Market Basket Analysis

A study of Instacart groceries orders

Mafalda Fragoso

Final Project | Data Analytics Bootcamp
17/04/2020

Agenda

1. Research purpose
2. Methodology
3. Main findings

Research purpose

1. **Exploratory analysis on customers buying behavior of groceries**

2. **Segmentation of customers using clustering analysis**

3. **Understand which sets of products are frequently bought together**

Methodology

Data obtained from the Instacart public dataset

- "The Instacart Online Grocery Shopping Dataset 2017"
- Accessed on 08/04/2020 from https://www.instacart.com/datasets/grocery-shopping-2017
- Set of files describing customers' orders over time (data from 6 csv files)

instacart

Grocery delivery platform

Business model

Products you love

Find 1,000's of products from the stores you already shop at.

Same-day delivery

We make deliveries in cities like Los Angeles, Miami, New York City, Chicago, Austin, Washington D.C, Houston, Atlanta and many more.

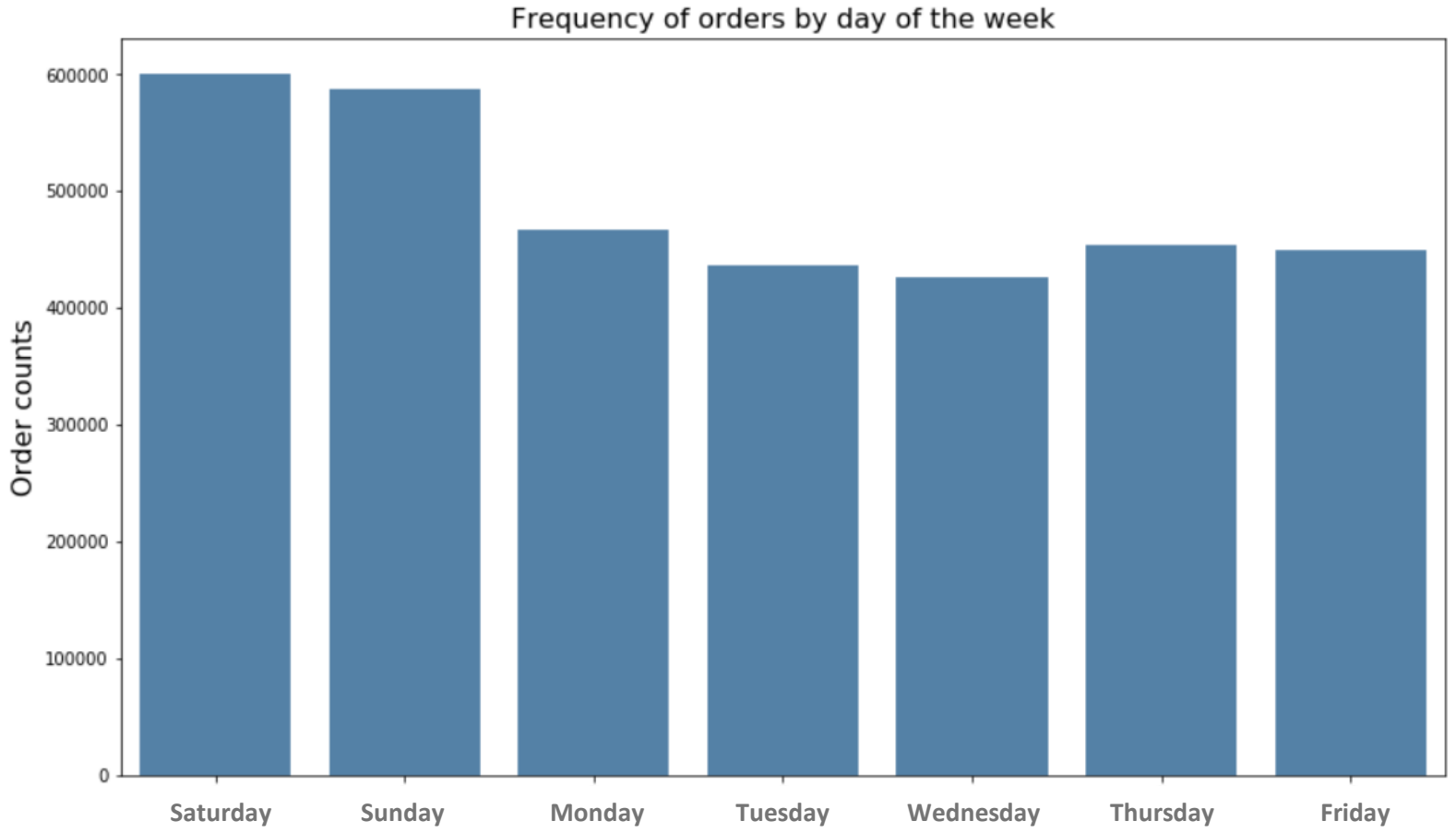
Save time & money

Find exclusive deals on popular products — delivered to your front door!

Main findings: exploratory analysis

When do customers place online orders?

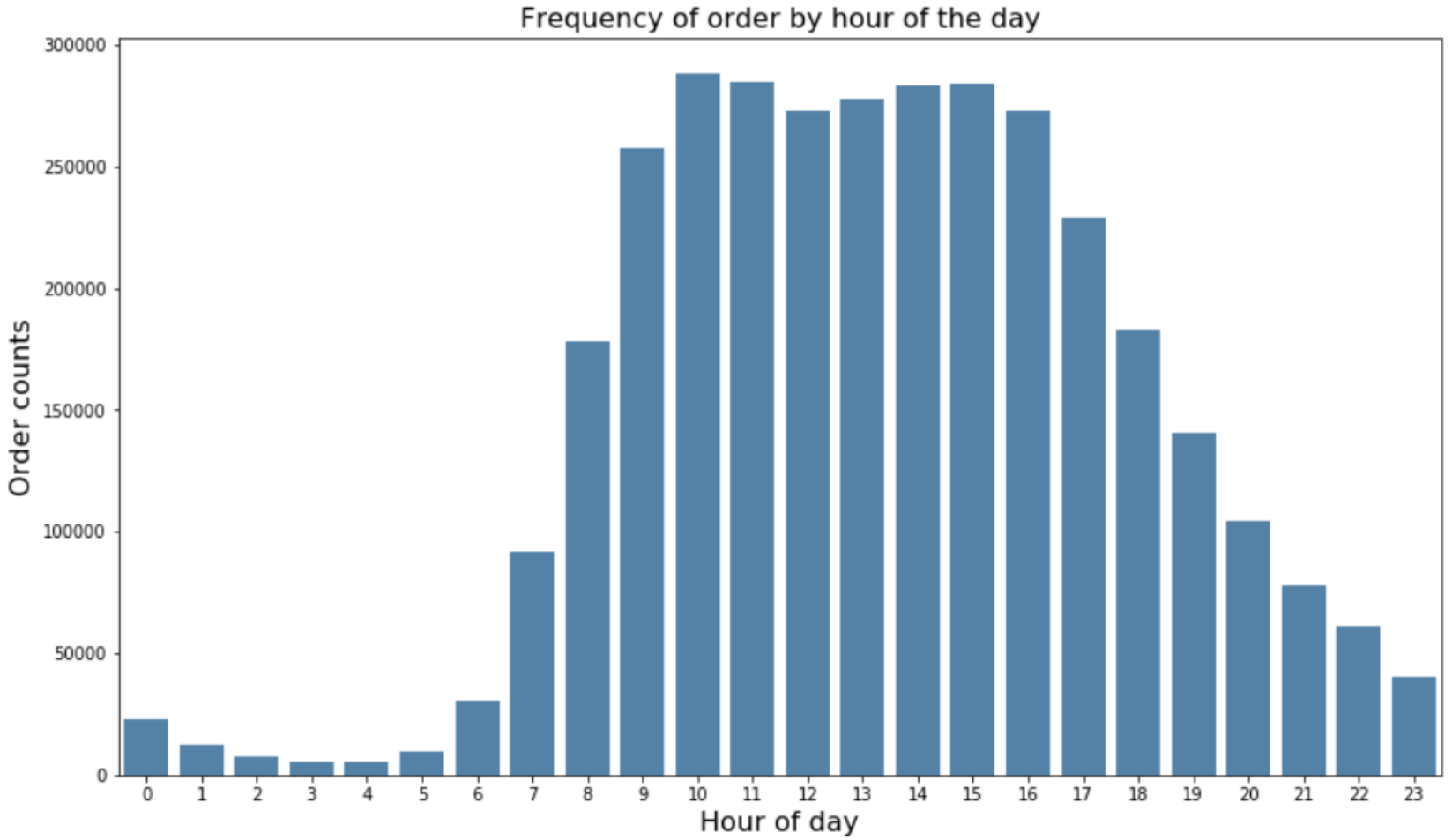
Saturday and **Sunday** are the most frequent days for ordering online



