Recommendation of next item in the cart

By Jyothirmayee Nagireddy

Instacart Market Basket Analysis

The problem

What is the next product that the user might add to cart?

Instacart - online grocery delivery service

Make tailored recommendations based on user's purchase history :

- By segmenting users using K-Means clustering into various clusters.
- Applying product association rules within those clusters.

Challenge:

There is no data pertaining to what is already in the cart for the current order of each user (test data set).

The Data

- 200,000 Instacart users
- Over 3 million grocery orders
- 4 to 100 orders per each user.

Relational set of files describing customers orders over time with the following files :

- Aisles - with aisle id and aisle name.
- Products - product id, product name and aisle id to identify which aisle each product belongs to.
- Departments - department id and name.
- Order products - contains 2 files. One with prior orders and the other contains the latest order for part of the users. Contains one product per row.
- Orders - Contains order numbers for each user. Also has information related to time of day and day of week.

Data Wrangling

Aisles : There are 134 aisles with unique aisle id and aisle name.

Departments : 21 departments with department id and department name.

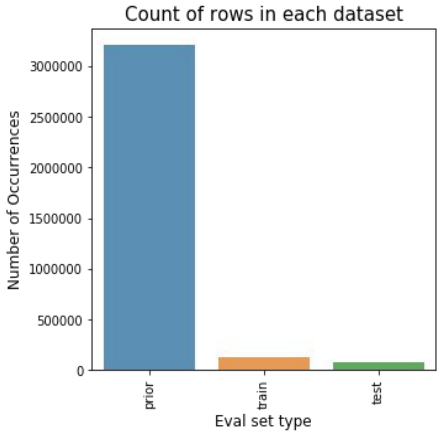
Order Products (Prior and Train) : Prior table has 32,434,489 rows and train table has 1,384,617 rows. Contains one product per order. Columns : Order id, product id, add to cart order, reordered (1 for yes and 0 for no).

Products : 49688 distinct products with product id , product name, aisle id and department id.

The data set has no missing values.

Orders : It has 3,421,083 rows with each row representing a unique order. It has 7 columns :

- Order id
- User id
- Eval set
- Order number
- Day of week
- Hour of day
- Days since prior order



Data Visualization

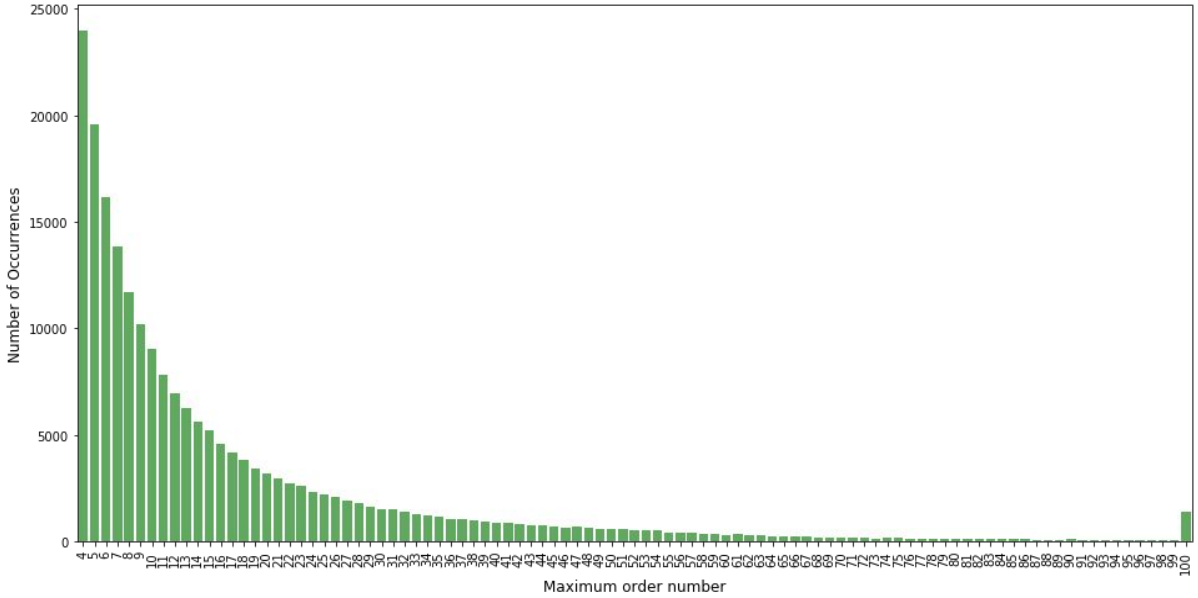
The main focus for Exploratory data analysis is the following files:

- Orders
- Order products prior
- Order products train

After analyzing both the files, all the files are merged on product id, aisle id, department id and user id.

Number of orders by each customer :

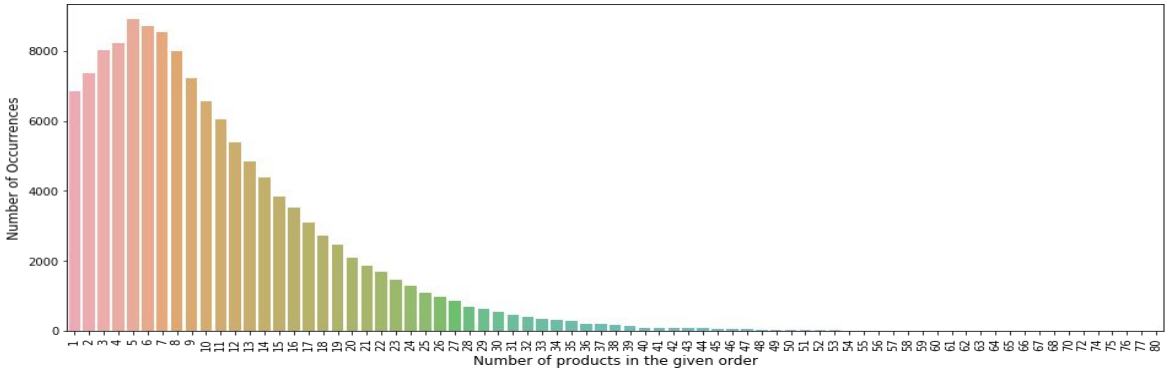
Each customer made between 4 and 100 orders.



