Instacart 3 Million Order Challenge

Understand the users behind every click

Team: London, Jonathan, Saurabh, Meng

Intro to Instacart

- Grocery delivery service
- Personal shoppers shop in various retails
- Can be delivered same-day
- Slightly more expensive than in-store prices and require delivery fees

Intro to Our Project

- Goal: understand user behavior and segmentation through order history on Instacart
- Data:
  - Orders (include train and test data):
    - What time of the day order is placed
    - What day of the week order is placed
    - Days since prior order
    - If the item in the order is a reordered product
  - Departments ID - 21 departments
  - Aisle ID - 134 aisles
  - Products ID - 50k products

Today’s Agenda

- Descriptives
  - Tableau
- Exploratory & modeling
  - Python pandas, numpy, scikit-learn
- Q&A

Descriptives

Distribution within Datasets

In the dataset, we have 3,214,874 orders from 206,209 users. Among the users, 131,209 users' data were used to train our model but only 75,000 users' data were used to test our model.

Distribution of Orders

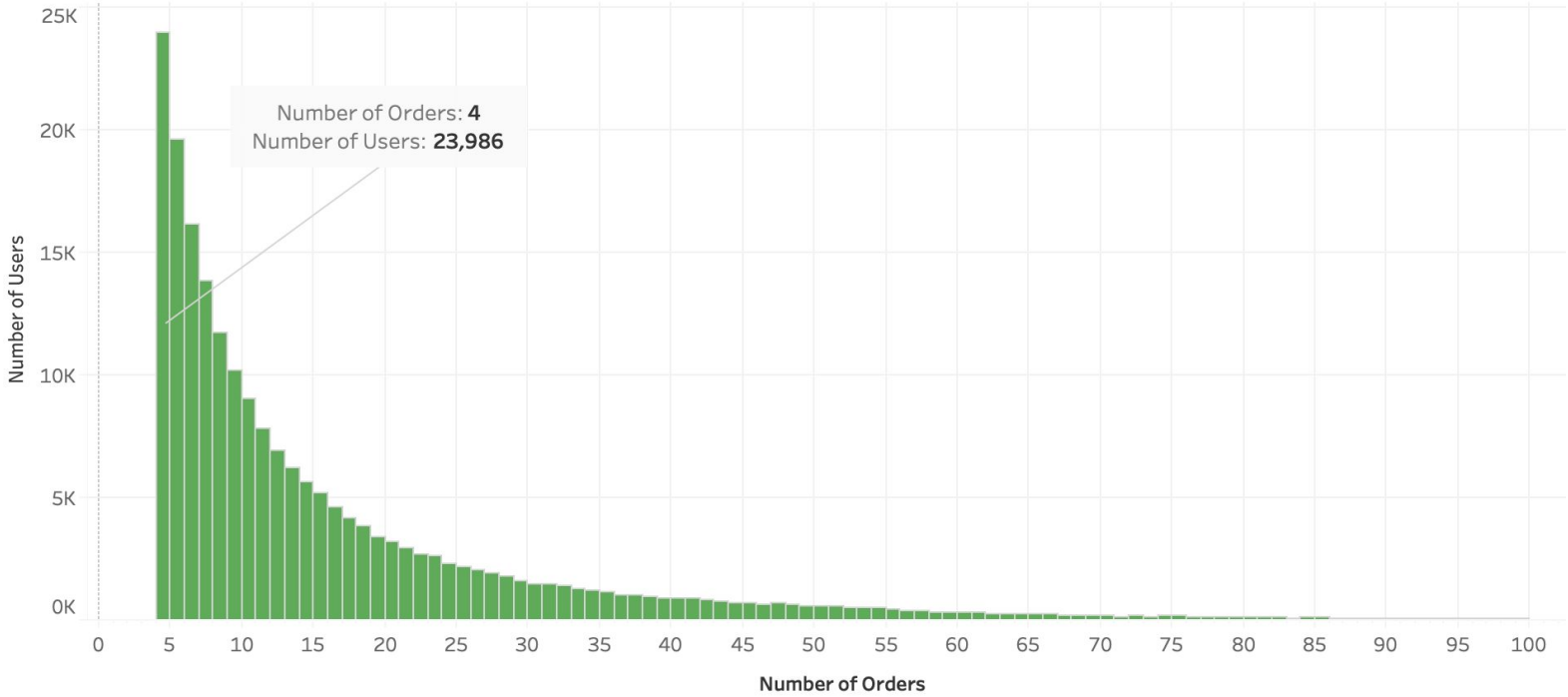
Eval Set	
prior	3,214,874
test	75,000
train	131,209