Main findings: exploratory analysis

At what time do customers place orders?

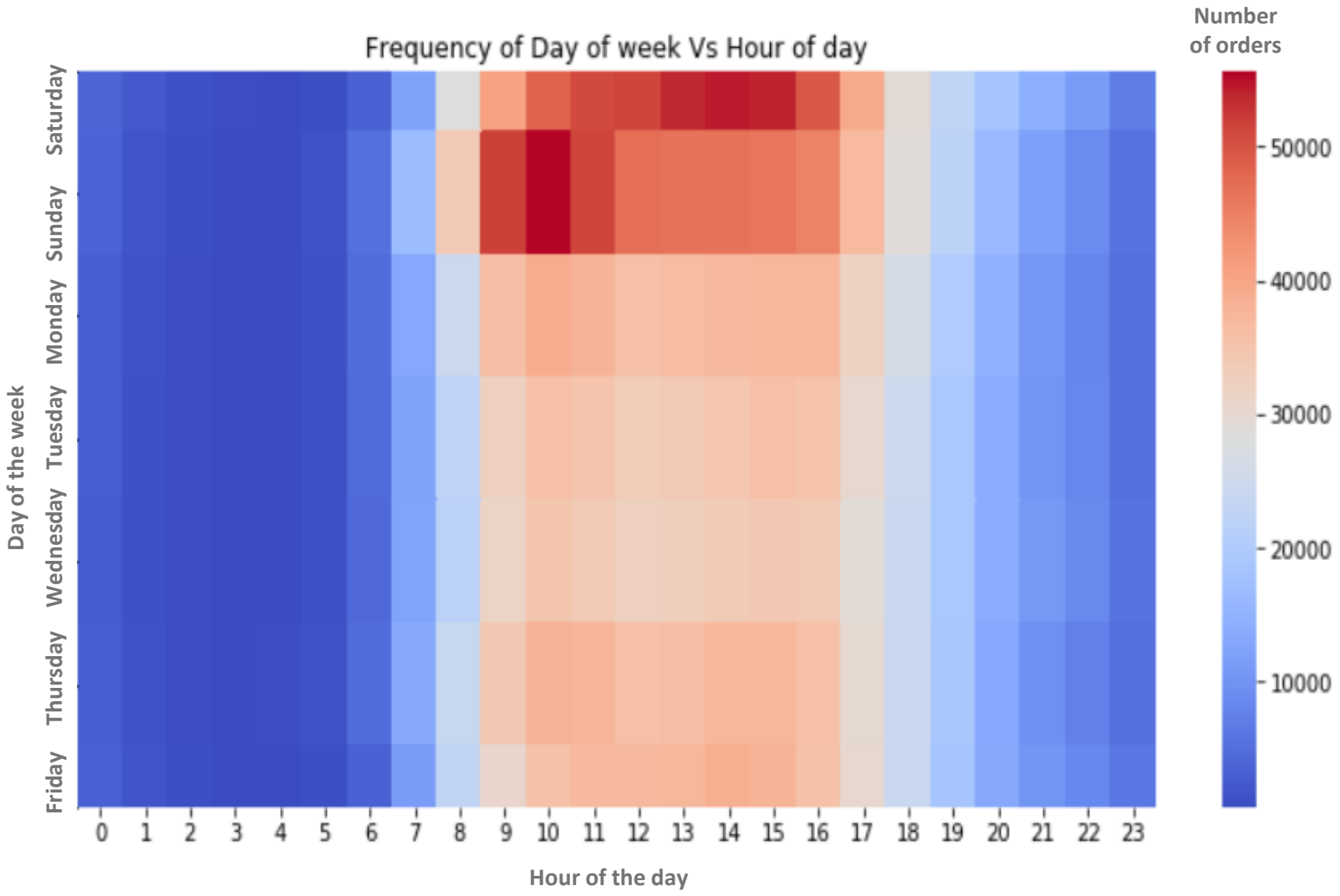
The majority of orders are placed between 10am and 16pm



Main findings: exploratory analysis

What is the effect of hour of the day and day of the week on placing an order?

