Análisis de precios de Capital Federal

Informe TP1 - Data Mining

Ing. Christian Jorge Marcusa

Maestría en Data Mining, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires
Presentado el 3 de Junio de 2019

Abstract

La fluctuación de precios en los mercados porteños es un factor fundamental en la vida de los ciudadanos. Con variables y factores externos influyentes en ello, como la inflación, la oferta y demanda, el mercado cambiario, e incluso la elaboración nacional y extranjera de los mismos, se hace muy difícil encontrar un precio acorde a cada producto. Incluso a esto, sumamos que si el precio es justo para el mismo, si conviene comprar otro o mismo comprar en otra cantidad, si conviene comprar en uno u otro mercado (Ya sea minorista, mayorista o hipermercado). Todos estos factores son algunos de los que se analizan en este desarrollo, en el cual se intenta brindar un análisis objetivo al respecto.

Keywords: Data Mining, Precios, Descubrimiento de Conocimiento, KDD

Introducción

Al hablar de precios, y su análisis, debemos siempre recurrir a datos objetivos e históricos de los mismos. Por lo tanto se ha buscado información confiable provista por el gobierno de la ciudad para realizar este análisis.

Se ha planteado una estructura de datos sencilla y durante su desarrollo se irá simplificando aún más para análisis específicos.

Tabla 1: Sucursales (length = 837)

Columna	Tipo
sucursalTipo	string
direccion	string
provincia	string
banderaId	numeric
localidad	string
banderaDescripcion	string
lat	numeric
lng	numeric
comercioRazonSocial	string

sucursalNombre	string
comercioId	numeric
sucursalId	string
id	string

Tabla 2: Precios (length = 1.584.661)

Columna	Tipo
_id	string
producto	string
sucursal	string
precio	numeric
fecha	date
medicion	string

Tabla 3: Productos (length = 1.000)

Columna	Tipo
nombre	string
marca	string
presentacion	string
id	string

Lo primero que podemos observar es la cantidad de campos no numéricos, que trataremos en la parte de preprocesamiento.

Es importante saber que cada precio corresponde a una medición (1 a 10), y que también corresponde a un producto y a una sucursal determinada. Esto nos da una idea del diagrama de datos que poseemos, y podemos tratar a nuestra colección de precios como una **tabla de hechos**.

Es importante también hacer un análisis de la distribución de los datos crudos, así podemos darnos una idea clara del dominio.

Preprocesamiento

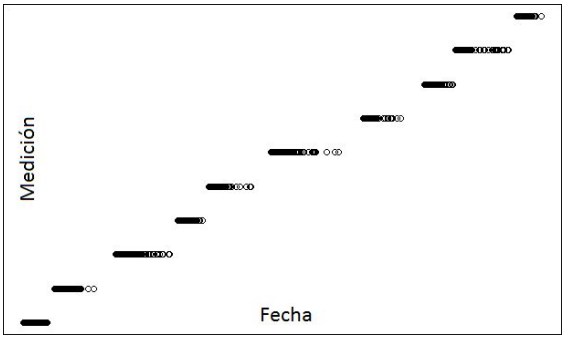
Formateo y selección

Como se mencionó anteriormente, los datos tienen muchas variables de tipo texto, las cuales no son compatibles con algunos métodos que utilizaremos en el desarrollo. En principio los que más debemos transformar en factores son los ids (Que si bien pueden no tener correlación en el análisis, nos servirá incluirlos, por si aportan algún tipo de información, por ejemplo, en los outliers).

Es importante también identificar qué columnas de texto nos aportan información a la metodología. Por ejemplo, en Productos, marca y presentación aportan al desarrollo pero nombre posiblemente no. Por esto mismo, elegimos un conjunto de datos de texto que creemos que no nos aportan en primera instancia, y los eliminamos de los gráficos y tablas.

