E-Commerce Customer Segmentation

Applying K-Means Clustering Based on Aisle-Level Transactions

Presented by Yanan Xie on Sun 6 Sep, 2020

The Problem Scope

- **Customer segmentation based on**
  - Customer information
  - **Customer behaviour**
    - Datetime of transactions
    - **Items in transactions**

      ➡ Product / **Aisle** / Department
- **Algorithms**
  - Supervised learning
  - **Unsupervised learning**

Table Of Contents

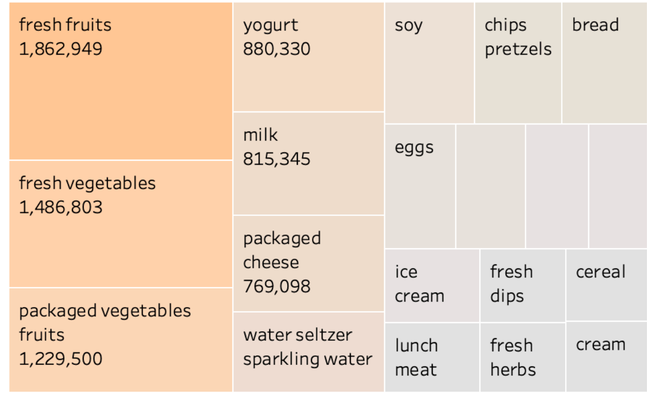
1. Data Exploration

departments	21
aisles	134
products	49,688
users	206,209
orders	3,421,080

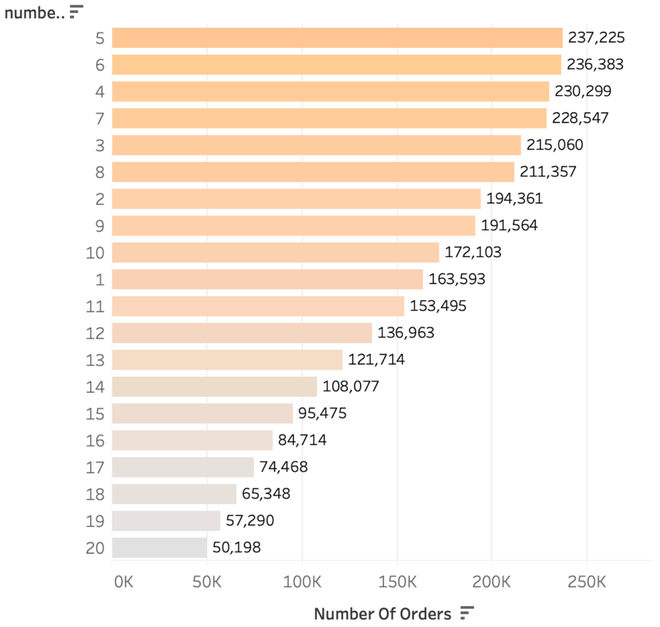
Customers' favourite departments <Top6>

department	department_id	
produce	4	9,888,378
dairy eggs	16	5,631,067
snacks	19	3,006,412
beverages	7	2,804,175
frozen	1	2,336,858
pantry	13	1,956,819

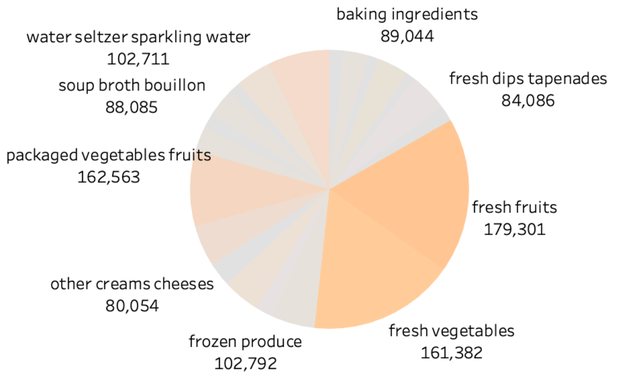
Customers' favourite aisles <Top20>



Number of products people usually order <Top 20>



Number of customers in each aisle



2. Feature Engineering

Data Transformation

How to interpret the difference between customers who buy an aisle 1000 times and 100 times?

- Non transformation
- Log10 transformation
- One-Hot encoding

aisle	user_id	air fresheners candles	asian foods	baby accessories	baby bath body care	baby food formula	bakery desserts	baking ingredients	baking supplies decor	beauty	...	spreads	tea	tofu meat alternatives	tortillas flat bread	trail mix snack mix	trash bags liners
0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0
1	2	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	1.0	1.0	1.0	0.0	0.0	0.0
2	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	1.0	0.0	0.0	0.0	0.0
3	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	1.0	0.0	0.0
4	5	1.0	1.0	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
206204	206205	0.0	0.0	1.0	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	206206	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	...	1.0	0.0	0.0	0.0	0.0	1.0
206206	206207	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	1.0	1.0	0.0	1.0	1.0	0.0
206207	206208	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	1.0	0.0	0.0
206208	206209	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0

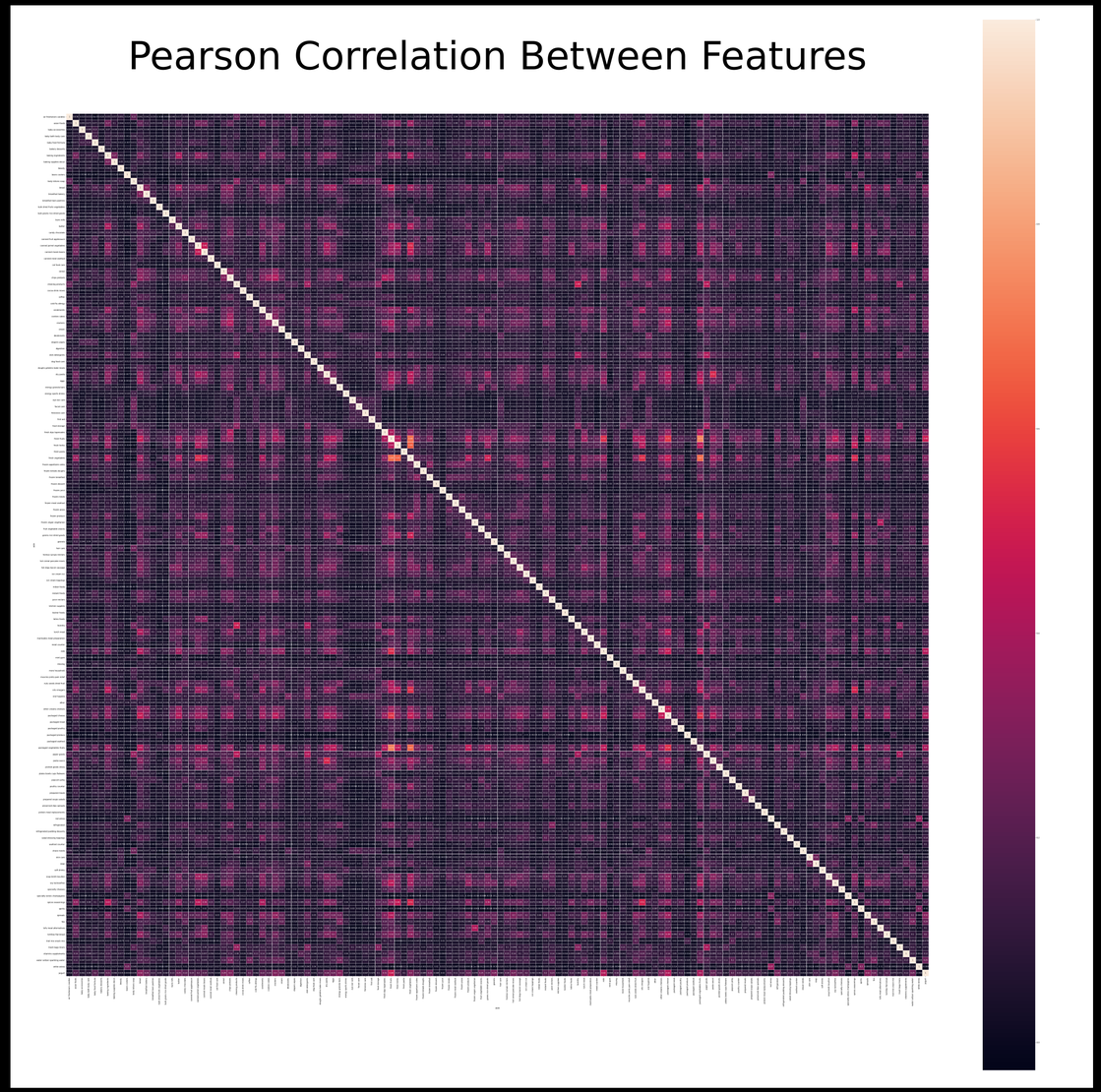
206209 rows × 135 columns

2. Feature Engineering

Data Transformation

Dimension Reduction

- Pearson Correlation Between Features
- PCA - **75** out of 134 principle components covering **72.6%** information (non-transformation example)



This Pearson Correlation is created from data without transformation.

3. Building K-Means

from sklearn.cluster import KMeans

- algorithm{“auto”, “full”, “elkan”}, default=“auto”

optimal of k ?

4. Model Evaluation

A Quantitative Perspective

- SSE
- Elbow Criterion - the "elbow" on the arm is the value of optimal **k**

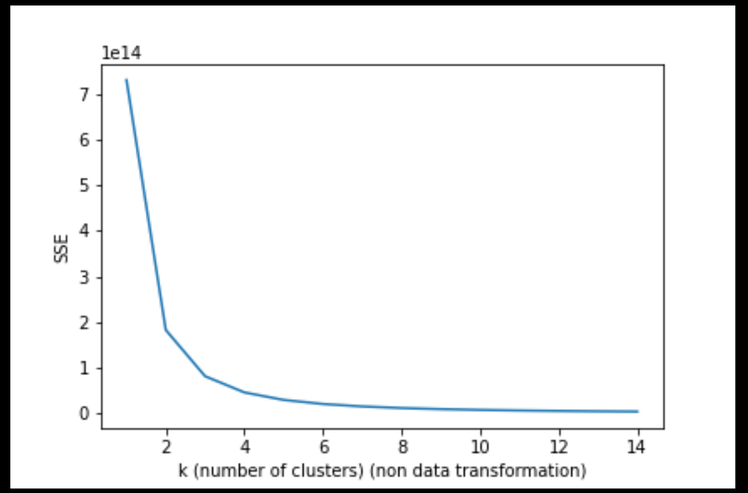
A Business Perspective

- Common favourite aisles within a cluster?
- Different favourite aisles among clusters?

4. Model Evaluation

Non Data Transformation

- 4 clusters by elbow criterion
- Calculating and sorting absolute sum values of each aisle
- Same top 13 favourites
- Minor differences after top 13
- Consistency with overall favourite aisles in whole dataset



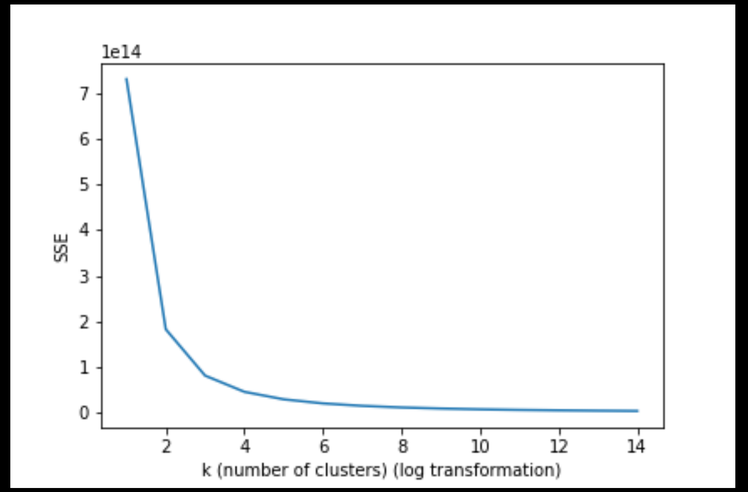
	cluster0	cluster1	cluster2	cluster3
0	fresh fruits	fresh fruits	fresh fruits	fresh fruits
1	fresh vegetables	fresh vegetables	fresh vegetables	fresh vegetables
2	packaged vegetables fruits	packaged vegetables fruits	packaged vegetables fruits	packaged vegetables fruits
3	yogurt	yogurt	yogurt	yogurt
4	packaged cheese	packaged cheese	packaged cheese	packaged cheese
5	milk	milk	milk	milk
6	water seltzer sparkling water	water seltzer sparkling water	water seltzer sparkling water	water seltzer sparkling water
7	chips pretzels	chips pretzels	chips pretzels	chips pretzels
8	soy lactosefree	soy lactosefree	soy lactosefree	soy lactosefree
9	bread	bread	bread	bread
10	refrigerated	refrigerated	refrigerated	refrigerated
11	frozen produce	frozen produce	frozen produce	frozen produce
12	ice cream ice	ice cream ice	ice cream ice	ice cream ice
13	crackers	energy granola bars	energy granola bars	crackers
14	eggs	eggs	crackers	eggs
15	energy granola bars	crackers	eggs	energy granola bars
16	lunch meat	lunch meat	lunch meat	frozen meals
17	frozen meals	frozen meals	frozen meals	lunch meat
18	baby food formula	cereal	cereal	baby food formula
19	fresh herbs	fresh herbs	baby food formula	fresh herbs

4. Model Evaluation

Non Data Transformation

Log10 Transformation

- 4 clusters by elbow criterion
- Calculating and sorting log10 sum values of each aisle
- Only same top 3 favourites
- Distinct differences since top7
- Optimised compared to non data transformation



	cluster0	cluster1	cluster2	cluster3
0	fresh fruits	fresh fruits	fresh fruits	fresh fruits
1	fresh vegetables	fresh vegetables	fresh vegetables	fresh vegetables
2	packaged vegetables fruits	packaged vegetables fruits	packaged vegetables fruits	packaged vegetables fruits
3	yogurt	yogurt	yogurt	paper goods
4	water seltzer sparkling water	packaged cheese	packaged cheese	packaged cheese
5	packaged cheese	milk	milk	chips pretzels
6	milk	chips pretzels	chips pretzels	yogurt
7	chips pretzels	bread	soy lactosefree	water seltzer sparkling water
8	soy lactosefree	frozen produce	water seltzer sparkling water	milk
9	ice cream ice	soy lactosefree	bread	soft drinks
10	refrigerated	crackers	frozen produce	bread
11	bread	water seltzer sparkling water	eggs	ice cream ice
12	soft drinks	eggs	refrigerated	crackers
13	packaged produce	fresh herbs	ice cream ice	cleaning products
14	frozen produce	lunch meat	crackers	cereal
15	eggs	refrigerated	fresh herbs	refrigerated
16	crackers	soup broth bouillon	lunch meat	eggs
17	frozen meals	ice cream ice	fresh dips tapenades	juice nectars
18	energy granola bars	baking ingredients	soup broth bouillon	laundry
19	cereal	canned jarred vegetables	energy granola bars	baking ingredients

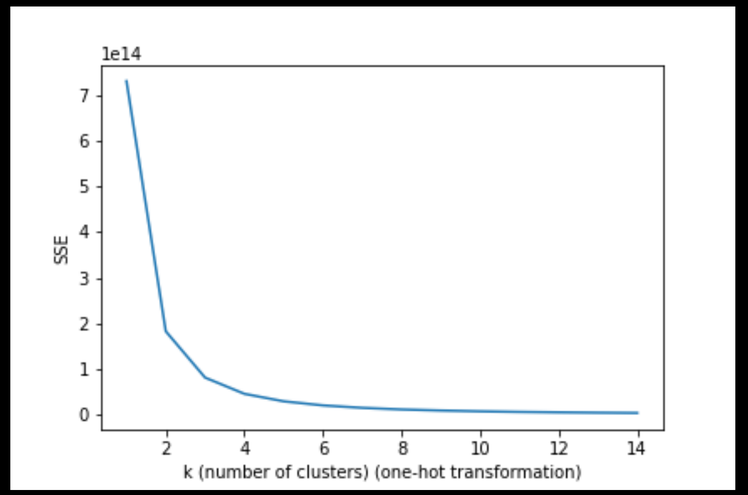
4. Model Evaluation

Non Data Transformation

Log10 Transformation

<u>One-Hot Encoding</u>

- 4 clusters by elbow criterion
- Calculating and sorting sum of one-hot of each aisle
- Difference since top2
- Rare aisles showing in top (e.g. condiments)
- Totally not considering quantities



	cluster0	cluster1	cluster2	cluster3
0	fresh fruits	fresh fruits	fresh fruits	fresh fruits
1	fresh vegetables	fresh vegetables	packaged vegetables fruits	packaged vegetables fruits
2	packaged vegetables fruits	packaged vegetables fruits	fresh vegetables	packaged cheese
3	packaged cheese	packaged cheese	water seltzer sparkling water	fresh vegetables
4	yogurt	yogurt	chips pretzels	paper goods
5	milk	frozen produce	packaged cheese	chips pretzels
6	bread	chips pretzels	milk	bread
7	chips pretzels	bread	yogurt	yogurt
8	frozen produce	milk	packaged produce	crackers
9	soy lactosefree	soup broth bouillon	soft drinks	ice cream ice
10	eggs	baking ingredients	soy lactosefree	milk
11	soup broth bouillon	oils vinegars	bread	cleaning products
12	water seltzer sparkling water	soy lactosefree	refrigerated	water seltzer sparkling water
13	fresh herbs	condiments	crackers	soft drinks
14	ice cream ice	crackers	ice cream ice	cereal
15	crackers	fresh herbs	frozen produce	condiments
16	fresh dips tapenades	canned jarred vegetables	juice nectars	baking ingredients
17	oils vinegars	dry pasta	nuts seeds dried fruit	eggs
18	refrigerated	spices seasonings	cereal	other creams cheeses
19	baking ingredients	spreads	oils vinegars	refrigerated

5. Conclusion

k=4 clusters based on elbow criterion of SSE

Trade-off of data transformation strategies

	Non transformation	Log10 transformation	One-hot encoding
Inter-cluster difference	Low	Medium	High
Inner-cluster individuality	Low	Medium	High
Impact of quantity	High	Medium	Low

Future work: Dropping columns of frequent aisles? Potential risks?

THE END