Market Basket Analysis

Business Report

Yedupati Venkata Yamini
17-Sep-2023

Table of contents:

List of tables

List of figures

Agenda

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. Below are the tasks done

- **Data set exploration and exploratory analysis.**
- **Association rule learning and formulation.**
- **Inferences from association rules, framing recommendations and offers/discounts.**

Exploratory Data Analysis

Executive Summary of Data

- The dataset_group dataset has 20641 rows and 3 columns.
- The dataset contains the following columns:
  - **Date**: Date on which the order was placed.
  - **Order_id**: ID (or) Number of the order
  - **Product**: Product ordered as part of the specified Order_id.
- First and last 5 rows of the 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

	Date	Order_id	Product
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

Figure 1: First and last rows of the dataset

Executive Summary of Data

- Basic information of the dataset:

```
<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
```

Figure 2: Basic information of the dataset

- As seen in the above table, there are no null values in the dataset.
- The dataset has features with 1 integer data type, 1 categorical data type, 1 datetime data type.

Executive Summary of Data

- There are no duplicate rows in the dataset.
- The dataset contains records for 1140 unique orders.
- Basic summary of unique products ordered in the store:

```
count       20641
unique         37
top       poultry
freq          640
Name: Product, dtype: object
```

Figure 3: Summary of products ordered from the store

- Unique products ordered by customers are 37. Poultry related products have been ordered the most.

Top 10 ordered products

- All products in the store are comparably ordered. Of all the products, poultry are highest ordered.

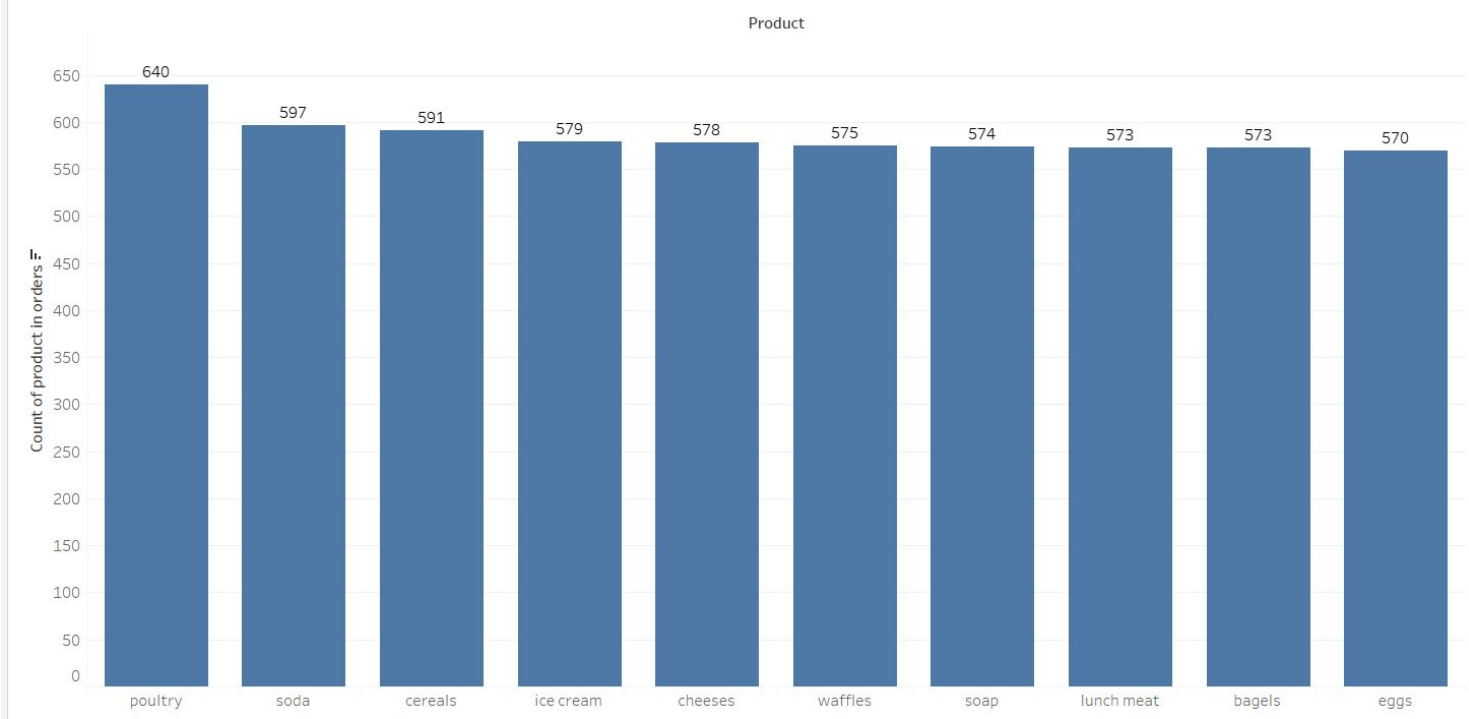


Figure 4: Top 10 ordered products

Bottom 10 ordered products

- All products in the store are comparably ordered. Of all the products, hand soap is least ordered.

