BUILDING A RECOMMENDER SYSTEM

Caixuan Sun

Springboard Capstone Project 2
March 2019

RECOMMENDER SYSTEMS ARE EVERYWHERE

THE TASK:

Build a simple recommender system using Instacart open source data

DATA DESCRIPTION

- A set of six relational files describing customers' orders over time
- A sample of over **3 million orders** from more than **200,000 users**
- *orders.csv*

  - each row represents an unique order

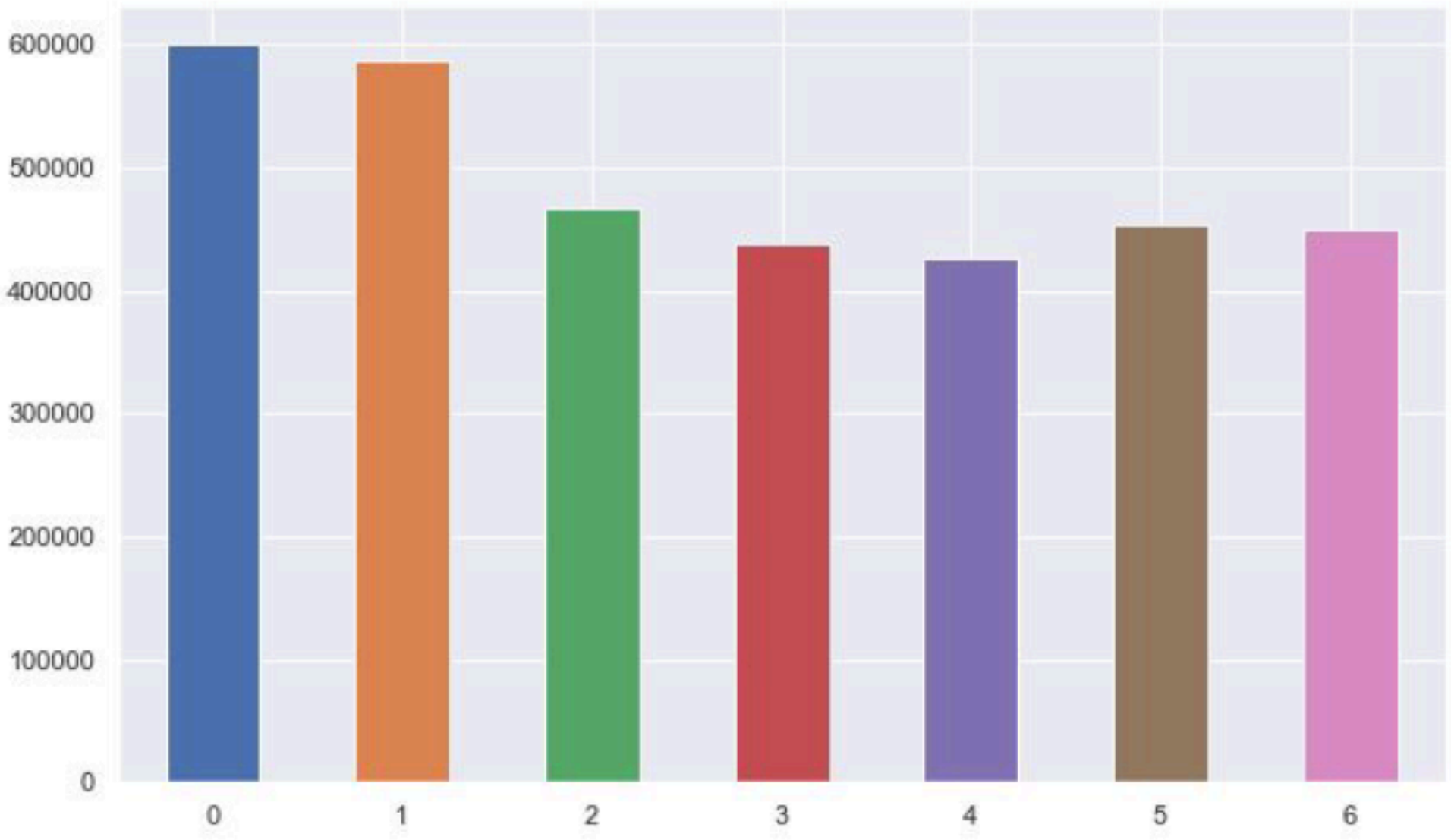
  - columns: user_id, the order sequence number, day of week and hour of day the order placed, days since the last order

- *order_products.csv*

  - products bought in each order, the sequence each product was added to cart, the product was reordered or not