Distribution of Users

Eval Set	
prior	206,209
test	75,000
train	131,209

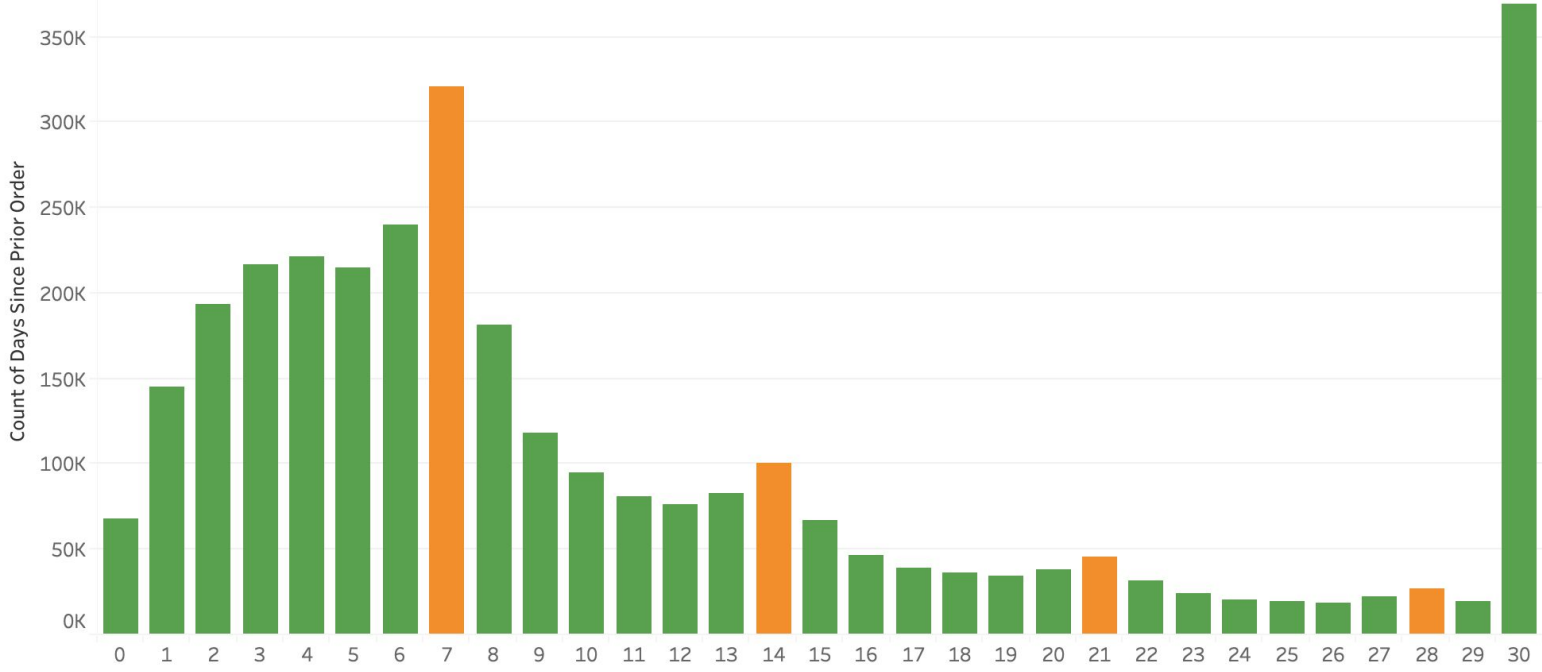
Order Summary

Amongst the 3,214,874 orders, each user ordered 4-100 times in the past. Average number of orders per user placed is ~17 times. Most people ordered 4 times in the dataset.



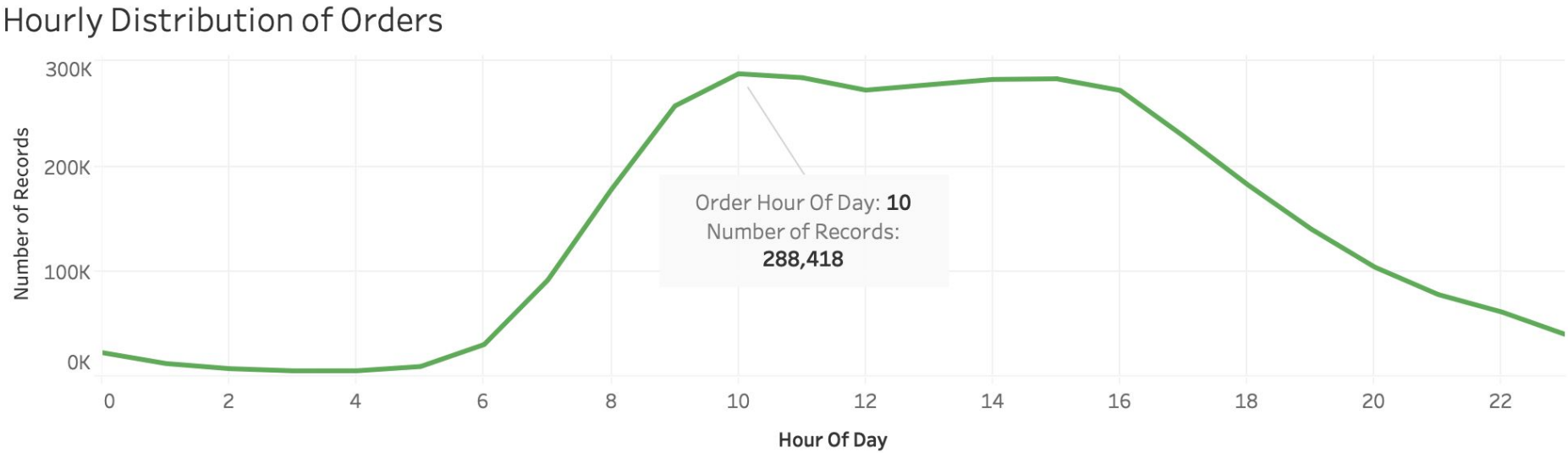
How Frequently Did People Reorder?

Most people tend to reorder a week later after the first order, and there's also a pattern of reordering every 7 days (day 14, 21, 28). A large group of people reorder 30 or more days apart.



