INSTACART MARKET BASKET ANALYSIS

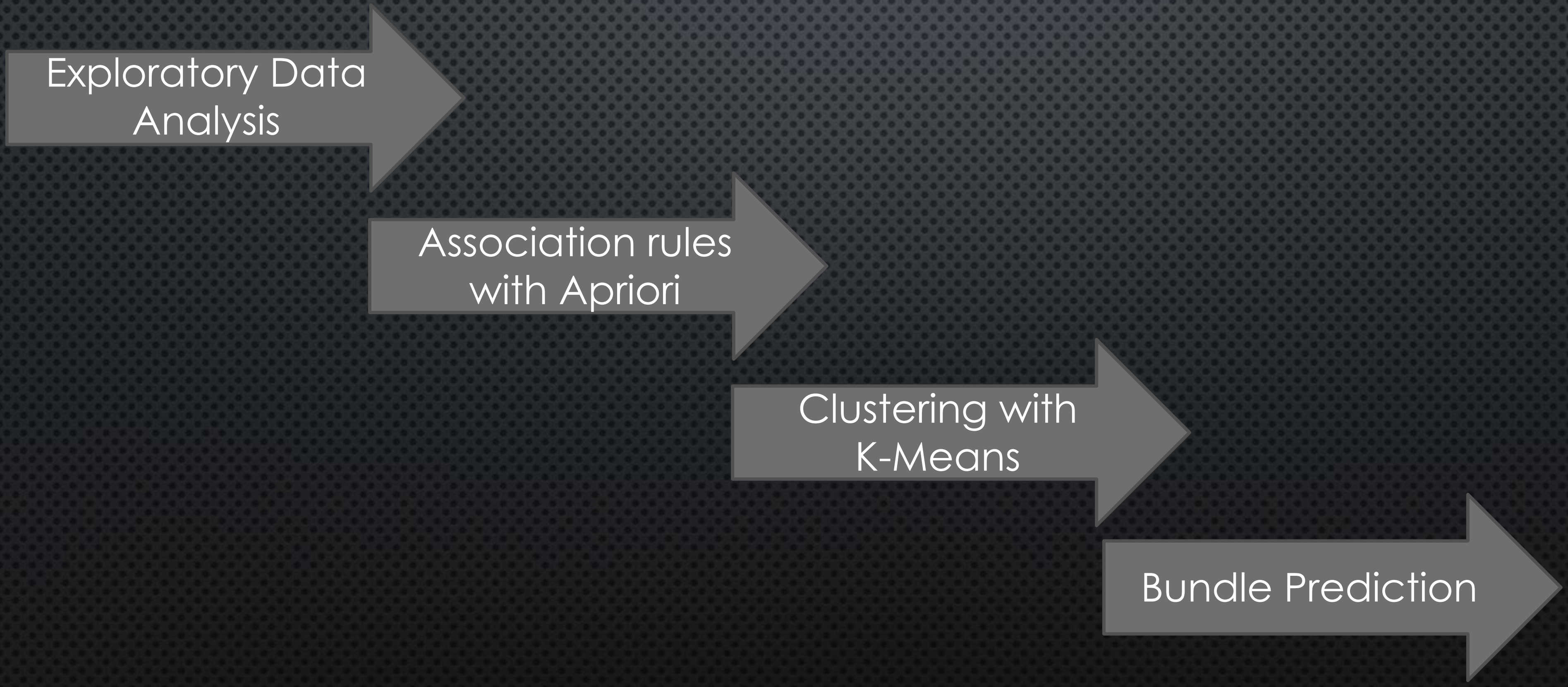
By: Ari Sagherian, Suhasini Kalaiah Linagiah,

Bharath kumar karre, and Sumanth Pobala

Date: 12/15/2019

Class: cis 5570

PROJECT PIPELINE



Exploratory Data Analysis

Association rules with Apriori

Clustering with K-Means

Bundle Prediction

MOTIVATION

What is Instacart?

- Same-Day Grocery delivery service

Goal of project

For Instacart:

1. Increase sales
2. Improve customer satisfaction
3. Gain more customers

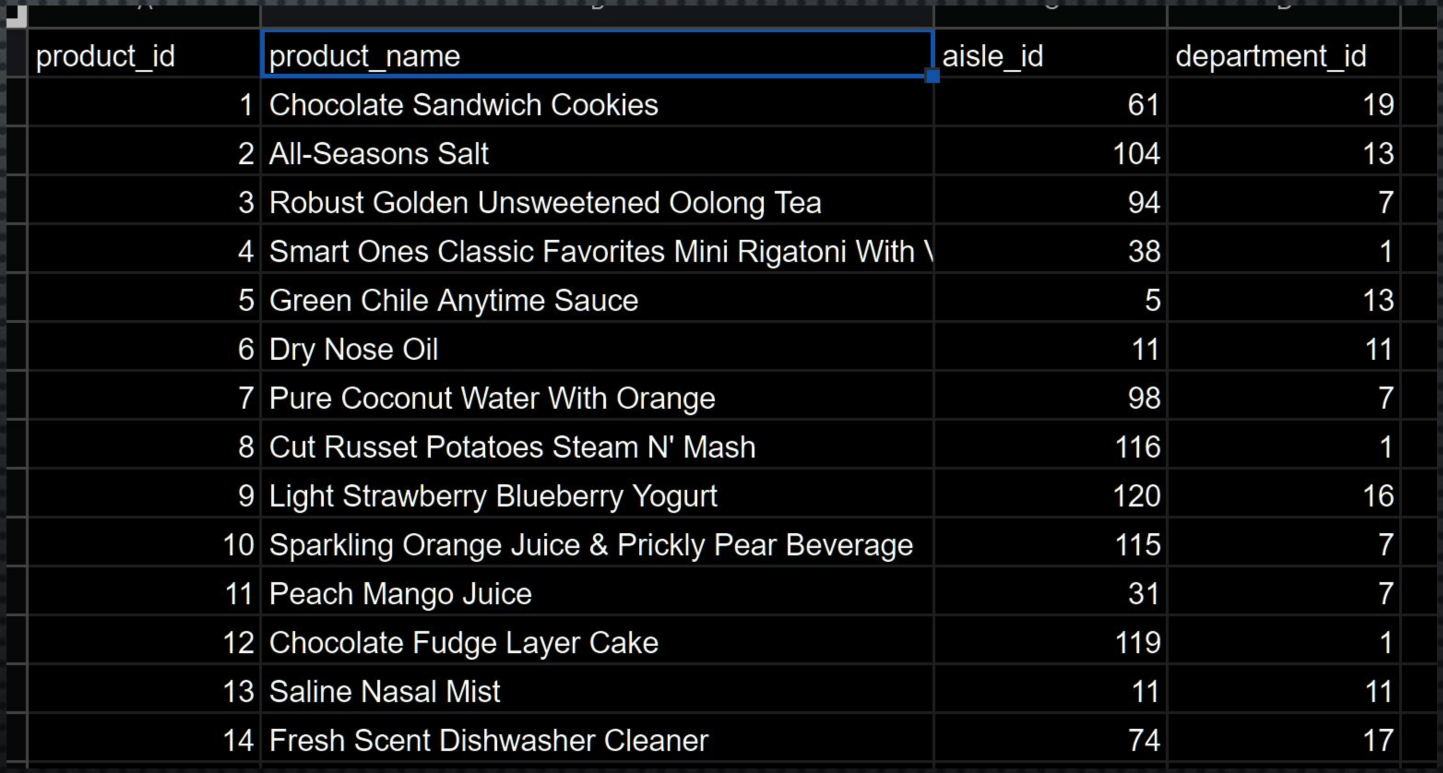
For Users:

1. Save time and money from going to grocery stores
2. Make the shopping experience better

DATASET

Instacart online grocery shopping dataset

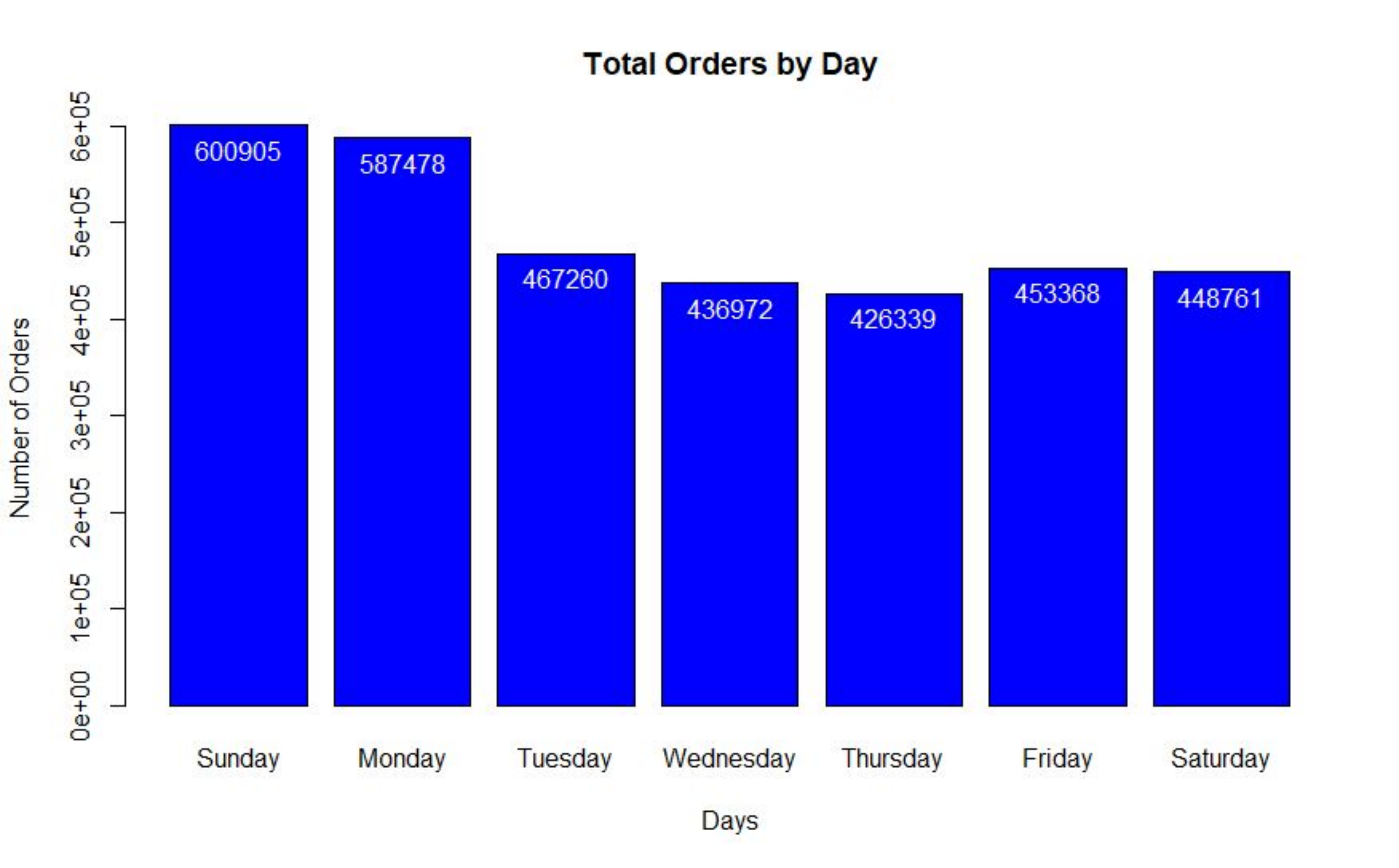
- Over 3 million orders
- Nearly 50,000 different products
- Files showcasing user habits



