InstaCart Online Grocery Basket Analysis

By Irina Shneider

PROJECT GOALS

How can I identify next best offer for each customer?

How can I do the product offerings more personalized?

How can I avoid overstocking or understocking of certain items?

How can I promote healthy food and wellness?

WHERE DID I GET MY DATA

Instacart is a grocery ordering and delivery app, aims to make it easy to fill your refrigerator and pantry with your personal favorites and staples when you need them.
After selecting products through the Instacart app, personal shoppers review your order and do the in-store shopping and delivery for you.

Data source: Kaggle

- Instacart open sourced this data.
- This is an anonymized data on customer orders over time.

DATA DESCRIPTION

<u>Dataset consists of some tables:</u>

- Aisles.csv
- Department.csv
- Order_product.csv
- Orders.csv
- Products.csv

<u>Final table has columns:</u>

- Order_id
- User_id
- Product_name
- Aisle
- Department
- Food / non-food
- Healthy / unhealthy
- Healthy product share
- Healthy basket

”Food/non-food”, “Healthy/unhealthy” - created manually based on aisle names.
“Healthy basket” - if the percentage of healthy products in the order exceeds 80%, the basket is healthy..

DOES CUSTOMER HAVE HEALTHY OR UNHEALTHY PRODUCT BASKET?

To predict whether a product basket is healthy or unhealthy based on the product names, I utilized the Bag of Words vectorizer to split the product names and implemented three classification algorithms:

Train: 91%
Test: 90%

Logistic Regression

Train: 95%
Test: 77%

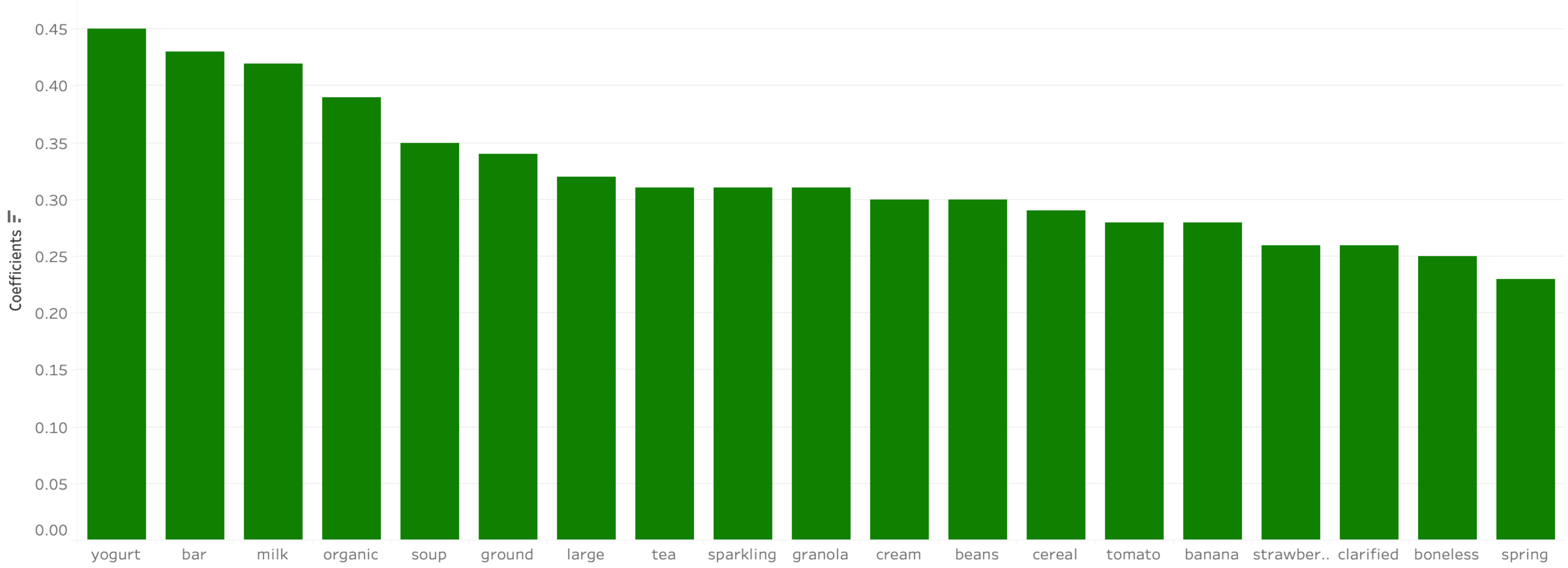
Decision Tree

Train: 69%
Test: 66%

KNN Model

WHAT WORDS IDENTIFY HEALTHY AND UNHEALTHY BASKETS? (1/2)

