Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

Data Mining

Trabajo Práctico N°1

Preprocesamiento de Datos, Integración
y Gestión de Datos mediante una DB NoSQL

Lucas Tomasini

3 de Junio de 2019, Comisión 1

Introducción

A lo largo del presente documento se detalla el análisis realizado sobre el sitio web de Precios Claros [1] de la Presidencia de Nación, una plataforma que se encarga de comparar los precios de los productos entre sucursales de cualquier firma. Con dicha información y herramientas de Minería de Datos se irán contestando las siguientes preguntas, que enumeramos a continuación para futura referencia:

A. ¿Qué sucursales son las más caras y cuáles las más económicas en general teniendo en cuenta la totalidad de los productos?

B. ¿Cómo es la variación de precios de los productos en general entre sucursales de la misma bandera (Vea, Jumbo, Disco, Coto, etc.)?

C. ¿Cuáles son los productos con mayores y menores variaciones de precios entre los distintos puntos de venta?

D. ¿Existe una dependencia entre la variación del precio de un producto entre las distintas sucursales y la categoría a la que corresponde (almacén, congelados, perfumería, etc)?

E. ¿Existe una variación de precios entre los diferentes barrios porteños?

F. ¿Qué productos son los que varían más y menos sus precios a lo largo del tiempo?

G. ¿Qué categorías son las que más y menos aumentan sus precios con el tiempo?

H. ¿Cómo aumenta el precio de la canasta básica a lo largo del tiempo entre las diferentes banderas y en comparación con la suba del dólar?

Fuente de Datos

El proceso de relevamiento de precios fue generado de manera automática mediante la técnica de *web crawling*. Consistió en la generación de consultas sobre la página de Precios Claros para relevar información de sucursales, productos y precios. El período para llevar a cabo un relevamiento completo lleva entre una a tres semanas debido a la gran cantidad de datos y consultas que se deben hacer sobre la aplicación. Este proceso fue repetido sucesivamente a lo largo del tiempo generando una secuencia de diez mediciones de precios desde Noviembre de 2018 hasta Febrero de este año sobre los 1000 productos más frecuentes que se venden en supermercados e hipermecados de la Ciudad Autónoma de Buenos Aires (CABA). Para integrar con el relevamiento de precios, también se extrajo información de barrios y comunas de CABA del portal de datos abiertos de la ciudad [2] y de la cotización del dólar proveniente del Banco Nación Argentina (BNA) durante el período de medición considerado [3].

Los 5 archivos de base a partir de los cuales se integrará y procesará toda la información para luego ser analizada son:

- *sucursales.json*: todas las sucursales relevadas de comercios del rubro de supermercados, hipermercados y autoservicios ubicados en CABA.
- *productos.json*: los 1000 productos de diferentes categorías con más frecuencia en el relevamientos de precios en CABA. Es importante aclarar que se agregó a este archivo el campo *categoria* donde, por simple inspección en cada uno de los productos, se determinó la categoría a la que pertenece (frescos, almacén, bebés, limpieza, etc).
- *precios.json*: mediciones de precios de lista sobre los 1000 productos seleccionados en las sucursales de tipo supermercados e hipermercados de CABA (se excluye a la categoría autoservicios).
- *barrios.geojson*: información geoespacial de los barrios de CABA que contiene para cada barrio el conjunto de puntos que lo delimitan (polígono), además de la comuna a la que pertenece, área y perímetro.
- *cotizacion_dolar_bna.csv*: cotización día a día del dólar para la compra y para la venta según el BNA durante el período de medición de precios.

Preprocesamiento e Integración y de Datos

El siguiente paso es integrar la información de los 5 archivos fuente en un único dataframe que se utilizará de base para los análisis y transformaciones de datos posteriores. Pero antes hay que adaptar ciertos campos de las tablas individuales de manera tal que sus formatos sean compatibles a la hora de realizar el *join.*

Primero convertimos el dataframe *Sucursales* a tipo *SpaciaPoints* (conjunto de puntos), donde indicamos que los campos *lng* y *lat* serán las coordenadas (longitud y latitud) de cada sucursal representada como un punto con atributos. Establecemos el mismo Sistema de Coordenadas Geográficas en los polígonos de *Barrios* y en los puntos de *Sucursales* para poder generar consultas geoespaciales entre ambos. Para eso utilizamos el paquete *sp* de R, que mediante la función *over* va a encontrar, para cada sucursal representada como un punto, qué polígono (barrio) la contiene. De esa manera ya tenemos integrada la información de sucursales con la de barrios y comunas en una sola tabla: *Sucursales_Barrios*.

En segundo lugar observamos que varios *IDs* de productos en la tabla *Precios* contienen por error ceros a su izquierda, con lo cual a la hora de hacer el join con la tabla *Productos*, no encuentran un match y los productos para ese precio quedan con valores *NA*. Por eso eliminamos todos los ceros a izquierda en el campo *producto* de la tabla *Precios* mediante una expresión regular y, a continuación, procedemos a realizar el *join*. Tenemos así la tabla de precios con la información de los productos, y con otro join agregamos la información de las sucursales y barrios.

Por último, transformamos el campo *fecha* de la tabla *Precios* (de la forma yyyy-mm-dd) y de la tabla *Dolar* (de la forma dd/mm/yyyy) a un formato común tipo *Date* de R para poder hacer el join entre ambas tablas a partir de este atributo. Como resultado final de la integración de datos nos queda un dataframe que llamaremos *Precios_Full* de 1,584,661 filas y 25 columnas que contiene toda la información de los 5 archivos base. Algunos de los campos más relevantes de esta tabla son: *producto* (id)*, sucursal* (id)*, precio, fecha, medicion, nombre* (del producto)*, marca, categoria, tipo de sucursal, bandera, razon social, barrio, comuna, cotizacion del dólar*, etc. Comprobamos que no quedaron datos faltantes: cada precio tiene su sucursal con el barrio correspondiente, su producto y la cotización del dólar para esa fecha.

Preguntas A, B, C y D

Transformación de datos y generación de nuevas variables

Para esta sección generamos una matriz numérica a partir del dataframe principal, en donde cada fila representa una sucursal y cada columna un producto determinado. Las celdas contienen el precio promedio de un producto en una dada sucursal. Se tomó el promedio de las diferentes mediciones, ya que el objetivo es comparar precios de productos entre las diferentes sucursales, eliminando el factor variación en el tiempo. La mayoría de los productos tienen 10 mediciones en cada sucursal, pero algunos pueden tener 9 u 8. De todas maneras, al promediar estamos tomando un valor representativo de ese producto en esa sucursal. La matriz tiene 175 filas (correspondientes a las 175 sucursales para las cuales tenemos mediciones) y 1000 columnas (correspondientes a los 1000 productos principales).

A continuación observamos que de las 175,000 celdas hay 7,689 que contienen valores faltantes, lo que representa un 4.4% del total. Entre otras técnicas vamos a aplicar un Análisis de Componentes Principales sobre esta matriz, razón por la cual necesitamos imputar de la mejor manera posible estos valores faltantes (PCA no funciona si la matriz contiene algún *NA*). La alternativa simplificada de imputar con ceros no sería correcta en este caso, ya que aumentaría el coeficiente de correlación entre las columnas, alterando los resultados del PCA que dependen fuertemente de este hecho. Como los precios de sucursales de una determinada bandera son similares para los diferentes productos, una primera imputación posible es asignarle a cada dato faltante la media de todos los precios correspondiente a la misma bandera (ej: si para un dado producto en una sucursal de Vea no existe medición de precio, se le imputa el promedio de los

precios de ese producto en el resto de las sucursales Vea). Luego de hacer esto vemos que aún quedan *NAs*, ya que hay casos donde no hay precio de un producto en ninguna de las sucursales de esa bandera. En ese caso, y como una segunda estrategia de imputación más global, imputamos el valor de acuerdo al promedio de los precios de ese producto dentro de todas las sucursales correspondientes a la misma Razón Social (ej.: si no hay precio para ninguna sucursal de Vea de un determinado producto se le imputa el precio promedio de ese producto en todas las sucursales de Jumbo Retail Argentina S.A., razón social que contiene a los Jumbo, Disco y Vea). Por último, eliminamos todas las columnas (productos) que hayan quedado con algún dato faltante todavía sin imputar. Como resultado nos queda una matriz sin datos faltantes de 175 filas y 821 columnas (179 productos se descartaron por contener algún *NA* en alguna sucursal). Para futura referencia llamamos a esta matriz *Suc_Prod*.

A partir de la matriz *Suc_Prod* podemos calcular el coeficiente de variación o CV (es decir el cociente entre el desvío estándar y la media) de cada producto respecto de las sucursales. Éste nos da una medida del porcentaje de variación que sufre cada producto entre las diferentes sucursales sin importar la empresa. Agregamos ese valor a la tabla *Productos* como una nueva columna. Repetimos este mismo procedimiento pero ahora calculando el CV de cada producto respecto de las sucursales de una determinada bandera. Con esto podemos obtener, por ejemplo, el porcentaje de variación de cada producto entre las diferentes sucursales de Vea, o de Carrefour Express, etc. Agregamos a la tabla *Productos* una columna por cada bandera con esta información. Son 9 columnas nuevas en total: 1 con los CVs globales entre sucursales sin importar la bandera, y otras 8 con los CVs dentro de cada bandera. Walmart, ChangoMas y Josimar quedaron excluidos de este análisis ya que no hay suficientes sucursales (sólo 1 o 2) como para computar el CV por producto.