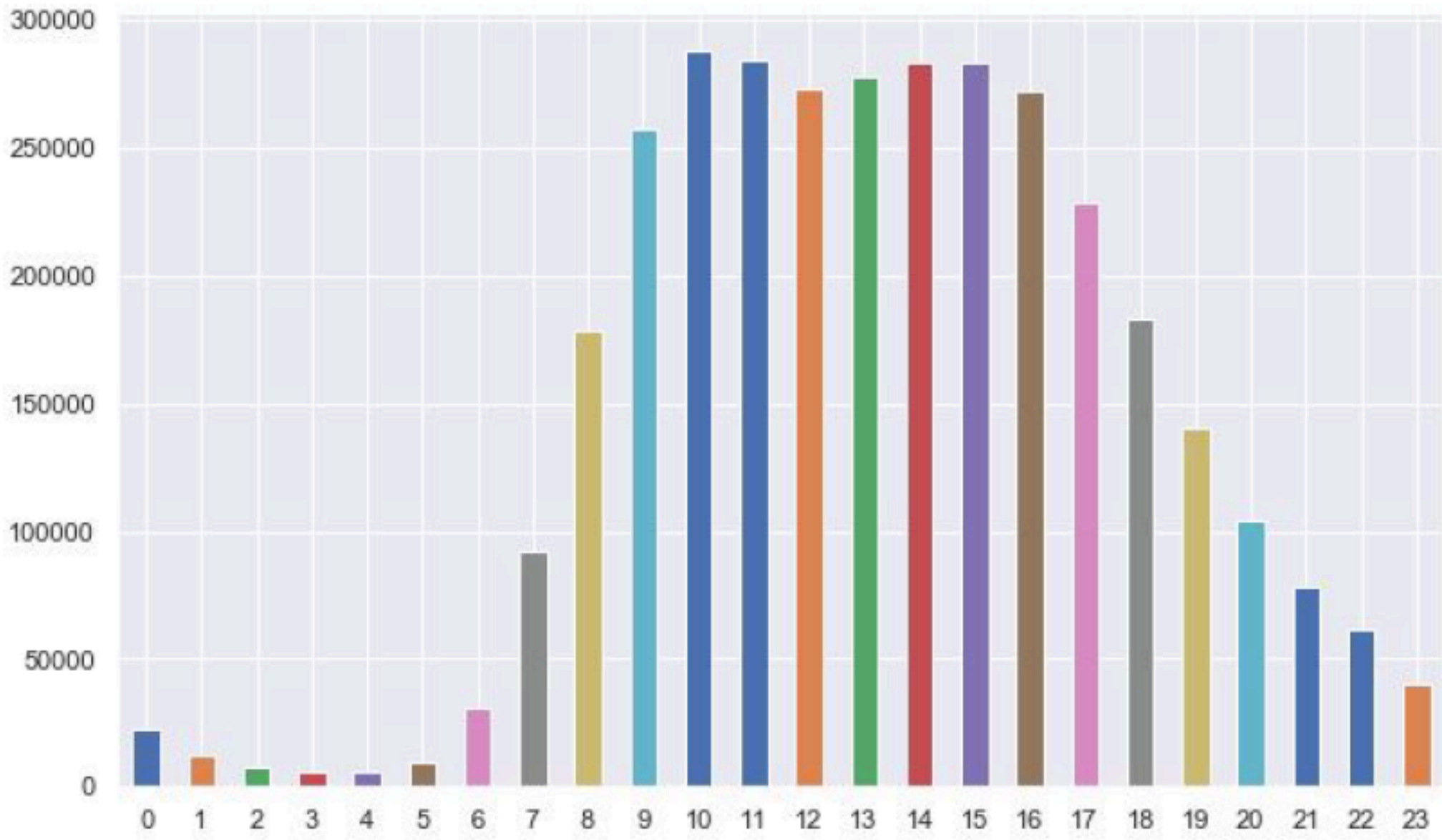
EDA: SHOPPING HABITS (1)

