Precios CABA

Informe TP2 - Datamining

Ignacio Chiapella, Juan Knebel

Maestría en Data Mining, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

Entregado el 9 de Julio de 2019

1. Abstract

En este trabajo se incorpora el uso de reglas de asociación mediante el uso del algoritmo apriori y su implementación en R. En el desarrollo del mismo intentaremos explicar el comportamiento de la variación de los precios de un conjunto de productos de supermercado. En el primer capítulo daremos una introducción al procesamiento de los datos para luego comenzar el análisis y uso de las reglas de asociación y los resultados y conclusiones obtenidas.

Keywords: Datos, reglas, asociacion, plot, R Studio, transformación, resultado.

2. Introducción

En el presente trabajo nos disponemos mediante el mismo set de datos del trabajo practico pasado el estudio de reglas de asociación que se podrán obtener mediante el enriquecimiento del data set original y el estudio de distintas variables que nos permitirán decidir sobre si una regla es o no conviene sobre otra.
Un punto interesante es que en el primer trabajo obtuvimos ciertas conclusiones por medio del estudio de los precios de ciertos productos, y evidenciamos cierto comportamiento tanto en los precios de productos particulares como en conjuntos de ellos, desafiaremos esos resultados ahora viendo si podemos afirmar algunas de esas conclusiones mediante las reglas.

La resolución del presente trabajo se realizó utilizando como lenguaje de programación R y como reservorio de datos MongoDb. El código y los script de la base están a disposición de los docentes por si quieren revisarlo.

3. Datos

3.1 Análisis de datos

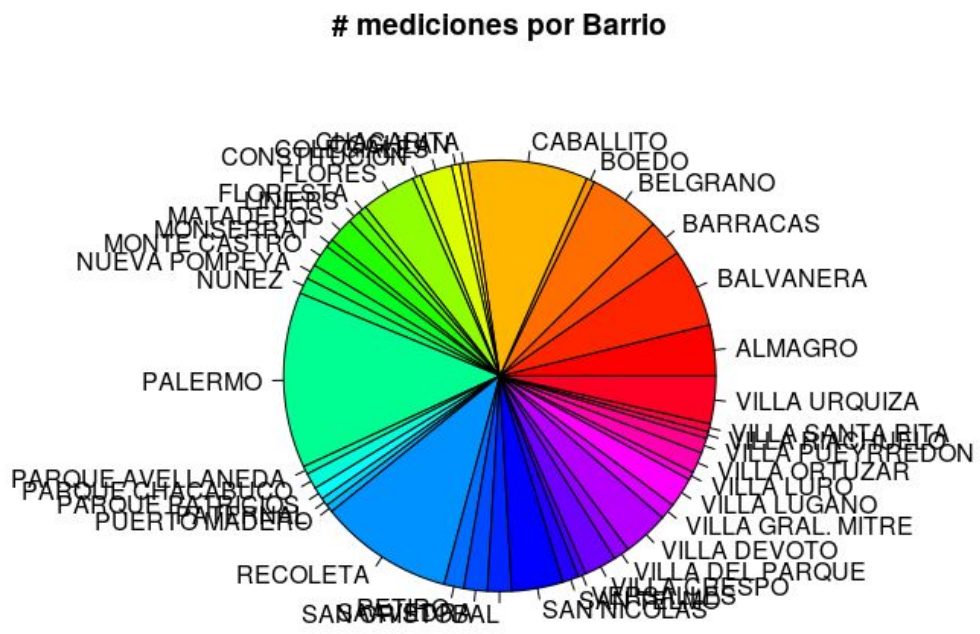
Para comenzar con el análisis de los precios en la Ciudad Autónoma de Buenos Aires se volvió a partir de la base de los datos del TP1. Adicionalmente a los datos recibidos sumamos otros conjuntos de datos. Uno de ellos contiene los polígonos de referencia para cada uno de los barrios de la Ciudad de Buenos Aires para poder identificar a qué barrio pertenece cada una de las sucursales, otro contiene una discretización de los precios y las variaciones de los mismos, separados en distintos intervalos fijos de tiempo lo que nos permite una discretización de dichos valores, y por último un conjunto de datos de presencia/ausencia.

