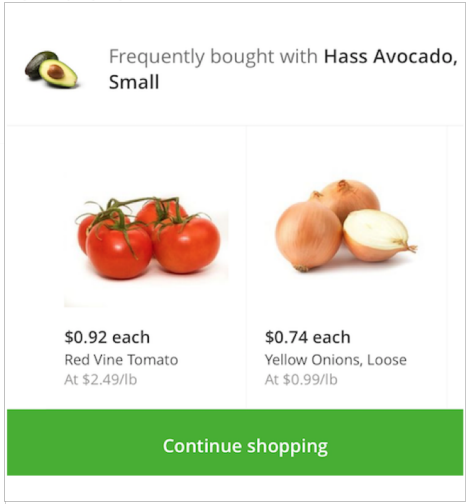
# Market Basket Analysis Using Apriori Algorithm

Yunjin Bak

# Introduction

- Analyzing customers' purchase habit
- Which items are bought together frequently
- Recommendation system for shopping websites

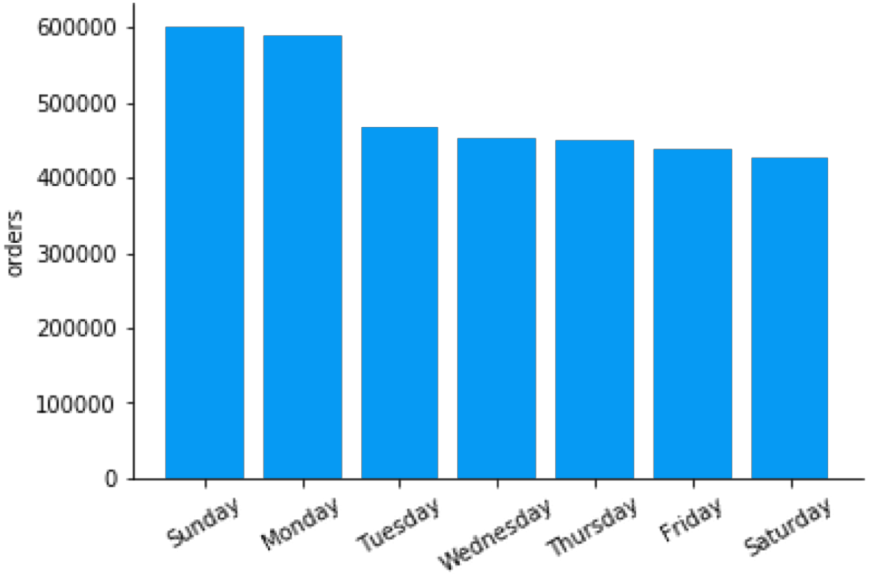



## Dataset

- Instacart orders data
- Contains over 3 million grocery orders  from more than 200,000 Instacart users
- Includes information about order time and day, whether reordered or not, days since prior order, aisles and departments and so on