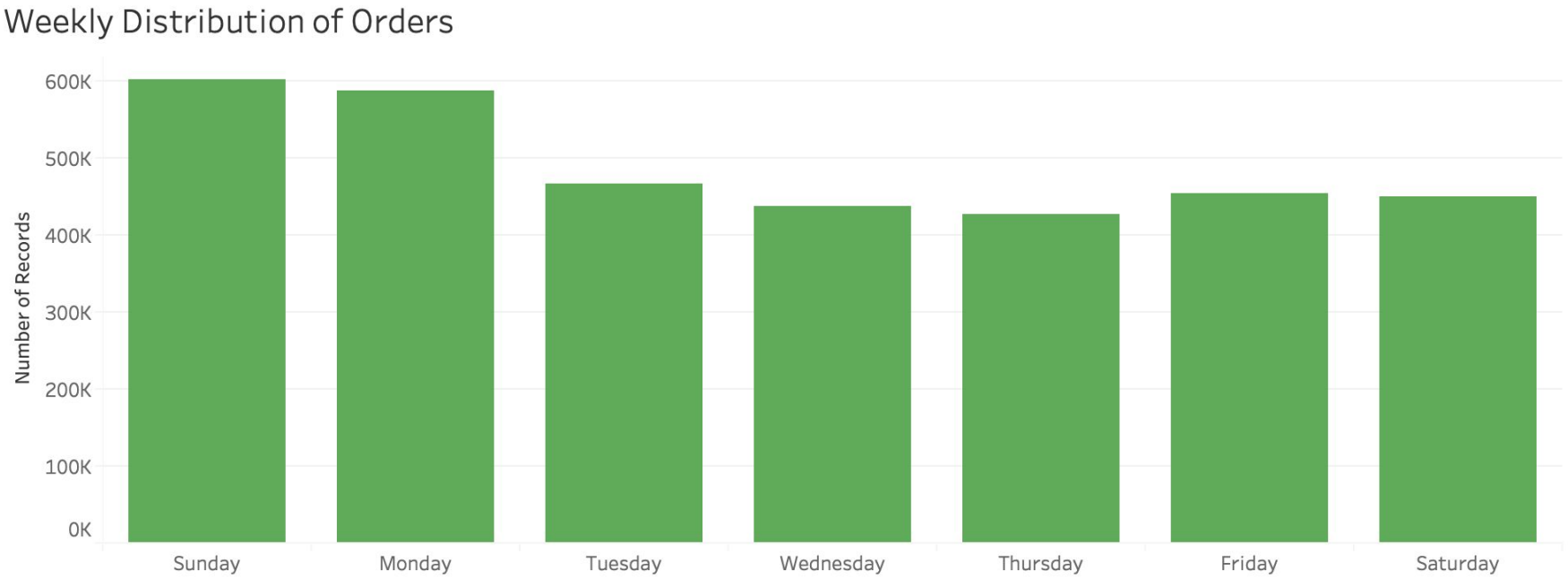
When Are Orders Placed Over the Day and Over the Week? (1 of 3)

Most orders are placed between 9am to 4pm.

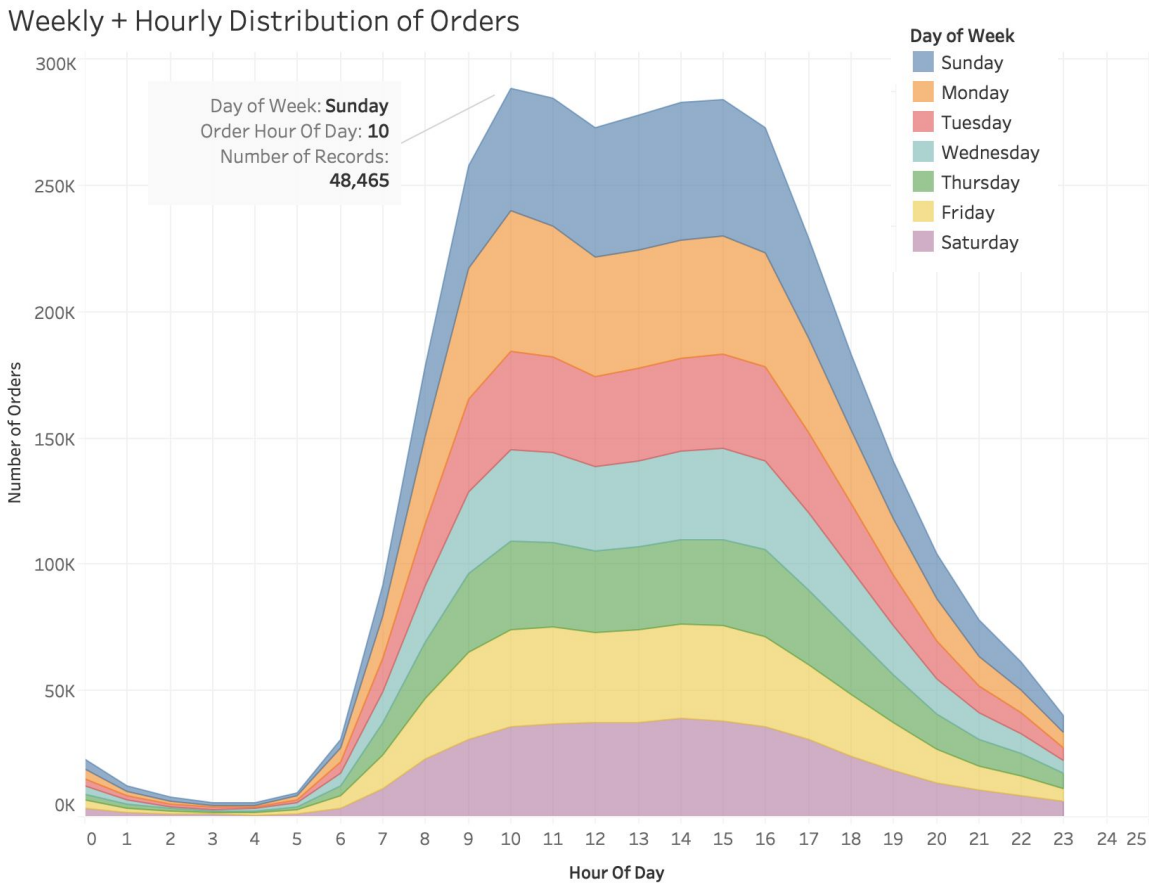


When Are Orders Placed Over the Day and Over the Week? (2 of 3)

