Problem Statement:

Categorize the customer into various segment according to their buying behavior. We are having sales data of various products since last 3 years and would be predicting the future sales of various products according to different customer requirement i.e. recommending the relevant product to the customer according to the past behavior of various customers lying in that particular segment (Building a Recommendation System)
This will enhance the sales and efficiency of our client through better categorization and recommendation of products to their customers.

Dataset Description:

Containing Bill number & Smart Card as customer unique ids with billing date & time.
Products details are given under Brick (Main category of product), Class (Sub-category), Material (Product's name) along with their price (MRP-Values) and their Segments.
Company's Gross Sales (Rs.) is also been provided along with the plant details and region of that plant.

B23 112894

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Bill Number	Billing Date	Brick	Brick.1	Calendar Month	Class	Class.1	Discounts	Family	Family.1	Gross Sales (Rs)	MRP-Values	Material
2	112851	01.12.2015	101020204	CREAM BISCUIT	12.2015	101020200	BISCUIT	0	101020000	PROCESSE	100	100	1007376
3	112851	01.12.2015	101150112	OLIVE OIL	12.2015	101150100	EDIBLE OIL	0	101150000	EDIBLE OII	699.02	699.02	1254544
4	112851	01.12.2015	101190101	APPLES & PEARS	12.2015	101190100	FRESH FRUIT	0	101190000	FRESH FRU	241.98	241.98	1030524
5	112853	01.12.2015	104050801	CONC-FOOT WEAR	12.2015	104050800	FOOT WEAR CONC	0	104050000	FOOTWAF	149	149	2094879
6	112853	01.12.2015	101020402	BLOCK CHOCOLATE	12.2015	101020400	CONFECTIONARY	0	101020000	PROCESSE	10	10	1007710
7	112853	01.12.2015	101020402	BLOCK CHOCOLATE	12.2015	101020400	CONFECTIONARY	0	101020000	PROCESSE	10	10	1007749
8	112853	01.12.2015	101170108	CONC-DRY FRUIT	12.2015	101170100	DRY FRUIT	0	101170000	DRY FRUIT	103.94	103.94	1247631
9	112853	01.12.2015	101170108	CONC-DRY FRUIT	12.2015	101170100	DRY FRUIT	0	101170000	DRY FRUIT	203.98	203.98	1247632
10	112853	01.12.2015	101190102	CITRUS	12.2015	101190100	FRESH FRUIT	0	101190000	FRESH FRU	114.47	114.47	1030530
11	112853	01.12.2015	103011001	POOJA AIDS	12.2015	103011000	POOJA NEED	0	103010000	HOME	25	25	1062478

B23 112894

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Family.1	Gross Sales (Rs)	MRP-Values	Material	Material.1	No.of Bills	Plant	Plant.1	Region	Sales Qty	Segment	Smart Card	Till Number	Time
2	PROCESSE	100	100	1007376	BRITANNIA BOURBON CREAM 60g-75g	1	H001	MD	T	10	FOOD	2.00001E+11	1	13:49:43
3	EDIBLE OII	699.02	699.02	1254544	FARRELL OLIVE POMACE OIL 1l	1	H001	MD	T	1	FOOD	2.00001E+11	1	13:49:43
4	FRESH FRU	241.98	241.98	1030524	APPLE ROYAL GALA LOOSE	1	H001	MD	T	1.22	FOOD	2.00001E+11	1	13:49:43
5	FOOTWAF	149	149	2094879	PLASTIC CHAPPELS	1	H001	MD	T	1	FASHION	2.0001E+11	1	13:52:51
6	PROCESSE	10	10	1007710	CADBURY 5STAR 21.5g	1	H001	MD	T	1	FOOD	2.0001E+11	1	13:52:51
7	PROCESSE	10	10	1007749	CADBURY DAIRY MILK 14g	1	H001	MD	T	1	FOOD	2.0001E+11	1	13:52:51
8	DRY FRUIT	103.94	103.94	1247631	STPL CONC KISHMISH YELLOW	1	H001	MD	T	0.21	FOOD	2.0001E+11	1	13:52:51

Tools Used:

Coding Language: Python 3.0
Libraries: Pandas, Numpy, Matplotlib, sklearn, Surprise, Apriori, mlxtend
Platform: PyCharm
Algorithm: Mini-Batch K-Means Clustering, Apriori, NMF, SVD, SVDpp

Solution:

Considering Smart Card as Unique Ids of the customers and extracting various features as required. User based filtering is done according to the various products (Bricks) and then Feeding it to the Mini-Batch K-Means Clustering to get clusters of customers and then on each cluster Recommendations are made through Association Rule and Collaborative Filtering.