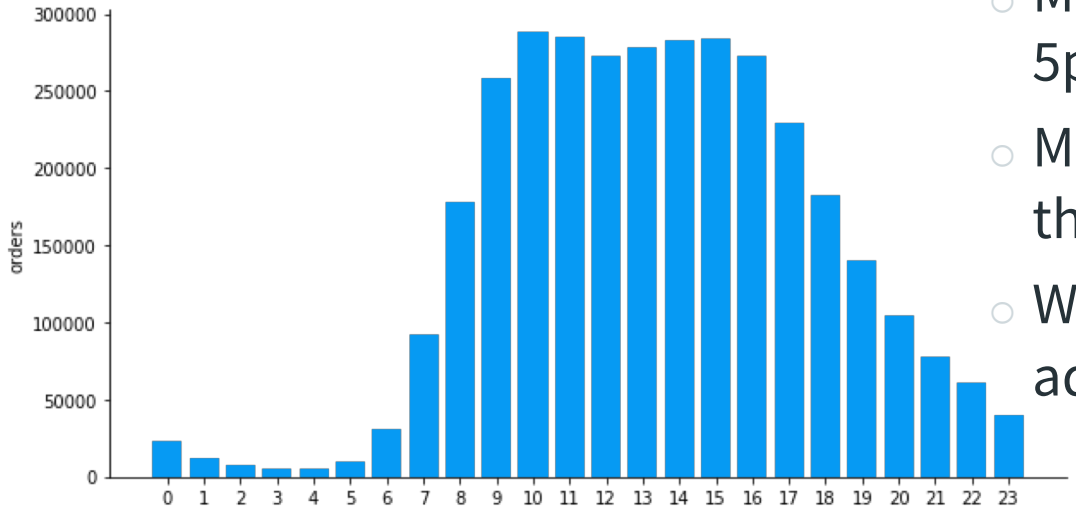
# 1.
# Exploratory Analysis

# Orders by Day of Week



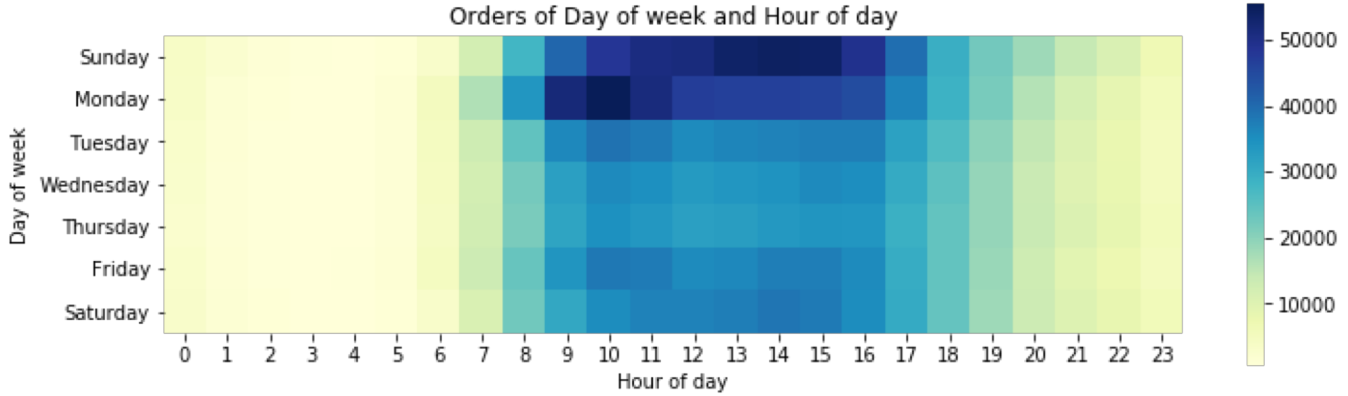

- Sunday and Monday are the busiest days of week
- Important for inventory management with other information

# Orders by Hours of Day



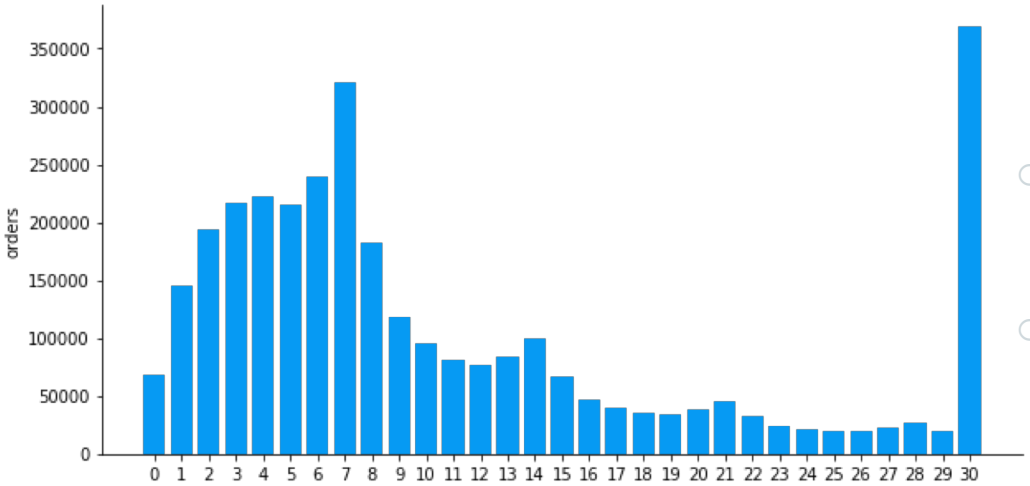

- Most of orders occur from 10am to 5pm
- Means there are most of access to the website
- We could use this information for ads on the website