Los mismos son:

- Sucursales -> Dirección: Disponemos de latitud y longitud si necesitamos hacer un análisis de geolocalización, y son mucho mejor para trabajar que el texto plano.
- Sucursales -> Provincia (Todos nuestros datos son de CABA, e incluso está en muchos casos formateado distinto)
- Sucursales -> Localidad (Idem anterior)
- Sucursales -> BanderaDescripción (Es texto plano)
- Sucursales -> Nombre / Ids (Estos campos estaría bueno unificarlos en uno solo. En principio los dejaremos y veremos con cual etiquetamos para cada caso particular).
- Precios -> _id (Se factoriza en numérico)
- Productos -> Nombre (Solo se utilizará para etiquetar en caso requerido, aunque con la marca y la presentación debería alcanzar en casi todos los casos)
- Precios -> Fecha (Fecha parece super importante, pero no aporta mucha información al análisis dado que ya tenemos medición, se detalla con un gráfico a continuación, donde se puede apreciar que la columna “medición” ya funciona como una discretización de la fecha, y por lo tanto aportan la misma información)



Gráfica 1: Asociación Medición-Fecha.

Datos Faltantes

Nuestro principal conjunto de trabajo es el de precios, el cual es nuestra tabla de hechos. Para asegurarnos que el análisis es correcto deberíamos asegurarnos que todos los datos tengan las 10 mediciones, sino la información tendrá un sesgo inductivo y los estadísticos podrían verse afectados. Sabemos que nuestra colección de precios va a tener información sobre los 1000 productos en las 837 sucursales. Las sucursales pueden no estar presentes porque no todas venden todos los productos y esa información es correcta, y para cada par de ellos, debería haber 10 mediciones.

	producto	sucursal	avg_precio	var	mediciones
1	0000040084107	10-1-220	39.77778	6.567491	9
2	0000040084107	10-1-26	40.61111	6.503738	9
3	0000040084107	10-2-116	40.61111	6.503738	9
4	0000040084107	10-2-118	40.88889	6.693176	9
5	0000040084107	10-2-150	48.71429	4.535574	7
6	0000040084107	10-2-151	40.61111	6.503738	9

Figura 1: Par ordenado Producto-Sucursal con mediciones faltantes. (54.188 resultados)

Y a su vez:

	producto	sucursal	avg_precio	var	mediciones
1	0000040084107	10-1-214	40.15	6.302777	10
2	0000040084107	10-1-219	40.40	6.497008	10
3	0000040084107	10-1-271	40.40	6.497008	10
4	0000040084107	10-1-3	40.15	6.302777	10
5	0000040084107	10-1-30	40.40	6.497008	10

Figura 2: Par ordenado Producto-Sucursal sin mediciones faltantes. (113.123 resultados)

Esto significa que de los 167.311 pares ordenados Producto-Sucursal en nuestra tabla de hechos, hay 54.188 resultados que no poseen las 10 mediciones.

Tenemos varias alternativas para trabajar con estas faltantes. La primera es eliminar de nuestro set de trabajo aquellos que no tengan las diez mediciones, ya que sesgan los estadísticos para el análisis (tanto vertical como horizontalmente). Por ejemplo, no es lo mismo trabajar de igual manera un producto que tiene 9 mediciones que uno de 10, y a su vez no es lo mismo trabajar una medición entera (todos los productos) cuando algunos no se encuentran.

Otra opción es utilizar alguna técnica para el llenado de esta información. Ya sea imputación por la media, por regresión, MICE o HotDeck. Usar la media es una mala decisión porque en las mediciones en los extremos (1 o 10) generaría un error muy grande, incluso si se utilizara la media para cada par ordenado “producto-sucursal”. Se considera que la regresión produciría el menor error cuadrático medio, para cada par ordenado.

Por lo tanto, lo que se quiere demostrar es que lo correcto sería hacer una regresión para cada medición faltante, con la información de las mediciones presentes (Ya que no hay productos con pocas mediciones). Como esto necesita una capacidad de procesamiento grande y una complejidad programática (iterar haciendo regresión 54.188 veces), se consideran los métodos automáticos, en particular MICE.

Para ello se ha utilizado en primer lugar las funciones expand.grid y merge, y obtenemos un dataset con todas las posibles combinaciones de los datos, con NA en los faltantes (precio).

Esto proporciona todas las combinaciones posibles entre producto, sucursal y medición, y al utilizar merge, se completan los valores presentes de precio, y se asigna NA para los registros ausentes.