producto	Descripción del producto
barrio	Barrio donde se ubica la sucursal que lo vende
precio_rel_per_1_disc	Muy caro | Medio caro | Levemente caro | Medio | Levemente barato | Medianamente barato | Muy barato |
precio_rel_per_2_disc	Muy caro | Medio caro | Levemente caro | Medio | Levemente barato | Medianamente barato | Muy barato |
precio_rel_per_3_disc	Muy caro | Medio caro | Levemente caro | Medio | Levemente barato | Medianamente barato | Muy barato |
precio_rel_per_4_disc	Muy caro | Medio caro | Levemente caro | Medio | Levemente barato | Medianamente barato | Muy barato |
precio_rel_medio_disc	Muy caro | Medio caro | Levemente caro | Medio | Levemente barato | Medianamente barato | Muy barato |
precio_var_per_1_disc	Disminución Fuerte | Disminución Media | Disminución Leve | Mantiene | Aumento Leve | Aumento Medio |Aumento Fuerte |
precio_var_per_2_disc	Disminución Fuerte | Disminución Media | Disminución Leve | Mantiene | Aumento Leve | Aumento Medio |Aumento Fuerte |
precio_var_per_3_disc	Disminución Fuerte | Disminución Media | Disminución Leve | Mantiene | Aumento Leve | Aumento Medio |Aumento Fuerte |
precio_var_total_disc	Disminución Fuerte | Disminución Media | Disminución Leve | Mantiene | Aumento Leve | Aumento Medio |Aumento Fuerte |
SUFIJO_elementoVocabulario_1	S | Na
...	S | Na
SUFIJO_elementoVocabulario_n	S | Na

Ejemplo de elemento de la colección de datos

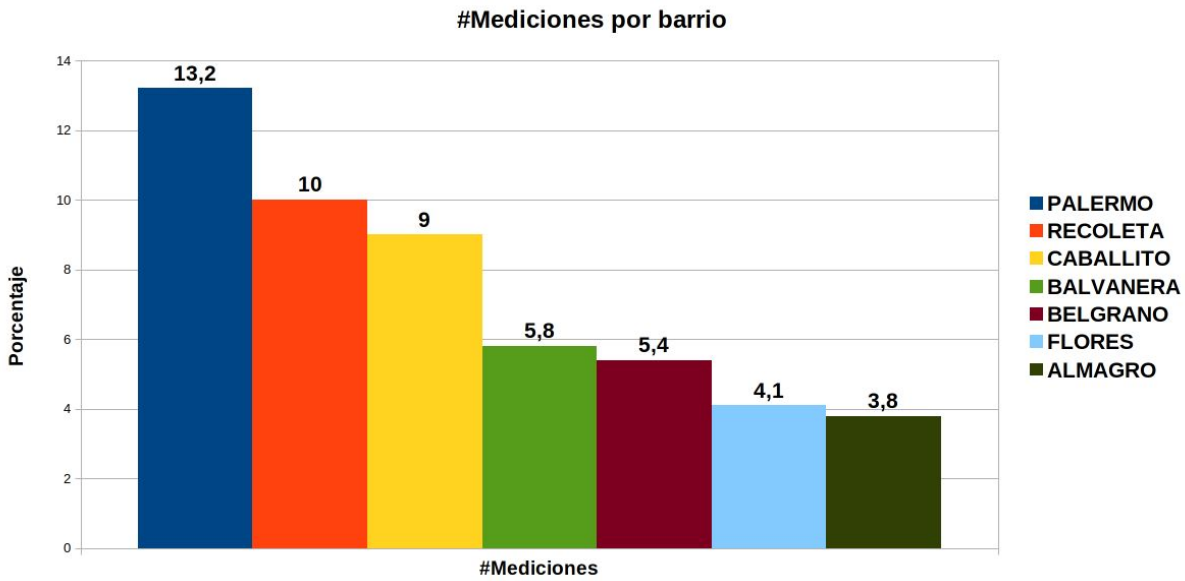

```
{
   "_id" : ObjectId("5d23676169d23974177526fe"),
   "producto" : "postre dulce  leche  pack    ",
   "barrio" : "PALERMO",
   "precio_rel_per_1_disc" : "Medianamente Barato",
   "precio_rel_per_2_disc" : "Medianamente Barato",
   "precio_rel_per_3_disc" : "Medio",
   "precio_rel_per_4_disc" : "Levemente Barato",
   "precio_rel_medio_disc" : "Levemente Barato",
   "precio_var_per_1_disc" : "Disminucion Fuerte",
   "precio_var_per_2_disc" : "Aumento Fuerte",
   "precio_var_per_3_disc" : "Disminucion Fuerte",
   "precio_var_total_disc" : "Aumento Fuerte",
   "SUFIJO_dulce" : "S",
   "SUFIJO_leche" : "S",
   "SUFIJO_pack" : "S",
   "SUFIJO_postre" : "S"
}
```


En un primer análisis antes de comenzar a buscar reglas de asociación y poder sacar conclusiones, buscamos entender cómo habían quedado repartidos los datos en los más de 160000 registros con los que contábamos.