# Orders by Day of Week x Hour



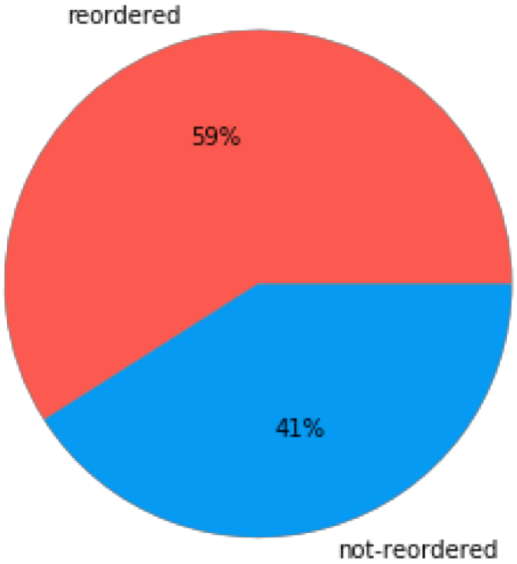

- By combining two factors, we can see which day and what time is busy

# Days Since Previous Order



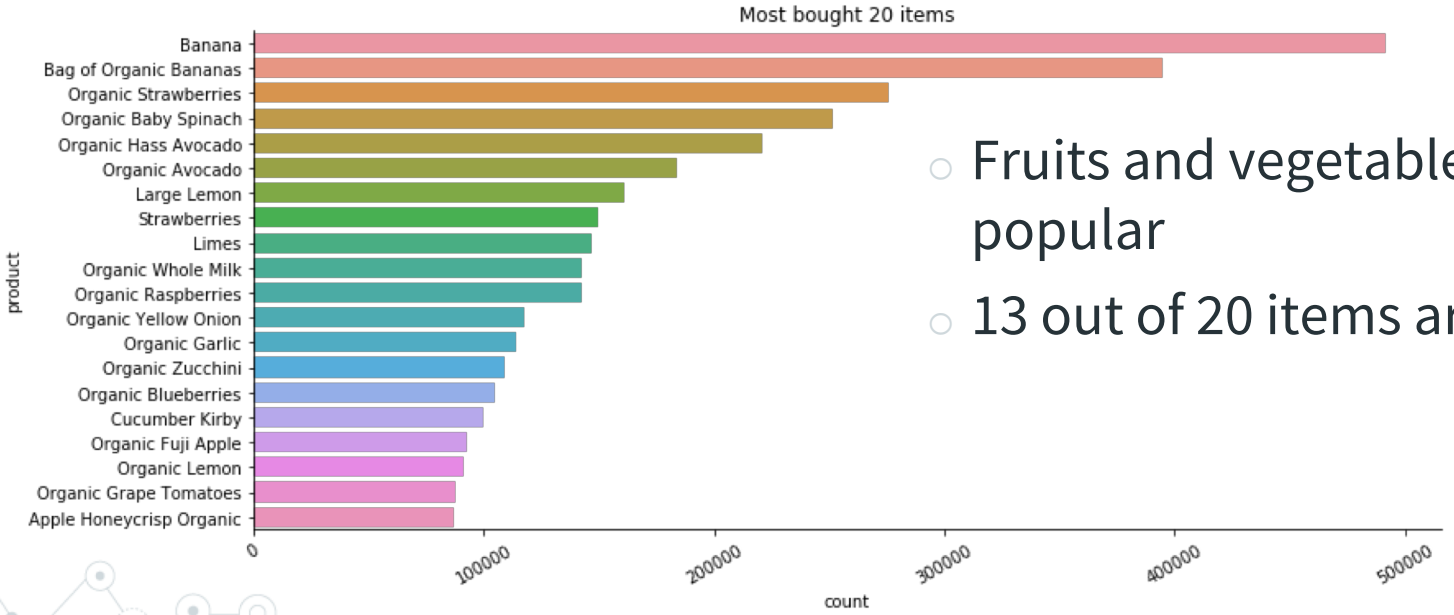

- Many orders are made weekly and monthly
- Small peaks at 14, 21, 28

# Percentage of Reorders



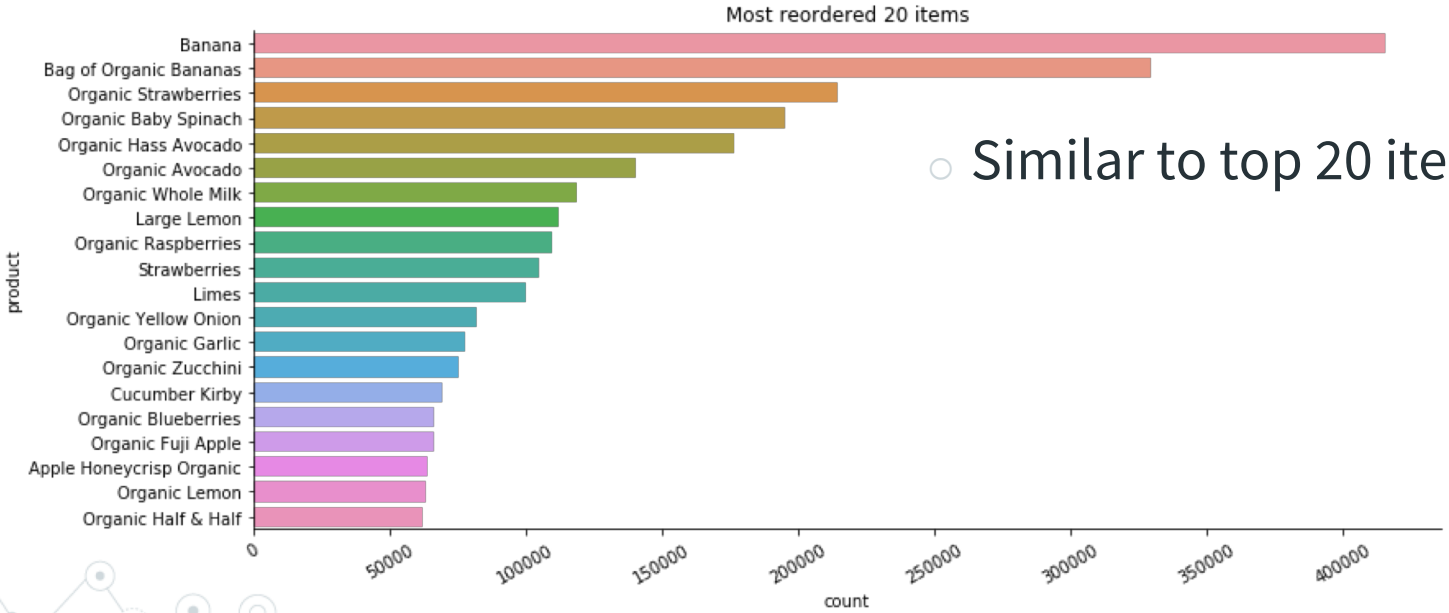

- 59% of products in an order are re-ordered by same customer