PART: 1 (Clustering)

Steps:

- Importing required Libraries.
- Merging the data from various years/months to a single dataframe.
- Loading the Data through pandas (as df).
- General overview of the data like checking for the shape of the data, df.info(), if any null value exits in the data (df.isnull().sum()), df.describe.
- Now converting the date-time columns to datetime datatype and extracting month and year from that and also converting categorical columns to numeric using Label Encoder.
- Now, extracting the frequency of transactions (Smart Card) made by customers per month per year into new dataframe (df_1) and then GroupBy with Smart Card taking mean of frequency.

	Smart Card	year	month	freq
0	200000002307	2017	3	1
1	200000002307	2017	12	4
2	200000002307	2018	1	29
3	200000002307	2018	2	15

- In new dataframe (df_4), GroupBy with Smart Card and Brick per month per year and summing up Gross Sales and Sales Quantity as aggregation function and then again new dataframe (df_5) GroupBy previous dataframe with Smart Card and Brick taking median of Gross Sales and Sales Quantity.

	Smart Card	Class	Sales Qty	Gross Sales (Rs)
0	200000002307	101020200	2.0	25.0
1	200000002307	101020300	4.0	468.0
2	200000002307	101020400	2.0	62.0
3	200000002307	101020800	4.0	698.0

- Now converting these Brick into features for clustering (as pivot table).

- Step 7, creates a problem of converting because of huge size of data so needed to be broken into batches and performing the required operation.
- Creating a list a unique Smart Card (if not created, some duplicates tends to add up in the data) and selecting first 2000 unique Smart Cards and converting to pivot table with Bricks as columns for Gross Sales and Sales Quantity, storing in new dataframes.
- Joining these split datarames to create final dataframe (df_c) for the model.
- Merging this new dataframe (df_c) with the initial dataframe (df2_1, containing frequency) and also merging this dataframe with another dataframe (df3, containing median values) and resetting the index and dropping down some unwanted columns.

	Gross Sales (Rs)_101020100	Gross Sales (Rs)_101020200	Gross Sales (Rs)_101020300	Gross Sales (Rs)_101020400	Gross Sales (Rs)_101020500	Gross Sales (Rs)_101020600	Gross Sales (Rs)_101020700	Gross Sales (Rs)_101020800	(
Smart Card									
200000143376	240.0	115.0	303.0	89.0	160.0	0.0	50.0	471.0	
200000143950	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
200000144472	0.0	60.0	0.0	17.5	0.0	0.0	0.0	0.0	
200000146647	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	
200000276489	0.0	0.0	0.0	0.0	70.0	0.0	0.0	0.0	

- As the final dataframe has huge size, K-Means clustering leads to high usage of memory, so the alternate solution was to do with Mini-Batch K-Means Clustering with batch size of 2000.
- Final dataframe has been developed. Now, it is to be find the optimized value of k (Number of Clusters).
- This optimized value of k is evaluated using Elbow Method and Silhouette Score. And the optimal value evaluated is: k = 19 clusters.
- After finding out the k value, predicting the clusters for each customer using the optimal k value and adding this new column (Cluster) to the final dataframe.

Smart Card	Cluster
2.00001E+11	0
2.0001E+11	0
2.00008E+11	0
2.00014E+11	0
2E+11	0
2.00013E+11	0
2.00008E+11	0
2.00014E+11	4
2.00012E+11	0
2.00006E+11	0
2.00012E+11	0
2.00001E+11	4

PART: 2 (Recommendations using Apriori)

Steps:

- Load Dataframe with frequency w.r.t Customer (Smart card) and Product along with the Cluster id for each customer.