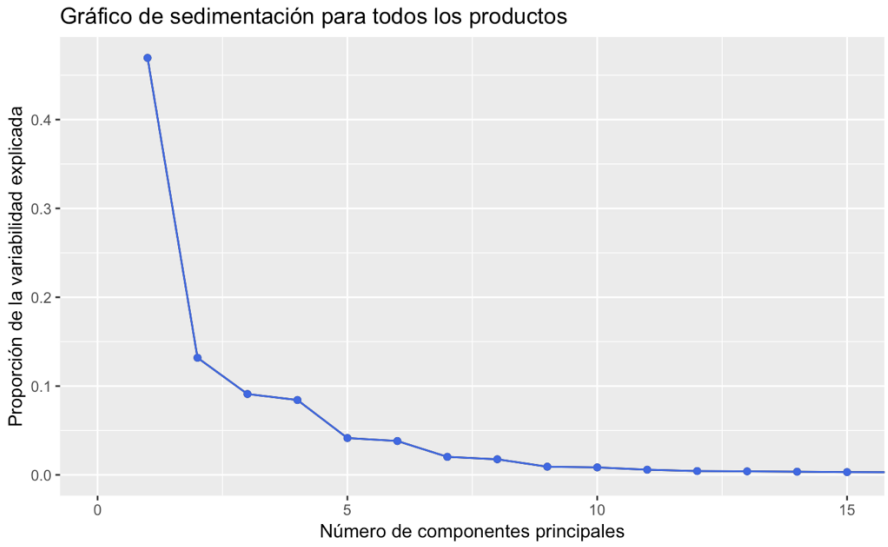
Una pregunta interesante que puede ser respondida a partir de estos datos es: ¿existe una dependencia entre la variación del precio de un producto entre las distintas sucursales y la categoría a la que corresponde (almacén, bebidas sin alcohol, frescos, etc)? Para eso clasificaremos a los productos en tres niveles de variabilidad posibles: baja (CV menor a 3%), media (CV entre 3% y 5%) y alta (CV mayor a 5%). Más adelante explicaremos por qué se seleccionaron dichos intervalos. Teniendo en cuenta esto, creamos un nuevo atributo categórico ordinal llamado *Variabilidad_Sucursales* en la tabla de *Productos*. La idea entonces es comparar el nivel de variabilidad de un producto con la categoría a la que corresponde aplicando un Análisis de Correspondencias. Construimos una tabla de contingencia de frecuencias absolutas entre ambas variables, donde las columnas representan *Variabilidad_Sucursales* y las filas *Categorías.*

Resultados preguntas A y B

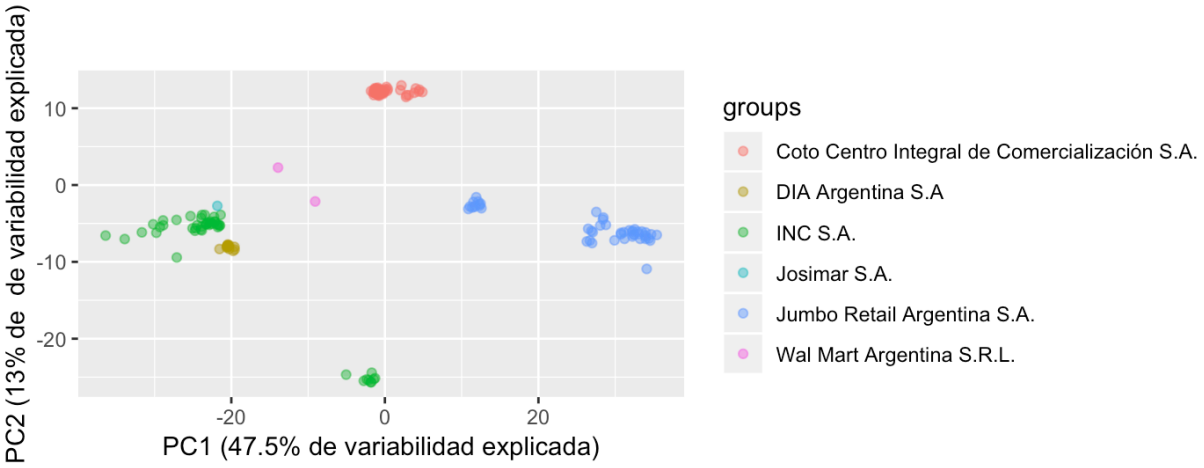
De la tabla *Productos* tomamos el promedio a lo largo de las columnas de todas las columnas que contienen el CV por bandera. Eso nos da una medida de la variabilidad promedio dentro de una determinada bandera. Ordenando los resultados de menor a mayor tenemos:

BANDERA	CV PROMEDIO (%)
DIA	0.25
COTO	0.47
CARREFOUR EXPRESS	0.79
VEA	0.81
DISCO	1.04
JUMBO	1.46
CARREFOUR MARKET	1.52
CARREFOUR HIPERMERCADO	1.83

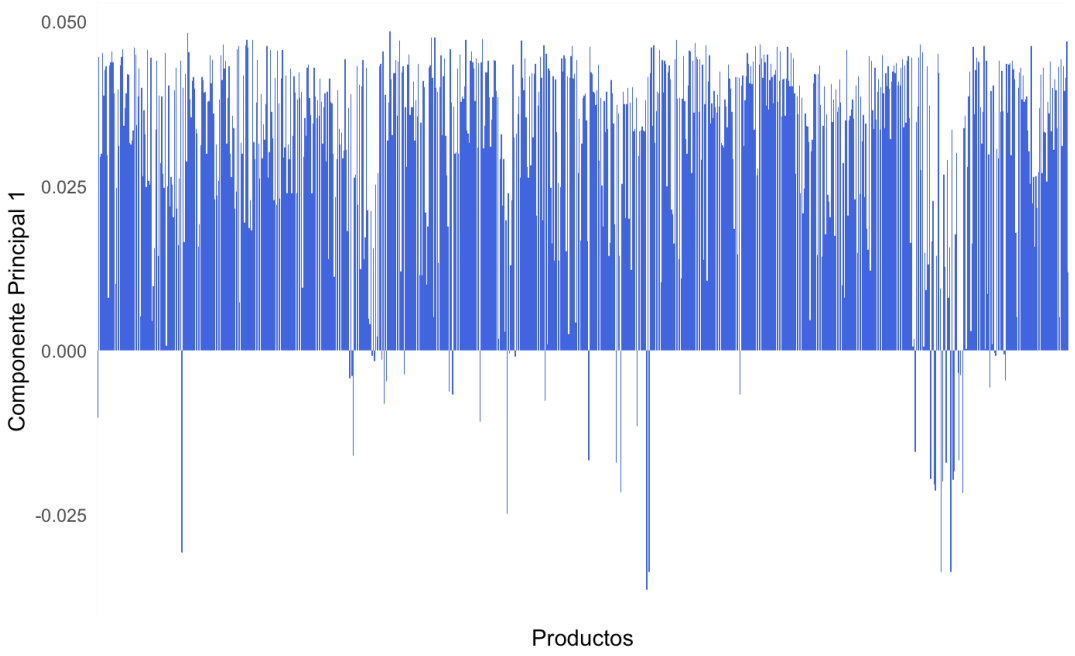
A continuación utilizamos la matriz *Suc_Prod* construida en la sección anterior para hacer un análisis de PCA sobre la misma, con el fin de visualizar cómo se agrupan las sucursales de acuerdo a los precios de los productos y qué características comparten. Recordemos que en este caso las variables son los productos y las filas sucursales, con lo cual se espera una alta correlación entre columnas y, por lo tanto, una reducción importante de la dimensionalidad si utilizamos la primera y segunda componente principal. El gráfico de sedimentación luego de aplicar PCA a la matriz con las columnas normalizadas (restando media y dividiendo por desvío estándar) resulta:



La primera componente explica el 47.5% de la variabilidad total, y la segunda el 13%. Ahora mostramos el biplot para la primera y segunda componente, proyectando las diferentes sucursales y asignándoles un color según la razón social a la que pertenecen:



La nube de puntos azules más concentrada corresponde a todos los Vea y la más dispersa a los Discos y Jumbos (mezclados). Asimismo, la nube verde concentrada son los Carrefour Express y la más dispersa los Carrefour Market e Hipermercados. La única sucursal de Josimar se encuentra pegada justo por arriba de la nube verde más grande. Ahora graficamos las cargas o *loadings* de la primera componente principales para tratar de entender qué representa la misma:



Conclusiones preguntas A y B

La primera componente principal tiene casi todas cargas positivas (con algunas excepciones que ya analizaremos), por lo cual se la considera una componente de tamaño. Es decir que una sucursal tendrá un valor alto en esta componente si los precios de sus productos en su generalidad son altos y viceversa. Recorriendo de izquierda a derecha eje X del biplot, que representa dicha componente, podemos ordenar las sucursales de las más baratas a las más caras: primero vienen los Carrefour Market, Carrefour Hipermercados y Josimar; muy pegado le siguen los Día. Luego los Wal Mart (todavía baratos pero menos que los anteriores), y después los Coto con los Carrefour Express (a ambos se los puede considerar como de precio medio). Finalmente los Vea (más caros que la media), y los Disco y Jumbos, que resultan ser las sucursales más caras.