Most popular days to place orders is Sunday and Monday



When Are Orders Placed Over the Day and Over the Week? (3 of 3)

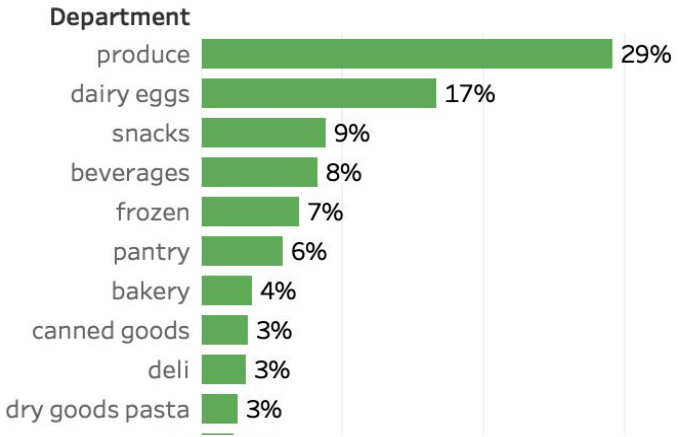


Sunday 10am and 3pm is most popular time to place orders. It is possible that people prefer to plan their week on Sunday and Monday, or it could be possible that due to delivery time, placing orders on Sunday or Monday can make sure that delivery is done by Friday.

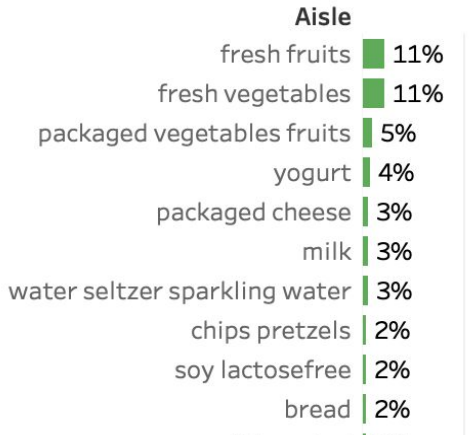
What are trending on Instacart?

There are 3 levels that describe ordered items in the dataset: departments, aisle, and product. Fresh fruits and veggies are the most popular items ordered. This data explains the pattern of shop every 7 days found in previous analysis. Produce and dairy are perishables that require shopping at least every 7 days for users to have fresh food available.

Most Popular Departments



Most Popular Aisles

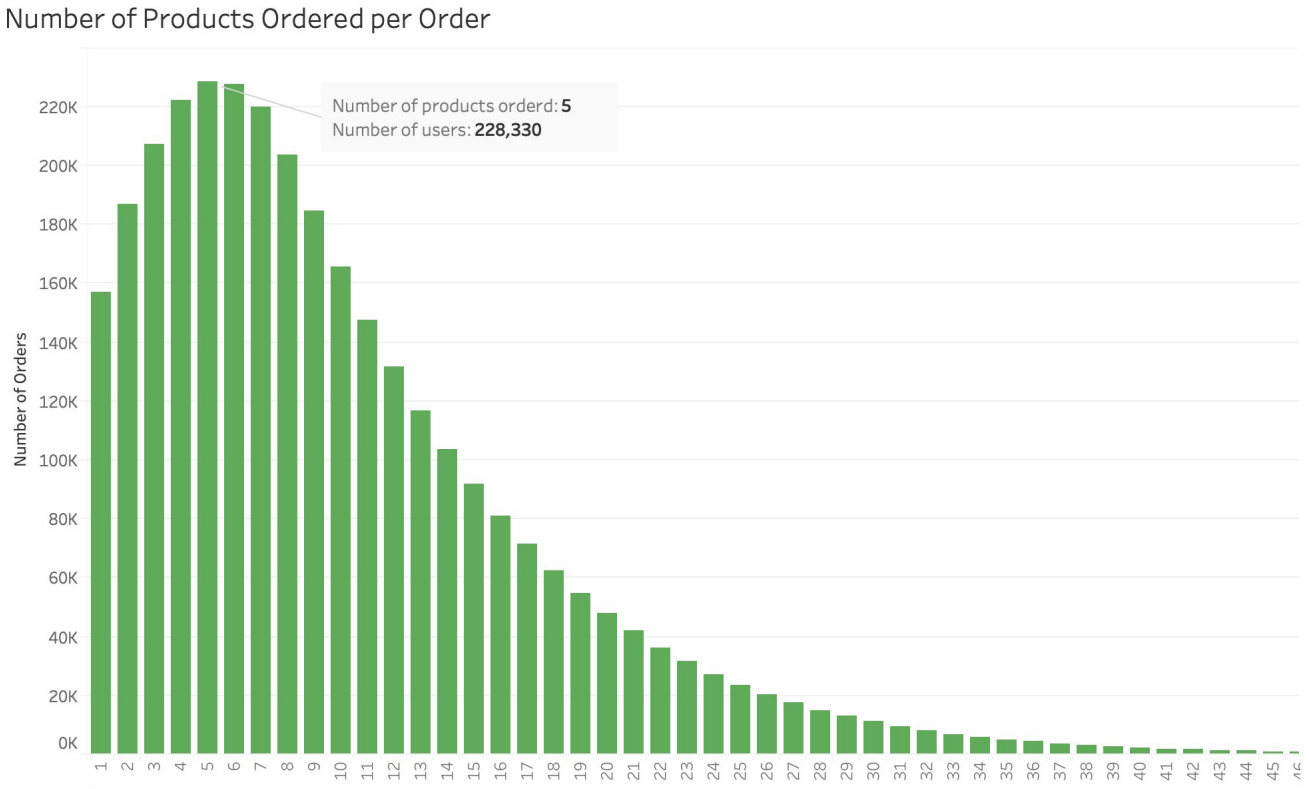


Most Popular Products Ordered

Product Name	
Banana	1.5%
Bag of Organic Bananas	1.2%
Organic Strawberries	0.8%
Organic Baby Spinach	0.7%
Organic Hass Avocado	0.7%
Organic Avocado	0.5%
Large Lemon	0.5%
Strawberries	0.4%
Limes	0.4%

How many products do people order per order?