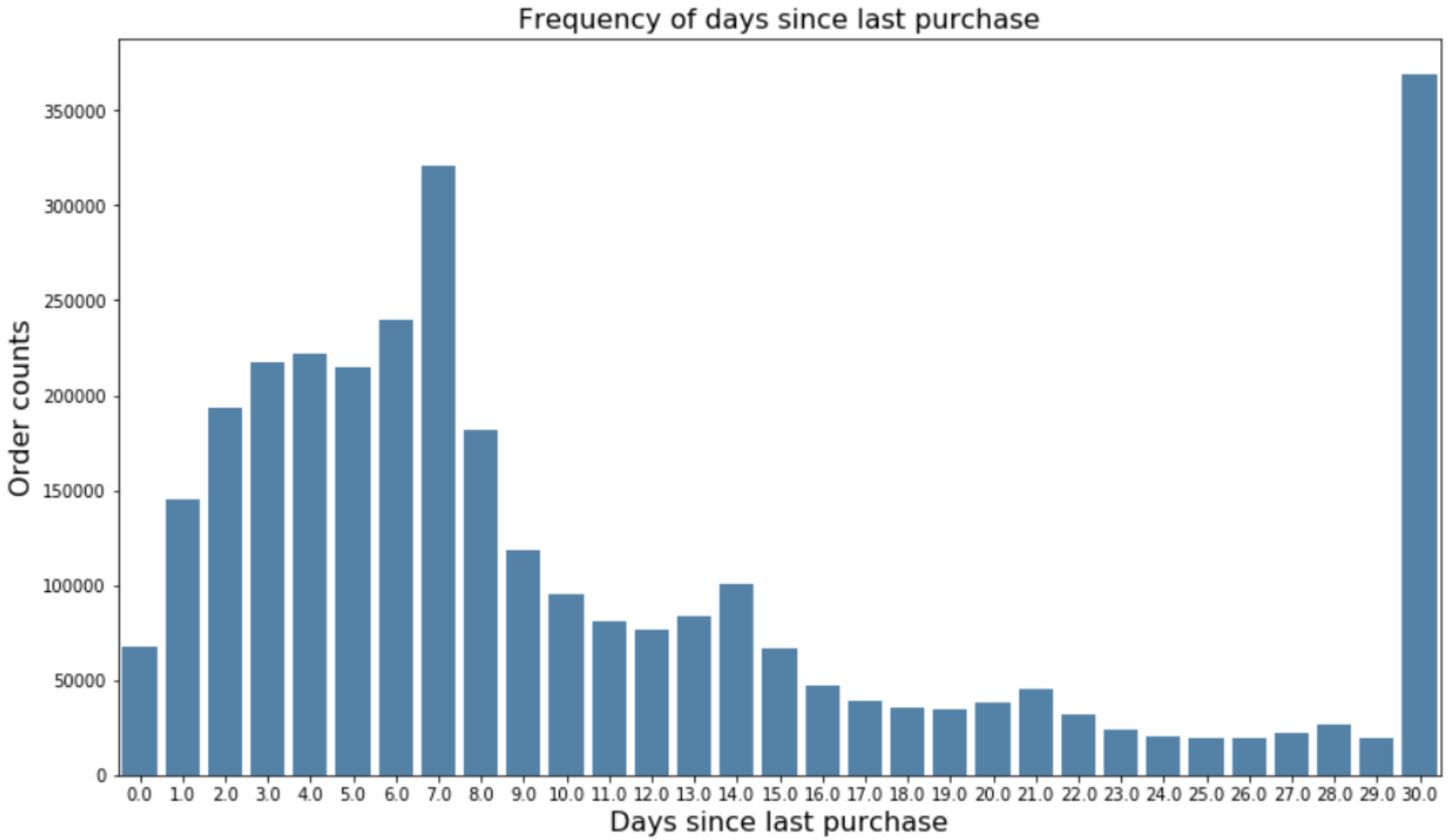
Prime time for orders are Saturday evenings and Sunday mornings



