Marketing and Retail Analytics Project – Part 2

CHETAN DUDHANE

PGP DSBA GRP 2 JULY B

06-JUN-2021

Problem Statement

Market Basket Analysis

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?

Synopsis

- Total No. of Sales Entries = 20641
- Total No. of Variables = 3
- No. of Missing Entries = 0
- No. of Duplicate Entries = 0
- Sales Data in the dataset is -
  - From 1st Jan 2018 to 26th Feb 2020
- 603 days of data

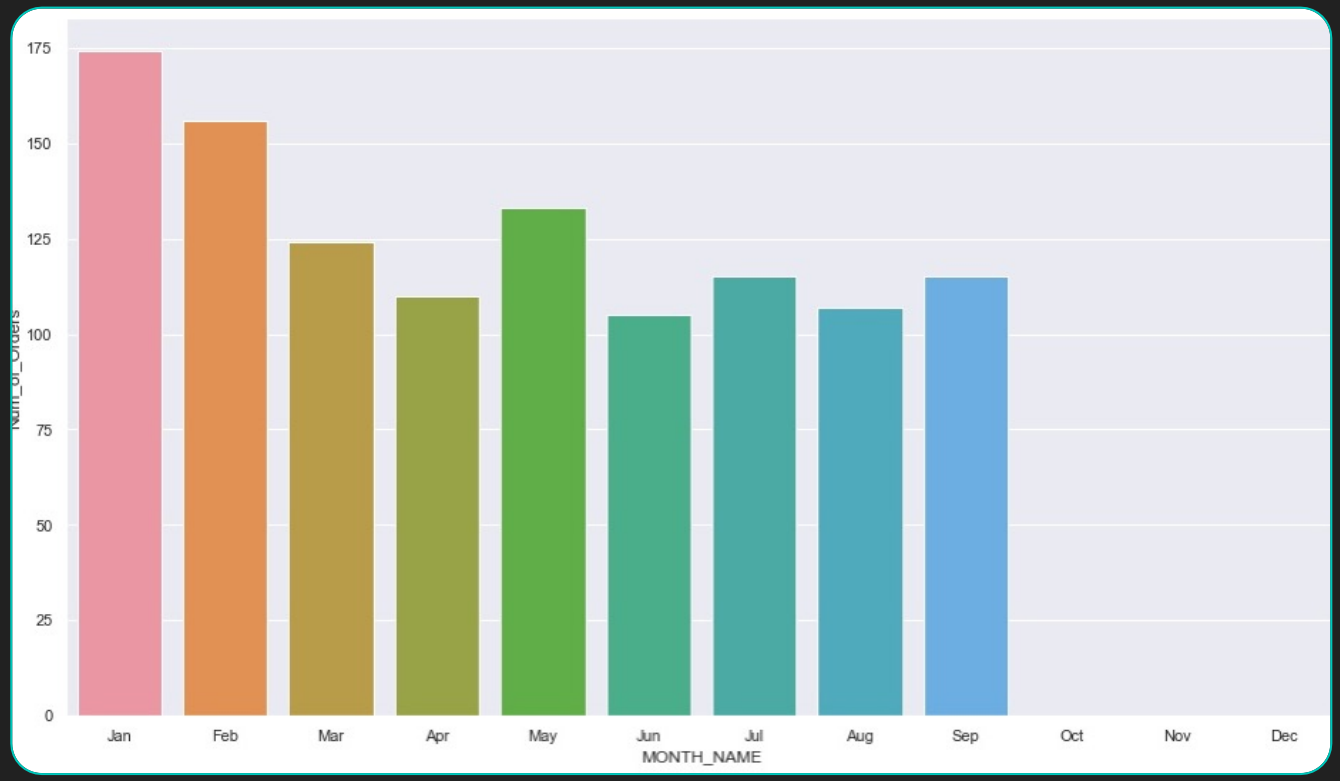
- Range of Products = 37
- Number of Orders (Invoices) = 1139
- Max Transactions occurred on 8th Feb 2019 – 183 transactions
- Avg no. of daily orders = 2
- Avg no. of Products per order = 18

Synopsis contd..

