Instacart Market Basket Analysis

An Exercise in Predicting Future Consumer Purchases

By

LaShawn Gaines

Why perform market basket analysis?

Allows organizations to understand consumer purchasing behaviors and patterns

Knowing which items are frequently purchased together enhances cross-selling and upselling opportunities

Allows organizations to optimize product placements and visibility

Allows organizations to better utilize targeted promotions and discounts to consumers

About the dataset

- Instacart published the data as part of a kaggle competition
- Dataset contains grocery orders for a population of over 200,000 users
- Several relational files were provided
  - An Orders table containing over 3.4 million records
  - A combined orders and products table containing over 30 million records, containing one record per item purchased in an order (includes the product id, days_since_prior_order, and add_to_cart_order)
  - Products (50,000 unique ids), aisles(134 unique ids) and department (21 unique ids) tables containing the categorical feature values
- All consumer demographic data was removed from the dataset to maintain privacy

Data Wrangling

the dataset was very clean with the appropriate data types established

completed a schema map to indicate all the relationships between the various relational files

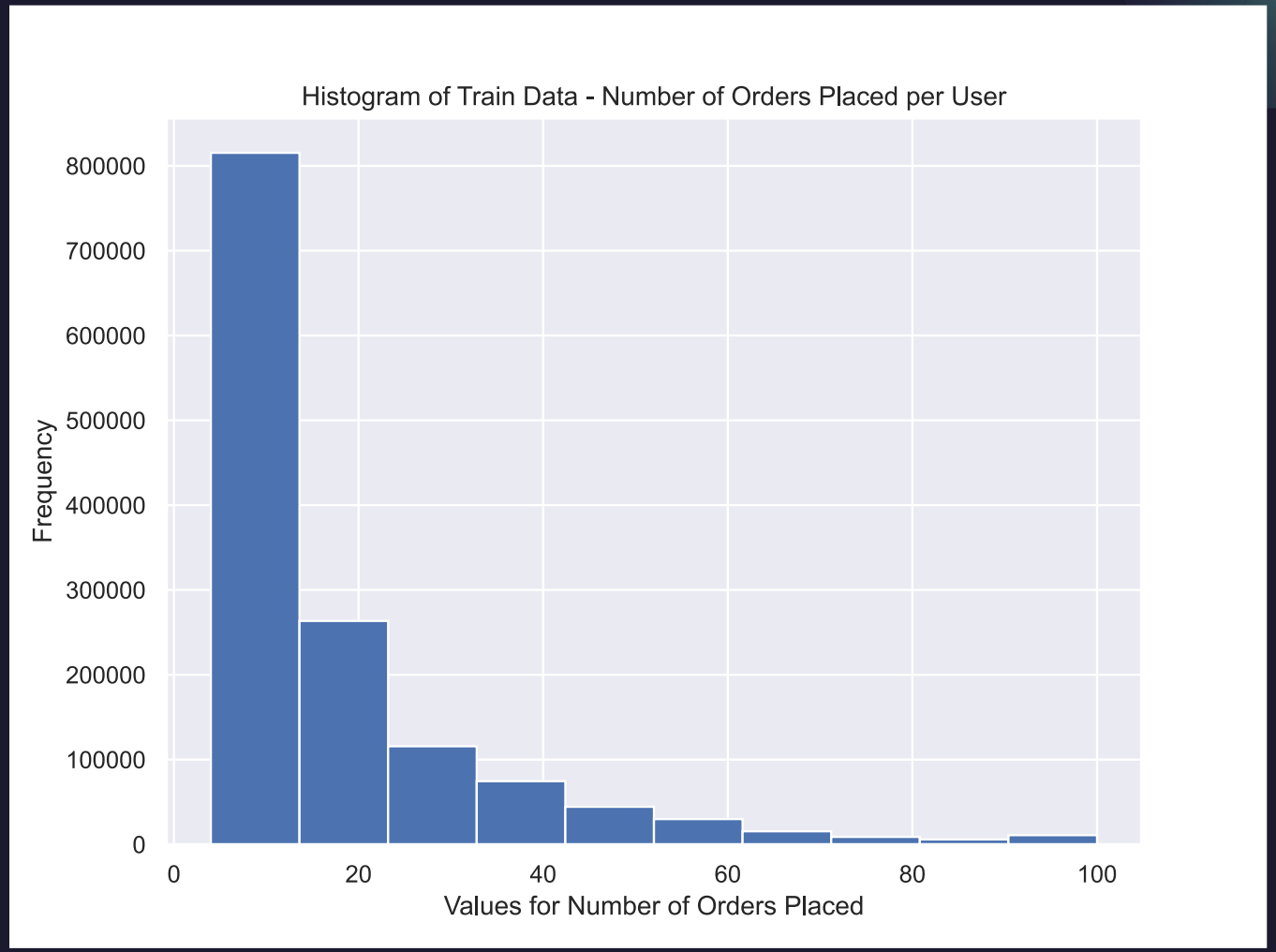
only the orders table contained missing data in the days_since_prior_order field

