Report on
Market Basket Analysis for E-commerce

TABLE OF CONTENTS

I. Introduction

1. Market Basket Analysis

Market basket analysis is a data mining technique that identifies customer purchasing habits and patterns to uncover associations between products frequently bought together. It involves analyzing transactional data from a retail or e-commerce environment to uncover associations between items bought in a single shopping session. It helps businesses understand consumer behaviour, optimize pricing strategies, and improve inventory management for increased profitability.

Major types of Market Basket Analysis

Association Rule Mining: Identifies frequent item sets and generates association rules that express the likelihood of one item being purchased with the purchase of another item.

Sequence Analysis: Identifies frequent item sequences and generates sequential association rules describing the likelihood of one item sequence being followed by another.

Cluster Analysis: Groups similar items or transactions into clusters or segments based on their attributes to identify customer segments with similar purchasing behaviours.

Objective of Market Basket Analysis

By identifying associations between products, retailers can strategically place related items near each other or offer personalized recommendations to increase sales and customer satisfaction. It helps improve inventory management, optimize product placement, and design effective marketing campaigns

2. Business questions

To uncover patterns in customer purchasing behaviour, we ask, "Which products tend to be bought together in a single transaction?", to identify frequently purchased product combinations, then suggest promotions and cross-selling strategies.

3. About the Dataset

This dataset is a retail e-commerce that sells groceries. This dataset includes 38765 entries and 3 columns. The dataset is for transactions from 2014-01-01 to 2015-12-30.

Overview of the dataset:

	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk
5	4941	14-02-2015	rolls/buns
6	4501	08-05-2015	other vegetables
7	3803	23-12-2015	pot plants
8	2762	20-03-2015	whole milk
9	4119	12-02-2015	tropical fruit

4. About this Project

As this project aims to identify the associations between itemsets, the Association Rule Mining approach is adopted.

There are several algorithms used in Market Basket Analysis. This project will use the Apriori Algorithm, one of the most widely used and well-known algorithms used in Market Basket Analysis. It helps to find frequent itemsets in transactions and identifies association rules between these items.

Apriori Algorithm

Some important metrics in Apriori Algorithm:

i. Support

Support is a measure of how frequently an item or itemset appears in the dataset.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

High support indicates that an item or itemset is common in the dataset, while low support indicates that it is rare.

ii. Confidence

Confidence is a measure of the strength of the association between two items.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

High confidence indicates that the presence of the first item is a strong predictor of the presence of the second item.

iii. Lift

Lift is a measure of the strength of the association between two items, taking into account the frequency of both items in the dataset. Lift is used to compare the strength of the association between two items to the expected strength of the association if the items were independent.

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

A lift value greater than 1 indicates that the association between two items is stronger than expected based on the frequency of the individual items. A lift value of less than 1 indicates that the association is weaker than expected and may be less reliable or less significant.

In this project, the minimum support threshold that an itemset must meet to be considered frequent is identified at 0.06% in a total number of transactions (0.06% of 14963, which is approx. 9).

Just because an item or itemset is common does not mean that there is a strong association between them. For example, the item "whole milk" may have a high support value, but that does not mean there is a strong association between coffee and any other particular item.

Therefore, we should also consider the lift metric when assessing the associations between items, as it takes into account the support of each item to measure the strength of the association between them.

lift{A,B} = 1: no relationship between A and B (A and B occur together only by chance).

lift{A,B} > 1: positive relationship between A and B (A and B occur together more often than random).

This project will set the minimum threshold for lift metric to 1.1

5. Tool Used

Jupyter Notebook: Using Python Programming language

6. Project Layout

The project will follow these steps:

1. Import and clean the dataset

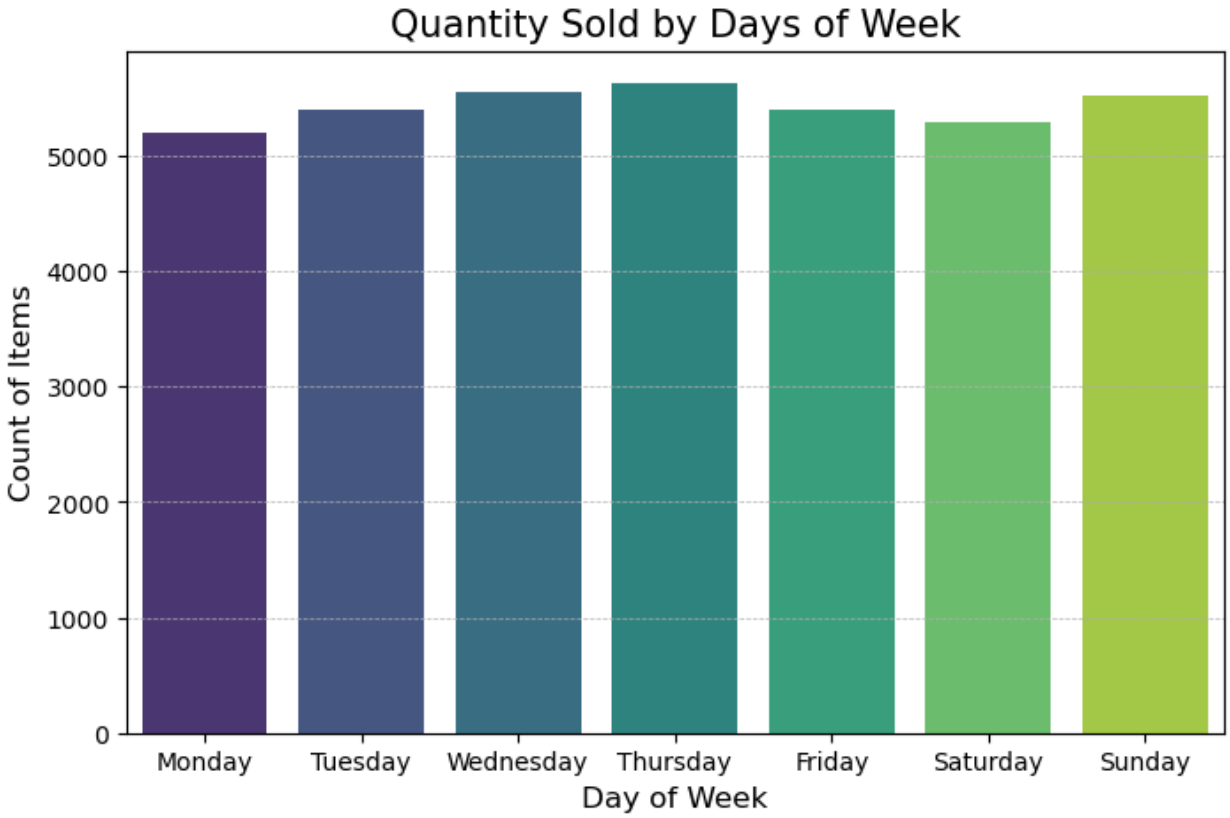
2. Conduct simple EDA and Visualizations