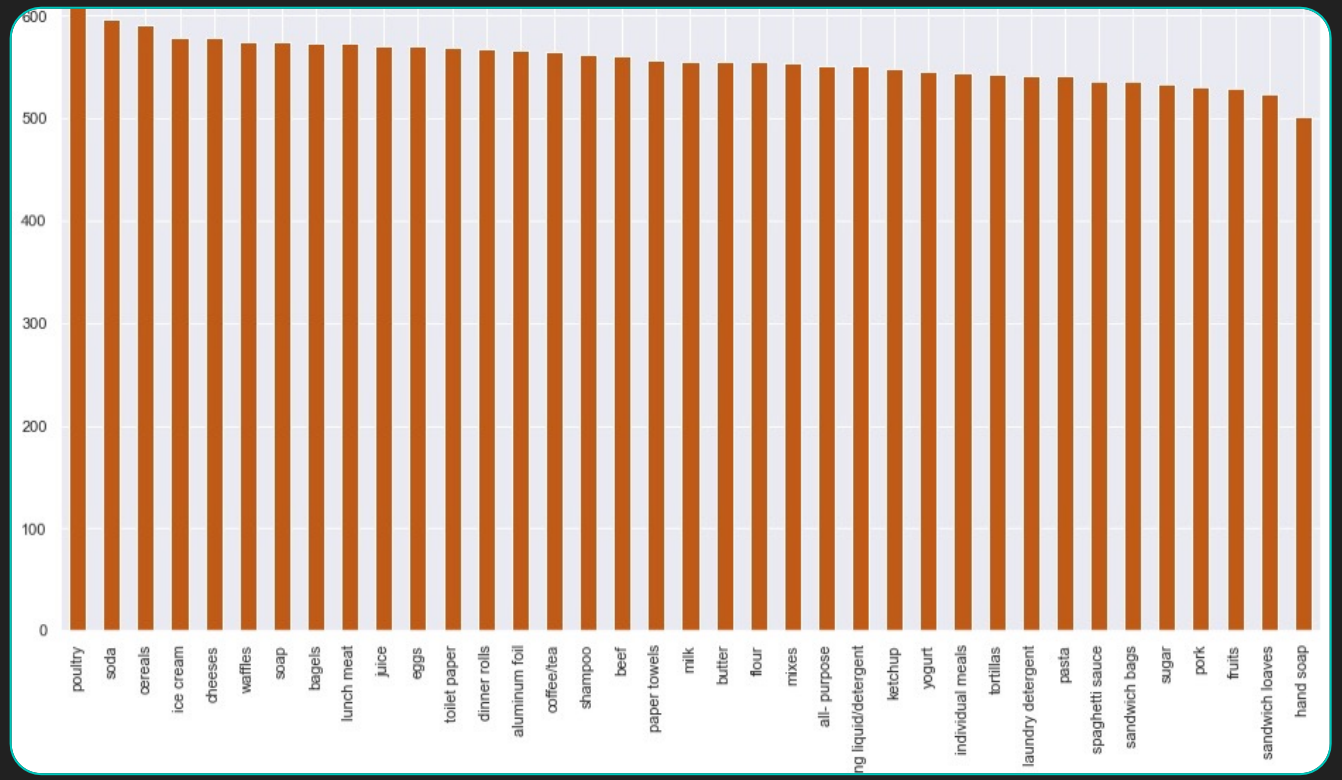
- Store hasn't submitted data for Oct-Nov-Dec of all years

Or

- Store remains closed in this period, possibly due to holidays
- Max transactions in Jan-Feb



Synopsis contd..



- All products are ordered with around the same frequency
- Poultry is ordered the most – 640 transactions
- Hand soap ordered the least

Synopsis contd..

- Tools Used in this Project –
  - Python – For basic exploration, EDA and Time Series
  - KNIME – For Market Basket Analysis
  - Excel – For Summary Pivots and Views

Raw Sales Data

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

Data Description

Numerical variables

	count	mean	std	min	25%	50%	75%	max
Order_id	20641	575.99	328.56	1	292	581	862	1139

Categorical variables

	count	unique	top	freq
Date	20641	603	08-02-2019	183
Product	20641	37	poultry	640

Summary Info

Raw data Summary Info

```
<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
```

