Online Grocery (Re)ordering

predicting reorders & restocking from Instacart data

Joel Mott

Flatiron School

Project goals

Present three business recommendations regarding:

- aspects of the Instacart data have the most significant correlations with popular, reordered groceries

This will help:

- inform how a grocery delivery ordering system can help **optimize restocking**
- refine **new/still-unordered product marketing** strategies

Business Context: Grocery Delivery & Curbside-Pickup

Online grocery delivery began in 1996.

Became mainstream in the mid-2010's through third-party service companies such as **Instacart** and **Shipt**.

During/after the pandemic, many US grocery chains started their own service:

Walmart Kroger FRESH FOR EVERYONE™ ALDI H-E-B amazon fresh

Besides Amazon Fresh, these stores also offer **curbside-pickup service**.

Data Overview

Instacart uploaded this dataset to Kaggle.com for a competition in 2017

- they challenged analysts to predict what users would order next

data contains over three million records of:

- order information
  - product names
  - whether it's a new item for the user or a reorder
- product details
  - with its department and aisle
- when orders were placed
  - time of day
  - day of the week
  - the number of days since the previous order

Project Overview

broadening scope of Instacart's dataset to inform **restocking** for grocery stores

based on exploratory & regression analysis

focusing on products shoppers order **often**

three business recommendations regarding **which aspects of the data correlate the most with reordering** to help inform restocking strategies

Project Steps

1. Exploratory Data Analysis (EDA)
   - popular products & departments
   - when & how often orders are made

2. linear regression analysis (using Python and StatsModels' OLS module)
   - isolate different aspects of the data
   - specify which correlate the most with reordering

3. business recommendations
   - based on EDA and regression findings

Preliminary findings

Produce and dairy/egg departments are by far the most popular

Customers tend to **add reordered products to the cart first**

Ordering groceries seems to be a **weekly phenomena**:

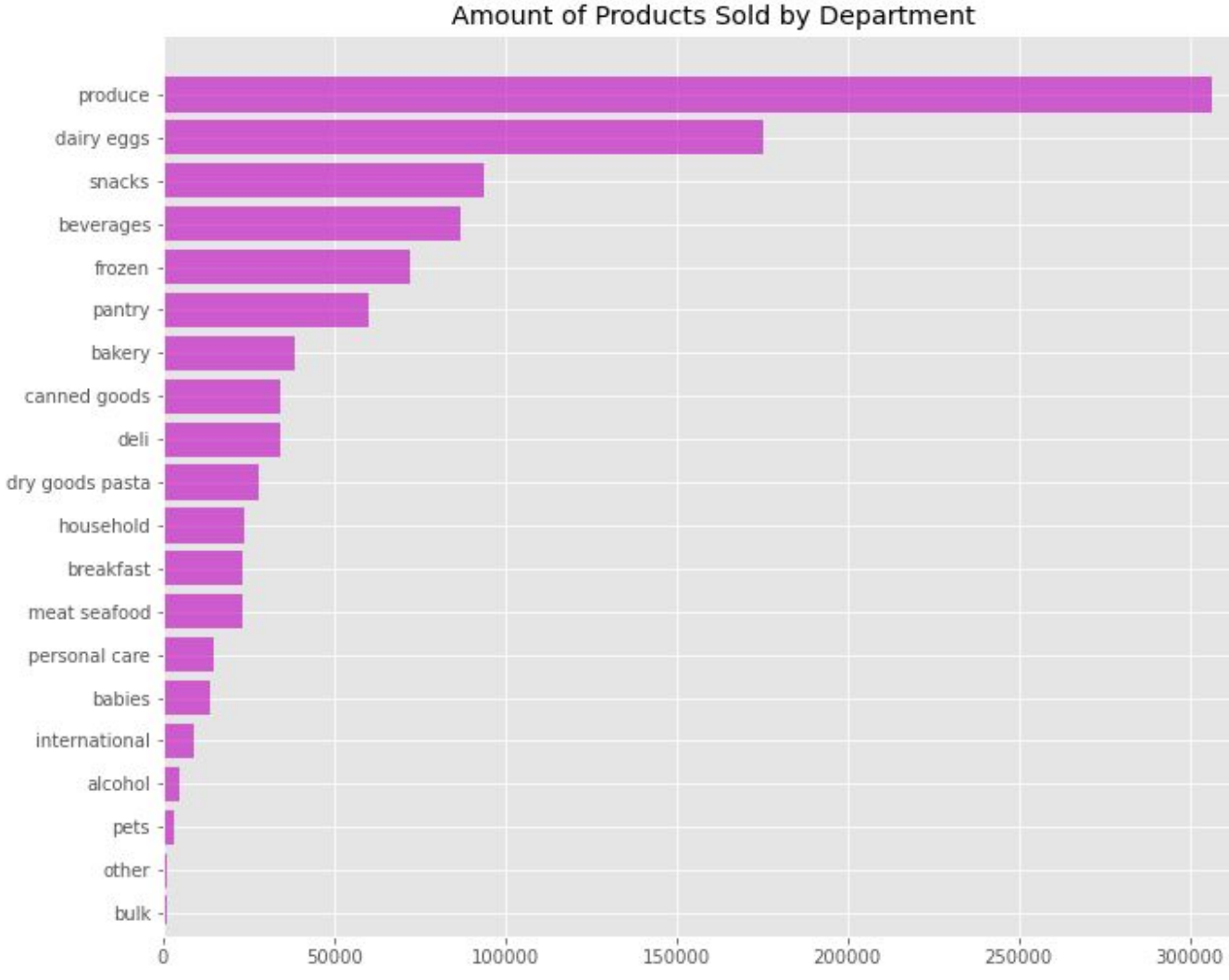
- The most common number of days between orders is **seven**.
- **Weekend orders** far outnumber weekdays.