Main findings: exploratory analysis

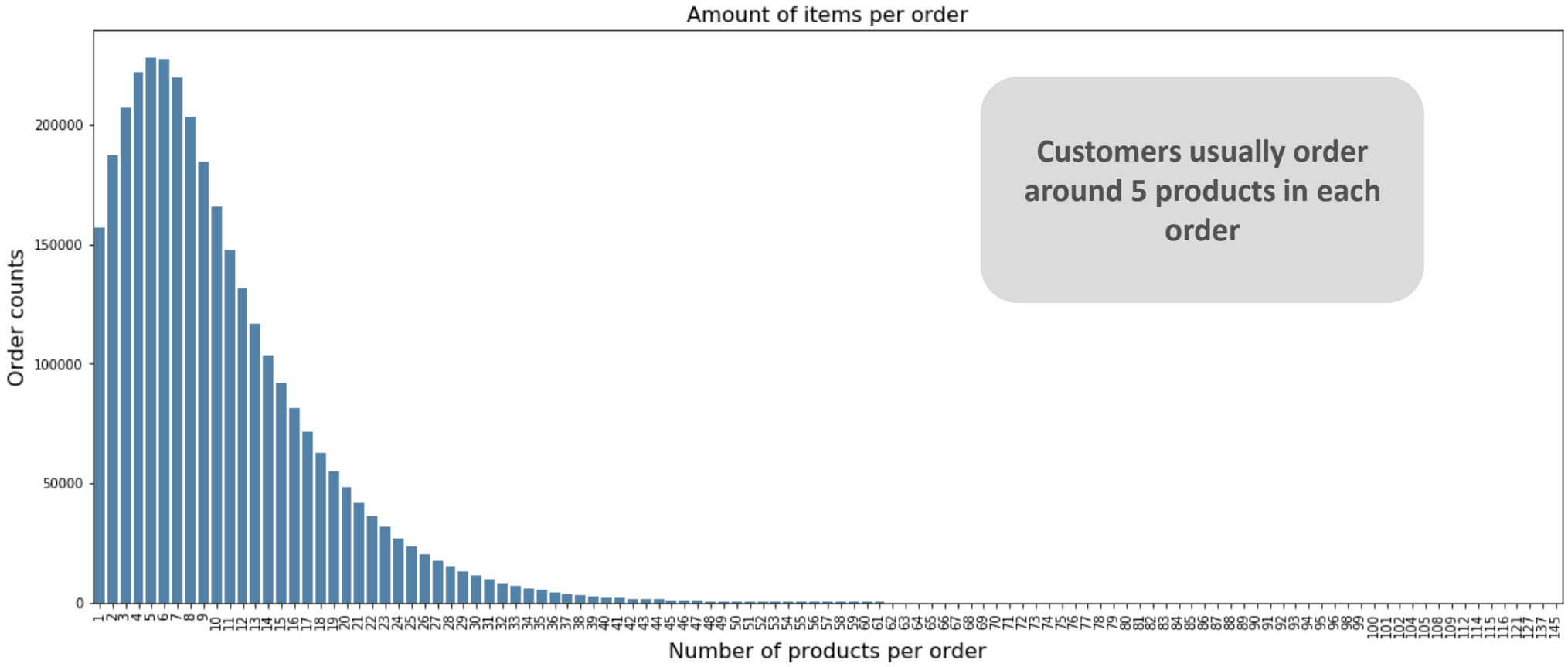
When do consumers order again?