	Smart Card	Class.1	freq
0	200000018405	BAKERY	1
1	200000018405	BISCUIT	5
2	200000018405	BROWN SPIRITS	4
3	200000018405	CARRY BAG WOVEN	4
4	200000018405	CEREALS	1

	Smart Card	Class.1	Cluster	freq
0	200000018405	BAKERY	1	1
1	200000018405	BISCUIT	1	5
2	200000018405	BROWN SPIRITS	1	4
3	200000018405	CARRY BAG WOVEN	1	4
4	200000018405	CEREALS	1	1

- Now, splitting the Dataframe for each Cluster id and storing these in list.
- Forming pivot table for each dataframe containing unique cluster using unique smart card, avoiding duplication of customers in pivot table.
- Now, a function is created in which data (for first cluster) is given to Apriori algorithm and associations rule are generated (support is calculated from frequency data) and then a loop is run to remove duplications of associations (redundancy). Then products are extracted from dataframe that are purchased by each customer and stored in a dataframe as set. Now, if antecedents are subset of purchased items by customer then consequents are added to the recommended list for that customer.

	antecedents	consequents	support	confidence	lift
141270	(SKIN CARE, UTENSIL CLEANER)	(EDIBLE OIL, MASALA & SPICE)	0.153846	1.0	5.571429
67697	(FLOOR CARE, INSECTICIDE)	(SKIN CARE, CARBONATED)	0.076923	1.0	6.500000
67758	(FLOOR CARE, CARBONATED, KITCHEN TOOLS METAL)	(UTENSIL CLEANER)	0.076923	1.0	5.571429
67757	(UTENSIL CLEANER, FLOOR CARE, KITCHEN TOOLS ME...	(CARBONATED)	0.076923	1.0	6.500000
67756	(UTENSIL CLEANER, FLOOR CARE, CARBONATED)	(KITCHEN TOOLS METAL)	0.076923	1.0	5.571429

- Then these recommended products are converted into sets and compared with others for same customer, removing repetition of products and then it is stored in dataframe w.r.t Smart Card as final output for that cluster.
- This function is executed for other dataframes containing different clusters ids.

PART: 3 (Recommendations using Collaborative Filtering)

Steps:

- Load Dataframe with frequency w.r.t Customer (Smart card) and Product along with the Cluster id for each customer (same as previous part).
- Removing customers with frequency lower than 50 and product frequency lower than 200, as these two does not have much impact on the data due very less transactions in long time period and recommendation may also be less efficient for these customers.
- Now, this data is divided into some quantiles on frequency and then based on this, grouping is done with different dataframes containing unique cluster id's (these groups represents rating) and this is final dataframe, this act as rating by each customer.