# Most Popular Items



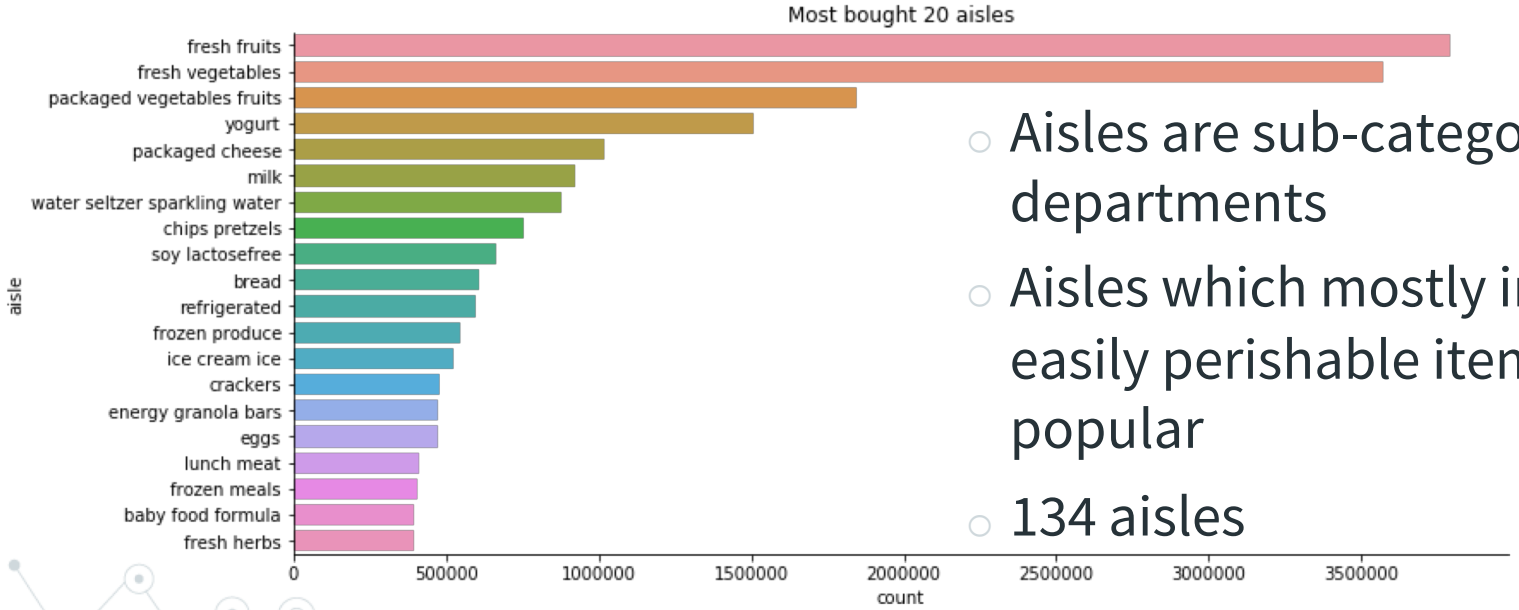

- Fruits and vegetables are most popular
- 13 out of 20 items are organic

# Most Reordered Items



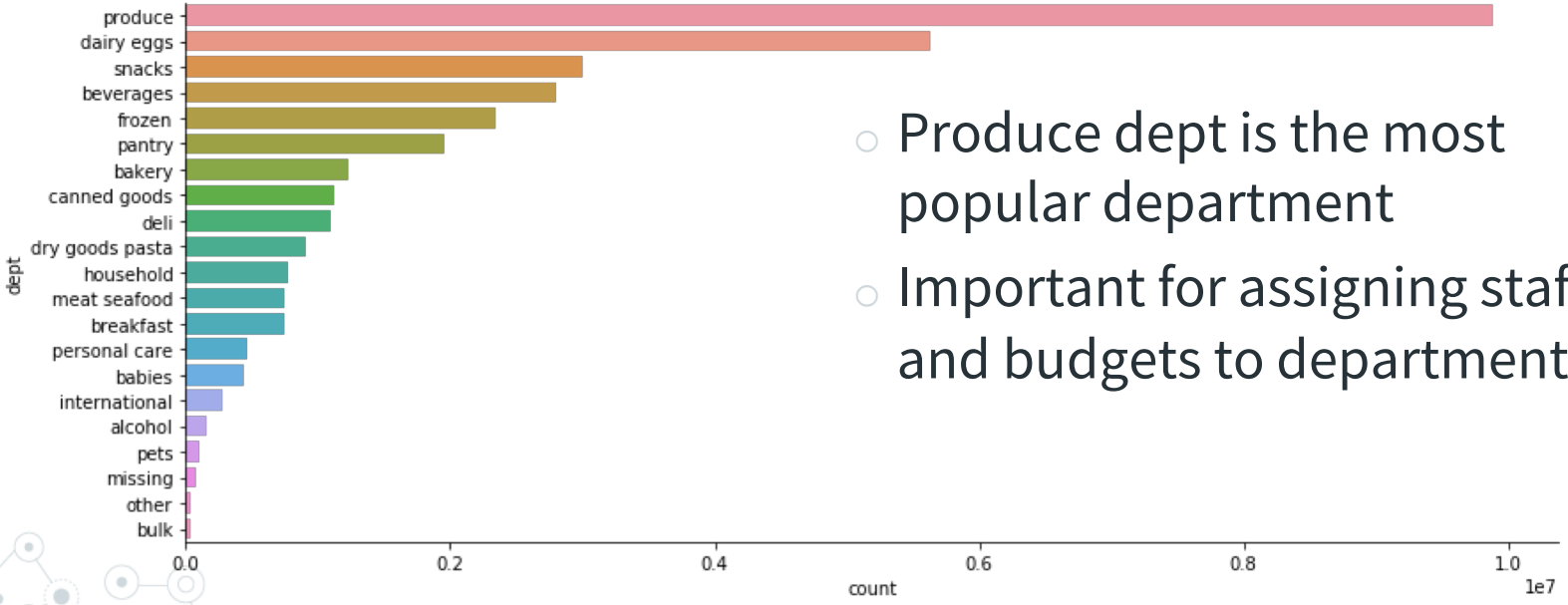

- Similar to top 20 items

# Most Popular Aisles




- Aisles are sub-categories of departments
- Aisles which mostly include easily perishable items are popular
- 134 aisles

# Departments by Order




- Produce dept is the most popular department
- Important for assigning staffs and budgets to departments

# 2.

# Association Rules Generation

From Frequent Item sets

# Metrics

- Support, Confidence, Lift
- Lift is commonly used metric
- Lift value for X->Y and Y->X is same

$$Support(Itemset) = \frac{Number\ of\ Transaction\ involving\ itemset}{Total\ Number\ of\ Transactions}$$

