MACHINE LEARNING I

Assignment I – Technical Report

Group Assignment

Daniel Bilitewski, Guillermo Chacón, Nisrine Ferahi, Jonas Hellevang, Mariana Narváez, Rabiga Shangereyeva

10th November 2019

Introduction

This technical report summarizes the Cluster analysis done to the stores of a retail company located in Bogotá. This data was given by a real retail company with the objective of classifying their stores in categories based on the demographic and sales data at hand, to better distribute the product portfolio in their stores to maximize sales. Also, this company made their own categorization in the past with 4 categories, the scope of this report is also comparing our cluster analysis with their personal categorization and provide insights if there are significant differences.

General Objective:

This analysis seeks to optimize the strategies to grow the different categories, understanding their behaviour according to their sales and the environment of the stores where they perform better.

Specific Objectives:

- Cluster stores according to their demographic behaviour
- Cluster stores according to their behaviour for sale
- Understand super category behaviour by crossing the clusters

Data sets:

4 data sets were provided for the analysis at hand, in the next lines we will explain the nature of each data set and the handling and preparation done to them before the analysis.

Sales by store and dictionary:

This first data set included the sales by store/category/month for all 2019.

Table 1: Sales by store raw data.

Type	Month	Store	Store ID	Product Cat	Sales 2019	Sales 2018	With or Without homes
REAL	01-01-2019	140 BOG	50WHX	ABARROTES COMESTIBLE	1.501	2.077	CON DOMICILIOS
REAL	01-01-2019	Andino BOG	50IIU	ABARROTES COMESTIBLE	0	0	CON DOMICILIOS
REAL	01-01-2019	ATABANZA	504EU	ABARROTES COMESTIBLE	2.535	0	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	ALIM. REFRIG. Y FRES	1.464	1.611	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	ARTS. DUMMY	0	0	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	ARTS. NO VENDIBLES	0	0	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	ARTS. PROMOCIONALES	3	0	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	BEBIDAS ALTERNATIVAS	5.238	6.071	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	BOTANAS	7.087	6.654	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	CERVEZA	9.457	8.464	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	CIGARROS	36.653	26.226	CON DOMICILIOS
REAL	01-01-2019	140 BOG	50WHX	COMPLEMENTOS REUNION	120	191	CON DOMICILIOS

The description of the variables is as follow:

Table 2: Sales by Store variable description.

Variable Name	Description
Type	Base column with only “Real” output
Month	Month of the year 2019
Store	Name of the store
Store ID	Unique ID of the Store for identification
Product Cat	Category of the product sold out of 31 possible

Sales 2019	2019 Sales of the month
Sales 2018	2018 Sales of the month
With or Without homes	Category column that indicates if sales include delivery to homes

A second file called “dictionary” was provided that included a broad categorization of products into 6 category types:

Table 3: Need vs Product category table

Super Grupo	Necesidad
BOTANAS	Antojo
DULCES	Antojo
GALLETAS	Antojo
HELADOS	Antojo
REPOSTERIA	Antojo
Abarrotes comestible	Diario y Reposición
ALIM. REFRIG. Y FRES	Diario y Reposición
HIGIENE HOGAR	Diario y Reposición
HIGIENE Y SALUD PERS	Diario y Reposición
LECHE	Diario y Reposición
MERCANCIAS GENERALES	Diario y Reposición
PAN TOST Y TORTILLAS	Diario y Reposición
FAST FOOD ALIMENTOS	Hambre
FAST FOOD BEBIDAS	Hambre
FRUTA Y VERDURA	Hambre
CIGARROS	Optimización
ENTRETENIMIENTO	Optimización
TELEFONIA	Optimización
CERVEZA	Reunión
COMPLEMENTOS REUNION	Reunión
HIELO	Reunión
VINOS Y LICORES	Reunión
AGUA EN GARRAFON	Sed
AGUA PURIFICADA	Sed
BEBIDAS ALTERNATIVAS	Sed
JUGOS	Sed
REFRESCOS	Sed
YOGHURT	Sed

The raw data was not suitable for the cluster analysis, so aggregation was needed to get the total sales of the year by store/product/category.

Using excel we made a cross between the broad categories and the original ones, creating a new column called “Need”:

Table 4: sales by store with “Need” Column

Type	Month	Store	Store ID	Product Cat	Sales 2019	Sales 2018	With or Without homes	Need
REAL	01-01-2019	140 BOG	50WHX	ABARROTES COMESTIBLE	1.501	2.077	CON DOMICILIOS	Diario y Reposición
REAL	01-01-2019	Andino BOG	50IIU	ABARROTES COMESTIBLE	0	0	CON DOMICILIOS	Diario y Reposición

There were also some “product cat” that didn’t fit any “need”

Table 5: sales by store, products with no need

Type	Month	Store	Store ID	Product Cat	Sales 2019	Sales 2018	With or Without homes	Need
REAL	01-01-2019	140 BOG	50WHX	ARTS. DUMMY	0	0	CON DOMICILIOS	#N/A
REAL	01-01-2019	140 BOG	50WHX	ARTS. NO VENDIBLES	0	0	CON DOMICILIOS	#N/A
REAL	01-01-2019	140 BOG	50WHX	ARTS. PROMOCIONALES	3	0	CON DOMICILIOS	#N/A

After analysing the sales volume, we decided to remove these products because their volume was significantly lower than the other categories and won’t affect the clustering analysis overall

Finally using pivot tables, we aggregated the whole sales by store in rows and product categories/need in columns:

Table 6: Example of aggregation using pivot table, "need" variables.

