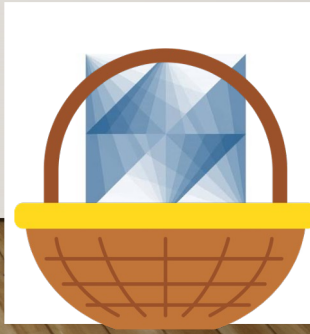
Recommendation using Market Basket Analysis



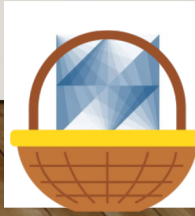
Data Description |

Data is divided across 6 files:

order_products [2 files]

- One for the basket of previously orders. [~33M rows]
- Another for the basket of the future orders.

	order_id	product_id	add_to_cart_order	reordered
0	1	49302	1	1
1	1	11109	2	1
2	1	10246	3	0



Data Description |

Aisles [~134 sub-department]

21 Departments

134 Aisles

49,688 Products

206,209 Users

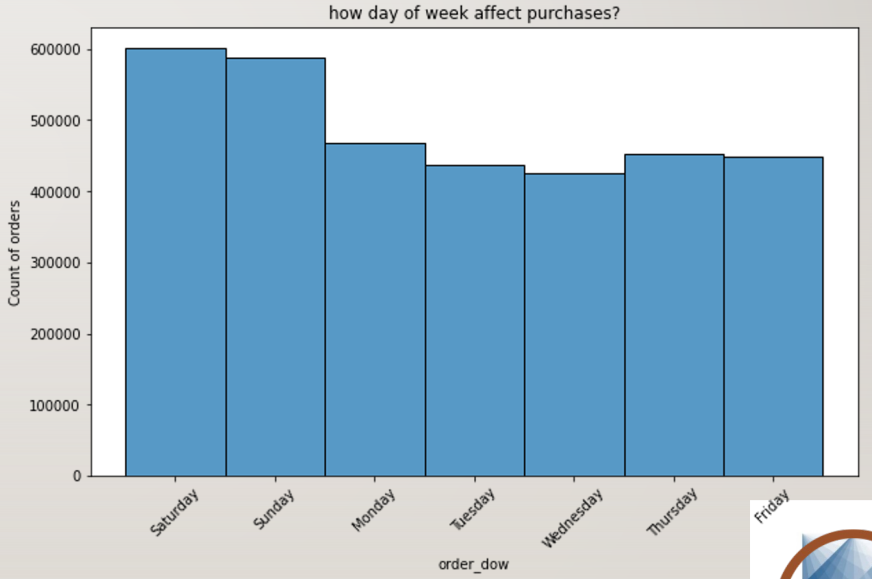
3,421,083 Orders…

Data Preprocessing

Days of week column only have numbers from 0-6

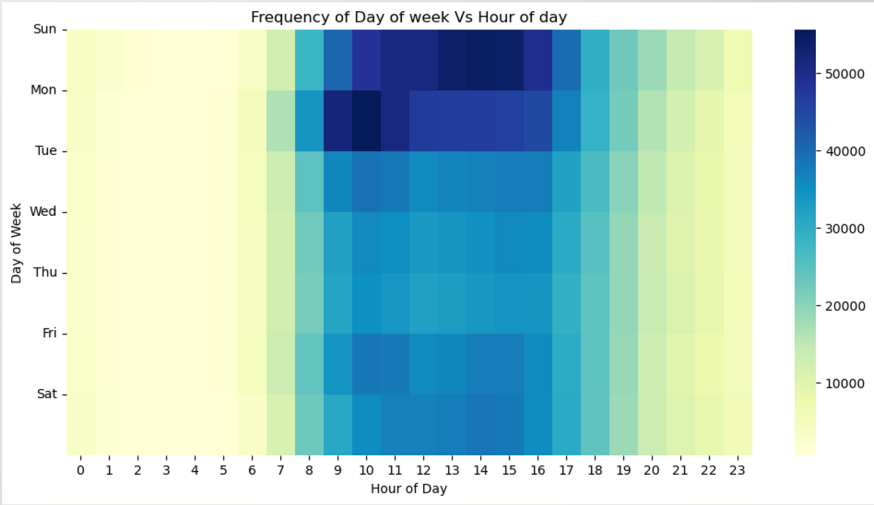
It didn't hold, which number correspond to which day of week

From the days of week histogram, Most orders were purchased on Day 0 and Day 1.



Thus we inferred that these 2 days seems to be the weekend.