The competition notes stated that nothing over 30 days was included in this column

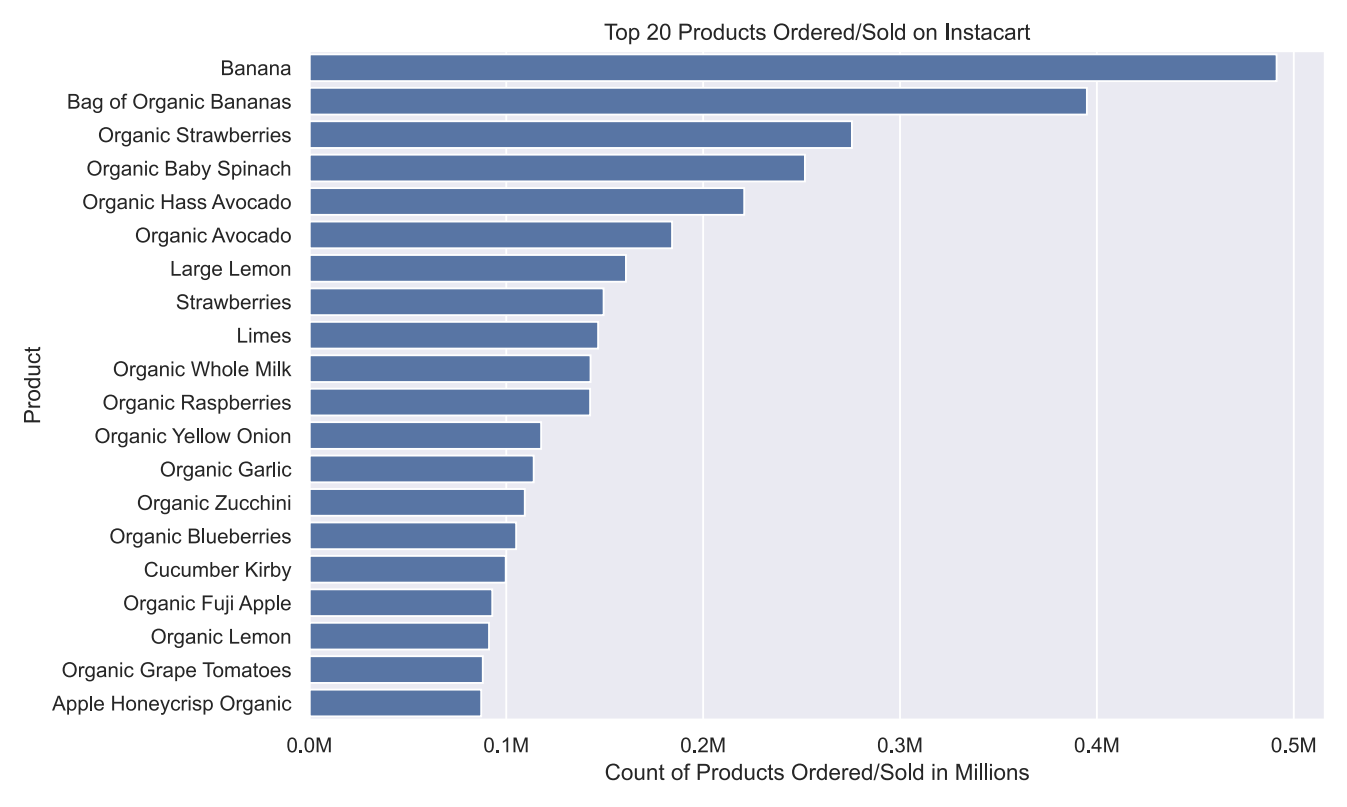
For an RFM analysis, a median value of 8 was used to replace the NaN values