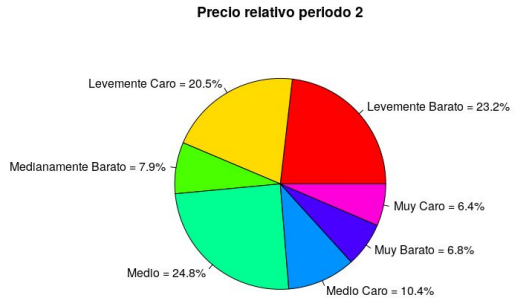
Cantidad de mediciones por barrio

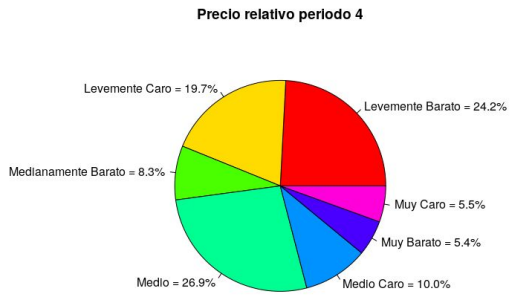
Como conclusión de este primer estudio vemos que tenemos muchas más mediciones en algunos barrios que en otros, haciendo un estudio sobre los 42 barrios estudiados, vemos que en los primeros 7 barrios se acumulan más del 50% de las mediciones totales, teniendo a Palermo como primero.



7 barrios acumulan más del 50% de las mediciones

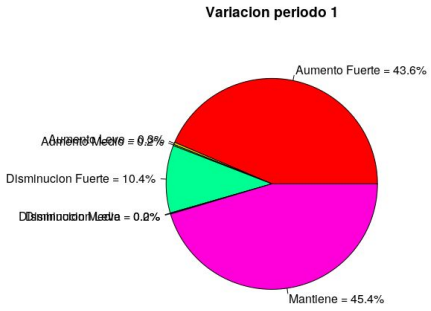
Consultamos el último informe que encontramos (2007) sobre la superficie de los barrios y la cantidad de habitantes, para ver si los datos que tenemos se relacionan de alguna manera con esas características.
Comenzamos, comparando el tamaño de los barrios y bien en el top7 Palermo seguía estando primero, aparecían segundo y tercero Villa Soldati y Villa Lugano, donde teníamos muy poca proporción de datos. Fue por ello, que intentamos estudiando la cantidad de habitantes por barrio y ahí fue donde encontramos la correlación. Segun la informacion que pudimos recolectar el top7 de barrios según sus habitantes son : Palermo, Recoleta, Caballito, Balvanera, Flores, Almagro, Belgrano, los cuales acumulan cerca del 50% de los habitantes de la ciudad. Coincidiendo con la cantidad de mediciones que obtuvimos para cada uno de ellos. Pudiendo generar una relación directa entre la cantidad de habitantes por barrio y la cantidad de mediciones en los mismos (ya que tienen más locales donde se realizan las mediciones).

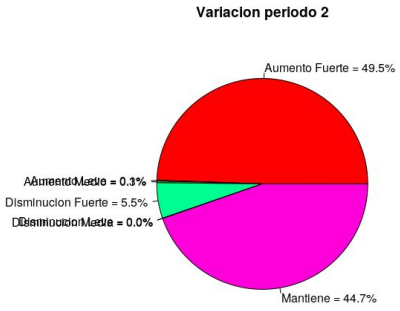


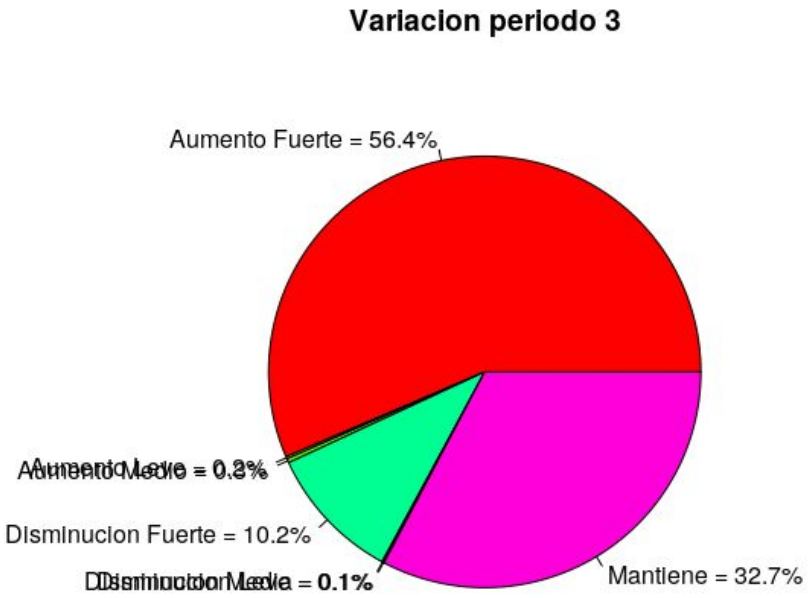


Discretización de precios relativos por periodo

Realizamos un estudio del resultado de discretizar los precios y separarlos en 4 periodos, al analizar la cantidad de mediciones discretizadas en una de las 7 categorías que se plantearon vemos que las proporciones se mantienen bastante estables en los 4 periodos y que en todos ellos hay una minoría de precios extremo (Muy Caro o Muy Barato siendo estos dos porcentajes prácticamente iguales), en todos los casos el precio Medio fue el que tuvo mayor proporción.







