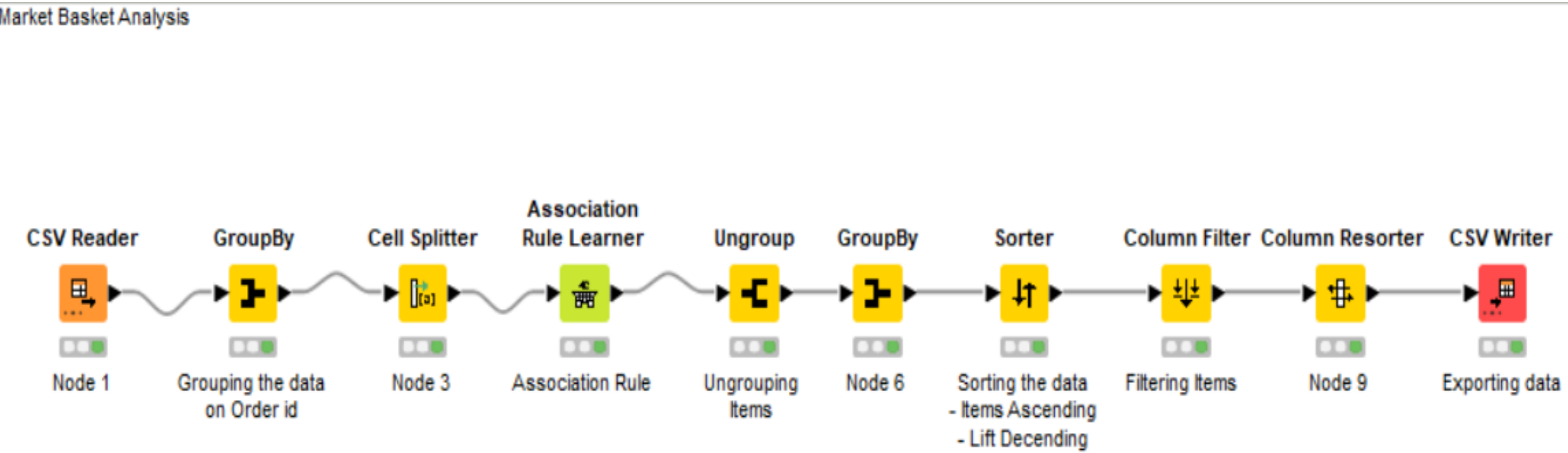
MARKET BASKET ANALYSIS

Market Basket Analysis



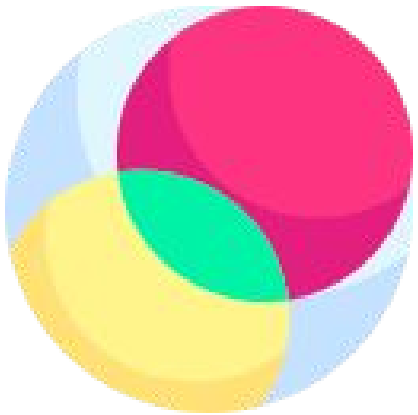
- **Definition:** Market Basket Analysis is a statistical technique that analyzes customer purchase patterns to identify associations between different products. It helps businesses understand which products are frequently purchased together and how customers' buying habits affect sales.
- **Data:** 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 WORKFLOW



