# MARKET BASKET ANALYSIS

**Association Rule**
**Data Mining**

Prepared By
**Evan Saju Mathew**

# Table of Content

# Introduction

## Overview

The Instacart Market Basket Analysis is a comprehensive study aimed at uncovering patterns, trends, and insights into customer purchasing behaviors using data from one of the leading online grocery platforms, Instacart. This project leverages transactional data to identify customer preferences, popular products, and opportunities for enhancing user experiences through data-driven strategies. Instacart, a platform enabling customers to order groceries online, provides datasets containing millions of anonymized transactions, product details, and customer ordering habits. Analyzing this data can yield actionable insights that benefit both the business and its customers.

## Objective

### Customer Behavior

- Examine order patterns across days and hours.
- Explore reorder tendencies and customer loyalty to products.

### Product & Order Insight

- Identify top-purchased and reordered products.
- Analyze popular aisles, departments, and order frequencies

### Basket Analysis & Recommendation

- Uncover frequently co-purchased products.
- Recommend complementary products to increase basket value.

# Data Overview

## Aisles

- This dataset provides information on the aisles such as aisle ID and aisle names, through which the products were organized.
- Shape : 134 Rows x 2 Columns
- Column Description:

| | |
|---|---|
| **aisle_id** | Labels the ID of the aisle |
| **aisle** | Mentions the aisle name in the retail stores |

## Departments

- This dataset provides information on the departments such as department names and department Id.
- Shape : 21 Rows x 2 Columns
- Column Description:

| | |
|---|---|
| **department_id** | Labels the ID of the departments |
| **department** | Mentions the department name in the retail stores |

## Orders Product Prior

- This dataset gives information on the orders, products, and reordered products
- Shape : 20526345 Rows x 4 Columns
- Column Description:

| | |
|---|---|
| **order_id** | Labels the ID of the order made by customer |
| **product_id** | Labels the ID of the products purchased by customers |
| **add_to_cart_order** | Sequence of the order placed in the cart |
| **reordered** | Denotes whether the products are reordered or not |

## Orders Product Train

- This dataset is same as order_products_prior and it is a trained dataset.
- Shape : 1384617 Rows x 4 Columns
- Column Description:

| | |
|---|---|
| **order_id** | Labels the ID of the order made by customer |
| **product_id** | Labels the ID of the products purchased by customers |
| **add_to_cart_order** | Sequence of the order placed in the cart |
| **reordered** | Denotes whether the products are reordered or not |

# Data Overview

## Orders

- This dataset has information about the customer orders like order ID, order number, week day of the order, hour of the order, user ID and days since prior order.
- Shape : 3421083 Rows x 7 Columns
- Column Description:

| | |
|---|---|
| **order_id** | Labels the ID of the order made by customers |
| **user_id** | Labels the ID of the users who made the purchase |
| **eval_set** | Categorizes the data into prior or test data |
| **order_number** | Denotes the order number made by the customer |
| **order_dow** | Denotes the day of the week, the order made by the customer |
| **order_hour_of_day** | Denotes the hour of the day, the order made by the customer |
| **days_since_prior_order** | Denotes the number of days since last order |

## Products

- This dataset gives information on the products such as product name, product ID, aisle and departments, which were sold to the customer
- Shape : 49688 Rows x 4 Columns
- Column Description:

| | |
|---|---|
| **product_id** | Labels the ID of the products purchased by customers |
| **product_name** | Denotes the product name purchased by the customer |
| **aisle_id** | Labels the ID of the aisles |
| **department_id** | Labels the ID of the departments |

# Data Cleaning

missing_values.ipynb

```
// Loop through each DataFrame to display missing values
file_names = ['aisles', 'departments', 'odpp', 'odtr', 'orders', 'products']
data = [aisles, departments, odpp, odtr, orders, products]

for i in range(len(data)):
    print(f"Missing Values Summary for: {file_names[i]}")
    missing_values = data[i].isna().sum()
    print(missing_values)

    # Optionally, show the total number of missing values
    total_missing = missing_values.sum()
    print(f"\nTotal missing values in {file_names[i]}: {total_missing}")
    print("=" * 50)
```

| | |
|---|---|
| aisles<br>**0** | departments<br>**0** |
| odpp<br>**0** | odtr<br>**206209** |
| orders<br>**206209** | products<br>**0** |

There were no null or empty values for the variables like aisle, departments, Order_product_prior, order_product_train and products datasets.

