Online Grocery (Re)ordering

predicting reorders & restocking from Instacart data

Joel Mott

Flatiron School

Business Context

Online grocery delivery used to be done primarily through third-party service companies such as Instacart and Shipt, both of which are still commonly used.

However, during and after the pandemic, many US grocery chains instituted their own delivery and curbside services, such as these popular grocers:

Data Overview

Instacart uploaded a dataset to Kaggle.com for a competition, challenging analysts to predict what users might order next.

This anonymized dataset contains over three million records that include:

- order information
  - product names and 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

I make three business recommendations regarding **which aspects of the data correlate the most with reordering** based on the products online shoppers order the most and how often they order them.

This slightly broader use of Instacart's dataset may help inform both customer habits *and* restocking now that many grocery chains operate their own delivery & curbside-pickup programs.

Project Steps

1. Exploratory Data Analysis (EDA)
   - Popular products
   - Popular departments
   - When customers order
   - How often they order

2. Linear regression analysis (using Python and StatsModels' OLS module)
   - Isolate different aspects of the data
   - Specify which aspects correlate with reordering

3. Business Recommendations
   - Based on EDA and regression findings

Preliminary findings

- The produce and dairy/egg departments are by far the most popular.

- Customers tend to add their reordered products to the cart earlier on during the ordering process; new items are usually added later.

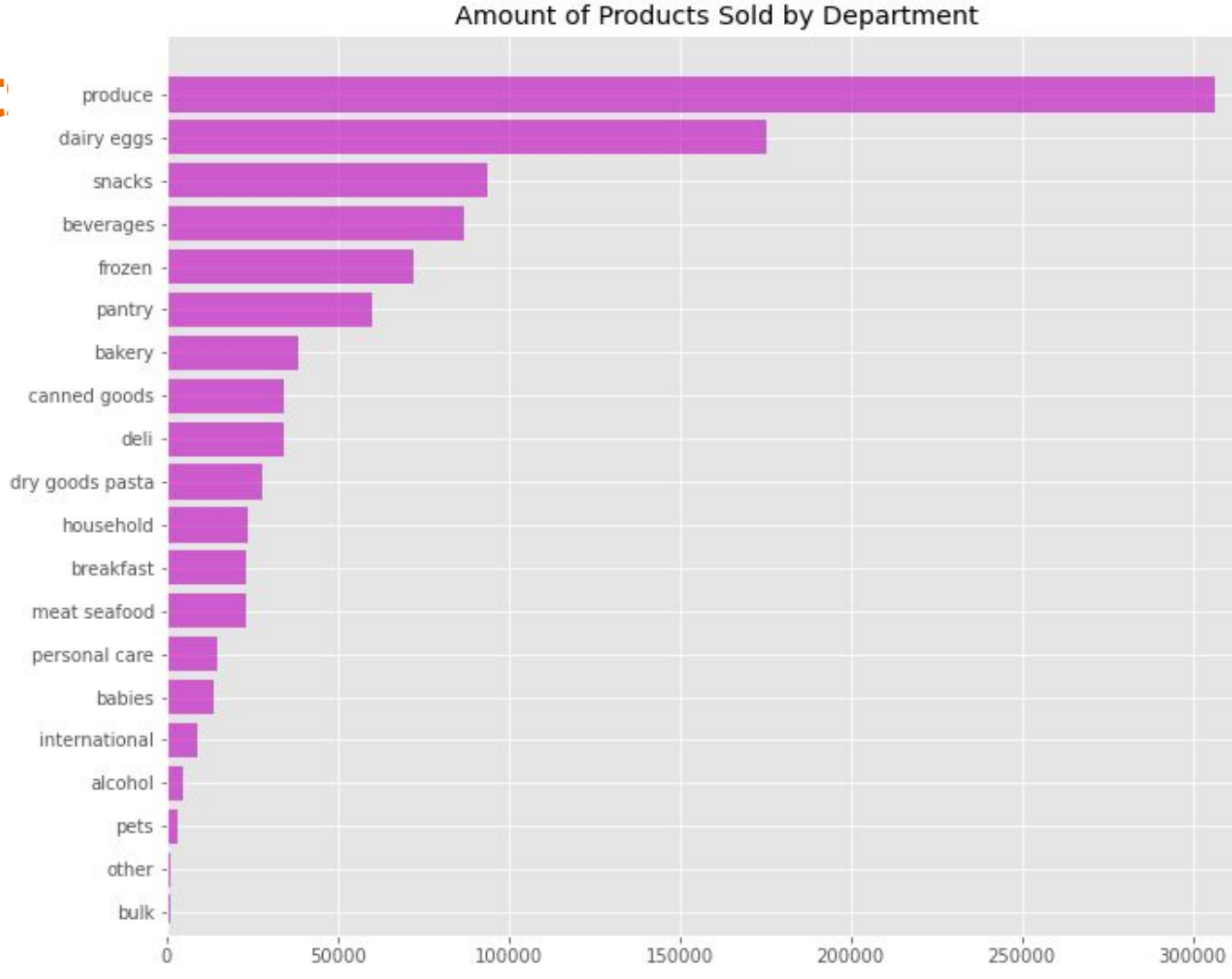
- Ordering groceries seems to be a weekly phenomena:
  - The most common number of days between orders is 7
  - Weekend orders far outnumber weekdays

