Grocery Recommendation System

Data from Instacart

Presented by : Melody Peterson
Flatiron School Data Science Capstone Project

Business Problem

- Client is a grocery store with online shopping capabilities
- Looking for unique ways to personalize marketing to customer base
- Particularly to provide product recommendations

Analyze grocery purchase data and group like customer segments together. Provide product recommendations on keyword search and to suggest additional items to add to cart.

Data

21 Departments

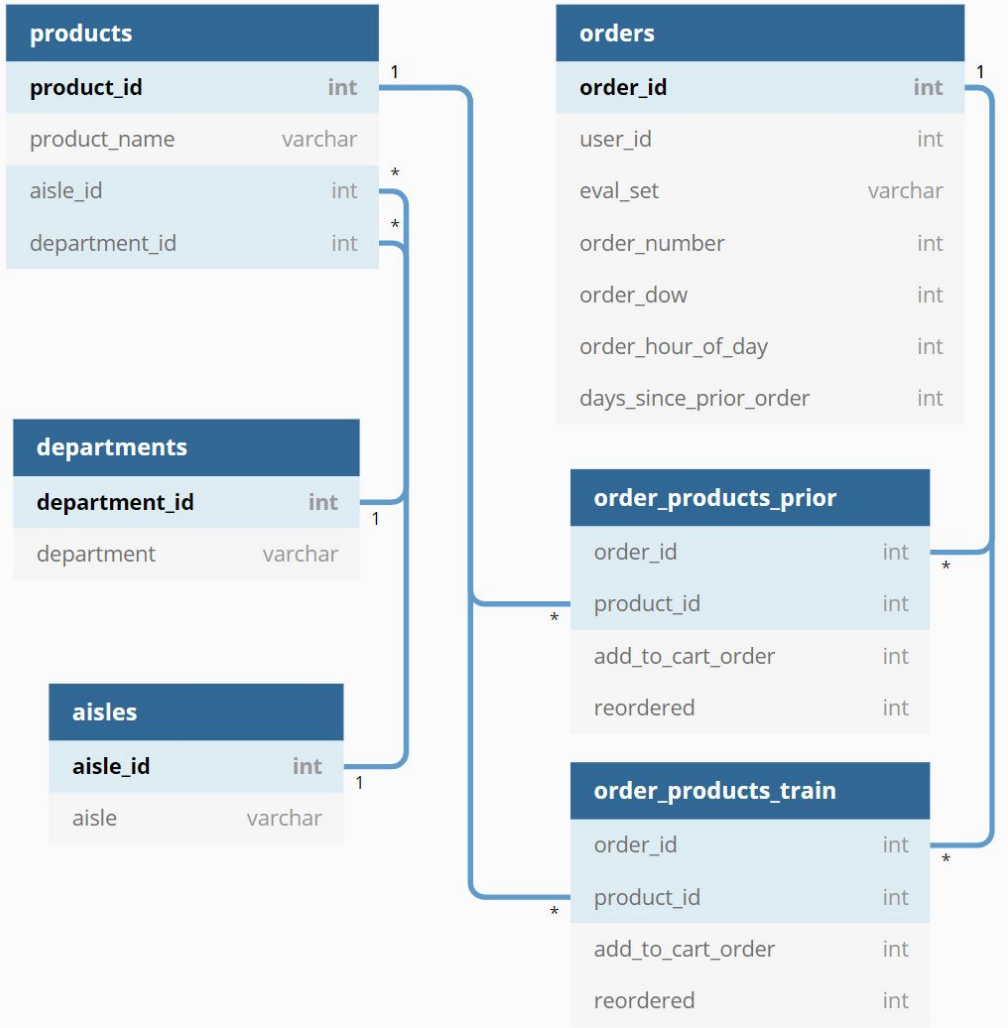
134 Aisles

49,688 Products

206,209 Users

3,421,083 Orders...