Figure 1. Distribution of Orders Over the Week



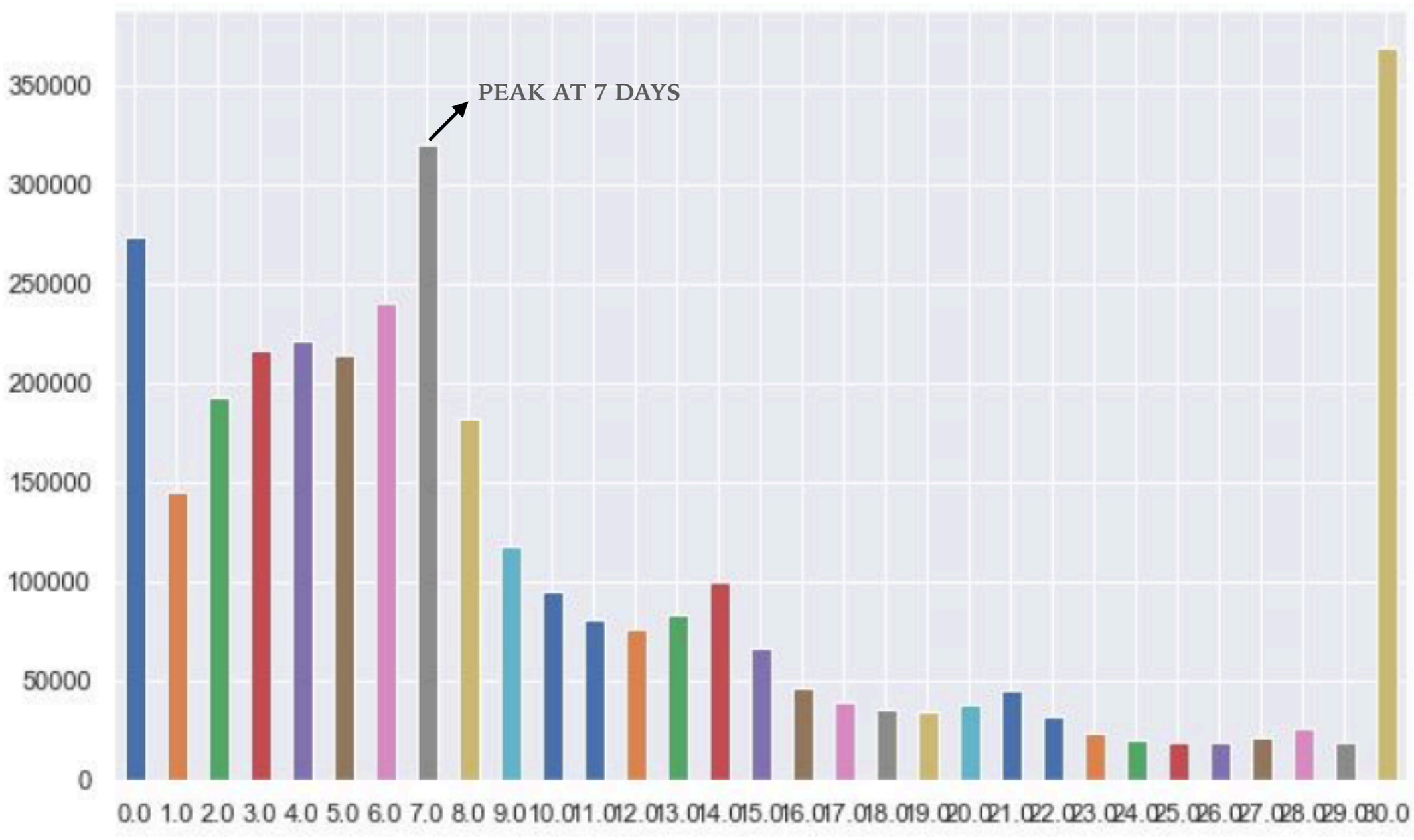
EDA: SHOPPING HABITS (2)

Figure 2. Order Distribution Over the Day



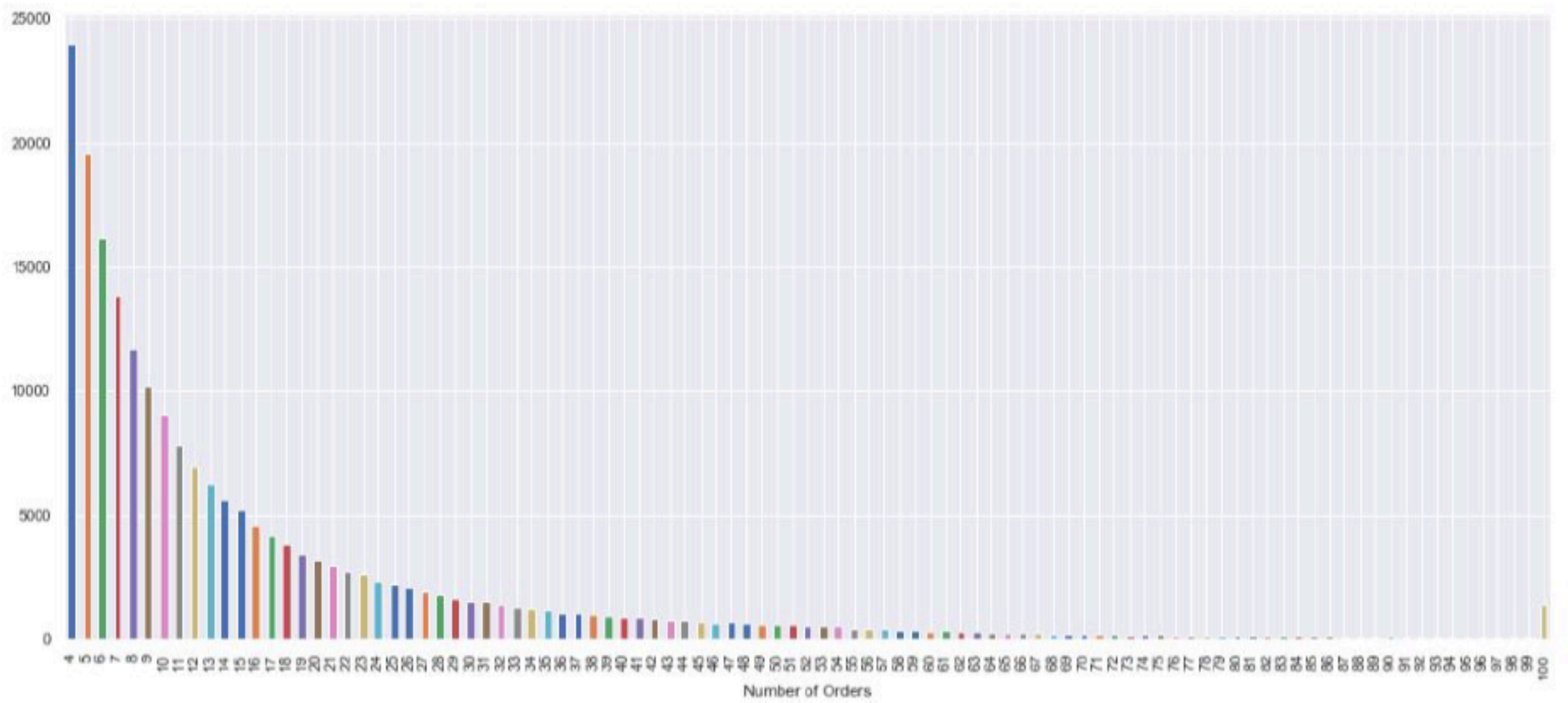
EDA: SHOPPING HABITS (3)