El data.frame resultante es el que será utilizado para aplicar la función MICE.

	producto	sucursal	medicion	precio
45960	7500435004664	10-3-400	10	NA
45961	7500435004664	11-4-1027	1	329.20
45962	7500435004664	11-4-1027	2	329.20
45963	7500435004664	11-4-1027	3	329.20
45964	7500435004664	11-4-1027	4	NA
45965	7500435004664	11-4-1027	5	329.20
45966	7500435004664	11-4-1027	6	NA
45967	7500435004664	11-4-1027	7	329.20

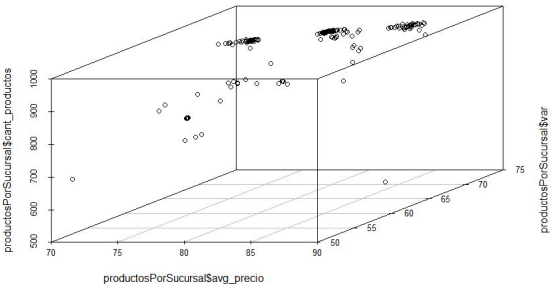
Figura 3: Dataframe con faltantes en NA. (Length=1.750.000)

Disclaimer: No se pudo hacer que el proceso pueda ejecutar MICE correctamente y finalice su ejecución. Se dejará para un análisis futuro. Y por lo tanto se eligió utilizar solo los datos completos.

Valores atípicos

Se han aplicado distintos métodos para detectar valores atípicos (outliers) en el dataset. El primer vistazo a la información nos indica que los productos de precio medio presentan grandes varianzas pero en muy pocos casos, mientras que los productos de mayor precio son más inestables, pero sus varianzas no son tan elevadas. No obstante no podemos considerar atípicos a los precios altos porque son importantes para el análisis.

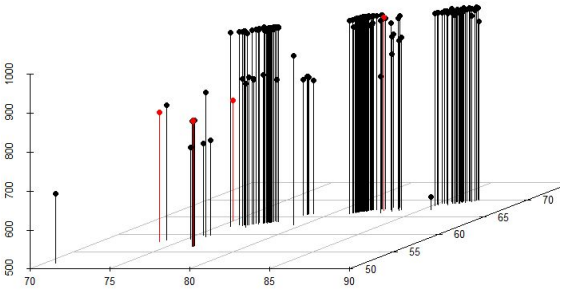
Para trabajar los outliers se han agrupado los datos en 3 subsets. Uno agrupados por sucursal, otro por producto, y otro por medición.



Gráfica 2: Precios, Varianza y Cantidad de productos, para cada sucursal.

Luego se utiliza el método LOF (Local outlier factor) para ver que outliers tenemos utilizando esa dimensionalidad.

También utilizamos la proyección sobre el plano Media-Varianza para visualizarlos mejor.

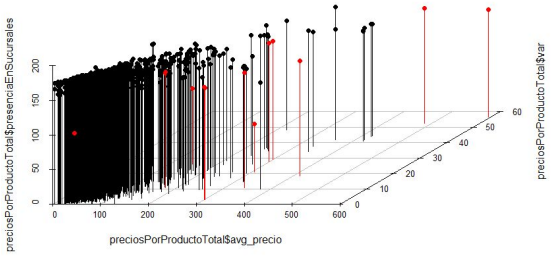


Grafica 3: Outliers con LOF en sucursales según Media-Varianza-Productos.

	sucursal	avg_precio	var	cant_productos	score	outlier
1	10-3-242	73.65915	57.99445	831	9.563357	TRUE
2	10-3-250	74.95397	63.95576	808	252.928028	TRUE
3	15-1-1072	76.57509	56.57905	822	6.021024	TRUE
4	15-1-1076	76.54945	56.58545	822	14.782855	TRUE
5	15-1-1092	76.58481	56.53721	822	6.021024	TRUE
6	9-1-700	82.76658	66.85750	995	8.457781	TRUE

Figura 4: Sucursales outliers.

Analizando los resultados elegimos conservar los outliers porque no parecen ser ofensivos para el análisis.