Popular departments

Produce dominates the other departments.

Dairy/eggs are a distant second, but still far from third.

These two departments sell oft-used products with relatively shorter shelf lives.

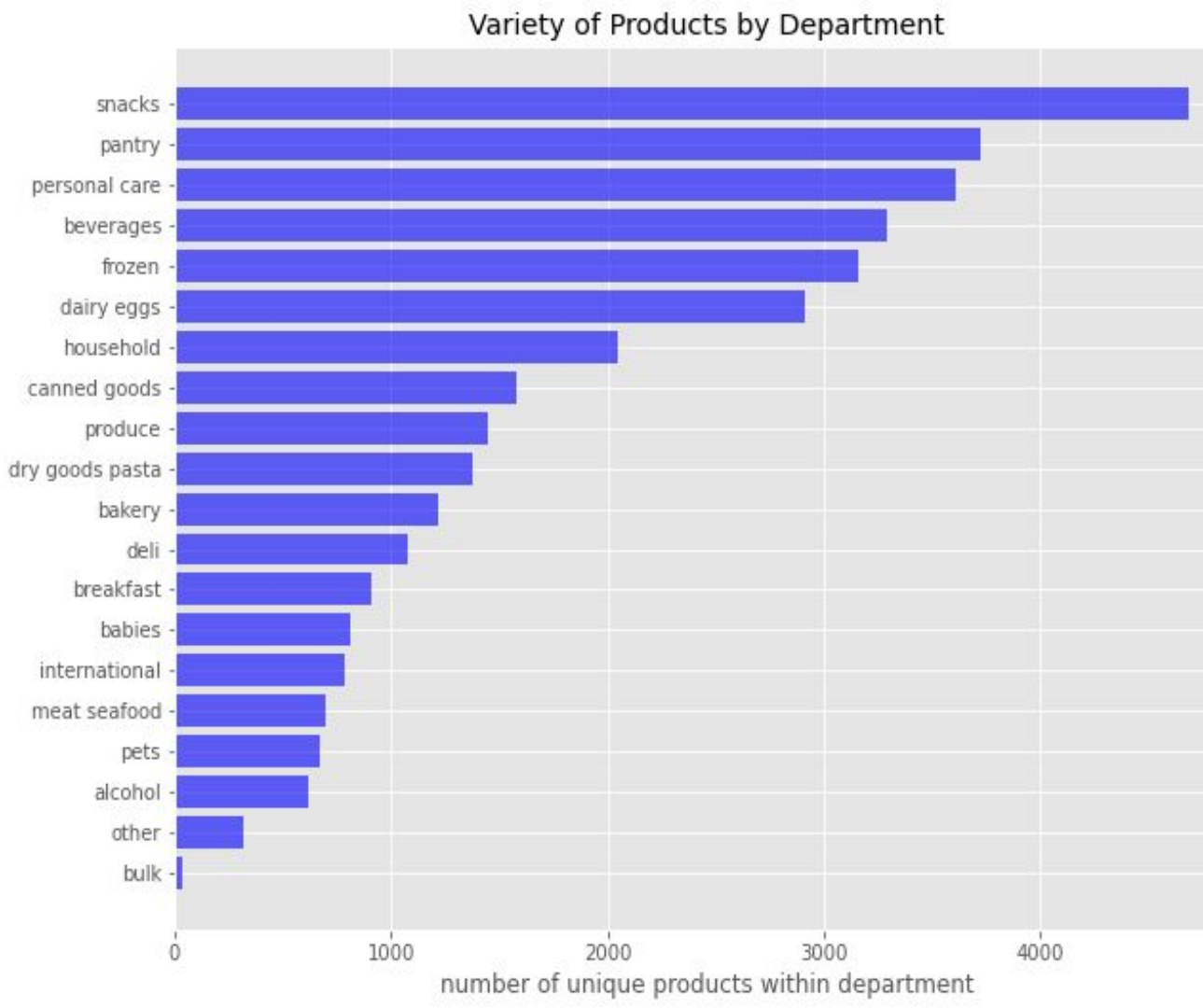


Department size by variety

Produce and dairy/eggs aren't necessarily the *largest* departments.

In fact, they sell fewer unique products than several others.

Scrutinizing the variety of stock for these popular, fresher, short-lived products seems especially vital.