Figure 3. Distribution of Days Since Prior Order



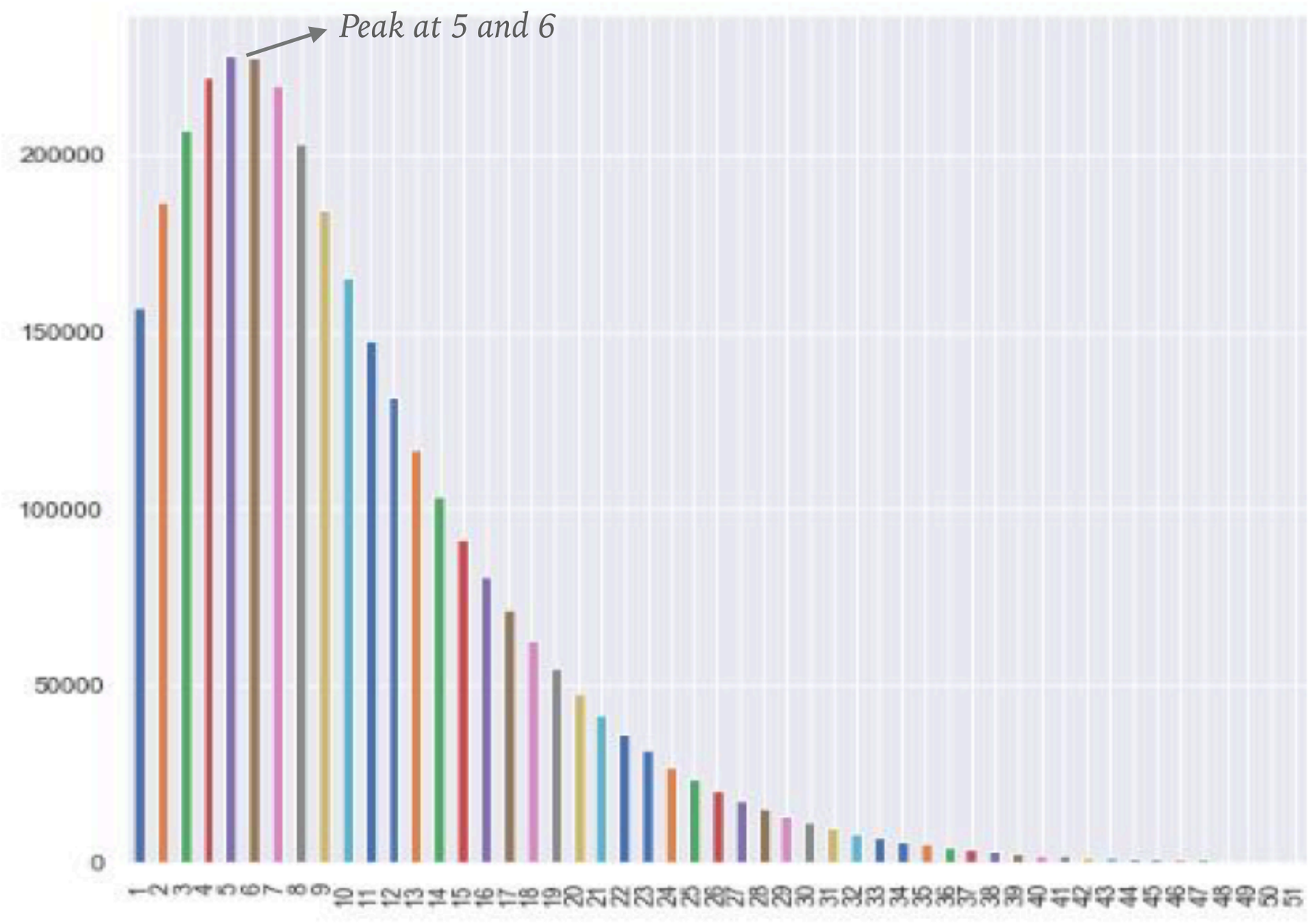
EDA: DISTRIBUTION OF ORDERS BY USERS ON INSTACART

Figure 4. Number of Orders Distribution



EDA: DISTRIBUTION OF NUMBER OF PRODUCTS IN EACH ORDER

Figure 5. Products per Order Distribution



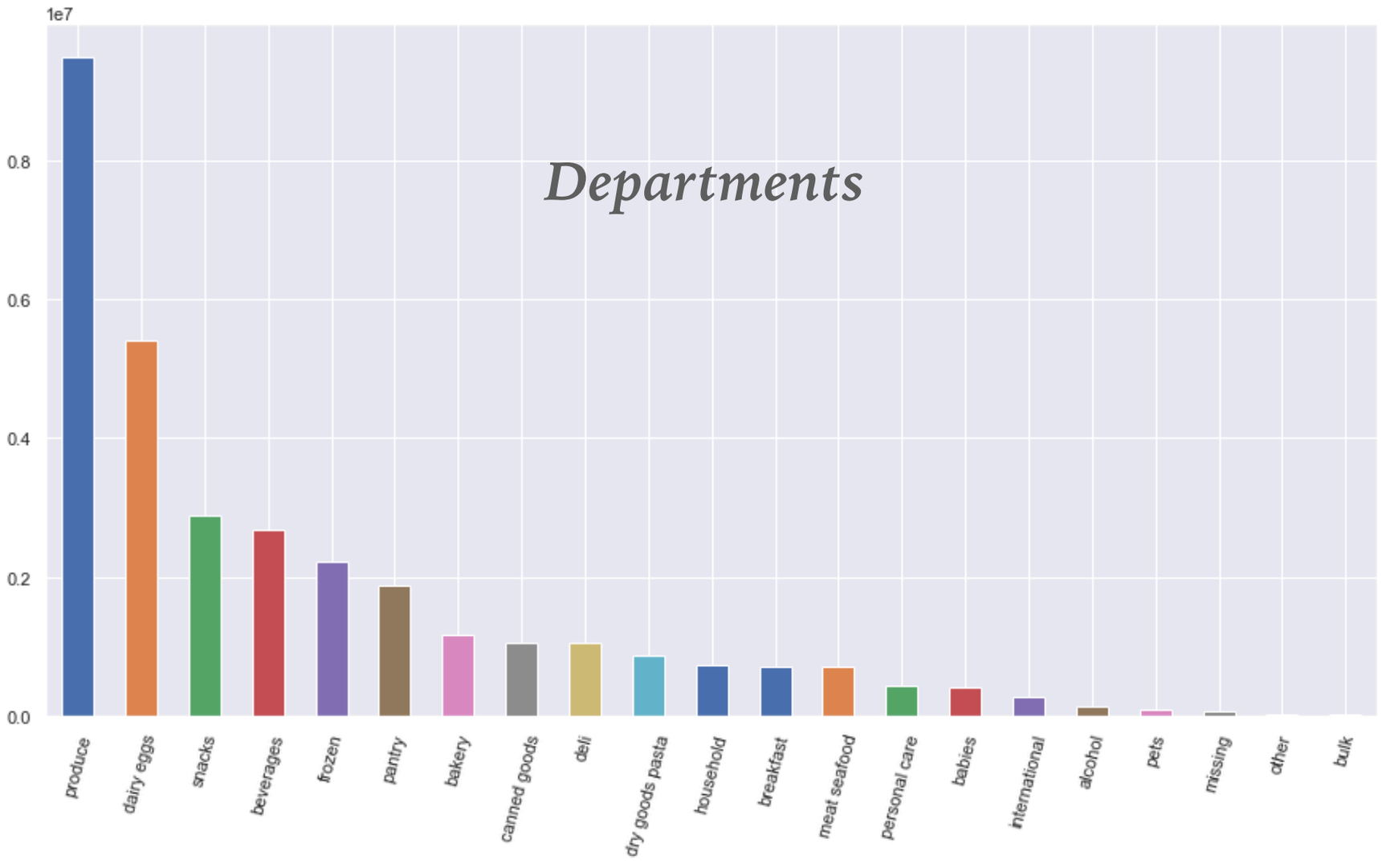
EDA: MOST POPULAR PRODUCTS

Banana	110916
Bag of Organic Bananas	78988
Organic Whole Milk	30927
Organic Strawberries	27975
Organic Hass Avocado	24116
Organic Baby Spinach	23543
Organic Avocado	22398
Spring Water	16822
Strawberries	16366
Organic Raspberries	14393
Sparkling Water Grapefruit	13733
Organic Half & Half	12676
Large Lemon	12316
Soda	11770
Organic Reduced Fat Milk	9885
Limes	9719
Half & Half	9528
Hass Avocados	9500
Organic Reduced Fat 2% Milk	9338
Raspberries	8885
Organic Fuji Apple	8762
Organic Blueberries	8740
Apple Honeycrisp Organic	8730
Organic Yellow Onion	8548
Unsweetened Almondmilk	8535