3. Apply Apriori Algorithm to implement Market Basket Analysis

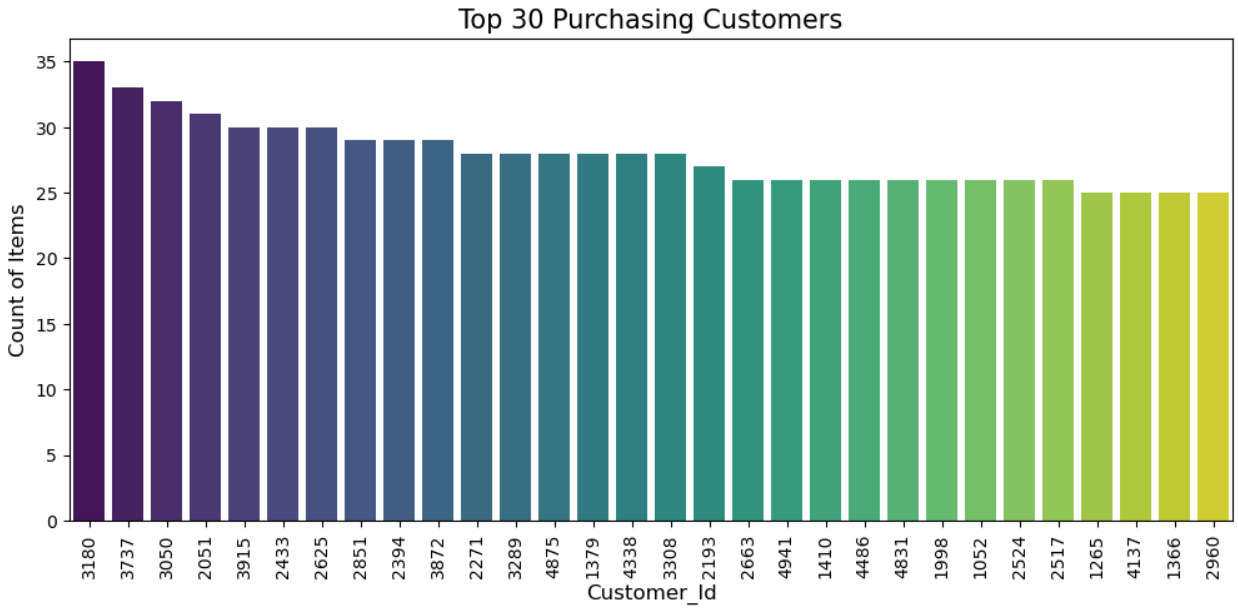
I

I. Data Exploration and Visualizations

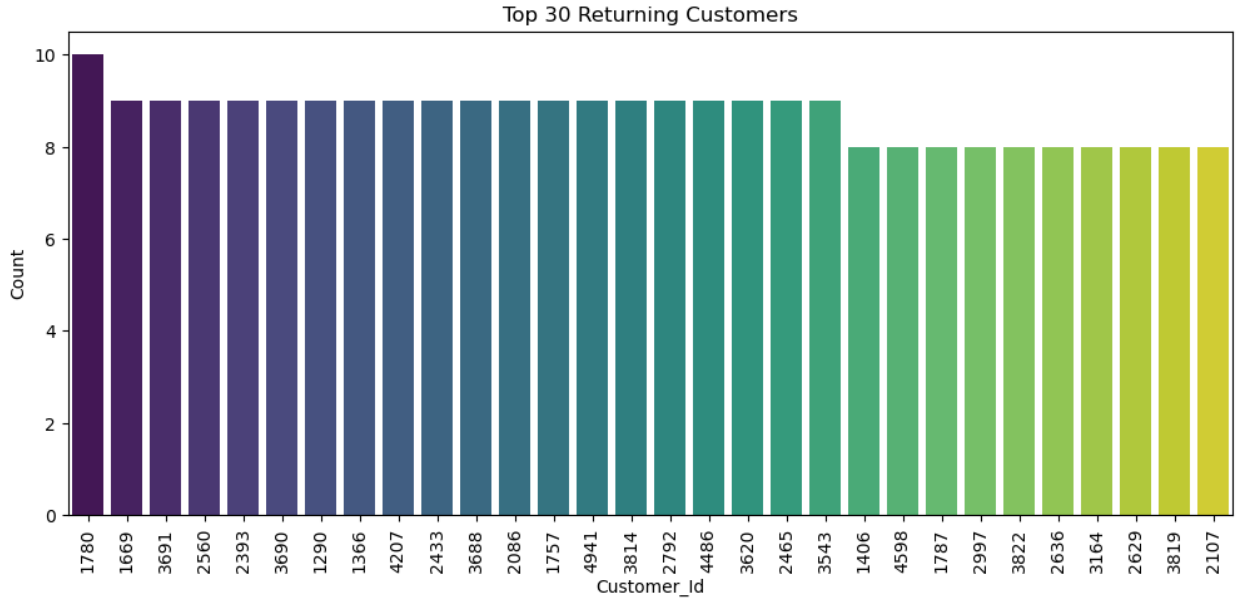
1. Quantity Sold by Days of Week



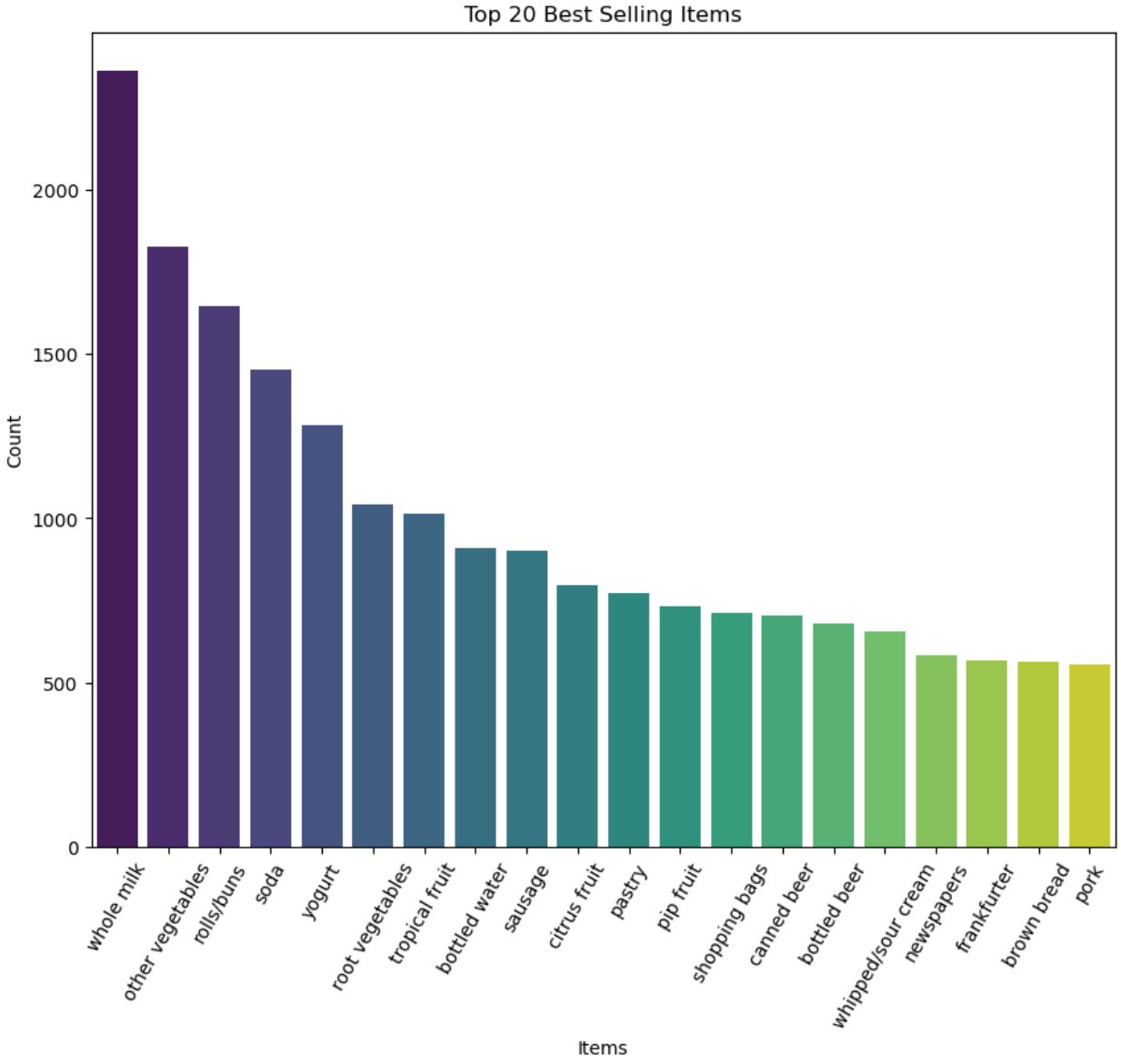
2. Top 30 Purchasing Customers



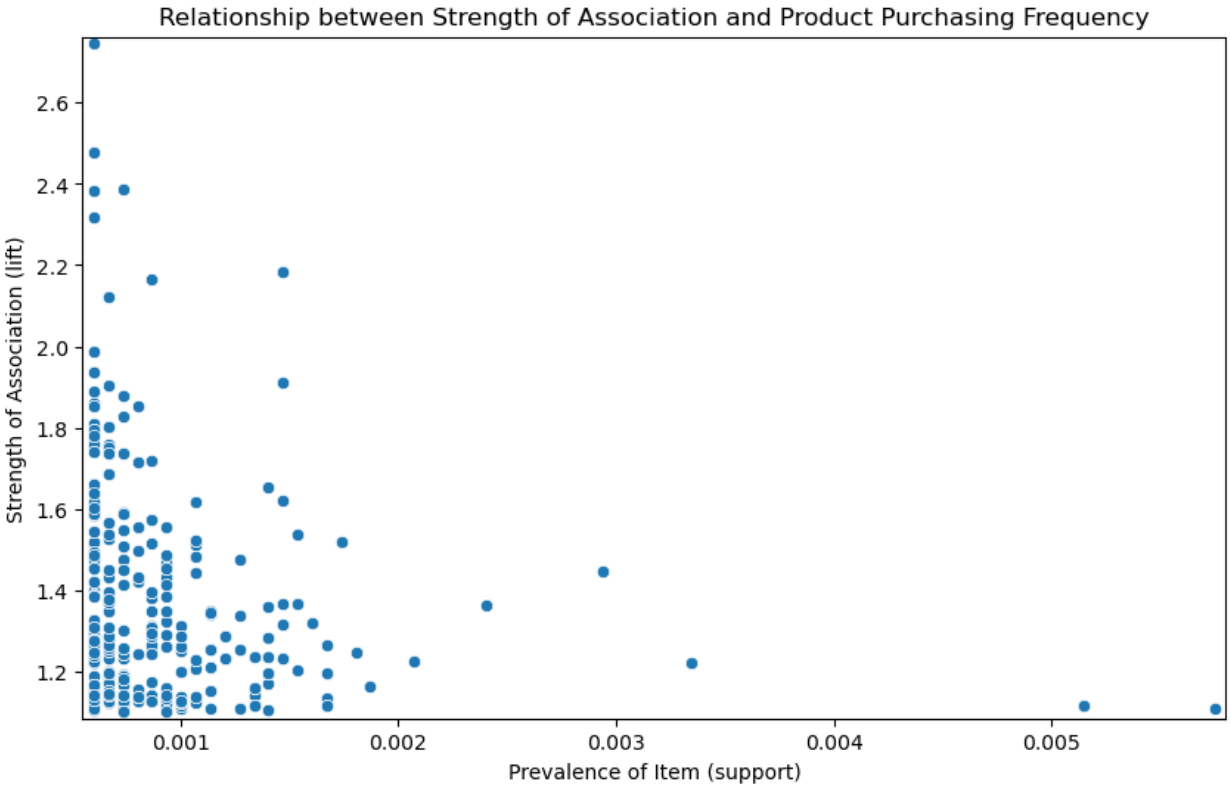
3. Top 30 Returning Customers



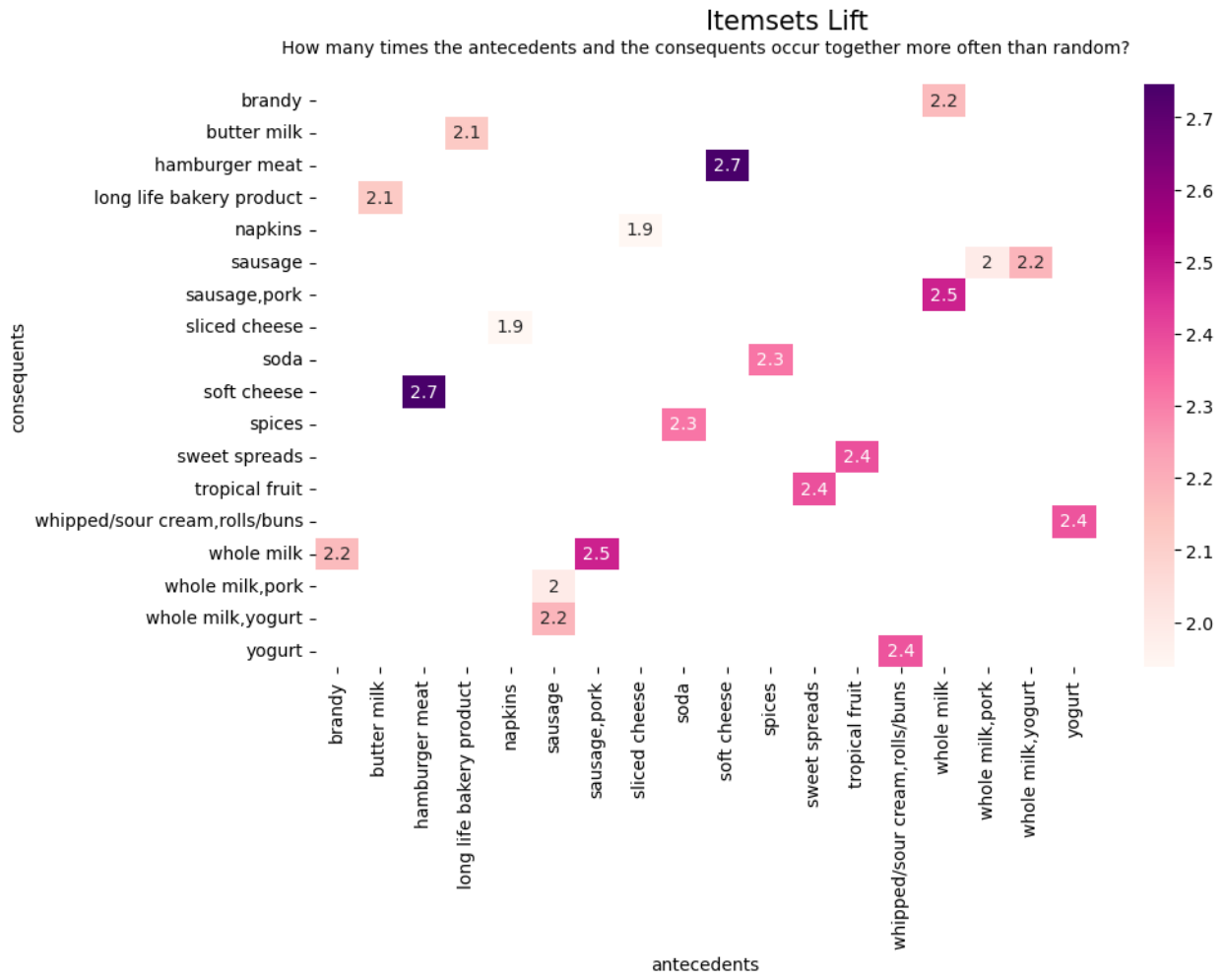
4. Top 20 Best Selling Items



5. Relationship between Strength of Association and Product Purchasing Frequency



6. Itemsets Lift



INSIGHTS

1. Although Thursday is the day with the highest sold items, the difference between the number of items sold on other days can be considered negligible, thus we cannot say Thursday is the busiest day.