product_id	product_name	aisle_id	department_id
1	Chocolate Sandwich Cookies	61	19
2	All-Seasons Salt	104	13
3	Robust Golden Unsweetened Oolong Tea	94	7
4	Smart Ones Classic Favorites Mini Rigatoni With V	38	1
5	Green Chile Anytime Sauce	5	13
6	Dry Nose Oil	11	11
7	Pure Coconut Water With Orange	98	7
8	Cut Russet Potatoes Steam N' Mash	116	1
9	Light Strawberry Blueberry Yogurt	120	16
10	Sparkling Orange Juice & Prickly Pear Beverage	115	7
11	Peach Mango Juice	31	7
12	Chocolate Fudge Layer Cake	119	1
13	Saline Nasal Mist	11	11
14	Fresh Scent Dishwasher Cleaner	74	17

EXPLORATORY DATA ANALYSIS

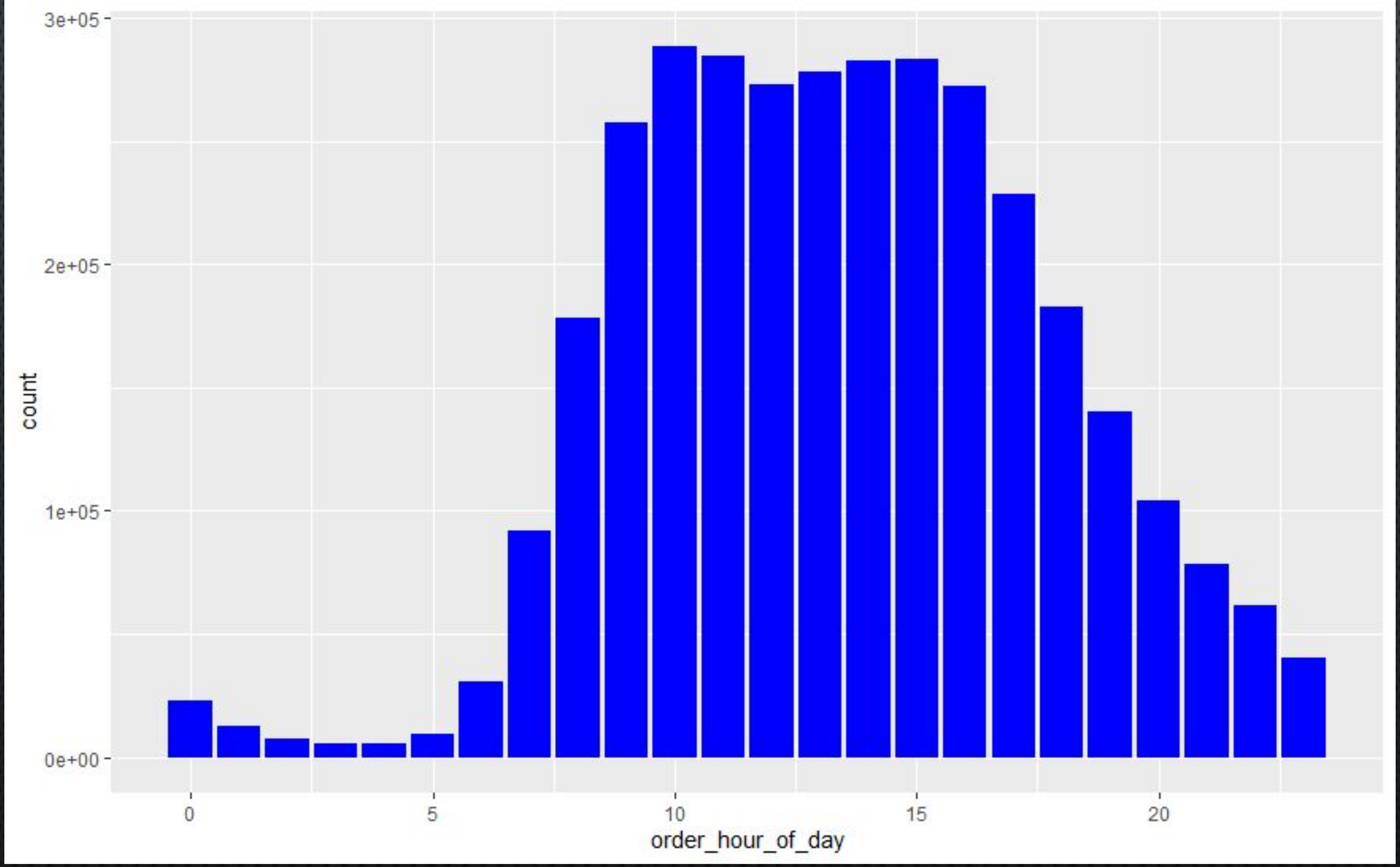
Number of Orders Per Day



Peak days;