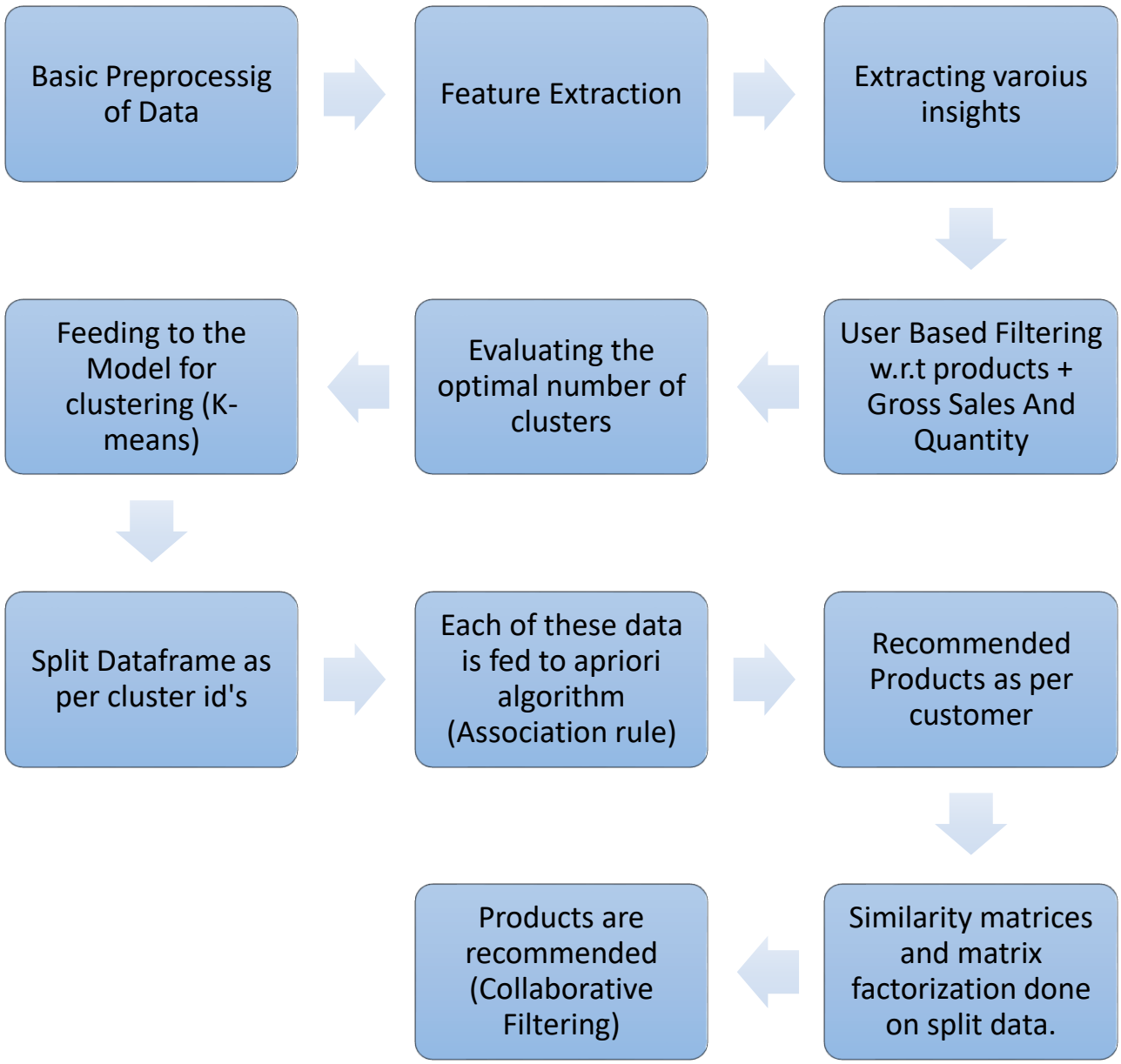
	Smart Card_	Class.1_	Cluster_	freq_min	freq_q2	freq_q3	freq_q4	freq_max	freq
0	200000025584	BATH	1	1	1	1	1	1	2.336207
1	200000025584	BISCUIT	1	1	1	2	3	4	2.336207
2	200000025584	BREAKFAST CEREAL	1	1	1	1	2	2	2.336207
3	200000025584	CARBONATED	1	1	1	1	2	2	2.336207

	Smart Card_	Class.1_	Cluster_	freq	group
0	200002496528	BAKERY	0	2.58	7
1	200002496528	BATHROOM AIDS	0	2.58	7
2	200002496528	BED	0	2.58	7
3	200002496528	BISCUIT	0	2.58	6

- Then Surprise library is imported with its various algorithms and final dataframe is fed to reader with rating from 1 to 7. And then a loop is run through various algorithm like NMF, SVD, SVDpp and KNNBasics etc. and most efficient algorithm is selected and then data is split into train and test dataset.
- These algorithm includes Matrix factorization, cosine similarity matrix, decomposing a matrix, using nearest neighbor etc. and NMF (matrix factorization) is selected algorithm for training the data for user-item matrix and splitting it to from individual matrix.
- Now, this is used for all customers to recommend products (as per similarity matrix) but the products which are already purchased by that customer are removed while recommending new products to that customers and recommend top 10 products.
- Then these recommendations are merged with smart card as final output.

	Smart Card	Recommended_Collaborative_Filtering
0	200002496528	BATH, COLOUR COSMETICS, FISH, GLASS WARE, GLUC...
1	200011646338	BATH, BED, CERAMIC WARE, FROZEN NON VEG, GLUCO...
2	200013885694	BATH, CERAMIC WARE, FASHION CONC, FROZEN NON V...
3	200013886073	EGG, FROZEN, FURNISHING, GLASS WARE, INDIAN WE...