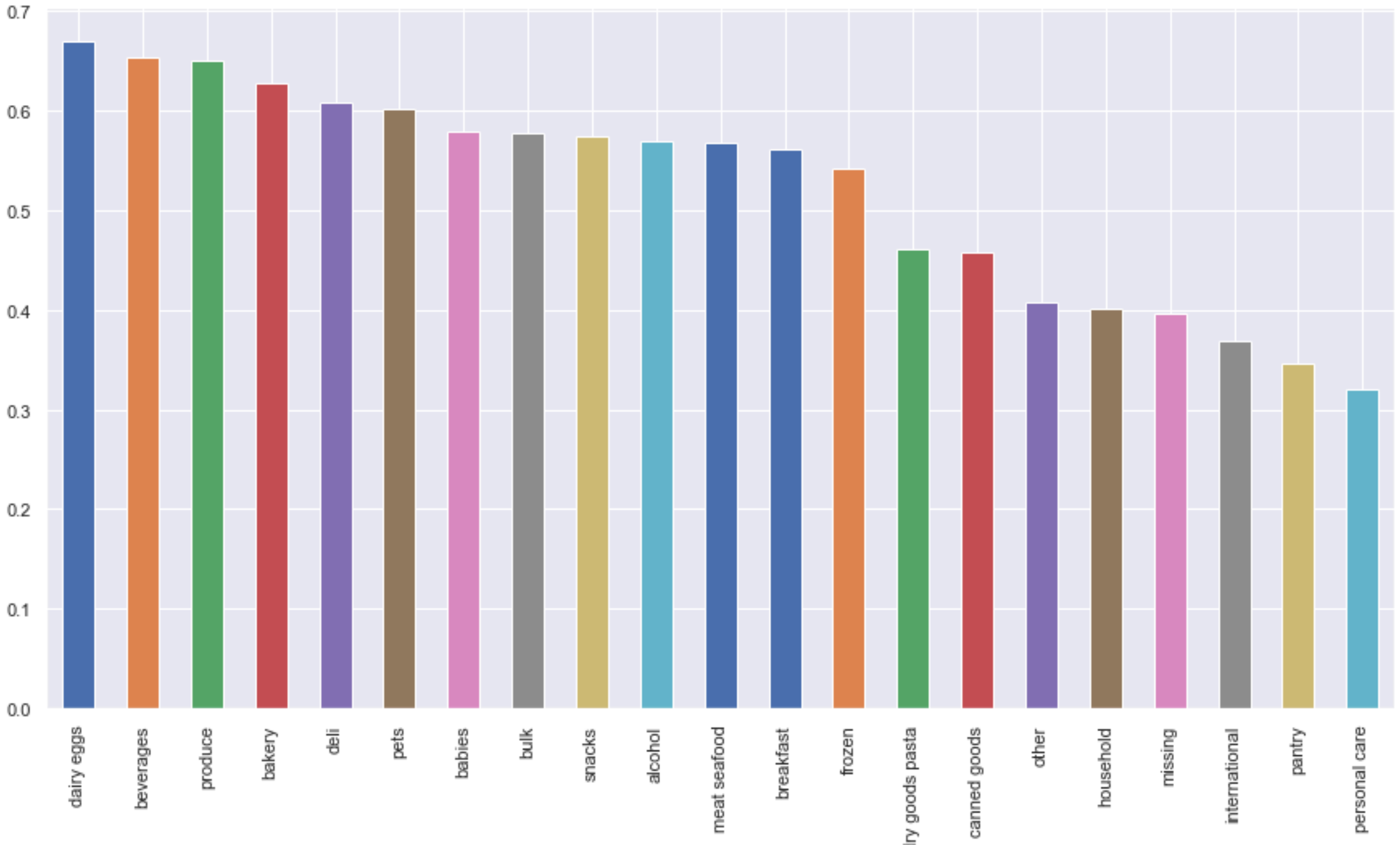
EDA: MOST POPULAR AISLES AND DEPARTMENTS

Aisles

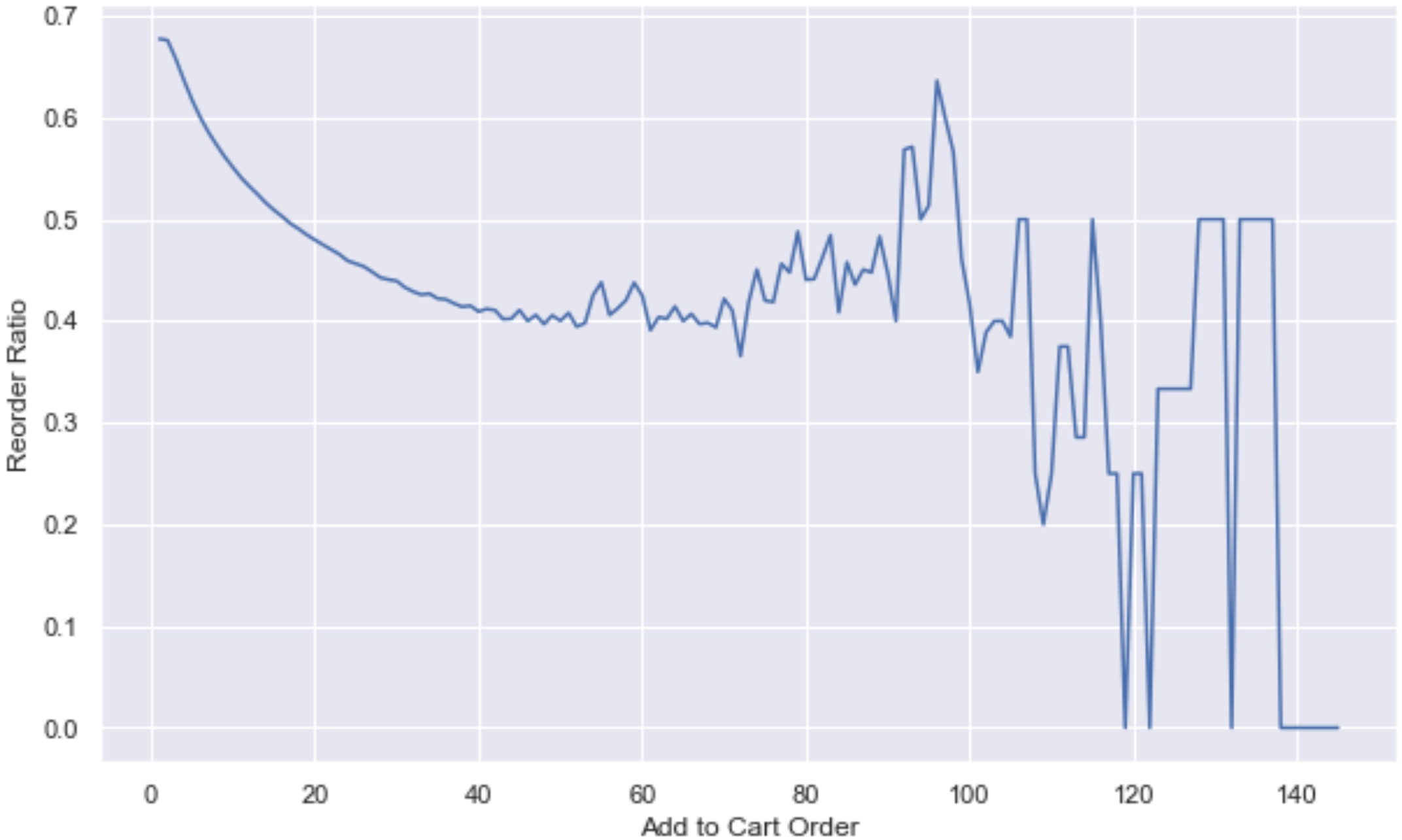
fresh fruits	3642188
fresh vegetables	3418021
packaged vegetables fruits	1765313
yogurt	1452343
packaged cheese	979763
milk	891015