EDA – preliminary findings

- Leveraged Ydata profiling for the initial exploration
- Correlations were obvious (reorder status is highly correlated to number of orders placed)
- Number of orders placed was skewed to the right
  - Will be addressed during RFM segmentation



EDA – most popular products



- Most frequently purchased items were in the produce department
  - Of the top 20 items sold on the platform, organic milk is the only item not categorized as produce
  - Bananas are the most frequently purchased item in this dataset
  - Produce items are more likely to be reordered than other goods
  - Organic items in produce are also more likely to be reordered than other goods
- The personal care items department has the most products offered for sale on the platform (however ranks only 14th in purchases)

EDA – Correlations found

No truly strong correlations were noted, other than the obvious (positive correlation between number of orders placed and reorder status)

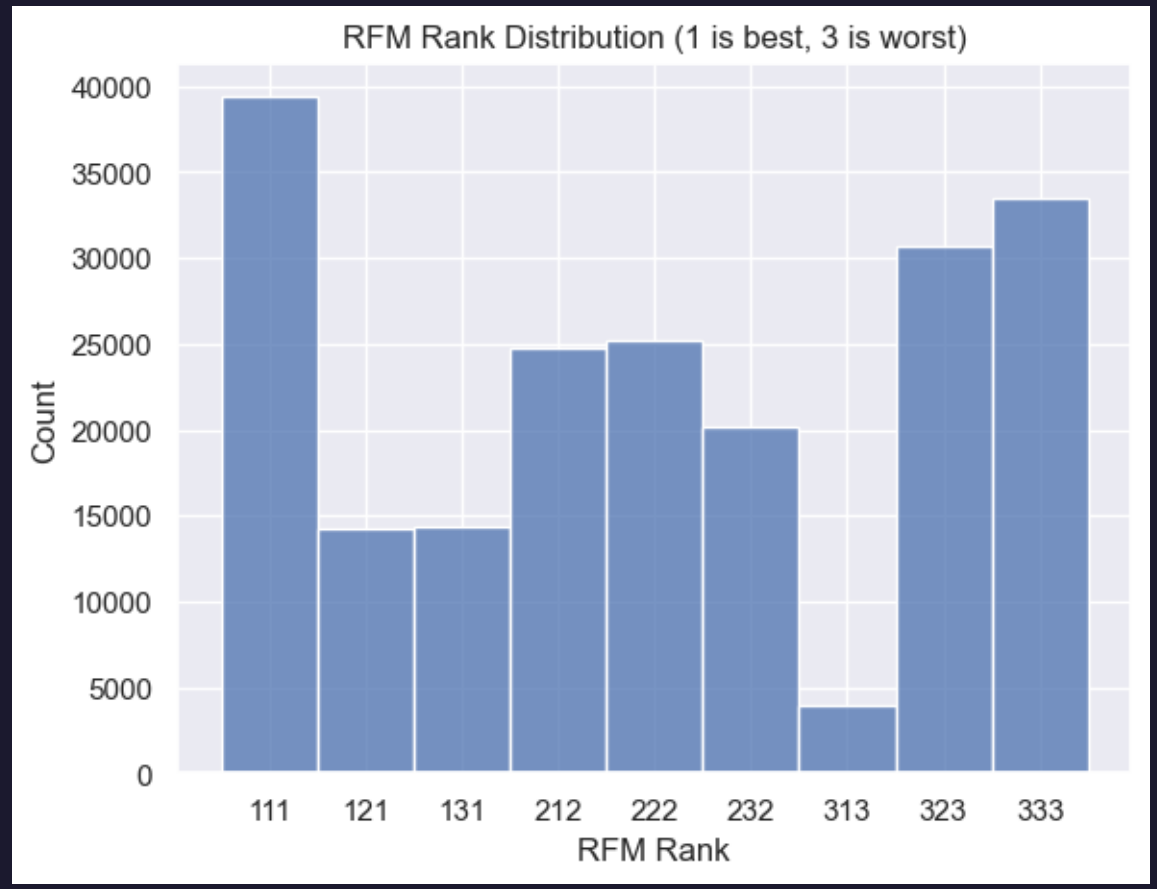
Items amongst the first 10 items added to a shopping cart have a weak positive correlation with reorder status