EDA



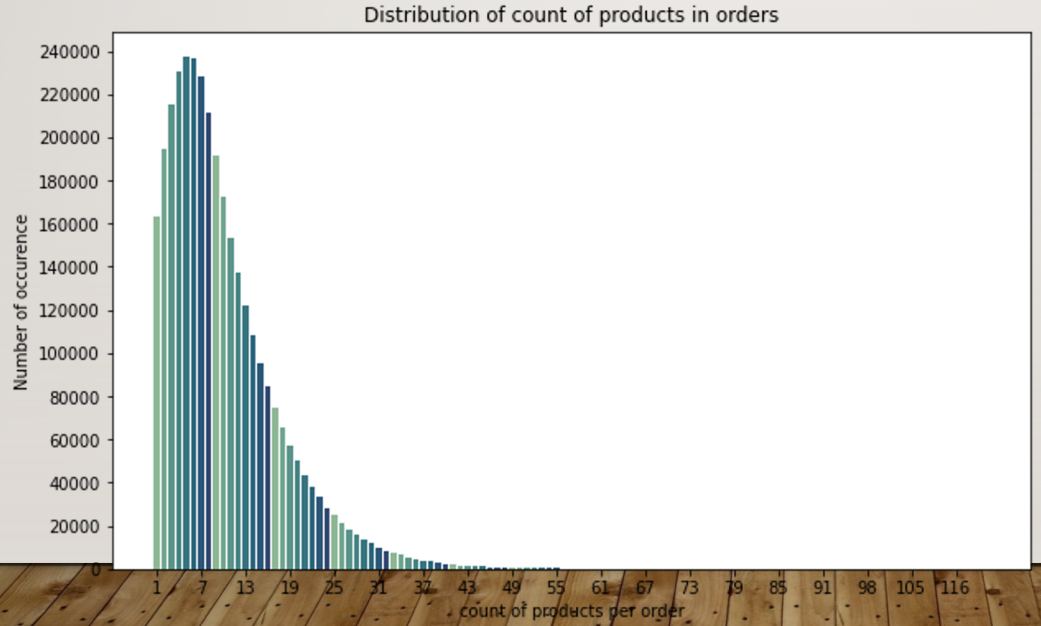
Peak ordering times are midday across all days, with evenings being the next busiest.

The least active hours are early morning (1-5 AM).

Sundays and Mondays show the highest overall ordering activity, indicating the start of the week as prime grocery shopping time.

EDA

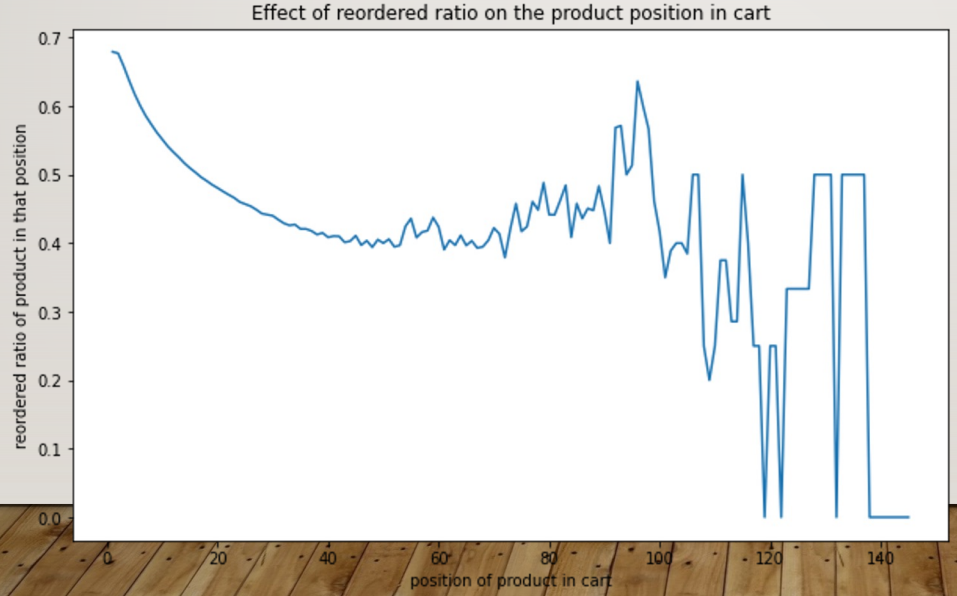
Most baskets contain from 5-8 products.



EDA

Intuitively, products placed first in cart are the products mostly reordered.

People tend to put first the products they already know.



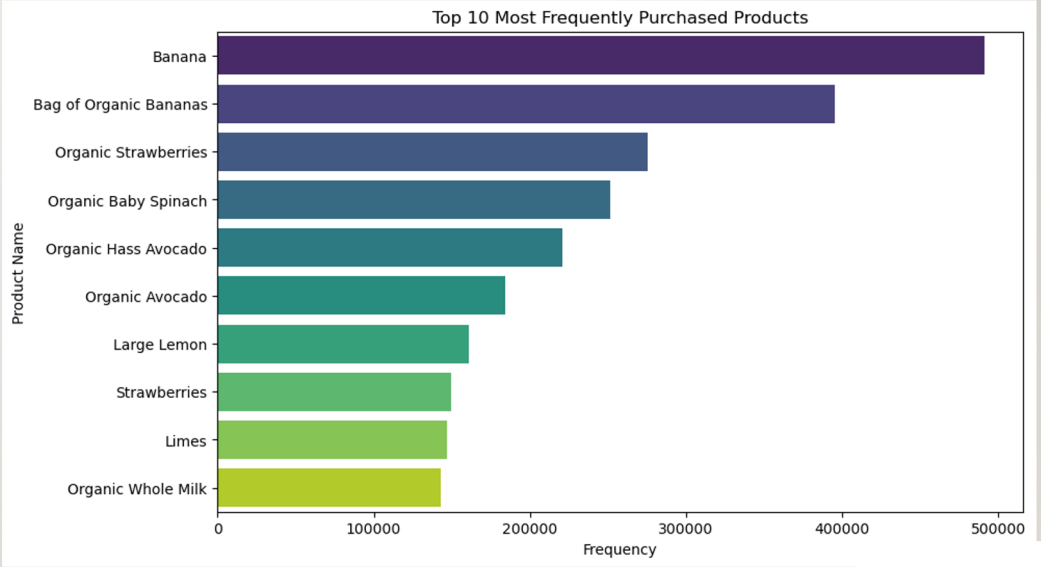
EDA

Analyzing products

5 Most Ordered Products

- Banana
- Bag of Organic Bananas
- Organic Strawberries
- Organic Baby Spinach
- Organic Hass Avocado

- 14% of all purchases are bananas.
- Organic products are frequently ordered.