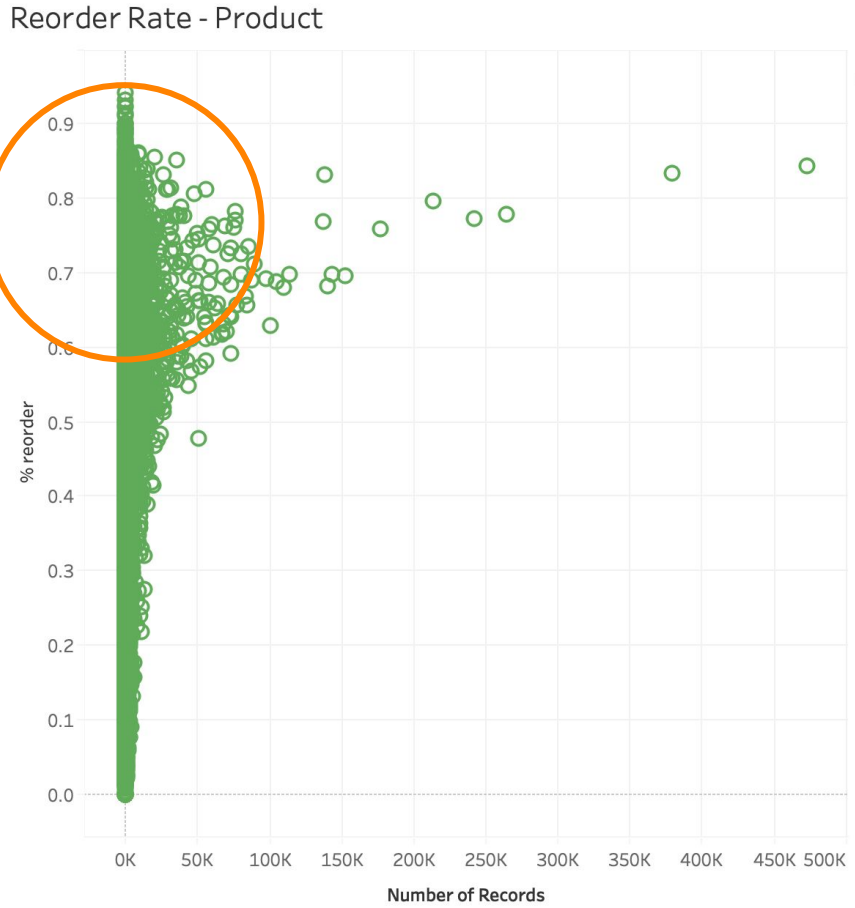
Most users order 5 products per order



How many aisles do people order from per order?