Flow Chart:


Basic Preprocessig of Data
Feature Extraction
Extracting varoius insights
User Based Filtering w.r.t products + Gross Sales And Quantity
Evaluating the optimal number of clusters
Feeding to the Model for clustering (K-means)
Split Dataframe as per cluster id's
Each of these data is fed to apriori algorithm (Association rule)
Recommended Products as per customer
Similarity matrices and matrix factorization done on split data.
Products are recommended (Collaborative Filtering)

Output:

By Collaborative Filtering

	Smart Card	Recommended_Collaborative_Filtering
0	200002496528	BATH, COLOUR COSMETICS, FISH, GLASS WARE, GLUCOSE & ENERGY, INDIAN WEAR TOPS, LAB, SHOE CARE, SPORT
1	200011646338	BATH, BED, CERAMIC WARE, FROZEN NON VEG, GLUCOSE & ENERGY, MEN'S BOTTOMWEAR, MILK FOOD, PACKD DRINKING WATER, READY TO EAT
2	200013885694	BATH, CERAMIC WARE, FASHION CONC, FROZEN NON VEG, FROZEN VEG, GLUCOSE & ENERGY, KITCHEN, MILK FOOD, SERVE/TABLE WARE
3	200013886073	EGG, FROZEN, FURNISHING, GLASS WARE, INDIAN WEAR TOPS, MEN'S INNERWEAR, PACKD DRINKING WATER, READY TO EAT, SHOE CARE
4	200000025584	COOKWARE METAL, DISINFECTANT AND MUL, ELECTRICAL EQUIPMENT, FASHION CONC, FROZEN NON VEG, GIFT CARD & VOUCHERS, GLASS WARE, KIDS BOYS BWEAR, KITCHEN
5	200000136756	ALLNCE & MISC INCOME, FROZEN NON VEG, GIFT CARD & VOUCHERS, GLASS WARE, GLUCOSE & ENERGY, KID'S BOYS TOP WEAR, KIDS BOYS BWEAR, KITCHEN, SERVE/TABLE WARE
6	200000166359	COLOUR COSMETICS, FASHION CONC, FROZEN NON VEG, KIDS BOYS BWEAR, KITCHEN, KITCHEN APPLIANCES, MEN'S BOTTOMWEAR, OTC FOOD, OTHER RTD
7	200000210933	ALLNCE & MISC INCOME, DISINFECTANT AND MUL, GIFT CARD & VOUCHERS, KID'S BOYS TOP WEAR, LAB, MEN'S INNERWEAR, OTHER RTD, SHOE CARE, SPORT
8	200000220444	BATH, COLOUR COSMETICS, FROZEN NON VEG, FURNISHING, KITCHEN, OTHER RTD, PACKD DRINKING WATER, SERVE/TABLE WARE, SHOE CARE
9	200000240642	BATH, CERAMIC WARE, FROZEN NON VEG, FURNISHING, INDIAN WEAR TOPS, MEN'S INNERWEAR, PACKD DRINKING WATER, SERVE/TABLE WARE, SYRUPS & CONCENTRATE

By Association Rule (Apriori)