Discretización de variaciones en los periodos

Al estudiar las variaciones entre los periodos, lo que observamos es que los dos grupos predominantes son Aumento Fuerte y Mantiene, pero a medida que avanzamos de periodo el Mantiene disminuye y el Aumento Fuerte cobra cada vez más importancia.

4. Reglas

4.1 Análisis descriptivo

Buscamos dos conjuntos de reglas, por un lado aquellas que nos describen el aumento relativo de los precios en el último inter periodo y por otro lado las que explican el no aumento de los precios relativos en el mismo periodo. Para buscar las reglas utilizamos los siguientes valores **support=0.034** y **confidence=0.5**.

	LHS	RHS	support	confidence	lift
[1]	{precio_var_total_disc=Disminucion Fuerte}	{precio_var_per_3_disc=Disminucion Fuerte}	0.046	0.671	6.578
[2]	{precio_rel_per_1_disc=Levemente Caro,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Mantiene,precio_var_total_disc=Mantiene}	{precio_var_per_3_disc=Mantiene}	0.036	1.000	3.057
[3]	{precio_rel_per_4_disc=Levemente Barato,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Mantiene,precio_var_total_disc=Mantiene}	{precio_var_per_3_disc=Mantiene}	0.036	1.000	3.057
[4]	{precio_rel_per_1_disc=Medio,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Mantiene,precio_var_total_disc=Mantiene}	{precio_var_per_3_disc=Mantiene}	0.037	1.000	3.057

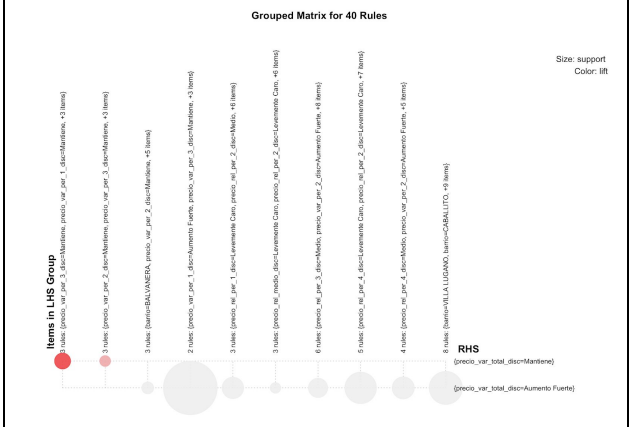
Cuatro reglas más robustas para explicar que los precios no aumentaron en el último periodo.

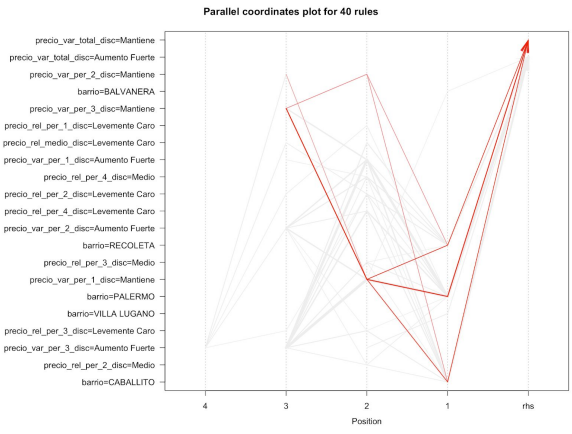
En cambio las reglas que quieren explicar el aumento de los precios, si bien obtuvieron valores de confianza cercanos a 1, su valor de lift también lo fue. Por tal motivo decidimos de no incluirlas como reglas representativas.

En la búsqueda de reglas que expliquen el comportamiento de los precios en relación a los barrios tuvimos que relajar los valores de soporte y confianza, utilizando los siguientes **support=0.01** y **confidence=0.4**.