Gráfica 4: Outliers de productos, según sucursal, media y varianza de precios.

Acá vemos que los que se separan para el lado de altas medias y varianzas son considerados outliers. Aquellos que se separan de la nube de puntos que

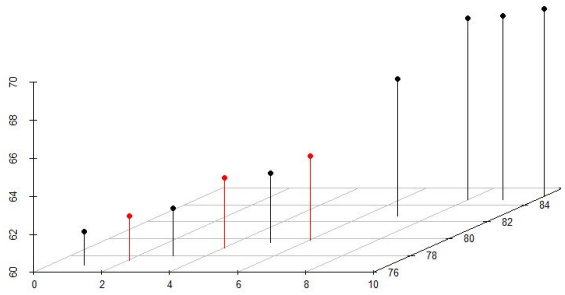
sigue una recta. También aquellos que están en pocas sucursales.

producto	avg_precio	var	presenciaEnSucursales	score	outlier
5000281040356	305.92917	20.830255	69	4.636117	TRUE
7500435004657	413.39159	18.497959	165	5.582373	TRUE
7500435004664	339.87663	10.914282	165	20.444540	TRUE
7622300847340	21.66963	4.427649	93	12.308856	TRUE
7790670050667	483.13956	52.647771	165	15.422083	TRUE
7790975000183	299.87377	2.968003	162	11.053754	TRUE
7790975000190	300.21372	3.095299	161	11.012874	TRUE
7791250001994	594.27462	56.640154	154	26.279287	TRUE
7791290011731	296.98339	29.188113	170	7.142730	TRUE
7791290011939	297.33794	27.830383	170	7.094632	TRUE
7791293033969	145.77432	26.218011	108	7.341405	TRUE
7791560000441	165.75268	12.429393	161	5.029557	TRUE
7794626007217	173.77841	10.993227	166	5.394588	TRUE

Figura 5: Productos Outliers.

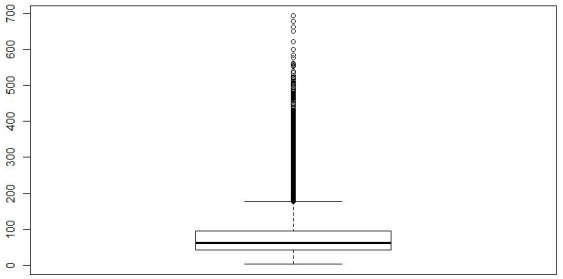
Se decide dejarlos en el análisis ya que aportan y no se consideran peligrosos.

En el caso de agrupar por medición, se considera que no es información útil, ya que el resultado es hacer una regresión de 10 puntos y sólo hace falta mirar los outliers para saber que la información debe ser tenida en cuenta.



Grafica 5: Outliers según medición.

El último análisis de outliers que haremos es considerar todos los precios (lo cual no tiene mucho sentido porque estaríamos dejando afuera los productos más caros. Aún así, se adjunta el resultado del análisis para conocer la distribución de los precios (sin contemplar medición, ni sucursal, ni producto).



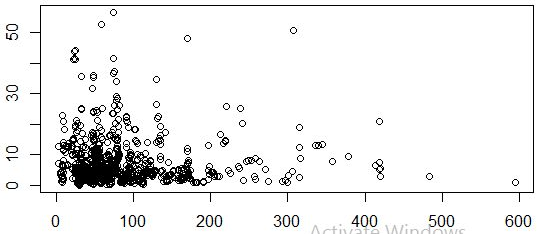
Grafica 6: Boxplot de precios.

Se observa que existen muchos outliers por precio. Más adelante se particionan estos datos para hacer un análisis discriminatorio entre los valores dentro de Q1 y Q3, y los outliers de precios.

Metodología

En esta sección utilizaremos los datos preprocesados anteriormente para hacer un análisis detallado y poder responder preguntas concretas sobre los datos en cuestión.

¿Cuáles productos variaron más con el tiempo?
Para llevar a cabo este análisis se hizo una agrupación de las mediciones por producto, calculando la varianza y el desvío estándar para cada producto.



Gráfica 7: Media y desvío estándar para cada producto en el tiempo.