**Orders** dataset has some null values in days since prior order variable and only 5% of the values were found to be missing and this has been rejected since the count is very low to be a significant issue.

# Data Cleaning

## Filling Those NaN Values with ‘0’

If we replace missing values with the mean (e.g., say the mean is 15 days), we are wrongly assuming that the customer waited 15 days before their first order’s which doesn’t make sense because they had no prior order at all!

Using ‘0’ makes sense because:
- It correctly represents first-time customers who don’t have a previous order.
- It keeps the meaning accurate (saying "0 days since the last order" means there was no previous order).
- It avoids introducing incorrect assumptions into the data.

handling_missing_values.ipynb

```
#Dealing with missing values
#There are missing values in the 'days_since_prior_order' column in the 'orders' DataFrame

orders['days_since_prior_order'] = orders['days_since_prior_order'].fillna(0)
orders.isna().sum()
```

snappify.com

## Merging Data-Frame’s

handling_missing_values.ipynb

```
# Merge the DataFrames

merge_orders_products = odpp.merge(products, on='product_id', how='left')
product_details = products.merge(aisles, on='aisle_id', how='left').merge(departments, on='department_id', how='left')
order_details = orders.merge(merge_orders_products, on='order_id', how='left')
```

# Data Analysis


Customer Behavior

Product Insight

Department & Aisle Analysis

Basket Analysis

Statistics Analysis

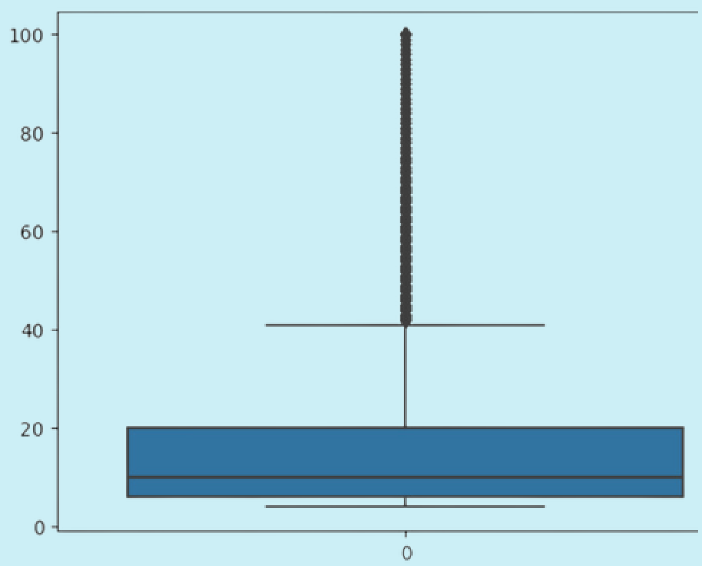
# Statistics Analysis

## Days Since Prior Order

| | |
|---|---|
| count | 3421083.00 |
| mean | 10.44 |
| std | 9.31 |
| min | 0.00 |
| 25% | 4.00 |
| 50% | 7.00 |
| 75% | 15.00 |
| max | 30.00 |

- According to **mean**, Customer reorders every 11 days.
- According to **median**, 50% of customers reorders every 7 days
- According to **75th percentile**, 75% of customers reorders every 15 days
- So only 25% customers reorders with gap greater than 15 days.

## Average Orders By Customers




- According to **mean**, average of times users has ordered 16 times.
- According to **median**, 50% of Users has kept 10 orders.
- According to 75% percentile, 75% of users has kept 20 orders.
- Maximum orders done by user bounded to 100.
- Box plot and **IQR** suggests that upper whisker = 41 and all the users who has done orders more than 41 are **outliers**.