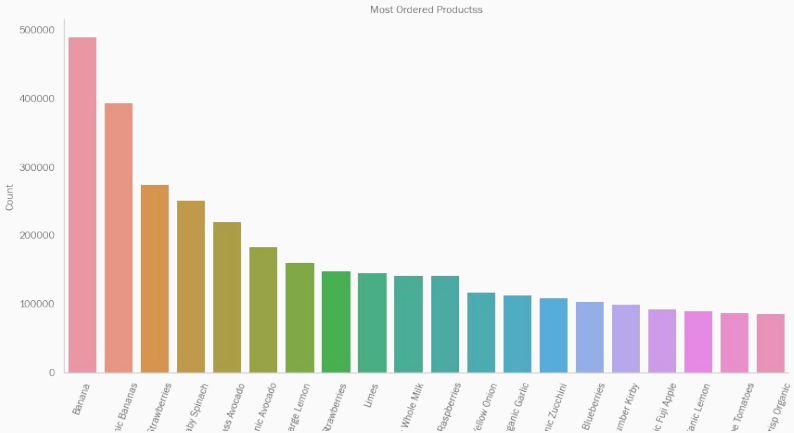
33,819,106 Ordered Products



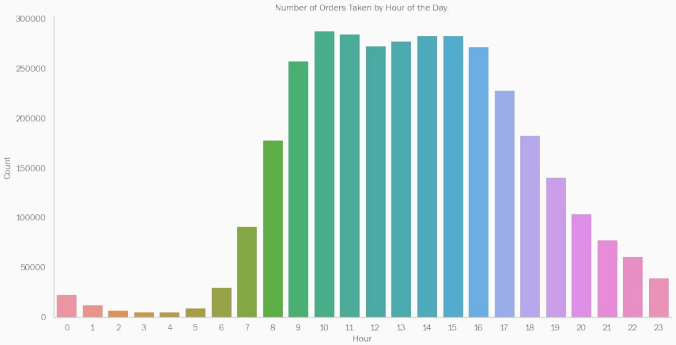
Fun Food EDA

Most Ordered Product - **Bananas**

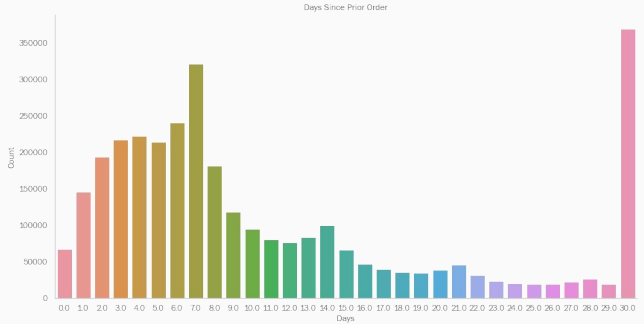
2nd Most Ordered Product - **Organic Bag of Bananas**



Order Hour of Day



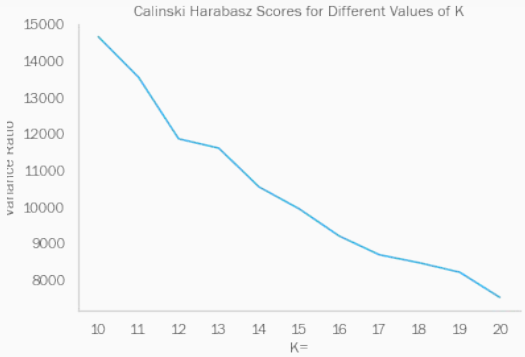
Days Since Last Order

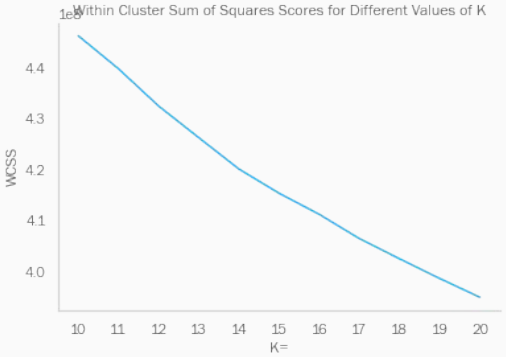


Building a multi-layered recommender

1. Cluster users for target marketing
2. Natural Language Processing search engine for products
3. SVD recommendation system
4. Market Basket Analysis to determine product lift
5. Deploy to web application