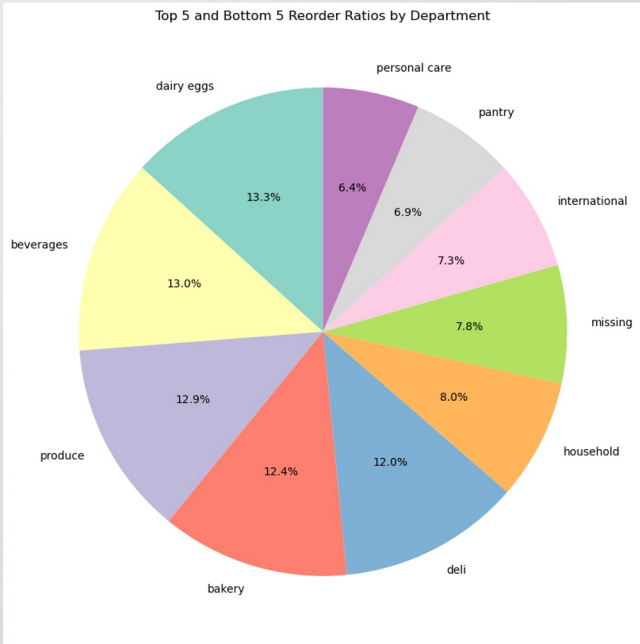
EDA

Departmental Reorder Ratios Analysis

It seems that personal contains very few reorder products (6.4% of the products).

Lowest Reorder Ratios: Departments such as international, pantry, and personal care have lower reorder ratios, which may point to less frequent needs or higher competition.

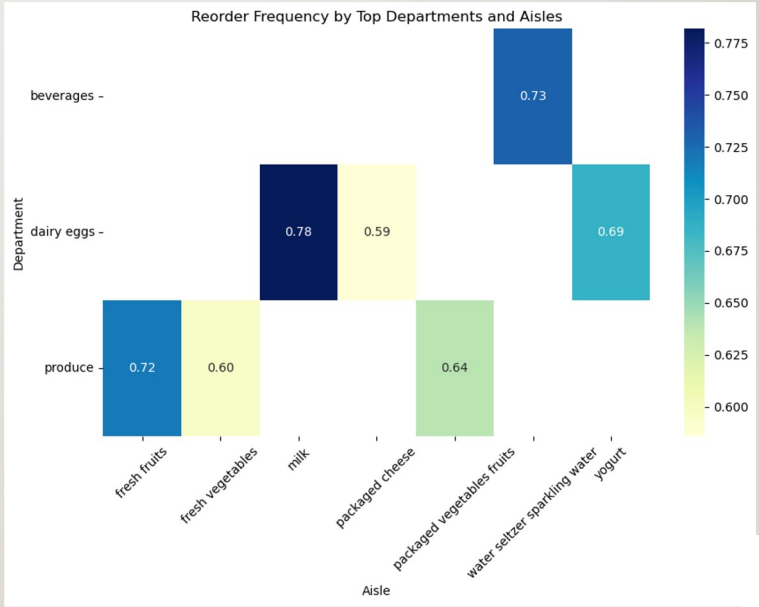
Dairy eggs and beverages departments have the highest reorder ratios, each constituting over 13% of reorders, indicating strong customer retention.



EDA

Reorder Frequency Heatmap

- Beverages and dairy products, particularly yogurt, show strong customer repurchase patterns.
- Produce items like fresh fruits and vegetables have varied reorder frequencies, indicating a mix of staple and occasional purchases.
- Prioritize stock replenishment for high-frequency items to maintain supply with demand.



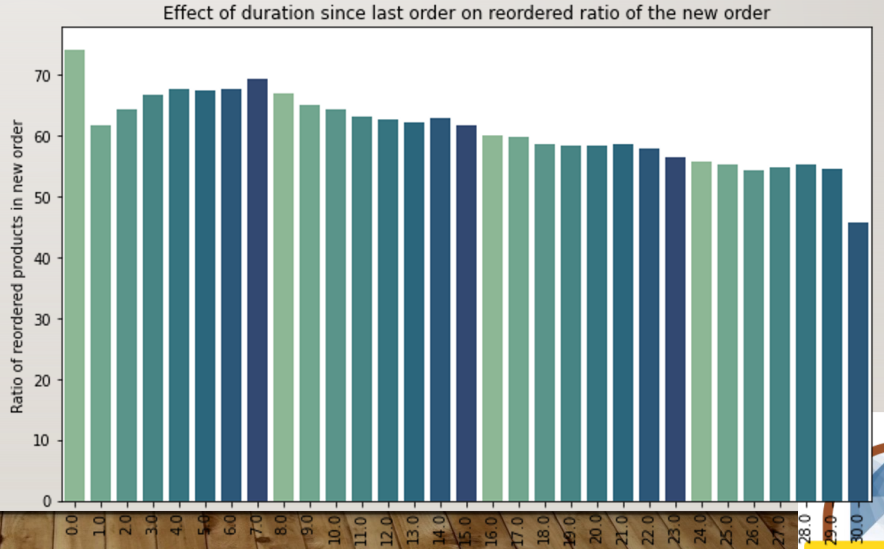
Business Questions

When to avoid recommending new products?