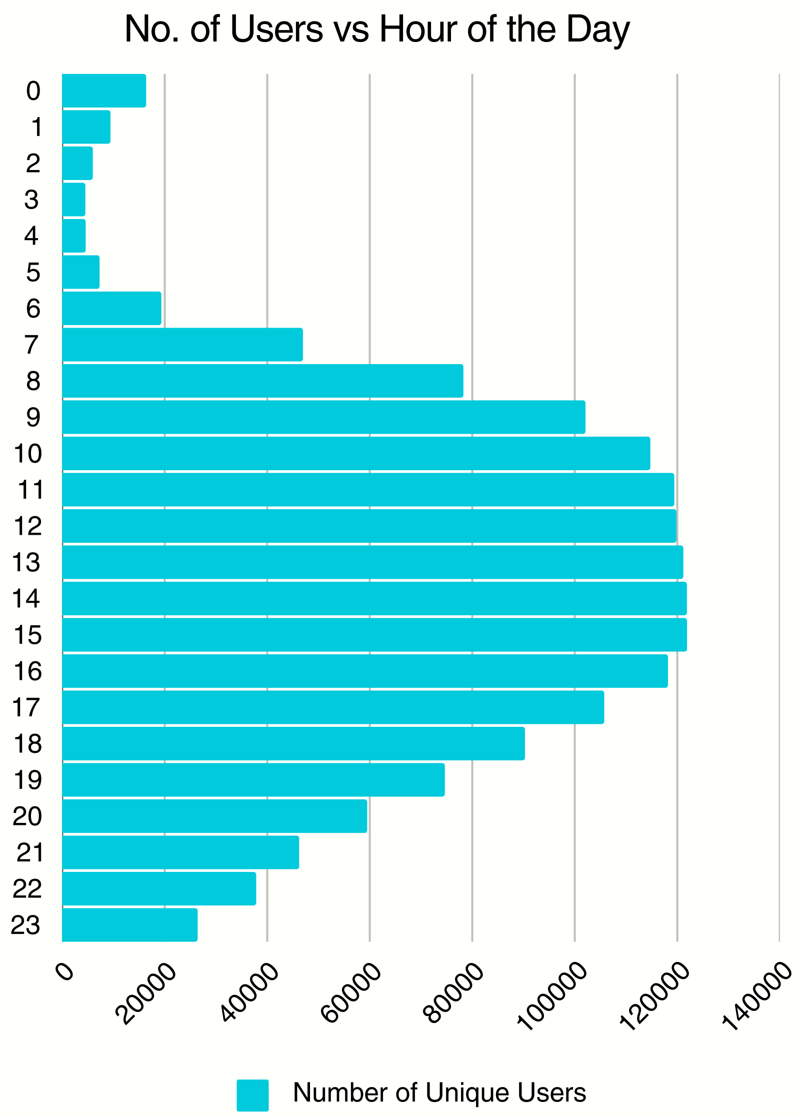
# Customer Rush

**Rush Hours**

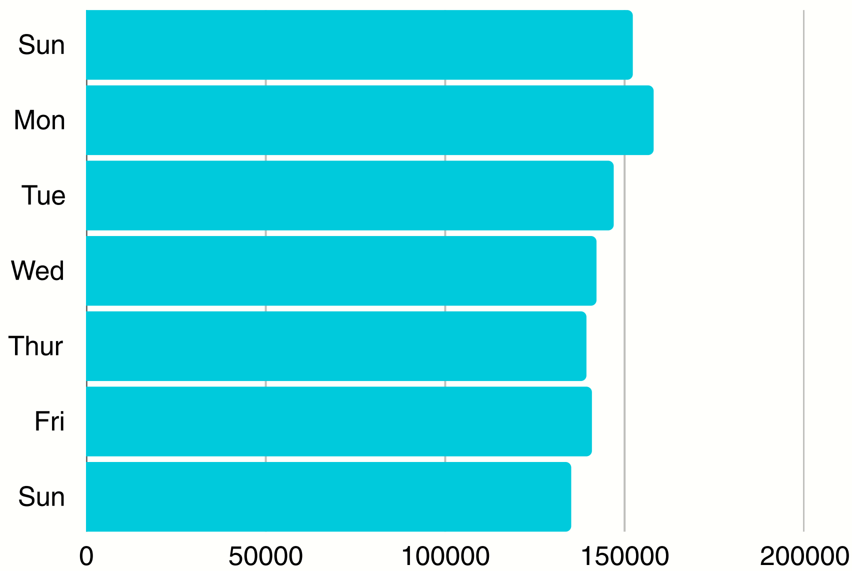
9 AM - 7 PM

**Rush Days**

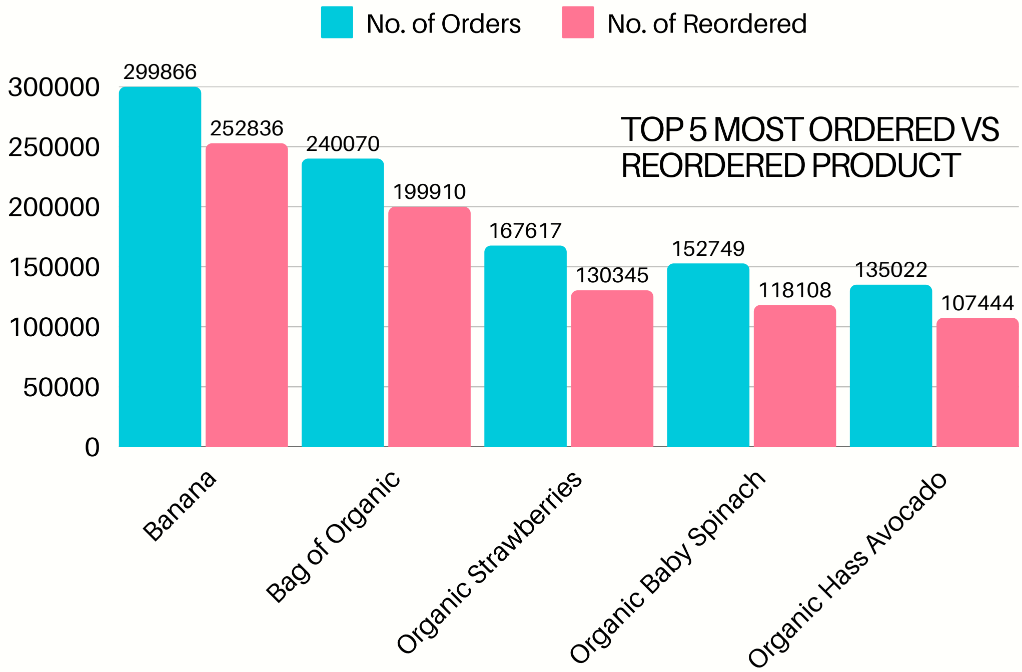
Sun - Mon




- **Rush Hours:**
  - The busiest hours for customers placing orders are typically between 9 AM and 7 PM.
  - The highest activity is usually observed around 10 AM to 3 PM, which suggests that users prefer placing their orders in the late morning and early afternoon
- **Rush Day's of Week:**
  - Sunday and Monday see the highest number of unique customers placing orders.
  - This suggests that users prefer restocking groceries and essentials at the beginning of the week.