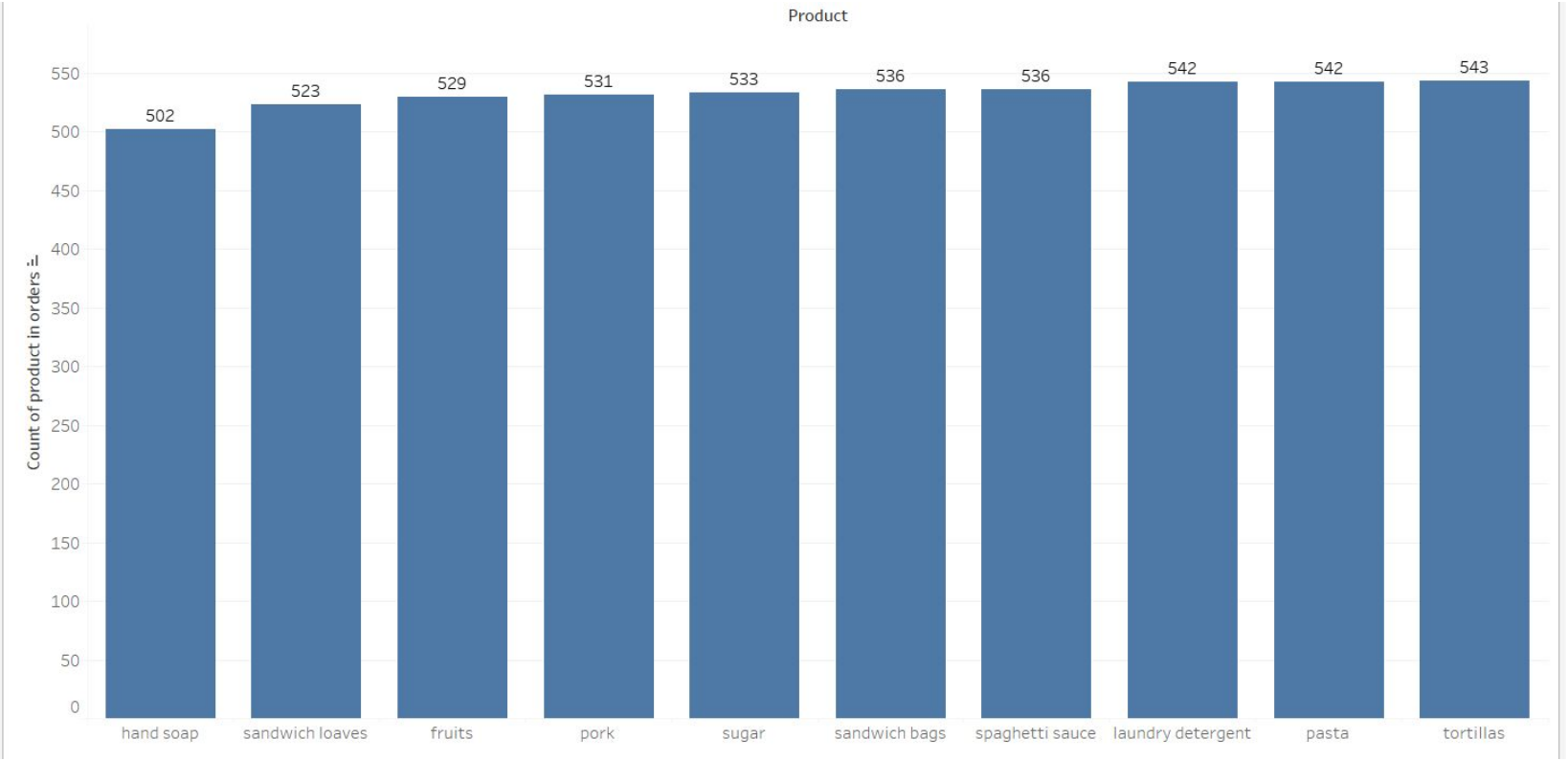


Figure 5: Least 10 ordered products

Trends of orders containing popular products in 2018

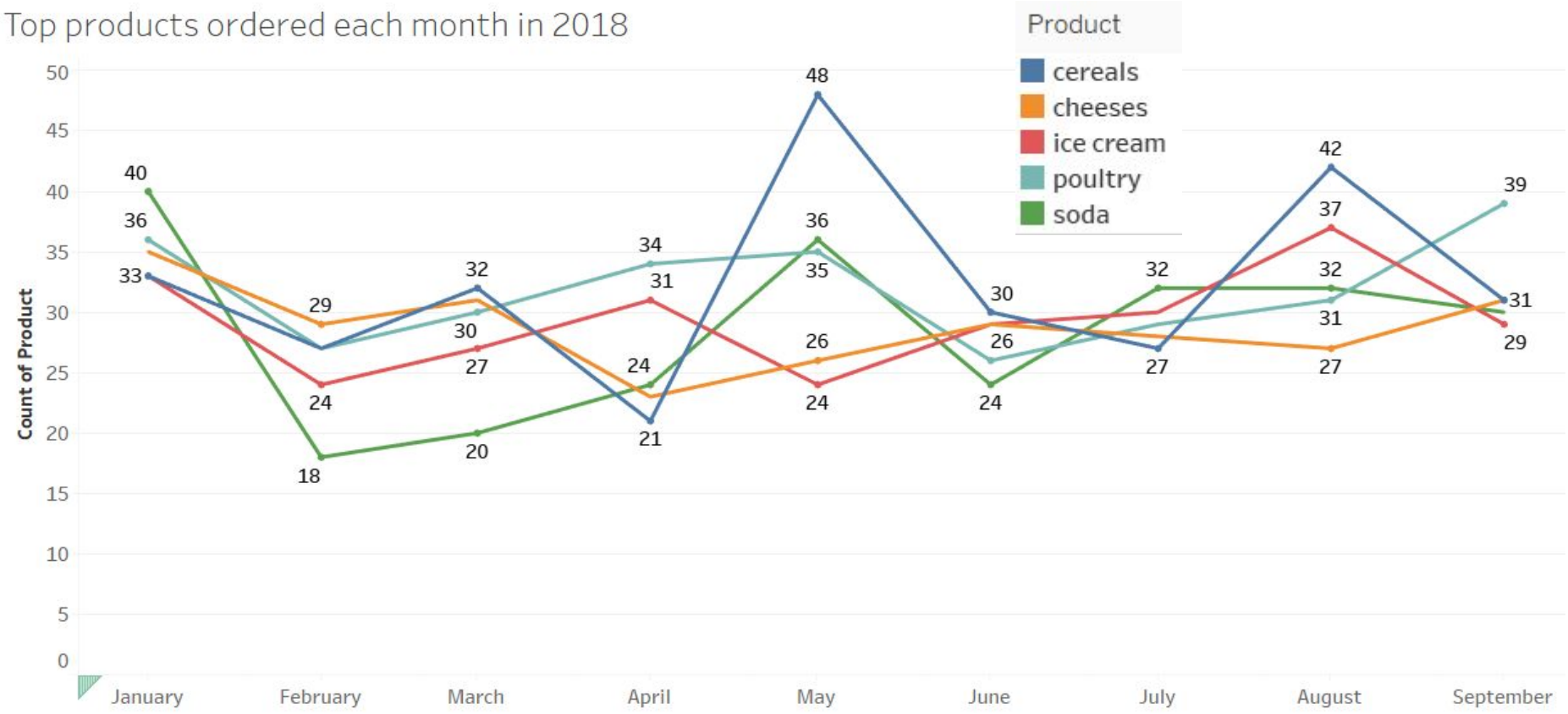


Figure 6: Trends of orders of top products in 2018

Trends of orders containing popular products in 2019

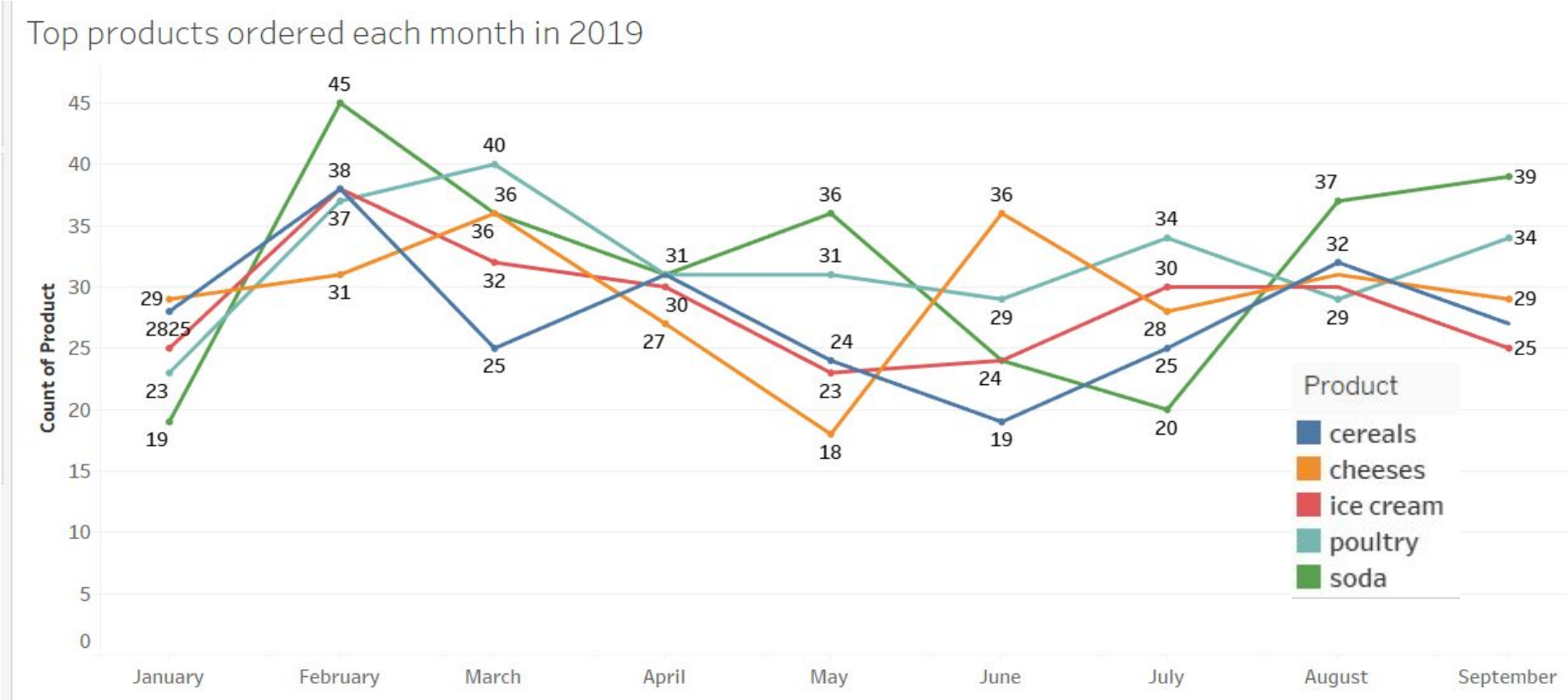


Figure 7: Trends of orders of top products in 2019

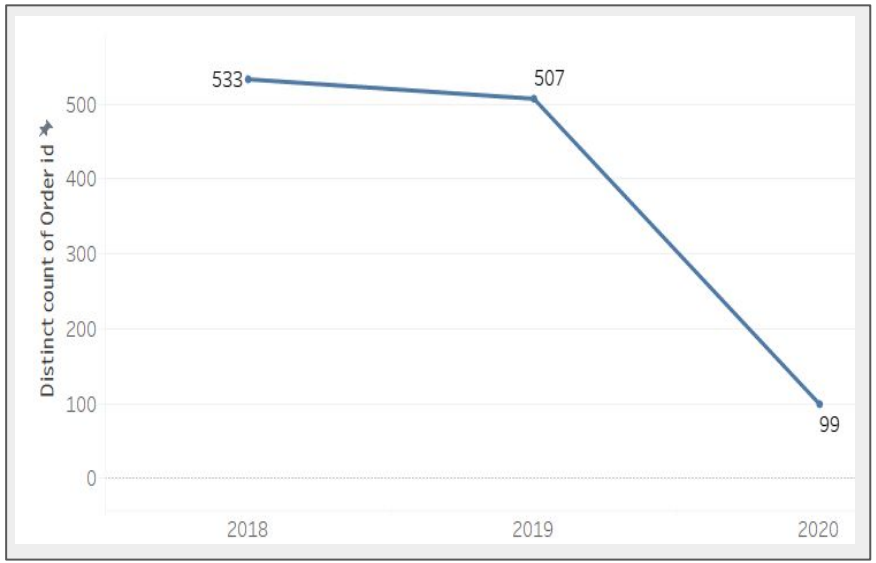
Insights and executive summary of top findings

- Cereals were ordered highest in May 2018 and February 2019
- Poultry items were ordered highest during September which is end of the year.
- This could probably be due to occasions and get togethers happening during the time of the year.