Contrariamente a la creencia popular, los Dia no resultan ser más económicos que los Carrefour Market cuando se trata de productos que no son marca Día, lo cual ocurre con todos los productos aquí analizados (recordemos que 179 productos fueron descartaron por contener algún *NA* en alguna sucursal. Entre ellos se encuentran los productos Dia).

Del gráfico de cargas también surge que existen algunas pocas excepciones de productos que tienen un comportamiento inverso al de la mayoría. Dichos productos tienden a ser más baratos en sucursales más caras y viceversa. En la tabla siguiente mostramos los 10 productos más raros, es decir, con mayor carga negativa:

NOMBRE	MARCA	CATEGORIA
Alfajor de Chocolate Terrabusi 6 Un 300 Gr	TERRABUSI	ALMACEN
Caramelos Menta Cristal Arcor 150 Gr	ARCOR	ALMACEN
Jugo Concentrado Naranja Carioca 1.5 Lt	CARIOCA	BEBIDAS SIN ALCOHOL
Alfajor de Chocolate con Mousse Mini Jorgito 6 Un	JORGITO	ALMACEN
Alfajor Chocolate Jorgito 6 Un	JORGITO	ALMACEN
Salsa 4 Quesos Deshidratada Knorr 35 Gr	KNORR	ALMACEN
Salsa Blanca Deshidratada Knorr 20 Gr	KNORR	ALMACEN
Salsa Golf sin TACC Hellmanns 250 Gr	HELLMANN'S	ALMACEN
Papas Fritas Corte Americano Krachitos 65 Gr	KRACHITOS	ALMACEN
Almidon de Maiz Maizena 520 Gr	MAIZENA	ALMACEN

Por ejemplo, tomando el precio de la sexta medición del alfajor Terrabusi de 6 unidades en Jumbo y Disco está $82.90 (sucursales más caras), y en Carrefour Market $103 (sucursal barata). Podemos notar que curiosamente hay 3 productos de alfajores por 6 unidades que se comportan distinto a la tendencia general: conviene comprarlos en los supermercados más caros.

Por último, observando la tabla con los coeficientes de variación de las distintas banderas ordenados de menor a mayor, vemos que los Dia, Coto, Carrefour Express y Vea son las sucursales que menos variaciones de precios tienen entre ellas, lo cual es también consistente con el gráfico de biplot, donde se observa que dichas sucursales están más concentradas. A la inversa, Disco, Jumbo, Carrefour Market y Carrefour Hipermercados son las sucursales con mayores variaciones de precios, también consistente con el biplot, en donde están representadas por las nubes verdes y azules dispersas.

Resultados preguntas C y D

De la tabla *Productos* se observa que el rango de valores que toma el CV entre las diferentes sucursales va desde 0.16% hasta 18.11%. Ordenando los productos de menor a mayor coeficiente de variación, presentamos a continuación los 10 productos que menos varían y los 10 productos que más varían en porcentaje sus precios entre sucursales:

NOMBRE	MARCA	CATEGORIA	CV (%)
Crema Corporal para Piel Normal en Botella Nivea Body Milk 250 Ml	NIVEA	PERFUMERIA Y CUIDADO PERSONAL	0.1665
Ravioles de Pollo y Verdura La Salteña 500 Gr	LA SALTEÑA	FRESCOS	0.2994
Ravioles de Queso La Salteña Linea Econo 500 Gr	LA SALTEÑA	FRESCOS	0.3639
Yogur con Cereales Descremado Sancor Yogs 162 Gr	YOGS	FRESCOS	0.3645
Vino Blanco Dulce Natural Trapiche Cosecha Tardia 750 Ml	BODEGA TRAPICHE	BEBIDAS CON ALCOHOL	0.5069
Queso Crema Untable Casancrem Light 500 Gr	CASANCREM	FRESCOS	0.5261
Queso Crema Untable Casancrem Clasico 500 Gr	CASANCREM	FRESCOS	0.5274
Jugo de Soja Natural Ades 1 Lt	ADES	BEBIDAS SIN ALCOHOL	0.5294
Jardinera Arcor 300 Gr	ARCOR	ALMACEN	0.5511
Jugo en Polvo Naranja Durazno Verao 10 Gr	VERAO	BEBIDAS SIN ALCOHOL	0.5648

NOMBRE	MARCA	CATEGORIA	CV (%)
Vino Espumante Blanco Frizze Classic 750 Ml	FRIZEE	BEBIDAS CON ALCOHOL	15.66
Vino Tinto Malbec Seleccion Los Arboles 750 Cc	LOS ÁRBOLES	BEBIDAS CON ALCOHOL	16.61
Vino Tinto Cabernet Sauvignon Los Arboles 750 Ml	LOS ÁRBOLES	BEBIDAS CON ALCOHOL	16.62
Sal Fina Modificada 66% Menos de Sodio Genser 90 Gr	GENSER	ALMACEN	16.96
Vino Espumante Blue Evolution Frizze 750 Ml	FRIZEE	BEBIDAS CON ALCOHOL	17.01
Granola de Maiz y Avena Azucarada con Miel Fortificada Kelloggs Kellness Muslix 300 Gr	KELLOGGS	ALMACEN	17.02
Limpiador Horno Aerosol Mr Musculo 360 Cc	MR MÚSCULO	LIMPIEZA	17.60
Sal Fina Modificada 66% Menos de Sodio Genser 300 Gr	GENSER	ALMACEN	17.82
Talco Efficient Rexona 100 Gr	REXONA	PERFUMERIA Y CUIDADO PERSONAL	17.98
Cepillo Dental Oral B Classic 1 Un	ORAL B	PERFUMERIA Y CUIDADO PERSONAL	18.11

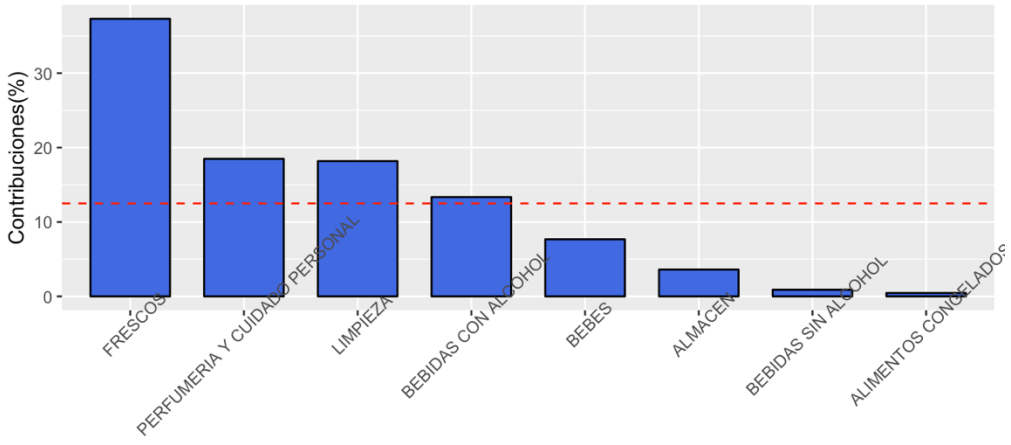
Mostramos la tabla de contingencia armada al principio, donde las columnas representan el nivel de variabilidad de precios entre sucursales y las filas las categorías de los productos (por ejemplo, hay 54 productos de la categoría *bebidas sin alcohol* que tienen una variabilidad media):

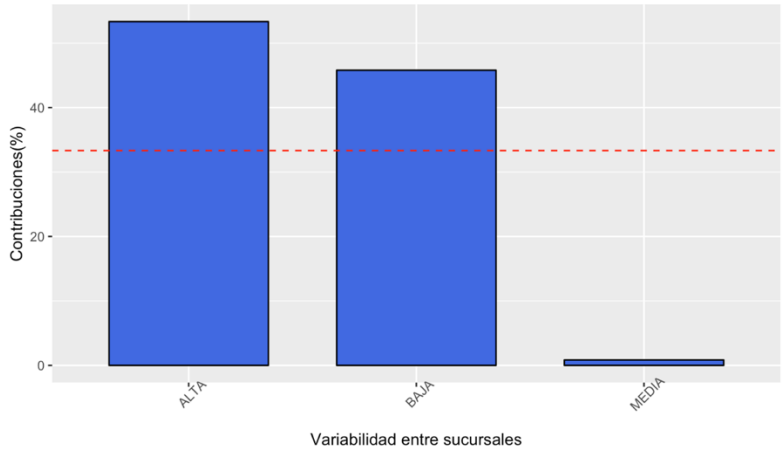
CATEGORIA/VARIABILIDAD	BAJA	MEDIA	ALTA
ALIMENTOS CONGELADOS	7	8	4
ALMACEN	148	113	130
BEBES	18	3	3
BEBIDAS CON ALCOHOL	33	17	60
BEBIDAS SIN ALCOHOL	49	54	36
FRESCOS	60	28	4
LIMPIEZA	12	27	63
PERFUMERIA Y CUIDADO PERSONAL	11	38	64