# Top Products




---

## TOP 5 MOST AISLE WHICH HAVE NO. OF UNIQUE PRODUCTS

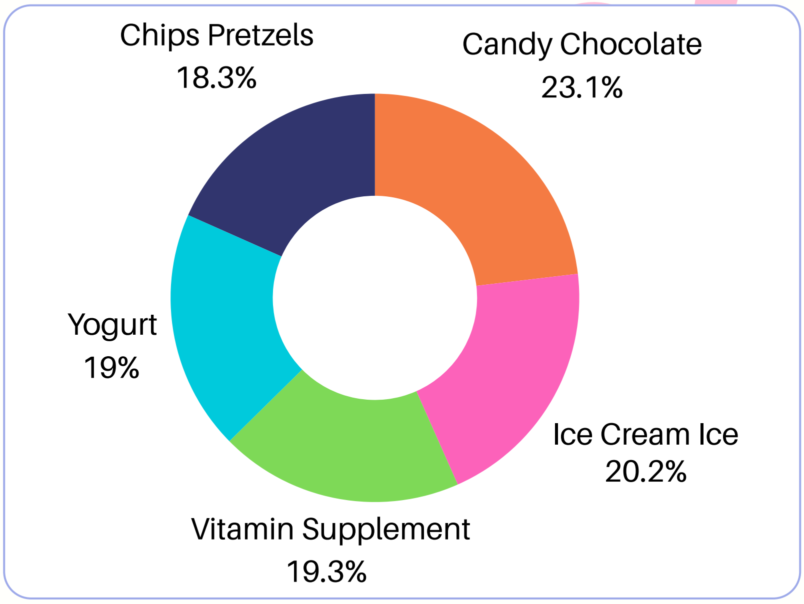
**1246**
Candy Chocolate

**1091**
Ice Cream Ice

**1038**
Vitamin Supplement

**1026**
Yogurt

**989**
Yogurt



# Aisle Insight

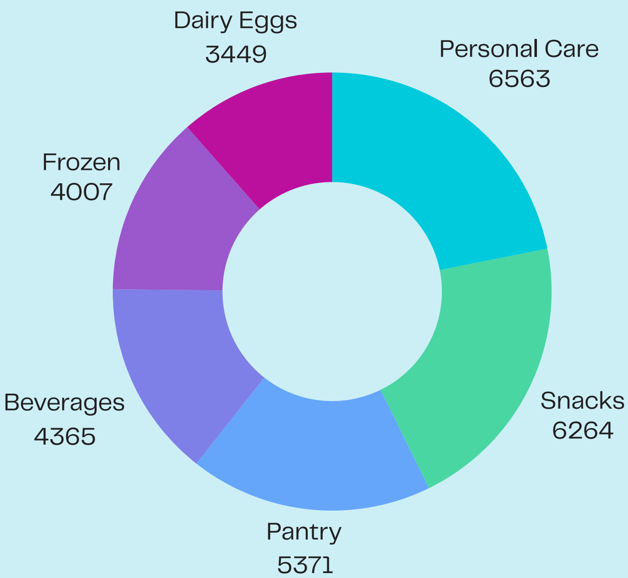
## Fresh vs Packaged Products in Aisles


## Top Most Aisle

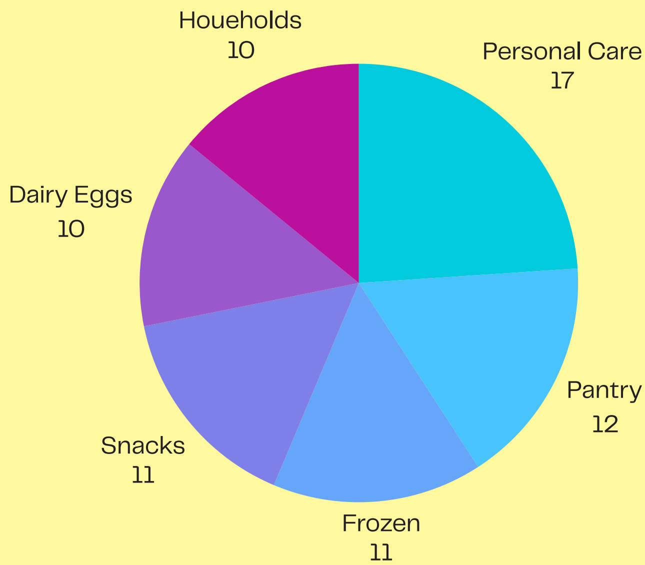
| | |
|---|---|
| Fresh Fruits | 2305892 |
| Fresh Vegetable | 2161455 |
| Packaged Fruits & Veg. | 1116489 |
| Yogurt | 918997 |
| Packaged Cheese | 619974 |
| Milk | 564516 |
| Sparkling Water | 533002 |
| Chips Pretzels | 457470 |
| Soy Lactose-Free | 403880 |