- Orders of these items were highest in May 2018 and February 2019.
- It could be probably be that since the store has started its sales in 2018, customers were in an experimenting phase.
- The items cereals, poultry, soda, cheese, ice cream were ordered least in January 2019.

Number of orders placed and products ordered per year

- Number of orders placed and total number of products ordered showed a slightly decreasing trend.
- The data has been recorded for 2 months in 2020 due to which the count for 2020 is low.

Order count per year



Number of products ordered per year

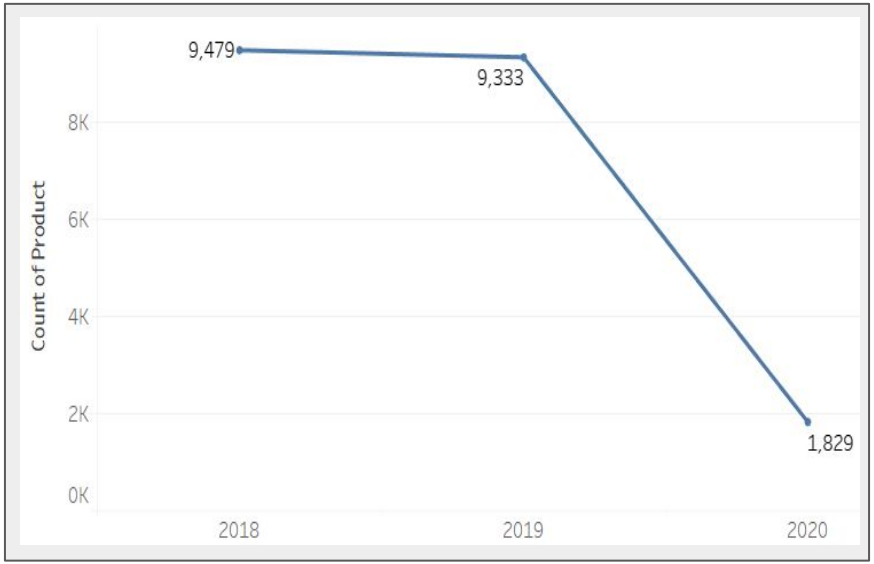


Figure 8: Yearly order count and product count trends

Number of orders placed ordered per quarter

- Around 175-180 orders were placed in each quarter in 2018.
- Order count was recorded less in the last quarter of 2019. This can be further investigated into.
- Note: Data exists till September of every year. Hence trends of 3 quarters are shown.

Figure 9: Quarterly order count trends

Number of products ordered per quarter

- Around 3100-3300 products were ordered in each quarter in 2018.
- Ordered product count was recorded less in the last quarter of 2019. This can be further investigated into.
- Note: Data exists till September of every year. Hence trends of 3 quarters are shown.

