Grocery Recommendation System

Data from InstaCart

Data

21 Departments

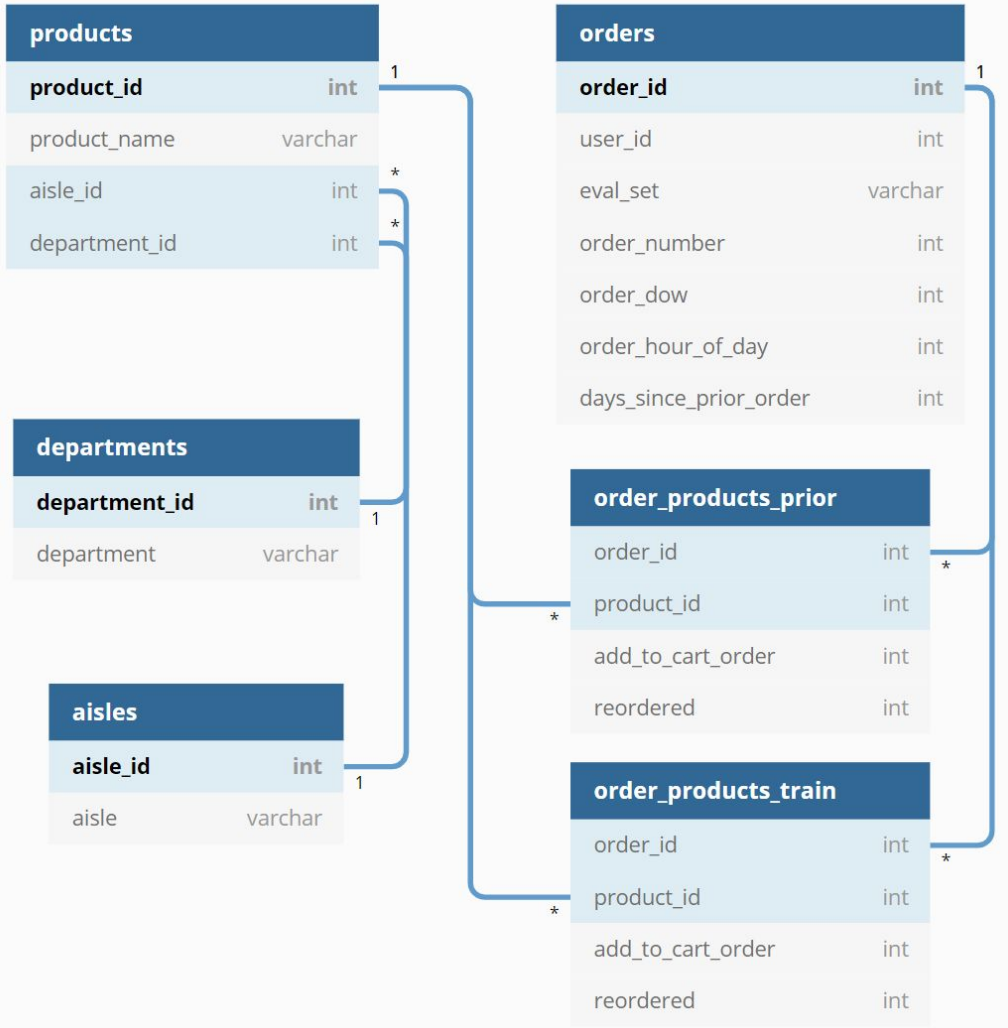
134 Aisles

49,688 Products

206,209 Users

3,421,083 Orders...