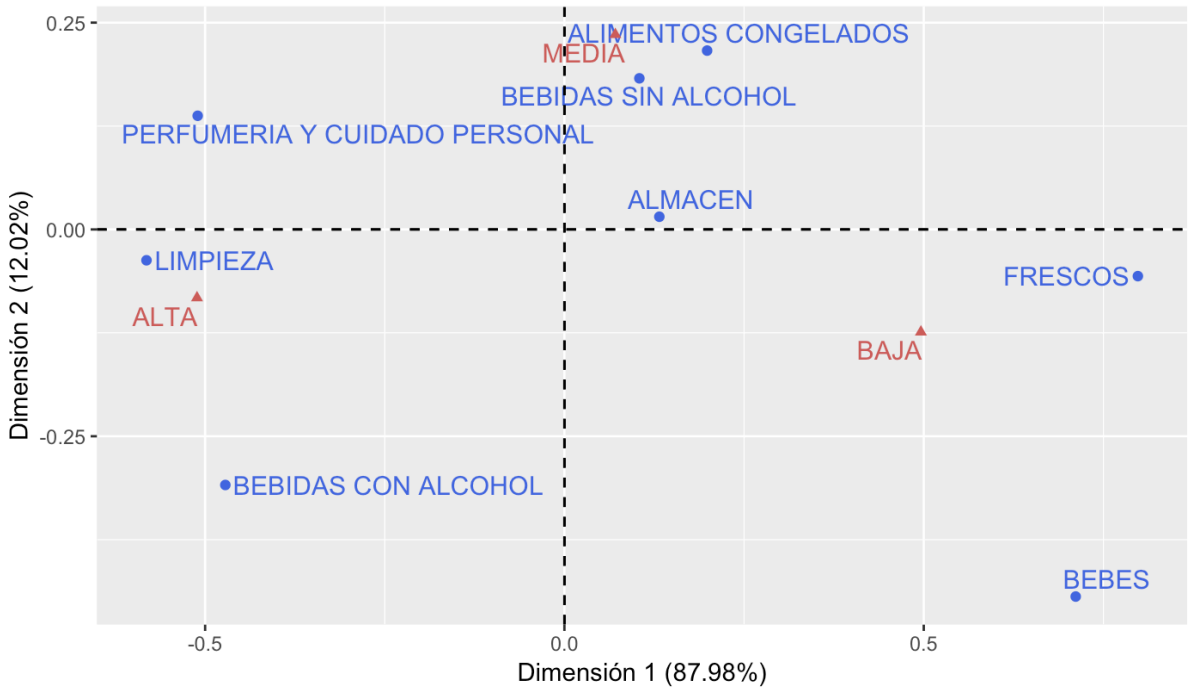
Antes de mostrar los resultados del Análisis de Correspondencias, primero aplicamos el test Chi-Cuadrado a ambas variables para determinar, con un nivel de significancia de 0.05, si hay una dependencia o no. El estadístico Chi-Cuadrado de Pearson permite cuantificar el apartamiento de la independencia de nuestras observaciones. Nótese que los productos de la categoría *mascotas* fueron excluidos de esta tabla, ya que la baja cantidad de ejemplos que tiene hace que este resultado no sea estadísticamente significativo. Para que sea válida su aplicación, es necesario que todas las frecuencias esperadas resulten superiores a 1 y a lo sumo el 20% de las mismas inferiores a 5. Por esta razón también elegimos los cortes en 3% y 5% para determinar los niveles de variabilidad, ya que con estos rangos tenemos muestras significativas para cada una de las categorías.

Los resultados del test son: Chi-cuadrado igual a 166.84, con grados de libertad 14 y un *p-value* menor a 2.2×10^{-16}. El valor para un nivel de significancia de 5% es 23.68, con lo cual se rechaza la hipótesis nula de que ambas variables son independientes. Sabiendo que hay una dependencia, procedemos ahora al Análisis de Correspondencias.

Las dos primeras figuras que siguen nos permiten visualizar las contribuciones a la inercia de las categorías y la variabilidad. En la siguiente página se muestra el biplot simétrico.







Conclusiones preguntas C y D

La dimensión 1 de este gráfico logra explicar el 87.98% de la inercia del conjunto. Vemos que los productos de limpieza son los que están más asociados a una variabilidad alta. Le siguen los productos de perfumería y cuidado personal y las bebidas con alcohol. De hecho se observa en la tabla que 7 de los 10 productos que más varían entre sucursales pertenecen a estas categorías.

En cambio, los alimentos frescos son productos de variabilidad más bien baja entre sucursales. De hecho 5 de los 10 productos que menos varían entre sucursales son de esta categoría. Los alimentos congelados, bebidas sin alcohol y productos de almacén tienen un comportamiento similar al promedio y se asocian a una variabilidad media de precios.

Pregunta E

Transformación de Datos y generación de nuevas variables

A partir de *Sucursales* generamos una tabla de frecuencias absolutas para inspeccionar cuántas sucursales con mediciones de precio hay por barrio. De ahí nos quedamos con aquellos barrios que tengan suficientes sucursales como para que el análisis sea significativo.

En base a los resultados de las preguntas A y B sabemos que en general los precios entre sucursales de una misma bandera no varían considerablemente, con lo cual para eliminar el factor variación de precios entre diferentes empresas analizamos cómo varían los mismos de un barrio a otro en sucursales de Coto, Disco y Jumbo. Tomamos Coto por un lado y Jumbo y Disco por otro, ya que sabemos que Jumbo y Disco tienen precios similares (sucursales representadas por la nube azul más grande en el gráfico de biplot de las preguntas A y B).

Para esta sección generamos otra matriz numérica a partir del dataframe principal, en donde cada fila representa una sucursal y cada columna un producto determinado. Las celdas contienen el precio promedio de las diferentes mediciones de un producto en una dada sucursal, tal cual se hizo en la primera parte. Luego filtramos únicamente aquellas sucursales que están dentro de los barrios que seleccionamos para el análisis. Esta vez no imputamos datos faltantes ni por bandera ni por Razón Social como se hizo anteriormente, de lo contrario estaríamos suavizando la

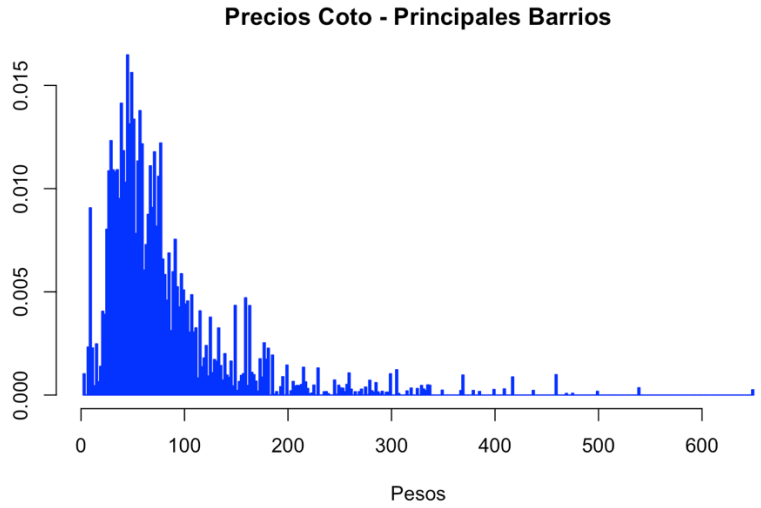
variación de precios entre sucursales de diferentes barrios, que es justamente el objetivo de esta sección. Simplemente descartamos las columnas que contienen *NA*s, ya que vamos a aplicar un análisis de PCA. Los dataframes resultantes, *Suc_Prod_Coto* y *Suc_Prod_JumboDisco*, son de 25x911 y 21x728 respectivamente.

Resultados pregunta E

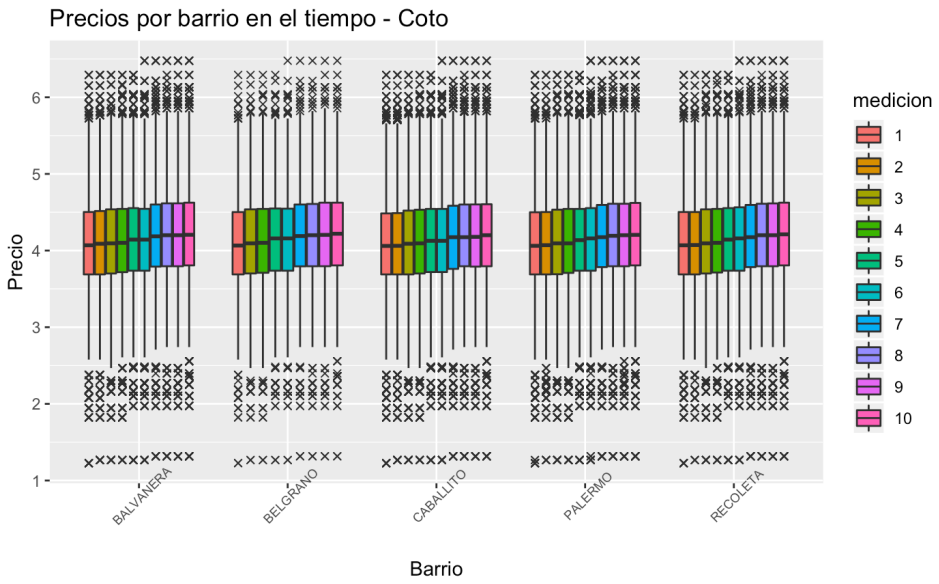
De la tabla de frecuencias de barrios tomamos los 5 con más cantidad de sucursales, que son: Palermo, Recoleta, Caballito, Balvanera y Belgrano.

