**Informe de Análisis y Visualización**

**Grupo 1:**

* **Dario Yvanoff**
* **Federico Sepulveda**
* **Gonzalo Torres**

# General

Se propone la **elaboración de un informe** que deberá entregarse en formato html o pdf y deberá apuntar a un público técnico, pero sin conocimientos específicos en el tema.  
La estructura del mismo puede seguir un esquema de secciones como el siguiente: **Introducción - Presentación del Dataset - Variables más relevantes o de interés - Respuestas a preguntas disparadoras - Relaciones interesantes entre variables - Principales conclusiones y posibilidad de futuros aportes**. No hay límites respecto a la extensión, pero se valorará el poder de síntesis y de comunicación, así como el uso de las herramientas pertinentes para comunicar. Pueden tomar, a modo de ejemplo, [**esta publicación de Open Data**](http://blog.opendatacordoba.org/estudio-sobre-el-aumento-de-precios-en-las-gondolas-entre-abril-y-mayo-del-2018/). *(Notar que el objetivo de este análisis es diferente al propuesto en este proyecto)*

# Introducción

Durante episodios inflacionarios los agentes económicos se ven afectados por la incertidumbre vinculada a los cambios frecuentes de precios relativos de los diferentes bienes y servicios. Estos cambios, dadas las estructuras productivas, pueden no ser transmitidos a lo largo de toda la cadena e una manera homogénea, por lo que un mismo producto podría encarecerse mas en una región que en otra.

El objetivo del presente informe es brindar una primera aproximación al fenómeno de variaciones de precios en contextos inflacionarios, intentando buscar si existe, a pesar del contexto que enfrenta a constantes cambios, alguna forma de agrupamiento de los productos que sea sistemáticamente más barata que las restantes (por localidad, por cadena de supermercado, por región).

# Presentación del Dataset

El dataset a utilizar para la realización del análisis consiste en extracciones periódicas semanales de precios de la página PreciosClaros obtenidos a partir de un scrapper (open source) disponibilizado por Open Data Córdoba (https://www.opendatacordoba.org/) y los csv fueron puestos a disposición en Kaggle (<https://www.kaggle.com/tinnqn/precios-claros-precios-de-argentina>).

Al momento de la elaboración de este proyecto ya hay cinco extracciones generadas, para las siguientes fechas: 12/04/20, 19/04/20, 26/04/20, 05/05/2020 y 18/05/2020.

# Variables de interés para el análisis

Las variables que dispone el dataset original son las siguientes:

|  |
| --- |
| precio |
| producto\_id |
| sucursal\_id |
| fecha |
| marca |
| nombre |
| presentacion |
| categoria1 |
| categoria2 |
| categoria3 |
| comercioId |
| banderaId |
| banderaDescripcion |
| comercioRazonSocial |
| provincia |
| localidad |
| direccion |
| lat |
| lng |
| sucursalNombre |
| sucursalTipo |
| nom\_provincia |
| region |
| cantidad |
| unidad |
| nombre\_sin\_um |

Del conjunto anterior se definió trabajar con las siguientes variables:

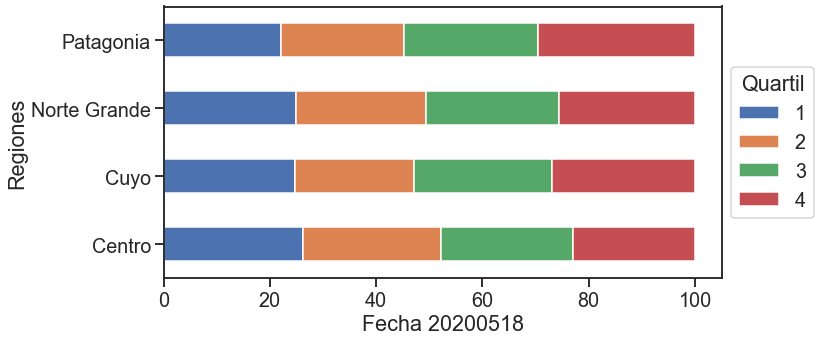
* Variables cuantitativas: Precio, Fecha
* Variables cualitativas: producto\_id, nombre\_sin\_um, marca, localidad, provincia,región, comercioRazonSocial, presentacion

