جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2 (Task 2)

Instacart Market Basket Analysis Dataset

Name	ID
Joud Ahmad Al-huthaly	444002970
Nehal Hamed Al-zahrani	444001073

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS UMM
AL-QURA UNIVERSITY

Table of content

Link to Google collaboration :https://colab.research.google.com/drive/1YQgGgKckprF4LIsrvpanpRYJ-4NdjHtn?usp=sharing

Link to dataset : https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Introduction

This competition involves a relational dataset describing customer orders over time, with the goal of predicting which products will appear in a user's next order. The database contains over 3 million grocery orders from more than 200,000 Instacart users, providing each user with between 4 to 100 of their orders, including the sequence of products purchased in each order. The dataset also includes information about the week and hour of the day when the order was placed, as well as a relative measure of time between orders.

explanation of the important files and their columns:

order_products__train.csv

- **order_id:** A unique identifier for each order.
- **product_id:** A unique identifier for each product.
- **add_to_cart_order:** The order in which items were added to the cart, providing insights into shopping behavior.
- **reordered:** A binary value indicating whether the product was reordered (1) or not (0), helping understand customer preferences.

Department

- **department_id:** A unique identifier for each department.
- **department:** Describes the type of products in the department (e.g., "frozen," "other," "bakery"), useful for analyzing sales trends by department.

aisle

- **aisle_id:** A unique identifier for each aisle.
- **aisle:** A description of the aisle's contents, including product categories like "prepared soups salads," "specialty cheeses," and "energy granola bars," aiding in understanding product organization within the store.

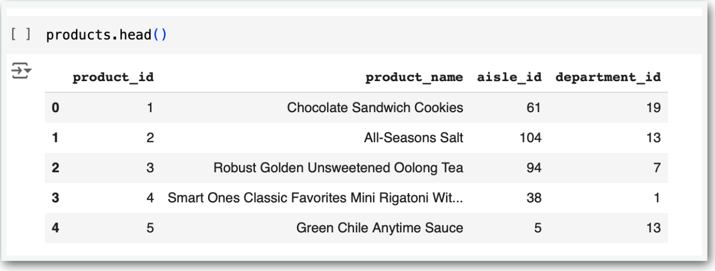
Orders

- **order_id:** A unique identifier for each order (same as in order_products__train.csv).
- **user_id:** A unique identifier for each user, helping track individual customer behaviors.
- **eval_set:** Indicates the set to which the order belongs (e.g., "prior" for previous orders).
- **order_number:** The sequential number of the order for each user, helping track order history.
- **order_dow:** Day of the week when the order was placed (0=Sunday, 1=Monday, etc.), revealing patterns in shopping behavior based on the day.
- **order_hour_of_day:** The hour of the day when the order was placed, providing insights into shopping times.

Products

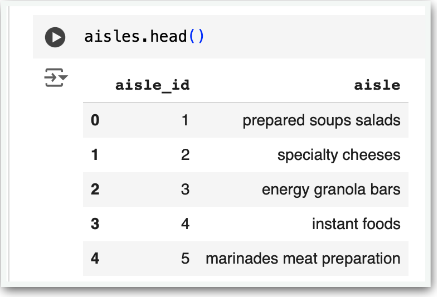
- **product_id:** A unique identifier for each product.
- **product_name:** The name of the product.
- **aisle_id:** The identifier for the aisle in which the product is located.
- **department_id:** The identifier for the department to which the product belongs.

The dataset



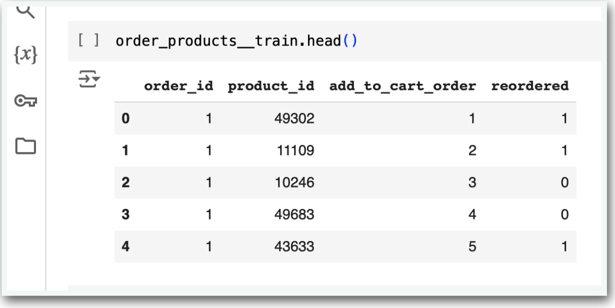
```
products.head()
```

	product_id	product_name	aisle_id	department_id
0	1	Chocolate Sandwich Cookies	61	19
1	2	All-Seasons Salt	104	13
2	3	Robust Golden Unsweetened Oolong Tea	94	7
3	4	Smart Ones Classic Favorites Mini Rigatoni Wit...	38	1
4	5	Green Chile Anytime Sauce	5	13



