Market Basket Analysis

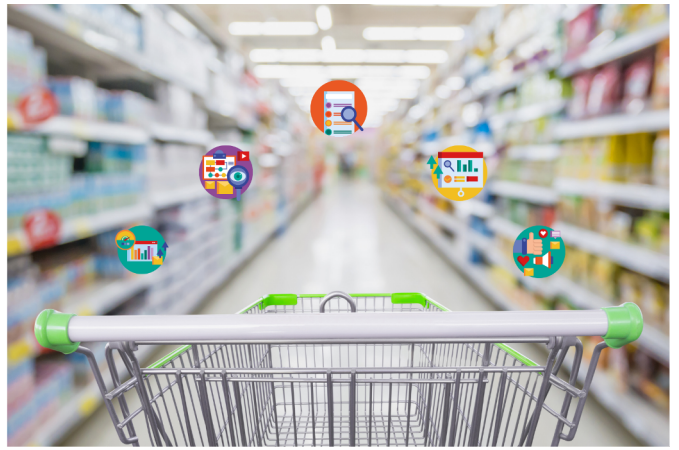
Group 5
12/09/2021

Market basket analysis

Story and Motivation

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

- Data mining technique used to better understand **customer purchasing patterns**
- Key techniques used to **uncover associations** between items
- **Analyzes customer buying habits** by finding associations between the different items that customers place in their “shopping baskets”



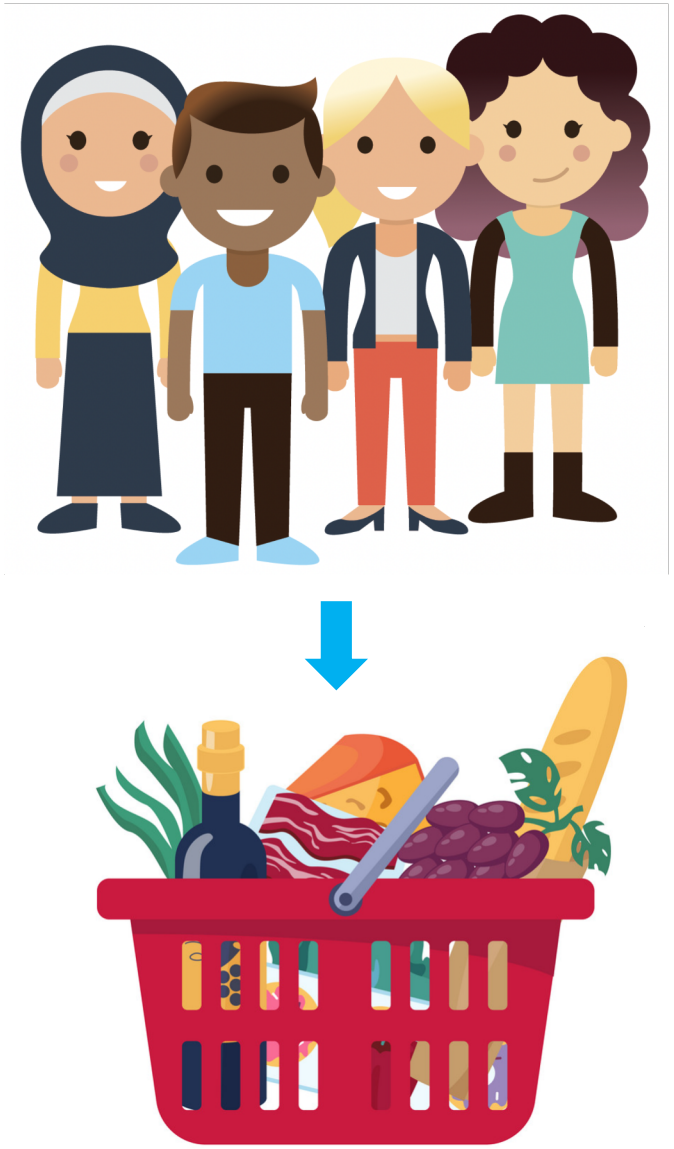
- ❏ Gives insight into which items are frequently bought together
- ❏ Helps improve marketing strategy
- ❏ Helps customer experience

Example: Instacart.com

Market basket analysis

The dataset

- Dataset used:
  - Instacart - market basket analysis
- Tables included:
  - Aisles
  - Products
  - Departments
  - Orders
  - Reorders product
  - Consumption habits (time)