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$


**Of transactions that included milk:**

- 71% included bread
- 43% included eggs
- 29% included toilet paper

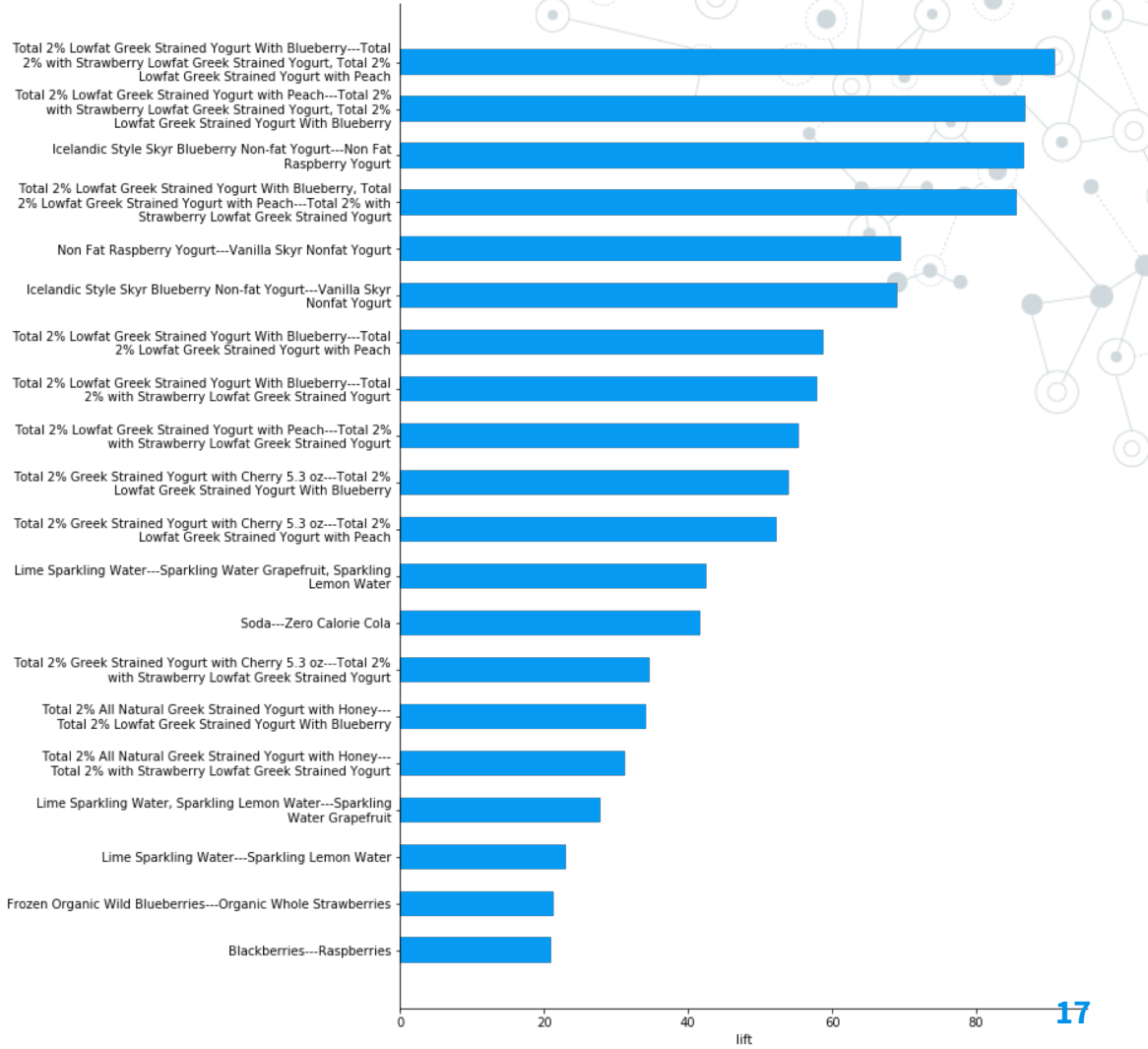
# Metrics

- mlxtend module
- Need to convert data into item lists of each transaction
- Returns dataframe including the metrics

| | antecedents | consequents | antecedent support | consequent support | confidence | lift |
|---|---|---|---|---|---|---|
| **1919** | (Total 2% Lowfat Greek Strained Yogurt With Bl... | (Total 2% with Strawberry Lowfat Greek Straine... | 0.005669 | 0.002300 | 0.209016 | 90.869349 |
| **1914** | (Total 2% with Strawberry Lowfat Greek Straine... | (Total 2% Lowfat Greek Strained Yogurt With Bl... | 0.002300 | 0.005669 | 0.515152 | 90.869349 |
| **1918** | (Total 2% Lowfat Greek Strained Yogurt with Pe... | (Total 2% with Strawberry Lowfat Greek Straine... | 0.005019 | 0.002718 | 0.236111 | 86.856600 |
| **1915** | (Total 2% with Strawberry Lowfat Greek Straine... | (Total 2% Lowfat Greek Strained Yogurt with Pe... | 0.002718 | 0.005019 | 0.435897 | 86.856600 |
| **935** | (Non Fat Raspberry Yogurt) | (Icelandic Style Skyr Blueberry Non-fat Yogurt) | 0.003996 | 0.004833 | 0.418605 | 86.618962 |