Items ordered between 8 and 14 days prior, also had a mildly positive correlation with reorder status

There was a slight negative correlation between the purchases of produce items and dairy items

Customer Segmentation Part 1: Recency Frequency & Monetary Value (RFM) Model

- RFM (Recency, Frequency, Monetary Value) model chosen due to lack of demographic data for segmenting customers
    - Revenue info was also absent from dataset (# of items purchased was used as an approximation)
    - Leveraged logarithmic and quantile transformations to normalize frequency and monetary values
- Largest RFM segment contains the most engaged, most frequent customers with the highest monetary value
- Second largest segment contains the least engaged, least frequent customers with the lowest monetary value

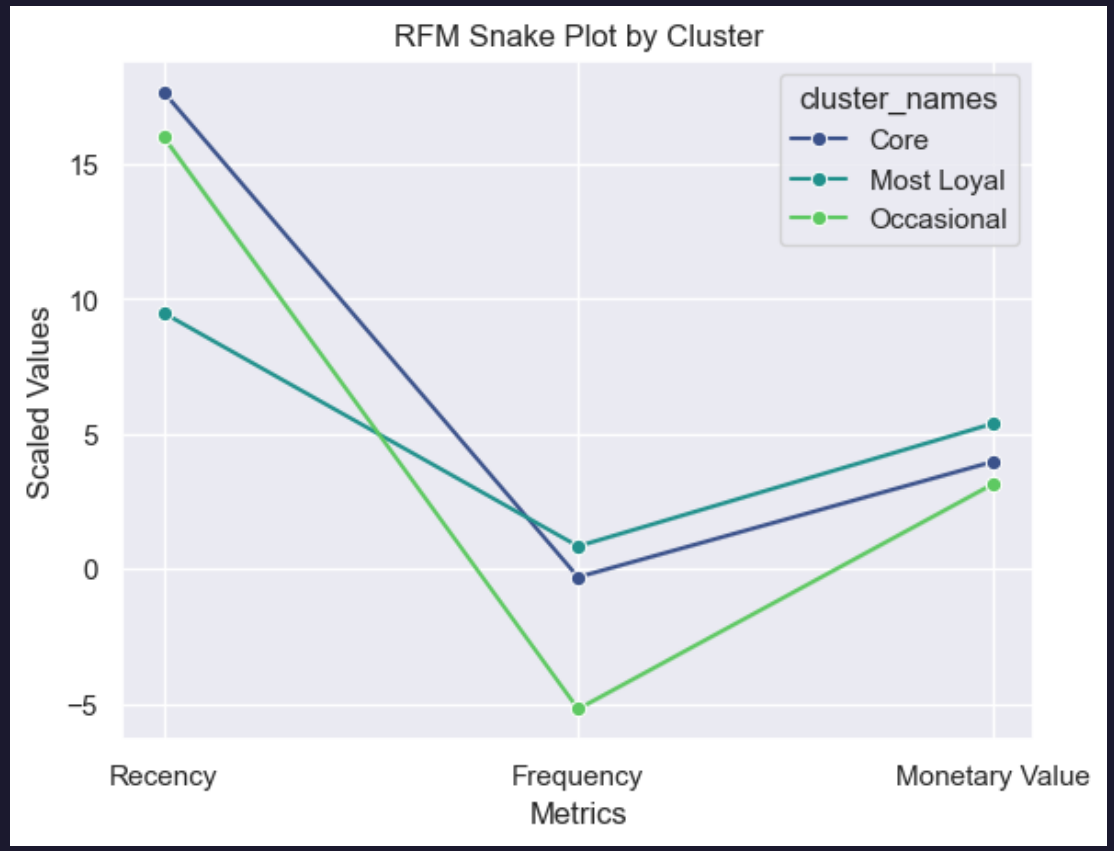


Customer Segmentation Part 2: K-means clustering

- Optimum number of clusters = 3 (elbow plot)
- Used t-SNE to reduce dimensionality of the clustered data
- Segments are clearly defined