Number of products bought in each order :

Most customers ordered between 1 and 15 products per order.

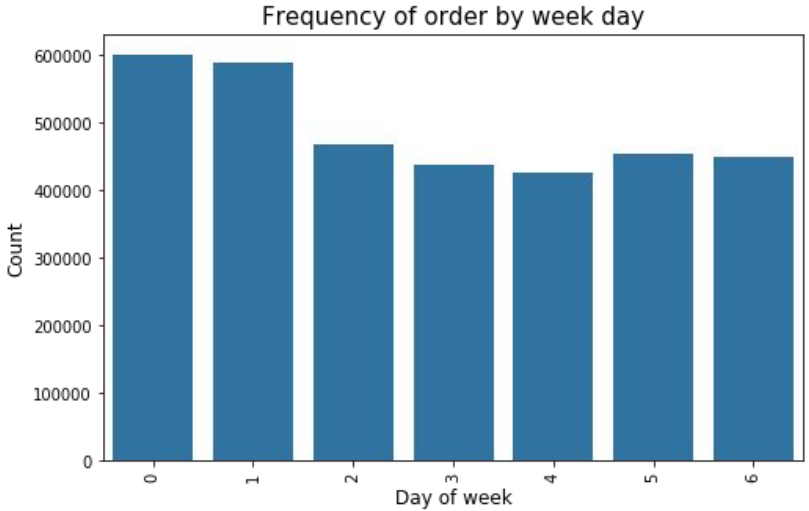
Peak at 5 products per order.

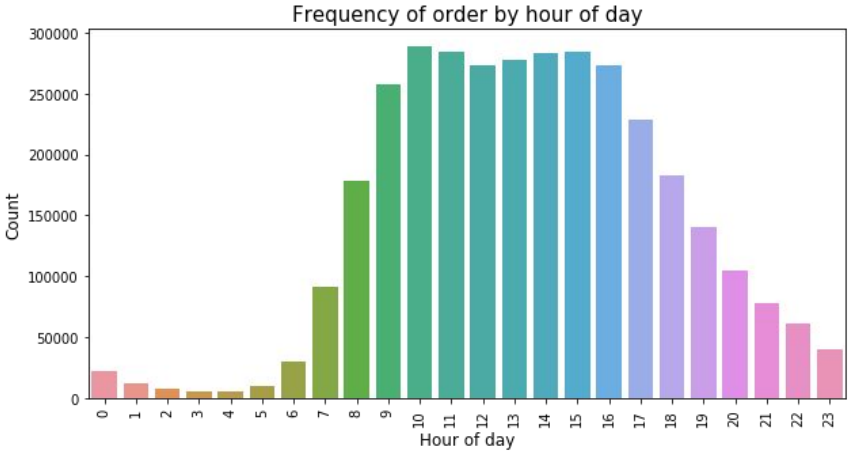


Ordering Distribution w.r.t. Day of week and Time of day

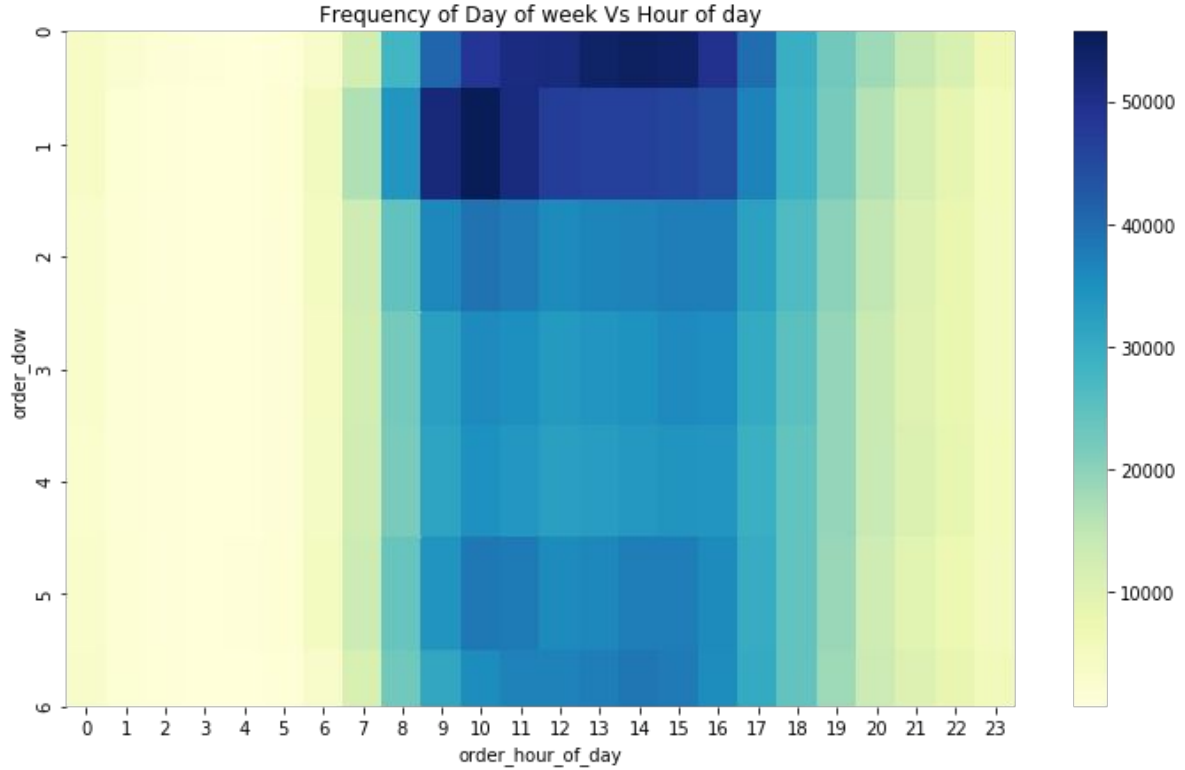
The orders are usually higher on saturdays and sundays and lowest on wednesday.

Most people order during day time. There are very few orders during night time.



