

**INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA**

**ESCUELA DE INGENIERÍA Y GESTIÓN**

Trabajo final Integrador

**AUTOR: Laciar, Mauro Jorge (Leg. Nº 105090)**

**TUTOR/ES: Denicolay, Gustavo**

**TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIA DE DATOS**

**BUENOS AIRES**

**PRIMER CUATRIMESTRE, 2022**

# Capítulo I: Introducción

## 1. Introducción

Actualmente existe un desconocimiento de los precios y de la variación de los productos disponibles en comercios por parte de los consumidores. La hipótesis de este trabajo es que los precios de determinados productos de consumo masivo varían según localización, tipo de comercio y día de la semana, por lo tanto, el objetivo es obtener datos de ciertos productos y almacenarlos en una base de datos para obtener insights.

El reconocimiento de patrones sobre la evolución de los precios de ciertos productos permitirá que el consumidor realice decisiones más eficientes en sus consumos.

Dicho proyecto se llevará a cabo mediante la realización de un estudio preliminar que comprende la extracción de datos mediante la creación de un algoritmo que alimente una base de datos a través de la API disponible en la web “precios claros”. Posteriormente se realizará la explotación y visualización mediante la herramienta Tableau. Finalmente, la información estará disponible para su consulta en un repositorio en Git Hub.

El presente trabajo pretende analizar, a través de datos obtenidos mediante un algoritmo, la variabilidad de los precios. La obtención de datos específicos sobre precios y productos generaría una mayor eficiencia en la toma de decisiones de los consumidores.

## 2. Estado de la cuestión

La variabilidad en los precios de los productos de una economía genera diversos impactos en los consumidores, uno de los más importantes es generar incertidumbre en los mismos, así como un aumento en el rango de los precios que los consumidores están dispuestos a aceptar (Murthi, B.P.S, Haruvy, E., & Zang, H., 2007).

Uno de los principales generadores de variaciones en los precios es la inflación. Esta se define como el aumento generalizado en los precios de los bienes y servicios de una economía en un período de tiempo. Uno de los indicadores más utilizados para medir la inflación es el índice de precios al consumidor (IPC). Si bien este indicador es útil como referencia, el mismo contiene una serie de supuestos que provocan que pueda no ser totalmente comparable con la canasta de consumo de un individuo, manteniendo su incertidumbre acerca del valor de los bienes que consume. Los principales se describirán a continuación.

Para la construcción de este índice se debe realizar un muestreo, lo que introduce el error de muestreo, que se explica como la diferencia entre los valores obtenidos si obtuvieran los valores del total de la población y los valores obtenidos a través del muestreo (Smith, P. A., 2021). Por otro lado, estos precios se toman de un promedio por región, cuando en realidad la inflación percibida de un individuo radica en donde realiza sus consumos, que por lo general lo hacen en comercios de cercanía. Asimismo, el índice se construye a partir de una canasta predefinida, sin embargo, según las preferencias del consumidor los productos incluidos en esta pueden estar en línea con sus hábitos de consumo o no. Por último, el IPC se publica a mediados del mes siguiente con respecto al relevado con lo que, en el mejor de los casos, un individuo posee como referencia el índice del mes anterior o de hace dos meses y debe inferir cómo se comportaron los precios en el futuro.

En economías con altos niveles de inflación, como la Argentina, que tuvo un incremento interanual del IPC en diciembre 2021 del 50,9%, los efectos descriptos anteriormente se exacerban añadiendo mayor incertidumbre sobre el valor en unidades monetarias de los productos de consumo.

Otra variable que aumenta la volatilidad de los precios proviene de la estrategia de venta de los comercios que pueden dividirse en dos grandes grupos: EDLP (every day low Price) o HILO (high Low) (Bell, D. R. & Lattin, J. M.,1998). La primera estrategia mencionada consiste en intentar poseer el menor precio posible en la totalidad de los productos ofrecidos en todo momento, y la segunda, como su nombre lo indica, consiste en añadir volatilidad a los precios de venta a fin de atraer clientes con ciertas ofertas que luego se compensan con compras adicionales que éstos realizan.