	Smart Card	Recommended_Association_Rule
0	200000379607	SWEETENER, FEMININE HYGIENE, VISUAL, TOILET CLEANER, FRESH VEG, MEN'S TOPWEAR, SAVOURIES, INSECTICIDE, AIR FRESHENER, UTENSIL CLEANER, STATIONERY, DRY FRUIT, PLASTICS & THERMOWAR, CONFECTIONARY, PATISSERIE, HAIR CARE, KITCHEN TOOLS METAL, ORAL CARE, CARRY BAG WOVEN, LAUNDRY, HEALTH DRINKS, ELECTRICAL EQUIPMENT, CEREALS, PULSES, FOOT WEAR CONC, BISCUIT, FLOUR, SKIN CARE, FOOD SERVICE, PACKAGED TEA, MASALA & SPICE, KIDS BOYS BWEAR, BREAKFAST CEREAL, SOUP, EDIBLE OIL, UTENSIL CLEANER, FRESH FRUIT, DAIRY, CLEANING AID, FLOOR CARE, POOJA NEED, JAMS/HONEY/SPREAD, NOODLE/VERMICELLI, READY TO FRY, GIFT CARD & VOUCHERS, BAKERY, DRINKS & JUICES, CARBONATED
1	200001364356	VISUAL, VISUAL, COOKWARE METAL, GIFT CARD & VOUCHERS
2	200002069379	SWEETENER, VISUAL, TOILET CLEANER, FRESH VEG, MEN'S TOPWEAR, SAVOURIES, INSECTICIDE, AIR FRESHENER, UTENSIL CLEANER, STATIONERY, DRY FRUIT, PLASTICS & THERMOWAR, CONFECTIONARY, HAIR CARE, PATISSERIE, DISPOSABLE, KITCHEN TOOLS METAL, ORAL CARE, CARRY BAG WOVEN, LAUNDRY, DRINKS & JUICES, HEALTH DRINKS, ELECTRICAL EQUIPMENT, BROWN SPIRITS, CEREALS, PULSES, FOOT WEAR CONC, BISCUIT, FLOUR, SKIN CARE, FOOD SERVICE, KIDS BOYS BWEAR, MASALA & SPICE, PACKAGED TEA, BREAKFAST CEREAL, SOUP, EDIBLE OIL, FRESH FRUIT, DAIRY, SALT, CLEANING AID, BATHROOM AIDS, ORGANIC FOOD, POOJA NEED, JAMS/HONEY/SPREAD, NOODLE/VERMICELLI, FLOOR CARE, MASALA & SPICE, BAKERY, READY TO FRY, CARBONATED, GIFT CARD & VOUCHERS
3	200002496528	SWEETENER, FEMININE HYGIENE, VISUAL, TOILET CLEANER, KETCHUP/SAUCE, FRESH VEG, MEN'S TOPWEAR, SAVOURIES, INSECTICIDE, AIR FRESHENER, UTENSIL CLEANER, STATIONERY, DRY FRUIT, PLASTICS & THERMOWAR, CONFECTIONARY, PATISSERIE, HAIR CARE, DISPOSABLE, KITCHEN TOOLS METAL, ORAL CARE, CARRY BAG WOVEN, LAUNDRY, DRINKS & JUICES, HEALTH DRINKS, ELECTRICAL EQUIPMENT, BROWN SPIRITS, CEREALS, PULSES, AUDIO, FOOT WEAR CONC, BISCUIT, DISINFECTANT AND MUL, FLOUR, SKIN CARE, FOOD SERVICE, BREAKFAST CEREAL, MASALA & SPICE, PACKAGED TEA, KIDS BOYS BWEAR, SOUP, EDIBLE OIL, UTENSIL CLEANER, FRESH FRUIT, DAIRY, SALT, COFFEE, CLEANING AID, BABY NEEDS, BATHROOM AIDS, ORGANIC FOOD, POOJA NEED, JAMS/HONEY/SPREAD, COOKWARE METAL, READY TO FRY, FLOOR CARE, NOODLE/VERMICELLI, BAKERY, GIFT CARD & VOUCHERS, CARBONATED
4	200003223479	SWEETENER, FEMININE HYGIENE, TOILET CLEANER, VISUAL, FRESH VEG, MEN'S TOPWEAR, SAVOURIES, INSECTICIDE, AIR FRESHENER, UTENSIL CLEANER, DRY FRUIT, PLASTICS & THERMOWAR, CONFECTIONARY, PATISSERIE, HAIR CARE, KITCHEN TOOLS METAL, ORAL CARE, CARRY BAG WOVEN, LAUNDRY, HEALTH DRINKS, ELECTRICAL EQUIPMENT, BROWN SPIRITS, CEREALS, PULSES, AUDIO, FOOT WEAR CONC, BISCUIT, FLOUR, FOOD SERVICE, SKIN CARE, BREAKFAST CEREAL, MASALA & SPICE, KIDS BOYS BWEAR, PACKAGED TEA, SOUP, EDIBLE OIL, FRESH FRUIT, DAIRY, COFFEE, CLEANING AID, BABY NEEDS, FLOOR CARE, ORGANIC FOOD, POOJA NEED, JAMS/HONEY/SPREAD, LAUNDRY, GIFT CARD & VOUCHERS, NOODLE/VERMICELLI, READY TO FRY, BAKERY, DRINKS & JUICES, CARBONATED