	LHS	RHS	support	confidence	lift
[65]	{barrio=CABALLITO,precio_var_per_1_disc=Mantiene,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Mantiene}	0.011	0.862	6.215
[97]	{barrio=RECOLETA,precio_var_per_1_disc=Mantiene,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Mantiene}	0.011	0.840	6.057
[131]	{barrio=PALERMO,precio_var_per_1_disc=Mantiene,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Mantiene}	0.015	0.820	5.912
[64]	{barrio=CABALLITO,precio_var_per_2_disc=Mantiene,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Mantiene}	0.011	0.554	3.994
[96]	{barrio=RECOLETA,precio_var_per_2_disc=Mantiene,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Mantiene}	0.011	0.544	3.920
[98]	{barrio=RECOLETA,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Mantiene}	{precio_var_total_disc=Mantiene}	0.011	0.473	3.410
[191]	{barrio=BALVANERA,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Aumento Fuerte,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.010	1.000	1.268
[193]	{barrio=RECOLETA,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Mantiene,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.011	1.000	1.268
[201]	{barrio=PALERMO,precio_var_per_1_disc=Mantiene,precio_var_per_2_disc=Mantiene,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.013	1.000	1.268
[188]	{barrio=PALERMO,precio_var_per_1_disc=Aumento Fuerte,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.031	0.999	1.266
[128]	{barrio=RECOLETA,precio_var_per_1_disc=Aumento Fuerte,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Aumento Fuerte}	0.013	0.998	1.265
[200]	{barrio=PALERMO,precio_rel_per_2_disc=Medio,precio_var_per_2_disc=Aumento Fuerte,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.011	0.995	1.262
[185]	{barrio=PALERMO,precio_var_per_1_disc=Aumento Fuerte,precio_var_per_3_disc=Mantiene}	{precio_var_total_disc=Aumento Fuerte}	0.019	0.995	1.261
[197]	{barrio=PALERMO,precio_rel_per_1_disc=Levemente Caro,precio_rel_per_2_disc=Levemente Caro,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.011	0.995	1.261
[192]	{barrio=CABALLITO,precio_rel_per_2_disc=Levemente Caro,precio_rel_medio_disc=Levemente Caro,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.010	0.994	1.260
[125]	{barrio=RECOLETA,precio_rel_per_4_disc=Medio,precio_var_per_1_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.011	0.994	1.260
[198]	{barrio=PALERMO,precio_rel_per_2_disc=Levemente Caro,precio_rel_medio_disc=Levemente Caro,precio_var_per_3_disc=Aumento Fuerte}	{precio_var_total_disc=Aumento Fuerte}	0.012	0.993	1.259

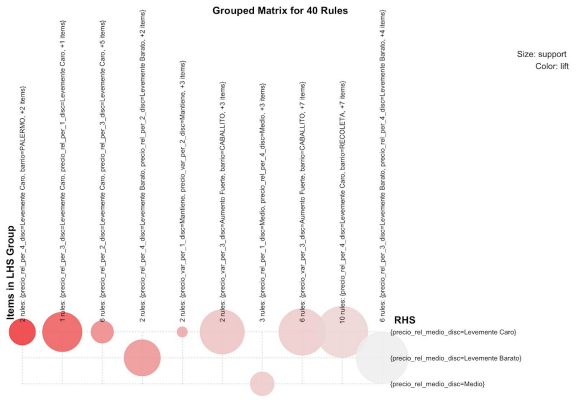
Reglas variación total de los precios por barrio.

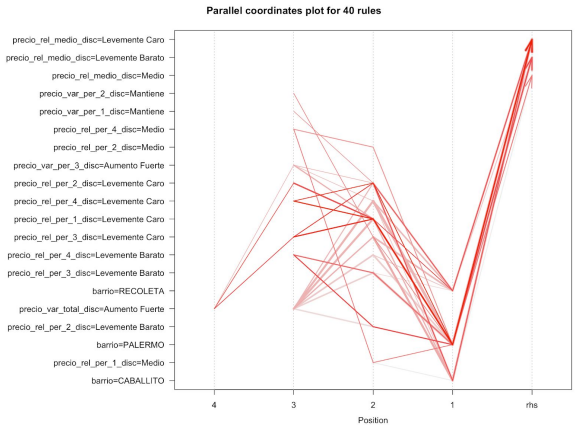




Variación de los precios entre barrios

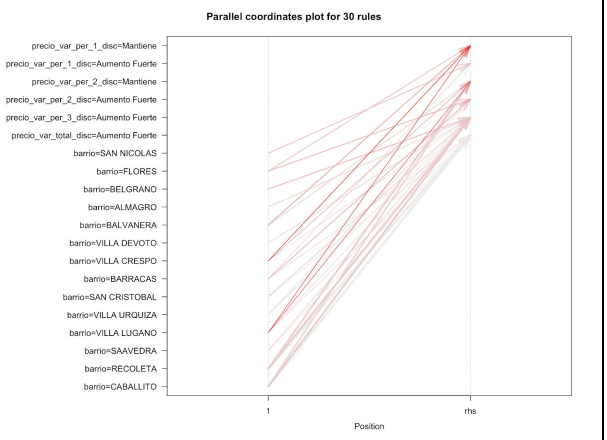
En estos gráficos podemos ver por un lado que los precios por barrio tuvieron un aumento fuerte o bien se mantuvieron, pero si bien el aumento se ve con más frecuencia, la confianza de las reglas es mayor en aquellas que explican que los precios se mantienen.
Otro dato que sale a la vista, es que los barrios que aparecen en las reglas son parte del conjunto de barrios que tenían la mayor cantidad de mediciones, lo cual guarda relación con la forma en que se están generando las reglas.