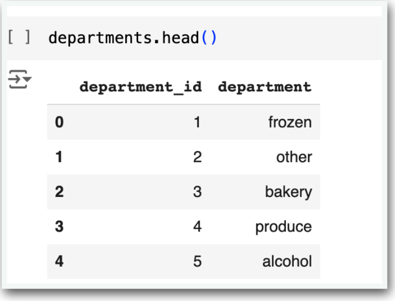
```
aisles.head()
```

	aisle_id	aisle
0	1	prepared soups salads
1	2	specialty cheeses
2	3	energy granola bars
3	4	instant foods
4	5	marinades meat preparation



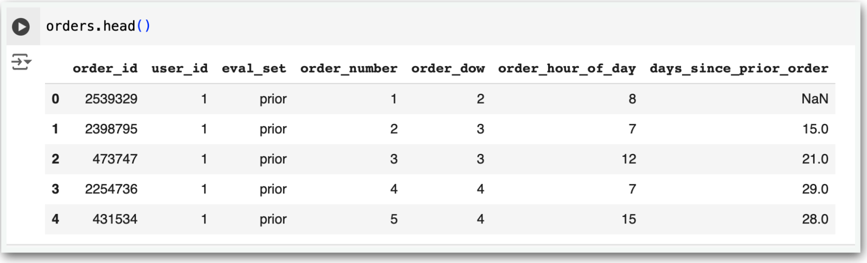
```
order_products__train.head()
```

	order_id	product_id	add_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0
3	1	49683	4	0
4	1	43633	5	1



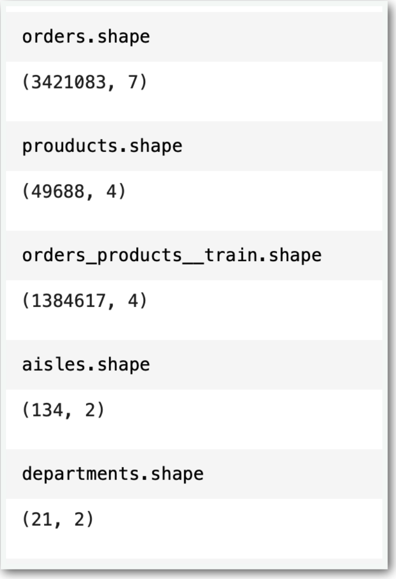
```
departments.head()
```

	department_id	department
0	1	frozen
1	2	other
2	3	bakery
3	4	produce
4	5	alcohol



```
orders.head()
```

	order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2539329	1	prior	1	2	8	NaN
1	2398795	1	prior	2	3	7	15.0
2	473747	1	prior	3	3	12	21.0
3	2254736	1	prior	4	4	7	29.0
4	431534	1	prior	5	4	15	28.0

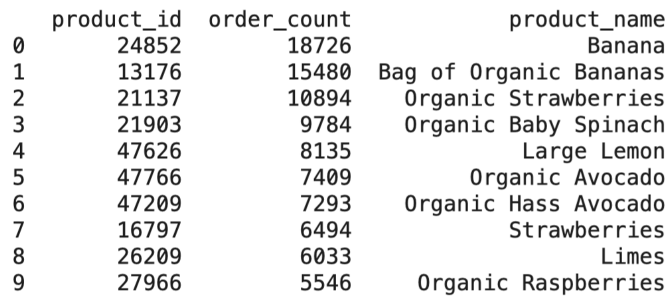


```
orders.shape
(3421083, 7)

prouducts.shape
(49688, 4)

orders_products__train.shape
(1384617, 4)

aisles.shape
(134, 2)

departments.shape
(21, 2)
```

Data cleaning

The dataset is clean and does not require any cleaning.