When it's most safe to recommend the user products they already know?

- 74 % of products bought at the same day of prev order, are reorders.
- 69% of products bought after one week from the previous order are reorders.

These are good timings to recommend products that user already knows.

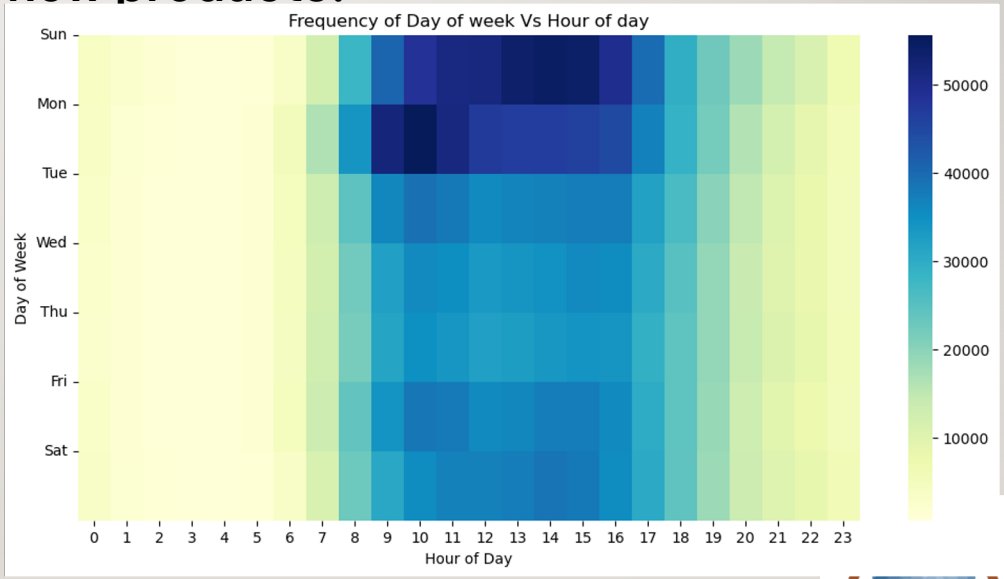


Business Questions

When to avoid recommending new products?

By more than 65%, People usually buy previously ordered products from 6:00AM to 8:00AM

Recommend previously ordered products at these hours, while avoiding recommending new products at these hours.



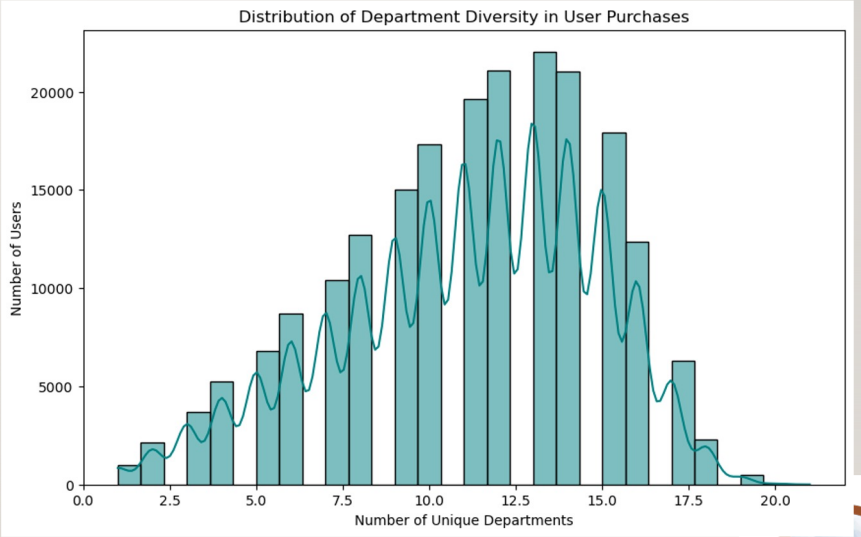
Business Questions

Analyzing Customer Purchase Diversity Across Departments

Most users purchase from a range of 10 to 15 unique departments, indicate a moderate to high level of diversity.

Fewer users shop at either a very low (0-5) or very high (15-20) number of unique departments.

Targeting users who shop from fewer departments could be a strategy to increase the breadth of their purchases by introducing them to a wider array of products.



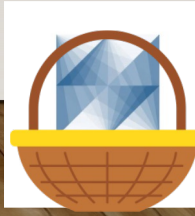
Business Questions

Who should we avoid recommending new products?

We found that 685 users always buy previously ordered products. starting from their 2nd order they never buy something new.

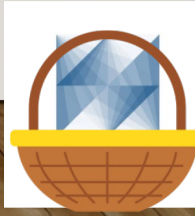
Users with **strong behavior!**

E.g. user_id 197064



Purchases of user_id 197064


```
Orders of user 197064:
Order number2981954:
['Organic White Onions', 'Natural Spring Water', '100% Natural Spring Water', 'Beef Short Ribs']
--------------
Order number1089895:
['Organic White Onions', 'Beef Short Ribs']
--------------
Order number1892685:
['Organic White Onions']
--------------
Order number1374661:
['Organic White Onions']
--------------
Order number1234679:
['Organic White Onions']
--------------
Order number849677:
['Organic White Onions']
--------------
Order number698794:
['Organic White Onions']
--------------
Order number66061:
['Organic White Onions']
--------------
Order number1923289:
['Organic White Onions']
--------------
Order number2685353:
['Organic White Onions']
--------------
Order number2401566:
['Organic White Onions', 'Beef Short Ribs']
--------------
Order number2139480:
['Organic White Onions']
--------------
Order number858962:
['Organic White Onions']
```



