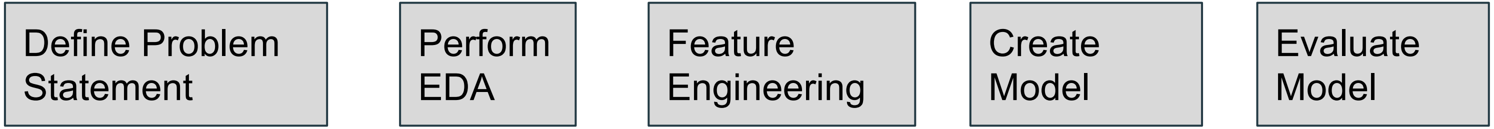
Term 4 - ML2

Rupinder Kohli
March 2020 cohort

An attempt has been made to find the clusters of items in the customer shopping basket

https://raw.githubusercontent.com/rupkohli/DS_ML_Research/main/clustering/groceries.csv



Define Problem Statement

Perform EDA

Feature Engineering

Create Model

Evaluate Model

Define Problem & Approach to solve

Problem Statement:

As the owner of the store, I need to understand the products the customers are buying

- Evaluate the clusters of shopping baskets.

Approach:

To understand consumer behavior and develop customer attributes or archetypes, we will need to use the clustering technique.

Since there is no target variable when evaluating the customer baskets, we will need to apply an unsupervised algorithm to find how data is logically grouped together.

The following algorithms can be applied to learn more about customer's baskets or shopping list -