Info after Date-Type formatted

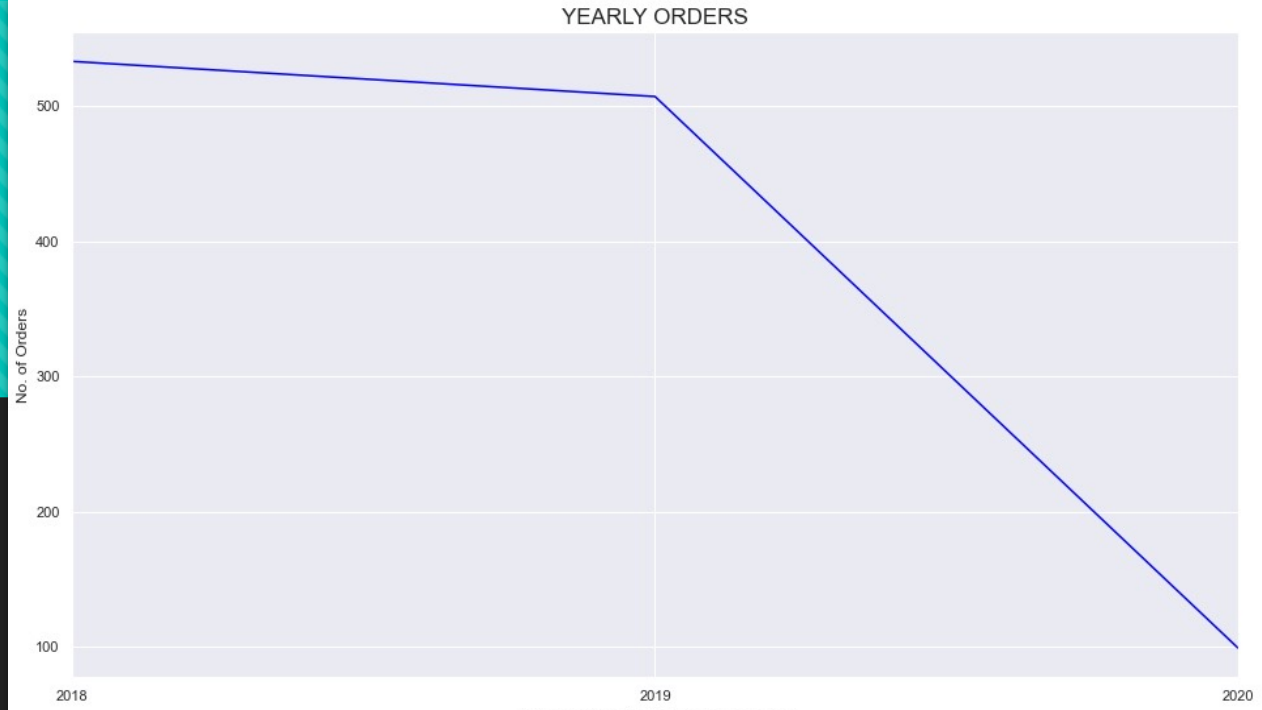
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20641 entries, 0 to 20640
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      20641 non-null  datetime64[ns]
 1   Order_id  20641 non-null  int64
 2   Product   20641 non-null  object
dtypes: datetime64[ns](1), int64(1), object(1)
memory usage: 483.9+ KB
```

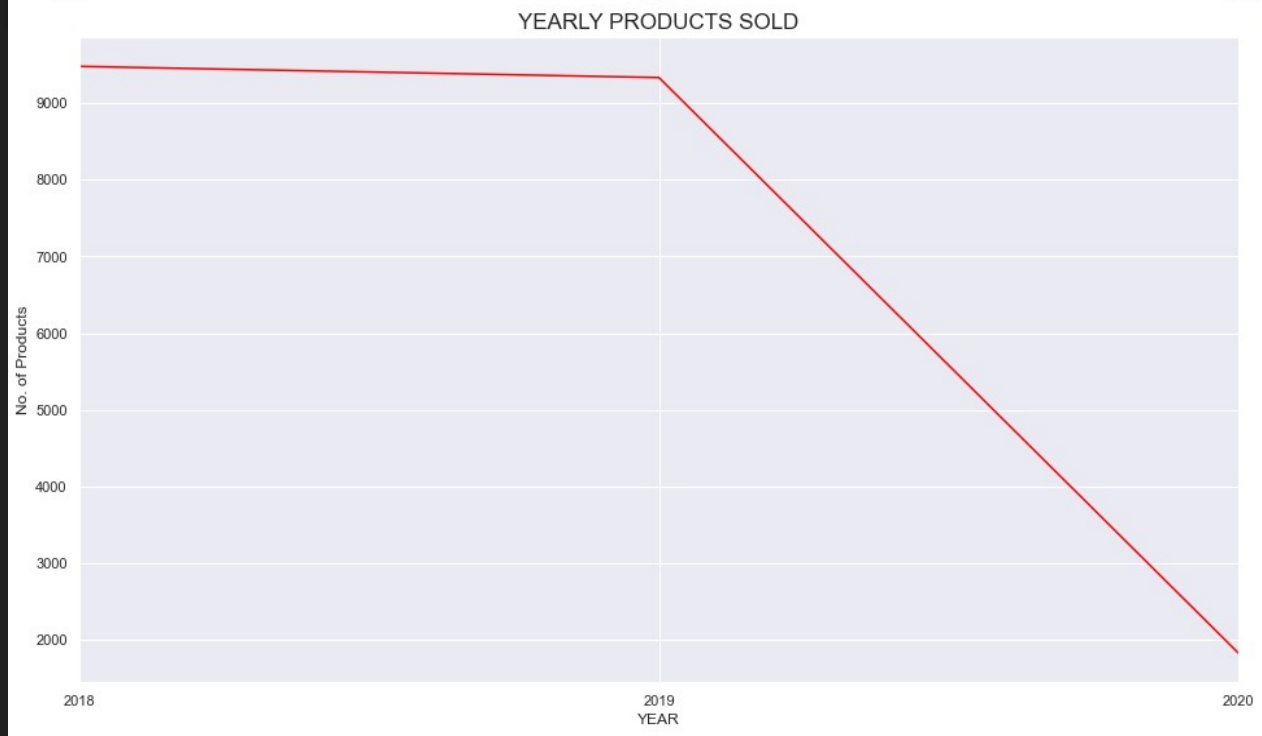
Yearly Trend

- Number of Transactions and Number of Products show a similar trend over the years

---

- Please note that the transactions are full years of 2018, 2019 and only first 2 months of 2020

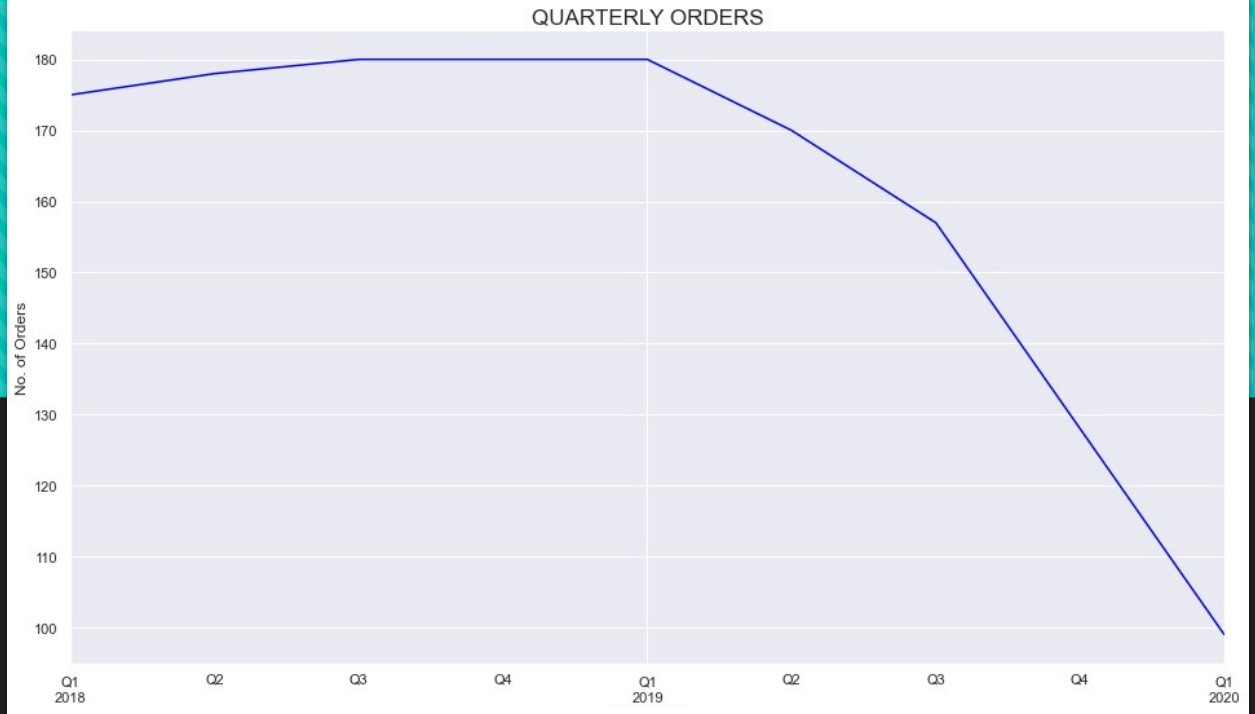


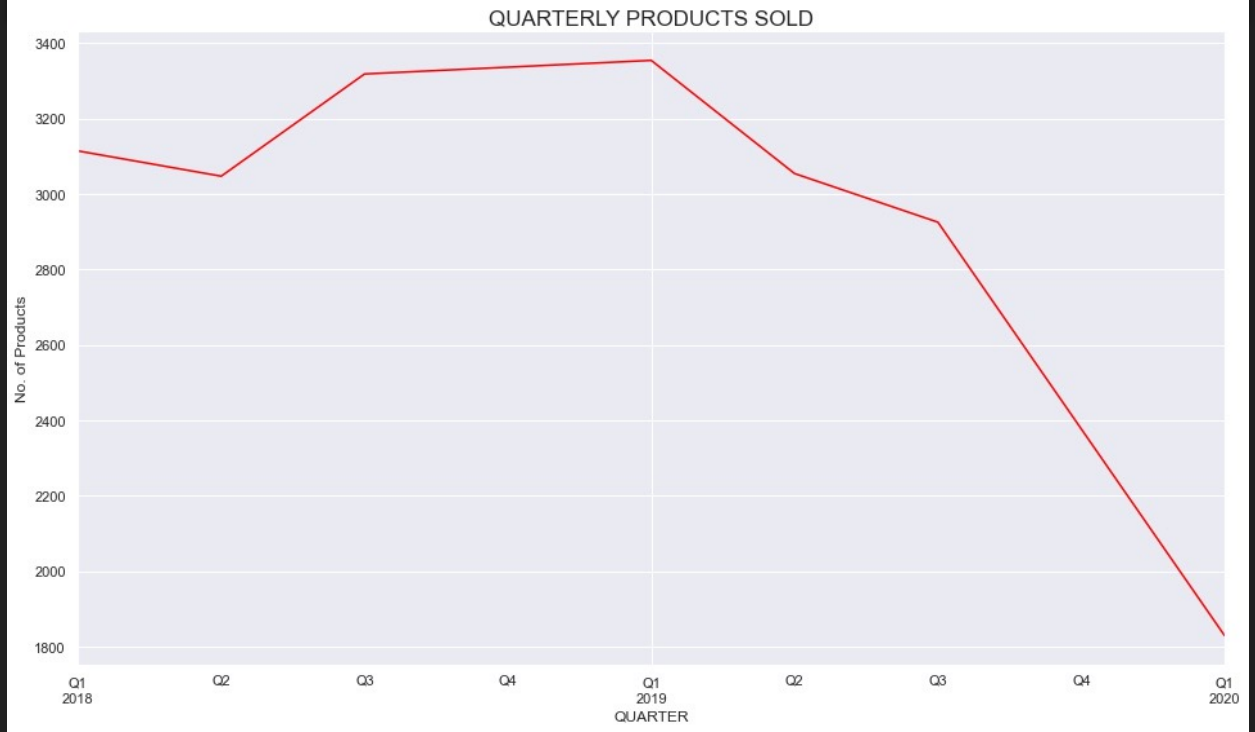


Quarterly Trend

- There seems to be a dip in Q2 every year from the high of Q1

- Please note that the transactions are full years of 2018, 2019 and only first 2 months of 2020

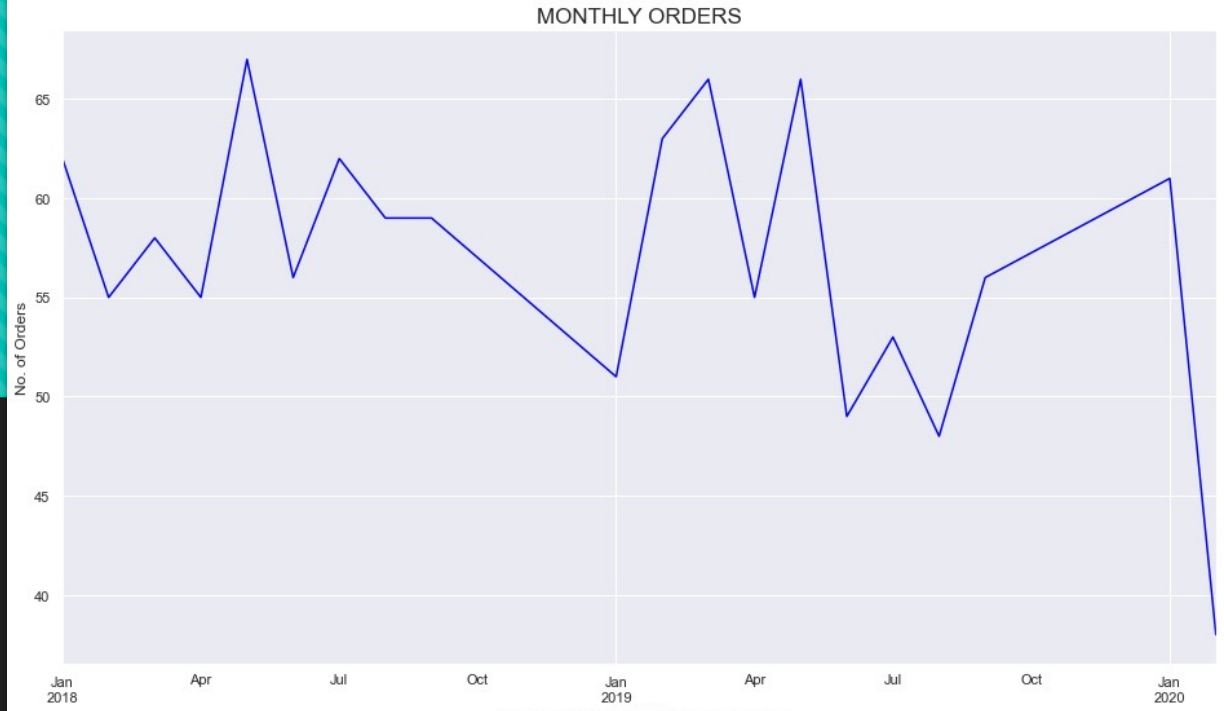




Monthly Trend

- There is no fixed monthly pattern

- Please note that the transactions are full years of 2018, 2019 and only first 2 months of 2020

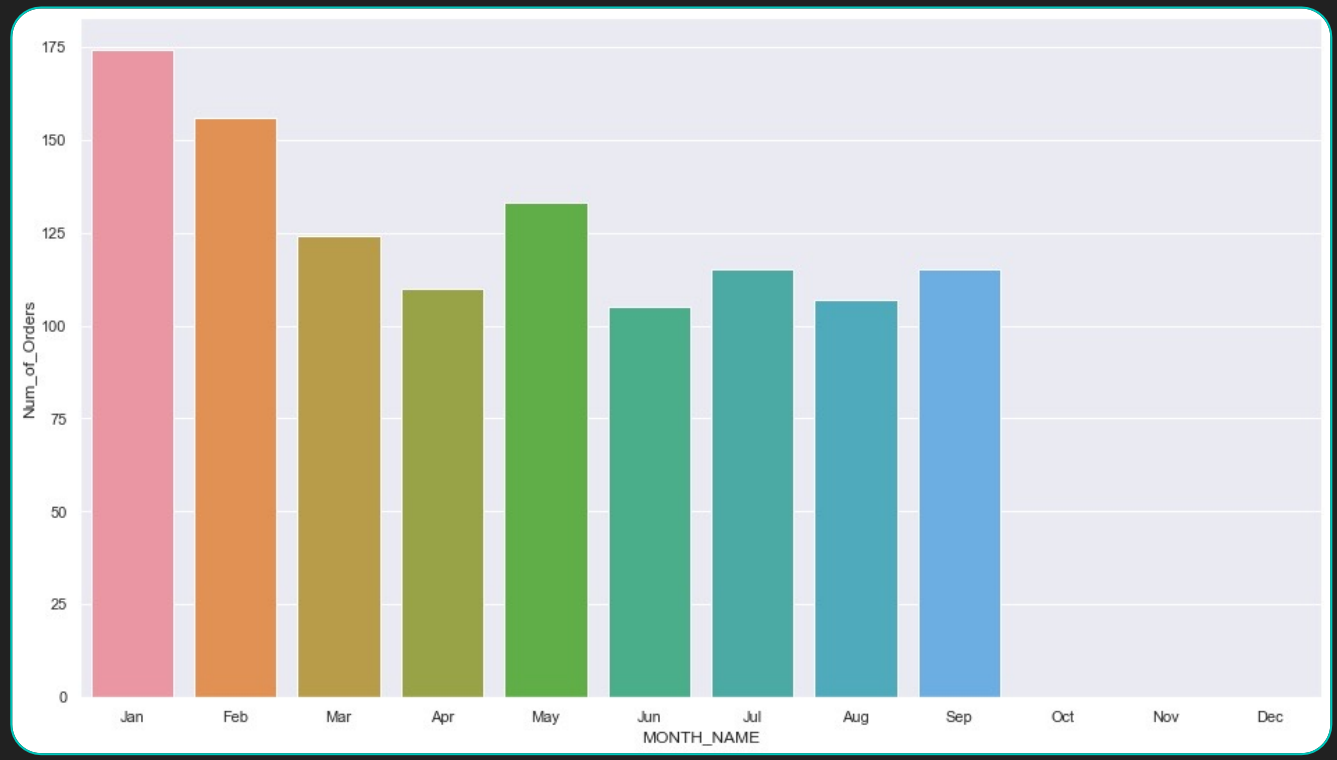


Monthly Transactions

- We notice, there are no transactions for Q4 i.e., Oct–Nov-Dec
- Transactions not submitted for analysis or Store closes every year, mostly due to holidays

---

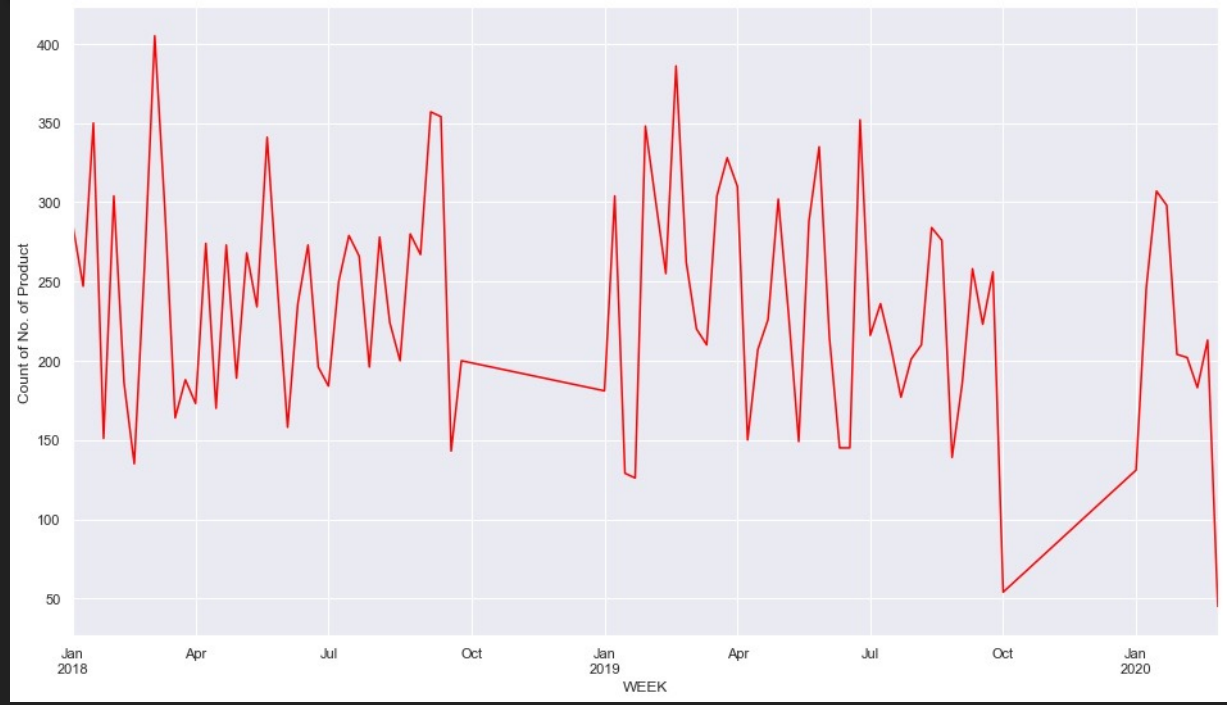
- Please note that the transactions are full years of 2018, 2019 and only first 2 months of 2020



Weekly Trend

- Weekly patterns shown

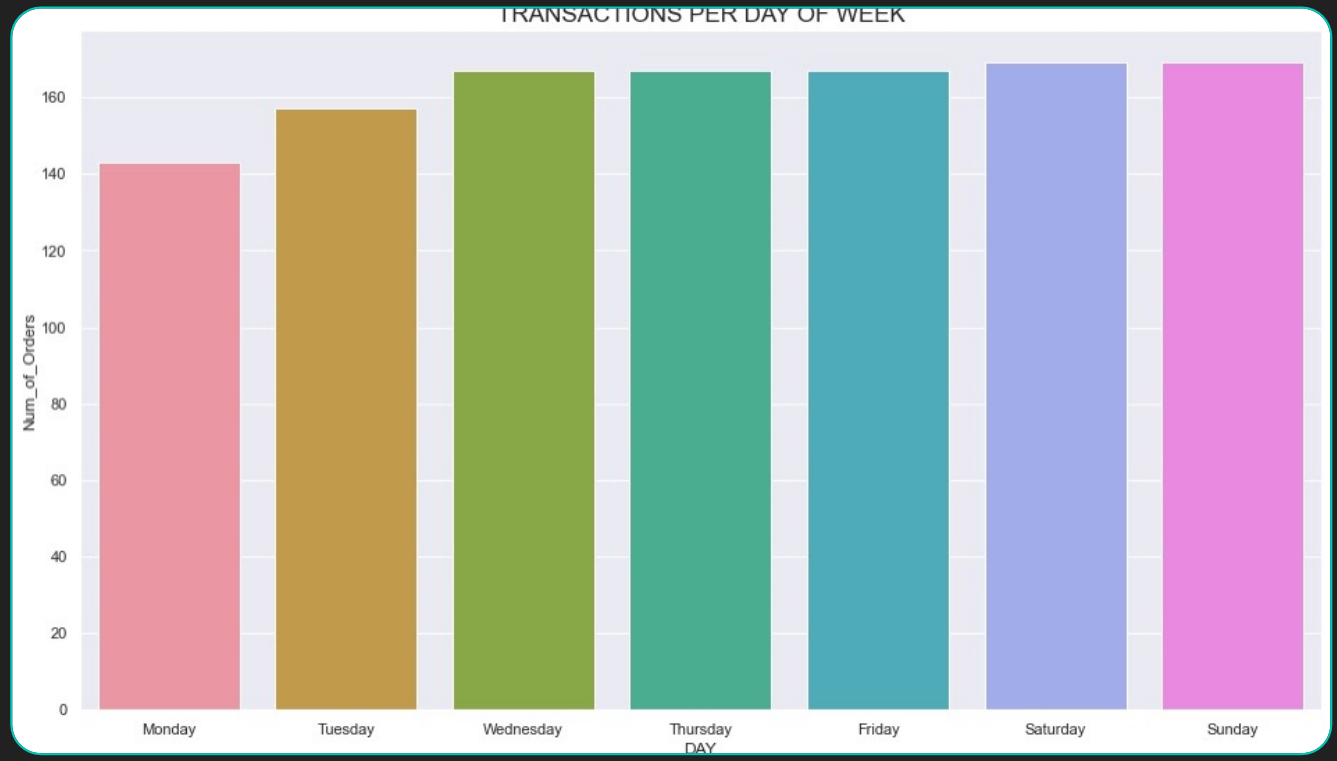