2. Top purchasing customers (customers that have bought the most goods) are not top returning customers. Why? How do we retain top purchasing customers?

3. Whole milk is the highest purchased item accounting for approximately 16% of items bought.

4. 3 fruits (tropical fruit, citrus fruit and pip fruit) are some of the best-selling items but looking at the least sold items, we can see a particular fruit, frozen fruit which has barely made sales, what could be the reason for this?

5. Something similar is also seen with root vegetables and other vegetables are one of the best-selling items but specialty vegetable is one of the least-sold items.

MARKET BASKET ANALYSIS

1. There are 520 association rules found with min_support = 0.0006 and min_lift = 1.1.

2. The itemset, sausage and yoghurt, have the highest support value indicating they are common in the dataset. They have high support values individually and together.

3. The scatterplot reveals that less common itemsets (those with lower support/occurrence rates) tend to have higher lift values. This means when those less frequent item pairs or groups appear together in purchases, they are more likely to be bought together intentionally rather than by chance. So less common itemsets can still have strong associations, even though they occur less frequently overall.

4. Hamburger meat and soft cheese have the highest lift value (2.7) indicating they are most frequently bought together, despite being a small percentage (2% and 1%) of total itemsets.

5. Whole milk (the most purchased item) has strong associations with sausage pork (lift 2.4) and brandy (lift 2.2).

RECOMMENDATIONS

1. Strategically placing items with high lift values near each other on shelves could increase joint purchases e.g. (hamburger meat and soft cheese, whole milk, sausage pork and brandy, and other strong associations)

2. Investigate why top purchasing customers are not returning customers

3. Investigate what makes frozen fruit different from the other types of fruit sold and maybe it should not be sold anymore

4. Investigate what makes specialty vegetables different from the other types of vegetables sold and maybe they should not be sold anymore