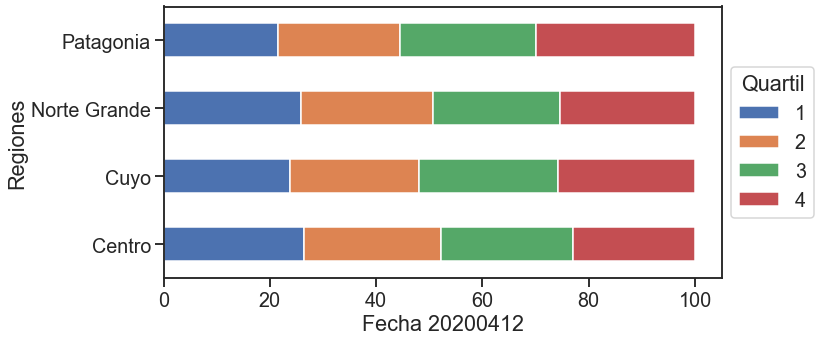
Las variables *categoría 1, 2 y 3*, si bien sumamente útiles para realizar un análisis más detallado, presentan valores NaN, por lo que se definió eliminarlas del conjunto a utilizar en el análisis. De la misma manera, se eliminaron todas las observaciones que presentaban valores faltantes en precio, o en las categorías de identificación del negocio (localidad, bandera, región), dado que al momento no encontramos una manera de realizar la re-ubicación de los datos.

Sobre la variable Precio se realizó una transformación (PrecioxUnidad), consistente en re-expresar los precios por unidades de medida estándares (kg, lts, mts), y posteriormente se realizó la limpieza de outliers. Se definió como outlier todo aquel valor que se encontraba a más de 3 desvíos estándar de la media de precio (por unidad de medida) de todo el país, para cada fecha particular. Se clasificaron 16.001 artículos como outliers, que representaban el 0,723% del total de valores disponibles[[1]](#footnote-1).Habiendo realizado la transformación mencionada sobre el precio, se estudió si la variable PrecioxUnidad se distribuye de manera normal, resultando negativo, tanto desde un punto de vista visual como a través de la prueba de Kolmogorov.

Posteriormente, nos interesamos en saber cual era la distribución de los precios condicionadas a variables categóricas, para intentar determinar si existía algún tipo de patrón.

En el caso de la variable región, se calcularon los cuartiles de precios y se segregó por región. Una primera conclusión, es que las regiones Norte Grande y Centro tienen mayor participación de precios “bajos” mientras que Patagonia tendría una mayor participación de precios pertenecientes al 4 cuartil (más altos).



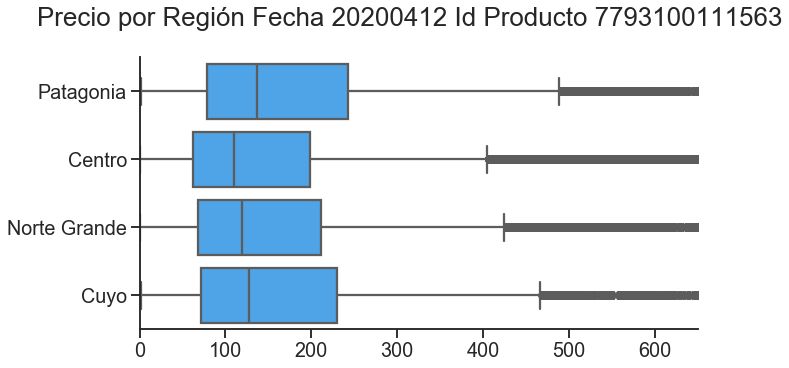


# Preguntas disparadoras

Para poder analizar los precios relativos de los diferentes bienes se definió al producto *7793100111563 Crema Dental Interdental Colgate* como el bien numerario que se consideraría, dado que es un artículo con presencia en las 24 provincias y cuyo precio fue relevado en todos los relevamientos.

Con respecto a la correlación de los precios nominales y relativos, los mismos deberían presentar correlación positiva, ya que los mismos se encuentran divididos por un valor independiente (exógeno).

La dispersión de precios la medimos teniendo en cuenta los cuartiles de precios por región lo cual se encuentra reflejado en el siguiente gráfico de caja (para el producto elegido bien numerario):