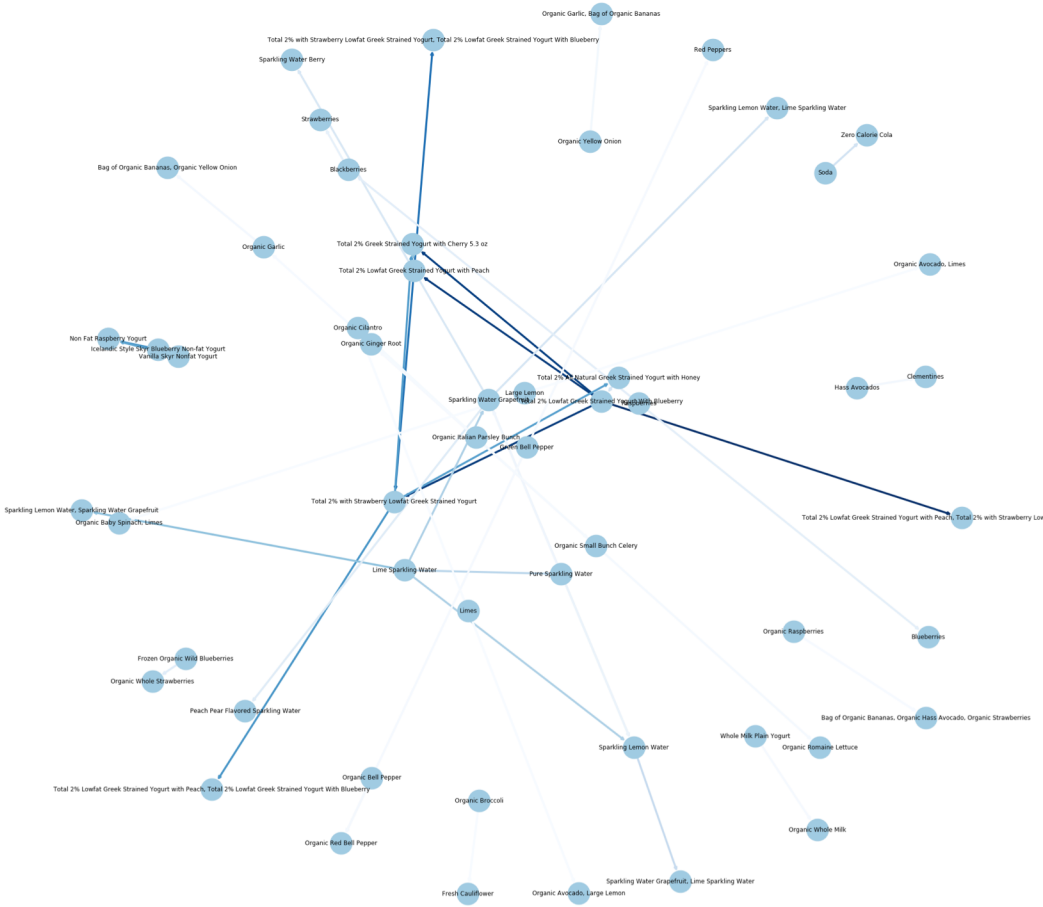
# 20 Popular Item Pairs by Lift

- Yogurt with different flavors are most bought together
- Similar items in same category are bought together



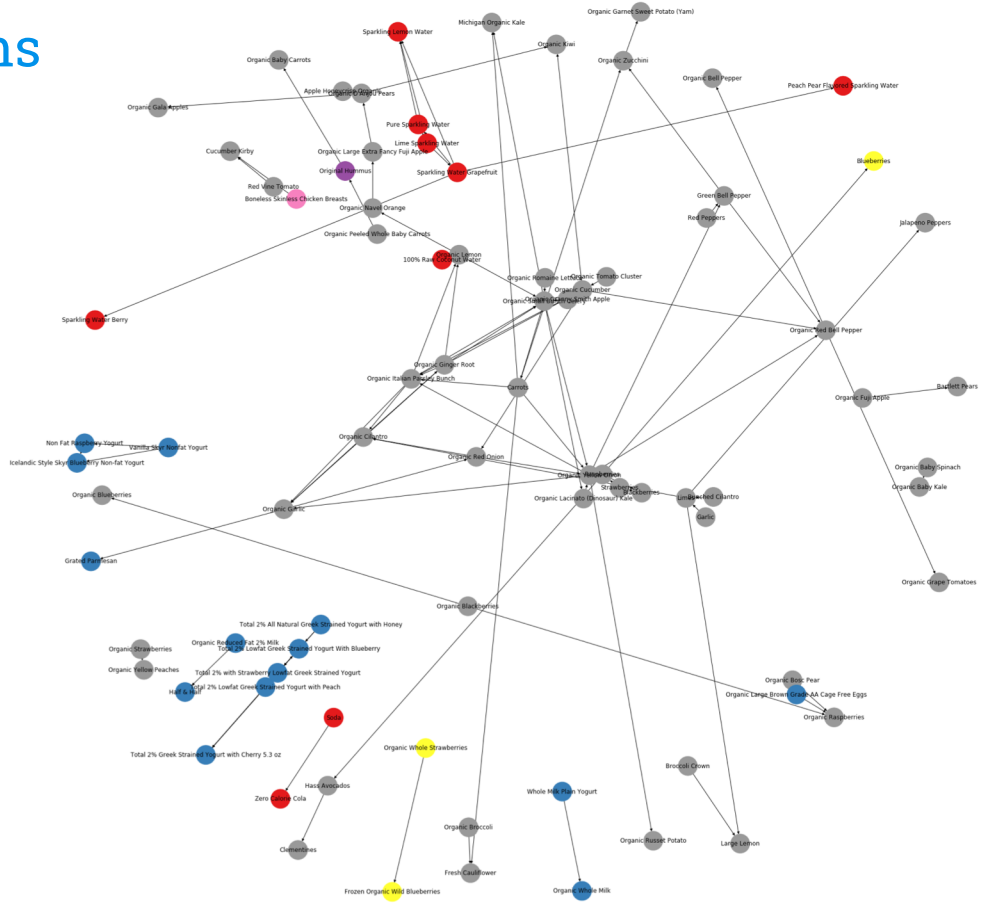
# Relationship Between Items

- Acyclic graph (directed)
- Edges are weighted by lift
- Top 50 item sets
- Organic items tend to be connected to each other with high lift value