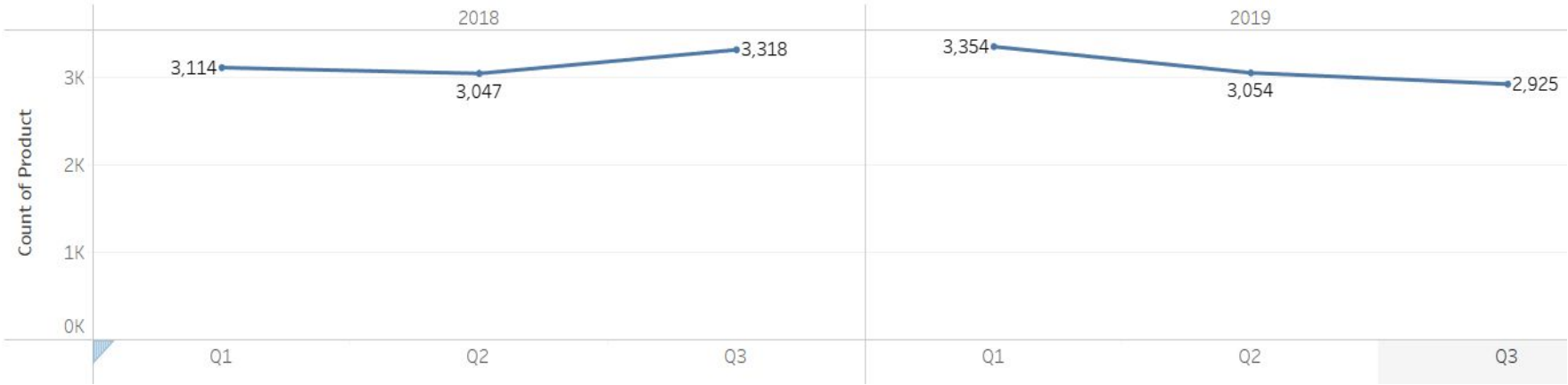


Figure 10: Quarterly product count trends

Number of orders placed per month

- Most of the orders were placed in the month of May in 2018 and 2019.
- Since the data only exists for 2 years there is no clear seasonality trend that can be deduced.

Figure 11: Monthly order count trends

Market Basket Analysis

Building association rules

Market Basket Analysis - Introduction

- Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.
- These patterns are formulated as association rules of the form if X then Y.
- For example
  - 60% of those who buy comprehensive motor insurance also buy health insurance.
  - 80% of those who buy books online also buy music online.
  - 50% of those who have high blood pressure and are overweight have high cholesterol.
- These rules are actionable in that they can be used to target customers for marketing, or for product placing, or more generally to inform decision making.
- Examples of areas in which association rules have been used include

  • Credit card transactions • Supermarket purchases • Telecommunication product purchases • Banking services • Insurance claims • Medical patient histories

Market Basket Analysis - Introduction

- Let S be the set of all possible purchases and let n be the number of transactions.
- Each transaction record is a subset of S.
- We consider rules of the form “(x1, x2, …, xj) implies (y1, y2, …, yk)” where x1, x2, …, y1, y2, … are elements of S.
- The collection (x1, x2, …, xj) is called an itemset.
- The support of the rule (x1, x2, …, xj) implies (y1, y2, …, yk) is defined as

$$\text{Supp}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{No. transactions containing x1, x2, … and y1, y2, …}}{n}$$

Market Basket Analysis - Introduction

- The confidence of the rule is

$$\text{Conf}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{Supp}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots))}{\text{Supp}(x1, x2, \ldots)}$$

- To consider a rule, we impose a minimum support, indicating a reasonable amount of data about the rule.
- The confidence measures how good a predictor the rule is.
- If we specify a minimum support s0 and a minimum confidence c0, then a strong rule is one which has Supp((x1, x2, …) implies (y1, y2, …)) > s0 and Conf((x1, x2, …) implies (y1, y2, …)) > c0.

Market Basket Analysis - Introduction

- High support or confidence does not always mean the rule is interesting.
- To measure the strength of a rule, we use an additional metric called ‘Lift’ or ‘improvement’ of the rule.

$$\text{Lift}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{Supp}(x1, x2, \ldots \text{ and } y1, y2, \ldots)}{\text{Supp}(x1, x2, \ldots)\ \text{Supp}(y1, y2, \ldots)}$$

- Lift > 1 indicates that the association between n (x1, x2, …) and (y1, y2, …) is due to more than just chance
- It indicates positive correlation between the events ‘purchased x1, x2,..’ and ‘purchased y1, y2, …’.
- Typically rules where the consequent items consist of a single item are the most useful.

The “A Priori” Algorithm

- Suppose there are a total of m items in S.
- The number of subsets of S is 2 power(m) , thus to check every transition record to see which sets it belongs to requires huge number of checks.
- Thus is computationally infeasible when m is even of moderate size. This is an instance of the “curse of dimensionality”.
- However, if we restrict ourselves to sets with support greater than s0 the search becomes feasible. We call these the frequent itemsets.
- This is because most sets have very small support, and because of the fact that for any y Supp(x1, x2, …, xk and y) is no greater than Supp(x1, x2, …, xk)
- Once all the sets with support greater than s0 have been found and their supports recorded, it is then a straightforward matter to calculate the confidence, lift and significance of all strong rules of the form “(x1, x2, …) implies (y1, y2, …)”, since all of these measures are calculated using the supports of various itemsets.