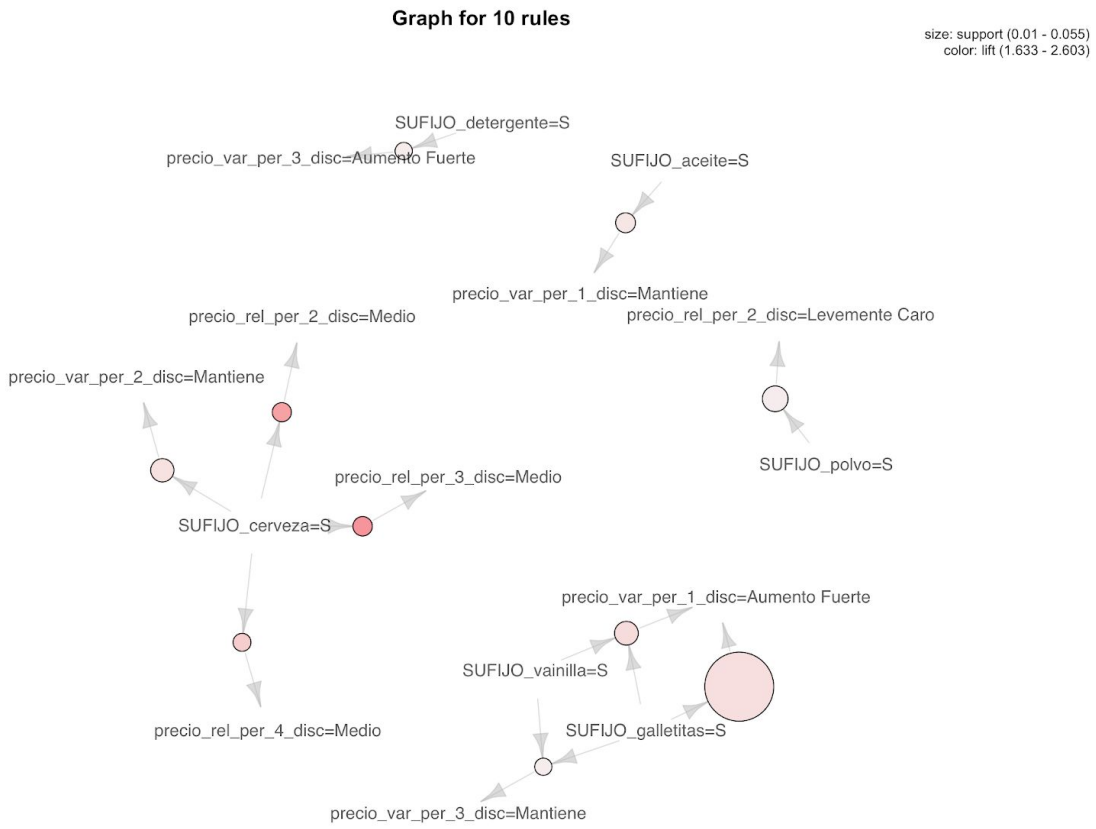
Precio relativo de los precios entre barrios

Observamos que en estos casos los precios relativos finales de la mayoría de las productos estuvo en un rango medio que va desde levemente caro a levemente barato. Lo cual nos indica que los precios se balancearon, ya que tuvieron picos de aumentos fuertes vs disminuciones fuertes.



Estudio de las reglas por barrio

En este gráfico podemos ver que la tendencia de los barrios es tener un aumento fuerte en los precios o mantenerse, en ningún caso vemos una baja importante en los precios (lo cual se condice con lo estudiado en el Tp1).



Productos estudiados que contienen el sufijo: vainilla, polvo, crema, light, galletitas, harina, aceite, turrón, cerveza, aerosol, carne, detergente.

Realizamos un estudio de algunos productos en particular, tomamos como base algunos productos estudiando en el TP1 (cerveza, aceite, galletitas) y otros que encontramos interesantes, y comprobamos mediante el grafo de la figura que los precios tienen a tener un aumento o a mantenerse, pero no a bajar como enunciamos anteriormente.

4.2 Análisis predictivo

Mantenemos los parámetros de la función a priori con los mismos valores que utilizamos en la sección anterior (support=0.01, confidence=0.4), que demostraron en nuestro análisis poder encontrar un conjunto de reglas suficientemente grande como para obtener las asociaciones que buscamos, siempre utilizando el valor del lift como medida de comparación entre dos reglas.