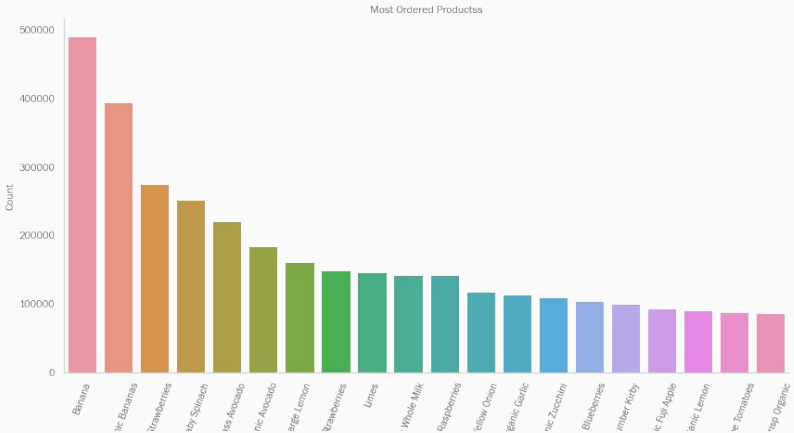
33,819,106 Ordered Products



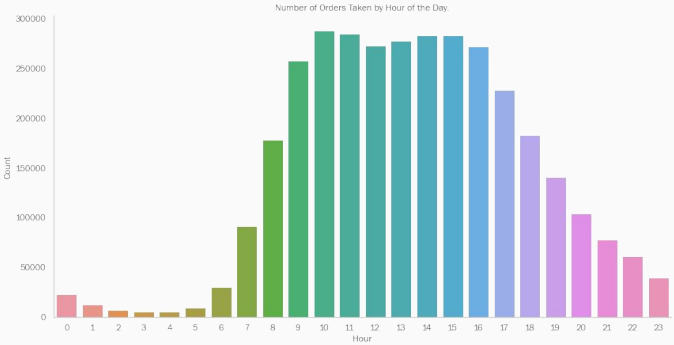
Fun Food EDA

Most Ordered Product - **Bananas**

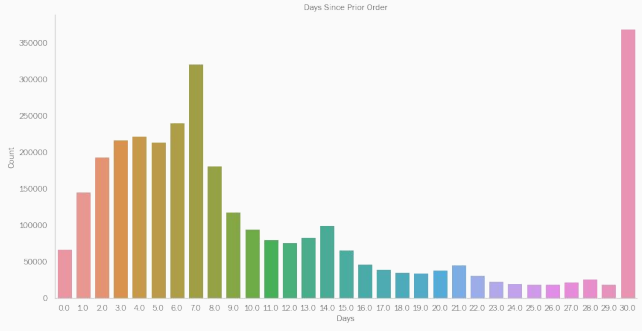
2nd Most Ordered Product - **Organic Bag of Bananas**



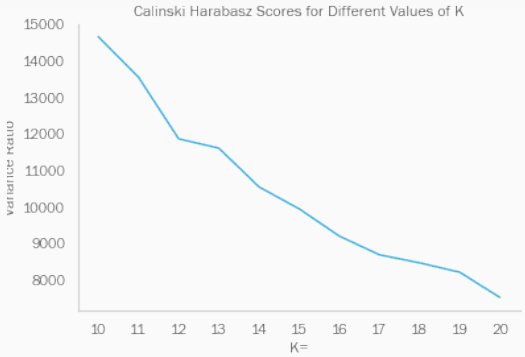
Order Hour of Day

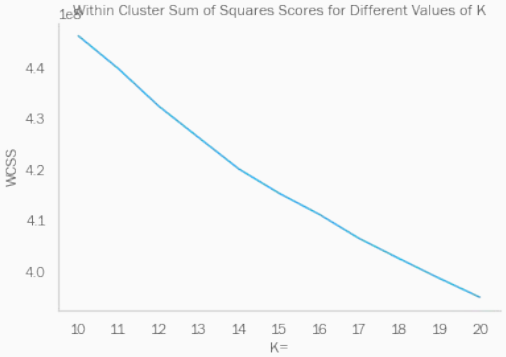


Days Since Last Order



KMeans Clustering





print(cluster_metrics[5]) # Lots of personal care / pharmacy type products

print(cluster_metrics[7]) # Baby products

print(cluster_metrics[8]) # Lots of orders, shortest days between orders, big buyers

print(cluster_metrics[9]) # Alcohol purchasers

print(cluster_metrics[11]) # Soap and skin care

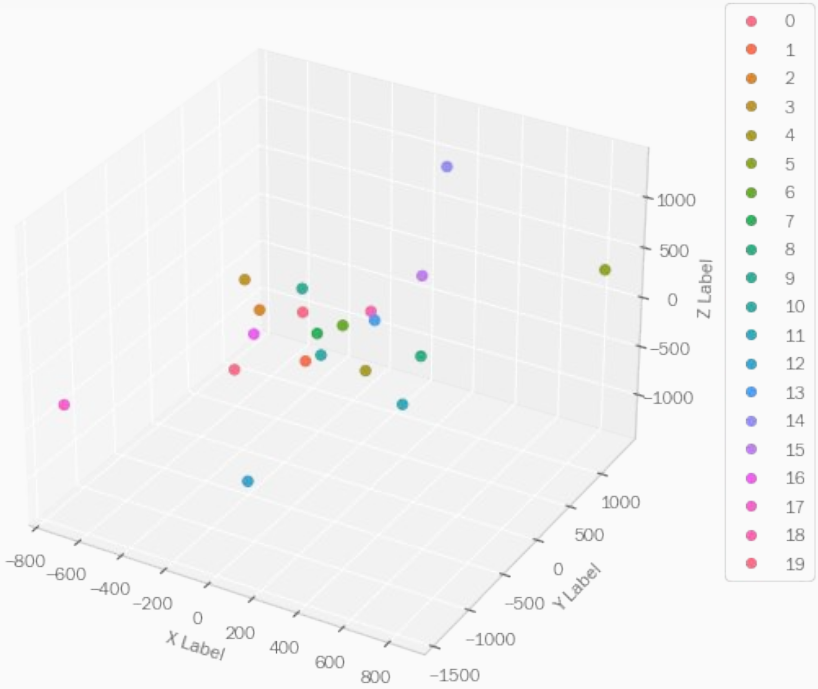
print(cluster_metrics[12]) # Very large cluster, with fewest number of orders and highest days between orders