Con/sin Domicilios	CON DOMICILIOS								
Mes	(Todas)								
Suma de Ventas Totales 2019 (Mil)		Necesidad							
CR	Tienda	Antojo	Diario y Reposición	Hambre	Optimización	Reunión	Sed	#N/A	Total general
500LJ	Avenida Jimenez BOG	176337,28	61001,14828	60189,40284	216239,6472	170813,3726	177092,4132	-18,11915	861655,145
5016I	PARQUE DE LOS NIÑOS	140728,8013	51306,73578	127917,7601	233827,2961	158889,4536	128074,9338	-391,52051	840353,4601
501YH	America Occidental BOG	137697,7706	38440,37268	57722,87189	169361,9775	82793,00395	99447,75326	277,84592	585741,5958
5021R	Cedis Bucaramanga							0	0
5029F	ARANZOQUE BOG	9264,91166	2084,33916	14080,4895	7567,62829	18032,25047	11059,64804	-1	62088,26712
503E4	Torre Americas BOG	112050,7332	1893,76024	94578,89566	39538,43634	19069,59788	78572,06263	-4	345699,486
503K5	GONZALEZ VALENCIA	208820,2866	85784,58912	291128,2283	169962,1305	83404,20933	224025,3227	16,41734	1063141,184
504CA	Chico 96 BOG	231293,6412	32766,5308	143205,9148	217621,1403	83169,5066	167580,8365	23,78751	875661,3577
504EU	ATABANZA	387039,8282	155672,9831	264624,7888	195211,478	244310,5565	216002,3621	4278,67617	1467140,673

As you can see in the aggregation of the data there were stores that didn't present any sales data at all like "Cedis Bucaramanga", stores like this were removed from the table.
The final table contained 37 columns and 93 stores.

Store profile:

A second file called "StoreProfile" was provided, this file contained demographic and geographic information of 65 stores in 96 columns:

Table 7: Storeprofile data example

StoreID-Nielsen	PDV	Store	HHs	POBLACION	ARA	CANASTAPOPULAR	CARULLA-EX	COLSUBSIDIO	EXITO	INDEPENDENT	JUMBO	JUSTO&BUENO	KOBA-D1	LA14	LAECONOMIA-ETICOS
111	UNIVERSITARIO	30719	5395	14.079	1	0	0	0	2	2	0	1	0	0	0
222	CALIMA	36158	1881	7.002	0	0	0	0	1	3	1	2	0	1	0
333	SAN ANDRESITO	35973	4013	14.659	0	0	0	0	0	9	0	1	0	0	0
444	INNOVA	36157	3771	12.673	2	0	0	0	0	2	0	1	1	0	0
555	KENNEDY	35972	7499	29.036	0	0	0	0	0	10	0	0	1	0	0
5688	140	16403	4354	14.571	1	0	1	0	1	4	0	2	1	0	0
5690	AUTOPISTA	15318	7129	24.930	0	0	0	0	0	5	0	0	3	0	0
5691	BRITANIA	14446	7345	26.506	0	0	0	0	0	3	0	1	1	0	0
5692	CALATRAVA	17992	2987	9.508	2	0	2	0	0	3	1	1	2	0	0
5693	CALLE 19	18620	4158	10.343	3	0	1	0	2	1	0	2	3	0	0
5694	CALLE 67	18485	4141	11.615	0	0	2	1	1	3	0	0	3	0	0
5695	LA CANDELARIA	18621	4027	10.789	0	0	0	0	1	2	0	0	0	0	0

The description of the variables is as follow:

Table 8: Storeprofile variable description.

Variable Name	Description
StoreID_Nielsen	Unique Id for each store (different from the one in Sales by store
PDV	Name of the Store
Store	Second ID of the store
HH's	Unique ID of the Store for identification
Población	Total population in a **500m radius**
Rows G to Z	# of competitors stores located in the same radius by competitor
Rows AA to AC	Population by Socioeconomic level (low, med, high)
Rows AD to AU	Gender and Age distribution of column population
Rows AV to BM	Family demographics (# kids, age, family members)
Rows BN to BP	Same as AA to AC
Rows BQ to CR	# of other business located in the same radius EX: Universities, banks, restaurants, pizza stores, bakeries, markets, pharmacies, cigarette stores.

The distribution of this table fitted the requirements for the store cluster, the problem was that there was no ID Store to match with "Sales by Store", so a manual creation of the store ID (Row C) was done to make the join afterwards.

Store features:

This last file included the data of some features of the store, the variable description was the following:

Table 9: Store features variable description.

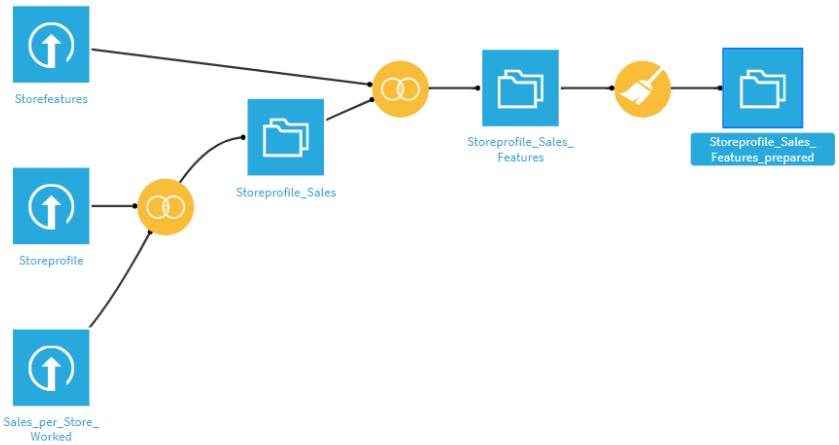
Variable Name	Description
CR	Store ID (same as Sales by Store)
Store	Name of the Store
Store_status	Active of Closed category of the store
Location	Broad location of the store
Segment	Classification made by the company of the store
Economic_model	Main business model generator category
TMCB	Code categorization for new or old stores
Store_type	Same as TCMB but in New or TMCB store
Open Date	Date that indicates the opening of the store
Surface	Square meter surface
Size	Size category
Estrato	Number category of the socioeconomic target
Refrigeration_type	Category of the type of freezing system
Doors	Number of Stores