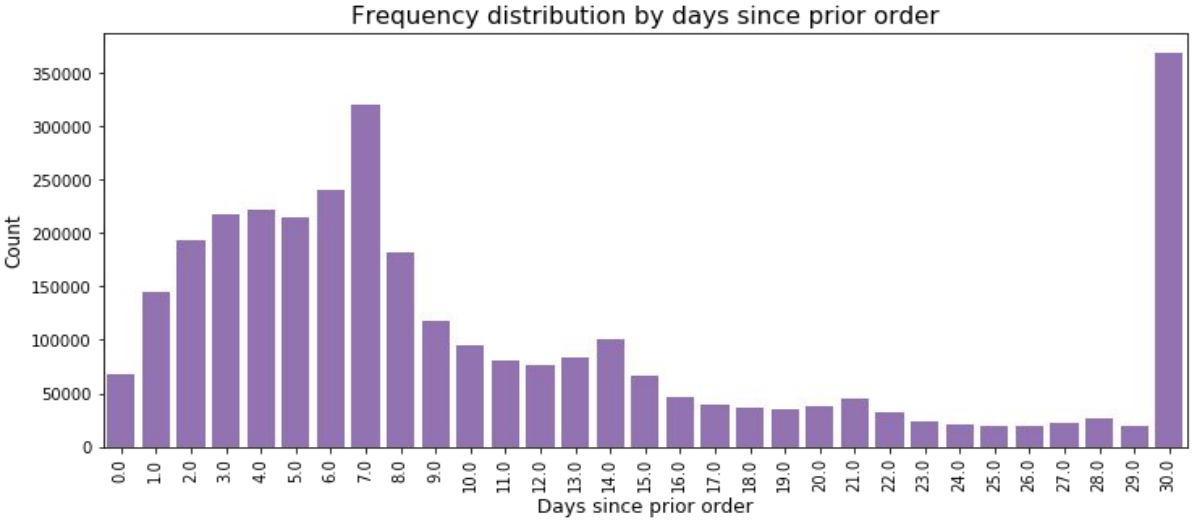


Heatmap of Time of day vs. Day of week



Prime time for orders is saturday night and sunday morning.

Time Interval between each order



Customers usually order once every 7 days or once every month.

(Peak at 7 days and 30 days)

There are also other peaks seen at 14 days, 21 days and 28 days.

Merging files

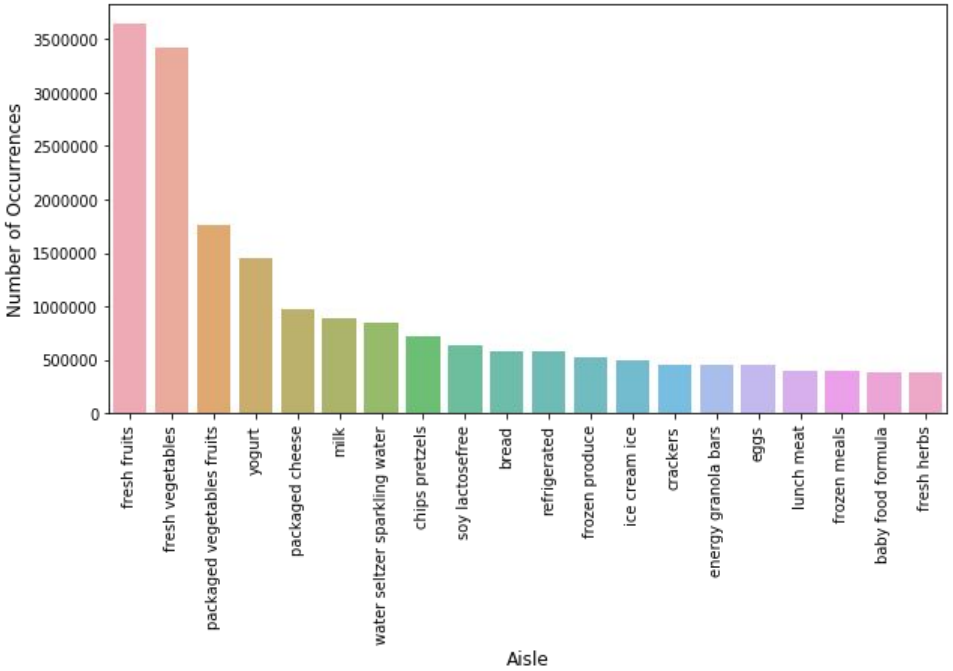
Product orders prior is merged with products, aisles and departments on product id, department id and aisle id for further EDA.

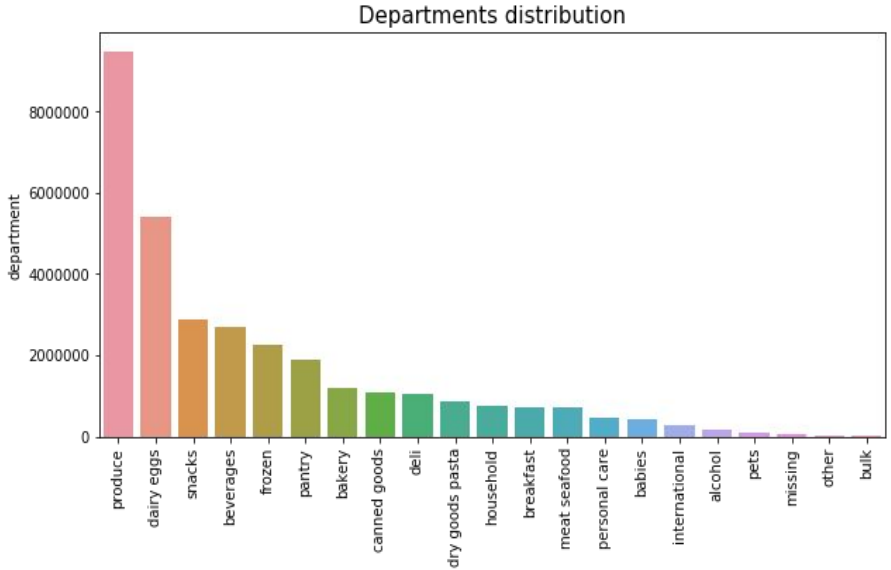
	order_id	product_id	add_to_cart_order	reordered	product_name	aisle_id	department_id	aisle	department
0	2	33120	1	1	Organic Egg Whites	86	16	eggs	dairy eggs
1	2	28985	2	1	Michigan Organic Kale	83	4	fresh vegetables	produce
2	2	9327	3	0	Garlic Powder	104	13	spices seasonings	pantry
3	2	45918	4	1	Coconut Butter	19	13	oils vinegars	pantry
4	2	30035	5	0	Natural Sweetener	17	13	baking ingredients	pantry