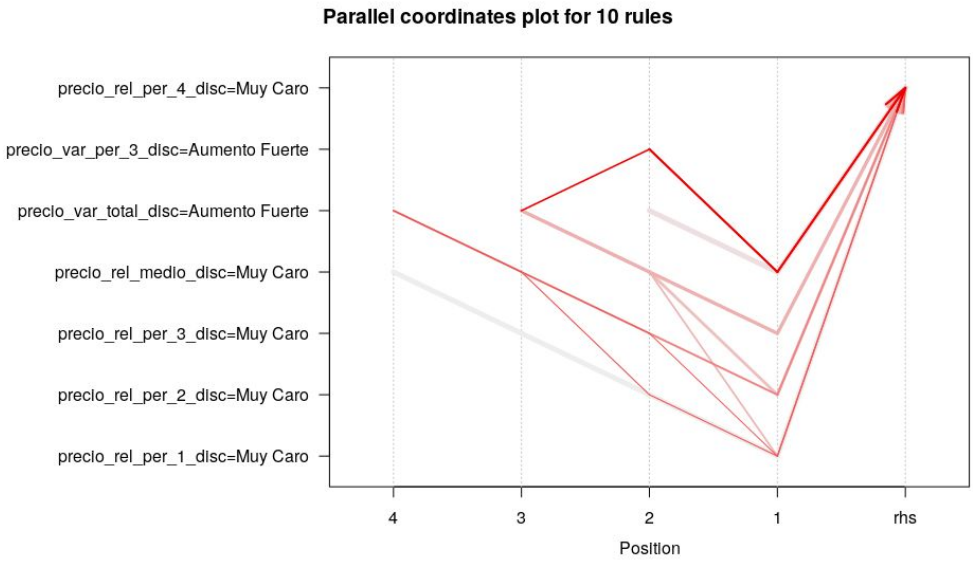
4.2.1 Estudio general de las reglas

Para generar estas 10 reglas, corrimos el algoritmo apriori pasándole el parámetro appearance los atributos identificatorios, en lhs, de los primeros 3 periodos. Al conjunto de reglas resultante, lo filtramos por las reglas que tuvieron en rhs el precio relativo del periodo 4, ordenamos los resultados por lift descendente.

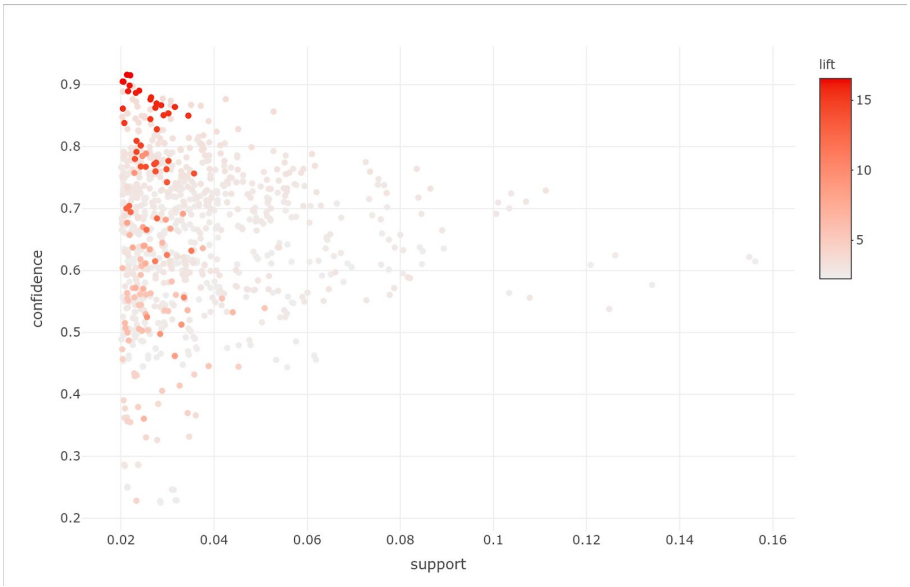
	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{precio_rel_medio_disc=Muy Caro,precio_var_per_3_disc=Aumento Fuerte,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.021	0.916	16.545	3,501.000
[2]	{precio_rel_medio_disc=Muy Caro,precio_var_per_3_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.022	0.915	16.532	3,619.000
[3]	{precio_rel_per_1_disc=Muy Caro,precio_rel_per_3_disc=Muy Caro,precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.020	0.905	16.348	3,347.000
[4]	{precio_rel_per_1_disc=Muy Caro,precio_rel_per_2_disc=Muy Caro,precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.021	0.905	16.342	3,381.000
[5]	{precio_rel_per_2_disc=Muy Caro,precio_rel_per_3_disc=Muy Caro,precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.022	0.899	16.233	3,593.000
[6]	{precio_rel_per_3_disc=Muy Caro,precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.024	0.890	16.083	3,933.000
[7]	{precio_rel_per_1_disc=Muy Caro,precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.022	0.889	16.066	3,540.000
[8]	{precio_rel_per_2_disc=Muy Caro,precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.023	0.887	16.018	3,816.000
[9]	{precio_rel_medio_disc=Muy Caro,precio_var_total_disc=Aumento Fuerte}	{precio_rel_per_4_disc=Muy Caro}	0.026	0.880	15.888	4,354.000
[10]	{precio_rel_per_1_disc=Muy Caro,precio_rel_per_2_disc=Muy Caro,precio_rel_per_3_disc=Muy Caro,precio_rel_medio_disc=Muy Caro}	{precio_rel_per_4_disc=Muy Caro}	0.026	0.876	15.825	4,328.000