- K-Means (K-Mode) clustering

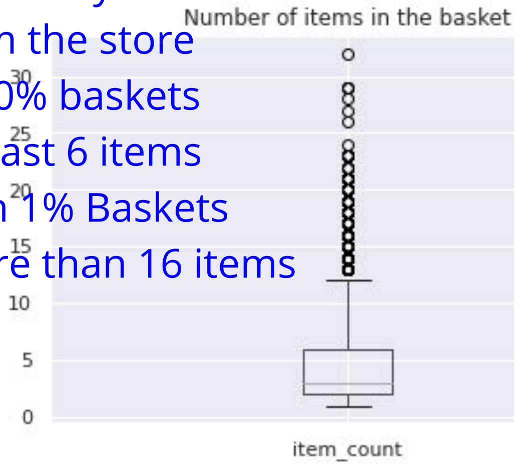
- Naive Bayes algorithm

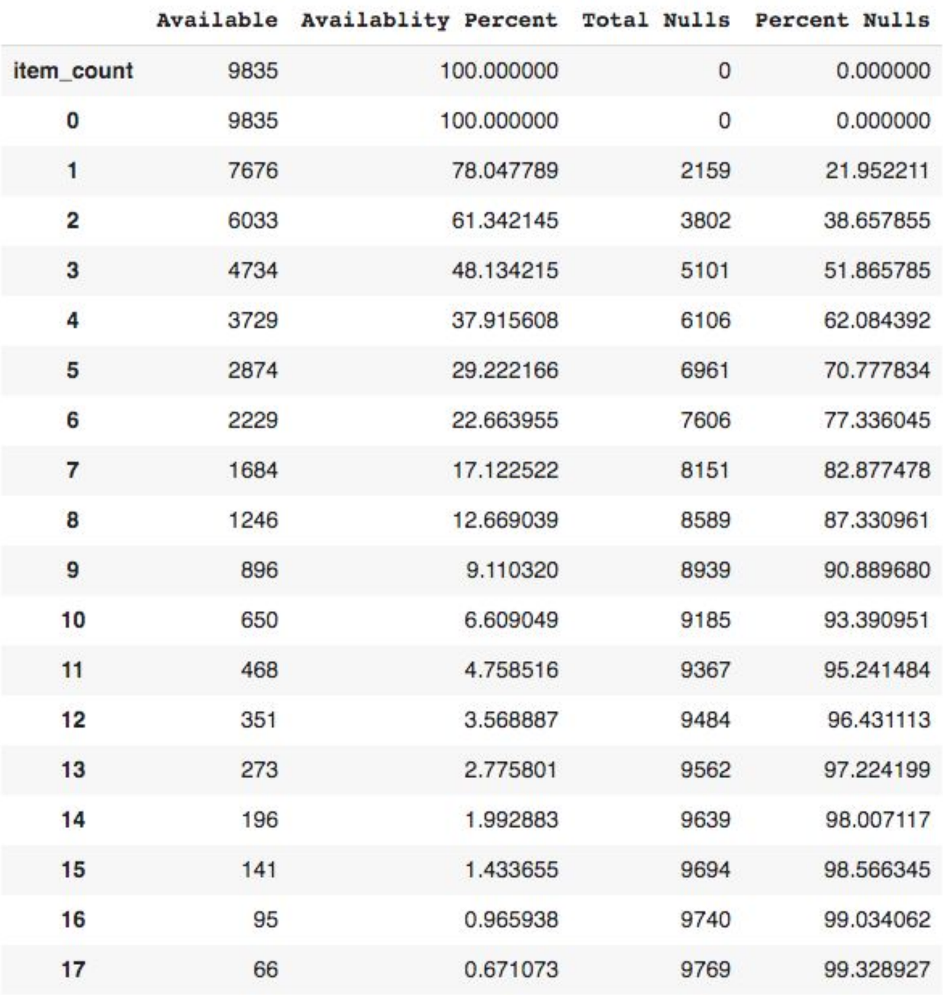
EDA

understanding the dataset

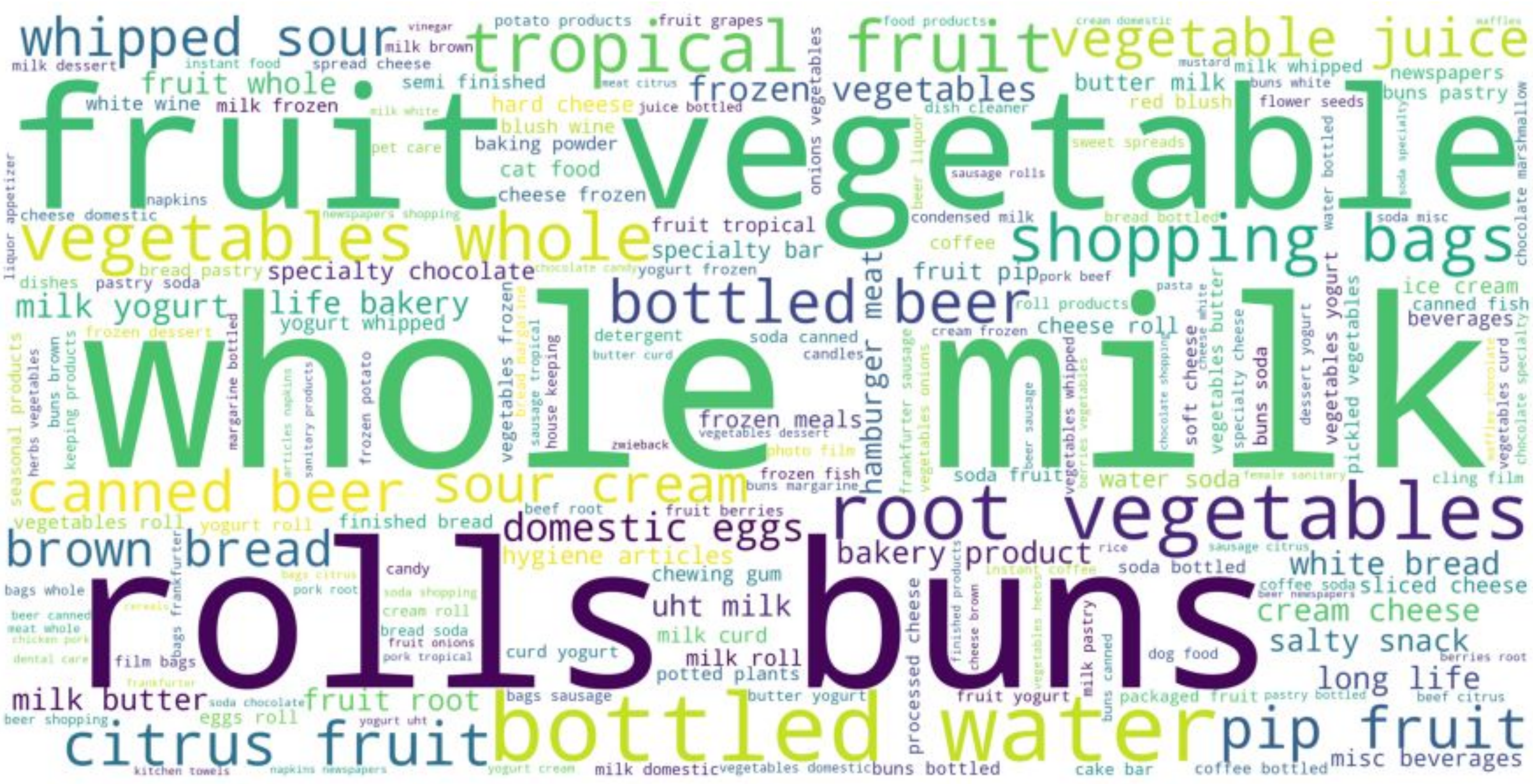
About Dataset

- Categorical dataset
- 9835 baskets
- 169 unique items
- Range of 1-32 items in baskets
- 463 repeated baskets
- Customers buy atleast 1 item from the store
- Atleast 30% baskets have atleast 6 items
- Less than 1% Baskets have more than 16 items



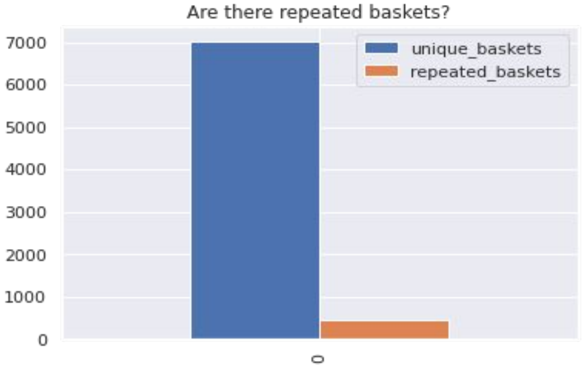


	Available	Availablity Percent	Total Nulls	Percent Nulls
item_count	9835	100.000000	0	0.000000
0	9835	100.000000	0	0.000000
1	7676	78.047789	2159	21.952211
2	6033	61.342145	3802	38.657855
3	4734	48.134215	5101	51.865785
4	3729	37.915608	6106	62.084392
5	2874	29.222166	6961	70.777834
6	2229	22.663955	7606	77.336045
7	1684	17.122522	8151	82.877478
8	1246	12.669039	8589	87.330961
9	896	9.110320	8939	90.889680
10	650	6.609049	9185	93.390951
11	468	4.758516	9367	95.241484
12	351	3.568887	9484	96.431113
13	273	2.775801	9562	97.224199
14	196	1.992883	9639	98.007117
15	141	1.433655	9694	98.566345
16	95	0.965938	9740	99.034062
17	66	0.671073	9769	99.328927