Top 20 Products

	product_name	frequency_count
0	Banana	472565
1	Bag of Organic Bananas	379450
2	Organic Strawberries	264683
3	Organic Baby Spinach	241921
4	Organic Hass Avocado	213584
5	Organic Avocado	176815
6	Large Lemon	152657
7	Strawberries	142951
8	Limes	140627
9	Organic Whole Milk	137905
10	Organic Raspberries	137057
11	Organic Yellow Onion	113426
12	Organic Garlic	109778
13	Organic Zucchini	104823
14	Organic Blueberries	100060
15	Cucumber Kirby	97315
16	Organic Fuji Apple	89632
17	Organic Lemon	87746
18	Apple Honeycrisp Organic	85020
19	Organic Grape Tomatoes	84255

Important Aisles and Departments





K-Means Clustering

And
Principal Component Analysis

Each user has different preferences when it comes to grocery shopping. This section primarily deals with:

- Merging all tables and preparing data for further analysis.
- PCA - feature reduction.
- Heatmap for aisle and department distribution for each cluster.
- Top aisles for each clusters.

Preparing Data for Further Analysis

Merging Tables :

Prior and train tables are concatenated together and the resulting table is merged with products, aisles, department and also with the orders table on product id, aisle id, department id and user id. And also, the clusters

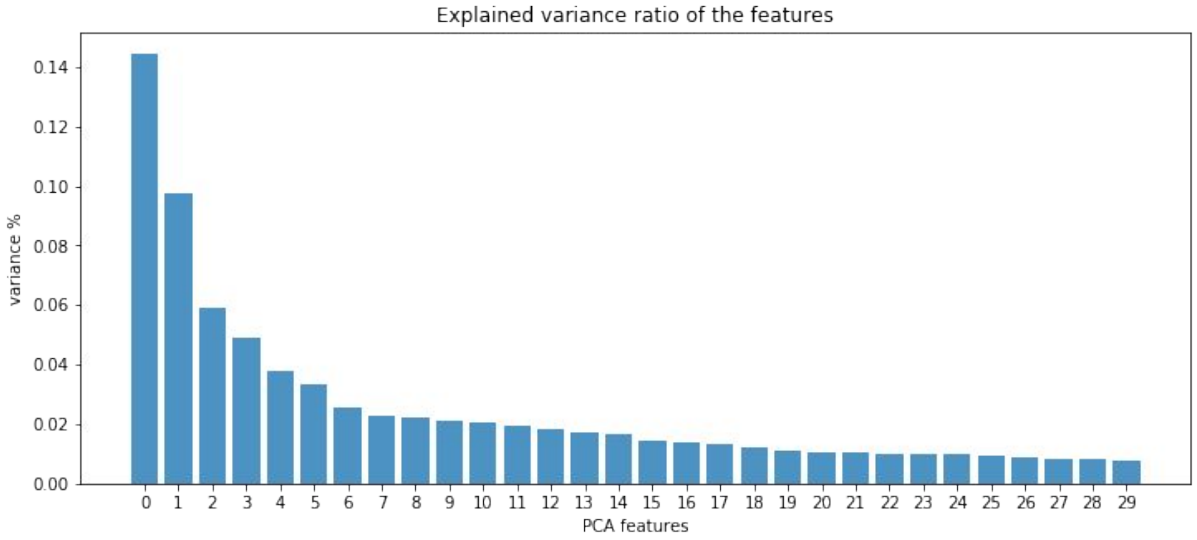
Pivot Table :

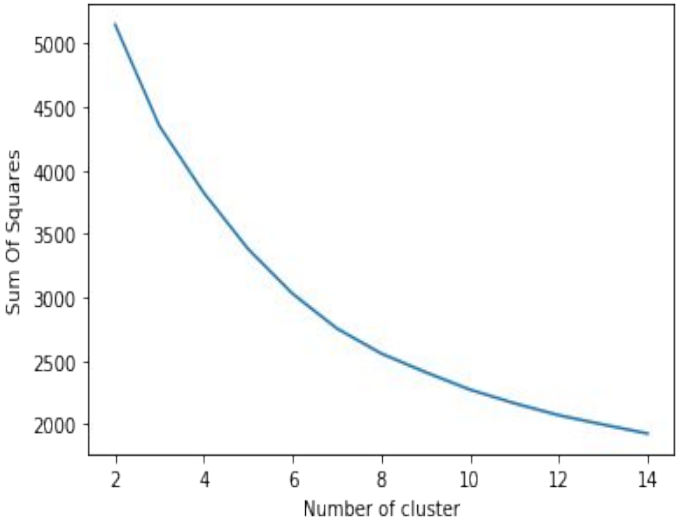
For the sake of PCA and K-Means clustering, the merged table is further pivoted to get user id in rows, aisles in columns and total number of products bought for each user for each aisle.

aisle	air fresheners candles	asian foods	baby accessories	baby bath body care	baby food formula	bakery desserts	baking ingredients	baking supplies decor	beauty	beers coolers	...	spreads	tea	tofu meat alternatives	tortillas flat bread	trail mix snack mix	trash bags liners
user_id																	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	5.0	0.0	0.0	0.0	0.0	0.0
2	0.0	23.0	0.0	0.0	0.0	0.0	9.0	0.0	0.0	0.0	...	50.0	7.0	15.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	15.0	5.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3.0	0.0	0.0
5	5.0	19.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
206205	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
206206	0.0	24.0	0.0	0.0	0.0	0.0	14.0	2.0	0.0	0.0	...	4.0	0.0	0.0	0.0	0.0	4.0
206207	0.0	0.0	0.0	0.0	23.0	0.0	0.0	0.0	0.0	0.0	...	16.0	44.0	0.0	25.0	14.0	0.0
206208	0.0	24.0	0.0	0.0	35.0	0.0	54.0	0.0	0.0	0.0	...	71.0	0.0	0.0	73.0	0.0	0.0
206209	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	6.0