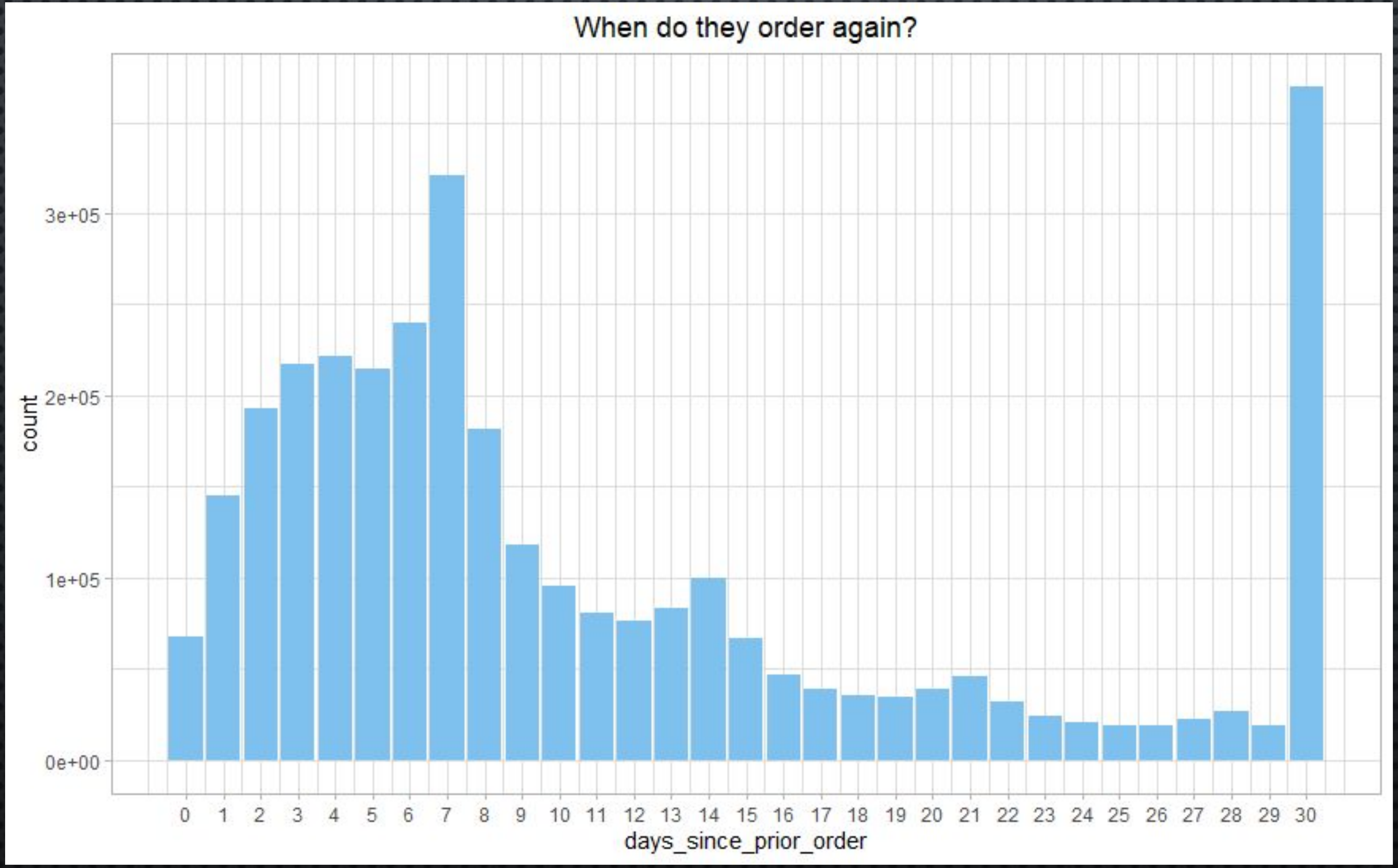
1. Sunday
2. Monday

NUMBER OF ORDERS PER HOUR



Peak hours;
9 A.M. – 5 P.M.