Relevance in this case

- The data set in the current scenario belongs to retail and fast moving consumer products like cereals, food items, household items etc.
- Forming association rules based on the products ordered together in the recorded transactions we can analyse the pairs of groups of frequently ordered items.
- These rules can help in placement of items in aisles to increase the count of products ordered.
- They can also help in framing combos
- They can also help in tagging promotions to the mostly ordered products in order to increase sales of multiple products.

KNIME workflow image

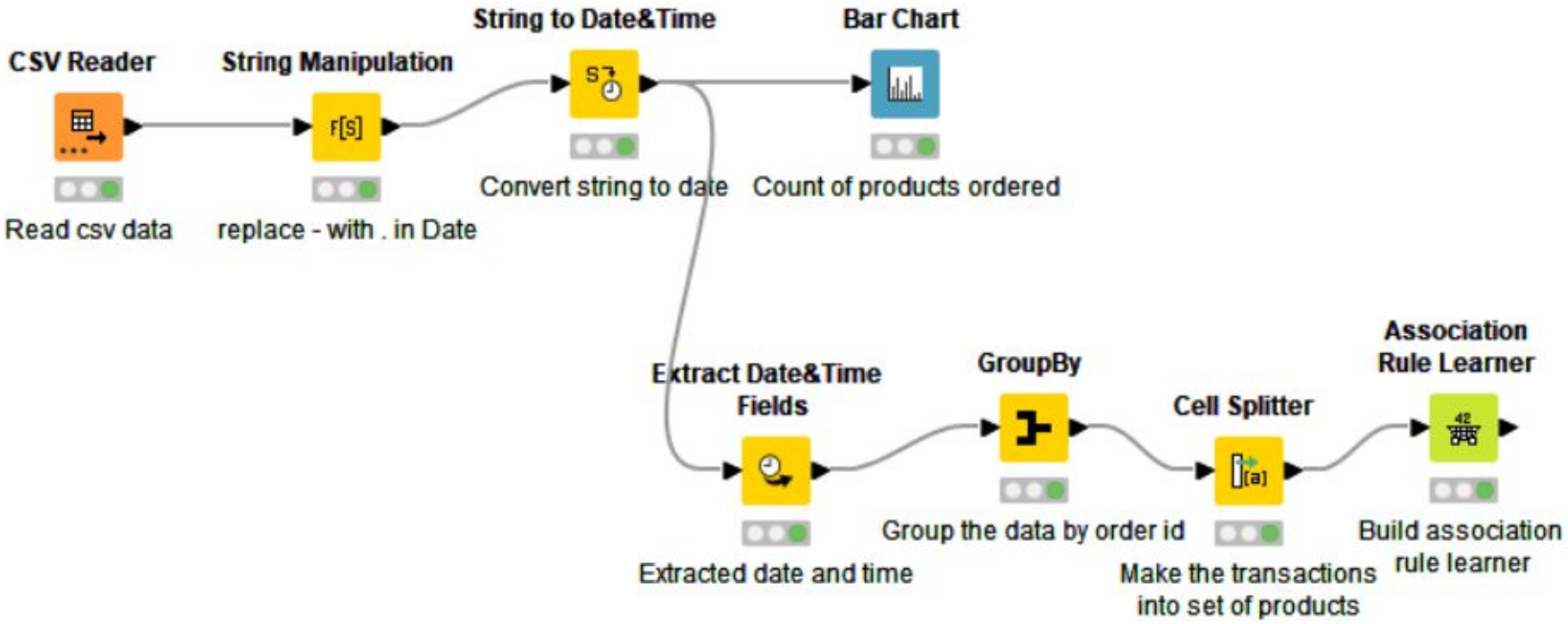


Figure 12: Market Basket Analysis - KNIME workflow

Steps performed in the KNIME workflow image

- Node1: The data has been read in csv format. The first row has been considered as column data.
- Glimpse of the data after reading in KNIME:

Table "default" - Rows: 20641 | Spec - Columns: 3 | Properties | Flow Variables

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

Table 1: Glimpse of the table read by KNIME

- Node2: The characters ‘-’ in Date column in above table have been replaced with ‘.’. This formatted string can be easily interpreted by next to convert this date in string format to date and time format.

Steps performed in the KNIME workflow image

- Node3: The date in string format has been converted into DateTime format.

Row ID	S Date	I Order_id	S Product	31 OrderD...
Row0	01-01-2018	1	yogurt	2018-01-01
Row1	01-01-2018	1	pork	2018-01-01
Row2	01-01-2018	1	sandwich bags	2018-01-01
Row3	01-01-2018	1	lunch meat	2018-01-01
Row4	01-01-2018	1	all- purpose	2018-01-01
Row5	01-01-2018	1	flour	2018-01-01
Row6	01-01-2018	1	soda	2018-01-01
Row7	01-01-2018	1	butter	2018-01-01

Table 2: Table with Date column extracted in Date format

- A bar chart has been drawn in KNIME to find out the number of times the product is included in an order.

Steps performed in the KNIME workflow image

Figure 13: Counts of products ordered - Bar chart in KNIME

Steps performed in the KNIME workflow image

- Extracted Date, Month, Year from Date column which might be helpful for further analysis.
- GroupBy: The data has been grouped by order id. The records contain order number and the list of products ordered in that order.
- Cell Splitter: The list of transactions above contain duplicates of products. We are interested in the occurrence of the product in the order not the count. Hence, removing the duplicate products in an order and making a set of items for the order.
- Association rule learner: The final rules are obtained by passing the above table into the Association rule learner node of KNIME.
- The support, confidence values applied for the above association rule learner are 0.05 and 0.05
- Support of 0.05 indicates that of all the transactions, there are 5% of the transactions that contain the item in the rule.
- Confidence of 0.05 indicates that of all the transactions that the base items, 5% of the items also contain the implied item.