206209 rows × 134 columns

Feature reduction and determining no. of clusters

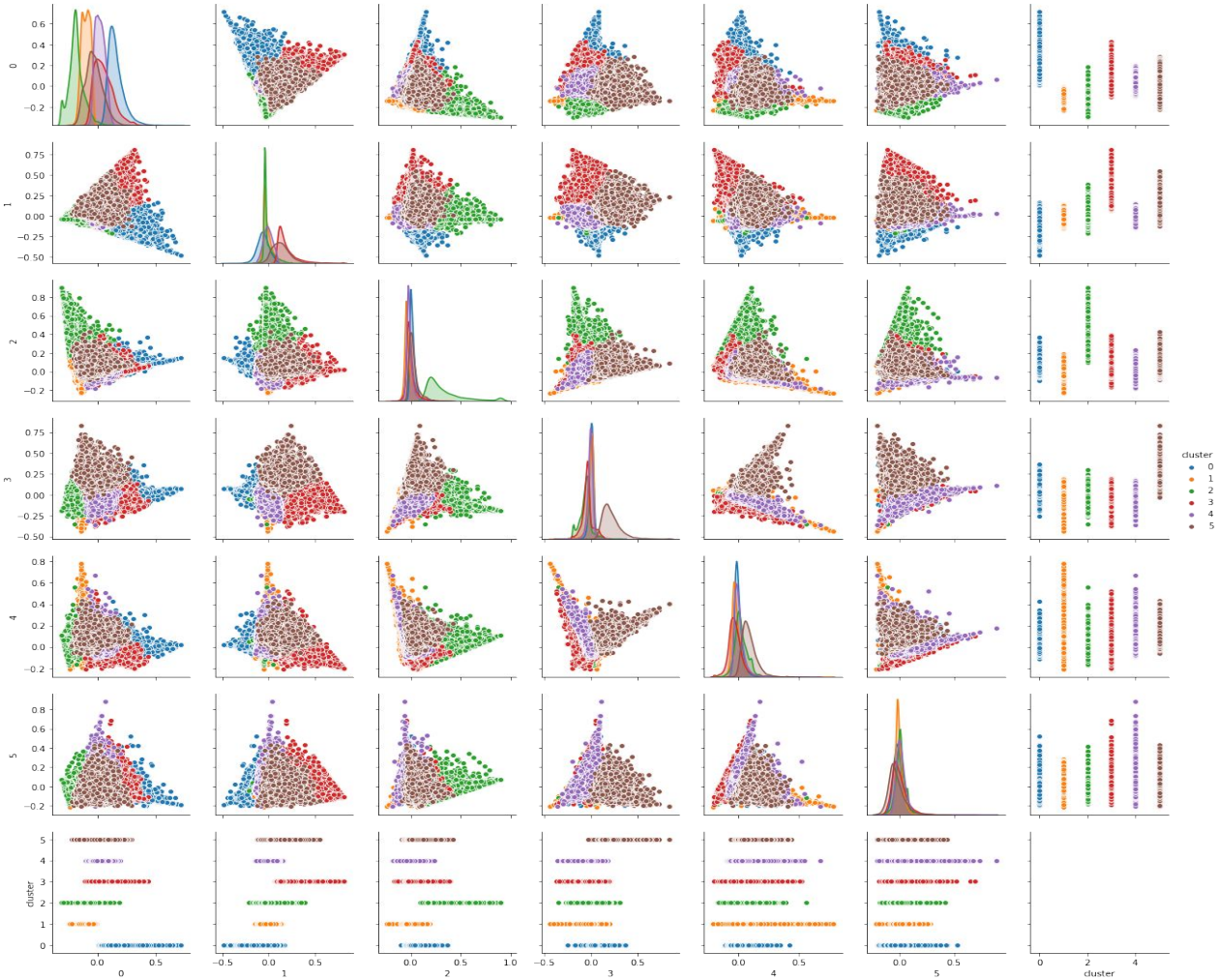




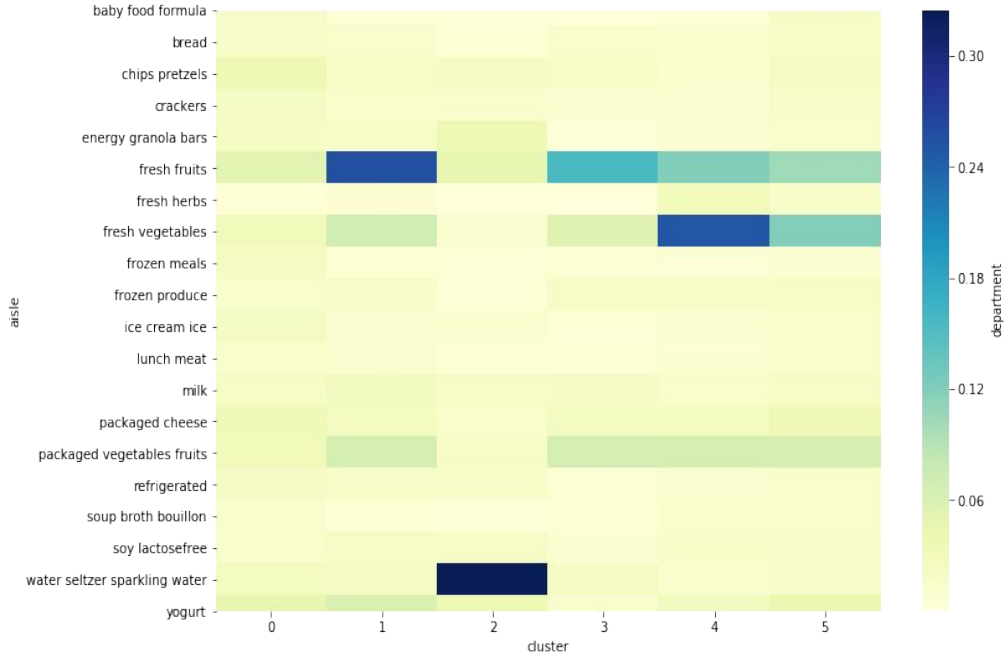
The number of features in this case would be 6 ie., after feature 5, the explained variance ratio is very low.