Sample Output Table

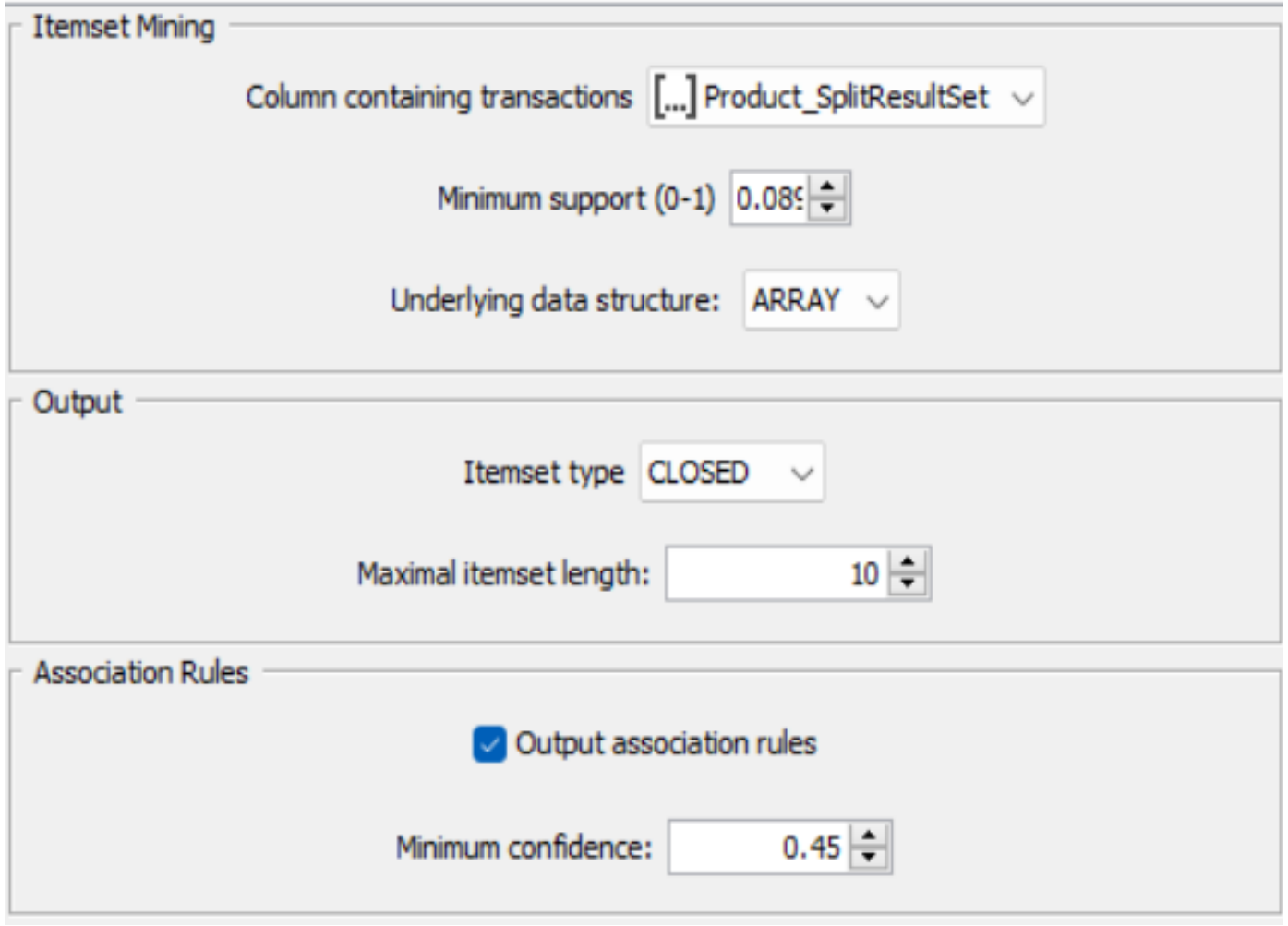
Row ID	S Date	I Order_id	S Product
Row0	01-01-2018	1	yogurt
Row1	01-01-2018	1	pork
Row2	01-01-2018	1	sandwich bags
Row3	01-01-2018	1	lunch meat
Row4	01-01-2018	1	all- purpose
Row5	01-01-2018	1	flour
Row6	01-01-2018	1	soda
Row7	01-01-2018	1	butter
Row8	01-01-2018	1	beef
Row9	01-01-2018	1	aluminum foil
Row10	01-01-2018	1	all- purpose
Row11	01-01-2018	1	dinner rolls
Row12	01-01-2018	1	shampoo
Row13	01-01-2018	1	all- purpose
Row14	01-01-2018	1	mixes
Row15	01-01-2018	1	soap
Row16	01-01-2018	1	laundry det...
Row17	01-01-2018	1	ice cream
Row18	01-01-2018	1	dinner rolls
Row19	01-01-2018	2	toilet paper
Row20	01-01-2018	2	shampoo
Row21	01-01-2018	2	hand soap
Row22	01-01-2018	2	waffles
Row23	01-01-2018	2	cheeses
Row24	01-01-2018	2	mixes
Row25	01-01-2018	2	milk



Associations Rule

Association Rule Parameters

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



Associations : KNIME Output

Row ID	S Conseq...	S implies	S Items (...	D Support	D Confide...	D Lift
Row85	poultry	<---	all-purpose	0.176	0.468	1.111
Row161	yogurt	<---	aluminum foil	0.177	0.461	1.199
Row49	ice cream	<---	aluminum foil	0.176	0.459	1.151
Row86	poultry	<---	aluminum foil	0.176	0.457	1.084
Row87	poultry	<---	beef	0.17	0.454	1.078
Row88	poultry	<---	butter	0.166	0.451	1.07
Row89	poultry	<---	cereals	0.181	0.457	1.084
Row50	ice cream	<---	cheeses, alu...	0.09	0.534	1.339
Row51	ice cream	<---	cheeses	0.179	0.458	1.15
Row90	poultry	<---	cheeses	0.181	0.463	1.098
Row152	waffles	<---	coffee/tea	0.172	0.454	1.151
Row91	poultry	<---	coffee/tea	0.175	0.461	1.093
Row92	poultry	<---	dinner rolls, ...	0.092	0.538	1.278
Row93	poultry	<---	dinner rolls, ...	0.09	0.543	1.287
Row94	poultry	<---	dinner rolls, ...	0.091	0.562	1.334
Row95	poultry	<---	dinner rolls, ...	0.09	0.557	1.321
Row38	eggs	<---	dinner rolls, ...	0.091	0.528	1.354
Row135	soda	<---	dinner rolls, ...	0.09	0.523	1.338
Row148	spaghetti sa...	<---	dinner rolls, ...	0.099	0.509	1.364
Row130	shampoo	<---	dinner rolls, ...	0.09	0.459	1.246
Row76	mixes	<---	dinner rolls, ...	0.09	0.464	1.235
Row66	laundry det...	<---	dinner rolls, ...	0.09	0.459	1.214
Row39	eggs	<---	dinner rolls, ...	0.091	0.468	1.202
Row9	cereals	<---	dinner rolls, ...	0.092	0.473	1.194
Row68	lunch meat	<---	dinner rolls, ...	0.091	0.468	1.186
Row81	pasta	<---	dinner rolls, ...	0.09	0.528	1.422
Row40	eggs	<---	dinner rolls, ...	0.095	0.554	1.421
Row96	poultry	<---	dinner rolls, ...	0.099	0.577	1.368
Row97	poultry	<---	dinner rolls	0.195	0.501	1.189
Row98	poultry	<---	dishwashing...	0.09	0.534	1.266
Row99	poultry	<---	dishwashing...	0.09	0.515	1.222