This file was ok for the model but with several variables missing or with the category "To confirm". In the next section we will explain how we dealt with this situation.

Data Preparation for the Cluster:

After doing the preliminary preparation of the data we used dataiku system to aggregate and clean the data, we used the following flow diagram:

Graphic 1: Dataiku Flow Diagram.



For the join between the databases we did an inner join between "StoreProfile" and "Sales per Store" as we believed this two tables were mandatory for the cluster analysis, this narrowed the number of stores down to 65. Then second joint was a left joint between "Storeprofile_sales" and "Storefeatures" that left some rows without features.

The cleaning included the fill up of all the missing and "por Confirmar" values and features, the steps were the following:

- In Doors, Estrato, Size columns, "por confirmar" values were cleared.
- Refrigeration type, Doors, Estrato, Size, Store Type, TMCB, Economic Model, Segment, Location and Store Status were filled with the mode value from their respective cells.

- Floor_size was filled with the average number.

Some extra preparations were made for the particular clusters we did, the details come in the next chapter along with the cluster execution.

Demographic Cluster:

The objective of the Demographic cluster is to provide the client with a new insight based on the location of their stores and the nature of their vicinity business, we will use this clustering as a compass to decide the ideal sales and store strategy that they should pursue and compare it with the current one they have.

Data Handling:

Table 10: Demographic data example.

PDV	CR	POBLAI	TOTALC	ARA	CANAS	CARULI	COLSUI	EXITO	INDEPE	JUMBO	JUSTO	KOBAD	LA14	LAECOI	METRO	OLIMPI	ROMI	SUPER	SURTIF	SURTIN	SURTIN	UNIME	ZAPATI	NSEBAJ	NSEME	NSEALT	WNSEBA
UNIVERSIT	50I1U	14079.0	8.0	1.0	0.0	0.0	0.0	2.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4149.0	9930.0	0.0	29.4
CALIMA	50C7J	7002.0	8.0	0.0	0.0	0.0	0.0	1.0	3.0	1.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0	4350.0	2635.0	0.2
SAN ANDR	50M9T	14659.0	10.0	0.0	0.0	0.0	0.0	0.0	9.0	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	0.0	14659.0	0.0	0.0
INNOVA	508Y9	12673.0	7.0	2.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	150.0	2938.0	9585.0	1.1
KENNEDY	50BJ8	29036.0	12.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	29036.0	0.0	0.0
140	50WHX	14571.0	11.0	1.0	0.0	1.0	0.0	1.0	4.0	0.0	2.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14571.0	0.0
AUTOPISTA	50AJF	24930.0	9.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	17168.0	7762.0	0.0
BRITANIA	50RBT	26506.0	6.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15620.0	10886.0	0.0
CALATRAV	50EKU	9508.0	11.0	2.0	0.0	2.0	0.0	0.0	3.0	1.0	1.0	2.0	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	9508.0	0.0
CALLE 19	50GAH	10343.0	15.0	3.0	0.0	1.0	0.0	2.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	10343.0	0.0	0.0
CALLE 67	50FYX	11615.0	13.0	0.0	0.0	2.0	1.0	1.0	3.0	0.0	0.0	3.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5696.0	5919.0	0.0
LA CANDEL	50FZY	10789.0	4.0	0.0	0.0	0.0	0.0	1.0	2.0	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	0.0	8078.0	2711.0	0.0	74.8
CEDRITOS	50HOU	17538.0	5.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17538.0	0.0
CHAPINER	50BUX	19825.0	24.0	2.0	0.0	0.0	0.0	3.0	10.0	0.0	2.0	4.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	10979.0	8846.0	0.0
COSMOS	50KZS	15780.0	17.0	1.0	0.0	1.0	1.0	1.0	6.0	0.0	0.0	3.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10971.0	4809.0	0.0
10MA SAN	50KUO	5883.0	4.0	0.0	0.0	0.0	0.0	3.0	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	0.0	0.0	2484.0	3399.0	0.0
HEROES	50HVJ	10563.0	14.0	2.0	0.0	1.0	0.0	3.0	3.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5587.0	4976.0	0.0
JAVERIANA	50JEO	9394.0	10.0	2.0	0.0	0.0	0.0	2.0	2.0	0.0	1.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1245.0	8149.0	0.0
JJ VARGAS	50JJJ	30049.0	8.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19897.0	10152.0	0.0
MAGDALE	50ECG	4206.0	4.0	1.0	0.0	0.0	0.0	1.0	1.0	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	0.0	1768.0	2438.0	0.0
MODERNC	50NBF	4746.0	9.0	1.0	0.0	2.0	0.0	4.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	389.0	4357.0	0.0
ORQUIDEA	50UXY	22023.0	8.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11140.0	10883.0	0.0
PALERMO	50UYL	20050.0	14.0	3.0	0.0	1.0	0.0	1.0	4.0	0.0	1.0	2.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	541.0	19509.0	0.0

Table 11: Demographic variable description.