AGRONOMIA	0	PATERNAL	1	VILLA DEL PARQUE	2	BARRACAS	5
BOCA	0	PUERTO MADERO	1	VILLA GRAL. MITRE	2	VILLA DEVOTO	6
PARQUE CHASS	0	VERSALLES	1	VILLA PUEYRREDON	2	VILLA URQUIZA	6
VELEZ SARSFIELD	0	VILLA LURO	1	MONTE CASTRO	3	ALMAGRO	7
VILLA SOLDATI	0	VILLA RIACHUELO	1	RETIRO	3	FLORES	7
BOEDO	1	VILLA SANTA RITA	1	SAAVEDRA	3	SAN NICOLAS	7
CHACARITA	1	LINIERS	2	SAN CRISTOBAL	3	BELGRANO	9
COGHLAN	1	NUEVA POMPEYA	2	VILLA ORTUZAR	3	BALVANERA	10
CONSTITUCION	1	NUÑEZ	2	COLEGIALES	4	CABALLITO	16
FLORESTA	1	PARQUE CHACABUCO	2	MATADEROS	4	RECOLETA	17
MONSERRAT	1	PARQUE PATRICIOS	2	VILLA CRESPO	4	PALERMO	23
PARQUE AVELLANEDA	1	SAN TELMO	2	VILLA LUGANO	4		

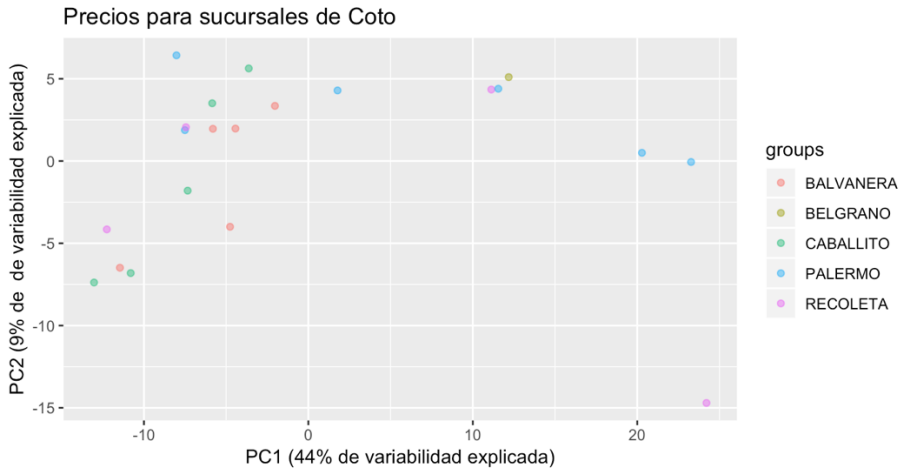
Del dataframe principal *Precios_Full* filtramos aquellas instancias que corresponden a sucursales de Coto ubicadas en estos 5 barrios y graficamos el histograma para ver la distribución estadística de los precios.



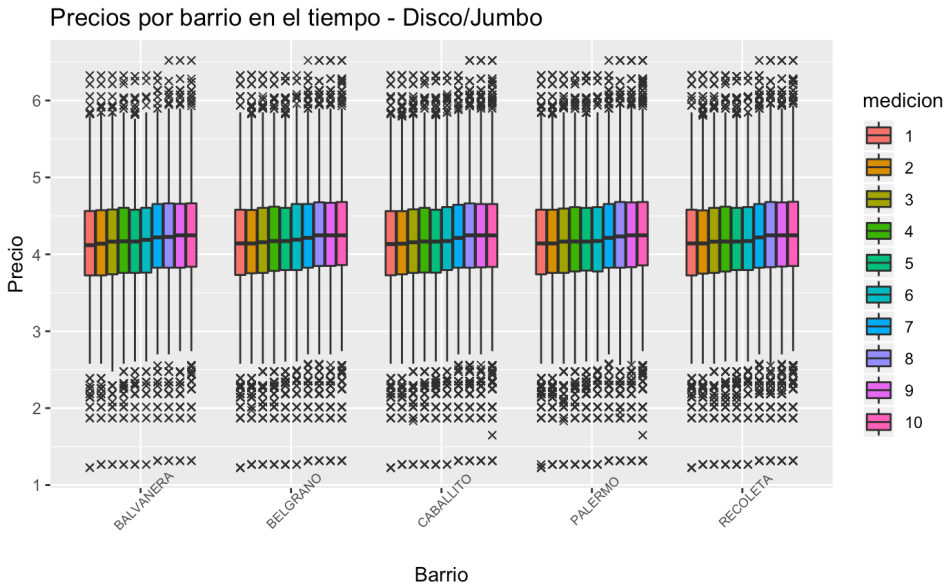
Luego mostramos los *box-plots* de los precios por cada barrio y agrupados por medición, de forma tal de apreciar la evolución de los precios en el tiempo dentro de cada barrio. Como predominan más bien los precios bajos (el histograma de arriba es sesgado a derecha), resulta conveniente para la comparación tomar el logaritmo de cada precio, que tiene el efecto de discriminar mejor los precios bajos que están amontonados y acercar más los precios altos (es decir, volver menos asimétrica la distribución):

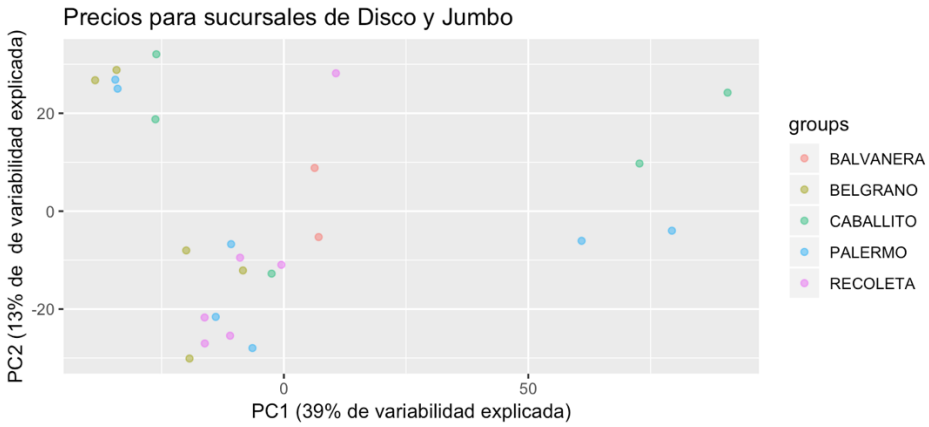


Para reforzar el análisis realizamos un PCA sobre la matriz *Suc_Prod_Coto* descrita más arriba y mostramos el biplot resultante:



Ahora mostramos los mismos gráficos para las sucursales de Jumbo y Disco:





Conclusiones pregunta E

Los grupos de *box-plots* de cada barrio son muy similares entre sí, tanto para Coto como para Jumbo y Disco. La distribución de *outliers* es prácticamente la misma en cada barrio y la mediana también. Incluso la variación de la mediana y del rango intercuartil a lo largo del tiempo es muy similar. No hay evidencia estadística significativa que indique que un barrio sea más caro o más barato que otro, dentro de una misma bandera. Esta conclusión queda confirmada con los gráficos de *biplot*, donde no se observa un agrupamiento o tendencia de los barrios sobre las componentes principales. Por ejemplo, hay sucursales de Disco y Jumbo en Palermo y Caballito con valores bajos y altos de ambas componentes principales. Lo mismo en los Coto de Caballito y Recoleta. Tal vez se puedan apreciar precios más bien bajos en las sucursales de Coto de Balvanera y Caballito y en las sucursales de Belgrano de Disco y Jumbo, pero son tendencias bastante sutiles.

Análisis a preguntas F y G

Transformación de Datos y Generación de nuevas variables

Para este análisis se generó otra matriz numérica a partir del dataframe principal, en donde cada fila representa una medición (numeradas del 1 al 10) y cada columna un producto determinado. Las celdas contienen el precio de un producto en una medición determinada promediado entre todas las sucursales. Se tomó el promedio de esa medición entre las sucursales ya que el objetivo es comparar precios de productos a lo largo del tiempo, eliminando el factor variación entre las diferentes sucursales. La matriz tiene 10 filas (una por cada medición) y 1000 columnas (correspondientes a los 1000 productos principales). En este caso no hay valores faltantes; las 10,000 celdas contienen un valor numérico representativo del precio del producto en un instante dado (que abarca un intervalo de 2 semanas aproximadamente). Llamamos a esta matriz *Med_Prod*. De nuevo, a partir de la misma podemos calcular el CV de cada producto respecto de las mediciones. Eso nos da una medida del porcentaje de variación que sufre cada producto a lo largo del tiempo. Agregamos ese valor a la tabla *Productos* como una nueva columna.