Popular department

Produce dominates the other departments.

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

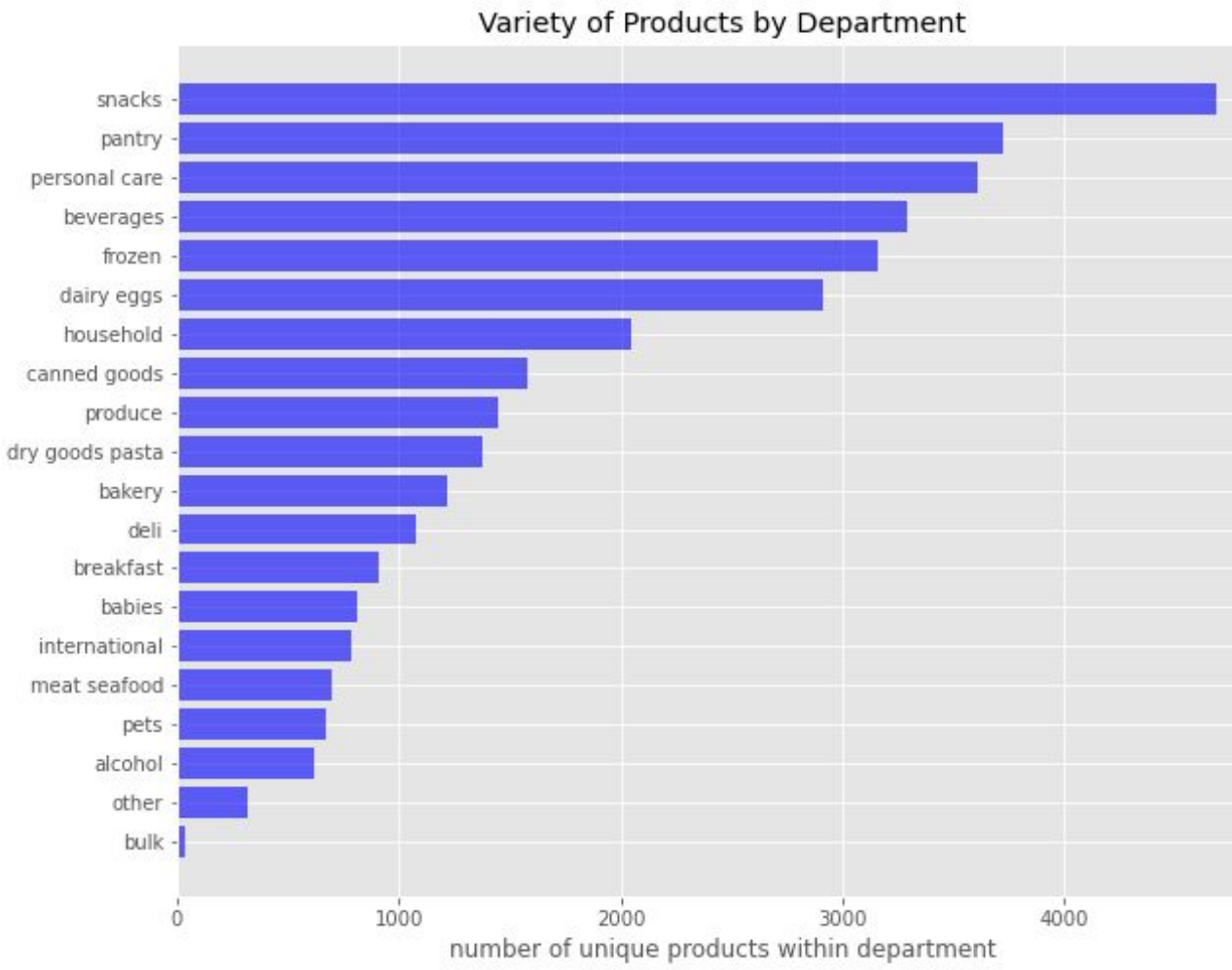
These two departments sell oft-used products with relatively shorter shelf lives.