- Please note that the transactions are full years of 2018, 2019 and only first 2 months of 2020



Transactions per Day of the Week

- Monday has the lowest transactions
- All other days have almost the same transactions per day

- Please note that the transactions are full years of 2018, 2019 and only first 2 months of 2020



Market Basket Analysis

- MBA is a strategy adopted by the retailers to gauge customer buying pattern
- It is all about understanding customer's basket behaviour
- It investigates general group of items; customers end up buying together
- MBA finds relationships between the items in a customer's shopping cart based on various metrics
- This level of understanding of the customer's shopping behaviour is used by the retailers in Target strategy and Recommendation systems

Association Rules

- Association Rule is the most important Data Mining technique used in Market Basket Analysis
- It tries to associate different items in a shopping cart with some others using some metrics
- Mainly, it is related to the statement "What goes with What"
- Association Rules give a result like "Set A ➔ Set B" –

  IF (items in Set A are bought)

  THEN (items in Set B will be bought)
- It is a directional rule, and the inverse does not necessarily hold true
- Here, Set A is called 'Precedent' and Set B is called 'Consequent'

Support, Confidence and Lift

- SUPPORT –
  - Support of A – is the fraction of transactions of A out of the total transactions
  - If item A is bought 100 times out of the total 1000 transactions of the store, then Support of A = 100/1000 = 0.1 (10%)