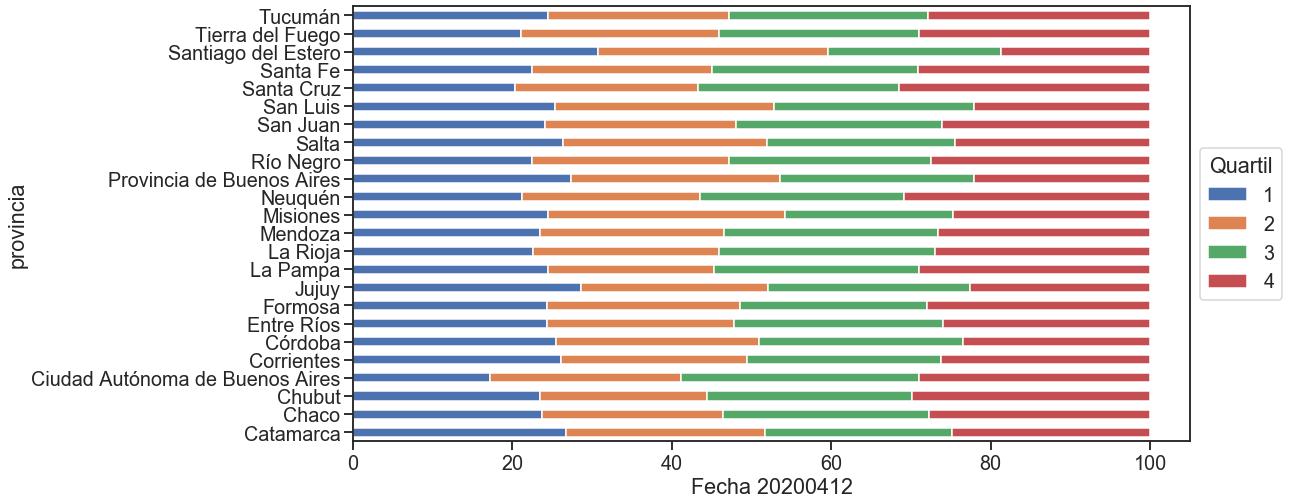


Los precios con mayor dispersión, mediana y valores máximos corresponden a las regiones de Patagonia y Cuyo.

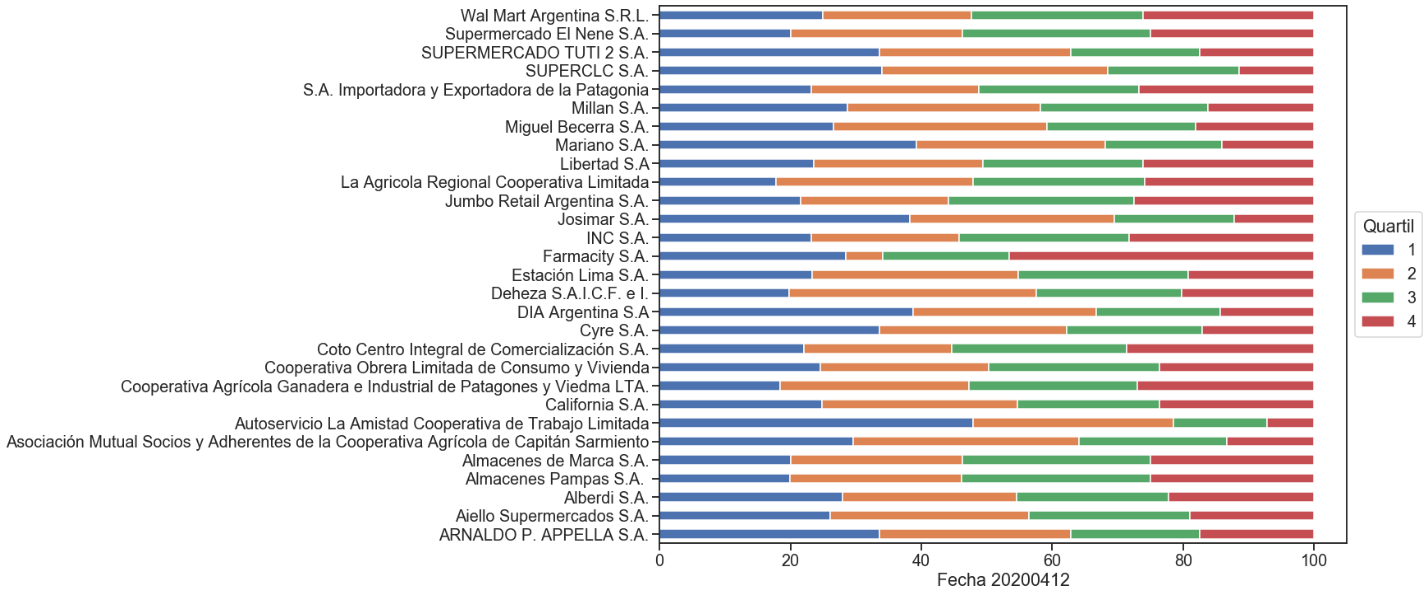
Con respecto a los costos particulares de productos, notamos que existen precios altos en chocolate y mascaras para pestañas, una vez que los mismos son llevados a una unidad de medida estándar, lo cual nos hace suponer que podría haber datos erroneos en la presentación de estos productos. Si se opta por excluir los del análisis, se observa que los productos con mayores valores corresponden a bebidas alcohólicas, y en cuanto a dispersión geográfica encuentran en la región Centro en la provincia de Buenos Aires. Por su parte, los productos mas baratos, son los elementos de higiene personal (shampoo y acondicionador), de puntos de venta en Región Centro – Provincia de Buenos Aires.

