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 5 de Agosto 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, se debe 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 se puede observar es la cantidad de campos no numéricos, que se tratarán 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 se poseen, y se puede tratar a la colección de precios como una **tabla de hechos**.

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

Preprocesamiento

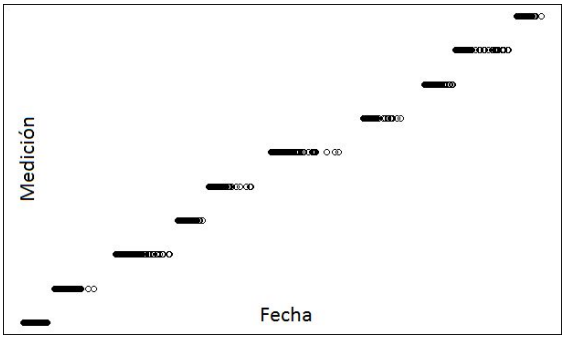
Formateo y selección

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

Es importante también identificar qué columnas de texto aportan información a la metodología. Por ejemplo, en Productos, marca y presentación aportan al desarrollo pero nombre posiblemente no. Por esto mismo, se elige un conjunto de datos de texto que parecen no aportar información en primera instancia, y se filtran.

Los mismos son:

- Sucursales -> Dirección: Se dispone de latitud y longitud para hacer un análisis de geolocalización, y son mucho mejor para trabajar que el texto plano.
- Sucursales -> Provincia (Todos los 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 se dejan y se tratarán en 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 existe 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

El principal conjunto de trabajo es el de precios, el cual es la tabla de hechos. Para asegurar que el análisis es correcto se debe verificar que todos los datos tengan las 10 mediciones, sino la información tendrá un sesgo inductivo y los estadísticos podrían verse afectados.

Se sabe que la 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 la tabla de hechos, hay 54.188 resultados que no poseen las 10 mediciones.

Se tienen varias alternativas para trabajar con estas faltantes. La primera es eliminar del 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). En este caso se ha optado por hacer una regresión por medición (para aproximar las mediciones faltantes) utilizando como modelo la media del producto contra la medición correspondiente.

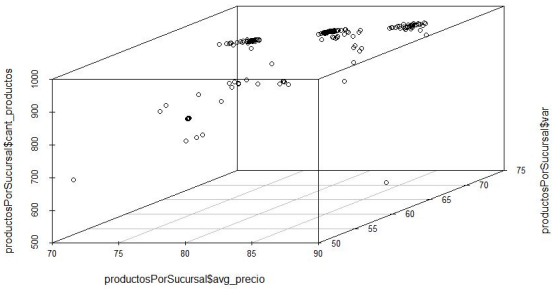
Esto da como resultado las 10 regresiones de la forma y=a.x = b siendo la variable independiente la media del producto, y la dependiente la medición en cuestión. Se obtienen 10 pares de coeficientes a y b que generan un modelo para cada medición.

Se han explorado otros métodos como MICE, el cual resultó poco performante, y perder los resultados o volver a generarlos por diferencias, tenía un alto costo de procesamiento, por eso se ha optado por utilizar las regresiones.

Valores atípicos

Se han aplicado distintos métodos para detectar valores atípicos (outliers) en el dataset. El primer vistazo a la información 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 se pueden considerar atípicos a los precios altos porque son importantes para el análisis.

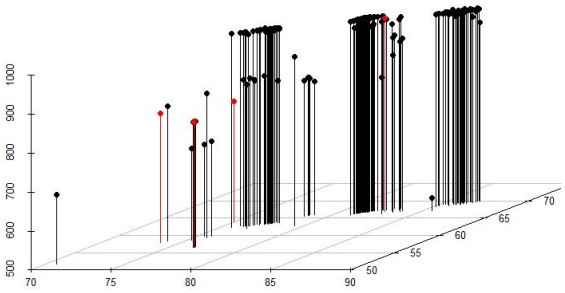
Para trabajar los outliers se han agrupado los datos en 2 subsets agrupándolos por distintas dimensiones. El primer subset es la media, varianza, y cantidad de productos, agrupados por sucursal. El segundo agrupados por producto (media, varianza y sucursales que lo comercian).



Gráfica 2: Media de precio, Varianza de precio y Cantidad de productos, agrupados por sucursal.

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

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

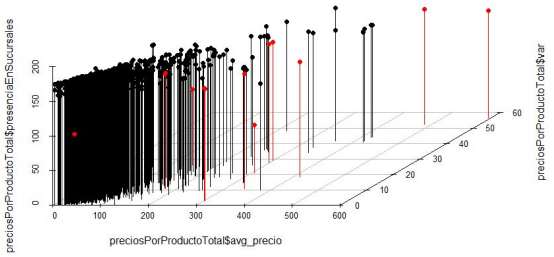


Grafica 3: Outliers con LOF en sucursales según Media-Varianza-Productos.

	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 se decide conservar los outliers porque no parecen ser ofensivos para el análisis.



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

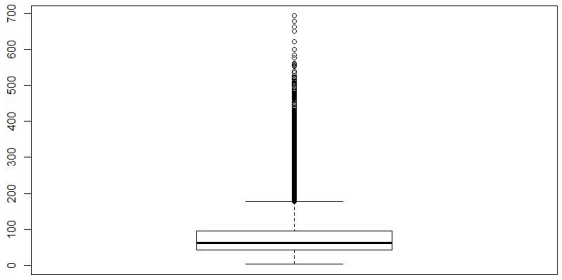
Acá se ve que los que se separan para el lado de altas medias y varianzas son considerados outliers. Aquellos que se separan de la nube de puntos que sigue una recta. También aquellos que están en pocas sucursales.

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.

El último análisis de outliers que se realiza 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.

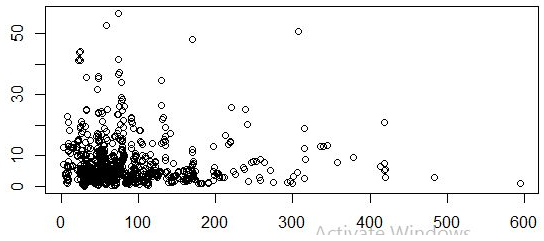
Se observa que existen muchos outliers por precio. Más adelante se particionan estos datos para hacer un análisis discriminatorio entre los valores dentro de Q1 y Q3, y los outliers de precios.

Metodología

En esta sección se utilizan 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.

Luego se calculó la media para las mediciones 1 y 10 (para sacar del juego la varianza entre las distintas sucursales) y se calculó un score, que es (m10-m1)/media. Este score representa la variación del precio agrupado por producto y tomando las medias de las mediciones, y tenemos el siguiente resultado:

avg_m1	avg_m10	media	score	_id.$oid	nombre	marca
48.58608	99.95097	70.72057	0.7263076	5cbc69be7af152186...	Hamburguesas de Carne ...	PATY
52.08078	96.22421	69.21573	0.6377658	5cbc69be7af152186...	Desodorante Antitranspir...	REXONA
29.55609	49.92365	36.40721	0.5594374	5cbc69be7af152186...	Gaseosa Pomelo Schwep...	SCHWEPPES
29.64682	48.30387	36.31456	0.5137624	5cbc69be7af152186...	Gaseosa Lima Limon Sprit...	SPRITE
115.66920	160.80871	142.43306	0.3169174	5cbc69be7af152186...	Acondicionador Blindaje ...	TRESEMMÉ
13.21346	18.04342	15.30609	0.3155583	5cbc69be7af152186...	Polvo para Mousse Choc...	ROYAL
18.26562	25.05018	21.80640	0.3111271	5cbc69be7af152186...	Agua Mineral sin Gas Kin ...	KIN
30.11654	40.92732	35.40419	0.3053532	5cbc69be7af152186...	Gaseosa Pomelo Light Pa...	PASO DE LOS TO...

Figura 6: Top productos que más aumentaron su precio.

¿Cuales marcas de productos sufrieron más modificaciones de precios?

Para llegar a este análisis se agruparon los productos por marca y se utilizó la misma tecnica que en el analisis pasado. Se agruparon teniendo en cuenta la media de las mediciones 1 y 10, y asignando un score comparado con la “media de las medias” (ya que las marcas pueden tener más de un producto).

El resultado se observa en la figura a continuación.

	marca	avg_m1	avg_m10	media	score
293	TRESEMMÉ	115.669200	160.808711	142.433056	0.31691738
292	SCHWEPPES	42.408471	57.581397	48.937937	0.31004426
291	PASO DE LOS TOROS	30.116536	40.927319	35.404194	0.30535318
290	LACTAL	36.455905	49.547519	43.640392	0.29998845
289	NUTRILÓN	36.456933	47.831183	42.297030	0.26891368
288	CAREFREE	130.971968	168.181205	143.676871	0.25897861
287	KINDER	45.241895	57.348979	48.014855	0.25215287
286	TERMIDOR	45.628199	58.273734	51.294457	0.24652829

Figura 7: Marcas que más subieron sus precios.

¿Cuales marcas de productos sufrieron menos modificaciones de precios?

Reutilizando el análisis anterior se puede llegar a la inversa.

	marca	avg_m1	avg_m10	media	score
1	HUGGIES	163.623053	147.561479	146.502509	-0.1096334413
2	NORTON	133.184075	128.043641	127.394593	-0.0403504921
3	GREEN HILLS	57.333318	55.831729	55.293454	-0.0271567205
4	NIETO SANETINER	190.328458	185.770827	187.944720	-0.0242498473
5	SADIA	304.839281	298.242217	309.846406	-0.0212914029
6	HARPIC	92.461599	90.627316	92.111678	-0.0199136862
7	PLAYADITO	105.151826	103.185682	104.868448	-0.0187486716
8	SUTER	63.003819	62.008917	62.416966	-0.0159395994

Figura 8: Marcas que menos subieron sus precios.

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

El primer análisis que se quiere 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.

Se puede ver que en general los hipermercados venden más cantidad distinta de productos que los supermercados (Teniendo 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.

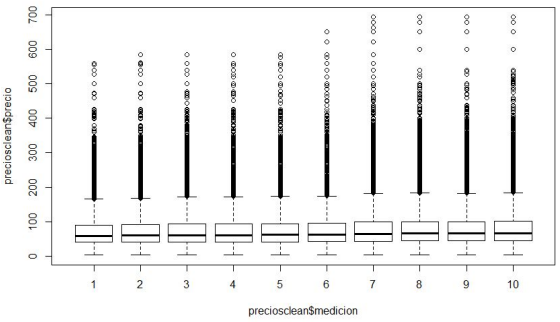
¿En qué período se produjo el mayor incremento de precios?

Para responder esta pregunta se consideran 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 mínimos (30%) tuvieron más fluctuación que los máximos (24%), y la creciente de las medias fue constante en el tiempo.



Gráfica 8: Boxplot de mediciones y precios

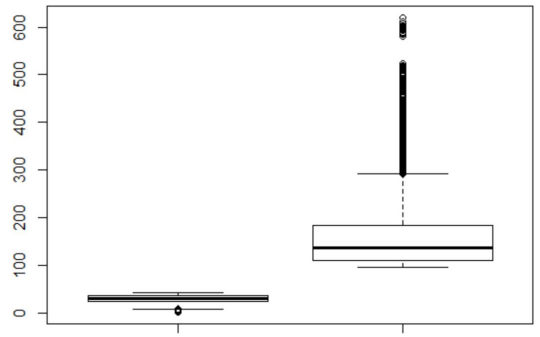
Se observa que hay movimiento en los outliers pero la tendencia de la mediana y el rango intercuartil casi no sufren modificaciones.

¿Son más volátiles los productos caros o los baratos?

Se decide dividir el dataset en O (Productos baratos, aquellos con medias menores al primer cuartil de medias) y Q (Productos caros, aquellos con media superior al tercer cuartil de medias)

```
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 medias de precios.



Grafica 9: Q vs O

Esto muestra una primera aproximación de los conjuntos a contrastar, pero para llegar a la conclusión es necesario estandarizarlos de alguna manera y que corresponden a distintas escalas.

Para ello se crea una transformación de Q y de O y se realiza un score agrupado por producto, y utilizando la media de sus mediciones primera y décima, y se divide por la media.

	producto	avg_m1	avg_m10	media	score
331	7790895010088	29.556093	49.923649	36.407209	0.55943744
330	7790895064173	29.646818	48.303873	36.314561	0.51376237
329	0000040084107	36.048964	52.289488	40.043732	0.40556968
328	7794000960275	30.800000	45.050000	41.393750	0.34425487
327	7622300871772	13.213457	18.043421	15.306089	0.31555831
326	7790895000201	18.265623	25.050185	21.806402	0.31112706

Grafica 10: Q scores

	producto	avg_m1	avg_m10	media	score
305	7791293033969	117.07562	164.58964	145.41382	0.32675035
304	7790010570626	90.81999	118.37500	103.27676	0.26680747
303	7501059273276	357.51261	469.42756	420.35791	0.26623729
302	7613032396350	358.39179	469.94822	420.48450	0.26530450
301	7790010616812	180.43623	232.99537	198.11264	0.26529927
300	7790010616775	180.04326	231.70420	197.39556	0.26171280
299	7791290013551	181.10785	234.49766	205.46819	0.25984464

Grafica 11: O scores

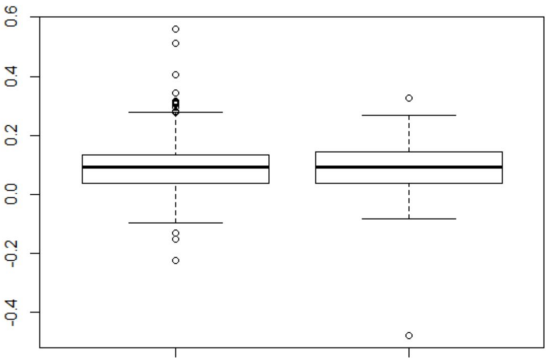
Se puede usar la media de esos scores para ver cuales fueron más volátiles

QRS

Min. 1st Qu. Median Mean 3rd Qu. Max.
-0.22388 0.03764 0.09299 0.09567 0.13392 0.55944

ORS:

Min. 1st Qu. Median Mean 3rd Qu. Max.
-0.47692 0.03759 0.09158 0.09158 0.14559 0.32675



Grafica 12: Q SCORES vs O SCORES

Se observa que sus variaciones siguen una distribución similar en cuanto a proporcionalidad.

¿Concuerdan los resultados obtenidos con las estadísticas del INDEC?

Primero que nada se necesita 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 arroja 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.

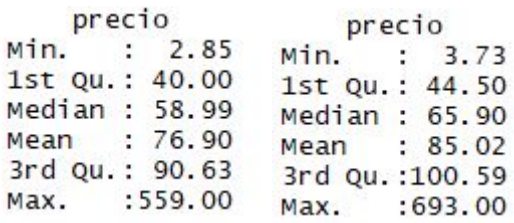


Figura 13: Comparación de medición 1 y medición 10.

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 se puede 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 se tienen 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).
- 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).
- Se observa que los productos caros y baratos tuvieron aumentos proporcionalmente similares.

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/