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

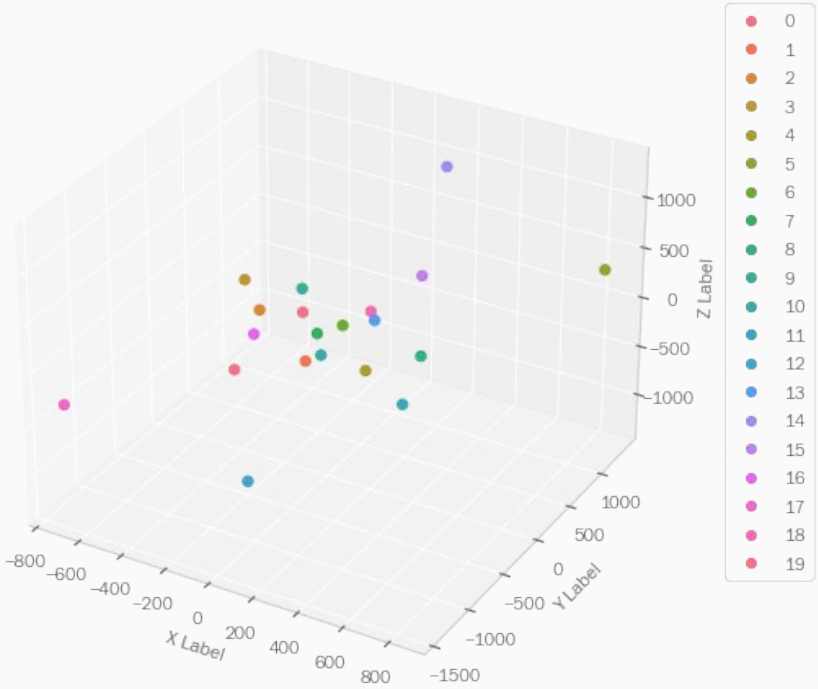
print(cluster_metrics[12]) # Very large cluster, with fewest number of orders and highest days between orders

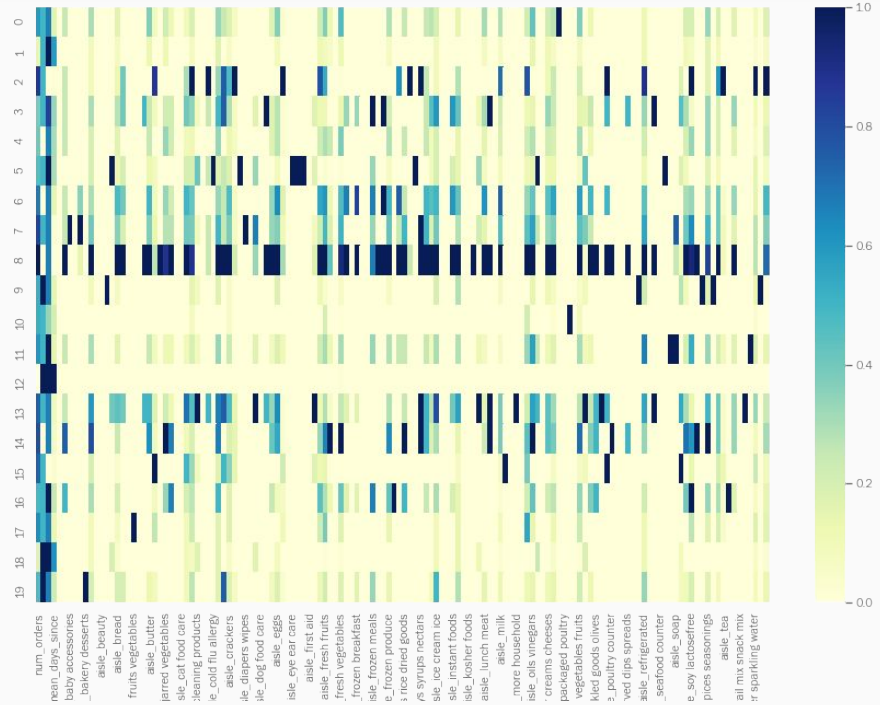
print(cluster_metrics[13]) # Household, laundry, cleaning products

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

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

Cluster visualizations





NLP 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 to existing product base.