- Customers usually reorder either after a week or after a month
- On average, around 59% of the products in an order are re-ordered products



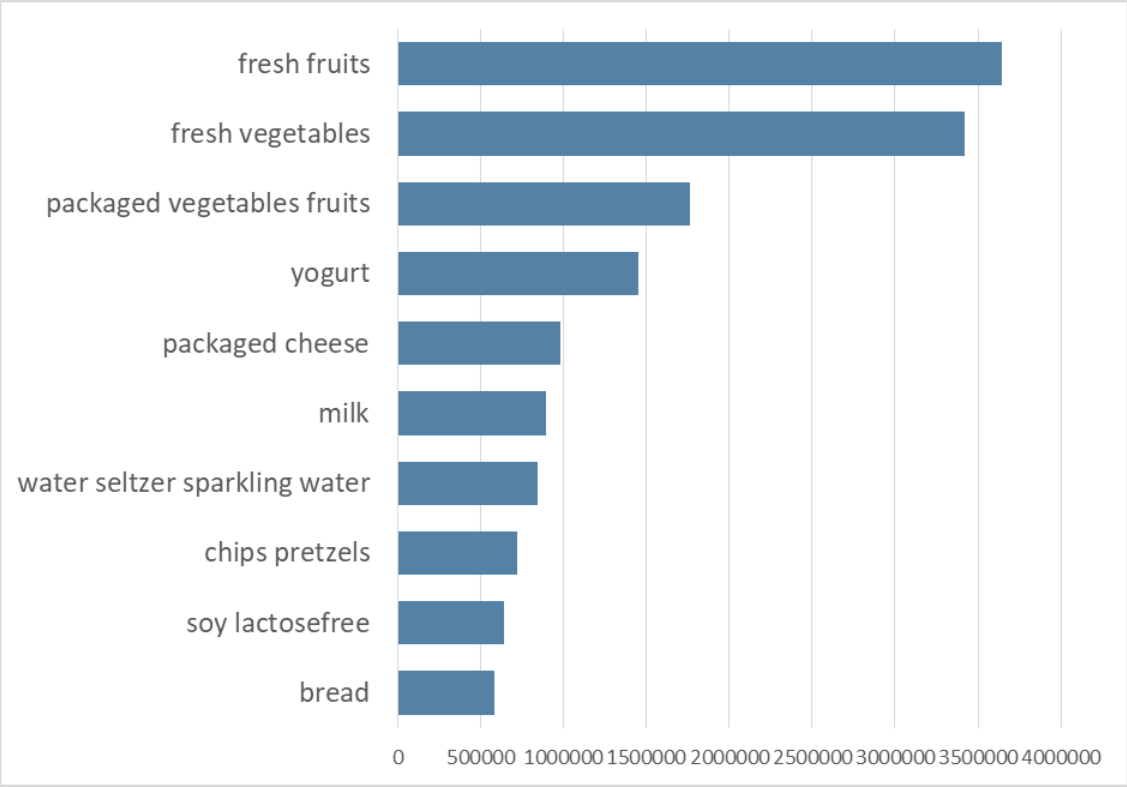
Main findings: exploratory analysis

How many items do people buy in each order?