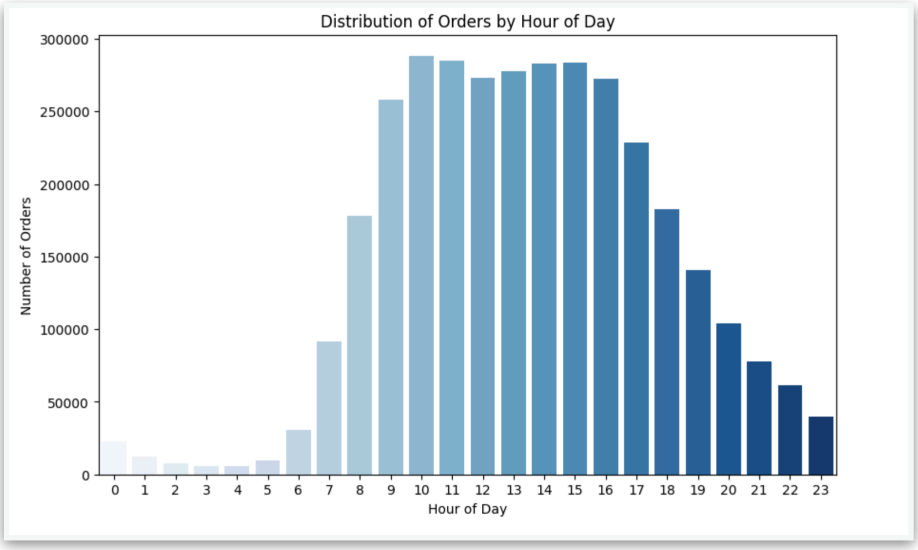
Exploratory Data Analysis



```
   product_id  order_count          product_name
0       24852        18726                Banana
1       13176        15480  Bag of Organic Bananas
2       21137        10894    Organic Strawberries
3       21903         9784    Organic Baby Spinach
4       47626         8135           Large Lemon
5       47766         7409       Organic Avocado
6       47209         7293    Organic Hass Avocado
7       16797         6494          Strawberries
8       26209         6033                 Limes
9       27966         5546     Organic Raspberries
```

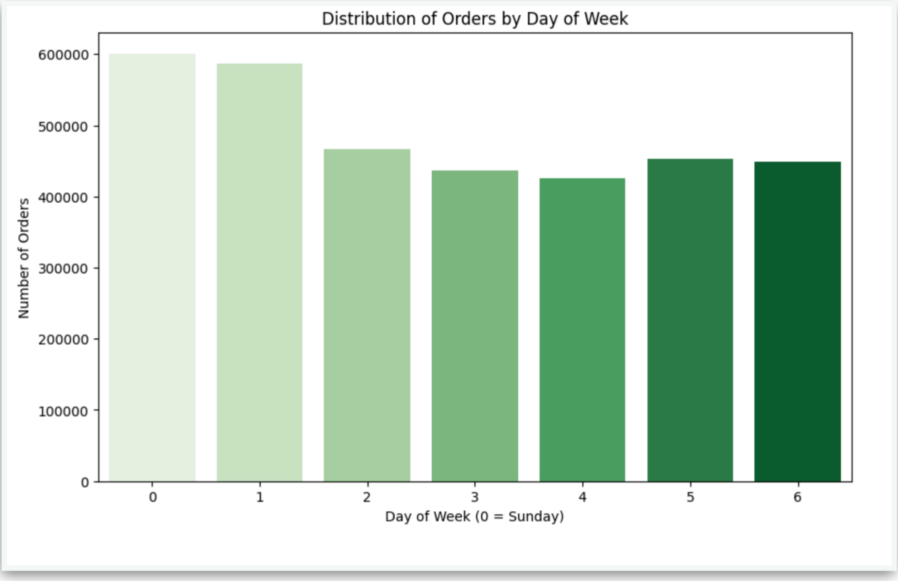
We identified the number of orders per product by aggregating orders based onWe identified the number of orders per product by aggregating orders based on product IDs and merging these results with product data for clarity.

Our results show that bananas and organic bananas are the most ordered products, indicating high demand. Organic items like organic strawberries and avocados also reflect a customer preference for healthier choices. Based on these insights, we recommend optimizing inventory management and focusing on fresh and organic products to better meet customer needs and enhance promotional strategies.



This chart led us to conclude that the highest number of orders occurred at 10 AM. This information is important as it indicates periods where the company can focus on enhancing service, such as:

- **Increasing Resources:** The number of staff can be increased, or logistics can be improved during this hour to meet the rising demand.
- **Marketing Strategies:** Marketing campaigns or special offers can be implemented during this time to boost sales.
- **Product Planning:** Decisions can be made regarding inventory based on peak times, ensuring product availability.



Order distribution information by day of the week enhances understanding of customer behavior, as the chart shows that the highest number of orders occurred on Sunday.