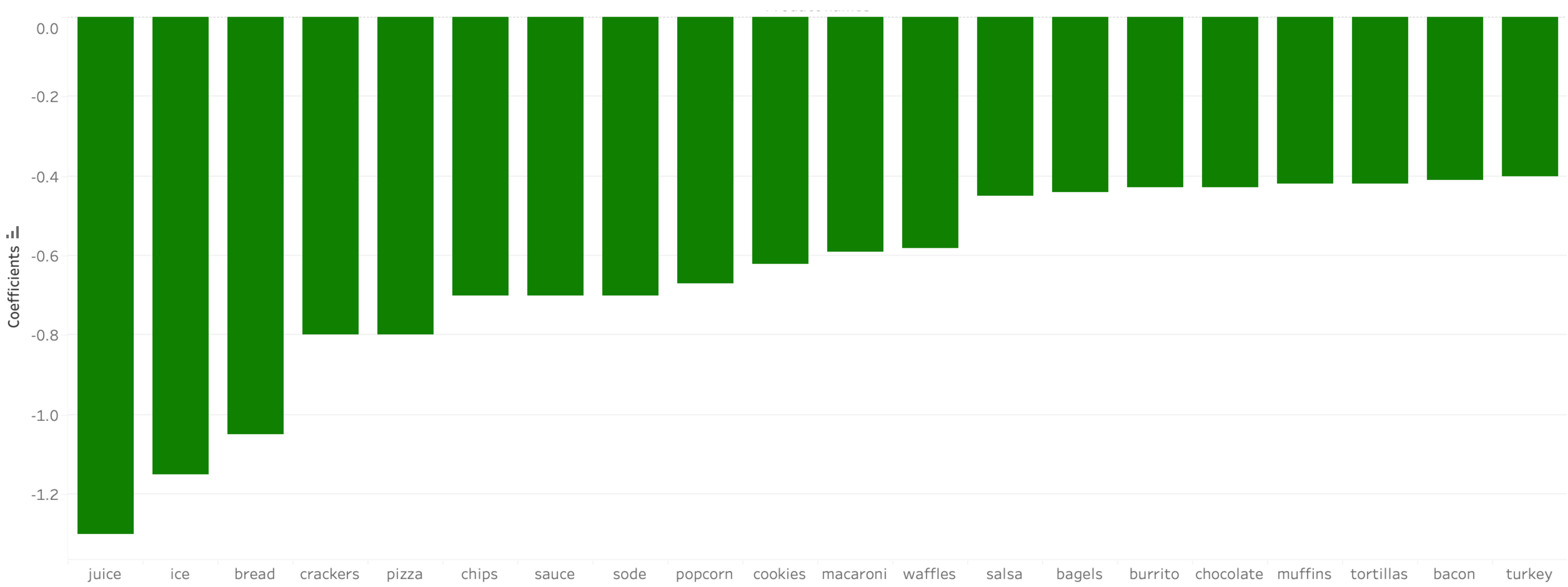
Top 20 coefficients indicate healthy product basket



Yogurt, bar, milk, organic and soup indicate healthy product basket.

WHAT WORDS IDENTIFY HEALTHY AND UNHEALTHY BASKETS? (2/2)

Bottom 20 coefficients indicate unhealthy product basket



Juice, ice (ice cream), bread, crackers and pizza indicate healthy product basket.

HOW TO GROUP CUSTOMERS BASED ON THEIR BEHAVIOUR? (1/2)

By using the KMeans algorithm I found out that the database could be divided into five distinct clusters. Although the silhouette score is low, indicating that the clusters are not well separated from each other, there are strong patterns in customer behaviour.