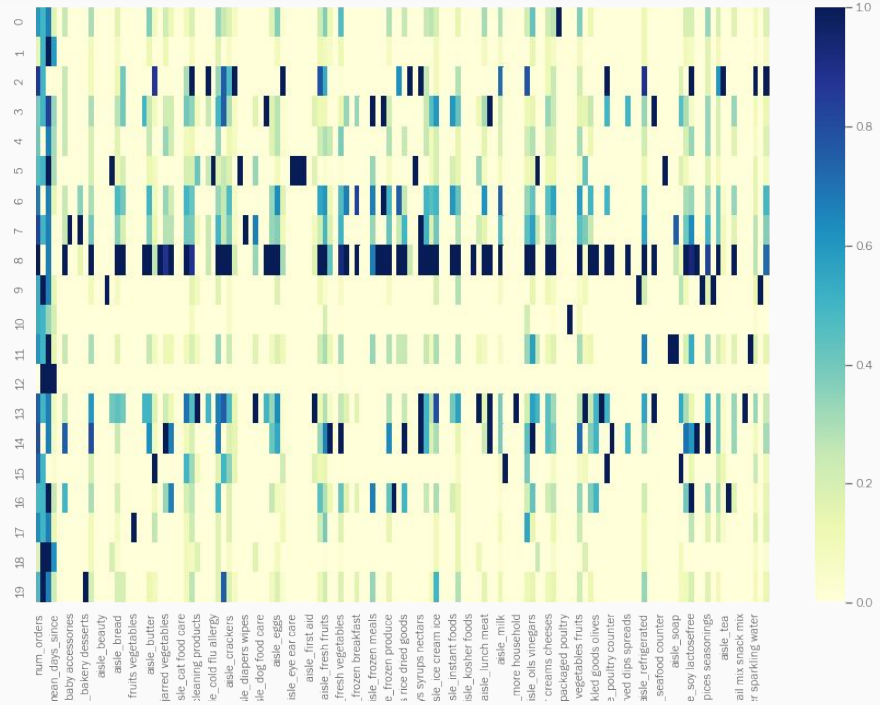
print(cluster_metrics[13]) # Household, laundry, cleaning products

print(cluster_metrics[15]) # Chocolate, gum and soft drinks, least veggies

print(cluster_metrics[16]) # Vegan products and tofu

Cluster visualizations





Metadata Recommendation

```
In [276]: vectorize_products_based_on_metadata('Oreos')

Out[276]: 22014              Thin Mint Crisp Oreos
          23995    Halloween Oreos Sandwich Cookies
          Name: product_name, dtype: object

In [249]: vectorize_products_based_on_metadata('Premium Almonds')

Out[249]: 49178                                    Premium Almonds
          24511    Condoms, Premium Latex, Ultra Thin, Premium Lu...
          44962                              Roasted Tamari Almonds
          7272                               Yogurt Covered Almonds
          5597                               Organic Tamari Almonds
          23466                          Platinum Premium Lubricant
          20405                            Roasted Unsalted Almonds
          21698    Pistachios, Premium Blend, Pomegranate, with A...
          18035                                Premium Horseradish
          25923                        Premium Lubricant Condoms Enz
          Name: product_name, dtype: object

In [253]: vectorize_products_based_on_metadata('Red Potatoes')

Out[253]: 13732                Red Potatoes
          5651          Organic Red Potatoes
          3492             Baby Red Potatoes
          44892                          Red
          10469                B Side Red
          47794                  Red Wine
          13706               Essential Red
          6877                   Red Blend
          14259    Organic Red Potatoes, Bag
          4739                   Decoy Red
          Name: product_name, dtype: object

In [277]: vectorize_products_based_on_metadata('randomword')

          No similar products found.  Please refine your search terms and try again
```

Natural Language Processing

Uses product names, aisle name, and department name with a Count Vectorizer and calculates cosine similarity.

Works like search engine optimization

Stemmed rather than lemmatized

Like to put stronger weight on nouns rather than adjectives

Recommendation System

Too large to perform memory/neighborhood based models

Surprise SVD Model with grid search

User Id, Item Id, Number of times purchased (1-100)

RMSE 3.25, but high values seems far off

Currently scaling number of times purchased to a 1-5 rating scale

Market Basket Analysis

Predict what a given user will order next. **Association rules** are normally written like this: ***{Diapers} -> {Beer}*** which means that there is a strong relationship between customers that purchased diapers and also purchased beer in the same transaction.

Support: Probability of buying X and Y products together: **Support(X, Y) = Freq(X,Y)/N**

Confidence: This says how likely item Y is purchased when item X is purchased.

Confidence(X, Y) = Freq(X,Y) / Freq(X)

Lift: Shows how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. **Lift = Support (X, Y) / (Support(X) * Support(Y))**

More Work

Front end application?

Jupyter Dash?

SQL tables