  - Similarly, if items A and B are together bought 50 times, then Support of A and B = 50/1000 = 0.05 (5%)

- $Support\ of\ A = \frac{Number\ of\ Transactions\ containing\ A}{Total\ Transactions}$

Support, Confidence and Lift

- CONFIDENCE –
  - Confidence (A ➔ B) – is the likelihood of a customer buying item A, will also buy item B
  - This is the Probability of B given that A has been bought
  - Out of the 100 times that A has been bought, if B is bought 50 times along with A, then,

    Confidence(A ➔ B) = 50/100 = 0.5 (50%)

- $Confidence\ of\ A \Rightarrow B =$ P(B | A)

$$\frac{Number\ of\ Transactions\ containing\ A\ and\ B}{Total\ Transactions\ containing\ A}$$

Support, Confidence and Lift

- LIFT –
  - Lift is the most important metric to consider when choosing an association rule
  - Given A is bought, then –

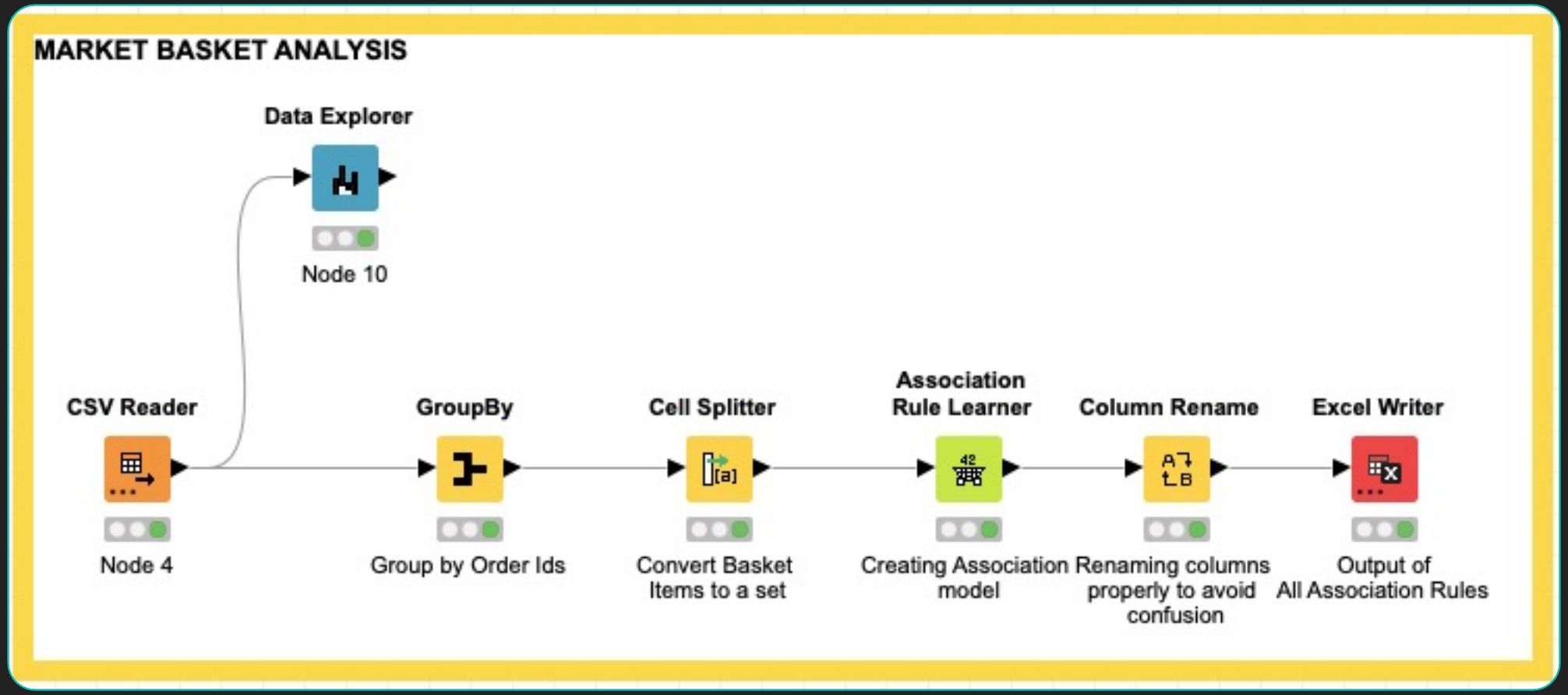
    - Lift is the % increase in chance of buying B
  - Lift(A → B) < 1 =➔ Presence of A has decreased the chance of buying B
  - Lift(A → B) > 1 =➔ Presence of A has increased the chance of buying B
  - For example, Lift = 1.57 =➔ Chance of buying B has increased by 57%

- $Lift\ of\ A \Rightarrow B = \dfrac{Confidence\ of\ A \longrightarrow B}{Support\ of\ B}$

MBA on Grocery Store Dataset

- We perform MBA using KNIME on the given Grocery Store Dataset
- We choose multiple thresholds for Min Support and Min Confidence to filter out less frequent and less appropriate rules
- Finally chosen thresholds for –