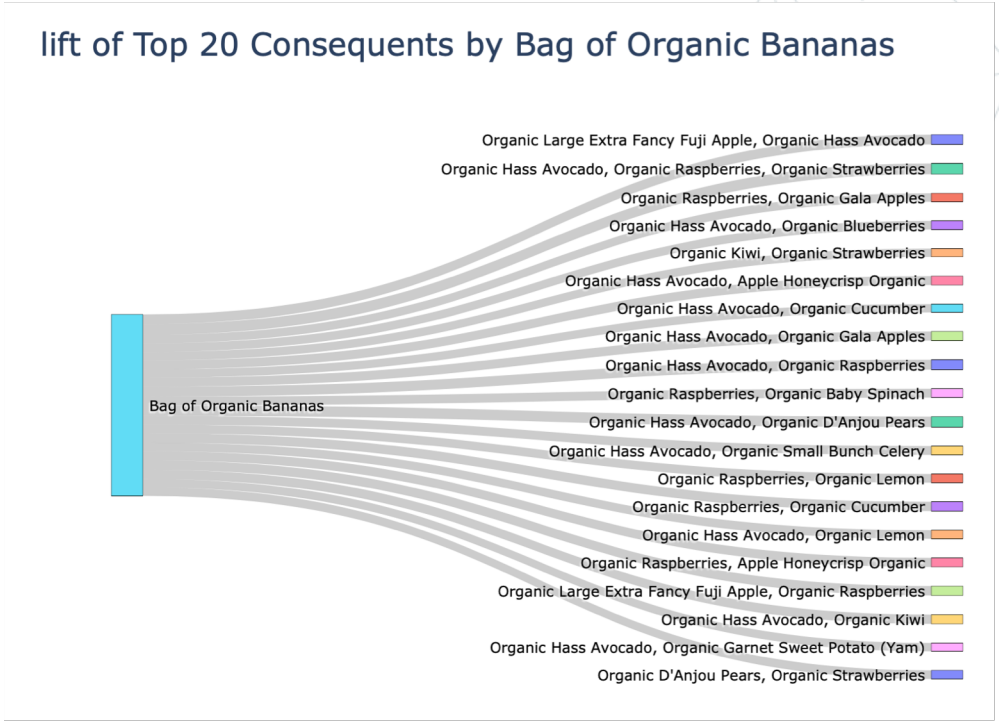
# Relationship Between Items with Categories

- Top 50 itemsets with department
- Items within same category tend to cluster



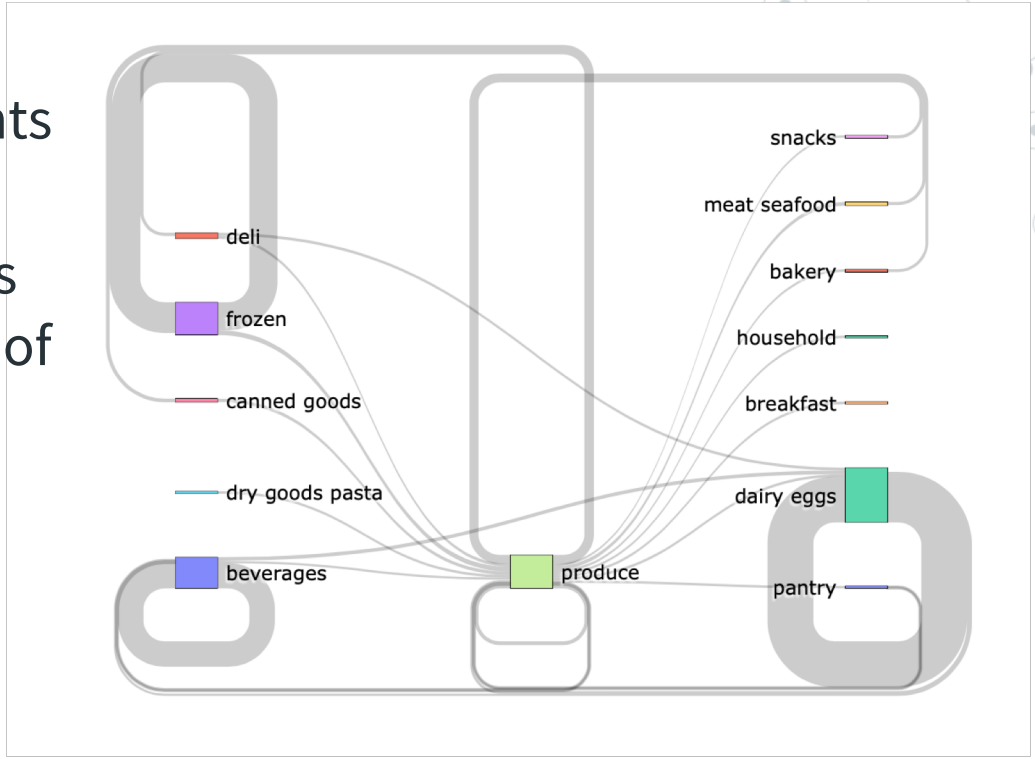
# Sankey Diagram for Top N Items Bought with Item A

- Interactive figure (demo)
- Prints out which items have highest lift value with specific item
- Algorithm for recommendation engine



# Sankey Diagram for Departments

- Shows which two departments have high lift value
- Dairy/eggs, frozen, beverages and produce depts have lots of self-loops
- Produce dept is the most connected with other depts



# Conclusion

- Association rules mining using Instacart orders
- By calculating lift value, which items are purchased together frequently were examined
- Could be used for recommendation engines and inventory management
- Information about individual customer’s buying pattern can improve a prediction about next purchases

# Thanks!

## Any questions?

- github.com/datasciyj
- yunjin.bak1@gmail.com