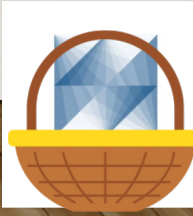
Analysis Model

Problem formulation

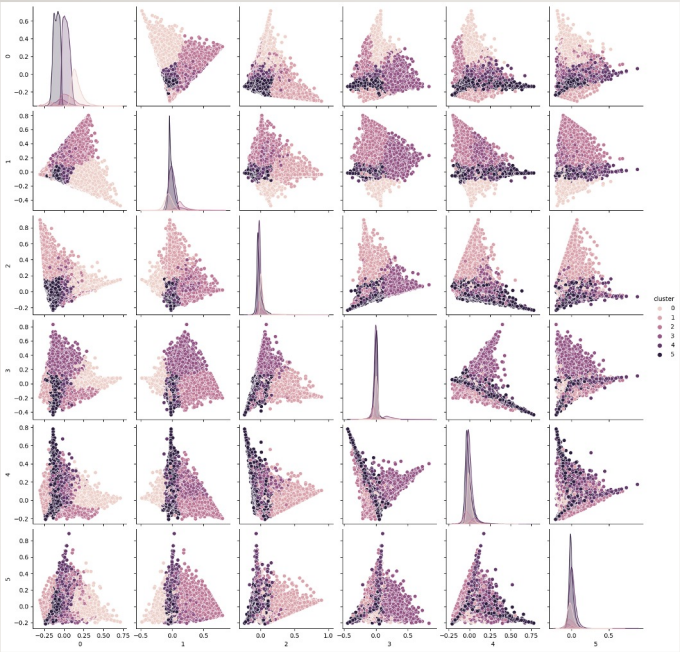
What do we want? We want to predict the products that will be in user's next future order.

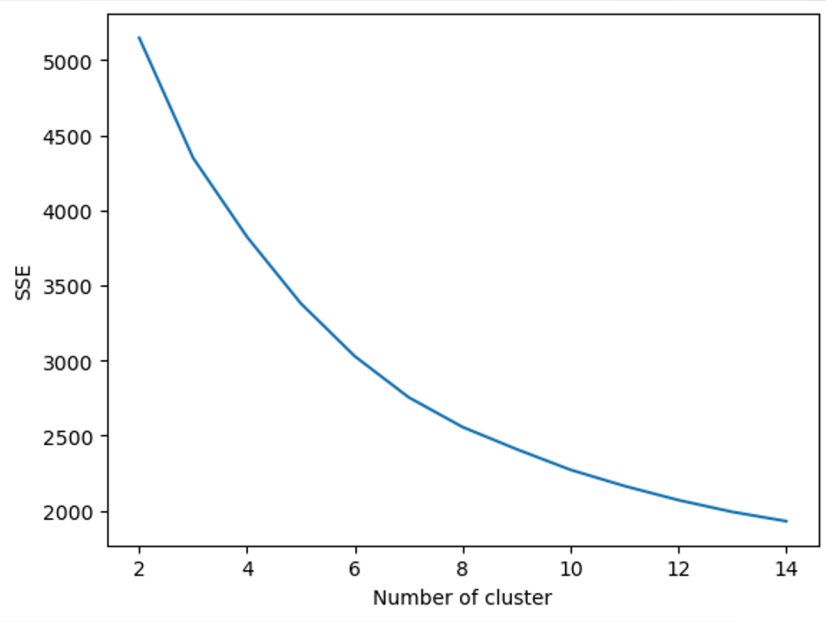
Forming the data: We take each user with his/her previously ordered products, and form each record to be a user-product pair.

Then using cluster and relation rules, we can see whether this user will order or not this product in his/her future order.
Let's extract features that relate to this user-product pair.



K-means Clustering|





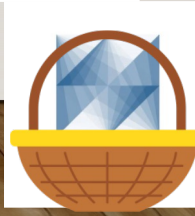
K-means Clustering |

Clustering users based on purchase behavior using K-means.

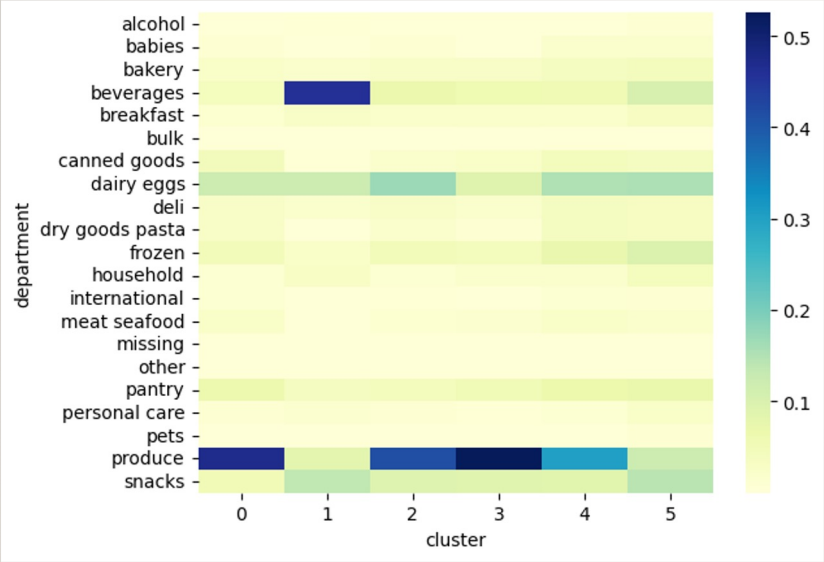
PCA components are used to define user segments.