CLUSTER 1

- Has small volume basket
- Tends to buy paper goods, cleaning products, and laundry items

CLUSTER 2

- Purchases a lot of food
- Family with a kids
- Tends to buy processed food and items that are considered less healthy

HOW TO GROUP CUSTOMERS BASED ON THEIR BEHAVIOUR? (2/2)

CLUSTER 3

- Has small volume basket
- Tends to buy alcohol and party-related items

CLUSTER 4

- Has large volume basket
- Family with a kids
- Tends to buy healthy products

CLUSTER 5

- Has small volume basket
- Single persons
- Hard-workers
- Tends to buy prepared foods

RELATED PRODUCTS THAT ARE FREQUENTLY BOUGHT TOGETHER

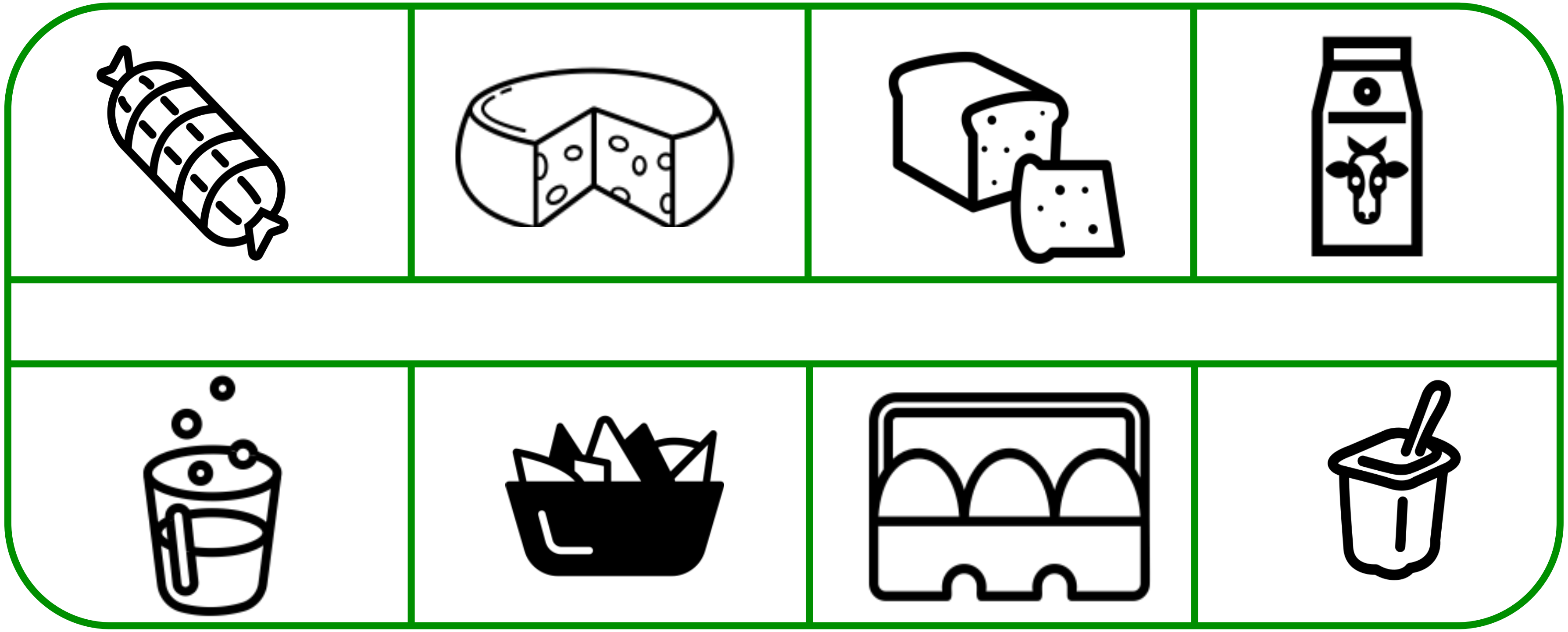
Together purchasing is **3** times frequently than separately