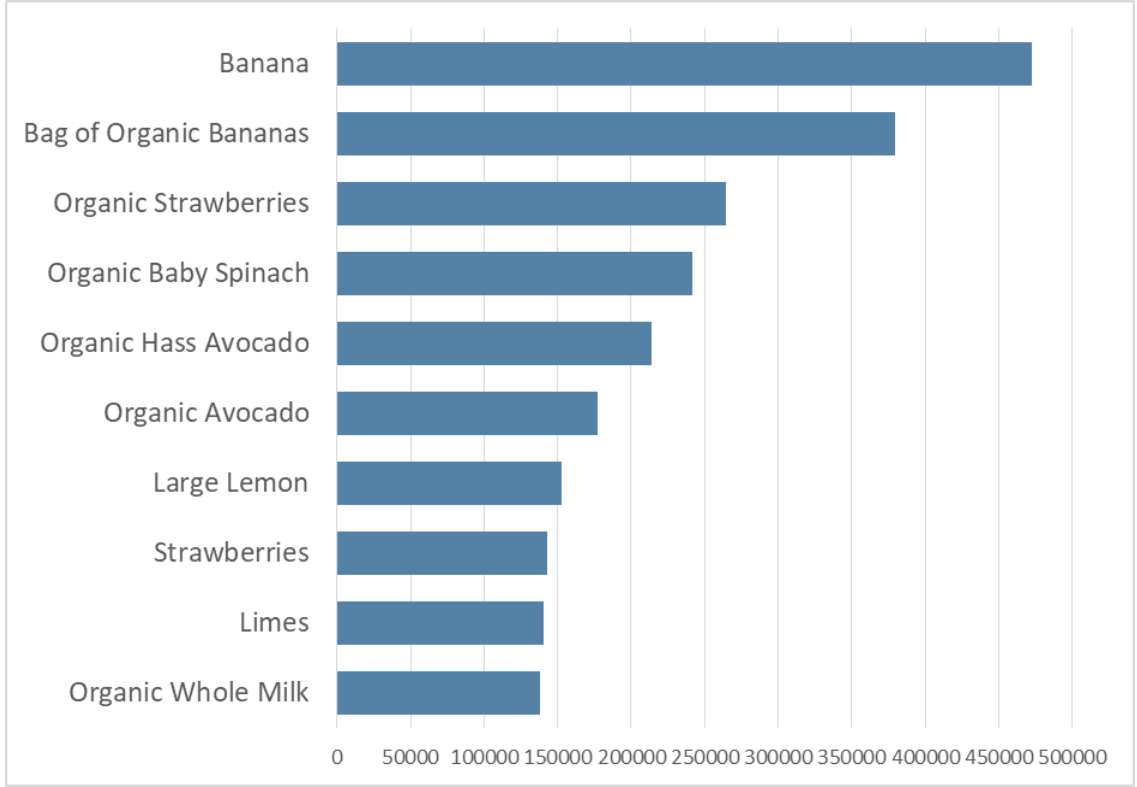


Main findings: exploratory analysis

What are the categories most ordered?



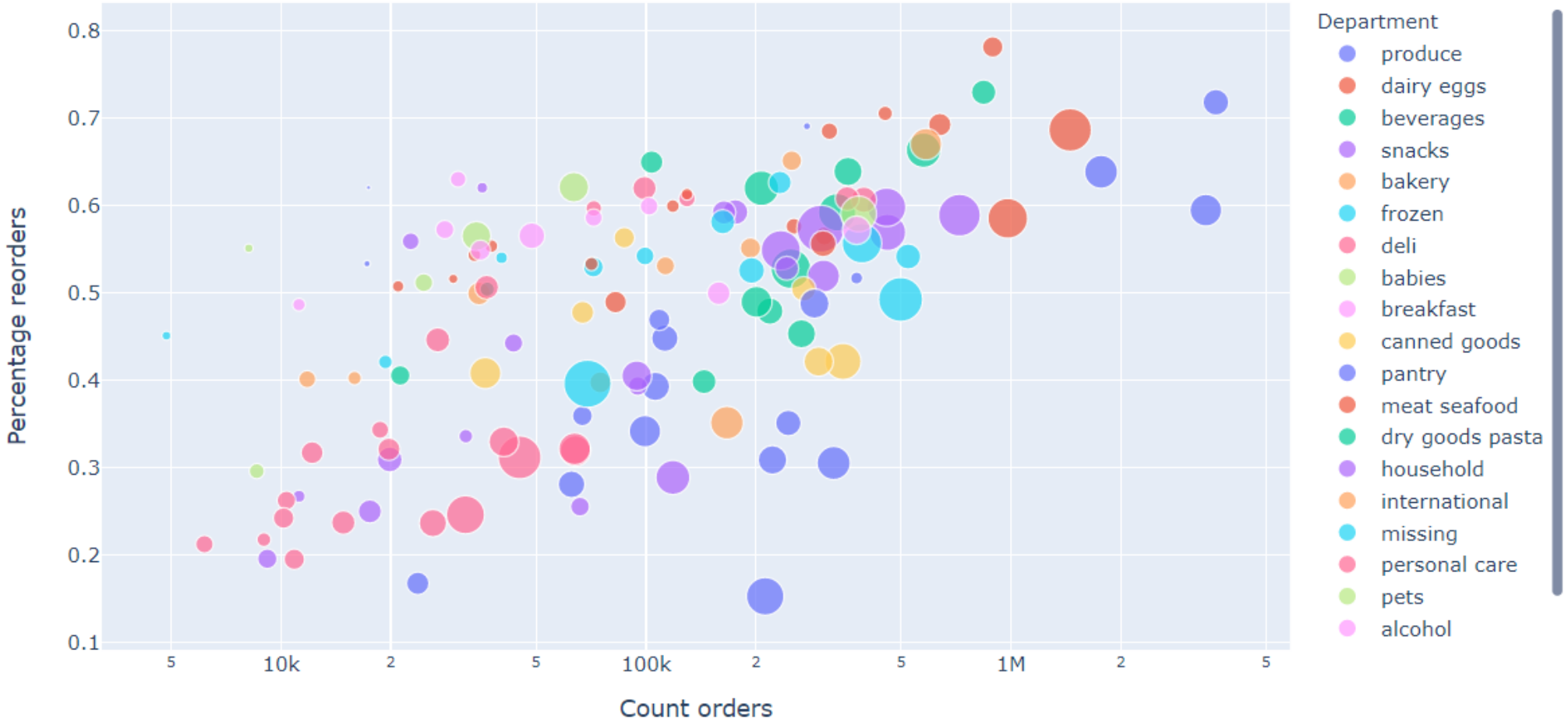
What are the products most ordered?