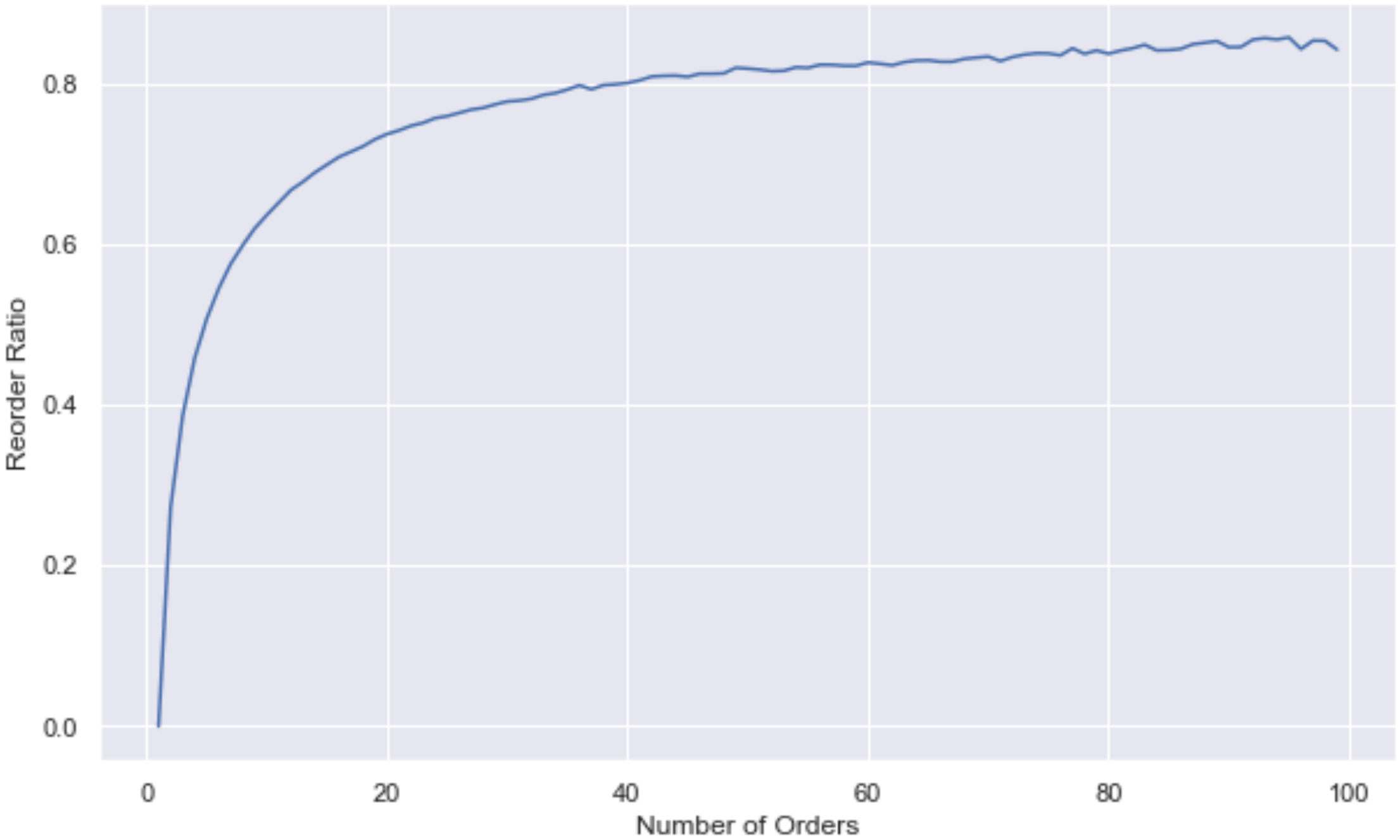
EDA: REORDER RATIO BY DEPARTMENT



EDA: REORDER RATIO VS ADD_TO_CART_ORDER



EDA: REORDER RATIO VS NUMBER OF ORDERS



MODELS USED IN THIS PROJECT

- Matrix Factorization

- Neighborhood-based Model

- Models used in this project:
  - *matrix factorization*
  - *neighborhood-based model*

- Evaluation metric: *recall@k*
  - *with all products included*
  - *without reordered products*

- Benchmark: *non-personalized popularity model*

EVALUATION OF BENCHMARK MODEL

	with reordered products	without reordered products
recall@10	0.070	0.027
recall@20	0.096	0.044
recall@50	0.154	0.080

User_id 100 actually bought:

	product_id	product_name
21136	21137	Organic Strawberries
21615	21616	Organic Baby Arugula
24851	24852	Banana
26368	26369	Organic Roma Tomato
27343	27344	Uncured Genoa Salami
38546	38547	Bubblegum Flavor Natural Chewing Gum
38688	38689	Organic Reduced Fat Milk
48627	48628	Organic Whole Wheat Bread

User_id 100 got recommended:

	product_id	product_name
13175	13176	Bag of Organic Bananas
16796	16797	Strawberries
21136	21137	Organic Strawberries
21902	21903	Organic Baby Spinach
24851	24852	Banana
26208	26209	Limes
27844	27845	Organic Whole Milk
47208	47209	Organic Hass Avocado
47625	47626	Large Lemon
47765	47766	Organic Avocado

MATRIX FACTORIZATION USING ALS: DATA PREPROCESSING

All Prior Orders → *Training* → User-item sparse matrix

Row: users

Col: items

Val: # of purchases

Most Recent Orders → *Test* → Dataframe

Row: users

Col: products bought in their most recent order

MF: MODEL PARAMETERS

- Alpha: used for the confidence matrix (input of model)

- Factors: the number of latent factors

- Regularization: to prevent overfitting

- Iterations: the number of ALS iterations to use when fitting data

MF: TUNING THE MODEL

- Alpha: [10, 15]
- Factors: [30, 40, 50]
- Regularization: [0.01, 0.1, 1.0]
- Iterations: [25, 50]

MF: MODEL EVALUATION

	with reordered products	without reordered products
recall@10	0.021	0.043
recall@20	0.035	0.072
recall@50	0.063	0.131

Non-personalized popularity model is performing better with reordered products than without. On the contrary, MF model is doing better when reordered products were excluded than included.

MF model is doing a better job in recommending new products to consumers.

NEIGHBORHOOD-BASED MODEL: DATA PREPROCESSING

All Prior Orders → *Training* → SFrame

Cols: user_id,
product_id,
# of purchases

Most Recent Orders → *Test* → Dataframe

Row: users

Col: products bought in their most recent order

NEIGHBORHOOD-BASED MODEL: EVALUATION

	Cosine		Jaccard	
	with reordered products	without reordered products	with reordered products	without reordered products
recall@10	0.018	0.039	0.018	0.037
recall@20	0.029	0.062	0.030	0.061
recall@50	0.05	0.105	0.053	0.110

- Two similarity measures were used: cosine and jaccard
- Performance are quite similar.

MODEL COMPARISON

	Benchmark		MF		Cosine		Jaccard	
	include	exclude	include	exclude	include	exclude	include	exclude
recall @10	0.070	0.027	0.021	0.043	0.018	0.039	0.018	0.037
recall @20	0.096	0.044	0.035	0.072	0.029	0.062	0.030	0.061
recall @50	0.154	0.080	0.063	0.131	0.05	0.105	0.053	0.110

- *Primary focus is to evaluate their performance excluding those reordered products.*
- *MF is the BEST!*

THANK YOU!

Caixuan Sun
Email: suncaixuan@gmail.com
Project Repository at Github: https://github.com/caixuansun/Springboard/tree/master/capstone_project_2