Al igual que hicimos con las categorías y la variabilidad entre sucursales, realizamos el mismo análisis para el caso de variabilidad en el tiempo. Para eso clasificamos a los productos en los mismos tres niveles de variabilidad que antes: baja (CV menor a 3%), media (CV entre 3% y 5%) y alta (CV mayor a 5%). Esto resulta en una distribución de la cantidad de productos en cada categoría que hace que la aplicación de Chi-Cuadrado sea estadísticamente significativa. Creamos un nuevo atributo categórico ordinal llamado *Variabilidad_Tiempo* en la tabla de *Productos*. La idea entonces es comparar el nivel de variabilidad de un producto con la categoría a la que

corresponde aplicando un Análisis de Correspondencias. Construimos una tabla de contingencia de frecuencias absolutas entre ambas variables, donde las columnas representan *Variabilidad_Tiempo* y las filas *Categorías*.

Para aportar más información a la respuesta de la pregunta F realizamos también un análisis de Regresión Lineal sobre el precio de los productos de una categoría dada en función de los días. Armamos una lista de R con 9 elementos (uno por categoría), donde cada elemento es un dataframe que contiene los precios promedio por día de la categoría correspondiente. El promedio es tomado entre todos los productos de esa categoría y entre todas las sucursales, en un determinado día. Aplicamos a cada conjunto una regresión lineal y comparamos qué categoría es la que más aumenta y la que menos aumenta en el tiempo. Para eso tomamos la pendiente de cada recta y la dividimos por la media, lo cual da una medida del porcentaje de incremento del precio por día para esa categoría.

Resultados preguntas F y G

De la tabla *Productos* se observa que el rango de valores que toma este coeficiente en este caso va desde 0.016% hasta 37.33%. Ordenando los productos de menor a mayor CV, presentamos a continuación los 10 productos que menos varían y los 10 productos que más varían en porcentaje sus precios en el tiempo:

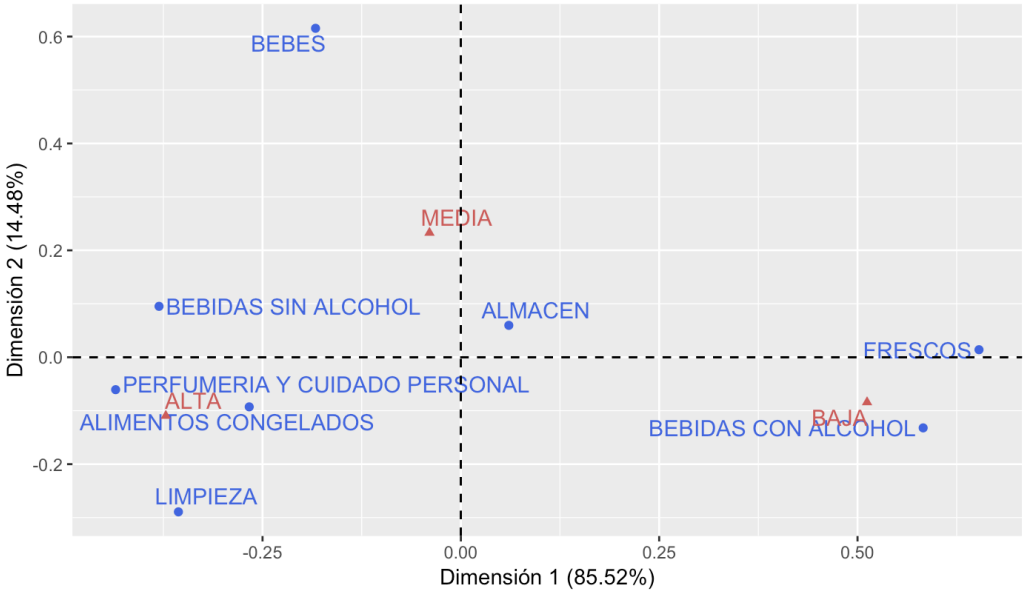
NOMBRE	MARCA	CATEGORIA	CV (%)
Ravioles de Pollo y Verdura La Salteña 500 Gr	LA SALTEÑA	FRESCOS	0.0162
Jugo de Soja Natural Ades 1 Lt	ADES	BEBIDAS SIN ALCOHOL	0.0406
Vino Tinto Cabernet Sauvignon Malbec Pinot Noir Valmont 750 Ml	VALMONT	BEBIDAS CON ALCOHOL	0.0437
Vino Espumante Extra Brut Chandon 750 Ml	CHANDON	BEBIDAS CON ALCOHOL	0.0619
Ravioles de Queso La Salteña Linea Econo 500 Gr	LA SALTEÑA	FRESCOS	0.0662
Canela Molida Dos Anclas 30 Gr	DOS ANCLAS	ALMACEN	0.1087
Flan para Preparar Vainilla Exquisita 60 Gr	EXQUISITA	ALMACEN	0.1437
Flan para Preparar Dulce de Leche Fortificado Exquisita 60 Gr	EXQUISITA	ALMACEN	0.1567
Jugo de Limon Minerva 500 Ml	MINERVA	ALMACEN	0.1706
Aceite de Girasol Natura 900 Ml	NATURA	ALMACEN	0.1710

NOMBRE	MARCA	CATEGORIA	CV (%)
Shampoo para Bebe Clasico Johnsons Baby 200 Ml	JOHNSON'S	PERFUMERIA Y CUIDADO PERSONAL	14.25
Acondicionador Blindaje Platinum Tresemme 400 Ml	TRESEMMÉ	PERFUMERIA Y CUIDADO PERSONAL	14.912
Desodorante Antitranspirante en Aerosol Rexona Invisible 150 Ml	REXONA	PERFUMERIA Y CUIDADO PERSONAL	15.52
Polvo para Mousse Chocolate Royal Light 40 Gr	ROYAL	ALMACEN	16.04
Gaseosa Pomelo Light Paso de los Toros 1.5 Lt	PASO DE LOS TOROS	BEBIDAS SIN ALCOHOL	20.26
Pañal G Active Sec Huggies 24 Un	HUGGIES	BEBES	20.53
Gaseosa Lima Limon Sprite Zero 1.5 Lt	SPRITE	BEBIDAS SIN ALCOHOL	26.28
Gaseosa Pomelo Schweppes Zero 1.5 Lt	SCHWEPPES	BEBIDAS SIN ALCOHOL	26.52
Desodorante Antitranspirante en Aerosol Rexona Antibacterial 150 Ml	REXONA	PERFUMERIA Y CUIDADO PERSONAL	35.44
Hamburguesas de Carne Vacuna Paty Express 4 Un 276 Gr	PATY	ALIMENTOS CONGELADOS	37.33

Mostramos la tabla de contingencia armada en la sección anterior, donde las columnas representan el nivel de variabilidad de precios entre sucursales y las filas las categorías de los productos:

CATEGORIA/VARIABILIDAD	BAJA	MEDIA	ALTA
ALIMENTOS CONGELADOS	4	5	10
ALMACEN	127	126	138
BEBES	3	14	7
BEBIDAS CON ALCOHOL	64	24	22
BEBIDAS SIN ALCOHOL	18	49	72
FRESCOS	54	26	12
LIMPIEZA	21	18	63
PERFUMERIA Y CUIDADO PERSONAL	15	32	66