  - Support = 0.03 (3%)
  - Confidence = 0.55 (55%)
- The above values indicate –
  - We want to create rules with only those items which appear in at least 3% of transactions
  - We want a minimum Confidence of 55% in the rule i.e Suggested item should be bought at least 55 out of every 100 times Basket items are bought
- We get an output of 10550 rules
- All rules - Average Lift, Confidence and Support is 1.48, 0.57 and 0.04 respectively

KNIME Workflow - MBA



Final Output Rules Table (Top 20 rules sorted by Lift)

Rule #	Support	Confidence	Lift	Suggested_Item	implies	Basket_Items
1	0.03	0.80	2.19	paper towels	<---	[eggs, ice cream, pasta, lunch meat]
2	0.03	0.78	2.16	paper towels	<---	[eggs, ice cream, pasta, cereals]
3	0.03	0.73	2.07	flour	<---	[dishwashing liquid/detergent, cheeses, waffles, soda]
4	0.03	0.74	2.04	paper towels	<---	[eggs, dinner rolls, ice cream, pasta]
5	0.03	0.72	1.99	paper towels	<---	[eggs, poultry, ice cream, pasta]
6	0.03	0.78	1.95	ice cream	<---	[paper towels, eggs, pasta, lunch meat]
7	0.03	0.76	1.95	soda	<---	[dishwashing liquid/detergent, cheeses, flour, waffles]
8	0.03	0.72	1.93	pasta	<---	[paper towels, dishwashing liquid/detergent, eggs, ice cream]
9	0.04	0.70	1.92	paper towels	<---	[all- purpose, individual meals, toilet paper]
10	0.03	0.71	1.91	spaghetti sauce	<---	[dinner rolls, poultry, laundry detergent, juice]

Final Output Rules Table contd…

Rule #	Support	Confidence	Lift	Suggested_Item	implies	Basket_Items
11	0.03	0.75	1.91	eggs	<---	[paper towels, dishwashing liquid/detergent, ice cream, pasta]