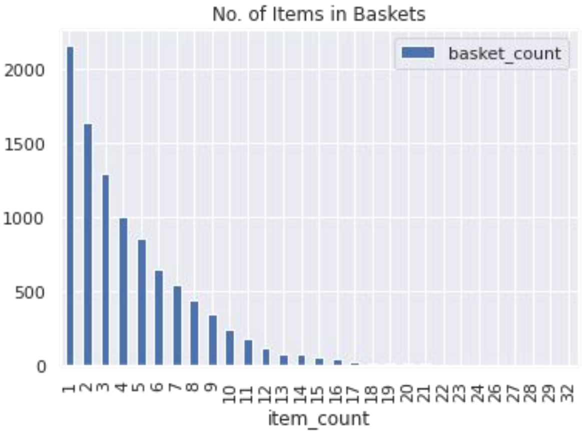


Fast going Items

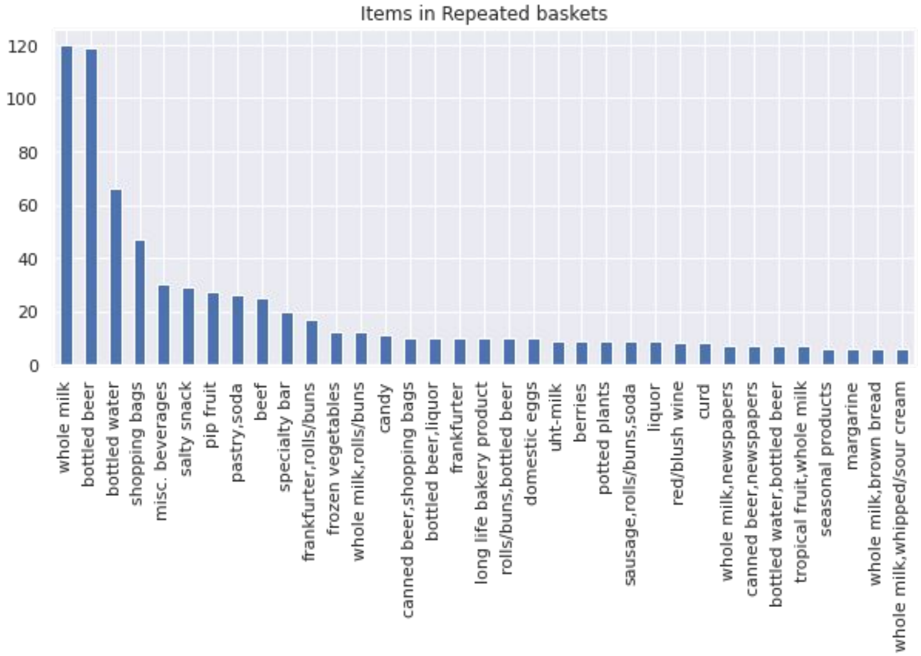
How does the customer baskets look like



Are there repeated baskets?