Variable Name	Description
PDV	Name of the Store
CR	Store ID (same as Sales by Store)
Población	Total population in a **500m radius of each store**
Rows E to X	# of direct competitors stores located in the same radius by competitor
Total Competitor	Sum of rows E to X
Rows Y to AA	Population by Socioeconomic level (low, med, high)
Rows AB to AD	Population as percentage by Socioeconomic level (low, med, high)
Rows AE to AI	Population by age groups
Rows AJ to AN	Population as percentage by age groups
Rows AO to BA	# of other business grouped by categories
Rows BB to CC	# of other business located in the same radius EX: Universities, banks, restaurants, pizza stores, bakeries, markets, pharmacies, cigarette stores.

As is indicated in the above table we make groups or categories for the different business located in the same radius of each store. The categories were determined as follows:

Table 12: Categories of other business

Categories	Stores by category
Comida de Barrio	Cafetería, Panadería
Restaurante	Restaurante
Comida Rápida	Asadero, Comida Rápida, Pizzería
Entretenimiento	Bares, Billar
Especializada	Pañalera, Papelería, Cacharrería
Droguería	Droguería
Kiosko	Kiosko
Perecederos	Fama, Frutería, Granero, Salsamentaria
Sitio de Afluencia	Bancos, Ent. Gobierno, Colegios y Universidades
Superete	Independientes, minimercados
Tienda de Barrio	Tienda de Barrio
Total Competitors	Cadena, Supercadena, Superete
Tienda de Barrio Especializada	Cigarerría, Confitería

Demographic cluster analysis:

1. We did the analysis having in consideration the following variables:

   - Level of Income
   - Competitors grouped by categories (e.g. fast food stores, supermarkets, pharmacies)
   - High traffic areas
   - Number of family members per house

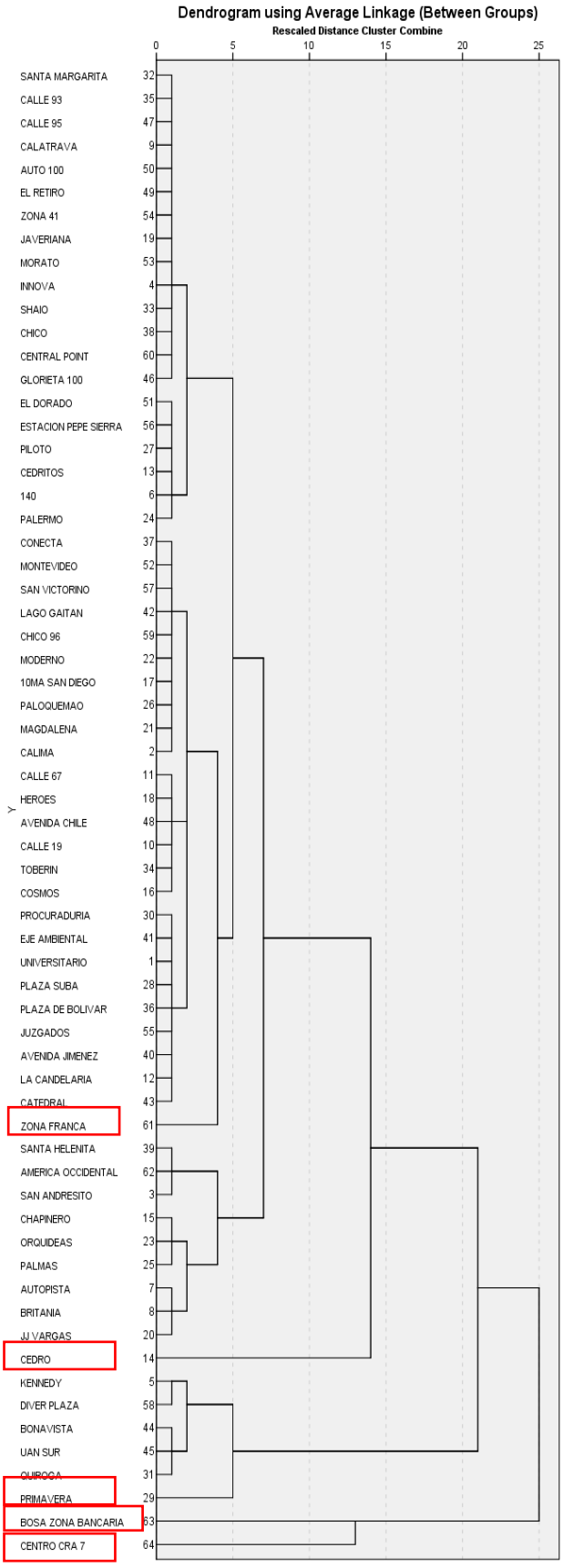
   However, we realized that the 4th variable didn't bring value to the analyses because was no cluster differentiation. We decide then, replace that variable with