12	0.03	0.69	1.91	paper towels	<---	[dishwashing liquid/detergent, eggs, ice cream, pasta]
13	0.03	0.71	1.90	pasta	<---	[paper towels, eggs, poultry, ice cream]
14	0.03	0.72	1.85	eggs	<---	[paper towels, poultry, ice cream, pasta]
15	0.03	0.69	1.84	pasta	<---	[paper towels, eggs, dinner rolls, ice cream]
16	0.04	0.64	1.83	sandwich loaves	<---	[all- purpose, flour, individual meals]
17	0.03	0.71	1.83	eggs	<---	[paper towels, dinner rolls, ice cream, pasta]
18	0.04	0.68	1.82	pasta	<---	[hand soap, soda, aluminum foil]
19	0.04	0.68	1.82	ketchup	<---	[butter, aluminum foil, soap]
20	0.04	0.63	1.81	sandwich loaves	<---	[paper towels, flour, individual meals]

Output Rules

- Rules consist of Precedent (Basket_Items) and Consequent (Suggested_Item), which gives us rules – IF Basket_Items are bought, THEN Suggested_Item are likely to be bought
- Rule #1 –
  - IF [eggs, ice cream, pasta, lunch meat] is bought then there is 2.19 times likelihood that [paper towels] will be bought with 80% Confidence
- Rule #10 –
  - IF [dinner rolls, poultry, laundry detergent, juice] is bought then there is 91% more likelihood that [sandwich loaves] will be bought with 71% Confidence
- Rule #17 –
  - IF [paper towels, dinner rolls, ice cream, pasta] is bought then there is 83% more likelihood that [eggs] will be bought with 71% Confidence

Recommendations

Suggested Items	Count
poultry	1330
cheeses	598
lunch meat	575
soda	549
eggs	499
yogurt	487
dinner rolls	427
ice cream	411
waffles	404
juice	393