Associations

The generated association rules serve as a recommendation system during customer shopping experiences.

Recommendation Logic: If a customer shows interest in or has items from Set A in their cart, they will be recommended products from Set B.
Priority Order: When multiple products (consequents) are associated with an item in Set A, the recommendation order is determined by the lift value. The product with the highest lift is recommended first, followed by others.
Example: As an illustration, consider the association rules pertaining to Yogurt

Item	Implies	Consequent	Support	Confidence	Lift
Yogurt	=>	Juice	0.176	0.459	1.218
Yogurt	=>	Aluminum foil	0.177	0.461	1.199
Yogurt	=>	Eggs	0.175	0.454	1.166
Yogurt	=>	Waffles	0.174	0.452	1.147
Yogurt	=>	Poultry	0.181	0.470	1.116

- If a customer purchases Yogurt, the first recommendation will be of Juice as it has a higher probability of being purchased along with Juice based on past data (higher lift).
- Then Aluminum Foil, Eggs in that order.

Recommendations

Recommendation - 1

Item	Implies	Consequent	Support	Confidence	Lift
Spaghetti sauce, Poultry	=>	Dinner rolls	0.099	0.579	1.490
Dinner rolls, Poultry	=>	Spaghetti sauce	0.099	0.509	1.364
Dinner rolls	=>	Poultry	0.195	0.501	1.189
Poultry	=>	Dinner rolls	0.195	0.463	1.189
Spaghetti sauce	=>	Dinner rolls	0.172	0.461	1.186
Spaghetti sauce	=>	Poultry	0.171	0.459	1.089

- **Smart Pairing**: **Combine Dinner Rolls, Spaghetti Sauce, and Poultry** to create attractive product bundles. **Offer** discounts for customers **buying all three or a "buy 2 get 1 free" deal**.
  - **For example**: purchasing Dinner Rolls and Poultry would include Spaghetti Sauce at no extra cost.
- **Demand Insights**: While Spaghetti Sauce demand is declining, Poultry and Dinner Rolls show steady or slightly increasing weekly demands, with Poultry being the top seller.
- **Boost Sales**: **Pairing Spaghetti Sauce with Poultry and Dinner Rolls** can rejuvenate Spaghetti **Sauce sales**. This approach not only drives sauce sales but also encourages customers to choose the combo, **boosting overall sales** and customer satisfaction.

Recommendation - 2

Item	Implies	Consequent	Support	Confidence	Lift
Ice Cream, Soda	=>	Waffles	0.090	0.536	1.361
Ice Cream, Waffles	=>	Soda	0.090	0.523	1.338
Waffles, Soda	=>	Ice Cream	0.090	0.510	1.279
Soda	=>	Waffles	0.177	0.454	1.152

- **Combo Deals**: Explore **bundled offers** for **Ice Cream, Soda, and Waffles**, allowing customers to purchase these products **together at a reduced price**. Consider implementing a "**buy 2 get 1" promotion,** where buying **Ice Cream and Soda includes a complimentary pack of Waffles**.
- **Demand Boost**: While Waffles face decreasing demand, Ice Cream and Sodas show steady to slightly increasing weekly demand. By pairing them together, you can **revitalize Waffles sales** and entice customers to opt for the combo.

Recommendations - Summary

Boost Sales with Smart Strategies:

- **Combo Deals**: Offer discounted schemes for product combinations: Offer a "Buy Two Get One Free promotion to encourage customers to purchase more items at once.
  - Dinner Rolls, Spaghetti Sauce & Poultry
  - Yogurt, Juice & Aluminum Foil
  - Ice Cream, Soda & Waffles
- **Promotional Sales**: Implement frequent sale offers for slower-selling products such as Hand Soap, Sandwich Loaves, and fruits. Create a "Paper Products Bundle" offer that includes paper towels, toilet paper, and/or tissues at a discounted price.

- **Leverage Associations**: Explore additional product associations to enhance sales:
  - Soda & Eggs
  - Dinner Rolls & Eggs
  - Ice Cream & Cheeses
  - Yogurt & Poultry
  - Lunch Meat & Poultry
- Consider bundled offers for these product pairs, providing customers with a discounted rate when purchasing them together. This strategy can boost sales for these products and create a mutually beneficial sales impact.
- 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.

Thank You