Each user is assigned a cluster number reflecting their purchasing pattern.

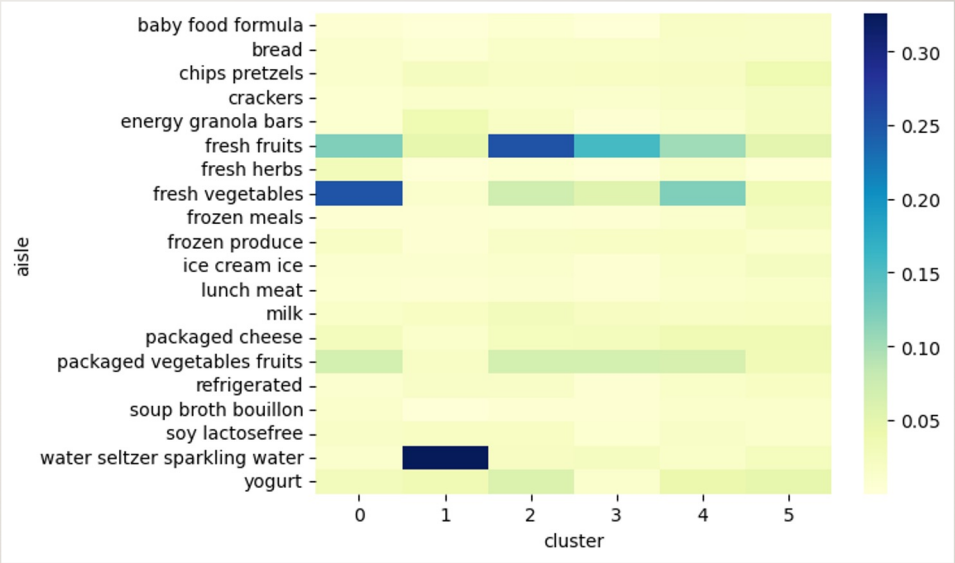
	user_id	0	1	2	3	4	5	cluster
0	1	-0.103161	0.048617	-0.031836	-0.005638	-0.052998	-0.055309	5
1	2	-0.077766	0.077983	-0.076150	-0.101671	0.081782	-0.040423	5
2	3	0.057496	0.085719	0.001779	-0.013190	-0.028506	0.113080	4
3	4	-0.055343	0.134731	-0.002393	-0.037706	-0.087800	-0.027367	2
4	5	0.133936	-0.019591	-0.020961	-0.003731	0.057733	0.138870	4
...	...	...	...	...	...	...	...	...
206204	206205	-0.021674	0.081783	-0.073479	-0.088437	0.155956	-0.022710	4
206205	206206	-0.040315	-0.036658	-0.024050	0.010060	-0.051011	0.000803	5
206206	206207	-0.022110	-0.032613	-0.009469	-0.015801	0.021487	0.020454	4
206207	206208	-0.002217	-0.010633	-0.037621	-0.007751	0.028734	0.044825	4
206208	206209	-0.072488	0.017515	-0.043591	-0.003229	-0.010472	0.027640	5



K-means Clustering |



Department Preferences by Customer Cluster



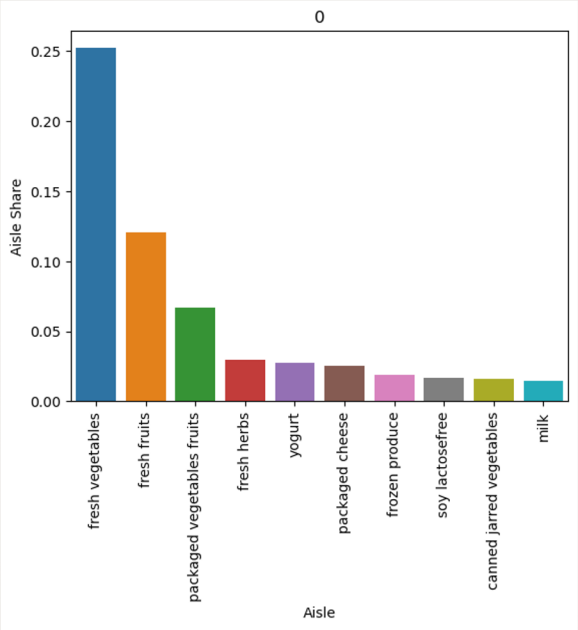
Aisle Preferences by Customer Cluster

K-means Clustering |

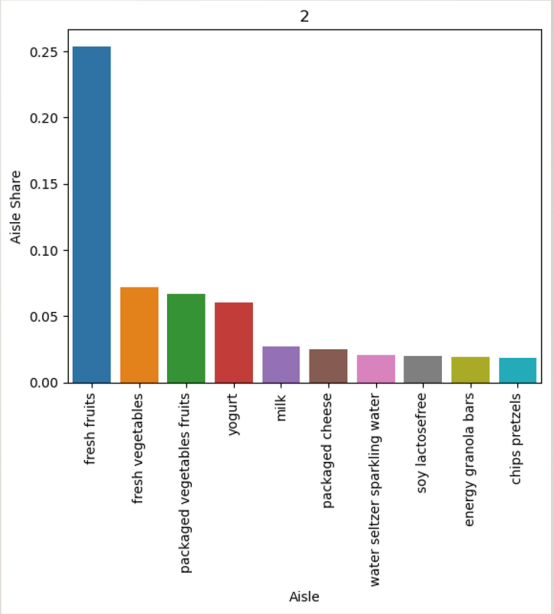
Top aisle purchases visualized for individual clusters.