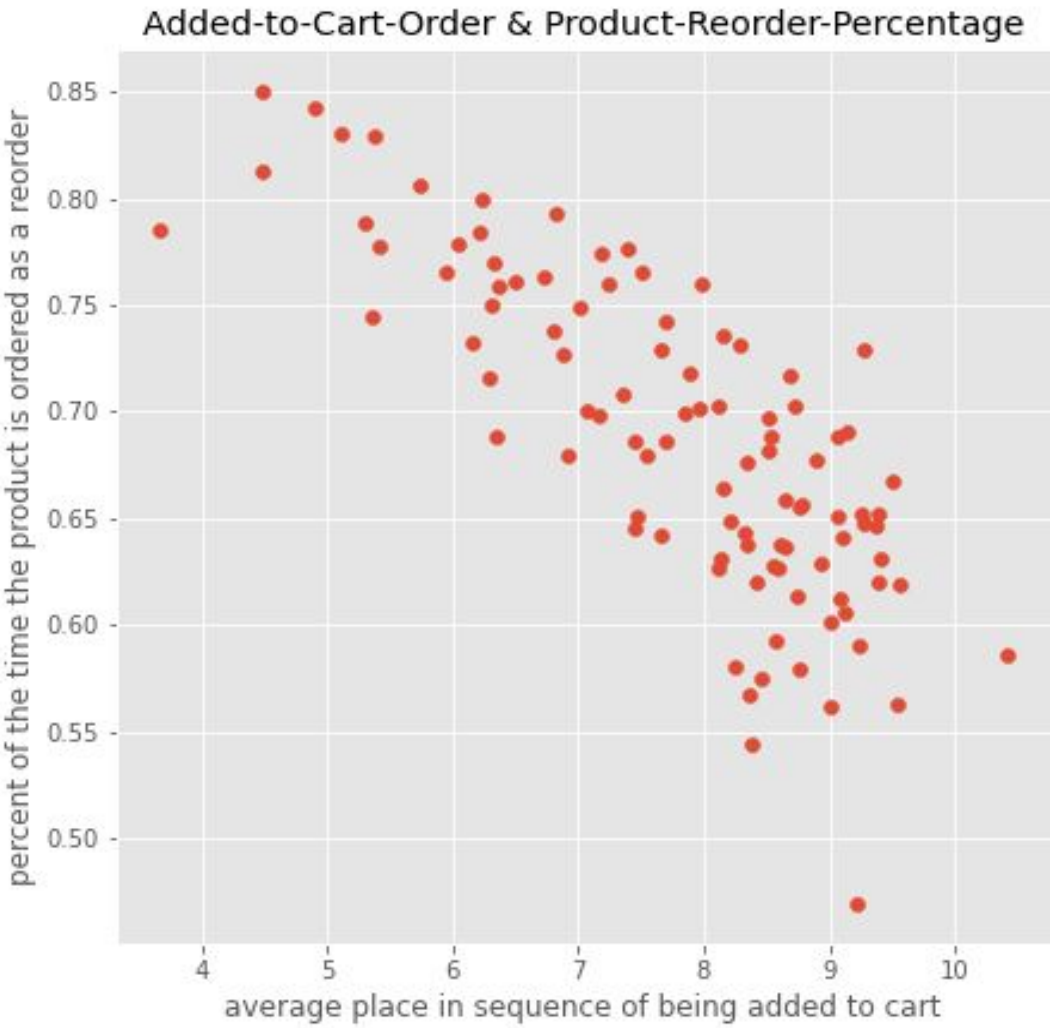
Product details and cart order

Reordered products are placed in the cart earlier in the order process.

This graph just shows the top 100 products.

Including all 35,449 unique products makes these patterns hard to see.

Later, regression analysis will help us specify the exact correlation with more (but not all) products.