## Top Most Department


## Departments of Most Products




## Department of Most No. of Aisle’s



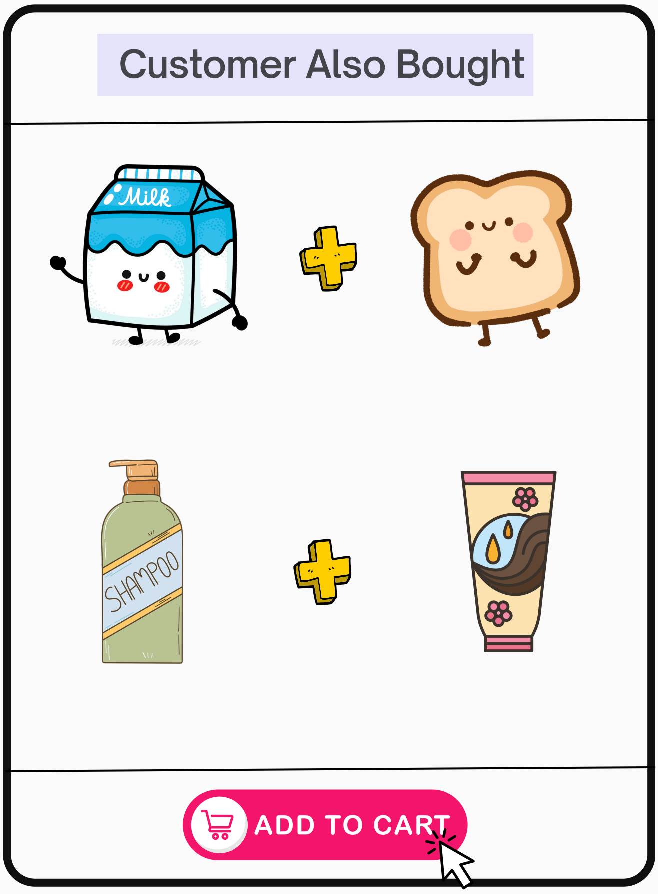
# Recommendation Products

## Product Recommendation Based on Association Rule Mining

Using **Association Rule Mining**, we can generate product recommendations based on customers purchase patterns. By analyzing frequent item-sets and strong association rules, we can suggest products that are often bought together.

In this analysis, we implemented the **Apriori Algorithm**, which finds frequent item-sets based on support, confidence, and lift:

- **Support**: The proportion of transactions containing a particular item or item-set.
- **Confidence**: The likelihood that a customer buying item A will also buy item B.
- **Lift**: Measures how much more likely item B is bought when item A is purchased compared to random chance.




## Key Findings

### High Confidence

High-confidence rules suggest strong product associations, meaning customers frequently buy certain items together.

### Lift Metrics

The lift metric confirms the significance of relationship-higher values indicate strong dependencies.

### Identifying Product Patterns

Identifying these patterns allows businesses to optimize cross-selling strategies, recommend complementary products, and improve store layout.

## Recommendation Strategy

### Cross-Selling Opportunities

- If a customer buys Product A, recommend Product B based on high confidence and lift values.
- Example: If a customer buys Organic Bananas, they are likely to purchase Almond Milk as well.

### Bundling & Promotions

- Retailers can create product bundles based on frequent item associations to drive sales.
- Example: Whole Wheat Bread + Peanut Butter + Organic Jam as a breakfast combo.

# Association Rule Mining(Steps)

## Data Preprocessing

Extracted order and product data.

Filtered products with support above the minimum threshold.

Removed orders with less than two items.

## Item Frequency & Support Calculation

Computed frequency (freqA, freqB) and support (supportA, supportB) for individual products.

## Generating Item Pairs

Created item pairs from transactions using combinations of frequently purchased products.

## Computing Pair Frequency & Support

Calculated how often product pairs appeared together (freqAB) and their support (supportAB).

## Association Rule Metrics Calculation

Confidence: Measures how likely a customer who buys Item A will also buy Item B (confidenceAtoB, confidenceBtoA).