Si analizamos como se distribuyen los precios considerando las variables Provincia o cadena de supermercado, observamos lo siguiente:

Al dividir la muestra por cuartiles, la mayor concentración de precios dentro del cuartil 1 se da en las provincias de Santiago del Estero, Jujuy, Catamarca, mientras que la mayor concentración de productos en el cuarto cuartil se da en la provincia de Santa Cruz, Chubut, CABA, Neuquén. Nuevammente observamos que los productos parecieran ser mas caros en la región Patagonia que en las restantes, a excepción de CABA, que logra incluirse en el grupo de jurisidiciones “caras”.

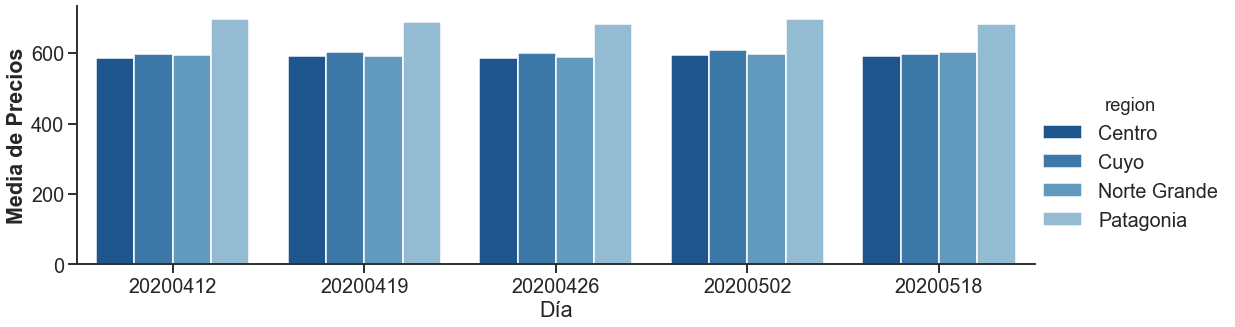


Por su parte, si analizamos las diferentes cadenas de supermercados, observamos que la cadena mas económica sería Autoservicio La Amistad y la más cara Farmacity. El grafico a continuación refleja la distribución de precios por cadena.



Volviendo nuevamente a las provincias, si analizamos el promedio de precios, notamos que la que presenta el menor promedio de precios es Catamarca, mientras que aquella con mayor promedio de precios es Tucumán.

Sin embargo, notamos que si el análisis se hace solo considerando productos que estén en todas las regiones, la región con mayor mediana de precios es la región Patagonia, lo cual coincide con el hecho que además presentó la mayor dispersión del bien numérico, mayor mediana y mayores cuartiles que los restantes.



Focalizando el análisis en la provincia de Córdoba notamos que es muy probable (aproximadammente 70%) que los precios de la provincia estén por debajo del promedio nacional.

# Aspectos adicionales y conclusiones

En lo que respecta a particularidades del conjunto de datos, encontramos los siguientes productos menos frecuentes

'Radicheta Atada La Anonima', 'Desodorante Atitranspirante Nivea Pearl & Beauty 43', 'Gel Exfoliante en Pomo Nivea Visage', 'Desmaquillador de Ojos Bifasico en Botella Nivea', 'Crema Facial Cuidado Radiante Nivea'

Mientras que las marcas menos frecuentes son las siguientes:

'La Anónima', 'Nevares', 'Top Line', 'Vauquita', 'Mentho Plus'

Además, las provincias y las cadenas de supermercados presentarían algún tipo de correlación, ya que de la prueba Chi Cuadrado rechaza la hipótesis nula de independencia de ambas variables.

Como conclusión del trabajo y posibles aspectos a profundizar encontramos los siguientes puntos:

1. Existiría al menos una región que sería mas económica que las restantes.
2. La región Patagonia sería la mas cara.
3. Los estadísticos convencionales se ven afectados por la disponibilidad de información por provincia/región. Al no disponer de categorías de productos (lo cual nos permitiría remplazar los datos faltantes o hacer un análisis por categorías) se podría sub-dividir el dataset en uno que contenga productos que estén en todas las provincias. Sin embargo, esto generaría una muestra mucho más pequeña.
4. Observamos que dadas las características de los establecimientos que presentan información, es entendible que se disponga de mayor variedad de precios en las regiones y provincias de mayor desarrollo que en las mas rezagadas.

1. A modo de ejemplo, uno de los artículos detectados es el siguiente: 0000042300083 - Crema Facial Cuidado Nivea – Precio por unidad: 1399,0 – Precio por unidad promedio del país (para la fecha 2020-04-12): 2139,98 [↑](#footnote-ref-1)