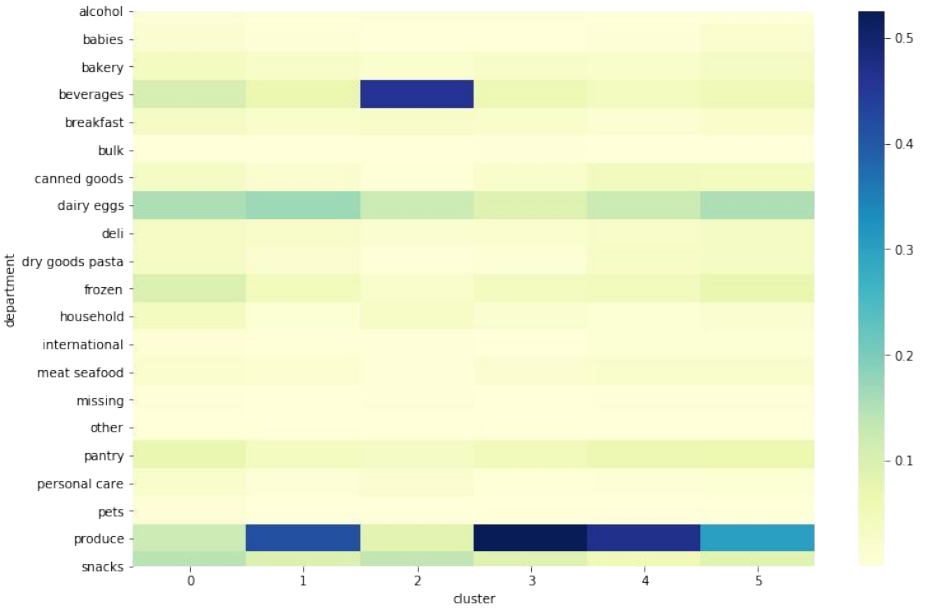
From the sum of squares curve, it seems that the curve starts to flatten. So the clusters are 0, 1, 2, 3, 4, 5.

Cluster Plot



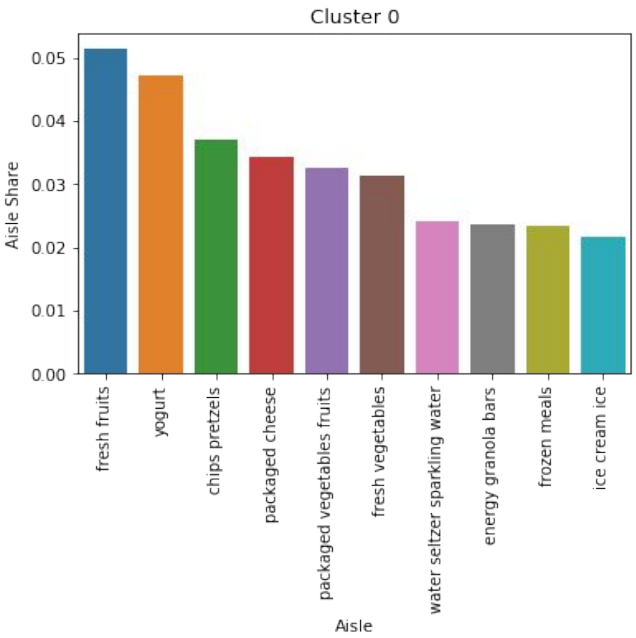
Heatmap for Aisle and department - clusters