La volatilidad en los precios ha demostrado tener impactos en las formas de consumo. Los compradores racionales que intentan minimizar el costo visitan las tiendas más frecuentemente y compran menores cantidades en promedio a medida que la fluctuación de precios del comercio aumenta (Murthi, B.P.S et al., 2007). Asimismo, a medida que los precios relativos se modifican a través del tiempo, los consumidores generalmente tenderán a descubrir que la estrategia que minimiza lo costos requiere cambios en su canasta de compras (Shapiro M. D. & Wilcox D. W., 1996).

De esta manera, se observa que los consumidores aun contando con información limitada intentan maximizar el rendimiento de su dinero, sin embargo, un contexto inflacionario sumado a las estrategias de precios de los comercios, provocan que los consumidores no cuenten con precios relativos de referencia para tomar una decisión racional y eficiente a la hora de realizar sus consumos.

## 3. Definición de problema

En las distintas regiones geográficas de la República Argentina diversos comercios ofrecen un mismo producto en un mismo barrio. Las diversas estrategias de precios adoptadas por los comercios generan un amplio margen de variabilidad de precios respecto a un mismo producto.

A esto se le suma la inflación e inestabilidad macroeconómica de la Argentina que generan aumentos sostenidos en el tiempo, que dependiendo de cuando son trasladados a precios por distintos oferentes puede ampliar esta volatilidad.

Estos dos principales factores generan un impacto en el consumidor, provocando un desconocimiento acerca de los valores de los productos disponibles en comercios, existiendo un amplio margen de precios entre productos idénticos en igual barrio.

## 4. Objetivos del estudio

**General**  
Desarrollar una herramienta que muestre visualizaciones e insights sobre precios obtenidos a través de una página web.

**Específicos:**

* **Obtención de los datos:** Generar un algoritmo de scraping a través del cual obtener los datos del sitio <https://www.preciosclaros.gob.ar/>
* **Guardar los datos en una base de datos:** A partir de un script se almacenarán los datos en una base a fin de poder tener datos históricos y poder relevar variaciones en las tendencias a lo largo del tiempo, o poder escalar el proyecto en un futuro.
* **Obtención de insights:** Una vez almacenados se examinarán las principales variables a fin de identificar tendencias e insights.
* **Generar visualizaciones:** Una vez analizadas las variables se generarán visualizaciones y reportes a fin de que los usuarios puedan observar de modo sencillo la variación y agregar racionalidad al realizar compras.

## 5. Justificación del estudio

El desarrollo del algoritmo otorgará a consumidores de ciertas áreas geográficas determinadas la posibilidad de tener un análisis detallado de las opciones de compra en locales de cercanía. A partir de esto, se podrán obtener tendencias generales acerca del comportamiento de los precios en los distintos tipos de comercios que posibiliten una mejora en la toma de decisiones.

## 6. Definiciones

IPC: Índice de precios al consumidor.

EDLP: Every day low price. Estrategia de ventas que consiste en intentar tener el mínimo precio todos los días para cada uno de los productos.

HILO: High Low. Estrategia de ventas que consiste en tener precios mínimos en algunos productos y obtener un mayor margen en otros.

## 7. Alcances del trabajo y limitaciones

Para este estudio se consideraron dos meses de historia y una muestra de la totalidad de los supermercados disponibles. Asimismo, de cada producto se consideraron algunas marcas seleccionadas al azar a fin de acotar el análisis.

La ejecución de código se realizó de manera local, con lo que la disponibilidad de los datos quedó acotada a los usuarios de los barrios predeterminados y a la actualización local de dicha información, es decir no se podrán realizar consultas ad-hoc.

# Capítulo II: Literatura de antecedentes de estudios similares

## Introducción

Se realizó una búsqueda por internet utilizando Google Scholar (https://scholar.google.com/)

## Cómo se obtienen los precios de bienes de consumo en la actualidad

Actualmente, existen diversos sitios donde se pueden obtener los precios de los bienes de consumo de una economía. Existen diversas tiendas que publican los precios de sus productos en la web a fin de posibilitar la compra online, encontrándose la información disponible las 24 horas del día y posibilitando al comprador comparar precios de diversos comercios.

Asimismo, en la Argentina, existe el sitio web gubernamental preciosclaros.gob.ar cuyo fin es brindar información de precios de una amplia variedad de productos de consumo de diversos comercios posibilitando al consumidor maximizar su beneficio.

Estos sitios web mencionados proveen información en tiempo real del valor de los bienes que un individuo puede consumir, sin embargo, pese a tener la posibilidad de comparar precios de distintos comercios, el consumidor no tiene la posibilidad de evaluar tendencias históricas y de esa manera determinar si su elección es apropiada.

## Obtención de insights y aplicación de técnicas de machine learning a la evolución de precios

Existen trabajos de investigación sobre la predicción de precios de los activos de una economía (<http://sedici.unlp.edu.ar/handle/10915/61106>), así como de la reacción de los consumidores ante la evolución de precios (<http://tesi.luiss.it/27580/>).

Por otro lado, existen análisis históricos acerca de la evolución de los precios de una economía (https://repositorio.cepal.org/handle/11362/45012).

# Capítulo III: Metodología

## Introducción

Para el desarrollo del proyecto se utilizará el modelo CRISP-DM (Cross Industry Standard Process for Data Mining). Este modelo proporciona una descripción normalizada del proceso de un proyecto estándar de análisis de datos, cubriendo las fases, tareas y relaciones entre tareas de este.

**Etapa 1: Comprensión del negocio.**

La primera fase aglutina las tareas de comprensión de los objetivos y requisitos del proyecto con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.

**Etapa 2: Análisis de los datos.**

La segunda fase comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis.

Para la recolección de datos se realizará un algoritmo de scraping desarrollado en Python, consultando la web preciosclaros.gob.ar.

**Etapa 3: Preparación de los datos.**

En esta fase, una vez efectuada la recolección inicial de datos, se procede a la preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente.

Una vez obtenidos los datos, estos se insertarán en una base de datos de postgres, permitiendo realizar una base que posea toda la historia que queremos evaluar.

**Etapa 4: Modelado**

En esta fase se seleccionarán las técnicas de modelado más apropiadas para el proyecto de Data Mining específico.

Aquí entra el objetivo de obtención de insights, se utilizarán distintas técnicas evaluando cuáles se adaptan mejor a nuestros datos.

**Etapa 5: Evaluación**

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema.

**Etapa 6: Despliegue**

Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones.

## Técnicas

Se utilizarán gráficos de línea de tiempo, boxplots, mapas identificando los valores por región geográfica, gráficos de barras y de dispersión, tanto para el análisis exploratorio de datos como para la obtención de insights.

## Herramientas

Para la obtención de datos e inserción en una base de datos se utilizará Python, escribiendo el código en Python plano. Los paquetes a utilizar son pandas, requests, urllib, datetime, json y os para la obtención de datos mientras que para la inserción se utilizará psycopg2 además de los mencionados.

La base de datos se realizará en postgres de manera local y la visualización y EDA se desarrollará en tableau de manera local. Por último, la estimación de precios se realizará en Python utilizando los algoritmos sklearn.

## Hipótesis

Los precios de determinados productos de consumo masivo varían según localización, tipo de comercio y día de la semana en las distintas zonas del país.

### Variables disponibles asociadas a la hipótesis:

Precios: únicamente aquellos correspondientes a comercios relevados por el ente gubernamental.

Productos:

- Yerba mate de las marcas: Taragüi, Mañanita, La Merced, Cruz de Malta y Playadito.

- Aceite de girasol de las marcas: Cocinero, Lira y Cañuelas.

- Harina de trigo de las marcas: Favorita, Morixe y Caserita.

Comercios: aquellos que informan sus datos al ente de contralor de barrios seleccionados.

Área geográfica (latitud y longitud) correspondiente a comercios seleccionados de ciertos barrios.

Fecha: períodos en los que se realizó el relevamiento de los precios.

## Recolección de los datos

Como fue mencionado anteriormente, los datos se obtuvieron de la web preciosclaros.gob.ar. Dicha web posee datos de todo el país y se actualiza diariamente.

Dicha web no posee datos históricos, por lo que se desarrollo un algoritmo en Python que fue ejecutado diariamente durante el período de 60 días, a fin de poder poseer los datos históricos. El script fue desarrollado en Python plano, y devuelve archivos json separados por regiones geográficas y productos consultados.

Una vez obtenidos los archivos json, los mismos son insertados en una base de datos de postgres. La ejecución se hace en dos etapas, dado que son frecuentes las interrupciones y los errores en la web consultada, y de esta manera se pueden identificar y eliminar fácilmente registros que únicamente traen errores.

## Análisis Exploratorio de Datos (EDA)

Variables obtenidas:

|  |  |  |
| --- | --- | --- |
| Nombre del campo | Descripción | Tipo de datos |
| KEY | Hace referencia a key del json que estamos llamando | Text |
| BANDERAID | ID correspondiente a la bandera del vendedor | Text |
| LAT | Latitud geográfica donde se encuentra el vendedor | Geo |
| LONG | Longitud geográfica donde se encuentra el vendedor | Geo |
| SUCURSALNOMBRE | Nombre de la sucursal correspondiente | Text |
| ID | ID del producto seleccionado | Text |
| SUCURSALTIPO | Cada sucursal puede corresponder a un hipermercado, supermercado, Autoservicio, Mayorista o puede que no esté definido | Text |
| PROVINCIA | Nombre código correspondiente a cada una de las provincias de Argentina | Text |
| PRECIOSPRODUCTO\_PROMO1\_DESCRIPCION | Se trata de promociones por compra de volumen. En este campo se indican los detalles de la promoción o la vigencia y condiciones. | Text |
| PRECIOSPRODUCTO\_PROMO1\_PRECIO | Precio unitario correspondiente si el cliente cumple los requisitos de la promoción | Float |
| PRECIOS\_PRODUCTO\_PRECIOLISTA | Precio de lista del producto | Float |
| PRECIOSPRODUCTO\_PROMO2\_DESCRIPCION | Promociones por utilización de un medio de pago específico, aquí se indican los requisitos para cumplir la promoción y/o la vigencia. | Text |
| PRECIOSPRODUCTO\_PROMO2\_PRECIO | Precio unitario correspondiente si el cliente cumple los requisitos de la promoción | Float |
| ACTUALIZADOHOY | Indica si el campo se encuentra actualizado por parte del vendedor en la base de datos de precios claros. | Boolean |
| DIRECCION | Dirección de la sucursal | Text |
| BANDERAID | Numero de ID correspondiente a la bandera de la sucursal. | Text |
| LOCALIDAD | Nombre de la localidad en la que se encuentra ubicado el comercio | Text |
| COMERCIORAZONSOCIAL | Razón social correspondiente al establecimiento vendedor | Text |
| COMERCIOID | Número de ID correspondiente al comercio | Text |
| MARCA | Marca del producto en venta | Text |
| NOMBRE | Nombre completo del producto a la venta | Text |
| PRESENTACION | Indica las unidades en las que se vende el producto y cantidad de unidades | Text |
| FECHA | Fecha en la que tomo el registro del precio | Date |

Partiendo de los datos mencionados en la base de datos, se desarrolló un dashboard utilizando la herramienta tableau con las visualizaciones pertinentes que nos permiten obtener insights de los datos relevados. La solución planteada consulta en vivo la base de datos desarrollada, por lo que, cada vez que actualizamos la misma nuestros gráficos contemplan esta nueva información.

Luego de evaluar alrededor de 1800 comercios en todo el país. El número de comercios que brindaron la información de los productos pertinentes por día es el siguiente:

Gráfico, Gráfico de barras

Descripción generada automáticamente

**Análisis por productos**

En cuanto a la distribución por provincia, no se observan diferencias sustanciales en el país, aquí en el boxplot del universo de yerbas no encontramos precios que llamen la atención siendo todos los gráficos de las distintas provincias similares.

Sí podemos mencionar la alta presencia de outliers en provincia de Buenos Aires y Capital federal, asimismo, contrario a lo que se podría pensar, estas urbes tienen los precios más bajos del país pese a que aquí el poder adquisitivo es mayor.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Al observar el universo de los aceites relevados, vemos un gráfico similar donde también vemos que los precios más bajos están en Capital Federal y Provincia de Buenos Aires. Es válido mencionar el outiler en Provincia de Buenos Aires con un precio de $669, más de un 30% por encima de los máximos subsiguientes:

Gráfico

Descripción generada automáticamente

Por último, observamos que el universo de harinas repite el mismo patrón, curiosamente vemos que la cadena que tenía el precio máximo en aceite tiene un outlier por precios bajos en harina. Esto nos hace pensar que esta cadena podría estar utilizando una estrategia de precios High-Low:

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

En cuanto a la evolución de los precios, se consideró la evolución tanto de los precios de lista como de las promociones. El primer punto a destacar es que la totalidad de los productos relevados sufrieron al menos un aumento en el período observado a excepción de la yerba Playadito.

El aumento promedio de la canasta relevada fue de un 13%. Observando por grupo de productos, vemos que las harinas aumentaron un 19% promedio mientras que los aceites lo hicieron en un 14% y por último las yerbas un 8%. El producto con mayor variación de precios al alza fue la harina Morixe con un incremento de un 32% en el período observado.

Asimismo, se observa una alta volatilidad en los precios, con productos con hasta seis aumentos en el período observado y con un rango elevado de precios por localidad:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Análisis por marca**

A continuación se muestran las distribuciones de los precios por marca de cada uno de los tipos de productos relevados. A partir del análisis, vemos que para un mismo día tomado como muestra no hay gran dispersión entre los precios. De un mismo producto.

**Yerbas**

**Cruz de malta**

**Gráfico

Descripción generada automáticamente con confianza media**

**La merced:**

**Aplicación

Descripción generada automáticamente con confianza baja**

Mañanita

Imagen que contiene Tabla

Descripción generada automáticamente

Playadito

Texto, Aplicación

Descripción generada automáticamente

Taragüi

Gráfico

Descripción generada automáticamente

Aceites

Cañuelas

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cocinero

Gráfico

Descripción generada automáticamente con confianza media

Lira

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Harinas

Caserita

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Favorita:

Gráfico

Descripción generada automáticamente

Morixe

Texto, Aplicación

Descripción generada automáticamente con confianza media

# Capítulo IV: Investigación

## Modelado

A partir del análisis exploratorio de datos encontramos que no hay gran dispersión en los precios de los productos relevados en un mismo día en las distintas zonas si los agrupamos por marca, no así cuando los agrupamos por tipo de producto. Por lo tanto, para realizar una evaluación de la evolución de los precios a nivel país se utilizó el precio de lista promedio por marca.

En primer lugar, se buscó relacionar la variación de los precios con la variación del dólar blue y comprender si poseíamos una correlación entre dichas variables.

Evolución cotización dólar blue:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## Conclusión

# Capítulo V: Resumen, resultados e implicaciones

## Introducción

## Resultados

## Implicancias

## Sugerencias para futuras investigaciones

# Referencias bibliográficas

1. Bell, D. R., & Lattin, J. M. (1998). Shopping behavior and consumer preference for store price format Why large basket shoppers prefer EDLP. *Marketing Science*, Vol. 17 No. 1, 66-88. doi: [10.1287/mksc.17.1.66](https://doi.org/10.1287/mksc.17.1.66).
2. Shapiro M. D. & Wilcox D. W. (1996). Mismeasurement in the consumer price index: an evaluation. *National bureau of economic research*, w.p. 5590, 94-142. doi: 10.3386/w5590.
3. Smith, P. A. (2021). Estimating Sampling Errors in Consumer Price Indices. *International Statistical Review*, Vol. 89 No. 3, 481-504. doi:10.1111/insr.12438.
4. Murthi, B.P.S, Haruvy, E., & Zang, H. (2007). The Effect of Variability in Price on Consumer Brand Choice. *Review of Marketing Science,* Vol. 5 No. 1, 1-24. doi: [10.2202/1546-5616.1071](https://doi.org/10.2202/1546-5616.1071).
5. Mohiuddin, A., Raihan S. & Syed M. S. I. (2020). The k-means Algorithm: A comprehensive survey and performance evaluation. *Electronics 2020*, 9, 1295; doi:10.3390.