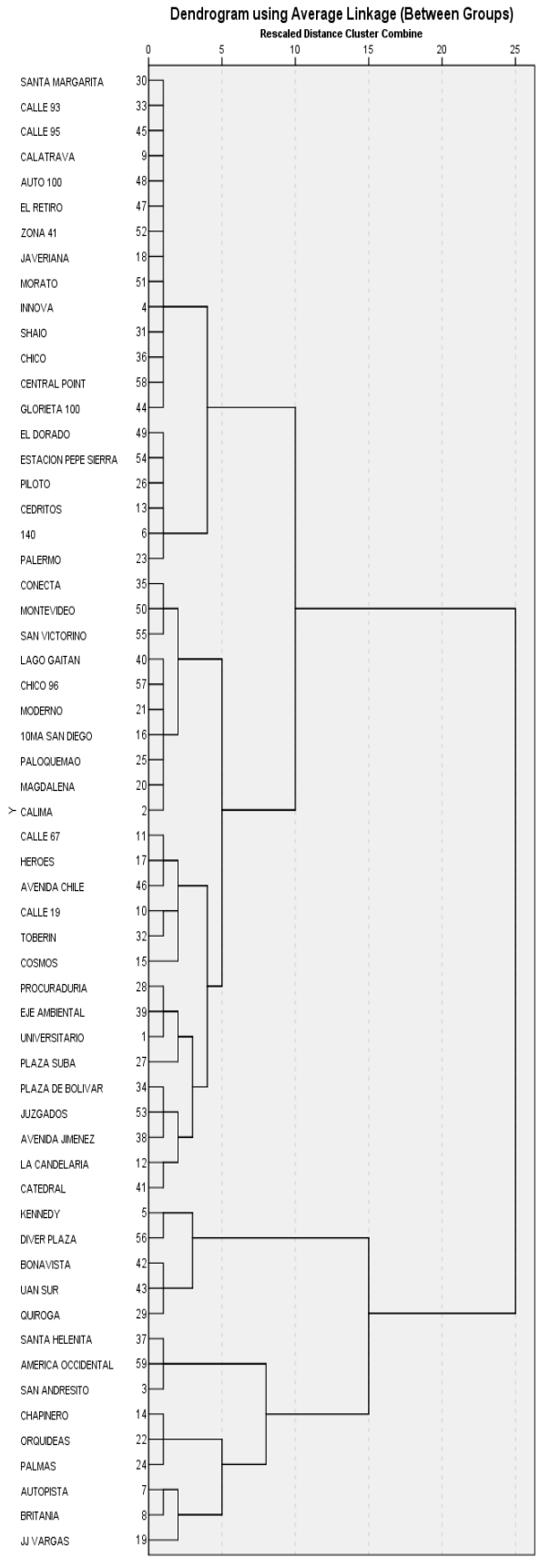
   - Age ranges

2. Taking out the outliers:

   When we run the second analysis with the new variable, we found we have 5 stores that were separate from any other major group, the stores classified as outliers where:

   - Zona Bosa Bancaria: Only have presence of one level on income
   - Cedro: Only have presence of one level on income
   - Centro Cra 7: Have a strong presence of small minimarkets
   - Zona Franca: It is a residential zone with low traffic of people. It is a store located in a place where four groups of competitors have no presence
   - Primavera: Only have presence of one level on income





3. We look for 4 different cluster analyses:

- 3 Cluster: Unbalanced (1st: 45 stores, 2nd: 9 stores, 3rd: 5 stores)
- 4 Cluster: More balanced, differentiated clusters (1st: 25 stores, 2nd: 9 stores, 3rd: 20 stores, 4th: 5 stores)
- 5 Cluster: Unbalanced, one cluster with only 3 stores (1st: 25 stores, 2nd: 3 stores, 3rd: 20 stores, 4th: 5 stores, 5th: 6 stores)
- 6 Cluster: Unbalanced, one cluster with only 3 stores (1st: 25 stores, 2nd: 3 stores, 3rd: 20 stores, 4th: 5 stores, 5th: 3 stores, 6th: 3 stores)

4. After careful consideration, we decided to go for a 4 clusters solution.

   First, we decide to make a heatmap with the clusters and the considered variables

Table 13: Heat map variables

	CLUSTER #1	CLUSTER #2	CLUSTER #3	CLUSTER #4
Average of WNSEBAJOPOBLACIÓN	15.86	0.00	0.06	1.77
Average of WNSEMEDIOPOBLACIÓN	56.46	72.19	5.16	98.23
Average of WNSEALTOPOBLACIÓN	27.67	27.81	94.78	0.00
Average of Total Competitors	9.00	11.78	9.20	14.00
Average of comida_rapida	41.56	29.44	23.65	38.60
Average of entretenimiento	36.32	30.78	15.10	39.40
Average of tienda_especializada	22.40	28.22	12.70	23.00
Average of droguerias	13.76	12.67	9.45	17.60
Average of perecederos	20.20	22.67	6.25	25.80
Average of super_ete	0.64	2.22	1.20	6.00
Average of kiosko	9.84	1.67	3.05	1.20
Average of especializada	41.52	28.78	11.10	59.40
Average of restaurante	132.84	72.22	55.15	47.20
Average of tienda_de_barrio	47.64	35.44	14.25	73.60
Average of sitio_de_afluencia	26.44	11.33	16.55	2.40
Average of WAge30orLess	11.90	11.74	9.09	12.47
Average of WAge3140	21.95	22.74	21.10	24.03
Average of WAge4150	26.63	27.59	27.50	28.22
Average of WAge5160	19.75	18.88	22.92	16.85
Average of WAge61orMore	19.74	19.05	19.39	18.44

However, because it was not easy to see a clear differentiation between the cluster when analysing the competitors, we decide to create an index to which group/s were over indexed for each cluster