Department size by variety

Produce and dairy/eggs aren't the largest departments; they sell fewer unique products than others.

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



Product details and cart order

The most popular products tend to appear in the cart earlier on during the order process.

Later, regression analysis will help us specify the exact correlation between these order frequency elements.

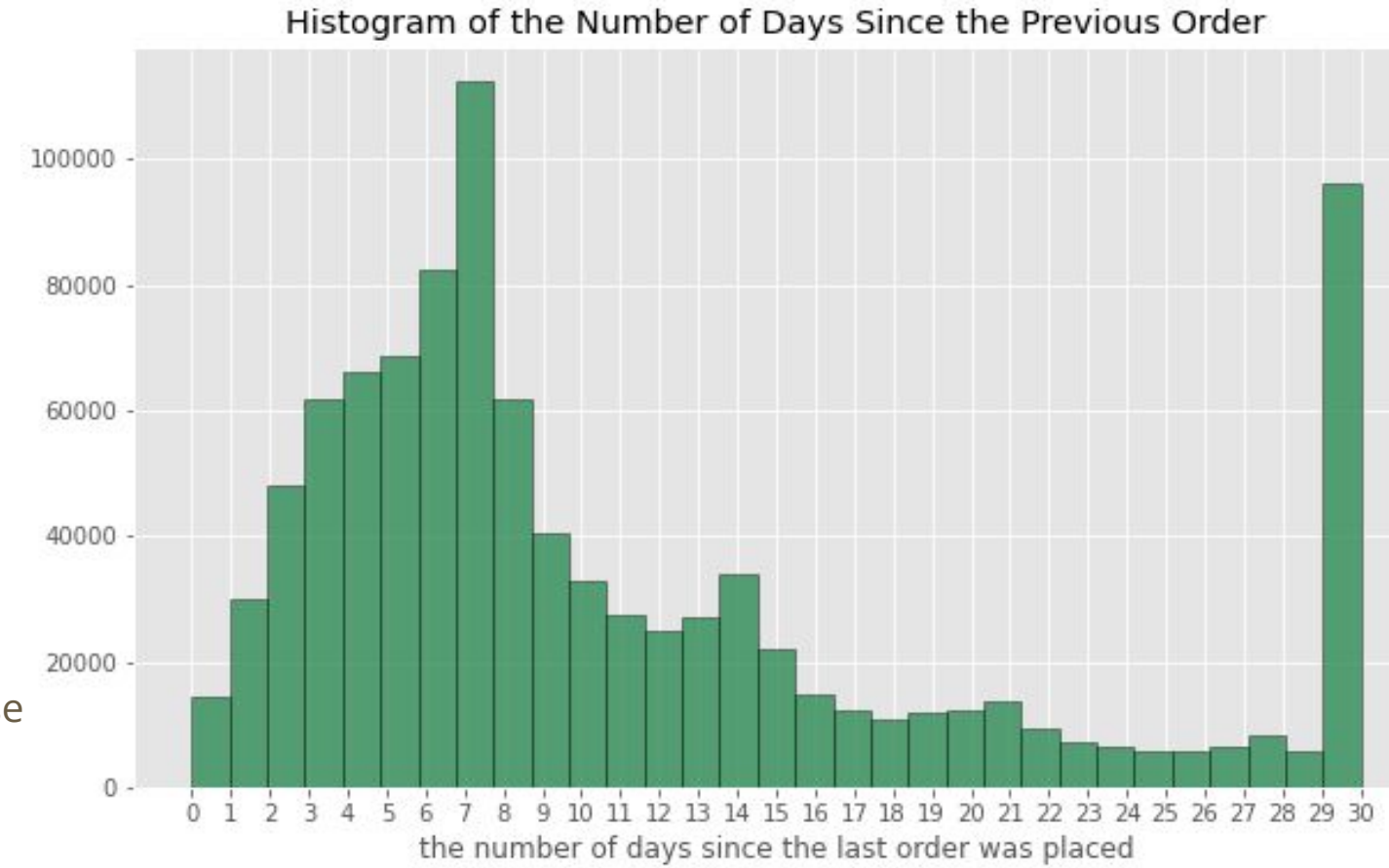
Frequency of orders shows a weekly trend

Many order groceries more often than once a week

There's a spike at seven days (and then smaller ones every week thereafter).