Se observa que los productos de mayores precios, y los menores (extremos) tienen poco desvío. El grueso de cambios estuvo en los productos entre 20 y 200 pesos de media. Detallaremos los primeros 10 para ver si podemos realizar una concordancia.

avg_precio	mediana	var	minimo	maximo	nombre	marca	presentacion
418.4880	417.9	41.42311	205.00	521.39	Cafe en Capsula...	NESCAFÉ	112.0 gr
418.8927	417.9	41.67967	135.60	521.39	Cafe en Capsula...	NESCAFÉ	16.0 un
420.4248	417.9	43.93817	299.00	527.00	Cafe en Capsula...	NESCAFÉ	8.0 un
420.4909	417.9	44.29963	205.00	527.00	Cafe en Capsula...	NESCAFÉ	100.0 gr
418.8838	417.9	41.13736	299.00	521.39	Cafe en Capsula...	NESCAFÉ	160.0 gr
483.1396	495.0	52.64777	246.00	539.90	Hamburguesas ...	PATY	960.0 gr
358.0989	339.0	41.67919	260.00	419.00	Whisky White H...	WHITE HORSE	750.0 cc
594.2746	559.0	56.64015	459.00	693.00	Whisky JyB 750 Cc	J&B	750.0 cc
221.2215	196.0	48.05221	196.00	368.00	Pan~al G Active ...	HUGGIES	24.0 un
307.5648	336.9	50.76581	201.49	345.90	Bocaditos de Po...	SADIA	900.0 gr

Figura 6: Top 10 productos que más aumentaron su precio.

Como podemos observar, hay determinada correlación. 5 de los productos son Cápsulas de NESCAFE, 2 son Whiskies, 2 son congelados, y el último pañales.

¿Cuales marcas de productos sufrieron más modificaciones de precios?

Para llegar a este análisis se agruparon los productos por marca y se utilizó la mediana de las varianzas como valor más adecuado.

Se observa una fuerte tendencia a las marcas de bebidas alcohólicas en este análisis, en particular bodegas populares de vino y whiskies de mediano precio.

	marca	avg_var
293	J&B	56.640154
292	SADIA	50.765805
291	WHITE HORSE	41.679188
290	NAVARRO CORREAS	36.809091
289	FOND DE CAVE	35.313775
288	DON DAVID	31.760726
287	SKIP	27.136458
286	TRESEMMÉ	26.218011
285	DRIVE	26.182480
284	KILLKA	25.885503

Figura 7: Marcas que más subieron sus precios.

¿Cuales marcas de productos sufrieron menos modificaciones de precios?

Reutilizando el análisis anterior podemos llegar a la inversa.

	marca	avg_var
1	VERAO	0.1661930
2	TANG	0.5124065
3	CACHAMAI	0.6836159
4	CARIOCA	0.7153275
5	VALMONT	0.7909528
6	CLIGHT	0.8315321
7	WHISKAS	1.0564571
8	PEDIGREE	1.1105305
9	DANONINO	1.1264332
10	DÍA	1.1293520

Figura 8: Marcas que menos subieron sus precios.

Se observa una fuerte presencia de productos de jugos, y alimento de mascotas. Así como productos marca DÍA que es una marca universal.

¿Los hipermercados ofrecen productos más baratos o más caros que los supermercados?

El primer análisis que queremos hacer es la cantidad de productos que ofrece cada tipo de sucursal.

	sucursalTipo	avg_mean	avg_var	promedioProductos
1	Hipermercado	81.10659	66.17089	994.2609
2	Supermercado	80.55753	64.93650	942.4419

Figura 9: Productos, posición y dispersión de precios de los tipos de sucursales.

Vemos que en general los hipermercados venden más cantidad distinta de productos que los supermercados (Tengamos en cuenta que este análisis está hecho sobre 1.000 productos, y la media es de 994,26).

A simple vista no se observa una diferencia significativa entre los precios, si bien la tendencia es un poco más elevada en los hipermercados.