Table 14: Index Creation heat map

	CLUSTER #1	CLUSTER #2	CLUSTER #3	CLUSTER #4
Average of WNSEBAJOPOBLACIÓN	15.86	0	0.06	1.77
Average of WNSEMEDIOPOBLACIÓN	56.46	72.19	5.16	98.23

Average of WNSEALTOPOBLACIÓN	27.67	27.81	94.78	0
Average of Total Competitors	0.91	1.19	0.93	1.41
Average of comida_rapida	1.28	0.91	0.73	1.19
Average of entretenimiento	1.3	1.1	0.54	1.41
Average of tienda_especializada	1.14	1.44	0.65	1.17
Average of droguerias	1.12	1.03	0.77	1.43
Average of perecederos	1.23	1.38	0.38	1.57
Average of super_ete	0.37	1.29	0.7	3.49
Average of kiosko	1.89	0.32	0.59	0.23
Average of especializada	1.31	0.91	0.35	1.88
Average of restaurante	1.56	0.85	0.65	0.55
Average of tienda_de_barrio	1.24	0.92	0.37	1.92
Average of sitio_de_afluencia	26.44	11.33	16.55	2.4
Average of WAge30orLess	11.9	11.74	9.09	12.47
Average of WAge3140	21.95	22.74	21.1	24.03
Average of WAge4150	26.63	27.59	27.5	28.22
Average of WAge5160	19.75	18.88	22.92	16.85
Average of WAge61orMore	19.74	19.05	19.39	18.44

Finally, we find the following conclusion for each cluster:

CLUSTER #1 -> OFFICE AREA – MIDDLE CLASS

- Strong presence of average-level income
- Have a high presence of informal commerce/stores
- High traffic places. More like offices and college areas
- Most people in the area (around 50%) are between 30 and 50 years old

CLUSTER #2 -> COMMERCIAL NEIGHBORHOOD – MIDDLE CLASS

- Strong presence of average-level income but without presence of low-level income
- Have high presence of stores selling house and daily basis products
- Low presence of high traffic places such as banks, colleges and universities. Residential areas
- Most people in the area (around 50%) are between 30 and 50 years old

CLUSTER #3 -> COMMERCIAL AREA - HIGH CLASS

- Strong presence of high-level income (more than 90%)
- Strong presence of medium and large retailers (direct competitors) in the area.
- Have a normal presence (not low, not high) of high traffic places. It is a mix between residential and high traffic areas.
- Low presence of young population (under 30) and more presence of people over 50

CLUSTER #4 -> RESIDENTIAL AREA - MIDDLE CLASS

- Strong presence of average-level income but without presence of high-level income
- Have the strongest presence of minimarkets and traditional stores. This cluster captures the stores with more direct competitors.
- The cluster with the lowest presence of traffic areas.
- Low presence of people over 50 and more presence of people under 30

Sales Cluster:

The objective of the sales cluster is to understand the current sales strategy of their stores, the idea is to contrast information with the current categorization of the company to see if this is having

impact on the sales of the company, more over on a second stage we will cross this clustering with the demographic one to see if the strategy of the store fits the demographic conditions.

Data Handling:

The main change that we applied to the dataset was that we calculated the store percentage of sales for each product and category, this was done to not bias the cluster with the size of the store (more size more sales) and only focus on the sales mix to cluster the store.

Table 14: Percentage sales calculation

PDV	CR	Total Sales	PORC_Cat_Antojo	PORC_Cat_Diario_Reposición	PORC_Cat_Hambre	PORC_Cat_Optimización	PORC_Cat_Reunión	PORC_Cat_Sed
CALLE 95	50554	1070672,323	23,48%	7,03%	15,10%	23,40%	11,62%	19,37%
AVENIDA JIMENEZ	500LJ	861673,2642	20,46%	7,08%	6,99%	25,10%	19,82%	20,55%
AMERICA OCCIDENTAL	501YH	585463,7499	23,52%	6,57%	9,86%	28,93%	14,14%	16,99%
CHICO 96	504CA	875637,5702	26,41%	3,74%	16,35%	24,85%	9,50%	19,14%
AVENIDA CHILE	506AW	1531691,227	20,93%	5,63%	7,93%	33,68%	15,63%	16,19%
EJE AMBIENTAL	5082N	1766784,331	16,37%	8,25%	9,50%	19,09%	32,32%	14,46%
GLORIETA 100	508PW	1231840,388	28,13%	5,86%	14,38%	23,42%	8,56%	19,64%
INNOVA	508Y9	61680,41336	25,82%	5,30%	29,00%	12,51%	7,34%	20,02%
ZONA FRANCA	509RX	1119070,494	33,22%	2,01%	17,57%	19,27%	4,22%	23,72%
EL DORADO	50A8C	1482163,616	25,96%	6,42%	14,84%	18,16%	16,37%	18,26%
AUTOPISTA	50AJF	1499144,043	19,99%	6,25%	8,22%	37,21%	13,10%	15,23%
AUTO 100	50AT6	1603729,893	25,31%	7,58%	11,31%	24,23%	11,71%	19,86%
CATEDRAL	50BFA	1192246,846	24,46%	9,02%	8,60%	23,86%	13,50%	20,56%

Sales cluster analysis:

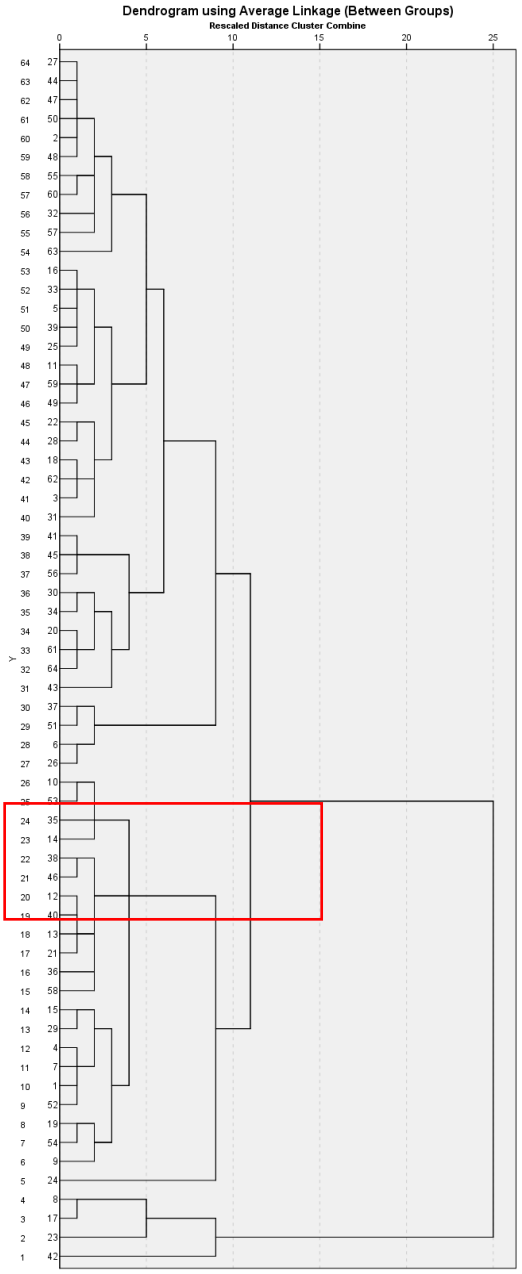
To begin the analysis, we decided to work with the broad category sales percentage calculated in the previous section:

- Porc_Cat_Antojo (craving)
- Porc_Cat_Diario_reposicion (Daily Home stuff)
- Porc_Cat_Hambre (Hunger)
- Porc_Cat_Optimization (Fun/Cigarretes)
- Porc_Cat_Reunion(Party)
- Porc_Cat_sed (thirst)

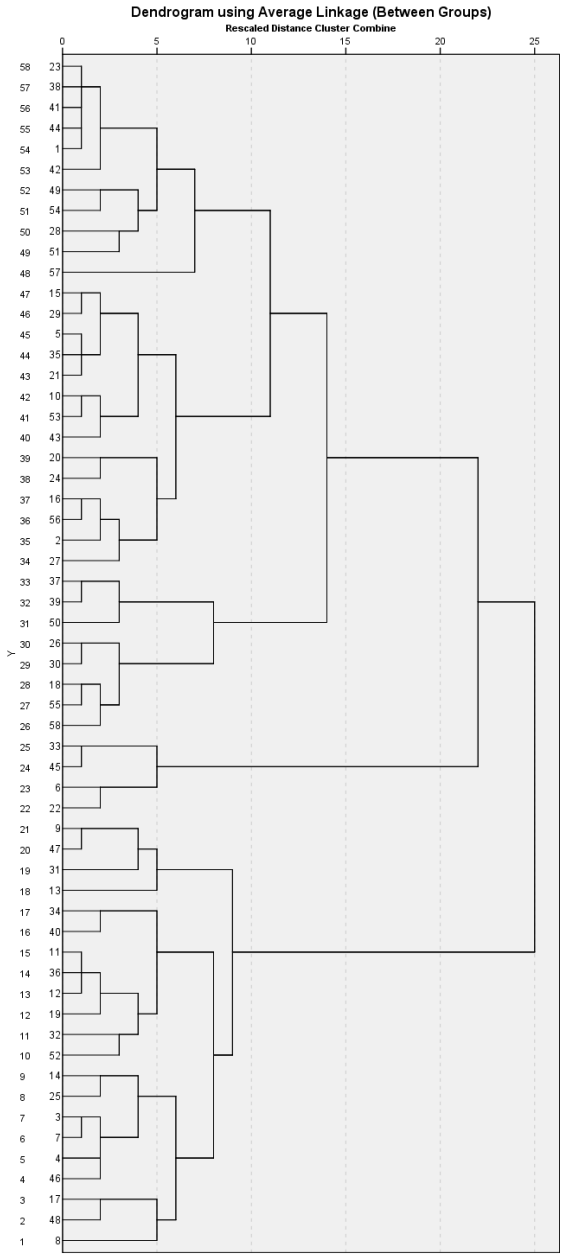
For the model after testing different combinations we decided that the ideal one was “between-group linkage” using “cosine” distance as a measure, opposed to the Euclidian distance.

Like the demographic cluster we also found out some sales outliers that were removed from the model.

With Outliers



Without Outliers



With this information we decided to test with 4, 5 and 6 clusters.

- 4 Cluster: Balanced, differentiated clusters (1st: 25 stores, 2nd: 21 stores, 3rd: 5 stores, 4th: 8 stores)
- 5 Cluster: Splits the first cluster into 2 increasing the balance (1st: 11 stores, 2nd: 14 stores, 3rd: 21 stores, 4th: 5 stores, 5th: 8 Store)
- 6 Cluster: Splits the third cluster (21 stores) into 2 clusters (17 and 4) with no significant difference.

We decided to stay with the 5-cluster analysis and analyse the heatmap:

Table 15: Sales Cluster heat map.