Again 5 is the magic number that most people order from 5 aisles per order

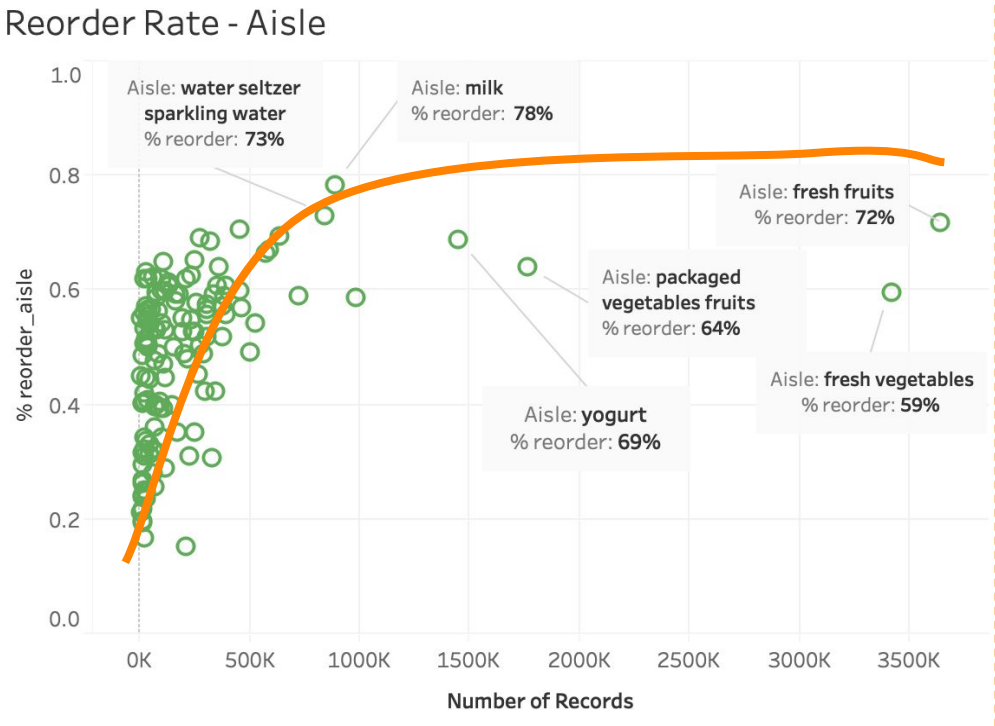
Reorder rate of product

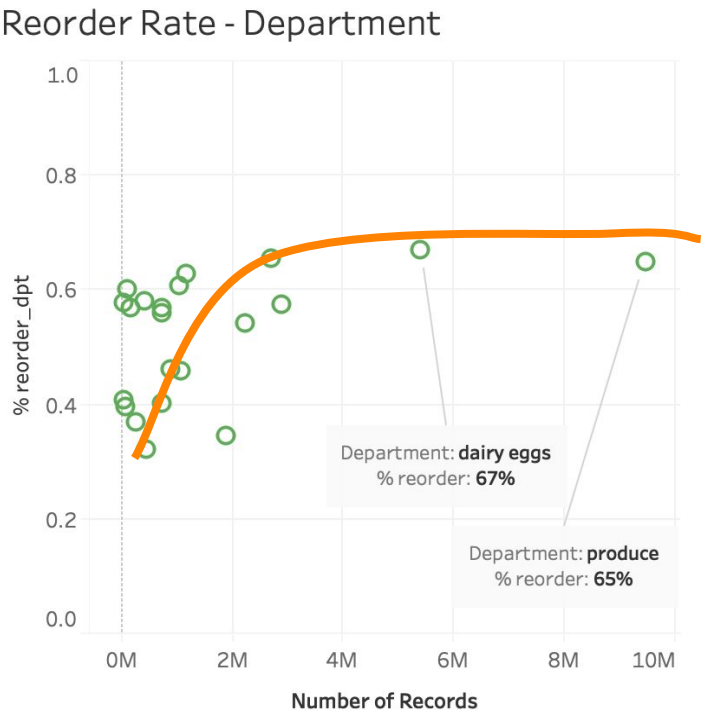


Only 59% or products have been reordered. There's correlation between the times of a product being ordered and the times of a product being reordered, but the relationship is not linear (e.g., the product has the highest reorder rate may not be the most ordered product). This is crucial to building models for predicting products ordered in the future as linear model may not give the best accuracy.

Reorder rate of aisle and departments

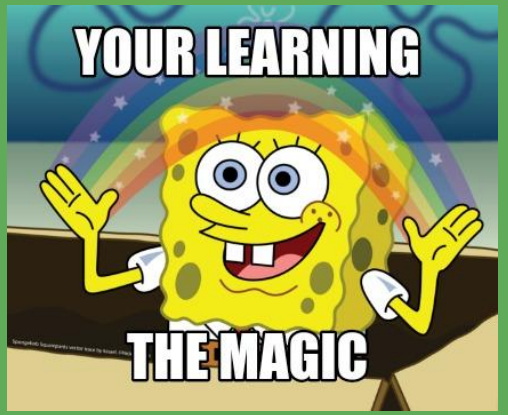
Aisle and departments has a more clean logarithmic trendline. We decided to focus on aisle as it has less noise but still doesn't lose too much information





Exploratory and Modeling

Do users have different preference over which aisle to order from?



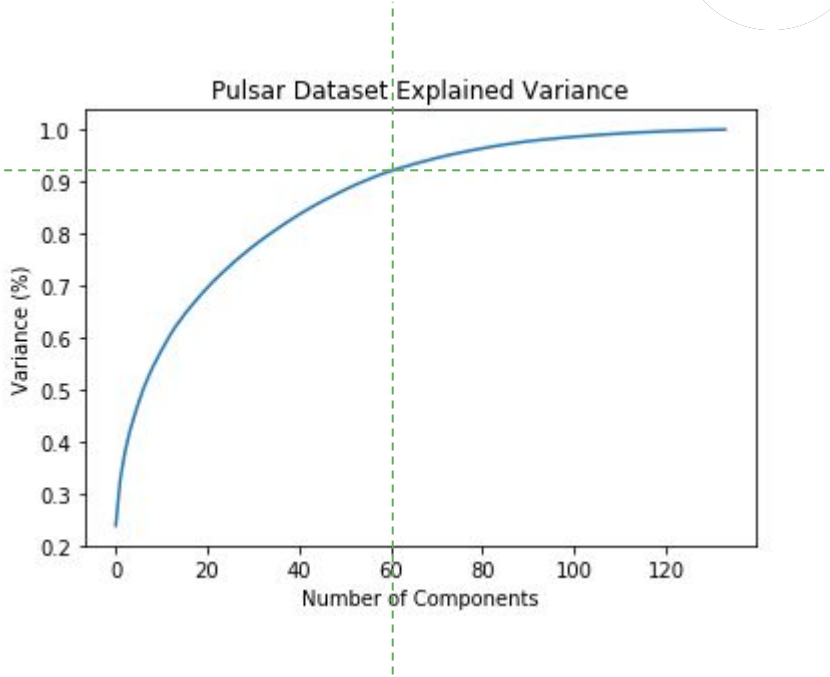
Data Structure

user_id	aisle_1	aisle_2 ….	aisle_134
1	0	5 ...	0
...	...	...	...
131,209	3	0 ...	3

How many times ordered from certain aisle in the training dataset

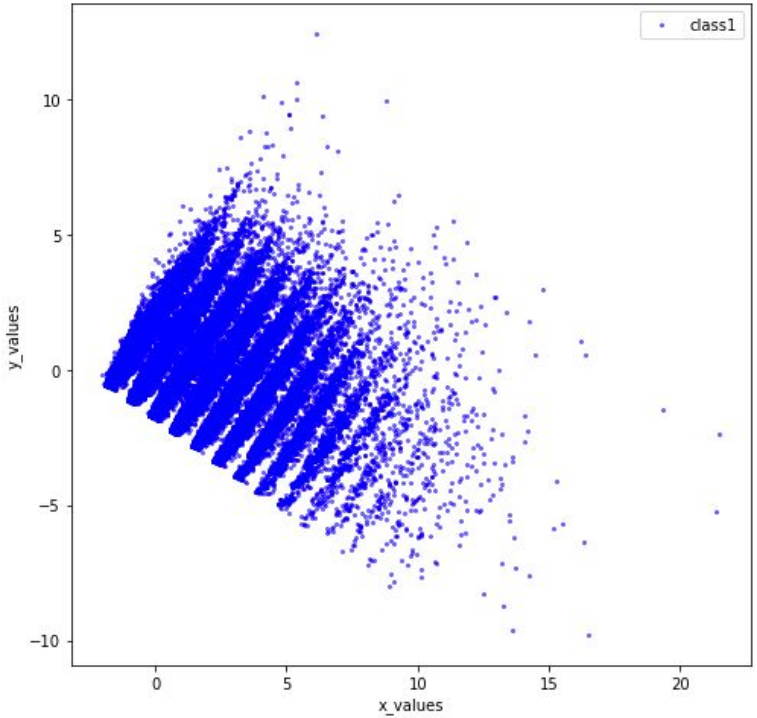
Principal Components Analysis (PCA)