Customer Segmentation Part 3: Customer Insights and Opportunities

- Most Loyal customers – shop most often, with larger carts and monetary value
  - Incentives may include suggested product offerings while shopping to generate cross-selling or upselling opportunities
- Core customers – still generate monetary value, likely through larger orders
  - Incentives for this segment may include reorder reminders to increase their frequency and monetary value
- Occasional customers – need incentives to both drive frequency and order size to increase their monetary value



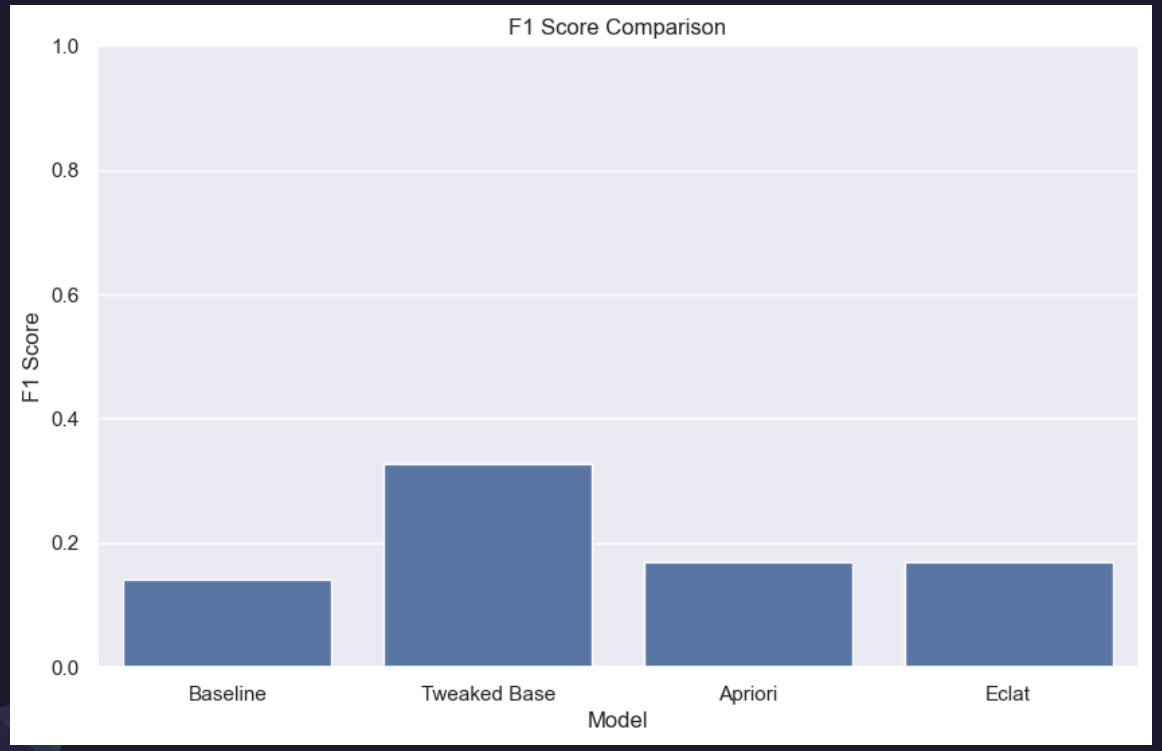
Market Basket Analysis: The Challenges

- Size of the combined orders/ products transaction table contained over 30 million rows
- Reducing the dataset to focus solely on Most Loyal customers still resulted in a dataset of 25 million rows
- Various customer sample sizes were attempted (10%, 5% and 2.5%) memory issues persisted
    - ending in sample size of 1% of the customers and their transaction records
- Reduced dataset was still too large to complete the FP Growth algorithm, leading to creation of a baseline model, tweaked baseline model and Apriori and ECLAT association rules models

Market Basket Analysis: The Models

- Baseline methodology: build a cart based on each user's average cart size (n) and selecting Instacart's Top n products
- Tweaked Baseline methodology: build a cart based on each user's average cart size (n) and selecting their specific Top n products ordered on Instacart
- Apriori & ECLAT Association Rules methods:
  - Build carts based on average user's cart size (n)
  - Initial item in the cart is the item each user purchases most frequently
  - Build a cart with n items, using the antecedents and consequents from the association rules

Market Basket Analysis: Recommended Model



- Recommend the tweaked baseline model with an F1 score of 32.7%
- Least computationally expensive of all the models
- Caveats
  - Due to CPU restraints these scores are based on a very small sample size
  - Use this only as a starting point, further algorithms should be pursued

Market Basket Analysis: What would I do differently?