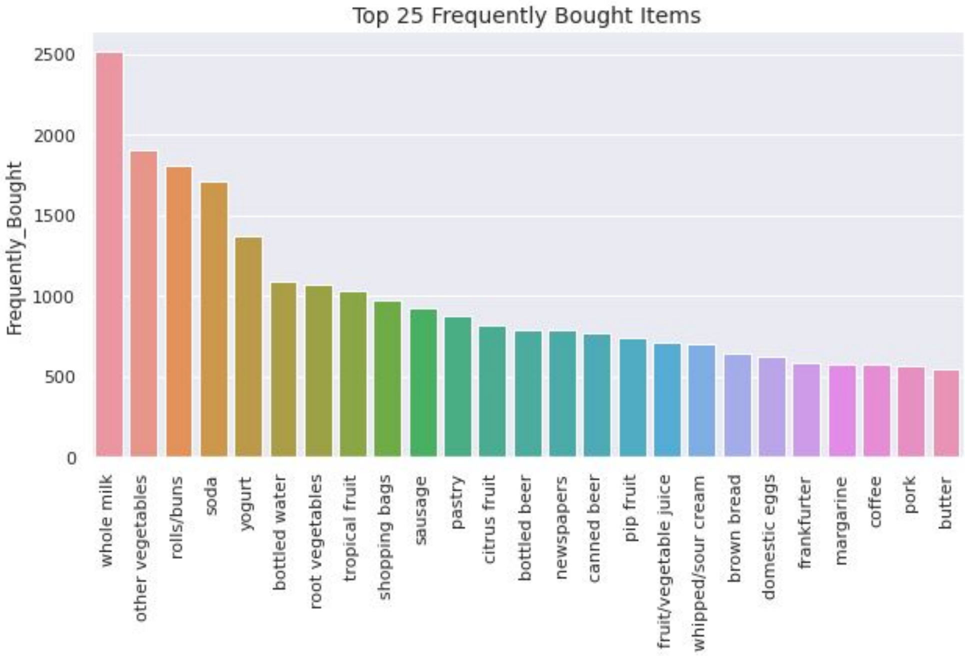


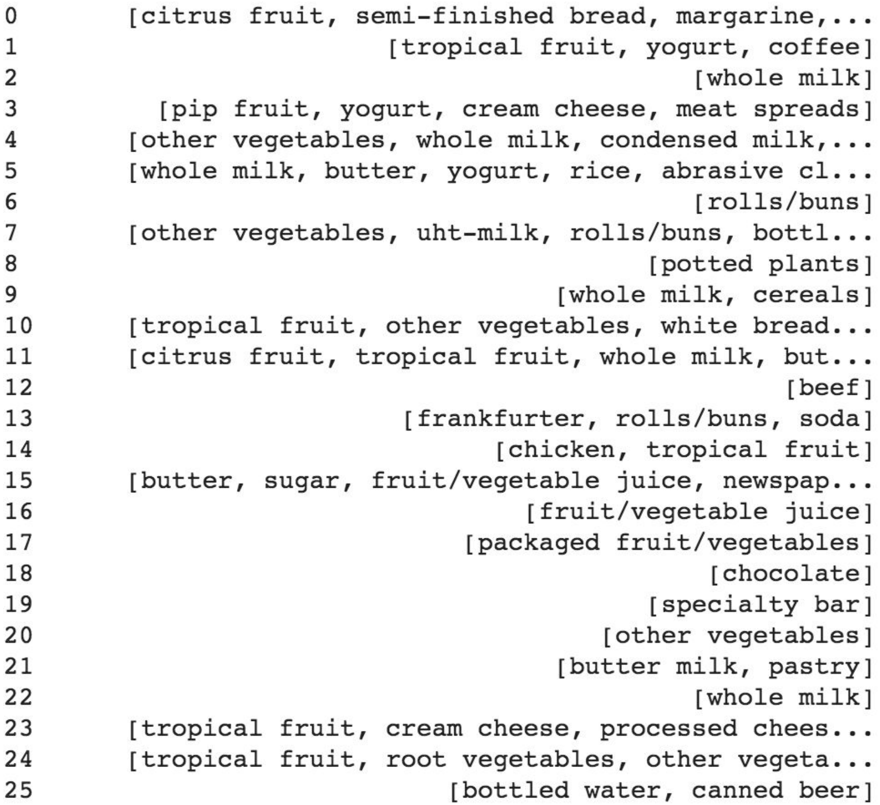
No. of Items in Baskets



Items in Repeated baskets

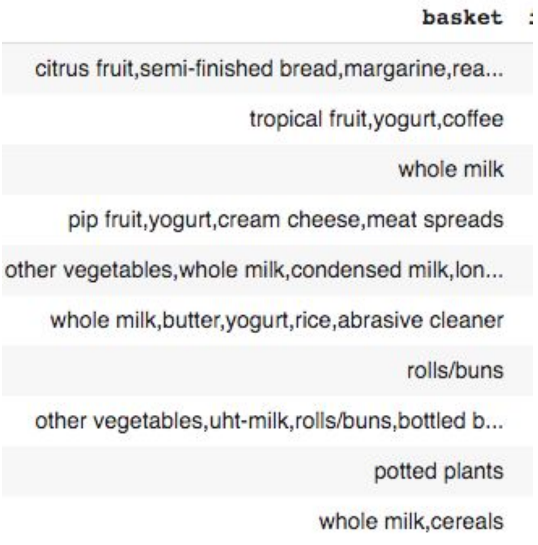
While building the cluster model, we should be analysing the baskets with atleast 9 items; we can confirm once we do a PCA (principal component analysis)



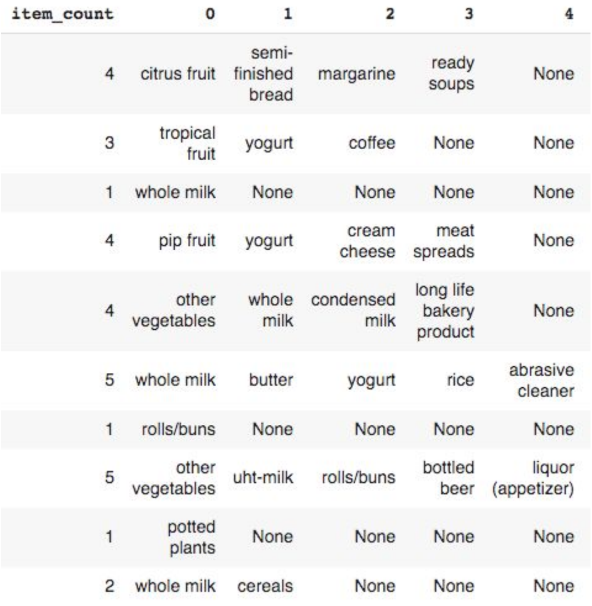


```
0         [citrus fruit, semi-finished bread, margarine,...
1                      [tropical fruit, yogurt, coffee]
2                                          [whole milk]
3          [pip fruit, yogurt, cream cheese, meat spreads]
4         [other vegetables, whole milk, condensed milk,...
5         [whole milk, butter, yogurt, rice, abrasive cl...
6                                          [rolls/buns]
7         [other vegetables, uht-milk, rolls/buns, bottl...
8                                       [potted plants]
9                                  [whole milk, cereals]
10        [tropical fruit, other vegetables, white bread...
11        [citrus fruit, tropical fruit, whole milk, but...
12                                               [beef]
13                      [frankfurter, rolls/buns, soda]
14                             [chicken, tropical fruit]
15        [butter, sugar, fruit/vegetable juice, newspap...
16                              [fruit/vegetable juice]
17                            [packaged fruit/vegetables]
18                                          [chocolate]
19                                       [specialty bar]
20                                    [other vegetables]
21                                 [butter milk, pastry]
22                                         [whole milk]
23        [tropical fruit, cream cheese, processed chees...
24        [tropical fruit, root vegetables, other vegeta...
25                          [bottled water, canned beer]
```