Associations Identified

Associations Identified

Row ID	Support	Confide...	Lift	Consequent	implies	Items
rule1142	0.055	0.649	1.791	paper towels	<---	[eggs,ice cream,pasta]
rule1141	0.055	0.643	1.731	pasta	<---	[paper towels,eggs,ice cream]
rule138	0.051	0.674	1.726	cheeses	<---	[bagels,cereals,sandwich bags]
rule51	0.05	0.64	1.7	juice	<---	[yogurt,toilet paper,aluminum foil]
rule135	0.051	0.63	1.678	mixes	<---	[yogurt,poultry,aluminum foil]
rule137	0.051	0.611	1.66	sandwich bags	<---	[cheeses,bagels,cereals]
rule616	0.054	0.642	1.651	dinner rolls	<---	[spaghetti sauce,poultry,laundry detergent]
rule289	0.052	0.641	1.649	dinner rolls	<---	[spaghetti sauce,poultry,ice cream]
rule55	0.05	0.62	1.645	juice	<---	[yogurt,poultry,aluminum foil]
rule292	0.052	0.686	1.628	poultry	<---	[dinner rolls,spaghetti sauce,ice cream]
rule298	0.052	0.634	1.627	eggs	<---	[paper towels,dinner rolls,pasta]
rule299	0.052	0.602	1.621	pasta	<---	[paper towels,eggs,dinner rolls]
rule141	0.051	0.63	1.621	dinner rolls	<---	[spaghetti sauce,poultry,cereals]
rule1140	0.055	0.63	1.616	eggs	<---	[paper towels,ice cream,pasta]
rule59	0.05	0.613	1.616	coffee/tea	<---	[yogurt,cheeses,cereals]
rule293	0.052	0.628	1.614	dinner rolls	<---	[spaghetti sauce,poultry,juice]
rule284	0.052	0.628	1.61	eggs	<---	[dinner rolls,poultry,soda]
rule618	0.054	0.598	1.603	spaghetti sauce	<---	[dinner rolls,poultry,laundry detergent]
rule146	0.051	0.604	1.589	milk	<---	[poultry,laundry detergent,cereals]
rule291	0.052	0.59	1.581	spaghetti sauce	<---	[dinner rolls,poultry,ice cream]
rule294	0.052	0.584	1.566	spaghetti sauce	<---	[dinner rolls,poultry,juice]
rule1139	0.055	0.624	1.565	ice cream	<---	[paper towels,eggs,pasta]
rule300	0.052	0.567	1.565	paper towels	<---	[eggs,dinner rolls,pasta]
rule60	0.05	0.588	1.564	mixes	<---	[dishwashing liquid/detergent,poultry,laundry deterg...
rule139	0.051	0.617	1.558	cereals	<---	[cheeses,bagels,sandwich bags]
rule619	0.054	0.656	1.556	poultry	<---	[dinner rolls,spaghetti sauce,laundry detergent]
rule49	0.05	0.594	1.544	aluminum foil	<---	[yogurt,toilet paper,juice]
rule56	0.05	0.588	1.528	yogurt	<---	[cheeses,cereals,coffee/tea]
rule57	0.05	0.594	1.52	cheeses	<---	[yogurt,cereals,coffee/tea]
rule145	0.051	0.574	1.518	laundry detergent	<---	[poultry,milk,cereals]
rule144	0.051	0.637	1.512	poultry	<---	[dinner rolls,spaghetti sauce,cereals]
rule140	0.051	0.58	1.505	bagels	<---	[cheeses,cereals,sandwich bags]
rule297	0.052	0.584	1.502	dinner rolls	<---	[paper towels,eggs,pasta]

Table 2: Associations identified using Market Basket analysis

Understanding support and confidence

- With increasing number of products in the market, it becomes difficult to analyze all possible combinations of products ordered and their implications.
- Therefore we consider threshold values of support and confidence in order to filter the strong association rules.
- High support and confidence indicate that the association rules we are considering have not occurred by chance.
- Along with support and confidence we also consider another metric called ‘Lift’
- Lift also measures the strength of rules and the strength of probability that if item A exists in the order then item B also exists.
- Lift > 1 indicates positive correlation and that item A implies item B in the order.
- Since we set low support and confidence values we have obtained as many as 26,400 rules.
- Out of these rules we are interested in the rules with high lift values i.e., > 1.

Understanding support and confidence

- Let S be the set of all possible purchases and let n be the number of transactions.
- Each transaction record is a subset of S.
- We consider rules of the form “(x1, x2, …, xj) implies (y1, y2, …, yk)” where x1, x2, …, y1, y2, … are elements of S.
- The collection (x1, x2, …, xj) is called an itemset.
- The support of the rule (x1, x2, …, xj) implies (y1, y2, …, yk) is defined as

$$\text{Supp}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{No. transactions containing x1, x2, … and y1, y2, …}}{n}$$

Understanding support and confidence

- The confidence of the rule is

$$\text{Conf}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{Supp}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots))}{\text{Supp}(x1, x2, \ldots)}$$

- To consider a rule, we impose a minimum support, indicating a reasonable amount of data about the rule.
- The confidence measures how good a predictor the rule is.
- If we specify a minimum support s0 and a minimum confidence c0, then a strong rule is one which has Supp((x1, x2, …) implies (y1, y2, …)) > s0 and Conf((x1, x2, …) implies (y1, y2, …)) > c0.