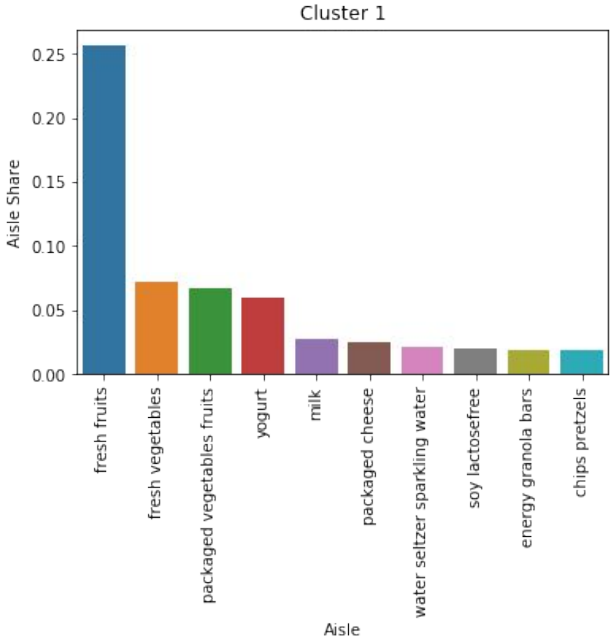


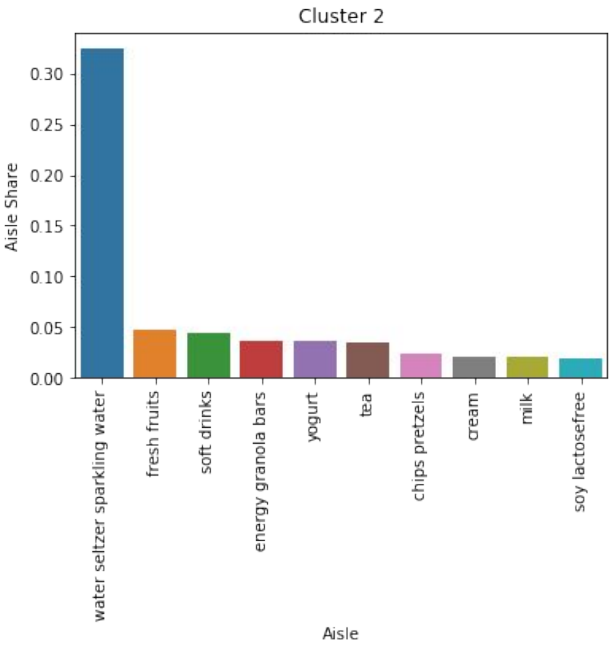


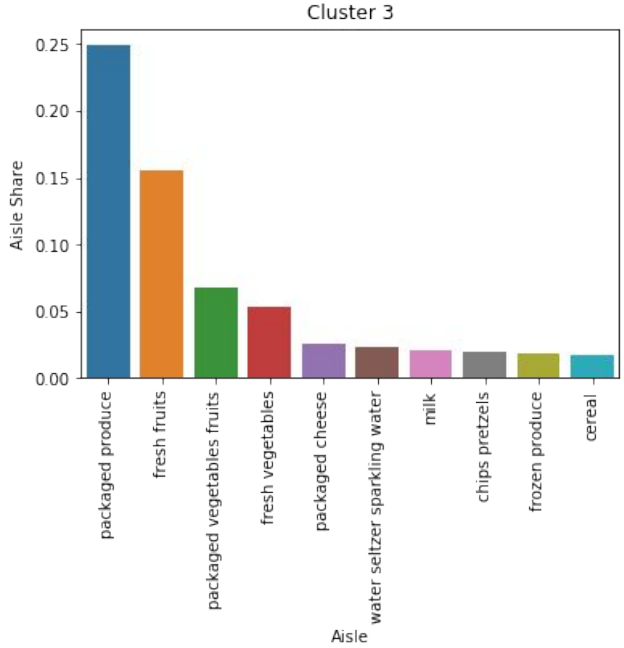
From, the above heatmap, 6 distinct clusters can be seen as obtained from PCA.

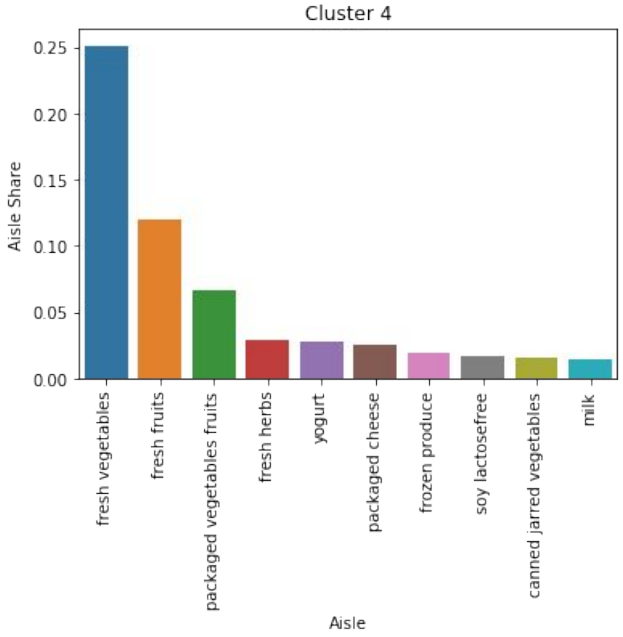
Top aisle distribution for each cluster

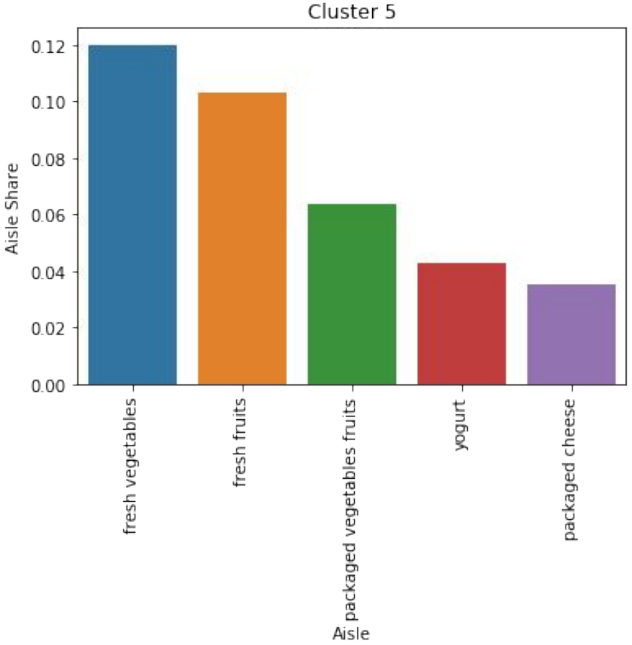







Cluster 3
Aisle Share
0.25
0.20
0.15
0.10
0.05
0.00
packaged produce
fresh fruits
packaged vegetables fruits
fresh vegetables
packaged cheese
water seltzer sparkling water
milk
chips pretzels
frozen produce
cereal
Aisle


Cluster 4
Aisle Share
0.25
0.20
0.15
0.10
0.05
0.00
fresh vegetables
fresh fruits
packaged vegetables fruits
fresh herbs
yogurt
packaged cheese
frozen produce
soy lactosefree
canned jarred vegetables
milk
Aisle


Cluster 5
Aisle Share
0.12
0.10
0.08
0.06
0.04
0.02
0.00
fresh vegetables
fresh fruits
packaged vegetables fruits
yogurt
packaged cheese
Aisle

Association Rule Mining

By using Apriori Algorithm

Apriori Algorithm

- It is used to find frequent item pairs using the “Bottom’s up” approach. It identifies the individual items that satisfy the minimum occurrence threshold.
- Then, it extends the item set, adding one item at a time and checking if the resulting item set still satisfies the specified threshold.
- Algorithms stop when there are no more items to add that meet the minimum occurrence requirement.

Key metrics when evaluating Association Rules

Support:

Support can be calculated as the fraction of orders that contain the item set.