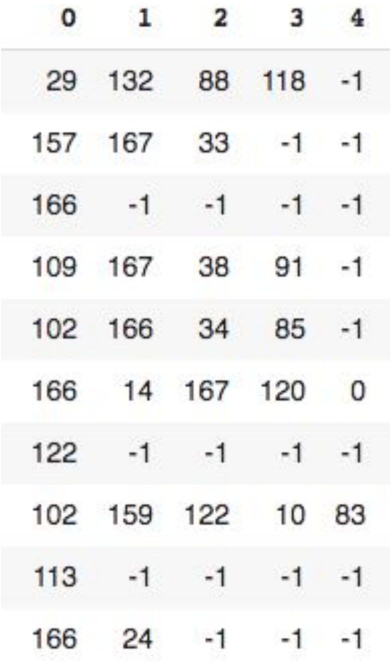
What does customer buy



basket
citrus fruit,semi-finished bread,margarine,rea...
tropical fruit,yogurt,coffee
whole milk
pip fruit,yogurt,cream cheese,meat spreads
other vegetables,whole milk,condensed milk,lon...
whole milk,butter,yogurt,rice,abrasive cleaner
rolls/buns
other vegetables,uht-milk,rolls/buns,bottled b...
potted plants
whole milk,cereals



item_count	0	1	2	3	4
4	citrus fruit	semi-finished bread	margarine	ready soups	None
3	tropical fruit	yogurt	coffee	None	None
1	whole milk	None	None	None	None
4	pip fruit	yogurt	cream cheese	meat spreads	None
4	other vegetables	whole milk	condensed milk	long life bakery product	None
5	whole milk	butter	yogurt	rice	abrasive cleaner
1	rolls/buns	None	None	None	None
5	other vegetables	uht-milk	rolls/buns	bottled beer	liquor (appetizer)
1	potted plants	None	None	None	None
2	whole milk	cereals	None	None	None



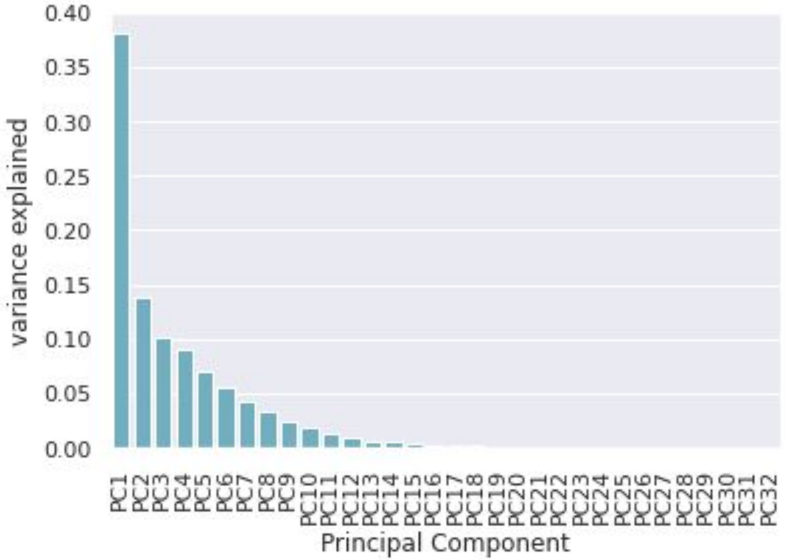
0	1	2	3	4
29	132	88	118	-1
157	167	33	-1	-1
166	-1	-1	-1	-1
109	167	38	91	-1
102	166	34	85	-1
166	14	167	120	0
122	-1	-1	-1	-1
102	159	122	10	83
113	-1	-1	-1	-1
166	24	-1	-1	-1

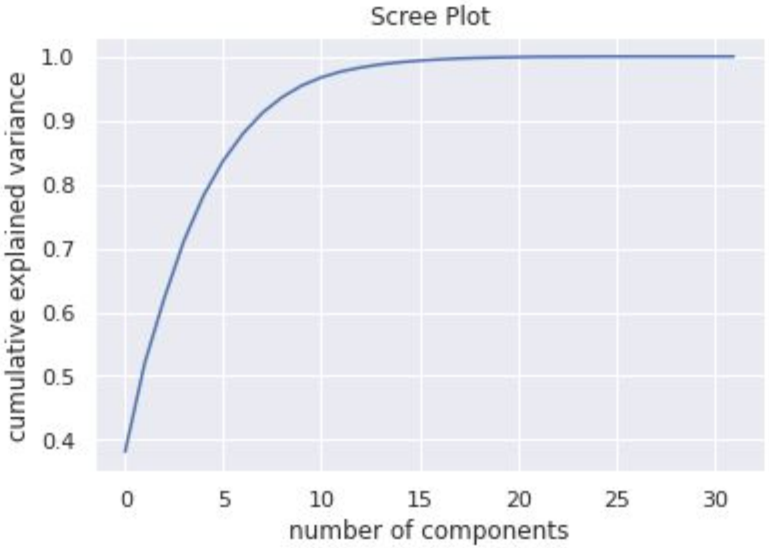
RAW ***ITEMISED*** ***ENCODED***

Data Setup - Encode the data

Principal Component Analysis

To confirm the number of items to be considered in clustering





Indicates the consideration should be given to baskets with atleast 9 items

PCA on the encoded dataset

K-Means Clustering

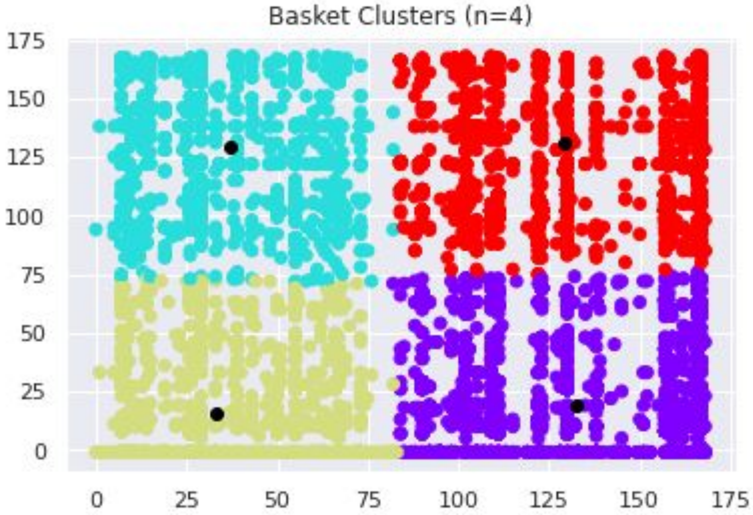
To cluster the items / baskets

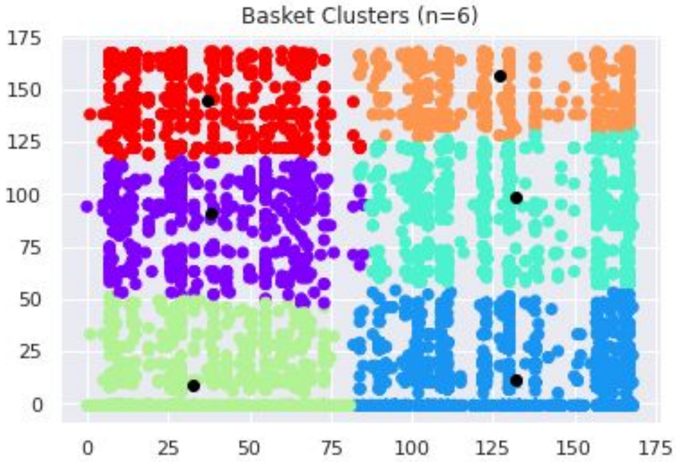
K-Means Clustering

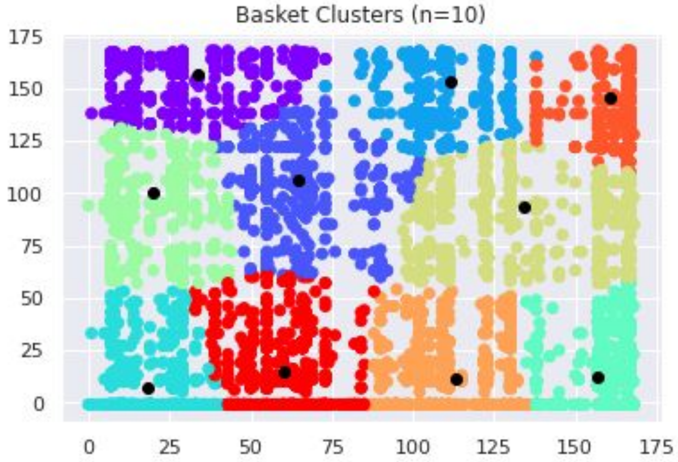
- Unsupervised learning algorithm

- Forms clusters of data based on similar instances

- Does not respect null values





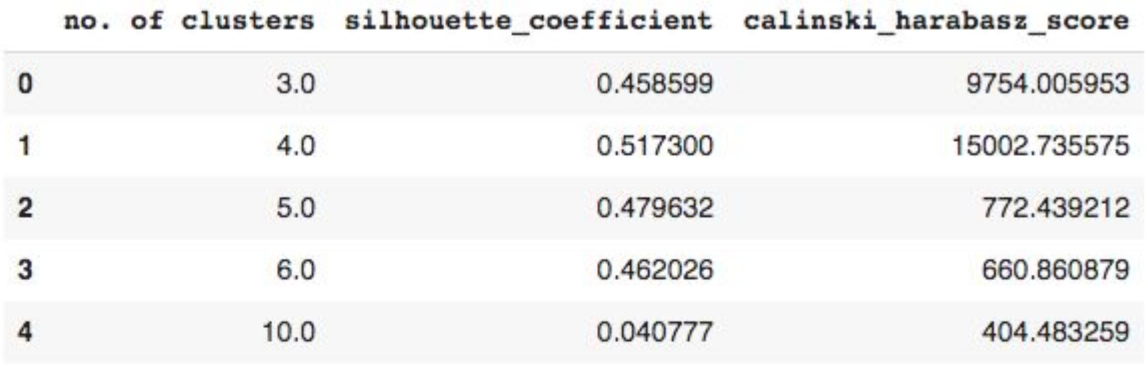


Assumptions -

- **No null values in dataset**
- **No categorical data**
- **For visualisations - considered baskets with 2 items**

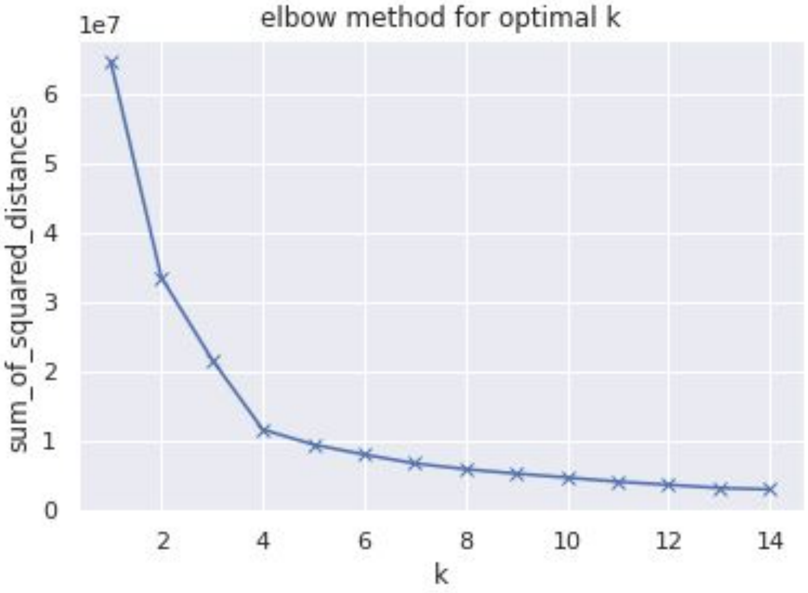
K-Means on the encoded dataset

K-Means - Evaluation



	no. of clusters	silhouette_coefficient	calinski_harabasz_score
0	3.0	0.458599	9754.005953
1	4.0	0.517300	15002.735575
2	5.0	0.479632	772.439212
3	6.0	0.462026	660.860879
4	10.0	0.040777	404.483259

Looking at the Cluster Scores both the silhouette coefficient and cv score are highest with 4 clusters, we can conclude that the optimum number of clusters should be 4 i.e. ***the optimal number of items in the basket are 4***

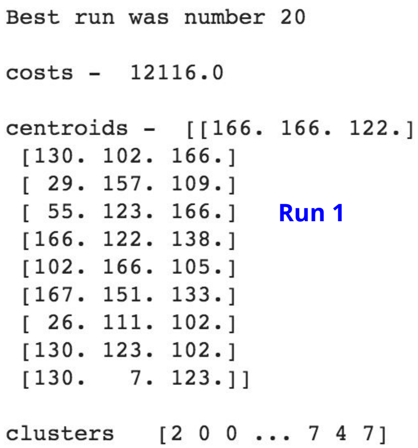


K-Mode Clustering

To view the clustered items

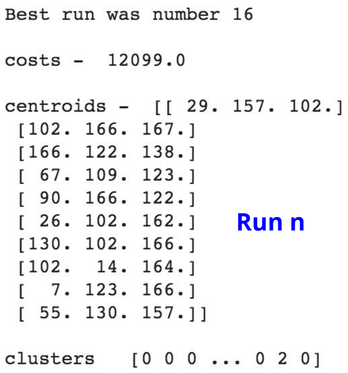
K-Mode Clustering

- Unsupervised learning algorithm
- Forms clusters of data based on similar instances
- Applicable only if there is categorical data
- The efficiency of the algorithm is based on the distance between 2 rows
- The algorithm was executed multiple times to reach an optimised cost
- First pass of the algorithm was run with 3 items per basket which was later extended to 10 items in basket



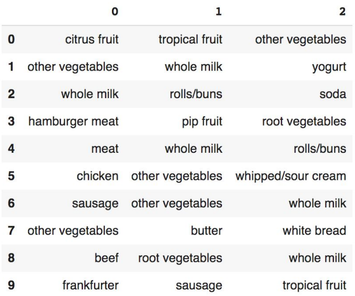
```
Best run was number 20

costs -  12116.0

centroids -  [[166. 166. 122.]
 [130. 102. 166.]
 [ 29. 157. 109.]
 [ 55. 123. 166.]
 [166. 122. 138.]
 [102. 166. 105.]
 [167. 151. 133.]
 [ 26. 111. 102.]
 [130. 123. 102.]
 [130.   7. 123.]]

clusters   [2 0 0 ... 7 4 7]
```

Run 1



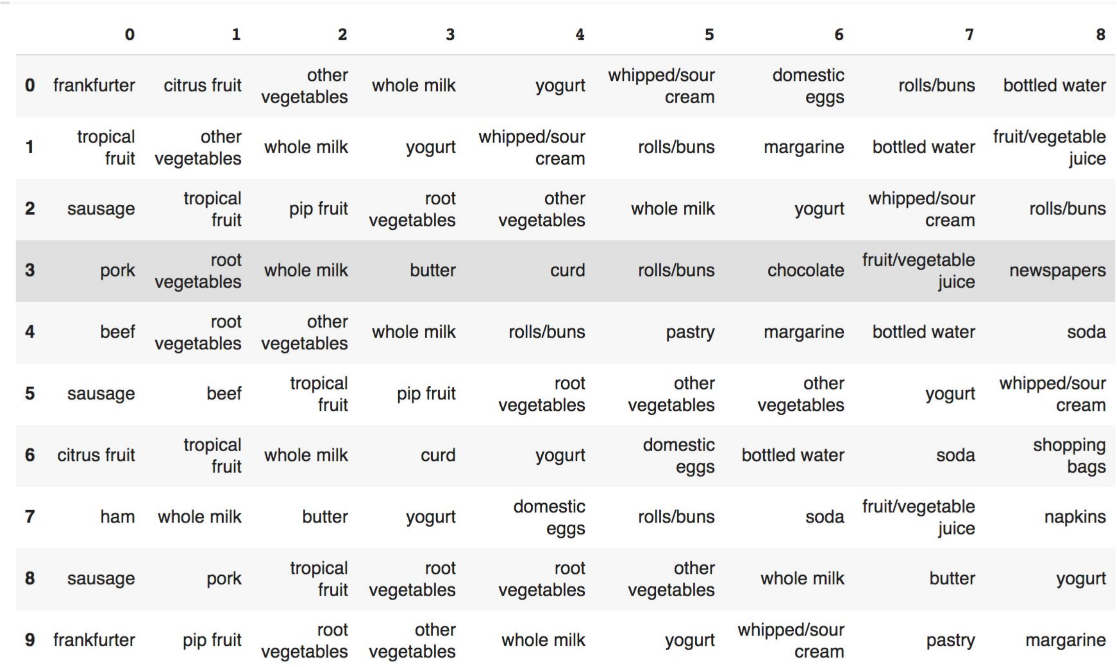
```
Best run was number 16

costs -  12099.0

centroids -  [[ 29. 157. 102.]
 [102. 166. 167.]
 [166. 122. 138.]
 [ 67. 109. 123.]
 [ 90. 166. 122.]
 [ 26. 102. 162.]
 [130. 102. 166.]
 [102.  14. 164.]
 [  7. 123. 166.]
 [ 55. 130. 157.]]

clusters   [0 0 0 ... 0 2 0]
```

Run n



	0	1	2
0	citrus fruit	tropical fruit	other vegetables
1	other vegetables	whole milk	yogurt
2	whole milk	rolls/buns	soda
3	hamburger meat	pip fruit	root vegetables
4	meat	whole milk	rolls/buns
5	chicken	other vegetables	whipped/sour cream
6	sausage	other vegetables	whole milk
7	other vegetables	butter	white bread
8	beef	root vegetables	whole milk
9	frankfurter	sausage	tropical fruit

Item Clusters



	0	1	2	3	4	5	6	7	8
0	frankfurter	citrus fruit	other vegetables	whole milk	yogurt	whipped/sour cream	domestic eggs	rolls/buns	bottled water
1	tropical fruit	other vegetables	whole milk	yogurt	whipped/sour cream	rolls/buns	margarine	bottled water	fruit/vegetable juice
2	sausage	tropical fruit	pip fruit	root vegetables	other vegetables	whole milk	yogurt	whipped/sour cream	rolls/buns
3	pork	root vegetables	whole milk	butter	curd	rolls/buns	chocolate	fruit/vegetable juice	newspapers
4	beef	root vegetables	other vegetables	whole milk	rolls/buns	pastry	margarine	bottled water	soda
5	sausage	beef	tropical fruit	pip fruit	root vegetables	other vegetables	other vegetables	yogurt	whipped/sour cream
6	citrus fruit	tropical fruit	whole milk	curd	yogurt	domestic eggs	bottled water	soda	shopping bags
7	ham	whole milk	butter	yogurt	domestic eggs	rolls/buns	soda	fruit/vegetable juice	napkins
8	sausage	pork	tropical fruit	root vegetables	root vegetables	other vegetables	whole milk	butter	yogurt
9	frankfurter	pip fruit	root vegetables	other vegetables	whole milk	yogurt	whipped/sour cream	pastry	margarine

- **Analysing 1246 baskets will 9 items each**
- **Outcome is a cluster of 10 baskets**

K-Mode clusters on dataset

Evaluation

Based on the generated Basket Clusters, the following products are fast flowing -

- **WHOLE MILK & ROLLS/BUNS & SODA** combinations were found in 87 baskets
- **WHOLE MILK & YOGURT** was part of 551 baskets
- **WHOLE MILK** was part of 2513 baskets
- **YOGURT** was part of 1372 baskets
- **WHOLE MILK & PASTRY** was part of 1372 baskets
- **FRANKFURTER** was part of 580 baskets

Conclusions

The Problem statement as identified for the project was **"An attempt has been made to find the clusters of items in the customer shopping basket"**

The Approach was to **"Identify the cluster of baskets with similar items"**

The Conclusion Based on the application of the K-Means and K-Mode algorithm

- **The baskets with 3, 4 and 9 items were clustered together successfully with the best possible cost**
- **Cross evaluated manually to understand if the clusters were correctly formed**
- **Had planned to apply the Naive Bayes algorithm to predict if buying "whole milk" what is the probability the customer will buy "yogurt" as well, due to the paucity of time it wasn't possible.**

THANK YOU!