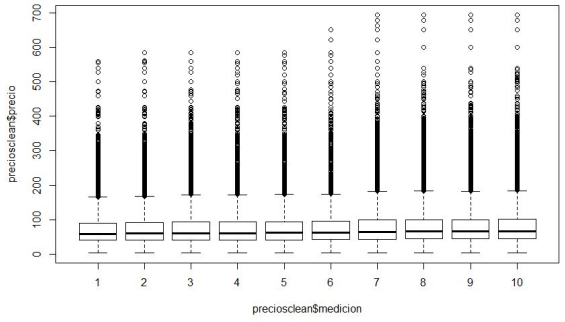
¿En qué período se produjo el mayor incremento de precios?

Para responder esta pregunta consideramos las medias, y los máximos y mínimos.

	medicion	avg_precio	max_precio	min_precio
1	1	76.90197	559	2.85
2	2	77.46766	584	2.85
3	3	77.98651	584	2.85
4	4	78.92510	584	2.85
5	5	79.53917	584	2.85
6	6	79.85231	650	2.85
7	7	82.66731	693	2.85
8	8	84.59351	693	2.85
9	9	84.63187	693	2.85
10	10	85.01748	693	3.73

Figura 10: Medias, máximos y mínimos de precios por medición.

En primer lugar se observa que los topes máximos tuvieron más fluctuación que los mínimos, y la creciente de las medias fue constante en el tiempo.



Gráfica 8: Boxplot de mediciones y precios

Se observa que hay movimiento en los outliers pero la tendencia de la mediana y el rango intercuartil casi no sufren modificaciones.

¿Son más volátiles los productos caros o los baratos?

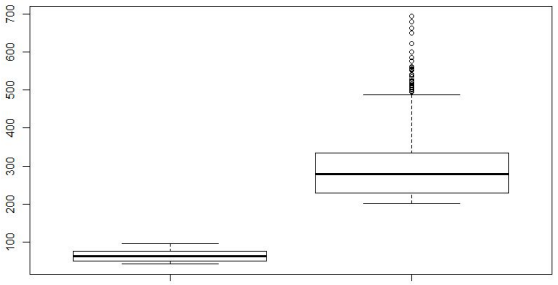
Para este análisis volvemos a la parte de detección de outliers. En la misma utilizamos un boxplot para ver los outliers según el precio (Sin importar la sucursal, la medición ni el producto). Retomaremos este análisis.

Se divide el data.frame en aquellos datos que pertenecen al rango intercuartil (Q1-Q3), y en aquellos que se consideran outliers en el preprocesamiento.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.85   42.00   62.90   80.79   95.99  693.00
```

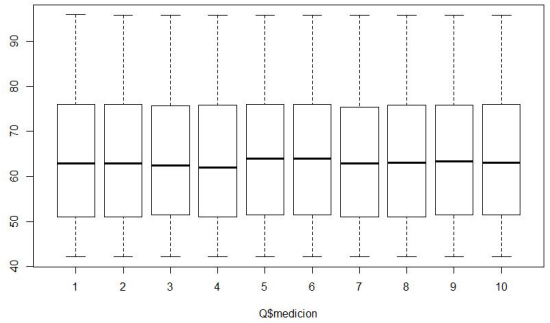
Figura 11: Cuartiles del boxplot de precios.

El conjunto al que llamaremos Q son aquellos valores que oscilan entre 42 y 95.99 (785.144 filas). Mientras que el conjunto que llamaremos O lo simplificamos en los valores mayores a 200 (77.816 filas).

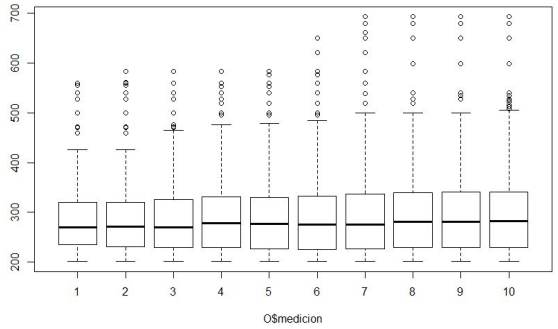


Grafica 9: Q vs O

Podemos observar que los datos del conjunto Q son concisos, mientras que los del conjunto O son más volátiles.



Grafica 10: Q precio por medición



Grafica 11: O precio por medición

Si bien no hay muchos cambios, se ve una tendencia más marcada en los productos más elevados de precio, y con alta presencia de outliers.

¿Concuerdan los resultados obtenidos con las estadísticas del INDEC?

Primero que nada necesitamos obtener los datos del indec para el período en cuestión. (5/11/2018 al 25/2/2019).

Nov-18	Dec-18	Jan-19	Feb-19
4.2	2.6	3.0	3.0

Figura 12: IPC (Índice de precios al consumidor) según INDEC.

Esto nos da un aumento promedio de 13.41997028 de aumento del IPC entre el 1 de Noviembre de 2018 y el 31 de Febrero de 2019.

La primer aproximación es medir las medias de las mediciones 1 y 10, y así llegar a un número que se pueda contrastar con el IPC.