Confidence:

Given two items, A and B, confidence measures the percentage of times that item B is purchased, given that item A was purchased. This is expressed as:

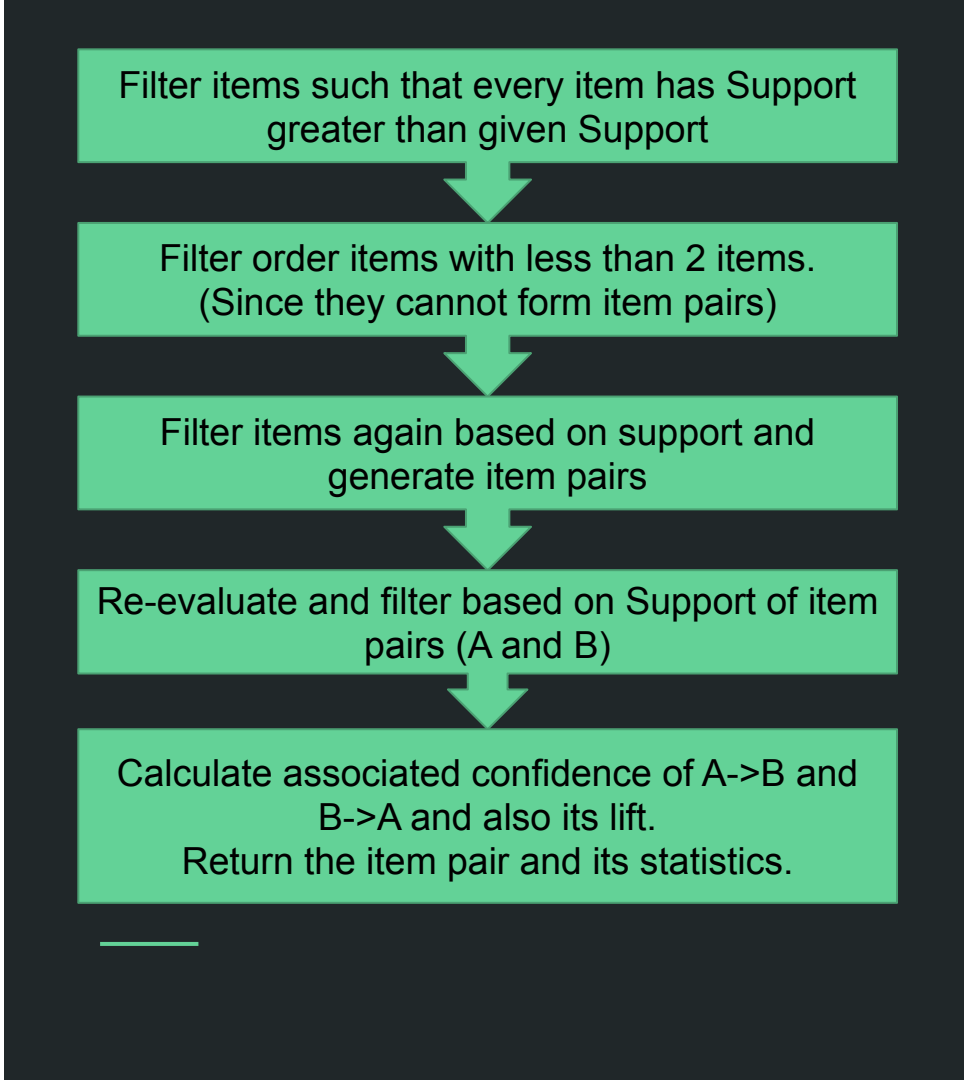
confidence{A->B} = support{A,B} / support{A}

Lift:

Given two items, A and B, lift indicates whether there is a relationship between A and B, or whether the two items are occurring together in the same orders simply by chance (ie: at random).

```
lift{A,B} = lift{B,A} =support{A,B}/(support{A} * support{B})
```

Association Rule



Association Rule Dictionary

	item_A	item_B	product_name_A	product_name_B	freqAB	supportAB	freqA	supportA	freqB	supportB	confidenceAtoB	confidenceBtoA	lift
0	20153	46949	Eat Your Colors Purples Puree Baby Food	Eat Your Colors Reds Puree Baby Food	132	0.000103	285	0.000223	227	0.000177	0.463158	0.581498	2610.737399
1	38652	29671	Yerba Mate Orange Exuberance Tea	Organic Bluephoria Yerba Mate	157	0.000123	294	0.000230	316	0.000247	0.534014	0.496835	2162.346142
5	38652	6583	Yerba Mate Orange Exuberance Tea	Oraganic Lemon Elation Yerba Mate Drink	130	0.000102	294	0.000230	293	0.000229	0.442177	0.443686	1931.027141
2	6583	29671	Oraganic Lemon Elation Yerba Mate Drink	Organic Bluephoria Yerba Mate	129	0.000101	293	0.000229	316	0.000247	0.440273	0.408228	1782.768631
14	8833	9497	Smoothie Fruits, Squished, The Green One, Over...	Smoothie Fruits Squished The Purple One Over 6...	196	0.000153	396	0.000309	377	0.000295	0.494949	0.519894	1679.884843
...	...	...	...	...	...	...	...	...	...	...	...	...	...
)91	21137	16797	Organic Strawberries	Strawberries	600	0.000469	155823	0.121779	61508	0.048070	0.003851	0.009755	0.080103
523	4605	22935	Yellow Onions	Organic Yellow Onion	154	0.000120	39411	0.030800	63689	0.049774	0.003908	0.002418	0.078505
)46	47626	5876	Large Lemon	Organic Lemon	175	0.000137	81233	0.063485	47843	0.037390	0.002154	0.003658	0.057617
336	47209	47766	Organic Hass Avocado	Organic Avocado	524	0.000410	125010	0.097698	100287	0.078376	0.004192	0.005225	0.053481
581	13176	24852	Bag of Organic Bananas	Banana	631	0.000493	191629	0.149762	239363	0.187067	0.003293	0.002636	0.017602

Product Recommender

Challenge : There is no data for the current order about what is already in the cart.

Based on past purchase history, recommendation for each customer in a cluster is given for the most bought product of that user.

Item pairs with lift lower than 1 are filtered out.

Item pair with highest confidence is returned.

Everything is compiled into a Data Frame.

Conclusion

- Users are divided into 6 clusters
- Item pairs along with confidence and lift are generated for each cluster
- Top product recommendation based on the purchase history.
- Based on EDA, weekly purchase of groceries is the usual trend.
- Top orders are usually organic vegetables and fruits.

Thank You