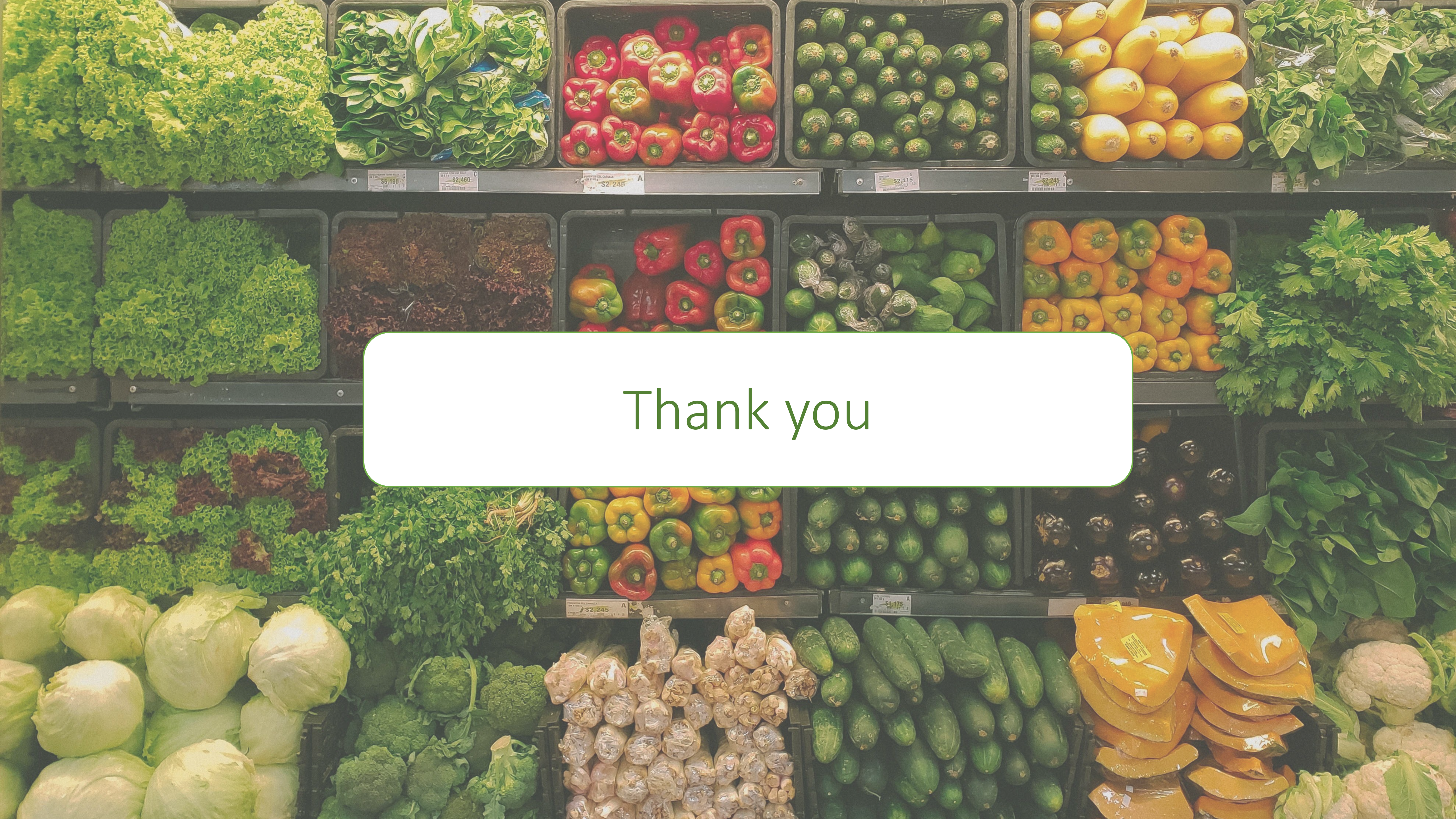
Together purchasing is **3,6** times frequently than separately

Together purchasing is **2** times frequently than separately

HOW TO PLACE AISLES / SUBCATEGORIES IN STORE OR ON THE WEBSITE?

By conducting a Market Basket Analysis based on aisle names, I discovered patterns in customer behaviour to find out aisles they visit sequentially. This information can be used to improve the placement of aisles both in-store and on the website.





Thank you