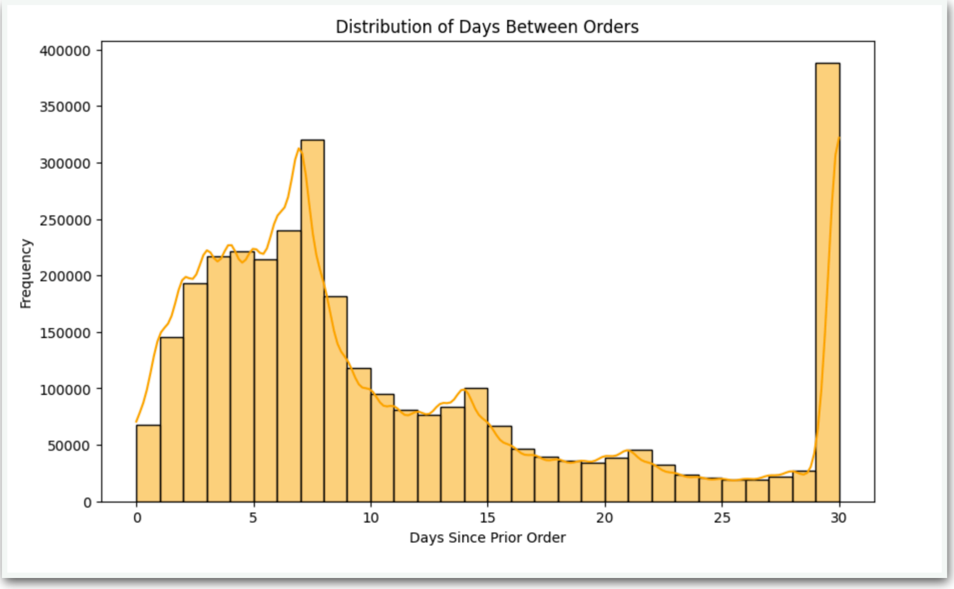
Using this information:

- **Consumer Behavior Analysis:** The increase in orders on Sunday indicates customer activity, suggesting opportunities for boosting sales through special offers.

- **Inventory Planning:** Inventory management can be improved by increasing product quantities available on Sundays.

- **Marketing Strategies:** Special marketing campaigns, such as promotional offers, can be developed to attract more customers.

- **Resource Allocation:** It is beneficial to allocate more staff on Sundays to enhance customer experience.



Understanding reordering periods is essential for grasping customer behavior. In this context, we see that people tend to reorder after specific time frames, mainly focusing on:

1. **After One Week**:
    - Many customers prefer to reorder in a shorter timeframe. This could be due to urgent needs or regular reliance on a specific product.
    - This group typically consists of customers who use the product frequently and want to ensure its availability.
2. **After One Month**:
    - Others opt to reorder after a longer period. This might be related to non-essential products, where customers feel they can wait longer before repurchasing.
    - This group generally includes customers who prefer to assess their experience with the product before deciding to buy again.

Importance of Understanding: Analyzing customer behavior helps refine marketing strategies, while insights into reordering patterns improve inventory management and enable tailored offers, enhancing overall customer service.

Association rule learning using Apriori Algorithm

Apriori Algorithm

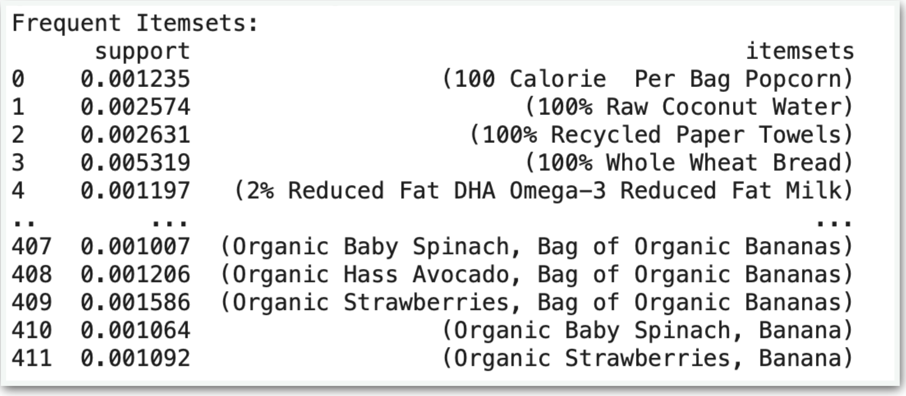
Apriori is a popular algorithm used to extract association rules from databases, particularly in analyzing consumer behavior. It is primarily applied in contexts like **basket analysis**.