- Effectively reduce the number of dimensions/features to be analyzed down to smaller set but still contains most of the information in the large set
- We reduced from 134 aisle down to 60 but still can explains ~90% variance of the original dataset

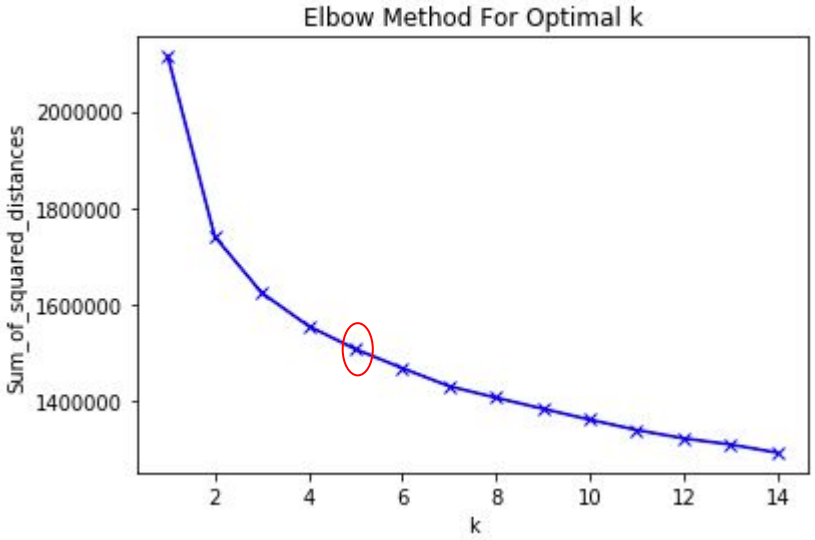


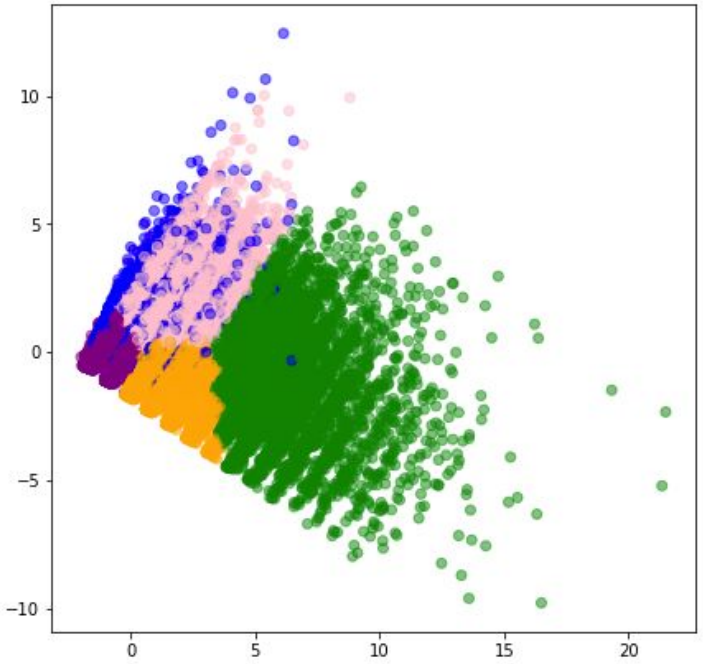
K-Means Clustering (1 of 2)

- Scatter plot of a pair of components (PC1, PC2) that is representative of 32% of the original dataset for visualization purpose only
- Ready for K-Means Clustering



K-Means Clustering (2 of 2)





How 4 clusters vary

Avg number of times an aisle was ordered from by a user

Cluster 1		Cluster2		Cluster 3		Cluster 4		Cluster 5	
aisle	avg	aisle	avg	aisle	avg	aisle	avg	aisle	avg
fresh vegetables	2.9	yogurt	3.3	fresh fruits	0.4	fresh vegetables	6.1	fresh fruits	3.2
fresh fruits	1.0	fresh fruits	1.4	water	0.3	fresh fruits	3.3	packaged vegetables fruits	1.2
packaged vegetables fruits	0.7	fresh vegetables	0.8	packaged vegetables fruits	0.3	packaged vegetables fruits	1.6	fresh vegetables	1.0
packaged cheese	0.4	packaged vegetables fruits	0.7	packaged cheese	0.2	yogurt	0.7	yogurt	0.5
yogurt	0.3	packaged cheese	0.6	chips pretzels	0.2	packaged cheese	0.6	packaged cheese	0.5
fresh herbs	0.3	chips pretzels	0.5	fresh vegetables	0.2	fresh herbs	0.6	milk	0.4
milk	0.2	milk	0.5	milk	0.2	frozen produce	0.4	chips pretzels	0.3

Persona Analysis (1 of 2)

Cluster2	Cluster 5	
aisle	aisle	avg
yogurt	fresh fruits	3.2
fresh fruits	packaged vegetables fruits	1.2
fresh vegetables	fresh vegetables	1.0
packaged vegetables fruits	yogurt	0.5
packaged cheese	packaged cheese	0.5
chips pretzels	milk	0.4
milk	chips pretzels	0.3