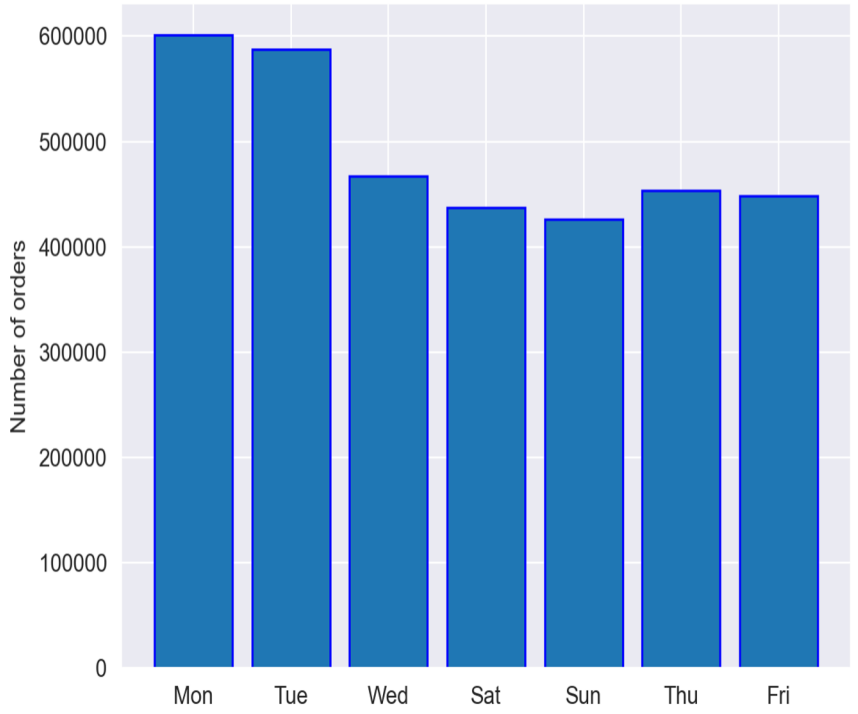
Market basket analysis

Exploratory Data Analysis

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

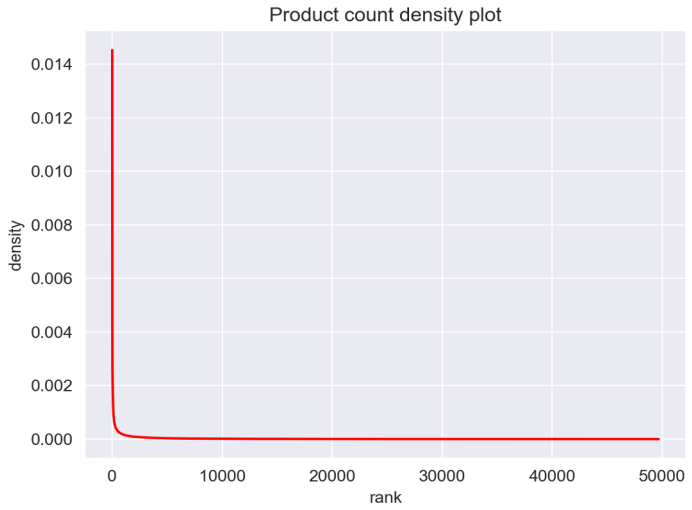
top 10 frequent products bought

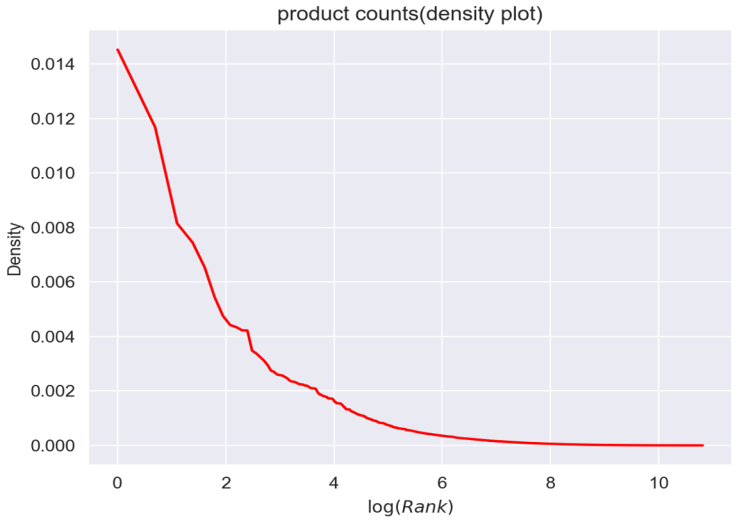
	product_name
0	Banana
1	Bag of Organic Bananas
2	Organic Strawberries
3	Organic Baby Spinach
4	Organic Hass Avocado
5	Organic Avocado
6	Large Lemon
7	Strawberries
8	Limes
9	Organic Whole Milk
10	Organic Raspberries

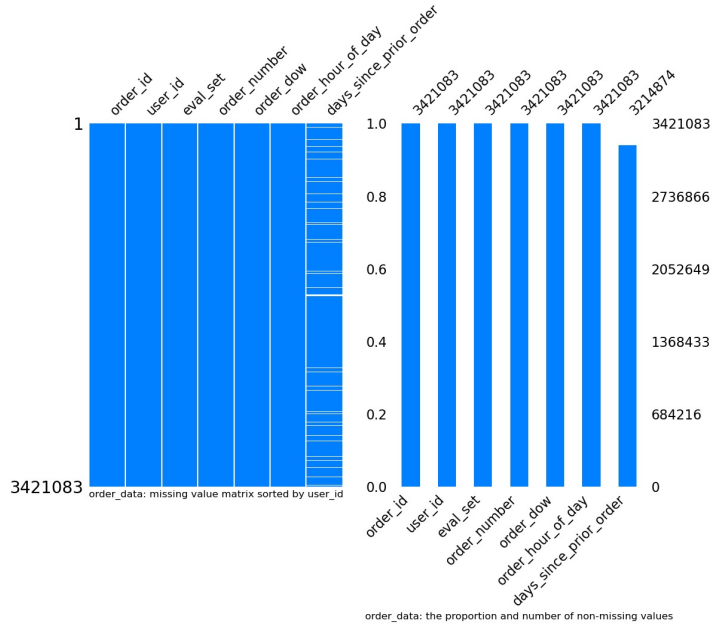


Market basket analysis

Exploratory Data Analysis







Check the missing data

Market basket analysis

Apriori/Association Rule

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

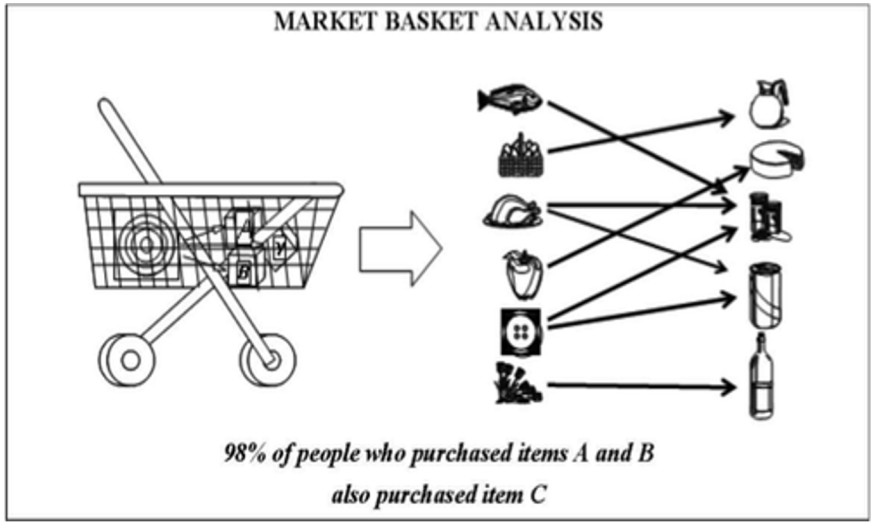
- Association rule mining:
  - Technique to identify underlying relations between various items
- Most common approach: Apriori algorithm
- Main idea: “All non-empty subsets of a frequent itemset must also be frequent”
- Bottom-up approach
  - Start from every individual item
  - Generate candidates by self-joining
  - Itemsets that contain infrequent subsets are pruned
  - Repeat until no more successful itemsets are formed

Association Mining Rule

Market basket analysis

Apriori/Association Rule

- One of the key techniques in Market Basket Analysis
- Market basket analysis tells you about items that are “frequently bought together”
- Eg., Amazon.com - “Customers also bought”,etc
- Idea of Market Basket Analysis: if item x is bought, item/ itemset y is bound to be or not be bought
- eg. , if one buys bread, chances of buying jam/butter is high



$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \rightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

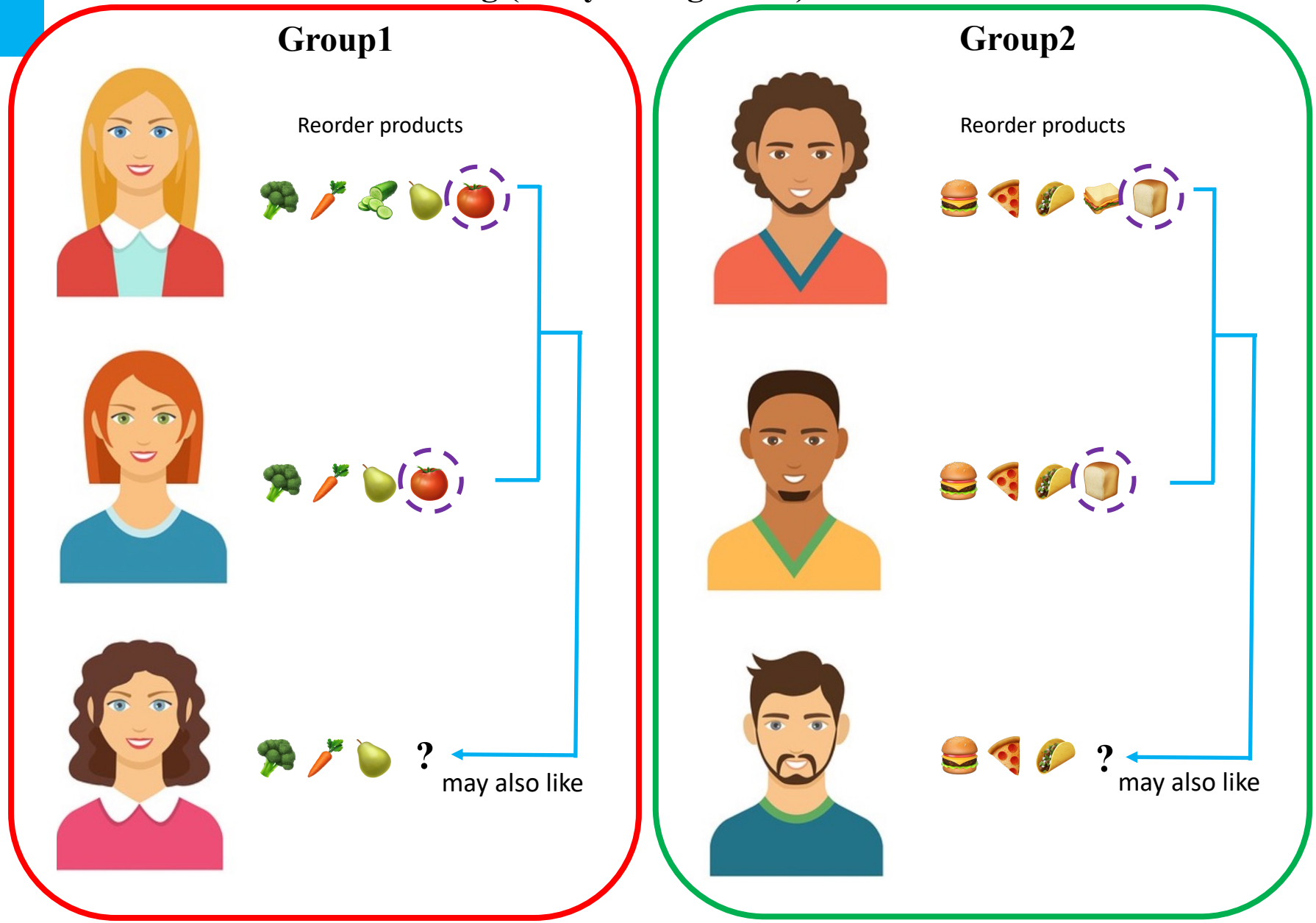
$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Market basket analysis dataset

LSH and KMeans clustering (Story Background)

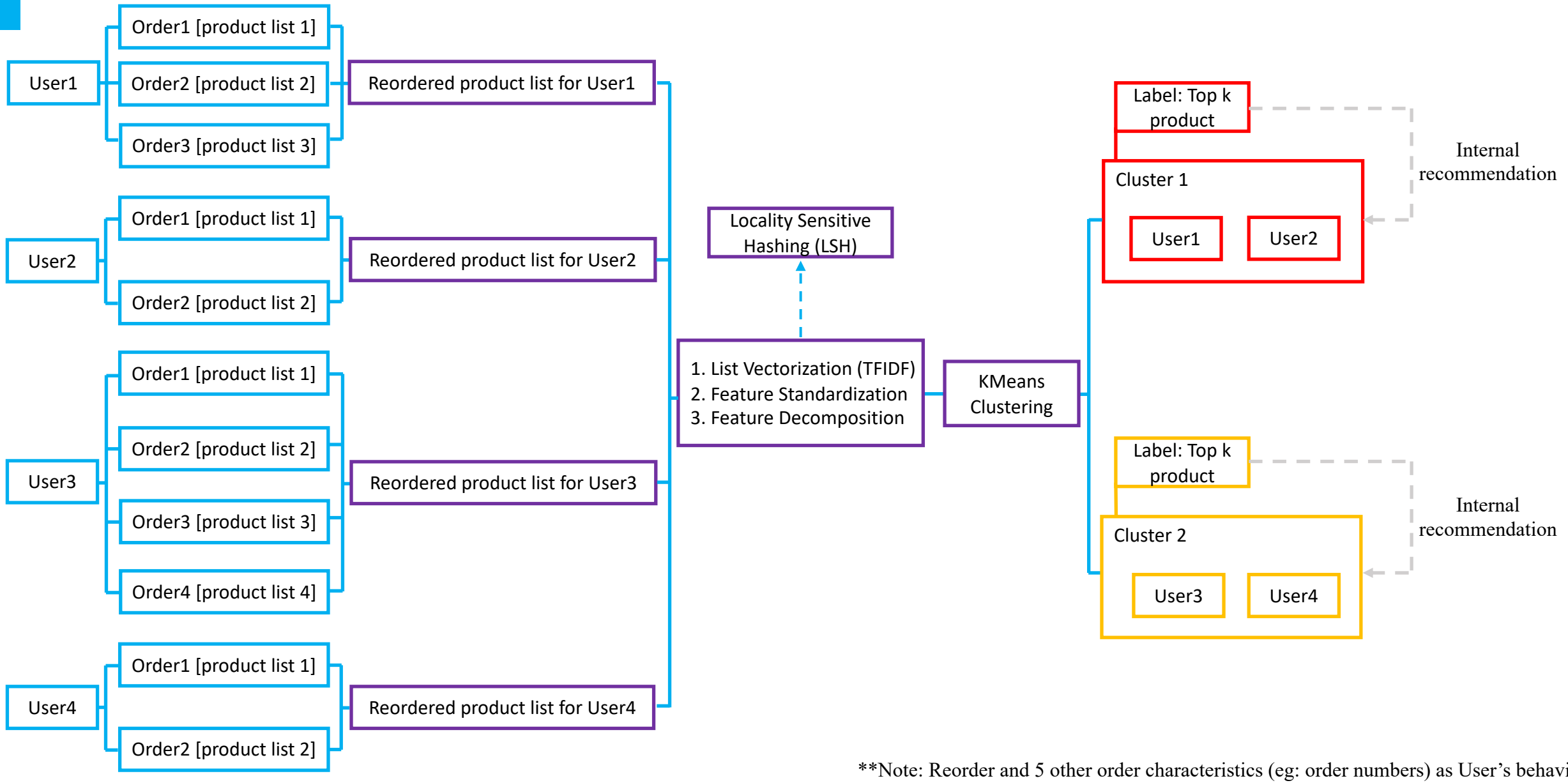


Each user has lots of orders.
The reorder products can be seen as user's favorites, which can **reflect user behavior**.

1. Based on the reorder product (user behavior) of each user, we can **classify user into different groups**;
2. Based on the **top-K favorite products of each group**, we can do an **internal recommendation**.

Market basket analysis dataset

LSH and KMeans clustering (workflow)



**Note: Reorder and 5 other order characteristics (eg: order numbers) as User's behavior

Market basket analysis dataset

LSH and KMeans clustering (results: Feature Engineering and LSH)

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Feature Generation

1000_tfidf_features

user_id	10	100	12	21	50	70	85	90	93	aa	acai	added	...	yellow	yerba	yo	yobaby	yoghurt	yogurt	yokids	yukon	zbar	zbars	zero	zucchini
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.183683	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.114357	0.0	...	0.0	0.0	0.0	0.0	0.140415	0.123297	0.0	0.0	0.0	0.0	0.000000	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.206131	0.0	0.0	0.0	0.0	0.000000	0.0

[5 rows x 1000 columns]

LSH_similar_pairs_result_hashbucket

	hash_values	similar_user_sets
0	7c1bec451b36871428c8b1af0e9ebd04	[1, 20, 24, 57, 81, 83, 107, 117, 122, 138, 18...
1	39ce64aef33c638d2fe05234a4a08caf	[2, 3, 7, 10, 47, 49, 59, 70, 101, 103, 144, 1...
2	6fe27aeda00b1f35f01a604e97674ae0	[4, 19, 28, 61, 116, 291, 377, 389, 419, 431, ...
3	69a4d6c6d0a2778f051127c13d99c671	[5, 48, 51, 54, 73, 76, 128, 153, 160, 198, 21...
4	74b870fa26e3d49d62a395a322f223a1	[6, 44, 74, 84, 100, 133, 186, 216, 217, 231, ...

1000_tdidf_features_plus_5_user_behavor_features

user_id	order_dow	order_hour_of_day	days_since_prior_order	order_num	product_num	10	100	12	21	...	yobaby	yoghurt	yogurt	yokids	yukon	zbar	zbars	zero	zucchini
1	-0.147273	-1.123285	0.591456	-0.299025	-0.333873	-0.109229	-0.485404	-0.15738	-0.144355	...	-0.119367	-0.162370	-0.682866	-0.182337	-0.153452	-0.083181	-0.08657	4.191568	-0.27855
2	-0.892007	-1.248441	0.186580	-0.116386	-0.007183	-0.109229	-0.485404	-0.15738	-0.144355	...	-0.119367	3.605954	1.547926	-0.182337	-0.153452	-0.083181	-0.08657	-0.181701	-0.27855
3	-1.696397	1.140663	-0.584408	-0.177265	-0.245918	-0.109229	-0.485404	-0.15738	-0.144355	...	-0.119367	-0.162370	-0.682866	-0.182337	-0.153452	-0.083181	-0.08657	-0.181701	-0.27855
4	1.080598	0.586981	0.775148	-0.786065	-0.585174	-0.109229	-0.485404	-0.15738	-0.144355	...	-0.119367	-0.162370	-0.682866	-0.182337	-0.153452	-0.083181	-0.08657	-0.181701	-0.27855
5	-1.479201	1.560088	0.025855	-0.664305	-0.503501	-0.109229	-0.485404	-0.15738	-0.144355	...	-0.119367	-0.162370	3.046620	-0.182337	-0.153452	-0.083181	-0.08657	-0.181701	-0.27855

[5 rows x 1005 columns]

decomposited_1005features_for_KMeans

user_id	PC1	PC2
1	-4.237737	3.268695
2	3.068878	2.148310
3	0.112483	-1.852136
4	-3.775791	-0.272410

Be further used for KMeans

```
3 user_product_ori[user_product_ori.index==1].iloc[0, 0]
```

'Soda, Original Beef Jerky, Pistachios, Organic String Cheese, Bag of Organic Bananas, Soda, Original Beef Jerky, Pistachios, Soda, Pistachios, Original Beef Jerky, Organic String Cheese, Cinnamon Toast Crunch, Soda, Original Beef Jerky, Pistachios, Organic String Cheese, XL Pick-A-Size Paper Towel Rolls, Organic Half & Half, Zero Calorie Cola, Organic String Cheese, Soda, Pistachios, Original Beef Jerky, Soda, Original Beef Jerky, Aged White Cheddar Popcorn, Soda, Zero Calorie Cola, Organic String Cheese, Pistachios, Cinnamon Toast Crunch, Original Beef Jerky, Original Beef Jerky, Soda, Pistachios, Organic String Cheese, Soda, Original Beef Jerky, Pistachios, Organic String Cheese'

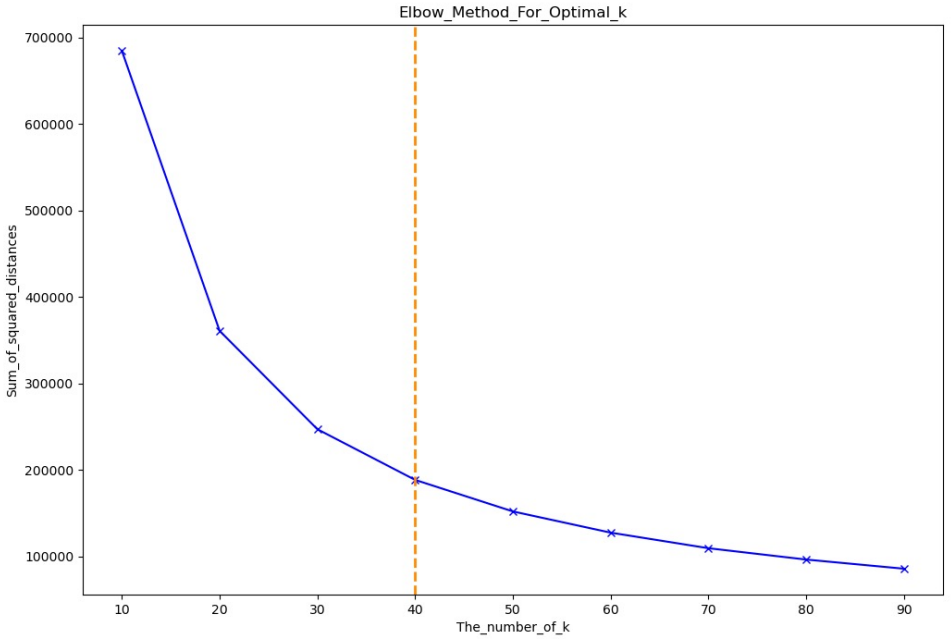
```
1 user_product_ori[user_product_ori.index==57].iloc[0, 0]
```

'Grated Parmesan Cheese, Whipped Cream Cheese, Whipped Cream Cheese, Roma Tomato, Grated Parmesan Cheese, Baked Aged White Cheddar Rice and Corn Puffs, Organic Roasted Dandelion Root Herbal Tea, Banana, Whipped Cream Cheese, Lemon Zinger Herbal Tea, Baked Aged White Cheddar Rice and Corn Puffs, Spaghetti, Organic Roasted Dandelion Root Herbal Tea, Whipped Cream Cheese, Banana, Distilled Water, Red Leaf Lettuce, 90% Lean Ground Beef, Artichoke Hearts, Whipped Cream Cheese, Baked Aged White Cheddar Rice and Corn Puffs, Banana, Whipped Cream Cheese, Roma Tomato, Organic Roasted Dandelion Root Herbal Tea, Lemon Zinger Herbal Tea, Whipped Cream Cheese, Lemon Zinger Herbal Tea, Red Leaf Lettuce, Banana, Whipped Cream Cheese, Grated Parmesan Cheese, Whipped Cream Cheese, Banana, Organic Roasted Dandelion Root Herbal Tea, Lemon Zinger Herbal Tea, Cucumber Kirby, Russet Potato, Whipped Cream Cheese, Grated Parmesan Cheese, Artichoke Hearts'

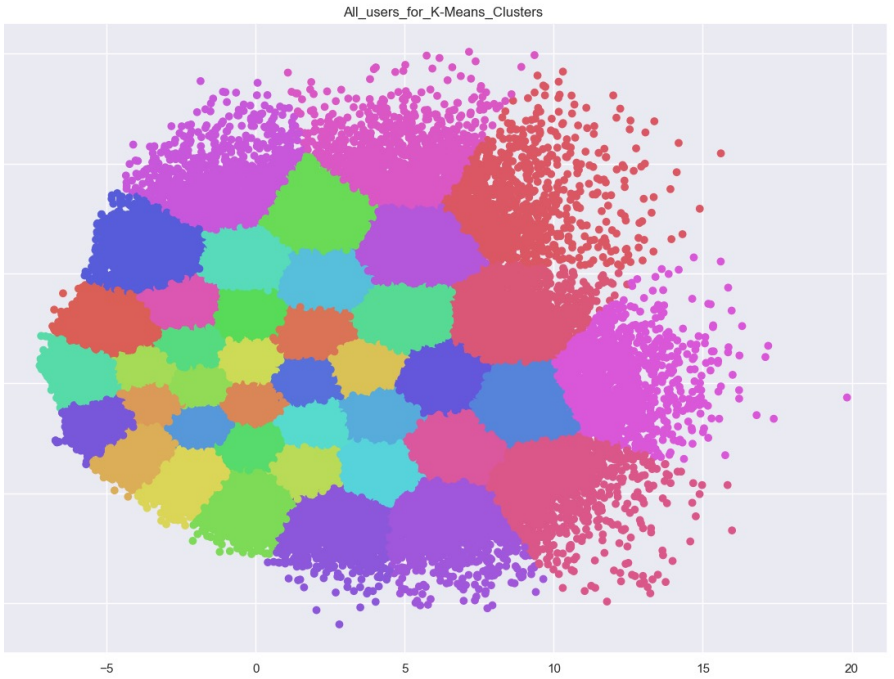
Roughly analysis by **LSH**:
The word “organic” and “cheese” may be two word put User1 and User57 together.

Market basket analysis dataset

LSH and KMeans clustering (results: Optimal K and KMeans Clustering)



Elbow Method: Screen for optimal-K



KMeans clustering of the users

Since we use **unsupervised Kmeans**, we have no idea about the true clustering, which means **how many clusters we need**.
We use **elbow method to look for the optimal K:**

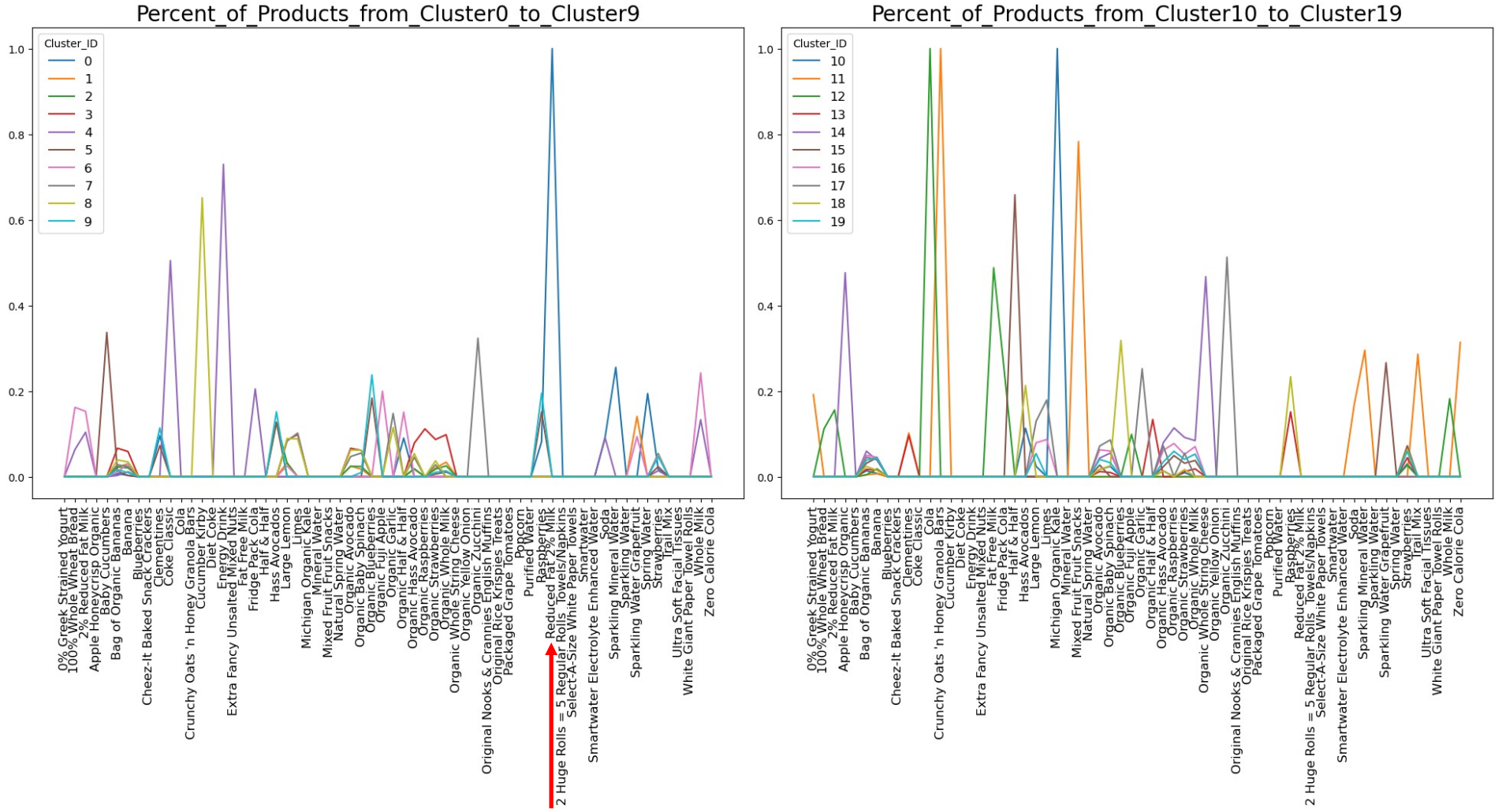
1. Calculate the **Within-Cluster-Sum of Squared Errors** (WSS: predict cluster center) for different k;
2. Choose the k for which **WSS starts to diminish (elbow)**. It could be a reasonable k for clustering.

Note: **Silhouette Analysis (the difference between its mean distance from points in the nearest cluster and its mean distance from points in its own cluster).

3. Finally, we **classify the users into 40 clusters** based on their behavior.

Market basket analysis dataset

LSH and KMeans clustering (results: TopK product for Internal Recommendation)



Top10 favorite products in each cluster comparing to other clusters

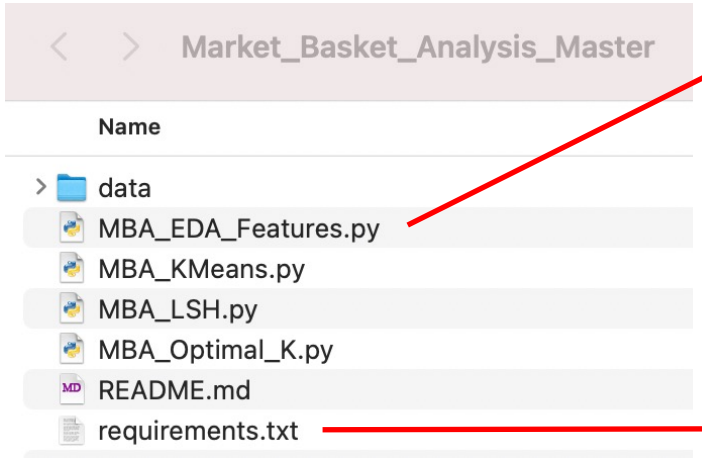
For example: In cluster 0: Reduced Fat Milk is the top 1 product, we could recommend this product to the other users of cluster 0 who have not bought it yet.

Market basket analysis dataset

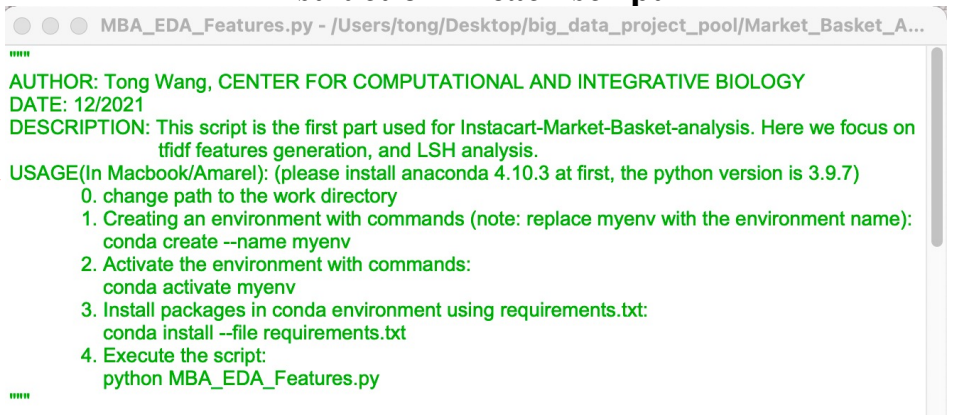
LSH and KMeans Scripts

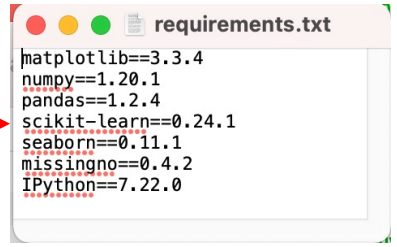
RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

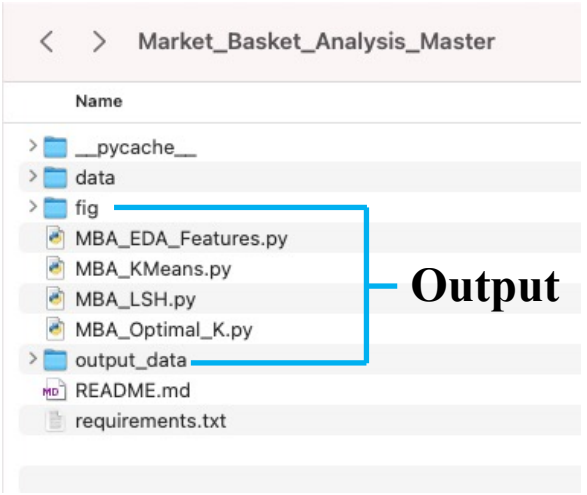
Working directory for the Scripts



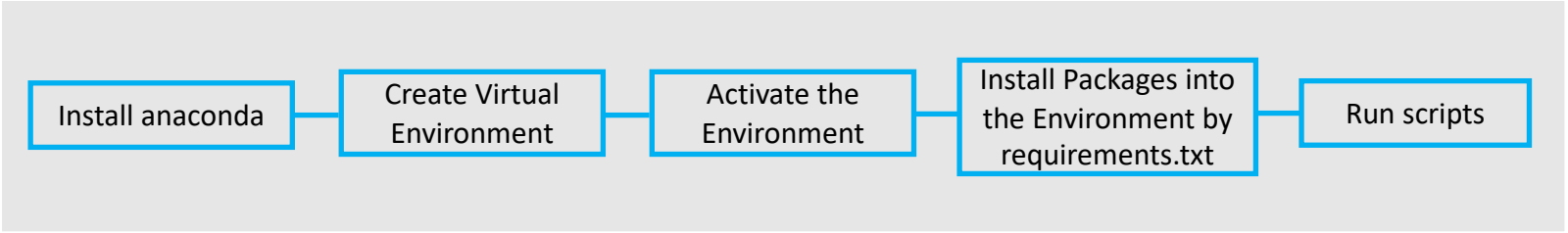
Instruction in each script







Run the demo in the shell:



More details of the script instruction, please check the following Github link:

https://github.com/wtcruijff/Market-Basket-Analysis-Big-Data-Course/tree/main/Market_Basket_Analysis_Master

Market basket analysis dataset

Cloud Computing (Amarel Cluster)

Through Phase 4 (late 2020):

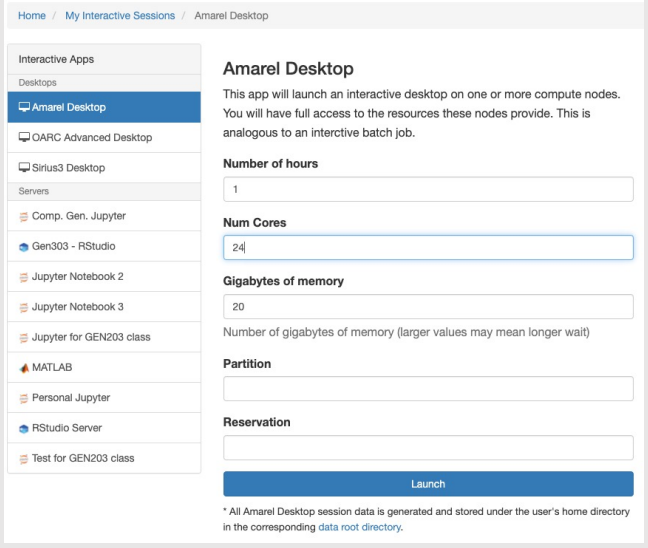
- 681 compute nodes
  - 19,972 Intel Xeon cores
  - 166 GPUs
  - Open OnDemand servers
  - InfiniBand FDR & EDR fabric

Amarel (owned by Rutgers) is a shared community-owned advanced computing environment available to any investigator or student with projects requiring research computing resources.

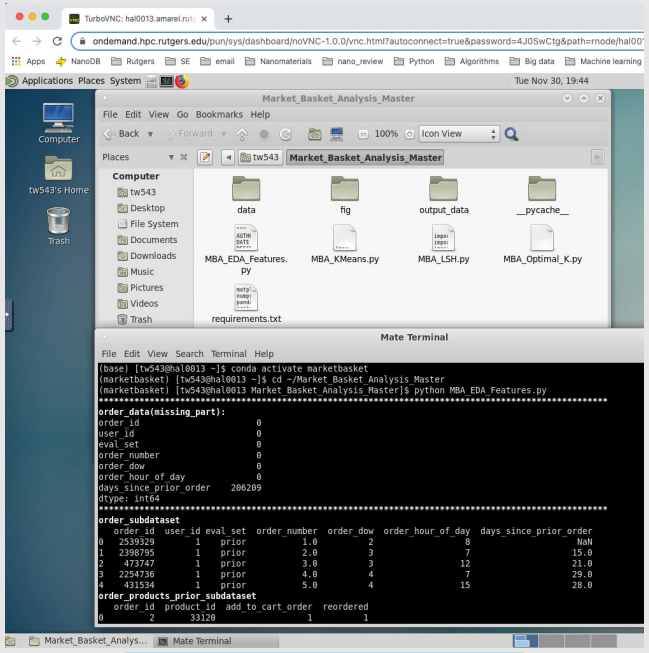
https://oarc.rutgers.edu/resources/amarel/

Cloud Computing workflow on Amarel:

1. Upload the data and script on the work directory
2. Install anaconda and setting up environment by the shell



3. Create a VM with certain parameters (cores and memory)



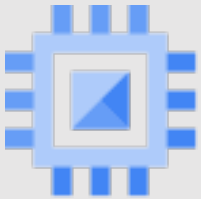
4. Run the script on Amarel

Market basket analysis dataset

Cloud Computing (Google Cloud Platform: GCP)

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Local User



GCP Compute Engine

GCP Virtual Machine by Remote desktop

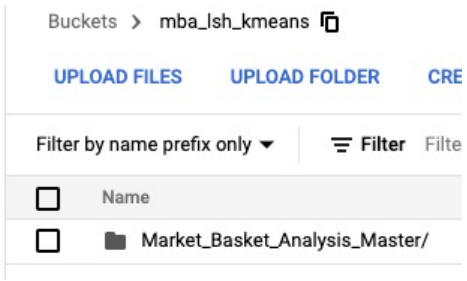
Create working environment by Anaconda

```
(MBA) snp116@demo:~/Market_Basket_Analysis_Master$ ls
MBA_EDA_Features.py  MBA_KMeans.py  MBA_LSH.py  MBA_Optimal_K.py  Miniconda3-latest-Linux-x86_64.sh  README.md  data  requirements.txt
(MBA) snp116@demo:~/Market_Basket_Analysis_Master$ python MBA_EDA_Features.py
**********************************************************************************************************
order_data(missing_part):
order_id                      0
user_id                       0
eval_set                      0
order_number                  0
order_dow                     0
order_hour_of_day             0
days_since_prior_order    206209
dtype: int64
**********************************************************************************************************
order_subdataset
    order_id  user_id eval_set  order_number  order_dow  order_hour_of_day  days_since_prior_order
0  2539329.0      1.0    prior           1.0        2.0                8.0                     NaN
1  2398795.0      1.0    prior           2.0        3.0                7.0                    15.0
2   473747.0      1.0    prior           3.0        3.0               12.0                    21.0
3  2254736.0      1.0    prior           4.0        4.0                7.0                    29.0
4   431534.0      1.0    prior           5.0        4.0               15.0                    28.0
order_products_prior_subdataset
   order_id  product_id  add_to_cart_order  reordered
0       2.0     33120.0                1.0        1.0
1       2.0     28985.0                2.0        1.0
2       2.0      9327.0                3.0        0.0
3       2.0     45918.0                4.0        1.0
4       2.0     30035.0                5.0        0.0
products_subdataset
                                          product_name  aisle_id  department_id
product_id
1.0                              Chocolate Sandwich Cookies      61.0           19.0
2.0                                       All-Seasons Salt     104.0           13.0
3.0                          Robust Golden Unsweetened Oolong Tea      94.0            7.0
4.0         Smart Ones Classic Favorites Mini Rigatoni Wit...      38.0            1.0
5.0                                    Green Chile Anytime Sauce       5.0           13.0
**********************************************************************************************************
1000_tfidf_features
             10  100   12   21   50   70   85   90   93   aa      acai  ...  yerba   yo  yobaby  yoghurt    yogurt  yokids  yukon  zbar  zbars      zero  zucchi
ni
user_id                                                                 ...
1.0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.000000  ...    0.0  0.0     0.0  0.000000  0.000000     0.0    0.0   0.0    0.0  0.183683       0
.0
2.0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.114357  ...    0.0  0.0     0.0  0.140415  0.123297     0.0    0.0   0.0    0.0  0.000000       0
.0
3.0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.000000  ...    0.0  0.0     0.0  0.000000  0.000000     0.0    0.0   0.0    0.0  0.000000       0
.0
4.0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.000000  ...    0.0  0.0     0.0  0.000000  0.000000     0.0    0.0   0.0    0.0  0.000000       0
.0
5.0      0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.000000  ...    0.0  0.0     0.0  0.000000  0.206131     0.0    0.0   0.0    0.0  0.000000       0
.0

[5 rows x 1000 columns]
**********************************************************************************************************
```

MBA_EDA_Features.py

Machine type:
n-standard-4
CPU: 4vCPUs
Memory: 15GB



Live coding instruction, please check the following YouTube links:

1. https://www.youtube.com/watch?v=CM-wXdwAI5E
2. https://www.youtube.com/watch?v=ijxXGw66Urc
3. https://www.youtube.com/watch?v=r1n6LMsV_RY&t=240

Acknowledgement: Many Thanks to Dr. Shende for the kind guidance!

Group Member:

Iswarya Madhu Desikan

Shreya Patel

Tong Wang

Arkit Pawar