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

Many customers only use Instacart every once in a while.



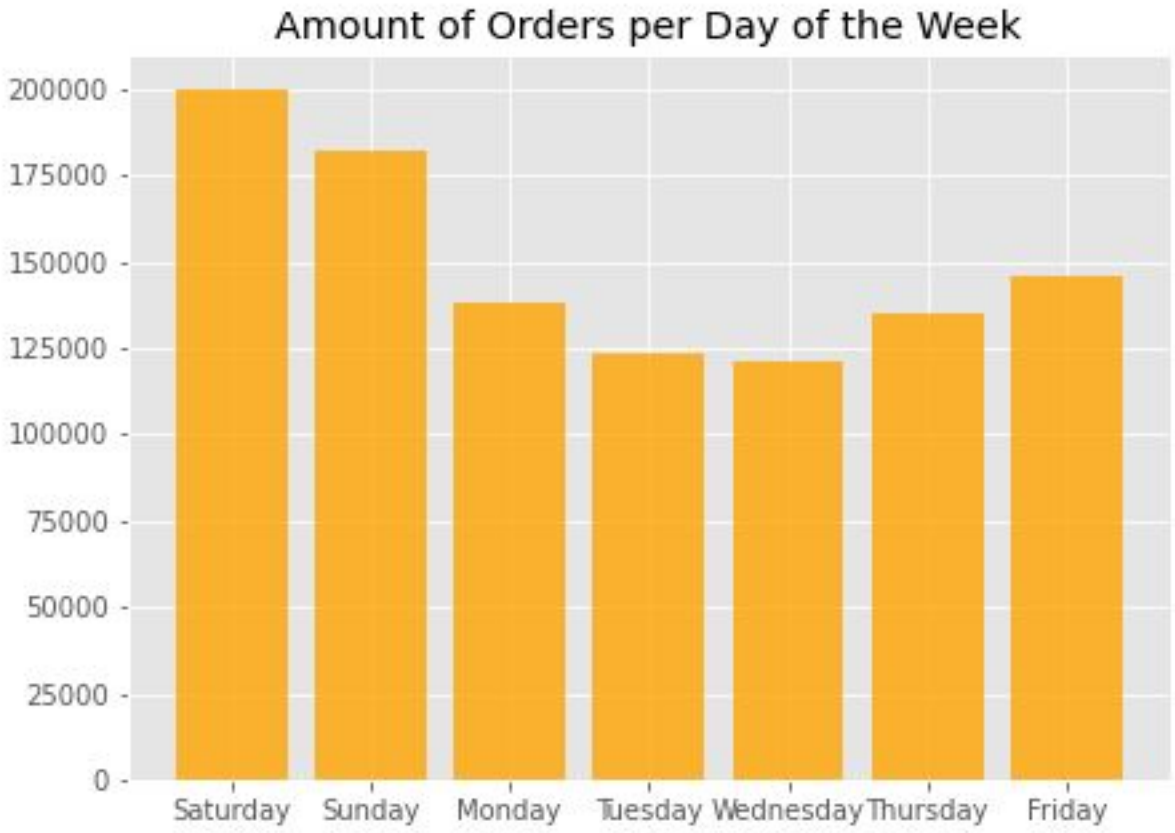
Variance among days of the week

Weekend grocery shopping is markedly 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: 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 problems

An R-Squared score is a measurement of how well correlations explain differences among the data. It ranges from 0 to 1.

At first, even the strongest correlation produces a weak model that does not explain much of the data (R-squared score of 0.022).

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

The variance among order amounts makes it harder to predict reordering.

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

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

This happens because the remaining, more popular items tell us more about reordering habits in the first place.

Starting to improve on the baseline model

We can add more data aspects (“features”) as long as they:

- make sense from a logical, business standpoint
- follow a few regression 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 up to **0.383**)

Approaching the final model

The next two strongest aspects regard order frequency:

- order **hour of day** and **day of the week**
- however, we cannot include both because they are too interrelated
- in other words, including both makes it hard to see their individual relation to reorders
- 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**

Excluding times ordered and reordered

The same scenario applies to the **amount of times a product was ordered or reordered**: these two aspects are quite interrelated.

However, we can't use *either* one here because their distributions are non-normal.

This means regression analysis won't be able to explain how they relate to the likelihood of a product being reordered.

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

Recommendation 1:

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

Subsequently, an advertisement for an as-of-yet unordered product for a customer 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%**.

Keep in mind that 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

joel.mott8@gmail.com

Flatiron School