Cluster 3	
aisle	avg
fresh fruits	0.4
water	0.3
packaged vegetables fruits	0.3
packaged cheese	0.2
chips pretzels	0.2
fresh vegetables	0.2
milk	0.2

Clus

-
-
-
-

ogurt fa

te Instacart
tremely on
moderately
ring to yogu
gnificantly le
aby food!
al character
parents, pi
ss of grocer

Fruits lover

rate Instacart user
moderately on Inscart for
fruits, yogurt, and milk
paring to fruits, shop less
es but still more than yogurt
c
tial characteristics: young
ng professionals,
rately tech savvy, cook
times, or may rely on other
delivery service

Perso of 2)

Cluster 1	
aisle	avg
fresh vegetables	2.9
fresh fruits	1.0
packaged vegetables fruits	0.7
packaged cheese	0.4
yogurt	0.3
fresh herbs	0.3
milk	0.2

Cluster 4	
aisle	avg
fresh vegetables	6.1
fresh fruits	3.3
packaged vegetables fruits	1.6
yogurt	0.7
packaged cheese	0.6
fresh herbs	0.6
frozen produce	0.4

Clus

-
-
-
-
-

ter

raditional

es, esp

rking

ely tech savvy,

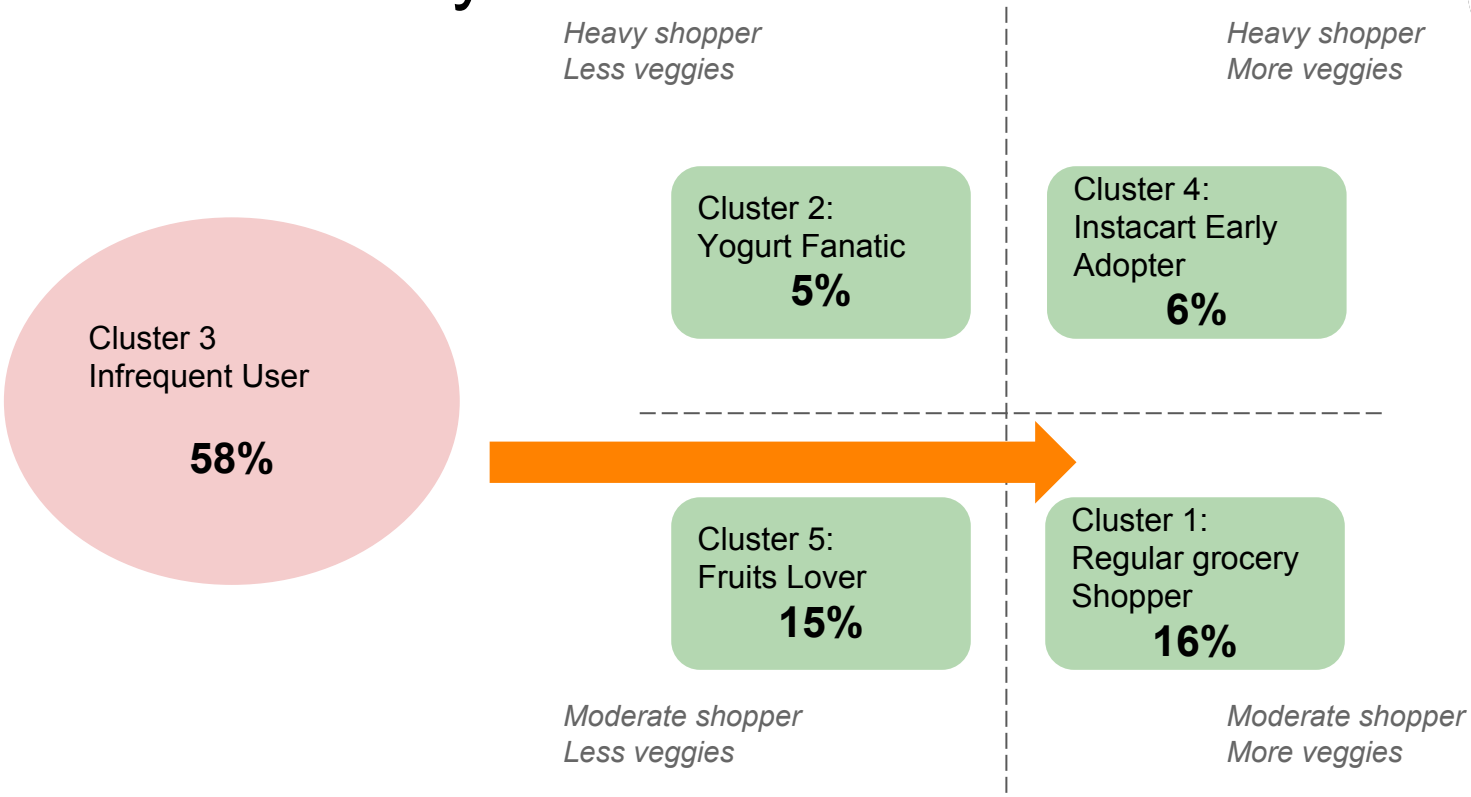
Cluster 1 Reg

- Moderate
- Rely mode in addition
- Shop fruits
- Lactose in
- Potential profession savvy, coo

opping

ely tech

Clusters Summary



Next steps?

- General:
  - Analyze internal user demographics data to better understand persona
- Infrequent users:
  - Conducting qualitative studies (focus group or in-person interviews) with infrequent users to understand their use cases and why Instacart is not appealing
  - Draft plans to convert infrequent users to regular grocery shoppers (e.g., more promotions, lower fees, including more popular retailers, marketing on freshness of food etc)
- Yogurt fanatics:
  - Conducting qualitative studies to understand their personas better, especially about their marital status (why baby food is ordered more frequently than other user groups)

Thank you

Questions?