Removed single-use words to limit size for FLASK functionality

Works like search engine optimization

Stemmed rather than lemmatized

Recommendation System

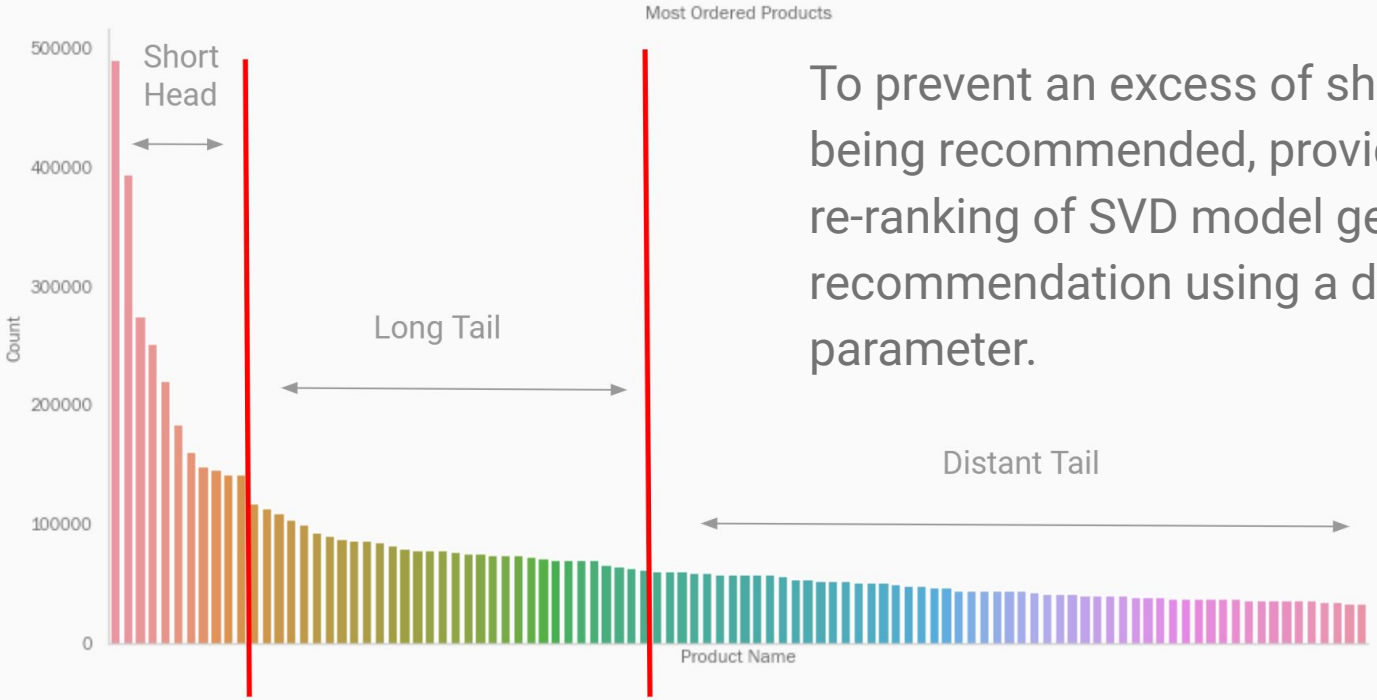
Original Surprise SVD Model with grid search

- User Id, Item Id, Number of times purchased (1-100)
- RMSE 3.25, but high values seems far off

Rescale purchase ratings to a 1-5 rating scale

- RMSE reduced to 1.26
- Add functionality to accept rating from new users

Popularity Bias - 80/15



To prevent an excess of short head items from being recommended, provide personalized re-ranking of SVD model generated recommendation using a diversity weighting parameter.

FLASK app

Live demonstration

Market Basket Analysis

Association rules indicate a strong relationship between item that customers purchased in the same transaction.

Frequency: Probability of buying a product or pair of products

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))**

Lift by Product - Cluster 19

	item_A	item_B	product_name_A	product_name_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confAtoB	confBtoA	lift
0	12191	29169	Kettle Cooked Original Potato Chips	Sea Salt & Cracked Pepper Potato Chips	3	0.000122	3	0.000122	3	0.000122	1.000000	1.000000	8189.333333
1	2202	47716	98% Fat Free Condensed Soup Cream of Chicken	98% Fat Free Condensed Soup Cream Of Celery	3	0.000122	3	0.000122	3	0.000122	1.000000	1.000000	8189.333333
850	2753	21985	Blueberry Drinkable Whole Milk Yogurt	Vanilla Whole Milk Drinkable Yogurt	3	0.000122	3	0.000122	3	0.000122	1.000000	1.000000	8189.333333
823	6907	42569	Chips Ahoy! White Fudge Chunky Chocolate Chunk...	Cinnamon Bun Sandwich Cookies	3	0.000122	3	0.000122	3	0.000122	1.000000	1.000000	8189.333333
10	5909	42436	Lemon Verbena Hand Wash	Hand Wash, Lavender Fields	3	0.000122	3	0.000122	3	0.000122	1.000000	1.000000	8189.333333
...	...	...	...	...	...	...	...	...	...	...	...	...	...
40530	13176	43965	Bag of Organic Bananas	Glazed Buttermilk Doughnuts	3	0.000122	3716	0.151254	117	0.004762	0.000807	0.025641	0.169523
29662	5450	47209	Small Hass Avocado	Organic Hass Avocado	5	0.000204	387	0.015752	1994	0.081162	0.012920	0.002508	0.159186
9782	6729	21137	Cookie Tray	Organic Strawberries	4	0.000163	302	0.012292	2496	0.101596	0.013245	0.001603	0.130370
9847	16797	21137	Strawberries	Organic Strawberries	11	0.000448	1069	0.043512	2496	0.101596	0.010290	0.004407	0.101284
2428	13176	24852	Bag of Organic Bananas	Banana	21	0.000855	3716	0.151254	3236	0.131716	0.005651	0.006489	0.042905