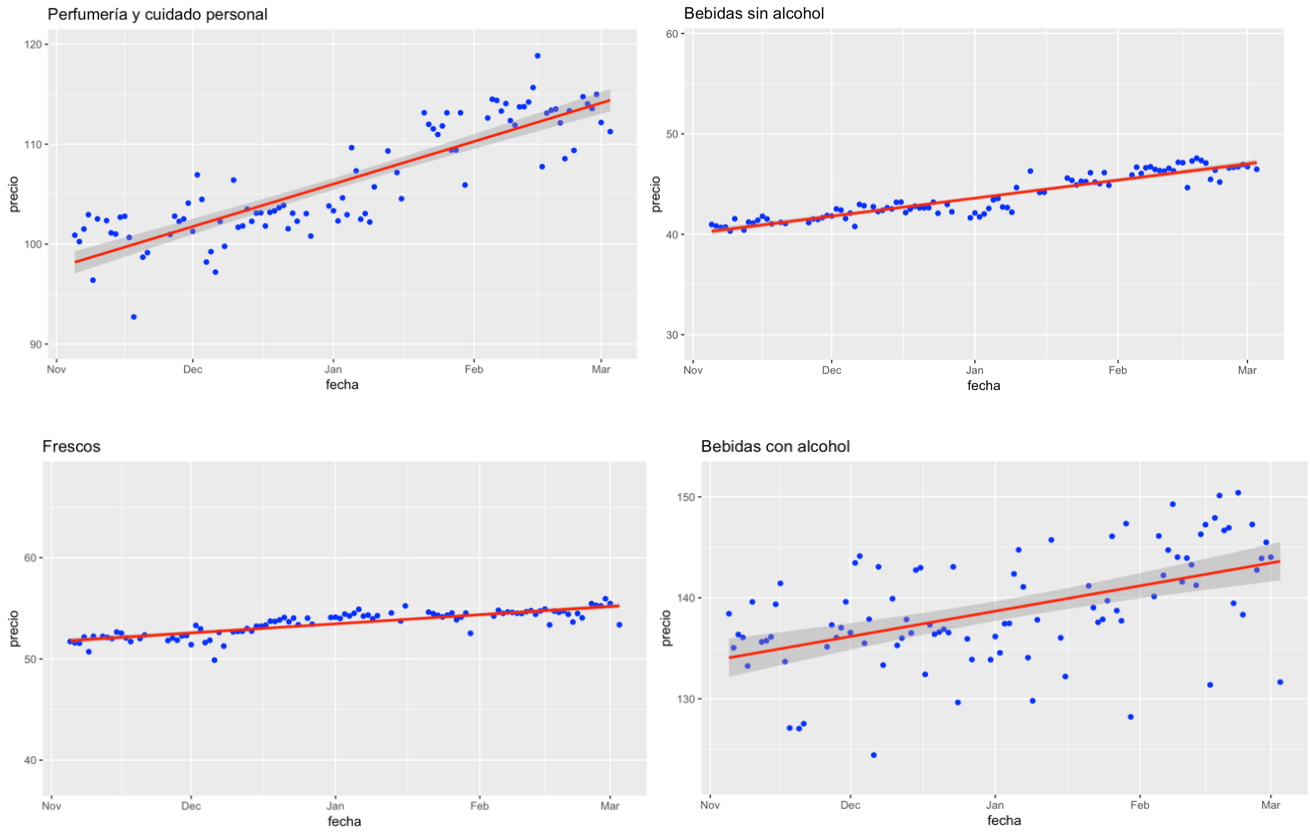
El estadístico Chi-Cuadrado es igual a 157.42, con grados de libertad 14 y un *p-value* menor a $2.2x10^{-16}$. El valor para un nivel de significancia de 5% es 23.68, con lo cual se rechaza la hipótesis nula de que ambas variables son independientes. Continuamos con el Análisis de Correspondencias mostrando el biplot simétrico:



A continuación volcamos los resultados de la regresión lineal aplicada a los precios de cada categoría en función de la fecha. El incremento está expresado en *porcentaje de suba de precio por mes* y ordenado de menor a mayor.

CATEGORIA	% INCREMENTO/MES
FRESCOS	1.63
BEBIDAS CON ALCOHOL	1.75
MASCOTAS	2.79
BEBES	2.82
ALMACEN	2.83
ALIMENTOS CONGELADOS	3.49
LIMPIEZA	3.51
PERFUMERIA Y CUIDADO PERSONAL	3.88
BEBIDAS SIN ALCOHOL	3.97

Finalmente, y a modo ilustrativo, visualizamos para las dos categorías que más y menos incrementan sus precios, los gráficos de dispersión de precios promedio por día junto con la recta que mejor ajusta dichos valores.



Conclusiones preguntas F y G

La dimensión 1 del biplot logra explicar el 85.52% de la inercia del conjunto. Vemos que los productos de limpieza, perfumería y cuidado personal, bebidas sin alcohol y alimentos congelados son los que están más asociados a una variabilidad alta en el tiempo. De hecho se observa en la tabla que 8 de los 10 productos que más varían en el tiempo pertenecen a estas categorías. En cambio, las bebidas con alcohol y los alimentos frescos son productos de variabilidad baja en el tiempo. Los productos de almacén tienen un comportamiento similar al promedio, con una variabilidad media. Notar que, a diferencia del análisis anterior, las bebidas con alcohol tienen variabilidad alta entre sucursales, pero baja en el tiempo. Algo parecido ocurre con los productos para bebés, cuya variabilidad entre sucursales es baja y en el tiempo está entre media y alta.

Las conclusiones anteriores son consistentes con los resultados que arroja el análisis de regresión lineal. Las cuatro categorías con mayor porcentaje de incremento de precio mensual son alimentos congelados, limpieza, perfumería y bebidas sin alcohol (en orden de menor a mayor). Son las mismas cuatro categorías asociadas a una variabilidad alta en el biplot. Lo mismo pasa con las dos más bajas: frescos y bebidas con alcohol. Es importante aclarar que una mayor pendiente en las regresiones lineales no implica mayor variabilidad porcentual en el tiempo ya que, por ejemplo, un producto de $10 que aumenta $1 peso por mes tiene mayor incremento porcentual que un producto de $100 que aumenta $5 por mes. Por se debe normalizar dividiendo por la media (como en nuestro caso) o la ordenada al origen.

Pregunta H

Transformación de Datos y Generación de nuevas variables

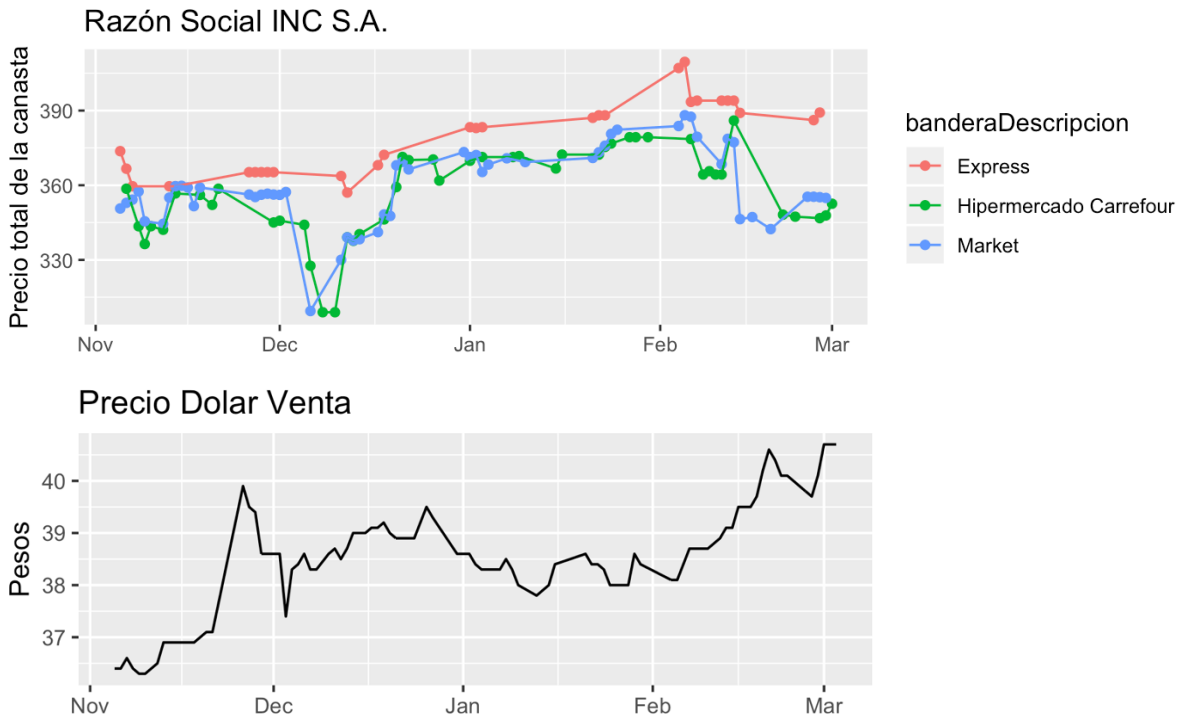
En esta sección realizamos un análisis de variación a lo largo del tiempo no para todos los productos en general sino aplicado a un grupo de 7 productos de la canasta básica que son muy comunes y se encuentran presentes en todos los puntos de venta. Dichos productos son:

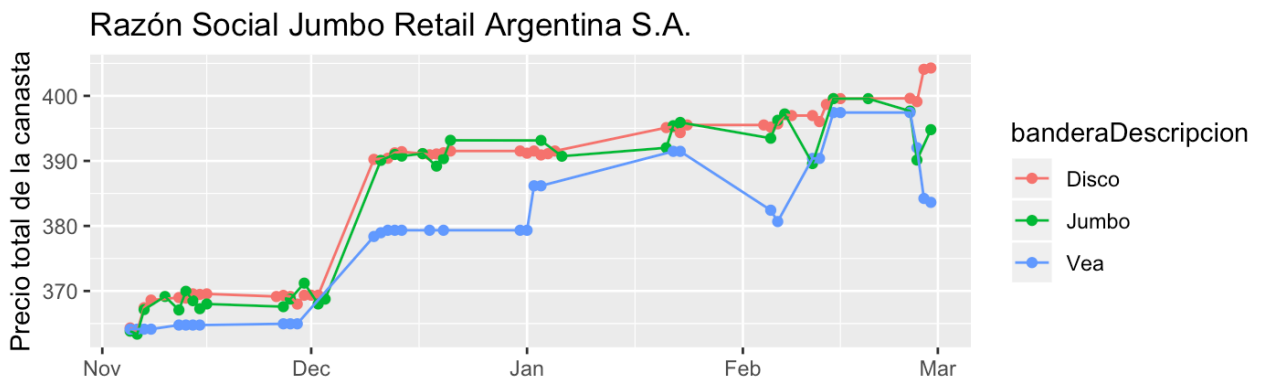
- *Galletitas Sandwich Traviata Pack 3 Un 303 Gr*
- *Mermelada de Durazno Diet BC La Campagnola 390 Gr*
- *Leche Parcialmente Descremada en Sachet Armonia 1 Lt*
- *Agua Mineral sin Gas Villavicencio 1.5 Lt*
- *Aceite de Girasol Cocinero 1.5 Lt*
- *Arroz Largo Fino en Bolsa Lucchetti 1 Kg*
- *Galletitas Surtido Seleccionadas Bagley 400 Gr*