Understanding support and confidence

- High support or confidence does not always mean the rule is interesting.
- To measure the strength of a rule, we use an additional metric called ‘Lift’ or ‘improvement’ of the rule.

$$\text{Lift}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{Supp}(x1, x2, \ldots \text{ and } y1, y2, \ldots)}{\text{Supp}(x1, x2, \ldots)\ \text{Supp}(y1, y2, \ldots)}$$

- Lift > 1 indicates that the association between n (x1, x2, …) and (y1, y2, …) is due to more than just chance
- It indicates positive correlation between the events ‘purchased x1, x2,..’ and ‘purchased y1, y2, …’.
- Typically rules where the consequent items consist of a single item are the most useful.

Interpreting the association rules

- Consider the rule eggs, icecream, pasta → paper towels
- Support of the rule is 0.055
- Confidence for the rule is 0.649
- Lift for the rule is 1.791

Interpretations:

- Support: Number of transactions containing eggs, icecream, pasta and paper towels / total number of transactions.
- A value of 0.055 indicates that around 63 out of 1140 orders contain the 4 items together.

Interpreting the association rules

- Confidence for the rule is 0.649
- Out of every 100 orders that contain eggs, ice cream and pasta together, around 65 of the orders contain paper towel also.
- It indicates the probability that the implied item is also present in the order given the actual supporting items.
- Lift of the association rule is 1.791. It indicates that the occurrence of paper towels along with eggs, ice cream and pasta is merely not by chance and can be considered as a thoughtful purchase by the customers.
- However, there are some rules that can arise which have high lift and confidence values but do not make more sense while interpreting.
- The rules might be implying a totally unrelated item purchased along with the supporting items.

Interpreting the association rules

- Of all the rules calculated by the algorithm, we pick the rules that have high lift values, desired support and confidence values.
- These rules can be categorized into three types
  - **Useful** rules are the ones we want, with high quality actionable information.
  - **Trivial** rules will already be known by anyone familiar with the business.
  - **Inexplicable** rules are those which have no apparent explanation and do not suggest a course of action. An example of the latter is the famous “Men who buy nappies on Thursdays also buy beer” rule. There is no automatic way of identifying trivial or inexplicable rules. In practice one needs

Interpreting the association rules

- Top 10 association rules are
  - paper towels<---[eggs, ice cream, pasta] → Useful
  - pasta<---[paper towels, eggs, ice cream] → Useful
  - cheeses<---[bagels, cereals, sandwich bags] → Trivial
  - juice<---[yogurt, toilet paper, aluminum foil] → Inexplicable
  - mixes<---[yogurt, poultry, aluminum foil] → Inexplicable
  - sandwich bags<---[cheeses, bagels, cereals] → Trivial
  - dinner rolls<---[spaghetti sauce, poultry, laundry detergent] → Trivial
  - dinner rolls<---[spaghetti sauce, poultry, ice cream] → Trivial
  - juice<---[yogurt, poultry, aluminum foil] → Inexplicable
  - poultry<---[dinner rolls, spaghetti sauce, ice cream] → Inexplicable

Suggestions of possible combos with lucrative offers

Recommendations

- Since poultry products are most ordered by the customers, more varieties in poultry products can be introduced which will boost the sales since there is already positive sign of sales for such products.
- Significant orders contain paper rolls and dinner rolls even though they are not related to other products in the order.
- This indicates that they can be placed in the place where more kitchen cookery related items and food items are placed.
- People often have bought food items like bagels, cereals, sandwich bags, cheeses together.
- Detergents like dishwashing liquids, laundry detergents can be placed together as people who have bought one in the past have also bought the other.
- Items like spaghetti sauce are bought along with poultry items. This indicates that people tend to purchase items that are useful for cooking meat and eggs when they are purchased.
- Some offers on ice-cream and spaghetti sauce can be introduced as they are included in orders.
- Top occurring combinations of products are
    - [eggs, ice cream, pasta, paper towels], [bagels, cereals, sandwich bags, cheeses], [spaghetti sauce, poultry, dinner rolls, ice cream], [yoghurt, aluminium foil, juice], [poultry, laundry detergent, dishwashing liquid, mixes]

Discount offers or combos based on association rules

- A small bottle of spaghetti sauce can be sold as free or a trial pack when any poultry item is purchased → This can improve the sales of poultry.
- A generic offer of buy one get one can be placed on paper towels and dinner rolls as they are bought frequently by customers.
- Combo of dish washing liquid and laundry detergent can be sold at a special combo price → This will motivate customers who buy one to also buy another.
- When bread type items like sandwich bags, bagels and sandwich loaves are purchased, either cheese, juice or cereals are purchased along with them. → Bread type items can be placed in between milk products and juice items.
- This promotes people who prefer such breakfast buy the items together.
- All combinations of offers on breakfast type items like yoghurt, eggs can be placed as customers tend to purchase them most often and also together.