DAYS SINCE PREVIOUS ORDER



Bimodal distribution

First peak: 7 days

Second peak: 30 days

Possible insight:
- Customers typically ord
in weekly or monthly
amounts

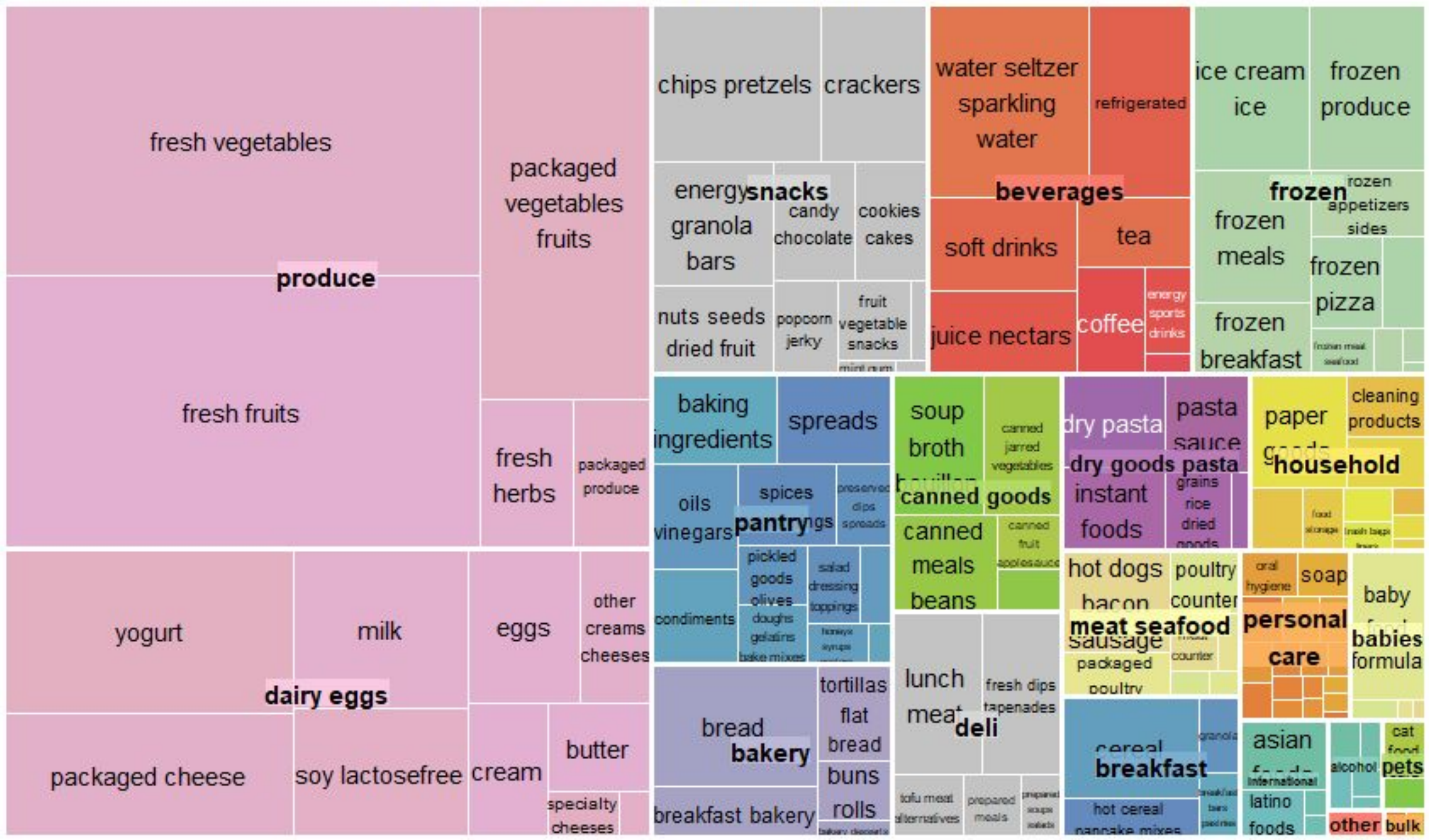
SALES PER DEPARTMENT

Number of sales
Reflected by relative
size of boxes

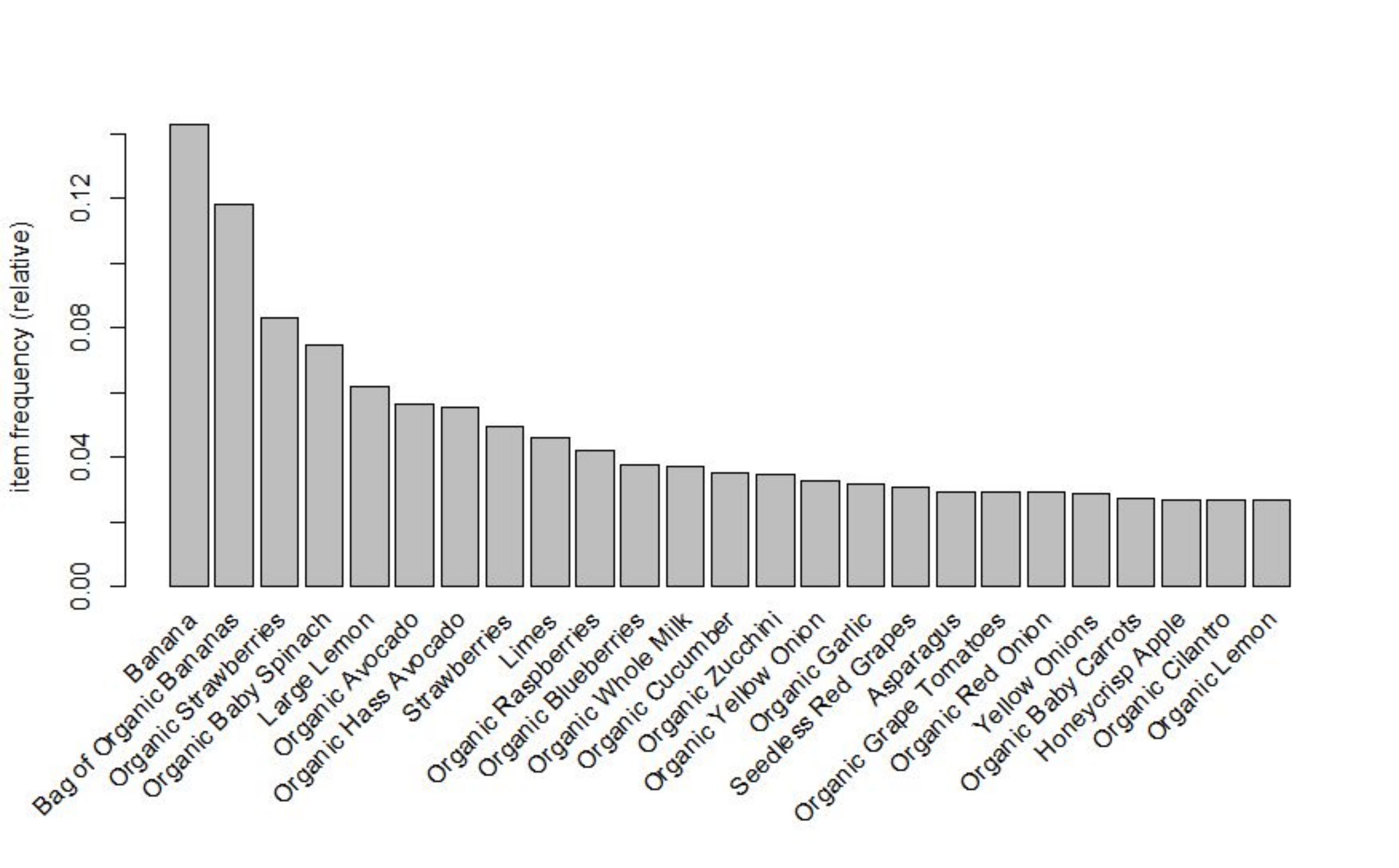
Best selling departments:

1. Produce
2. Dairy

Number of Sales from the Department/Aisle



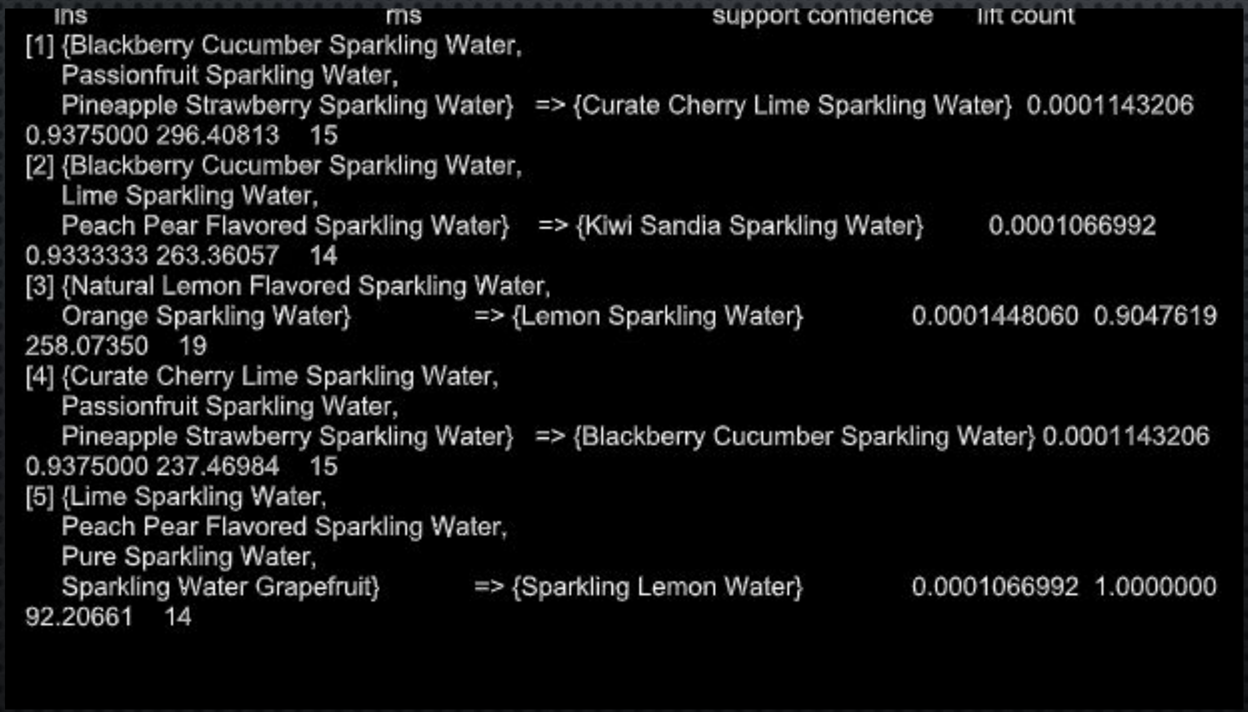
MOST FREQUENT ITEMS BOUGHT



ASSOCIATION RULES

GOAL: DEVELOP ASSOCIATION RULES BETWEEN ITEMS FOR FUTURE BUNDLING

ALGORITHM USED: APRIORI



```
    lhs                            rhs                              support confidence     lift count
[1] {Blackberry Cucumber Sparkling Water,
     Passionfruit Sparkling Water,
     Pineapple Strawberry Sparkling Water}   => {Curate Cherry Lime Sparkling Water}  0.0001143206
0.9375000 296.40813    15
[2] {Blackberry Cucumber Sparkling Water,
     Lime Sparkling Water,
     Peach Pear Flavored Sparkling Water}    => {Kiwi Sandia Sparkling Water}        0.0001066992
0.9333333 263.36057    14
[3] {Natural Lemon Flavored Sparkling Water,
     Orange Sparkling Water}               => {Lemon Sparkling Water}           0.0001448060  0.9047619
258.07350    19
[4] {Curate Cherry Lime Sparkling Water,
     Passionfruit Sparkling Water,
     Pineapple Strawberry Sparkling Water}   => {Blackberry Cucumber Sparkling Water} 0.0001143206
0.9375000 237.46984    15
[5] {Lime Sparkling Water,
     Peach Pear Flavored Sparkling Water,
     Pure Sparkling Water,
     Sparkling Water Grapefruit}           => {Sparkling Lemon Water}           0.0001066992  1.0000000
92.20661    14
```

ASSOCIATION RULES REFRESHER

FORM: (PROD A -> PROD B)

SUPPORT = FREQ(A, B) / N

CONFIDENCE = FREQ(A, B) / FREQ(A)

Classification of the Users

Goal: optimize recommendations

How we did it: classify users into groups based on habits

Algorithm Used: K-means using Elbow method

Result: created 40 clusters from which purchasing patterns were identified

CLUSTERING THE USERS

In order to cluster the users, we considered these features:

HABITS:

- Hours of a day in which user places orders
- Day of week n which user places orders
- order interval (TIME WHEN LAST ORDER IS PLACED)
- total number of orders placed

USER PREFERENCES:

- NAMES OF THE PRODUCTS CUSTOMERS BOUGHT
- Number of total products

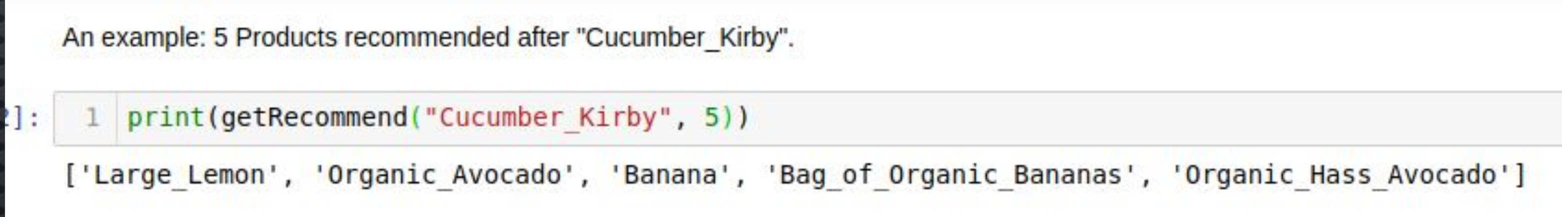
Popular Products in Each Group

Recommending Bundles of Items

BUNDLES OF ITEMS WERE RECOMMENDED AFTER USERS WERE CLUSTERED AND ASSOCIATION RULES FORMED

RECOMMENDED ITEM BASED ON BIGRAM FREQUENCY (ITEM 1, ITEM 2)

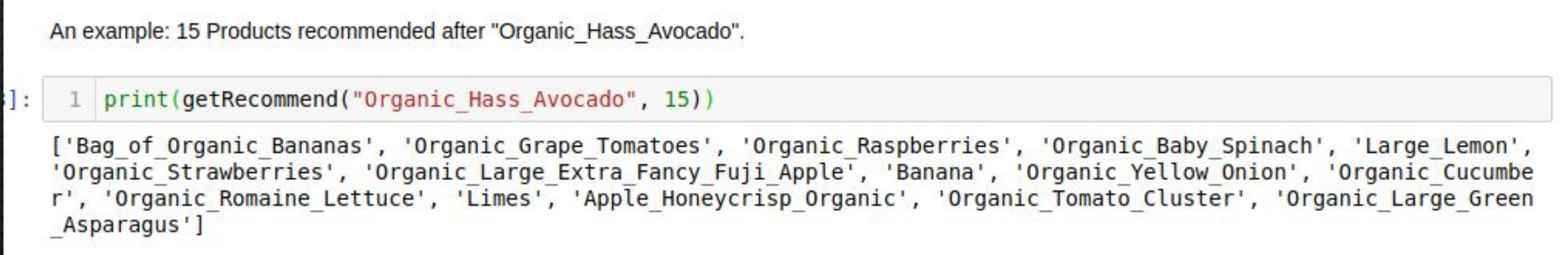
FORMED BUNDLES OF 5, 10, AND 15



An example: 5 Products recommended after "Cucumber_Kirby".

```
print(getRecommend("Cucumber_Kirby", 5))
```

```
['Large_Lemon', 'Organic_Avocado', 'Banana', 'Bag_of_Organic_Bananas', 'Organic_Hass_Avocado']
```



An example: 15 Products recommended after "Organic_Hass_Avocado".

```
print(getRecommend("Organic_Hass_Avocado", 15))
```

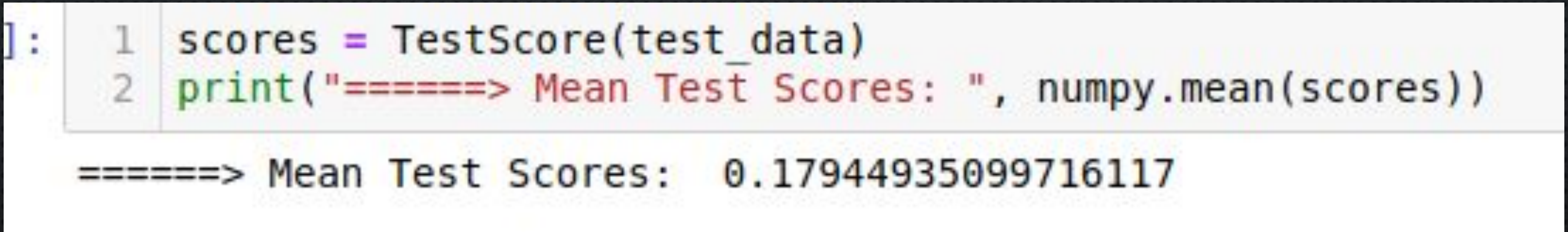
```
['Bag_of_Organic_Bananas', 'Organic_Grape_Tomatoes', 'Organic_Raspberries', 'Organic_Baby_Spinach', 'Large_Lemon',
'Organic_Strawberries', 'Organic_Large_Extra_Fancy_Fuji_Apple', 'Banana', 'Organic_Yellow_Onion', 'Organic_Cucumbe
r', 'Organic_Romaine_Lettuce', 'Limes', 'Apple_Honeycrisp_Organic', 'Organic_Tomato_Cluster', 'Organic_Large_Green
_Asparagus']
```

Results

Process:

1. conducted accuracy measures for recommendation bundles of 5, 10, and 15 items
2. Averaged accuracies for final result

Average accuracy: 17.94%



```
scores = TestScore(test_data)
print("======> Mean Test Scores: ", numpy.mean(scores))
```

```
======> Mean Test Scores:  0.17944935099716117
```