	1	2	3	4	5	Outlayer
Number of Clust	11	14	21	5	8	5
Craving	20,15%	19,97%	25,99%	15,67%	15,02%	22,81%
Daily Stuff home	6,58%	6,04%	6,36%	7,40%	7,13%	3,22%
Hunger	8,83%	10,61%	13,85%	8,85%	7,04%	28,73%

Party	22,03%	15,10%	12,35%	29,94%	21,73%	7,48%
Thirst	17,45%	17,05%	20,23%	14,37%	12,71%	21,25%
Fun/Cigg	24,97%	31,24%	21,23%	23,78%	36,37%	16,51%

With this information we decided to name our clusters the following way:
CLUSTER #1 -> MEETING CLUSTER:

- Average participation in most categories with an above sale in "Party" category.
- This cluster is frequented by people looking to buy stuff for a group meeting.

CLUSTER #2 -> BREAK CLUSTER:

- Close to average in Craving, Hunger and Thirst
- High Fun/Cigg purchases
- Low Daily Stuff and Party purchases
- This cluster is for people in work zones and universities that visit the store to buy something during their breaks

CLUSTER #3 -> EATING CLUSTER:

- High in Craving, Hunger and Thirst Categories
- Very low in party category
- This cluster encompasses stores dedicated to buy food related products

CLUSTER #4 -> PARTY CLUSTER:

- Top Party products consumption.
- Very low food product purchase
- As the name suggest the cluster to buy products for the party (probably excels at night purchases)

CLUSTER #5 -> CONVENIENCE CLUSTER:

- Excel at Daily stuff and Fun/cigg consumption
- Very low food related sales
- A cluster for stores designated to buy stuff on the go or while traveling

CLUSTER #6 -> OUTLIER CLUSTER:

- AKA super food cluster for the excel in those categories
- Might be an option to mix it with the food eating cluster

Recommendations by cross cluster:

SALES / m²	**COMMERCIAL AREA**	**COMMERCIAL NEIGHBORHOOD**	**OFFICE AREA**	**RESIDENTIAL AREA**	**Total general**
Break Cluster	$12.951	$15.098	$10.503	$19.009	$13.136
Convenience cluster	$22.964	$13.008	$12.266	$9.839	$13.487
Eating cluster	$15.912		$11.419	$438	$12.392
Meeting Cluster	$17.471		$16.045		$16.823
Party Cluster		$13.874	$19.767		$17.410
Total general	$16.308	$14.269	$13.439	$7.913	$14.060

When evaluating the average sale for each type of clusters, we obtain the following insights

By demographic cluster analysis:

1. We observed that the retail model has a very accepted value proposition in commercial areas with moderate traffic of people, mainly in areas where there is a "captive" shopper

who spends most of his time in a company, university or schools. In particular "Meeting" and "Convenience" excel in Commercial Areas, while "Break Clusters" seem to produce lower results.

2. It is important to see that the more residential areas have a different behavior, the retail value proposal still needs to be adapted so that the shoppers of these clusters achieve greater acceptance towards the retail.
3. Residential areas with low traffic and with a greater presence of low socio-economic levels, affect retail performance. The exception of the "Break Cluster" that we see with a high sales value belongs to only 1 Store that matches that category and therefore should be considered an outlier (further analysis should be made to understand better the success of this store in the residential area).

By sales cluster analysis:

1. The most successful stores are the ones in the Meeting and Party cluster, since their average sales/m^2 are correspondingly 19% and 23% above the total average of the stores.
2. Party-type stores excel in Office Zones, having the greatest performance out of all the combinations.
3. Break-type stores do not achieve acceptance in all areas, which leads us to think that they are the stores with the lowest tickets since they have a high dependence on the category of cigarettes, but they show interesting results in Commercial neighborhood zones.

By super categories cross cluster analysis:

Retail should focus its efforts on prices, assortment, spaces, promotions and communication by category having in consideration:

1. CRAVING: the category is going to have a better performance in high traffic areas and in Eating stores, where possibly are store with more space and therefore a wider assortment of this category.

   - Eating: Office Area
   - Eating: Commercial Area
   - Meeting: Commercial Area

2. DAILY STUFF HOME: the category is the one with the lowest weight in the client sales, so there is no area that stands over others. However, we see that its average sales are higher in Party stores, where probably considering the physical features of the store can be designed to give greater prominence to this category (eg, a larger store, more space in the store for this type of category, furniture or special displays for this category – bulk).

   - Party: Office Area
   - Party: Commercial Neighborhood
   - Meeting: Commercial Area

3. HUNGRY: the category has a better performance in areas of high traffic and Eating stores whose focus on sales is fast food and where surely is an adequate physical equipment for the operation (ovens, coffee machines, consumption area).

- Eating: Commercial Area
- Party: Office Area
- Eating: Office Area

4. FUN/CIGG: The category has a good rotation in general, but it is noteworthy mainly in convenience stores, which are characterized as stores that are over indexed in sales of Party that is a complementary category

   - Convenience: Office Area
   - Convenience: Commercial Area
   - Convenience: Commercial Neighborhood

5. PARTY: It has a good performance in areas of high traffic where offices, universities and social sites are present. In these areas the retail has more points of Party, with a space and portfolio more focused on this category

   - Party: Office Area
   - Party: Commercial Neighborhood
   - Convenience: Commercial Area

6. THIRST: It is characterized by having a shopper that travels around the area occasionally. This category is also complementary to categories of "meals" which shows a direct correlation in the sales performance of both categories.

   - Meeting: Commercial Area
   - Party: Office Area
   - Eating: Office Area