Identifies cluster-specific product affinities.

Helps tailor recommendations to cluster-specific popular products.



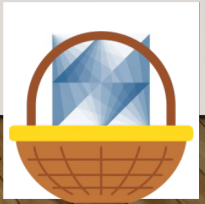
Bar Chart of Aisle Shares per Cluster 0



Bar Chart of Aisle Shares per Cluster 2

Association Rules

Association rule finds interesting association or correlation relationships among set of data items (products) which is used for decision-making processes.



Association Rules

Cluster-Specific Product Pairing

This step forms relationships between items within large datasets of transactions.

The get_item_pairs function iterates through each order to generate all possible item combinations, which are then used to find frequent itemsets.

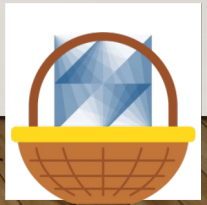
merge_item_stats function merges item frequency with the item pairs to calculate the support

```
Starting order_item:                    5576986
Items with support >= 0.0001:              7481
Remaining order_item:                   5300829
Remaining orders with 2+ items:          506419
Remaining order_item:                   5282284
Item pairs:                             4969306
Item pairs with support >= 0.0001:        91392

Starting order_item:                     218347
Items with support >= 0.0001:              4760
Remaining order_item:                    203034
Remaining orders with 2+ items:           37305
Remaining order_item:                    191853
Item pairs:                              218123
Item pairs with support >= 0.0001:        35444

Starting order_item:                    1753194
Items with support >= 0.0001:              6118
Remaining order_item:                   1640858
Remaining orders with 2+ items:          226422
Remaining order_item:                   1618518
Item pairs:                             1640144
Item pairs with support >= 0.0001:        45887

Starting order_item:                     541231
Items with support >= 0.0001:              1063
Remaining order_item:                    538140
Remaining orders with 2+ items:          100213
Remaining order_item:                    520682
Item pairs:                              151526
Item pairs with support >= 0.0001:        24498
```

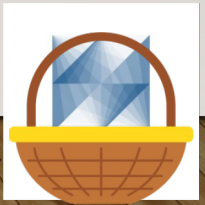


Association Rules

Cluster-Specific Product Pairing

	item_A	item_B	product_name_A	product_name_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	4376	44396	Coconut Kale & Cacao Organic Superfoods Bar	Organic Hazelnut Hemp Cacao Superfoods Bar	102	0.000102	221	0.000222	192	0.000192	0.461538	0.531250	2397.987981
1	15697	35208	Apricot Walnut & Coconut Granola Bar	Granola Bar, Fig, Cranberry & Hazelnut	132	0.000132	226	0.000227	252	0.000253	0.584071	0.523810	2312.092920
6	11224	39739	Organic Cashew Nondairy Blueberry Yogurt	Organic Nondairy Strawberry Cashew Yogurt	139	0.000139	208	0.000209	302	0.000303	0.668269	0.460265	2207.419400
8	29126	36361	Organic Strawberry Chia Lowfat 2% Cottage Cheese	Organic Cottage Cheese Blueberry Acai Chia	145	0.000145	301	0.000302	220	0.000221	0.481728	0.659091	2184.334567
9	3858	15692	2nd Foods Chicken & Gravy	2nd Foods Turkey Meat	158	0.000158	373	0.000374	201	0.000201	0.423592	0.786070	2102.289544
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8473	21903	23909	Organic Baby Spinach	2% Reduced Fat Milk	103	0.000103	19760	0.019808	23939	0.023997	0.005213	0.004303	0.217212
217	196	21137	Soda	Organic Strawberries	154	0.000154	22847	0.022903	31236	0.031312	0.006740	0.004930	0.215267
3943	23909	47209	2% Reduced Fat Milk	Organic Hass Avocado	111	0.000111	23939	0.023997	22921	0.022977	0.004637	0.004843	0.201801
280	16797	21137	Strawberries	Organic Strawberries	160	0.000160	33108	0.033189	31236	0.031312	0.004833	0.005122	0.154338
1772	13176	24852	Bag of Organic Bananas	Banana	215	0.000216	65224	0.065383	116854	0.117139	0.003296	0.001840	0.028140