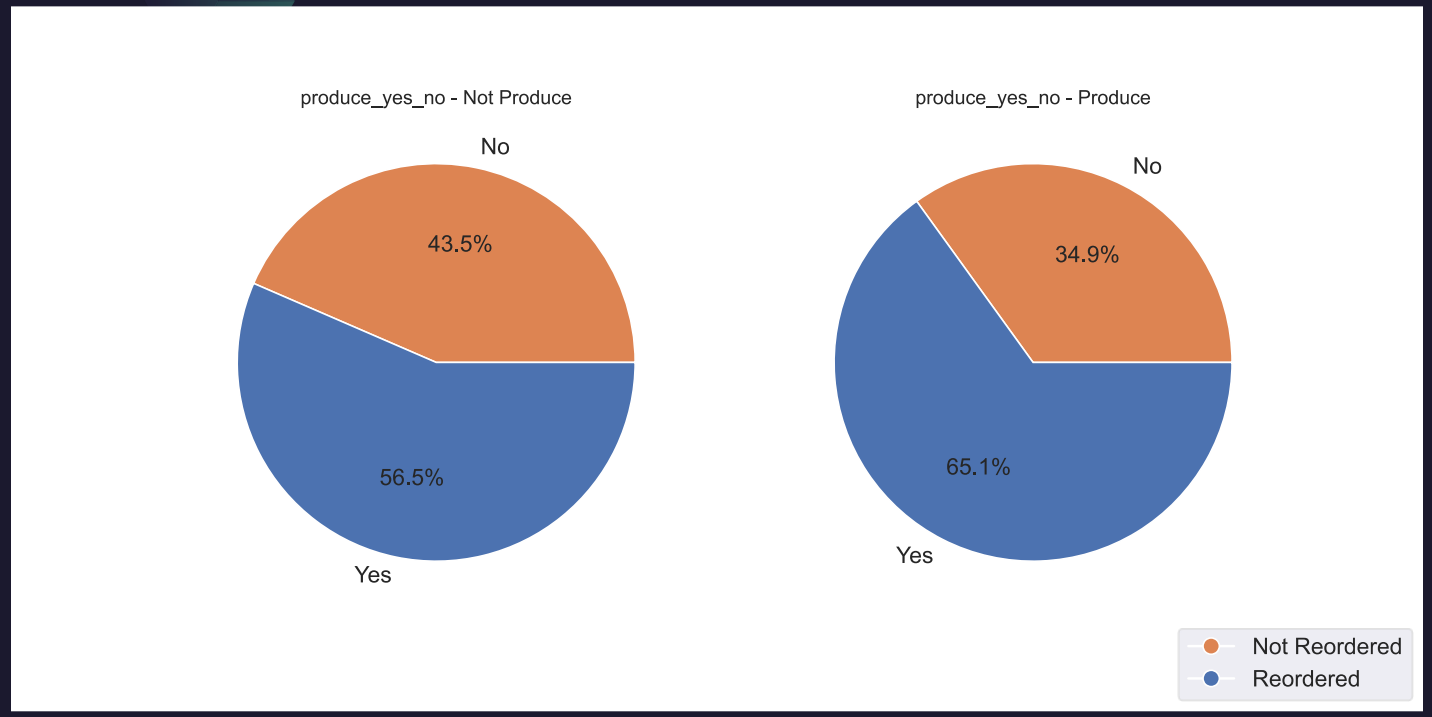
Use Apache Spark due to its scalability and efficiency when working with large datasets

Revisit Apriori and ECLAT models with larger sample sizes using mlxtend package for Spark

Revisit the FP Growth model (included within the Spark MLlib library)

Apply the XGBoost algorithm (included within the XGBoost4J-Spark library)

Business Recommendations



- Produce items and/ or organic items are the top selling items
  - Increase the visibility and of these items on the platform
  - Ensure these products are always readily available
  - Perhaps add reorder reminders and suggestions given the average time between orders
- Leverage the customer segmentation data to tailor promotions and discounts accordingly

Thank You!