How the Algorithm Works

1. **Identifying Frequent Items**:
   The algorithm identifies items that appear together frequently in the database based on a measure called **support**.

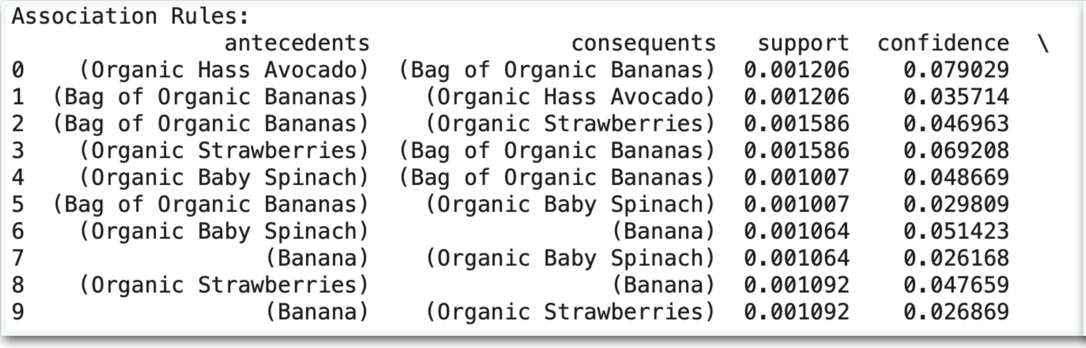
2. **Generating Rules**:
   After identifying frequent items, it generates rules based on two measures:
   - **Support**: The proportion of the dataset that contains a particular set of items.
   - **Confidence**: A measure of the reliability of the rule, indicating the likelihood of item A being present if item B is present.
3. **Filtering Rules**:
   The rules are filtered based on specified support and confidence thresholds to select the most significant rules.

Frequent Itemsets:

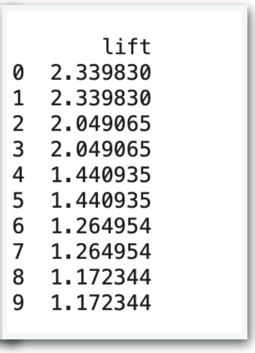


```
Frequent Itemsets:
      support                                         itemsets
0    0.001235                       (100 Calorie  Per Bag Popcorn)
1    0.002574                             (100% Raw Coconut Water)
2    0.002631                         (100% Recycled Paper Towels)
3    0.005319                             (100% Whole Wheat Bread)
4    0.001197    (2% Reduced Fat DHA Omega-3 Reduced Fat Milk)
..        ...                                              ...
407  0.001007   (Organic Baby Spinach, Bag of Organic Bananas)
408  0.001206   (Organic Hass Avocado, Bag of Organic Bananas)
409  0.001586   (Organic Strawberries, Bag of Organic Bananas)
410  0.001064                 (Organic Baby Spinach, Banana)
411  0.001092                 (Organic Strawberries, Banana)
```

Association Rules:

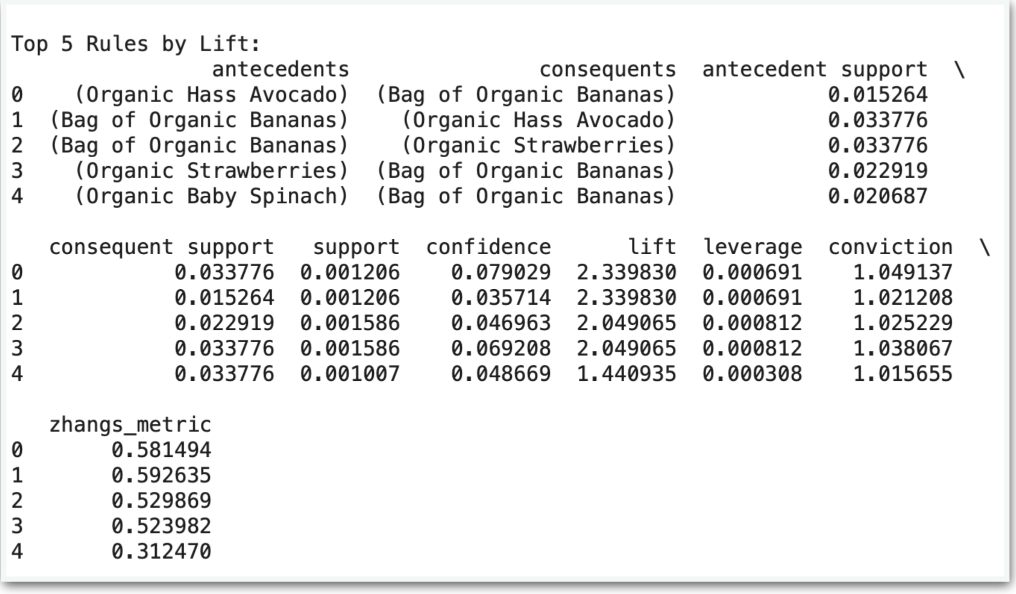


```
Association Rules:
                antecedents              consequents   support  confidence  \
0    (Organic Hass Avocado)  (Bag of Organic Bananas)  0.001206    0.079029
1  (Bag of Organic Bananas)    (Organic Hass Avocado)  0.001206    0.035714
2  (Bag of Organic Bananas)    (Organic Strawberries)  0.001586    0.046963
3    (Organic Strawberries)  (Bag of Organic Bananas)  0.001586    0.069208
4    (Organic Baby Spinach)  (Bag of Organic Bananas)  0.001007    0.048669
5  (Bag of Organic Bananas)    (Organic Baby Spinach)  0.001007    0.029809
6    (Organic Baby Spinach)                  (Banana)  0.001064    0.051423
7                  (Banana)    (Organic Baby Spinach)  0.001064    0.026168
8    (Organic Strawberries)                  (Banana)  0.001092    0.047659
9                  (Banana)    (Organic Strawberries)  0.001092    0.026869
```



```
       lift
0  2.339830
1  2.339830
2  2.049065
3  2.049065
4  1.440935
5  1.440935
6  1.264954
7  1.264954
8  1.172344
9  1.172344
```

Top 5 Rules by Lift:



```
Top 5 Rules by Lift:
                antecedents              consequents  antecedent support  \
0    (Organic Hass Avocado)  (Bag of Organic Bananas)            0.015264
1  (Bag of Organic Bananas)    (Organic Hass Avocado)            0.033776
2  (Bag of Organic Bananas)    (Organic Strawberries)            0.033776
3    (Organic Strawberries)  (Bag of Organic Bananas)            0.022919
4    (Organic Baby Spinach)  (Bag of Organic Bananas)            0.020687

   consequent support   support  confidence      lift  leverage  conviction  \
0            0.033776  0.001206    0.079029  2.339830  0.000691    1.049137
1            0.015264  0.001206    0.035714  2.339830  0.000691    1.021208
2            0.022919  0.001586    0.046963  2.049065  0.000812    1.025229
3            0.033776  0.001586    0.069208  2.049065  0.000812    1.038067
4            0.033776  0.001007    0.048669  1.440935  0.000308    1.015655

   zhangs_metric
0       0.581494
1       0.592635
2       0.529869
3       0.523982
4       0.312470
```

Frequent Itemsets Analysis

1. **Frequent Itemsets:**
   - **Most Frequent Items:** Products like 100% Whole Wheat Bread and 100% Raw Coconut Water show high support, indicating popularity.
   - **Low-Frequency Items:** Many combinations have support around 0.001, suggesting occasional purchases or niche buying patterns.
   - **Organic and Healthy Products:** A significant portion includes terms like organic or healthy, indicating a trend towards these products.
2. **Association Rules Analysis:**
   - **Top Rules by Lift:**
     - (Organic Hass Avocado) → (Bag of Organic Bananas) has a lift of 2.34, indicating strong likelihood of co-purchase.
     - (Bag of Organic Bananas) → (Organic Strawberries) with a lift of 2.05 shows a strong relationship between these fruits.
   - **Confidence Analysis:**
     - (Organic Hass Avocado → Bag of Organic Bananas) has a confidence of 7.9%.
     - Lower confidence for other pairs (e.g., 2.6% for Banana → Organic Strawberries) suggests they are related but not frequently purchased together.
3. **Business Implications:**
   - **Product Bundling Opportunities:** Bundle related products (e.g., avocados and bananas) for promotions to boost sales.
   - **Store Layout Optimization:** Place frequently associated items close together (e.g., Organic Baby Spinach and Bananas) to enhance shopping experience.
   - **Cross-Selling Suggestions:** High lift products (e.g., Hass Avocado and Bananas) are ideal for cross-selling in online carts or at checkout.
   - **Customer Segmentation and Personalized Offers:** Target health-conscious consumers with personalized offers based on organic product purchases to increase retention.