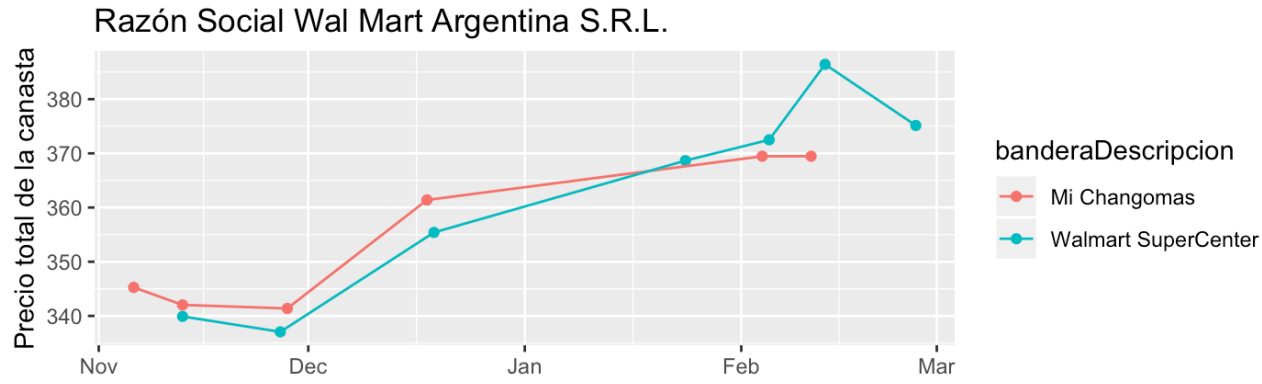
Partiendo del dataframe principal *Precios_Full*, filtramos únicamente los productos listados arriba, los agrupamos por *fecha*, *producto* y *bandera* y calculamos el promedio dentro de cada agrupación. Nos queda así una segunda tabla que contiene el precio promedio por día de cada producto de la canasta básica dentro de una determinada bandera (es decir, promediado también en todas las sucursales de esa bandera). Volvemos entonces agrupar por *fecha y por bandera*, tomando ahora la suma de los precios dentro de cada grupo. La tabla final contiene el precio total de la canasta básica por día para una determinada bandera. Por ejemplo, una fila puede indicar que el precio promedio de la canasta básica el día 21/12/2018 en Jumbo es de $393.

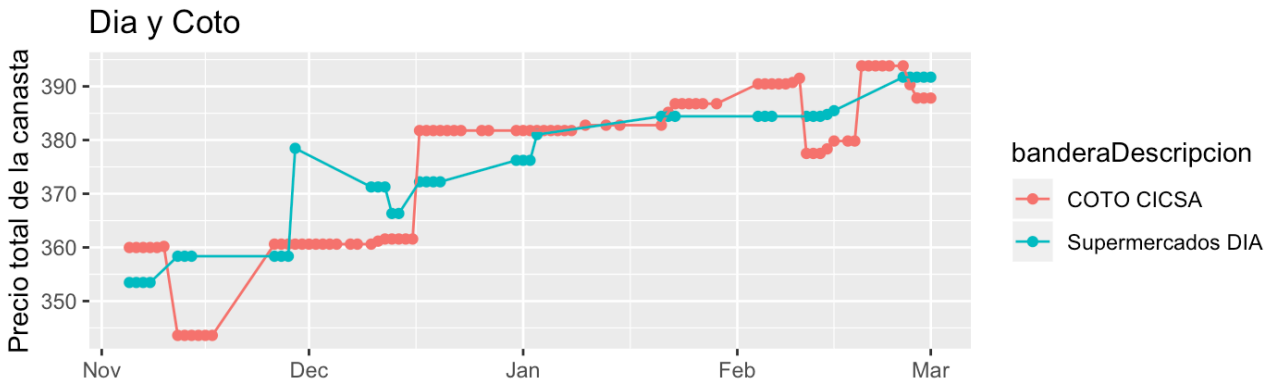
Resultados pregunta H

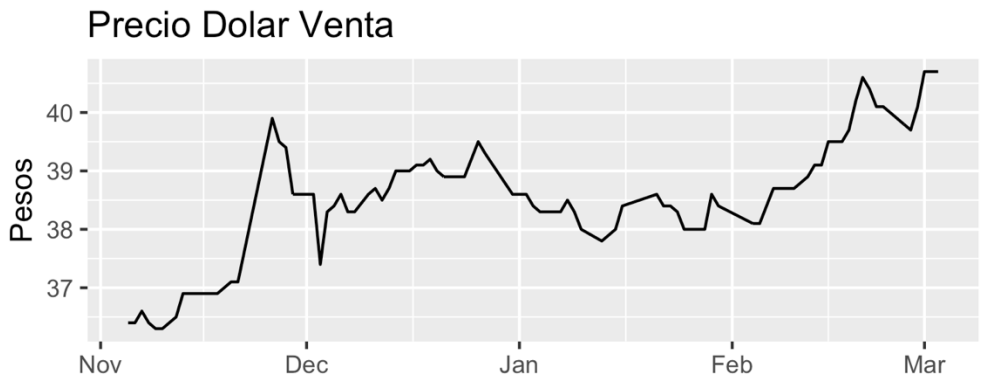
En base a la tabla construida, graficamos el precio de la canasta en función de la fecha, utilizando un gráfico por razón social y separando dentro de cada gráfico por bandera. La única excepción fue Día y Coto que, aunque sean razones sociales diferentes, las juntamos en un solo gráfico por conveniencia. Agregamos al final la cotización del dólar para la venta en función del tiempo, a efectos de comprar las diferentes estrategias de las empresas ante la devaluación del peso.




Razón Social Jumbo Retail Argentina S.A.
Precio total de la canasta
400
390
380
370
Nov
Dec
Jan
Feb
Mar
banderaDescripcion
Disco
Jumbo
Vea


Razón Social Wal Mart Argentina S.R.L.
Precio total de la canasta
380
370
360
350
340
Nov
Dec
Jan
Feb
Mar
banderaDescripcion
Mi Changomas
Walmart SuperCenter


Dia y Coto
Precio total de la canasta
390
380
370
360
350
Nov
Dec
Jan
Feb
Mar
banderaDescripcion
COTO CICSA
Supermercados DIA


Precio Dolar Venta
Pesos
40
39
38
37
Nov
Dec
Jan
Feb
Mar

Conclusiones pregunta H

Observamos primero que hubo una primera devaluación importante del peso a fines de Noviembre donde, una vez pasado el pico máximo, el precio del dólar quedó oscilante entre $38 y $39 durante el mes de Diciembre y Enero. El incremento fue aproximadamente del 4.3% respecto de Noviembre. También hubo una segunda gran devaluación que arranca en Febrero y se estabiliza temporalmente en Marzo, con un salto del 4.1%.

Del gráfico de Jumbo Retail S.A. se observa que los precios de la canasta de Disco y Jumbo van a la par a lo largo del tiempo, estando siempre por arriba de los de Vea. Antes de la primera devaluación Vea estaba apenas por debajo de Jumbo y Disco. Una o dos semanas después todas subieron sus precios: Jumbo y Disco aumentaron un 6.5% y Vea un 4.1%, para luego aumentar nuevamente a principios de Enero, quedándose otra vez ligeramente debajo del resto. Jumbo y Disco aumentan el valor de la canasta más abruptamente para recuperar la pérdida por desfasaje temporal, mientras que Vea aumenta más paulatinamente, con mayor cantidad de saltos pero más leves cada uno, alcanzando eventualmente la relación original entre precios de la canasta de Jumbo y Disco.

Aunque no hay suficientes puntos de medición de Wal Mart S.R.L., se ve que Mi Changomas y Waltmart Super Center van a la par y aumentan paulatinamente tras la devaluación para recuperar ganancias. Es interesante notar que los precios de la canasta de Coto y Dia se siguen de cerca a lo largo del tiempo, donde por momentos Coto está por encima de Dia, y por momentos sucede al revés. También se observa que Coto dio un salto abrupto del 5.5% a mitades de Diciembre mientras que Dia lo hizo antes, aunque con un salto menor, para luego ir aumentando paulatinamente.

Por último, se observa que los precios de la canasta de Carrefour Hipermercados y de Market se siguen de cerca y están siempre por debajo de los Carrefour Express. Los Carrefour Express sube paulatinamente el precio, sin grandes picos, mientras que los otros dos tienen un comportamiento más impredecible, donde se ven picos negativos esporádicos, aunque la tendencia va en aumento también.

Referencias

1. www.preciosclaros.gob.ar
2. data.buenosaires.gob.ar
3. https://www.cronista.com/MercadosOnline/moneda.html?id=ARS