- The table shows top 10 items suggested by the rules
- Most of these items are stored in a refrigerator
- *It is recommended to have Refrigerators lined up on a side wall which is accessible from all aisles*
- Poultry, Eggs, Ice Cream, Meat buyers tend to buy Paper Towels more often
- *It is recommended to keep Paper Towels for use and on a shelf to sell near Refrigerators*
- Looking at the cart mix of products, it seems customers very often come to pick up ingredients for pasta
- *It is recommended to make a Pasta Bag – containing eggs, cheese, pasta and spaghetti sauce*

Recommendations contd…

- Beef and/or Pork buyers are seen to have a high likelihood of also buying cleaning products such as soap, hand soap, shampoo and dishwashing liquid
- *It is recommended to have a small shelf of choicest cleaning agents near the deep freezers containing these meats*
- Discount Offers
  - Sandwich loaves and bags are sold less, though they are a common kitchen item
  - Make a bundle offer of BUY 2 GET 1 FREE of Sandwich Loaves with Sandwich bags – depending on the price margins
  - Similarly, Laundry Detergents and Hand Soap lie on a lower sales scale
  - They should be bundled into an offer of BUY 2 DETERGENTS AND GET 1 SOAP FREE
- Poultry is the most sold item – most customers seem to come primarily for this
- Maximise this by cross-selling other items with this like – BUY 2 POULTRY GET 50% OFF ON PAPER TOWELS or BUY 3 POULTRY GET $10 STORE CREDIT

Thank You

Marketing and Retail Analytics

Milestone 2 Project

By

CHETAN DUDHANE

PGP DSBA JULY B GRP 2