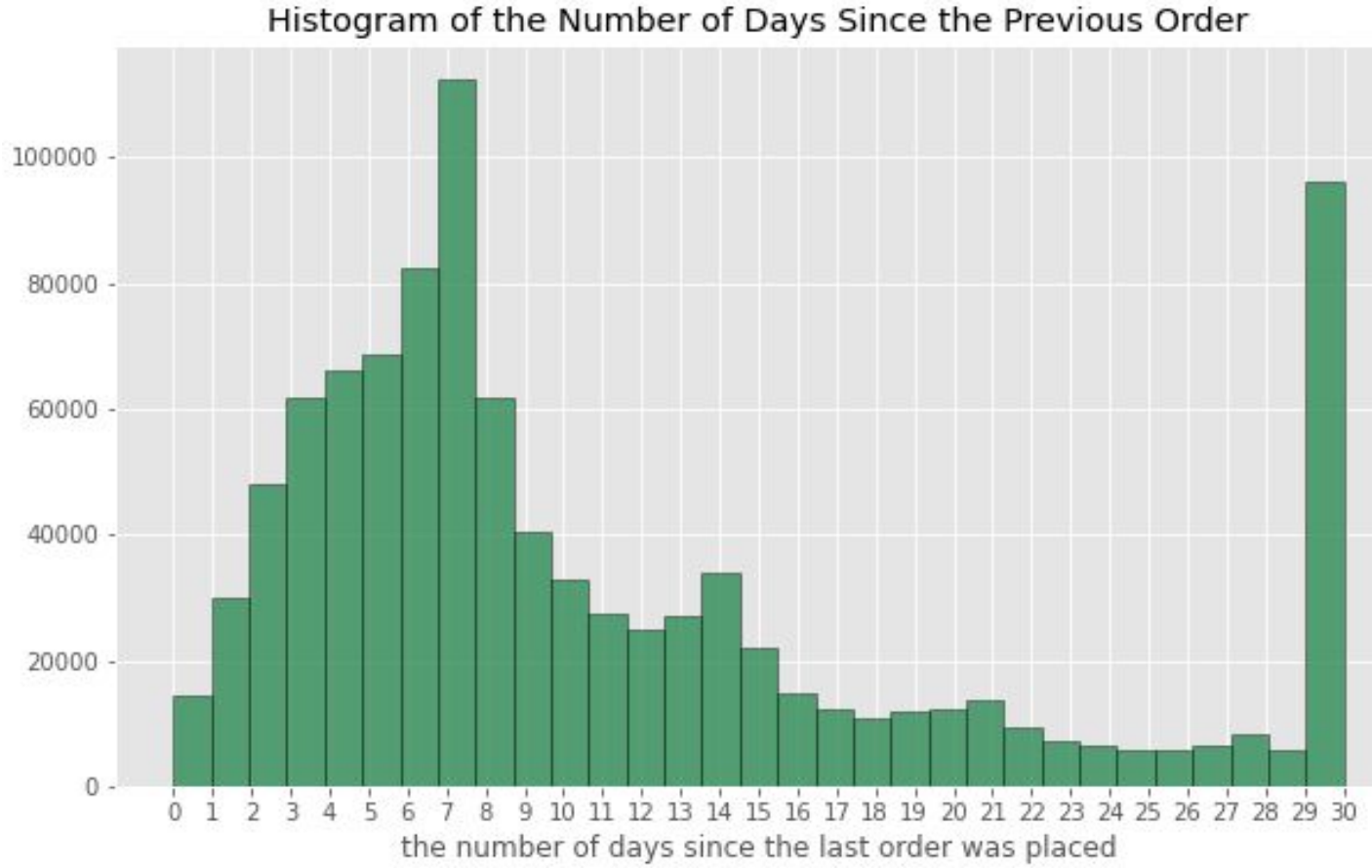


Frequency of orders shows a weekly trend

Many order groceries more often than once a week.

Largest spike is **every seven days** (and then smaller spikes every week thereafter).

“30” days is almost certainly a placeholder for “30 or more days between orders”.