Most of the top ordered products are organic fruits and vegetables

Main findings: exploratory analysis

The higher the number of orders of an item, the higher the percentage of reorders of that item



Milk has the highest reorder ratio per aisle while spices seasonings has the lowest

Main findings: cluster analysis

Clustering analysis using K-means

	Cluster 1	Cluster 2	Cluster 3
% customers	4%	78%	18%
% orders	22%	38%	40%
Avg. days between orders	7	17	11
Top ordered products (absolute values)	produce, dairy eggs and snacks	produce, dairy eggs and beverages	Produce, dairy eggs and beverages
For each category, what are the most representative clusters?	Fresh products mainly for breakfast and baking dairy eggs, breakfast, bakery Babies	Non perishable food household, alcohol, beverages, frozen	Fresh products produce, dairy eggs

Machine learning algorithm

Apriori Algorithm

- Apriori is an algorithm used for Association Rule Mining. It searches for a series of frequent sets of items in the datasets. It builds on associations and correlations between the itemsets.
- It is the algorithm behind **"You may also like"**

Procedure

1. **Hot encoding the data** (each row represents an order and each column represents the product_id)
2. **Build the model using the mlxtend library**:
    1. Apriori function to extract frequent itemsets for association rule mining
    2. Association rule function to generate association rules from frequent itemsets

product_name	100% Whole Wheat Bread	2% Reduced Fat Milk	Apple Honeycrisp Organic	Asparagus	Bag of Organic Bananas	Banana	Blueberries	Boneless Skinless Chicken Breasts	Broccoli Crown	Bunched Cilantro	...
order_id											
1	0	0	0	0	1	0	0	0	0	0	...
36	0	0	0	1	0	0	0	0	0	0	...
38	0	0	0	0	0	0	0	0	0	1	...
96	0	0	0	0	0	0	0	0	0	0	...
98	0	0	0	0	1	0	0	0	0	0	...

Metrics

1. **Support**: percentage of orders that contain the item set.
2. **Confidence**: the percentage of times that item B is purchased, given that item A was purchased.
3. **Lift**: indicates whether there is a relationship between A and B, or whether the two items are occurring in the same orders by chance

Main findings: predictive model

Which items are frequently bought together?

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
9	(Organic Large Extra Fancy Fuji Apple)	(Bag of Organic Bananas)	0.030831	0.165088	0.010377	0.336562	2.038677
6	(Organic Hass Avocado)	(Bag of Organic Bananas)	0.077777	0.165088	0.025808	0.331825	2.009985
77	(Organic Raspberries)	(Organic Strawberries)	0.059146	0.116180	0.017810	0.301118	2.591814
57	(Organic Cilantro)	(Limes)	0.037603	0.064340	0.010739	0.285593	4.438830
45	(Limes)	(Large Lemon)	0.064340	0.086757	0.017010	0.264379	3.047365
69	(Organic Blueberries)	(Organic Strawberries)	0.052960	0.116180	0.013533	0.255538	2.199491

Practical applications for the retail industry

Help retailers to **develop marketing strategies** by gaining insight into which items are frequently purchased together by customers:

- Changing the store layout according to trends
- Customer behavior analysis
- Cross-selling

Thank you