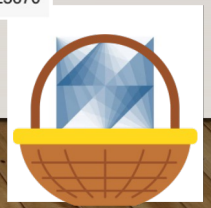
29115 rows × 13 columns



Association Rules

Cluster-Specific Aisle Pairing

	aisle_A	aisle_B	aisle_name_A	aisle_name_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	28	62	red wines	white wines	5794	0.005813	16460	0.016515	17320	0.017378	0.352005	0.334527	20.255522
55	28	134	red wines	specialty wines champagnes	1602	0.001607	16460	0.016515	6561	0.006583	0.097327	0.244170	14.784455
56	62	134	white wines	specialty wines champagnes	1643	0.001649	17320	0.017378	6561	0.006583	0.094861	0.250419	14.409945
6928	25	109	soap	skin care	592	0.000594	16142	0.016196	2808	0.002817	0.036675	0.210826	13.016971
7679	27	28	beers coolers	red wines	4228	0.004242	20495	0.020564	16460	0.016515	0.206294	0.256865	12.491077
...	...	...	...	...	...	...	...	...	...	...	...	...	...
46	57	62	granola	white wines	170	0.000171	31148	0.031253	17320	0.017378	0.005458	0.009815	0.314061
201	50	124	fruit vegetable snacks	spirits	288	0.000289	59925	0.060126	16022	0.016076	0.004806	0.017975	0.298958
171	42	124	frozen vegan vegetarian	spirits	127	0.000127	27312	0.027404	16022	0.016076	0.004650	0.007927	0.289252
165	92	124	baby food formula	spirits	169	0.000170	44413	0.044562	16022	0.016076	0.003805	0.010548	0.236702
217	57	124	granola	spirits	113	0.000113	31148	0.031253	16022	0.016076	0.003628	0.007053	0.225670

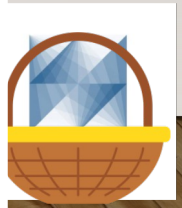


Association Rules

Cluster-Specific Aisle Pairing

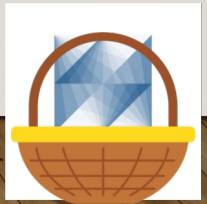
4:

	aisle_A	aisle_B	aisle_name_A	aisle_name_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	28	62	red wines	white wines	1760	0.001375	7001	0.005469	6458	0.005045	0.251393	0.272530	49.830867
1	27	62	beers coolers	white wines	774	0.000605	5513	0.004307	6458	0.005045	0.140395	0.119851	27.829078
2	6	62	other	white wines	325	0.000254	14294	0.011166	6458	0.005045	0.022737	0.050325	4.506874
3	41	62	cat food care	white wines	171	0.000134	9174	0.007167	6458	0.005045	0.018640	0.026479	3.694734
4	54	62	paper goods	white wines	774	0.000605	69409	0.054222	6458	0.005045	0.011151	0.119851	2.210401
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7560	2	6	specialty cheeses	other	539	0.000421	41906	0.032737	14294	0.011166	0.012862	0.037708	1.151868
7561	1	5	prepared soups salads	marinades meat preparation	720	0.000562	32185	0.025143	28852	0.022539	0.022371	0.024955	0.992538
7562	3	5	energy granola bars	marinades meat preparation	3044	0.002378	125964	0.098402	28852	0.022539	0.024166	0.105504	1.072176
7563	4	5	instant foods	marinades meat preparation	3124	0.002440	79529	0.062127	28852	0.022539	0.039281	0.108277	1.742824
7564	2	5	specialty cheeses	marinades meat preparation	1338	0.001045	41906	0.032737	28852	0.022539	0.031929	0.046372	1.416284



Recommendation

- Clustering Users: Segment users into clusters based on purchase behavior using K-means on PCA-reduced features.
- Mining Association Rules: Identify frequent itemsets and calculate support, confidence, and lift for each cluster.
- Building Rule Sets: Construct actionable rules from itemsets, filtering by confidence and lift thresholds.
- Developing Recommendation Engine: Create functions to match users to clusters and retrieve cluster-specific rules.
- Generating Recommendations: Recommend items with strong association rules to the input product for the user's cluster.
- Delivering Recommendations: Present a sorted list of recommended products, excluding the input item.

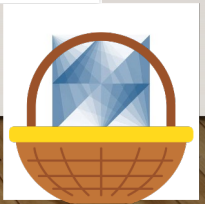


Recommendation


```
# results for users in the 6 different clusters

users = [10000,2631,954,101,481,721]
i = 0
for x in users:
    print('cluster ' + str(i) + str(pdp_recommender(x,39055,1,5)))
    i = i + 1
```

```
4
cluster 0['Thin & Light Tortilla Chips', 'Organic Large Brown Grade AA Cage Free Eggs', 'Organic Reduced Fat 2% Milk', 'Organic Large Grade AA Brown Eggs', 'Thick & Crispy Tortilla Chips']
5
cluster 1['Real Guacamole', 'Thin & Light Tortilla Chips', 'Original Hummus', 'Organic Reduced Fat 2% Milk', 'Thick & Crispy Tortilla Chips']
4
cluster 2['Thin & Light Tortilla Chips', 'Organic Large Brown Grade AA Cage Free Eggs', 'Organic Reduced Fat 2% Milk', 'Organic Large Grade AA Brown Eggs', 'Thick & Crispy Tortilla Chips']
4
cluster 3['Thin & Light Tortilla Chips', 'Organic Large Brown Grade AA Cage Free Eggs', 'Organic Reduced Fat 2% Milk', 'Organic Large Grade AA Brown Eggs', 'Thick & Crispy Tortilla Chips']
0
cluster 4['Thin & Light Tortilla Chips', 'Red Peppers', 'Organic Lemon', 'Organic Grape Tomatoes', 'Organic Small Bunch Celery']
4
cluster 5['Thin & Light Tortilla Chips', 'Organic Large Brown Grade AA Cage Free Eggs', 'Organic Reduced Fat 2% Milk', 'Organic Large Grade AA Brown Eggs', 'Thick & Crispy Tortilla Chips']
```



Conclusion

User shopping experience can be much enhanced if considered applying user-centric recommenders and associations.

From analysis, recommendation can differ according to time.