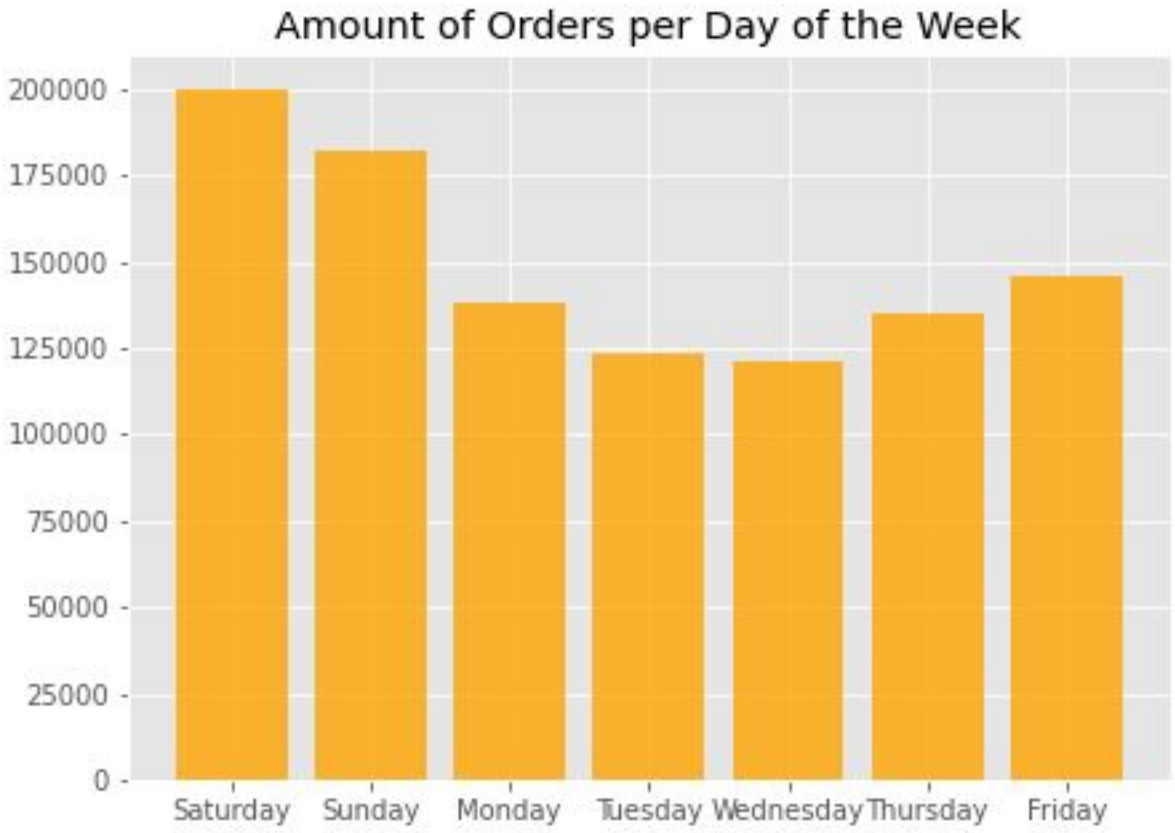
Variance among days of the week

Weekend grocery shopping is quite popular.

Monday-Wednesday may be ideal days for deeper restocking and lighter staffing.

Thursday already begins to ramp up to the weekend.

This pattern informs the peaks in our previous graph on days since the prior order.



Linear Regression Analysis

turning preliminary findings
into
business recommendations

Overall process:

1. Baseline model: only the single strongest aspect of our data that correlates with reorders
2. Investigate the pros & cons of that model
3. Improve on the baseline by adding and/or transforming more aspects of the data

product aspects considered:

1. product name
2. the product's department
3. the product's aisle
4. the average day of the week the product is ordered
5. the average hour of the day the week the product is ordered
6. the average amount of days between orders for this product
7. the average order in which this product is placed into the cart
8. the amount of times the product was ordered
9. the amount of times the product was reordered

strongest correlation

= aspects that show any kind of correlation with reordering

Baseline model

- How correlation is measured:
  - via an “R-Squared score”
  - measures how well correlations explain differences among the data
  - ranges from **0** (poor fit) to **1** (perfect fit)

- At first, even the strongest correlation produces a weak model
  - R-squared score of 0.022
  - This mean it does not explain the data well at all.
  - This has to do with the **amount of times each product was ordered**:
  - some were ordered over 150,000 times while others only once.

Baseline model solutions

Extreme variation among order *amounts* makes predicting difficult.

Eliminating rarely ordered products vastly improves the model to an R-squared score of **0.32**.

I define “rarely ordered products” as those with fewer than 100 orders.

The remaining, **more popular items tell us more about reordering habits** in the first place.

Improving the baseline model

We can add more data aspects to the model as long as they:

- make sense from a logical, business standpoint
- follow a few regression analysis requirements
- don’t introduce problems

The next strongest aspect of the data:

- **average number of days since a product was ordered**
- adding this improves the model (R-squared: **0.383**)

Approaching the final model

The next two strongest aspects regard order **frequency**:

- order **hour of day** & **day of the week**
- however, we cannot include both because they are too interrelated
- this means including both makes it hard to see their individual relation to reordering
- the **hour of day** aspect proves problematic for other reasons

Solution:

- exclude hour of day
- just keeping **day of the week** works well
- R-squared score improves to **0.434**

Final model results: three recommendations on what contributes to reordering products

Recommendation 1:

As the **average add-to-cart order** of a product increases by one, the likelihood of that product being ordered as a reorder **decreases by 4.7%**.

Subsequently, an advertisement for an unordered product may be more effective towards the end of the online ordering process.

Recommendation #2

As the **average number of days since a product was ordered** increases by one, the likelihood of that product being ordered as a reorder **decreases by 2.7%**.

However, orders tend to spike every seven days. This 2.7% decrease likely "resets" to an extent every week (I address this more in the third recommendation).

Nonetheless, focusing on the most frequently purchased products as opposed to rarely ordered items may seem obvious, but it can be helpful to see that the data validates this idea.

Recommendation #3:

Over the course of a week, as a product's average order **day of the week** increases *away from Saturday by one whole day*,
the likelihood of that product being ordered as a reorder **decreases by 15.2%**.

This speaks to the weekend's importance when it comes to stocking commonly-bought items. Mainstay products are ordered more frequently during the common once-a-week trips.

Thank You

any further questions welcome:

joel.mott8@gmail.com

Flatiron School