Análisis sobre reglas de los primeros 3 periodos que tienen impacto sobre el 4 periodo.

Lo que obtuvimos es un conjunto de reglas, en donde evidenciamos en todos los casos que tener precios relativos en los periodos 1, 2, 3 muy caros resulta en un precio relativo del periodo 4 como muy caro tambien.



Podemos observar que las 10 reglas llegan a un precio del periodo_4= Muy Alto, indistintamente de por cual periodo anterior pasaron.



Observamos que la mayor confianza la obtenemos con un lift más grande, esto las hace robustas, que son los casos donde en RHS tenemos un precio relativo del periodo 4 muy caro.

4.2.2 Estudio de productos en particular

En esta sección, lo que vamos a buscar es poder validar con casos y conclusiones que hallamos en el TP1 en base a los estudiado hasta el momento:

- Si bien hay aumentos en algunos casos y en otros se mantienen estables, la realidad como vimos en el TP1, es que los precios no tienen a la baja.
- Al igual que nos paso en el TP1, realizar el estudio de un artículo en particular o de un pequeño grupo de artículos en particular es engañoso ya que podríamos estar mirando justo un caso particular o hasta un dato atípico, sin embargo para los productos que habíamos estudiado en el TP1 (cerveza, aceite, etc) pudimos corroborar la evolución ascendente de los precios.
- Las 10 reglas encontradas, nos dicen que tener precios relativos altos en los primeros periodos generan un precio alto en el precio relativo del último periodo, lo cual es coherente con el aumento de precios constante y en otros casos aumentó con estacionalidad que estudiamos en el TP1.

5. Conclusiones

- Una de las primeras conclusiones que decantan del estudio que estuvimos haciendo, es que si bien el estudio de reglas de asociación en un caso de laboratorio en muy pequeño resulta sencillo, con un set de datos reales y un tamaño mediano pero sin conocer de antemano los resultados, el trabajo se vuelve mucho más difícil y poder llegar a conclusiones depende en gran parte del grado de conocimiento del dominio del problema al que se haya llegado.
- En la confección de este set de datos "potenciado" tuvimos que agregarle varias columnas y datos al conjunto de precios de los productos, lo que hizo que si bien no tuviéramos más cantidad de filas si incrementamos mucho la cantidad de columnas del dataSet. Sumado a esto, varias funciones y operaciones que recorren este set de datos haciendo comparaciones, sacando o modificando muestras hicieron que el tiempo de cómputo que tuviéramos fuera considerable.
  Como estrategia, realizamos bajadas de los dataSets intermedios a la base MongoDb con la que trabajamos para poder acelerar el proceso de desarrollo y experimentación, llegando a tener el dataSet final con el que se corría el algoritmo apriori completamente guardando en la base. El tiempo de búsqueda del dataSet en la base era considerablemente menor, lo que nos permitió no tener tiempos muertos en la confección del trabajo.
- En el trabajo anterior el estudio de los barrios se limitó a conocer precios asociados a barrios, o tendencias según barrios. Pero en este trabajo tuvimos que bajar un nivel más de entendimiento, donde datos como en qué barrios hay más mediciones, como es la proporción de locales por barrios y que barrios son los que mueven la vara de los precios en general (más locales y más mediciones), nos hicieron darle una vuelta de rosca más al análisis de los datos relacionados a los barrios.

6. Referencias bibliográficas

Cotización del dólar día a día: http://www.pullman.com.ar/es/historico-dolar/
Inflación mes a mes: https://www.indec.gob.ar/
Coordenadas de los barrios de Buenos Aires: https://data.buenosaires.gob.ar/dataset/barrios
Analisis de Canasta Básica:
https://www.argentina.gob.ar/sites/default/files/cbayt-gba-mar18.pdf
Características barrios CABA link