```
     precio                precio
Min.   :  2.85   Min.   :  3.73
1st Qu.: 40.00   1st Qu.: 44.50
Median : 58.99   Median : 65.90
Mean   : 76.90   Mean   : 85.02
3rd Qu.: 90.63   3rd Qu.:100.59
Max.   :559.00   Max.   :693.00
```

Figura 13: Comparación de medición 1 y medición 10.

Variable	1er M	10ma M	%
Min	2.85	3.73	30.87
1erQ	40	44.50	11.25
Mediana	58.99	65.90	11.17
Media	76.90	85.02	10.55
3erQ	90.63	100.59	10.98
Max	559.00	693.00	23.97

Tabla 4: Incrementos entre medición 1 y medición 10.

Como podemos observar, el resultado es similar a otros resultados que tuvimos en otros análisis. Hay mayor cambio en los extremos que dentro del rango intercuartil.

También cabe aclarar el sesgo inductivo del IPC (El cual incluye mucho más que los 1000 productos que tenemos en cuenta en este análisis, y a su vez contiene precios de servicios, públicos y privados).

Resultados

Los resultados de este informe demuestran varias hipótesis sobre el comportamiento de los precios.

- Los datos medidos corresponden con lo provisto por el IPC. Teniendo en cuenta el sesgo inductivo y los datos utilizados (Que si bien fueron muchos, no están en el mismo orden que el IPC).
- Como fue desarrollado a lo largo del informe, se observa que ha habido mayor variabilidad de los precios en los extremos (Más altos y más bajos).
- Se observa que no ha habido variaciones abruptas, ya que se decidió que los valores atípicos eran parte del problema y no errores de carga.
- Se puede observar que los Hipermercados venden una cantidad mayor de productos que los Supermercados, pero que sus precios son apenas más elevados (A diferencia de otras épocas, donde los pequeños mercados fundían por no poder competir contra sus precios tan bajos).
- Hay una tendencia a que las bebidas alcohólicas (En particular los vinos y los whiskies) y el café en cápsula ha aumentado su valor muy por encima que otros productos.
- Se ha observado que los jugos han variado poco su valor.

Discusión y trabajos futuros

Los resultados arrojan que es posible extraer mucha información de los datos de precios. Queda para futuras investigaciones cruzarla con otras fuentes de datos masivas.

Si bien los datos del IPC son a veces discutidos, y existe un gran malestar por aumentos desmedidos, se puede sacar conclusiones a partir de los mismos.

En este trabajo se eligió comparar con el IPC, otra buena herramienta es utilizar la latitud y la longitud para hacer un análisis geográfico del cambio de precios, utilizando servicios como GoogleMaps o Usig.

Conclusión

Se han descubierto cosas interesantes a partir de los datos, que puede ser utilizada tanto para futuros análisis, como para explicar fenómenos actuales.

Referencias

- Código fuente utilizado para este trabajo está disponible en GitHub, solicitar acceso. (https://github.com/Chaitooler).
- INDEC, IPC. https://www.indec.gob.ar/
- Dplyr-Documentation (https://dplyr.tidyverse.org/)
- R-Documentation (https://www.rdocumentation.org/)
- https://dmuba.github.io/
- MongoDB-Documentation https://docs.mongodb.com/