Lift by Aisle - Cluster 1

	aisle_A	aisle_B	aisle_name_A	aisle_name_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confAtoB	confBtoA	lift
0	28	62	red wines	white wines	1022	0.001372	4408	0.005917	3995	0.005363	0.231851	0.255820	43.233832
40	62	134	white wines	specialty wines champagnes	236	0.000317	3995	0.005363	1297	0.001741	0.059074	0.181958	33.930202
39	28	134	red wines	specialty wines champagnes	221	0.000297	4408	0.005917	1297	0.001741	0.050136	0.170393	28.796647
41	124	134	spirits	specialty wines champagnes	155	0.000208	3349	0.004496	1297	0.001741	0.046282	0.119507	26.583232
5415	27	28	beers coolers	red wines	808	0.001085	5438	0.007300	4408	0.005917	0.148584	0.183303	25.110871
...	...	...	...	...	...	...	...	...	...	...	...	...	...
157	91	124	soy lactosefree	spirits	264	0.000354	138303	0.185652	3349	0.004496	0.001909	0.078830	0.424608
5368	28	50	red wines	fruit vegetable snacks	83	0.000111	4408	0.005917	33070	0.044392	0.018829	0.002510	0.424164
158	120	124	yogurt	spirits	398	0.000534	211846	0.284373	3349	0.004496	0.001879	0.118841	0.417906
729	68	77	bulk grains rice dried goods	soft drinks	116	0.000156	3839	0.005153	60813	0.081633	0.030216	0.001907	0.370147
114	3	124	energy granola bars	spirits	111	0.000149	67726	0.090913	3349	0.004496	0.001639	0.033144	0.364572

Recommend by Product - Mild Salsa Roja

```
cluster 13 : user  125463
Crumbled Cheese, Traditional, Feta, Fat Free
Whole Wheat Pre-Sliced Mini Bagels
Fruit Squish'ems! Squeezable Fruit Pouch Apple
45 Calories & Delight Wheat Bread
Dark Chocolate, Intense Mint

cluster 14 : user  191053
Thin & Light Tortilla Chips
Boneless Skinless Chicken Breast
Organic Large Brown Grade AA Cage Free Eggs
Thick & Crispy Tortilla Chips
Organic Sour Cream

cluster 15 : user  133360
Wonderful 100% Pomegranate Juice
Frozen Limeade Concentrate
Triple Sec
Tortilla Chips Classic Yellow & Blue Family Value Pack
D'Noir Prunes

cluster 16 : user  9640
Thins Light & Tasty Snack Crackers Lightly Salted
Vegan Mac N Cheese
Lavender Harvest Aromatherapy Mist
Spinach And Cheese Ravioli
No Salt Added Organic Mixed Vegetables
```

Through the association rules, any products can be input into the model and it will generate recommendations based on the highest lift values for that product for each cluster.

In the case of salsa, we can clearly see associations with various tortilla chips.

Conclusions

- Cluster data is well-suited for market segmentation and targeted marketing
- It is important to look at the buying power of each cluster to determine how much effort to put into targeting
- The NLP search engine would be a useful tool for online shopping recommendations
- The SVD model created much better predictions when we lowered the rating scale of the products to 1-5
- Popularity bias skewed the SVD model but it can be controlled by personalized re-ranking of the recommendation

Next Steps / Future Work

- Cluster again allowing more than 20 clusters to make some of the big clusters smaller?
- Create a dashboard using DASH for graphical representation of the clusters
- Generate word embeddings for the search engine for more specific search results
- Create SQL tables of the data and load onto AWS
- Create Market Basket Analysis recommender in the FLASK app
- Use Heroku to push local FLASK app to the web

Thank you

Any questions?

You can find me at

GitHub / LinkedIn / melodygr@aol.com