Lift: Indicates the strength of the association between two products. A lift value > 1 suggests a strong correlation.

# Results

| Item A | Item B | freqAB | support AB (%) | freqA | support A (%) | freqB | support B (%) | confidence AtoB | confidence BtoA | Lift |
|---|---|---|---|---|---|---|---|---|---|---|
| Organic Strawberry Chia Lowfat 2% Cottage Cheese | Organic Cottage Cheese Blueberry Acai Chia | 306 | 1.02 | 1163 | 3.86 | 839 | 2.78 | 26.31% | 36.47% | 9.45 |
| Grain Free Chicken Formula Cat Food | Grain Free Turkey Formula Cat Food | 318 | 1.06 | 1809 | 6 | 879 | 2.92 | 17.58% | 36.18% | 6.03 |
| Organic Fruit Yogurt Smoothie Mixed Berry | Apple Blueberry Fruit Yogurt Smoothie | 349 | 1.16 | 1518 | 5.04 | 1249 | 4.14 | 22.99% | 27.94% | 5.55 |
| Nonfat Strawberry With Fruit On The Bottom Greek Yogurt | 0% Greek, Blueberry on the Bottom Yogurt | 409 | 1.36 | 1666 | 5.53 | 1391 | 4.62 | 24.55% | 29.40% | 5.32 |
| Organic Grapefruit Ginger Sparkling Yerba Mate | Cranberry Pomegranate Sparkling Yerba Mate | 351 | 1.16 | 1731 | 5.74 | 1149 | 3.81 | 20.28% | 30.55% | 5.32 |
| Baby Food Pouch - Roasted Carrot Spinach & Beans | Baby Food Pouch - Butternut Squash, Carrot & Chickpeas | 332 | 1.1 | 1503 | 4.99 | 1290 | 4.28 | 22.09% | 25.74% | 5.16 |
| Unsweetened Whole Milk Mixed Berry Greek Yogurt | Unsweetened Whole Milk Blueberry Greek Yogurt | 438 | 1.45 | 1622 | 5.38 | 1621 | 5.38 | 27.00% | 27.02% | 5.02 |
| Uncured Cracked Pepper Beef | Chipotle Beef & Pork Realstick | 410 | 1.36 | 1839 | 6.1 | 1370 | 4.55 | 22.29% | 29.93% | 4.9 |

## Products frequently bought together include:

- **Organic Strawberry Chia Cottage Cheese** + **Organic Cottage Cheese Blueberry Acai Chia**
  - (Lift: 9.45) -> Strongest association.
- **Grain-Free Chicken Formula Cat Food** + **Grain-Free Turkey Formula Cat Food**
  - (Lift: 6.02 -> High correlation among pet owners.
- **Organic Fruit Yogurt Smoothie Mixed Berry** + **Apple Blueberry Fruit Yogurt Smoothie**)
  - (Lift: 5.55) -> Yogurt lovers tend to buy both.
- **Unsweetened Whole Milk Mixed Berry Greek Yogurt** + **Unsweetened Whole Milk Blueberry Greek Yogurt**
  - Strong preference for similar yogurt flavors.

## Business Implications:

- These rules can enhance recommendation systems, cross-selling strategies, and inventory planning.
- High lift values indicate strong co-purchasing trends, suggesting bundling opportunities in marketing campaigns.

# Conclusion

## Summary

The Instacart Market Basket Analysis revealed key insights into customer shopping behavior, popular products, and purchase patterns. Peak shopping hours occur between 9 AM and 7 PM, with Sundays and Mondays being the busiest days. Fresh produce and dairy dominate sales, contributing 71.1% of total orders. Association Rule Analysis identified frequently bought-together products, which can enhance personalized recommendations and cross-selling strategies. These insights help optimize inventory management, targeted promotions, and customer experience, driving business growth through data-driven decisions.

---

## Reference


- https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/
- https://www.lexjansen.com/sesug/2019/SESUG2019_Paper-252_Final_PDF.pdf
- https://www.kaggle.com/code/datatheque/association-rules-mining-market-basket-analysis/notebook
- https://medium.com/towards-data-science/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce