Capstone Project

Customer Profiling & Market Basket Analysis

By,

Bhavya Tadikonda (71310099)

Haneesh Reddy (71310069)

Raj Kiran Swain (71310024)

Sweekar Tanugula (71310077)

Problem Statement

A leading retailer chain wants to expand their chain of stores. To do this in the current competitive retail scenario, it is very crucial to know and understand the customers in order to create the right offers and schemes to retain and grow the existing customer base, reduce customer attrition and increase influx of new customers.

Given the rich transactional data available of over one million transactions, we want to mine this data to **analyze and understand the profile of their current registered customers** based on their transaction patterns, shopping behaviour and association with the retail chain. Also **analyze the frequently bought items.**

Solution Proposed

Solution Approach and Rationale:

At the core of the solution is to profile the retail shoppers and define their shopping missions which will then be used by the retailers to target the right set of customers for the right programs. For example, loyalty programs may target customers who shop more frequently at the store in order to entice them to make more purchases. We also wish to provide Market-Basket analysis for cross-sell opportunity.

For Customer segmentation we propose to use Clustering methodology, and **for Market-Basket Analysis we will use Association Rules.**

Data Collection:

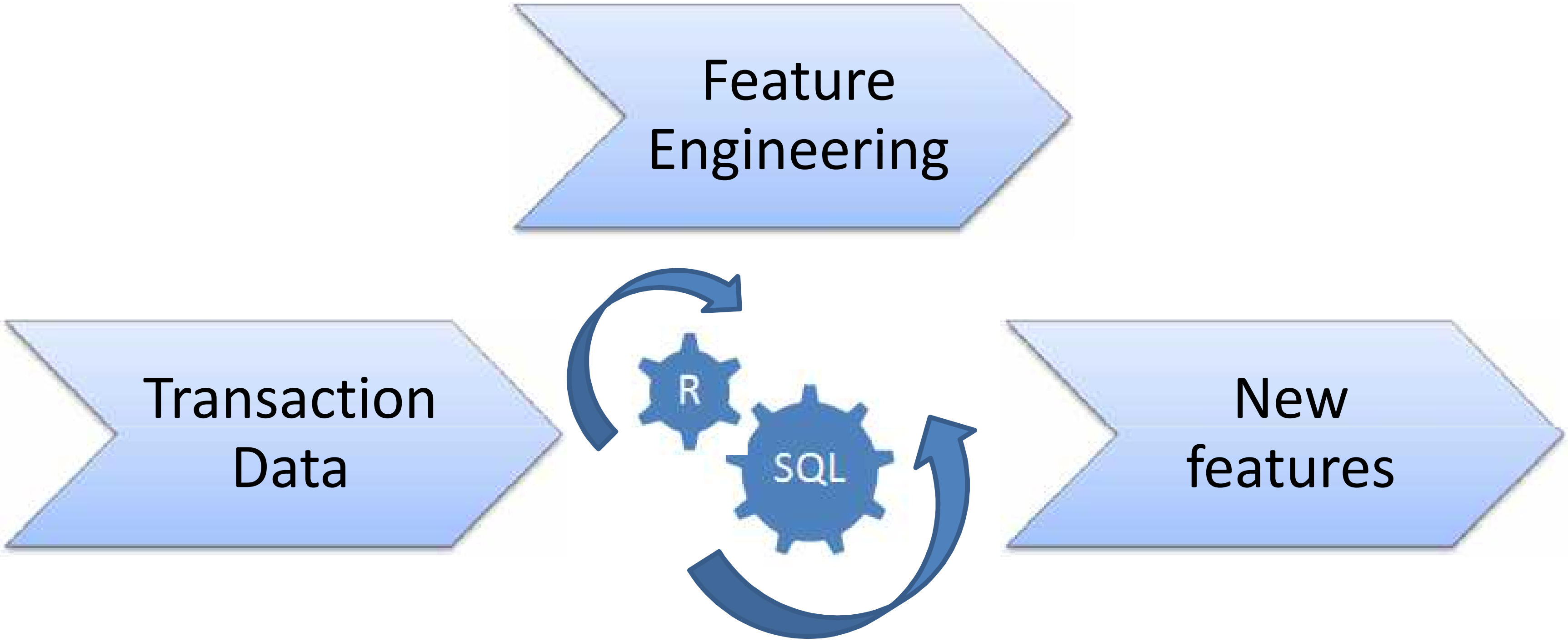
Data was provided by Marketelligent. 6 months of sample customer transaction data of registered customers of one store of the retailer was provided. Additionally provided us the product hierarchy used in the store. Data was shared to us in Access DB, we have imported the database through ftp and used for our analysis.

Software Used:

R and MS Access were used for our analysis. We leveraged SQL to connect between R and MS Access.

Customer Profiling

Feature Engineering



Feature Engineering

Transaction Data

New features

- Only transactional data was not sufficient for finding segments, so we have used R and SQL to create new features from the available data. We stored the new data in MS Access for further use.

- We have created 32 new features as part of Feature Engineering, the list of new features follows in the next slide.

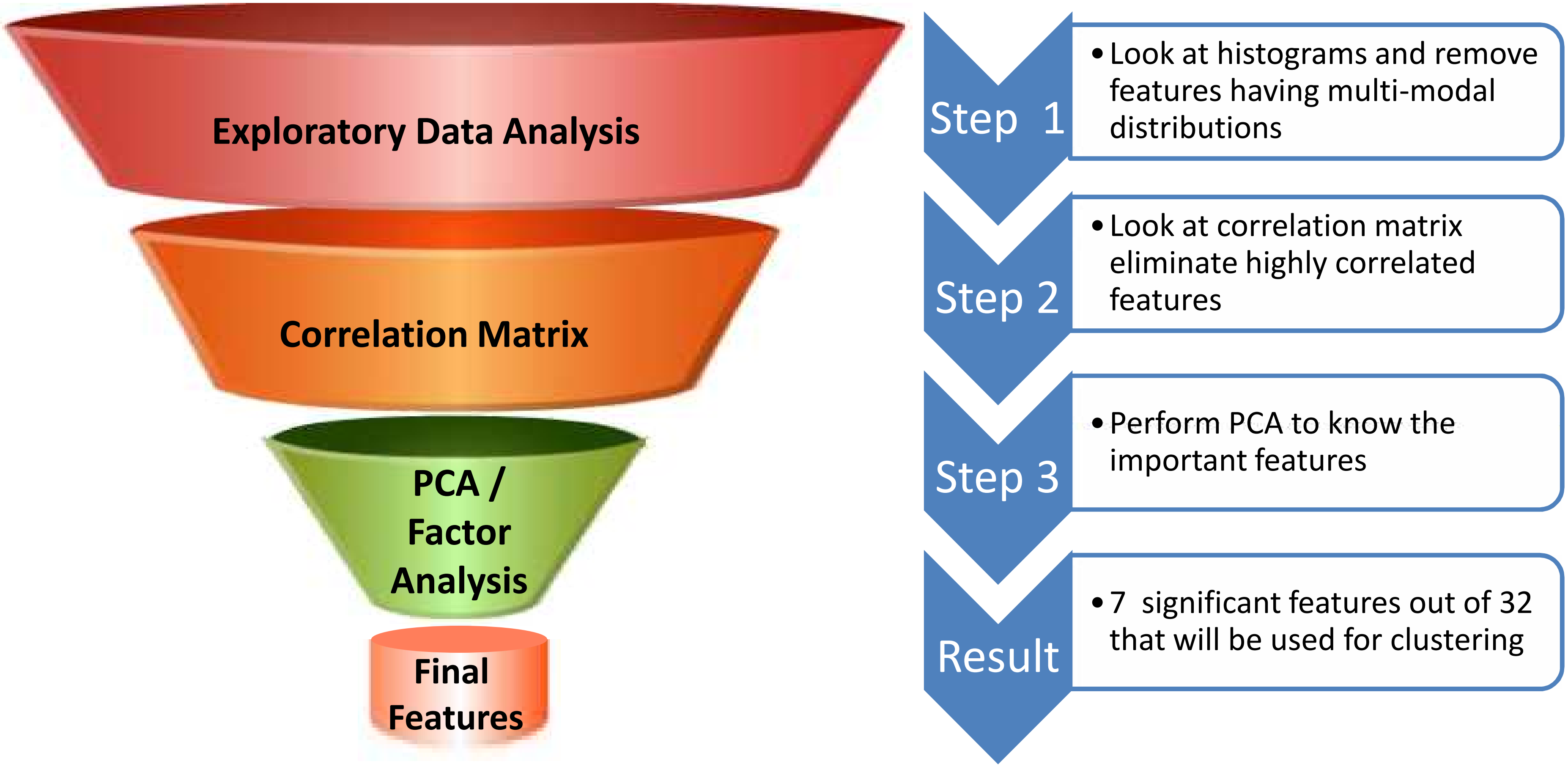
New Features

The new features that we create as part of our feature engineering exercise are

Features (1-16)
Avg Items brought in APPAREL Division
Avg Items brought in BAKERY_FRESH_KITCHEN Division
Avg Items brought in ELECTRONICS_APPLIANCES Division
Avg Items brought in FMCG Division
Avg Items brought in FRESH_FRUITS_VEGETABLES Division
Avg Items brought in HOME_NEEDS Division
Avg Items brought in LIQUOR_TOBACCO Division
Avg Items brought in NON_TRADING Division
Avg Items brought in NONVEG_DAIRY_FROZ Division
Avg Items brought in STAPLES Division
Avg Basket Size
Avg Basket Value
Avg Discount BasketSize
Avg Discount Basket Value
Avg Percentage Discount Basket Size
Avg Tax Free Basket Size

Feature (17-32)
Avg Tax Free Basket Value
Avg Percentage Tax Free Basket Size
Avg Free_Bee Basket Size
Avg Free_Bee Basket Value
Avg Percentage Free_Bee Basket Size
Number of Weekend Puchases
Number of Weekday Purchases
Number of Morning purchases
Number of Afternoon Purchases
Number of Evening Purchases
Number of Month start Purchases
Number of Month end Purchases
Average Days difference between consequent shopping
Avg Number of Unique Items
Number of Visits
Number of Visits before national holidays and Festivals

Variable Selection



Exploratory Data Analysis

Correlation Matrix

PCA / Factor Analysis

Final Features

Step 1

- Look at histograms and remove features having multi-modal distributions

Step 2

- Look at correlation matrix eliminate highly correlated features

Step 3

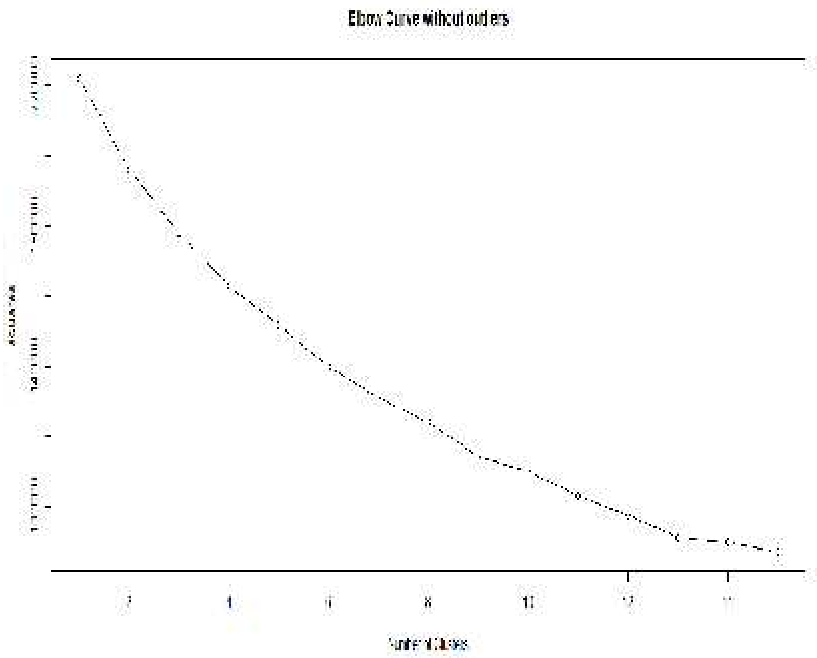
- Perform PCA to know the important features

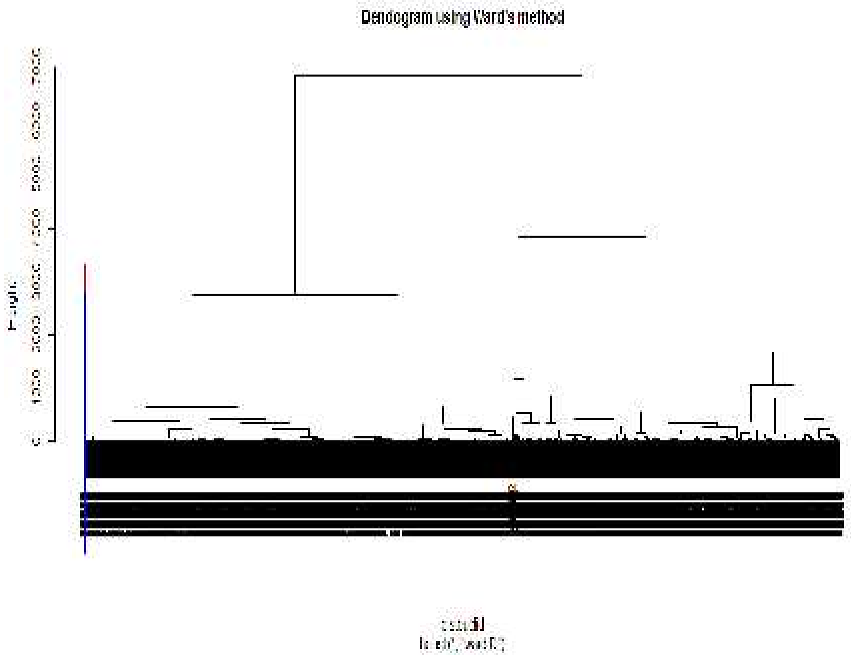
Result

- 7 significant features out of 32 that will be used for clustering

Avg Items brought in ELECTRONICS_APPLIANCES Division
Avg Items brought in LIQUOR_TOBACCO Division
Avg Items brought in FMCG Division
Avg Items brought in STAPLES Division
Avg Items brought in NON_TRADING Division
Avg Items brought in NONVEG_DAIRY_FROZ Division
Avg Items brought in FRESH_FRUITS_VEGETABLES Division

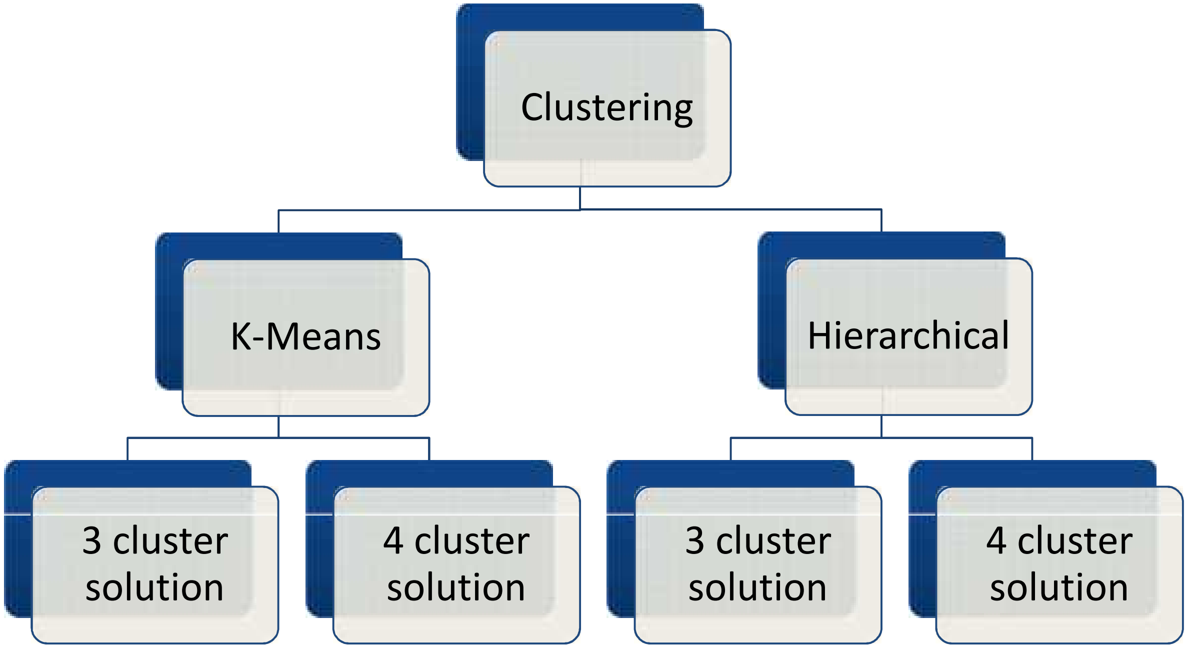
Number of Clusters





- Looking at the Dendogram and Elbow curve, we came to the conclusion that optimal number of clusters for our data are 3 (or) 4

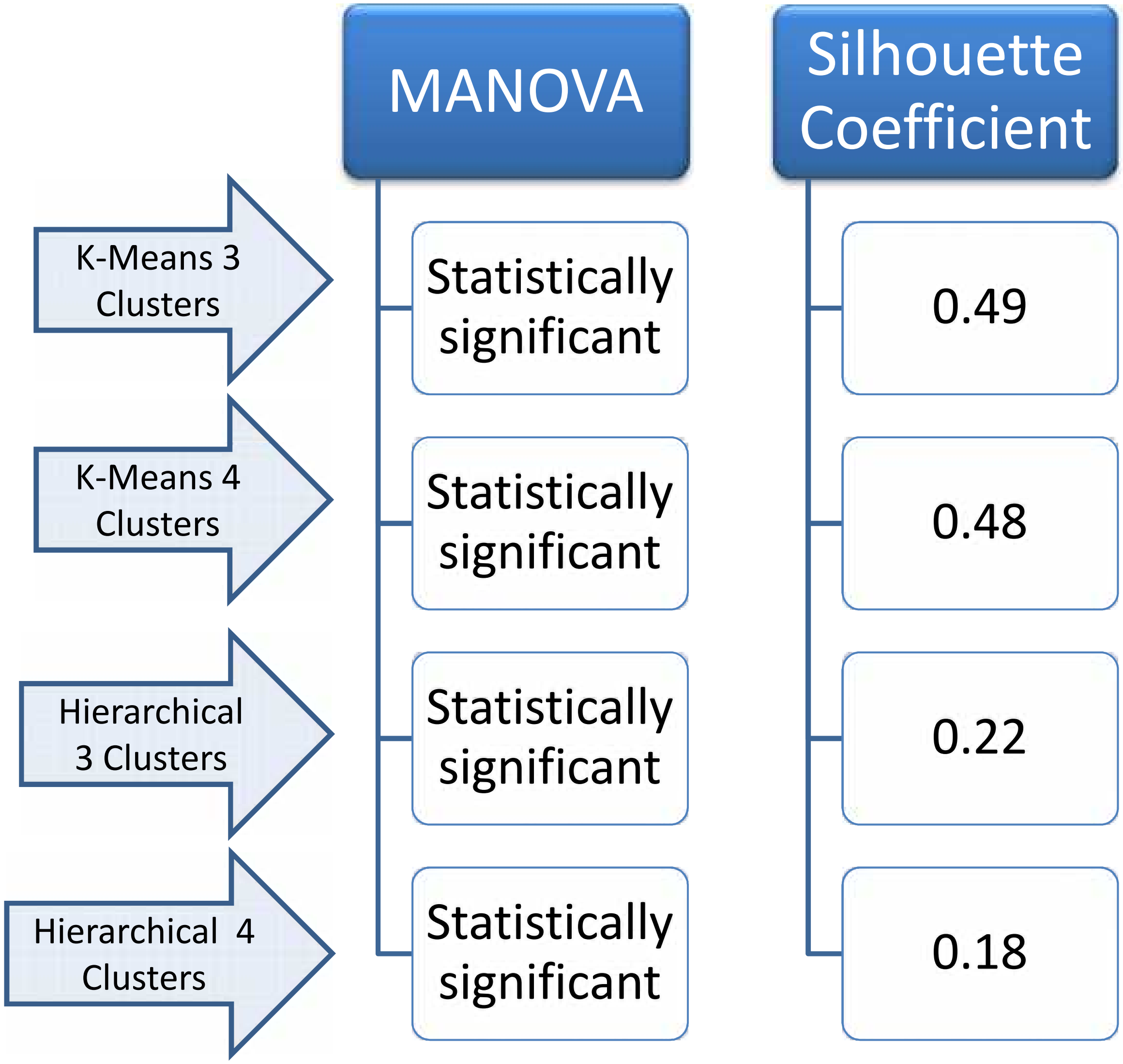
Clustering



- ***K-Means clustering*** was performed with initializing K = 3 & 4 i.e. Initial number of clusters as 3 & 4.

- ***Hierarchical clustering*** using Wards method was performed and both 3 cluster as well 4 cluster solution was saved.

Cluster Validation



	MANOVA	Silhouette Coefficient
K-Means 3 Clusters	Statistically significant	0.49
K-Means 4 Clusters	Statistically significant	0.48
Hierarchical 3 Clusters	Statistically significant	0.22
Hierarchical 4 Clusters	Statistically significant	0.18

- We checked if the clusters were statistically significant or not by applying MANOVA. All the cluster solutions were statistically significant.

- Then we calculated Silhouette coefficient for all the cluster solutions. The more the coefficient is closer to 1 , the better is the cluster solution.

- We chose the best cluster solution as K-Means 3 clusters going by the Parsimony rule.

- The best clusters are the ones got from :

K-Means 3 Clusters

Final Customer Profiles / Segments

High Value & Prodigal Shoppers

Daily needs & Monthly , Liberal Shoppers

Discount attractive , frugal shoppers

Segment 1

1) High value shoppers.
2) Shop MORE in APPAREL division
3) Shop MORE in BAKERY_FRESH_KITCHEN division
4) Shop MORE in ELECTRONICS_APPLIANCES, division
5) Shop MORE in HOME_NEEDS division
6) Shop MORE in LIQOUR_TOBACCO division
7) The average basket size is less but same average basket value when compared to segment 2. So, it means that they shop for items which are more costly. (Might be these are the customers who do NOT compromise on Quality for a price).
8) They are not so frequent visitors, may be monthly thrice.
9) Visit more before National Holidays and Festivals.

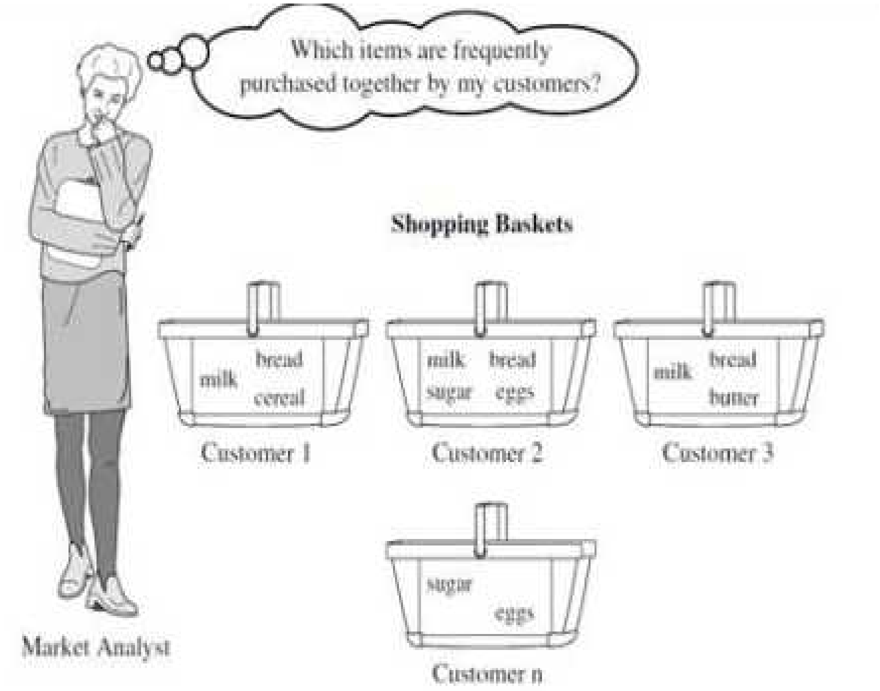
Segment 2

1) Medium value shoppers.
2) Shop MORE in FMCG division.
3) Shop MORE in FRESH_FRUITS_VEGETABLES division.
4) Shop MORE in STAPLES division.
5) The average basket size is very high, but the average basket value is same as cluster 1, so it means they shop for items which are less priced. (Might be these are the customers who compromise on Quality for a price).
6) These are the customers who buy MORE Tax free items.
7) They are the customers who visit less frequently. (May be monthly once or twice).

Segment 3

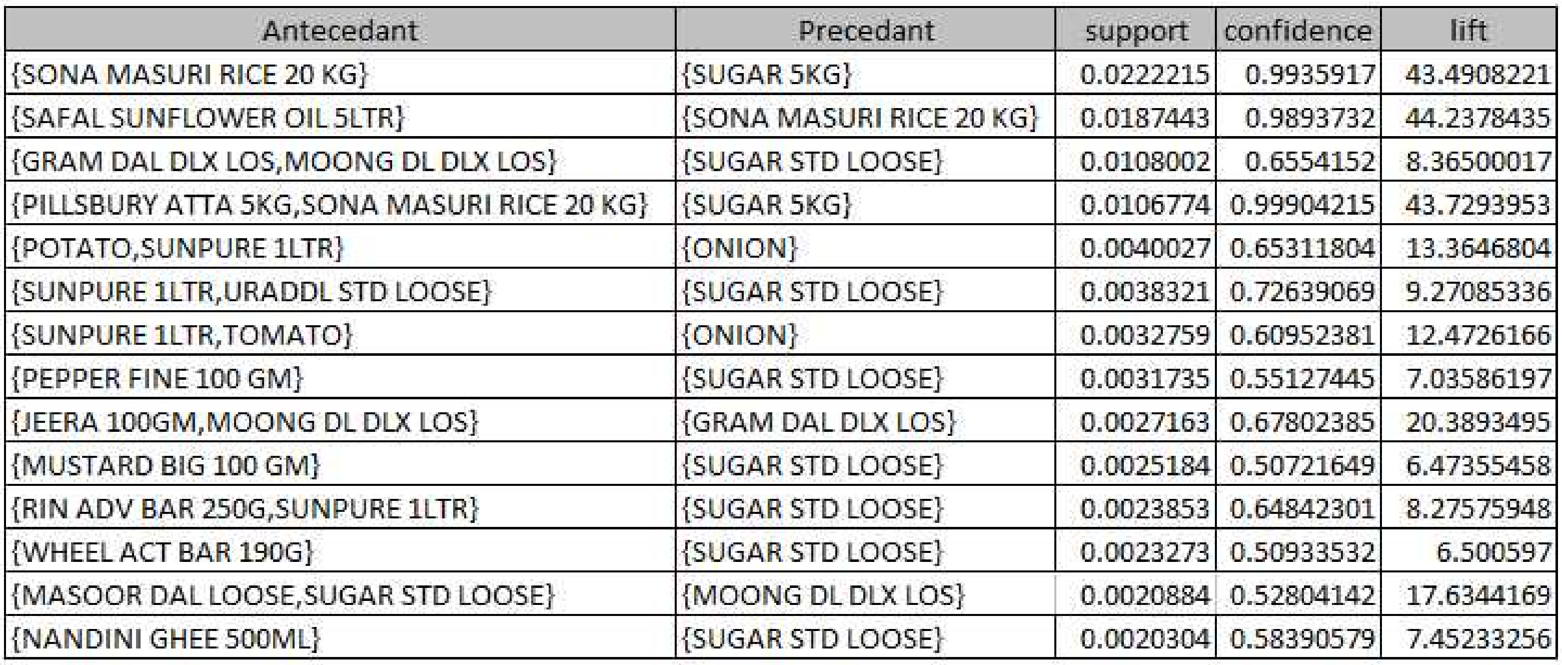
1) Low value shoppers.
2) They do not prefer shopping in ELECTRONICS AND APPLIANCES division.
3) These are the customers who are attracted by discounts i.e. they buy more discounted item.
4) These are also the customers who are attracted to freebies.
5) They are frequent visitors, may be weekly once.
6) When compared to other clusters, they buy equally in all the divisions.

Market Basket Analysis



Item / Product Association Rules

- Some of the top Item / Product association rules that we found are:



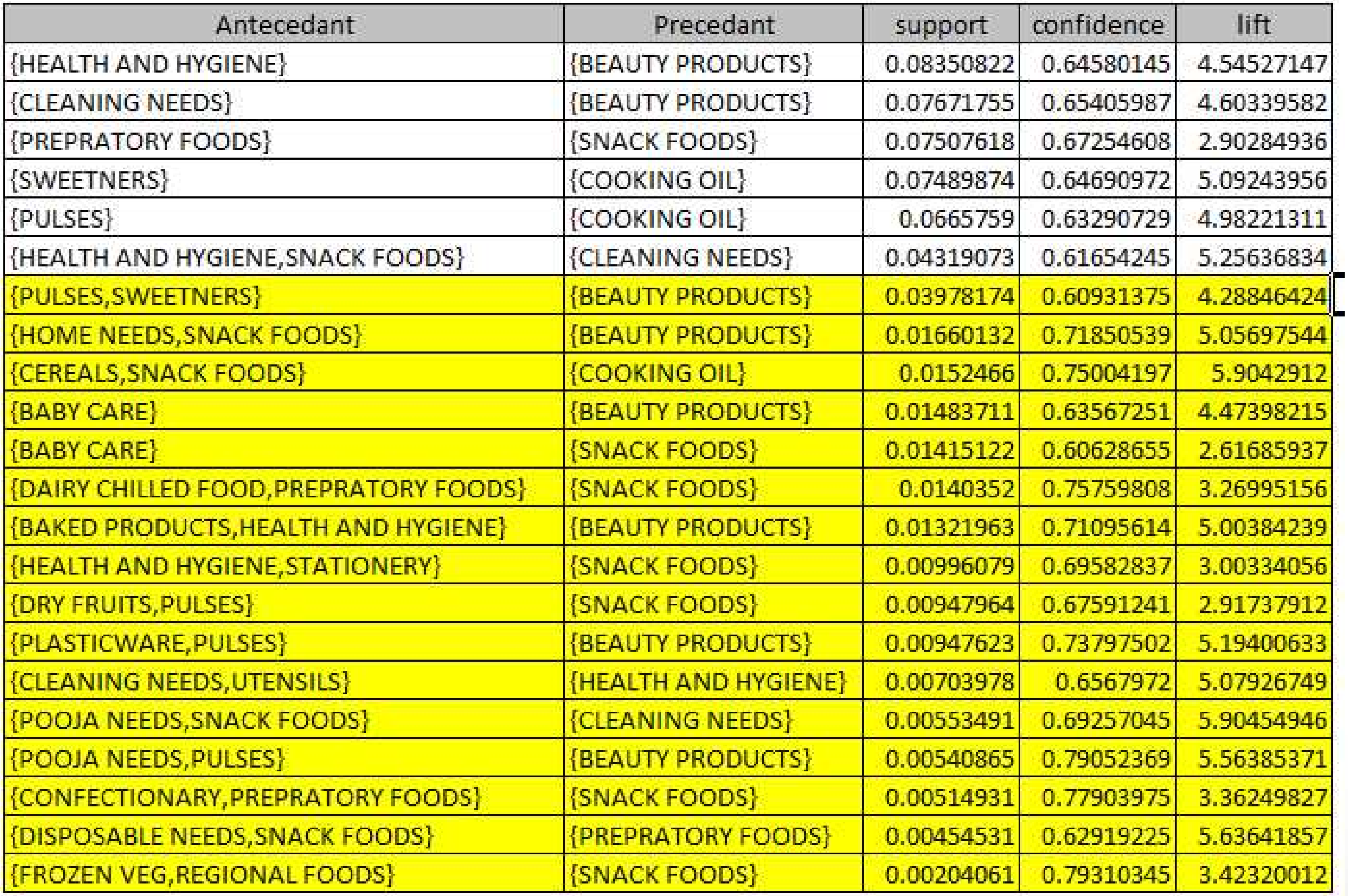
Antecedant	Precedant	support	confidence	lift
{SONA MASURI RICE 20 KG}	{SUGAR 5KG}	0.0222215	0.9935917	43.4908221
{SAFAL SUNFLOWER OIL 5LTR}	{SONA MASURI RICE 20 KG}	0.0187443	0.9893732	44.2378435
{GRAM DAL DLX LOS,MOONG DL DLX LOS}	{SUGAR STD LOOSE}	0.0108002	0.6554152	8.36500017
{PILLSBURY ATTA 5KG,SONA MASURI RICE 20 KG}	{SUGAR 5KG}	0.0106774	0.99904215	43.7293953
{POTATO,SUNPURE 1LTR}	{ONION}	0.0040027	0.65311804	13.3646804
{SUNPURE 1LTR,URADDL STD LOOSE}	{SUGAR STD LOOSE}	0.0038321	0.72639069	9.27085336
{SUNPURE 1LTR,TOMATO}	{ONION}	0.0032759	0.60952381	12.4726166
{PEPPER FINE 100 GM}	{SUGAR STD LOOSE}	0.0031735	0.55127445	7.03586197
{JEERA 100GM,MOONG DL DLX LOS}	{GRAM DAL DLX LOS}	0.0027163	0.67802385	20.3893495
{MUSTARD BIG 100 GM}	{SUGAR STD LOOSE}	0.0025184	0.50721649	6.47355458
{RIN ADV BAR 250G,SUNPURE 1LTR}	{SUGAR STD LOOSE}	0.0023853	0.64842301	8.27575948
{WHEEL ACT BAR 190G}	{SUGAR STD LOOSE}	0.0023273	0.50933532	6.500597
{MASOOR DAL LOOSE,SUGAR STD LOOSE}	{MOONG DL DLX LOS}	0.0020884	0.52804142	17.6344169
{NANDINI GHEE 500ML}	{SUGAR STD LOOSE}	0.0020304	0.58390579	7.45233256

Some Inferences that we can make from these rules are:

- Pulses sell more with oil.
- Also vegetables sell more with oils.
- Sugar was brought along with Detergent.
- Mustard , Jeera & Pepper were brought with Sugar.
- So if we can also place the sections Fabric Wash, Vegetables, Spices, Pulses and Oil together, then we can cross-sell more.

Category Association Rules

- We have seen Product Association rules, but we may get more insights when we find rules at Category level, let us have a look at them below:



Antecedant	Precedant	support	confidence	lift
{HEALTH AND HYGIENE}	{BEAUTY PRODUCTS}	0.08350822	0.64580145	4.54527147
{CLEANING NEEDS}	{BEAUTY PRODUCTS}	0.07671755	0.65405987	4.60339582
{PREPRATORY FOODS}	{SNACK FOODS}	0.07507618	0.67254608	2.90284936
{SWEETNERS}	{COOKING OIL}	0.07489874	0.64690972	5.09243956
{PULSES}	{COOKING OIL}	0.0665759	0.63290729	4.98221311
{HEALTH AND HYGIENE,SNACK FOODS}	{CLEANING NEEDS}	0.04319073	0.61654245	5.25636834
{PULSES,SWEETNERS}	{BEAUTY PRODUCTS}	0.03978174	0.60931375	4.28846424
{HOME NEEDS,SNACK FOODS}	{BEAUTY PRODUCTS}	0.01660132	0.71850539	5.05697544
{CEREALS,SNACK FOODS}	{COOKING OIL}	0.0152466	0.75004197	5.9042912
{BABY CARE}	{BEAUTY PRODUCTS}	0.01483711	0.63567251	4.47398215
{BABY CARE}	{SNACK FOODS}	0.01415122	0.60628655	2.61685937
{DAIRY CHILLED FOOD,PREPRATORY FOODS}	{SNACK FOODS}	0.0140352	0.75759808	3.26995156
{BAKED PRODUCTS,HEALTH AND HYGIENE}	{BEAUTY PRODUCTS}	0.01321963	0.71095614	5.00384239
{HEALTH AND HYGIENE,STATIONERY}	{SNACK FOODS}	0.00996079	0.69582837	3.00334056
{DRY FRUITS,PULSES}	{SNACK FOODS}	0.00947964	0.67591241	2.91737912
{PLASTICWARE,PULSES}	{BEAUTY PRODUCTS}	0.00947623	0.73797502	5.19400633
{CLEANING NEEDS,UTENSILS}	{HEALTH AND HYGIENE}	0.00703978	0.6567972	5.07926749
{POOJA NEEDS,SNACK FOODS}	{CLEANING NEEDS}	0.00553491	0.69257045	5.90454946
{POOJA NEEDS,PULSES}	{BEAUTY PRODUCTS}	0.00540865	0.79052369	5.56385371
{CONFECTIONARY,PREPRATORY FOODS}	{SNACK FOODS}	0.00514931	0.77903975	3.36249827
{DISPOSABLE NEEDS,SNACK FOODS}	{PREPRATORY FOODS}	0.00454531	0.62919225	5.63641857
{FROZEN VEG,REGIONAL FOODS}	{SNACK FOODS}	0.00204061	0.79310345	3.42320012

Category Association Rules (Contd.)

There were some interesting associations that we could find when we mined for rules at Category level. Interesting rules are marked in yellow in previous slide.

Some interesting insights are:

1) Baby Care & Beauty Products sell together
2) Baby Care & Snack foods sell together
3) Baked Products, Health & Hygiene & Beauty Products sell together
4) Dry Fruits sell more with Pulses
5) Plastic ware sell more with Beauty Products
6) Pooja Needs sell more with Cleaning needs & Beauty Products
7) Disposable Needs sell more with Snack foods & Preparatory foods
8) Confectionaries sell more with Snack foods.
9) Cooking Oil sells more with Pulses
10) Cereals and Cooking sell together

These categories stated above if placed side by side, we can achieve more cross-selling, thus increasing revenues.

Conclusion

Our analysis on the given dataset, gave us some interesting insights.

Some key benefits for the retailer by using the created customer segments and association rules are:

1) Targeting niche & potential Customer Segments
2) Improve targeted Marketing
3) Effective engagement and cross-sell/up-sell
4) To drive customer loyalty and sales

Thank you.