Predicción de la elasticidad de la demanda de un grupo de referencias del negocio de papel higiénico de una cadena de supermercados de Estados Unidos de América.

Estudio de caso

# Parte 1

# **Comprensión del Negocio**

Este análisis busca abordar varios objetivos de negocio clave relacionados con la comprensión y la predicción del impacto de las estrategias de precios en la demanda de productos.

1. Cuantificación Precisa de la Efectividad Promocional: Un objetivo principal es cuantificar con precisión la efectividad de las actividades promocionales para impulsar la demanda de productos. Este análisis buscó desarrollar una metodología más robusta para identificar y medir el verdadero impacto de las promociones.
2. Capacidades mejoradas de pronóstico de la demanda: Este análisis busca mejorar las capacidades de pronóstico de la demanda incorporando una comprensión más completa de la elasticidad precio. Al aprovechar datos sintéticos para ampliar el alcance temporal del conjunto de datos y abordar las limitaciones relacionadas con la dinámica del mercado, el análisis facilitará predicciones más precisas de la demanda futura en diversos escenarios de precios. Esto facilitará una mejor gestión del inventario, la planificación de la producción y la optimización general de las ventas.
3. Reducción de la incertidumbre en la toma de decisiones: Al proporcionar una comprensión más precisa y completa de la elasticidad precio, este análisis busca reducir la incertidumbre en la toma de decisiones relacionadas con precios y promociones. Esto permitirá a los gerentes tomar decisiones más informadas basadas en información basada en datos, lo que se traduce en mejores resultados comerciales.

# **Comprensión de los Datos**

El presente trabajo de monografía tiene como objetivo analizar la elasticidad precio de la demanda de productos de papel higiénico (Bathroom Tissue) utilizando el conjunto de datos Dominicks-Manual-and-Codebook\_KiltsCenter. Este dataset, proporcionado por el Kilts Center for Marketing de la Universidad de Chicago, contiene información real detallada sobre precios, volúmenes de venta y características de productos comercializados en la cadena de supermercados Dominick's.

El conjunto de datos cuenta con 2 conjuntos de datos:

* UPC (Universal product code) corresponde datos de información de los productos, códigos, descripción, tamaño. Contiene los datos maestro (upctti.cvs) tiene 128 filas y 6 columnas con un peso de 7KB
* Archivo de movimientos, contiene data de ventas semanales por UPC, de 5 años, precios, promociones, margen de ganancia, tienda, otros. Cuenta con 11 columnas y 1159016 filas.

Descripción de las variables:

* UPC: universal Product code numérico 8
* STORE: número de tienda numérico 3
* WEEK: número de semana numérico 3
* MOVE: Cantidad vendida numérico 8
* PRICE: Precio retail numérico 8
* QTY: Número de articulos juntos numérico 3
* PROFIT: Margen de ganancia numérico 8
* SALE: código de promoción (B(Bono),C(Cupón),S(reducción simple de precio)) Character 8
* OK: 1 para datos validos, 0 datos sospechosos numérico

Las variables numéricas cuentan con las siguientes características:

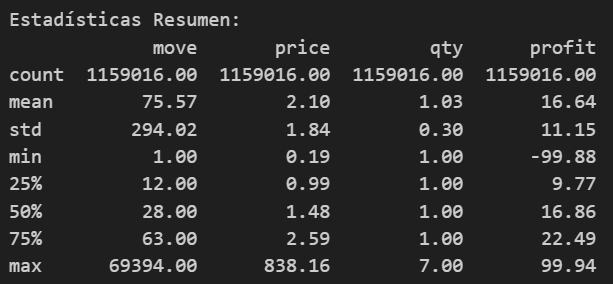


Ilustración 1 - Resumen de estadísticas descriptivas de variables numéricas

El conjunto de datos original presentaba limitaciones debido a su antigüedad y posible incompletitud en la captura de los efectos promocionales. Para abordar estas limitaciones y aumentar el conjunto de datos para un análisis más robusto, se implementó un proceso de síntesis de datos. Este proceso implicó la generación de datos sintéticos que reflejaran las propiedades estadísticas del conjunto de datos original, extendiendo así el alcance del conjunto de datos, particularmente en términos de cobertura temporal.

El proceso de síntesis se llevó a cabo utilizando Python y las bibliotecas pandas, numpy y scikit-learn. Los pasos clave de la metodología de síntesis de datos se describen a continuación:

* Inicialización del DataFrame Sintético. Se inicializó un DataFrame de pandas (synthetic\_data) con 1’000,000 filas para servir como base para los datos sintéticos.
* Generación de Variables Categóricas. Las variables categóricas *upc, sale y store* se generaron mediante el muestreo de las distribuciones empíricas observadas en el conjunto de datos original (*df\_move\_cleaned*). Esto aseguró que los datos sintéticos retuvieran la composición categórica de los datos reales.
* Generación de la Variable Temporal. Para extender el alcance temporal de los datos, la variable *week* se generó utilizando una distribución aleatoria uniforme, comenzando desde la semana inmediatamente posterior a la última semana en el conjunto de datos original. Este enfoque simuló semanas futuras, permitiendo el análisis de tendencias más allá del marco temporal del conjunto de datos original.
* Generación de Variables Numéricas. Las variables numéricas *move, price, qty y profit* se generaron condicionalmente en función de la variable sale y, en algunos casos, la variable week. Esta generación condicional tenía como objetivo capturar las relaciones entre las actividades promocionales y estas métricas clave.
* Se emplearon distribuciones de probabilidad específicas (p. ej., gamma, log-normal, normal) para simular la variabilidad y las características de distribución de estas variables. Los parámetros de estas distribuciones a veces se ajustaron para reflejar posibles tendencias temporales.
* Concatenación de Conjuntos de Datos. El conjunto de datos sintético (*synthetic\_data*) se concatenó con el conjunto de datos original (*df\_move\_cleaned*) para crear un conjunto de datos combinado (*combined\_data*). Este conjunto de datos combinado se utilizó para análisis posteriores.

# **Preparación de los Datos**

El conjunto de datos se sometió a varios pasos de preprocesamiento para garantizar su idoneidad para el análisis. Estos pasos incluyeron la limpieza de datos, el manejo de valores faltantes y la síntesis de datos.

* Limpieza de Datos. Se eliminaron las columnas irrelevantes o redundantes del conjunto de datos para agilizar el análisis y reducir la dimensionalidad. Se abordaron las inconsistencias en las variables categóricas estandarizando las representaciones. Específicamente, los valores nulos en la variable ‘*sale*’, inicialmente representados como '*nan*' y '*NaN*', se unificaron bajo la etiqueta "no\_aplica".

Se hizo una identificación inicial de valores extremos que permitió identificar mejor las distribuciones de las principales variables.

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 2.Distribución de variable price antes de limpieza

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 3 - Distribución de variable price después de limpieza

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 4 - Distribución de variable move (Cantidad ventas) después de limpieza

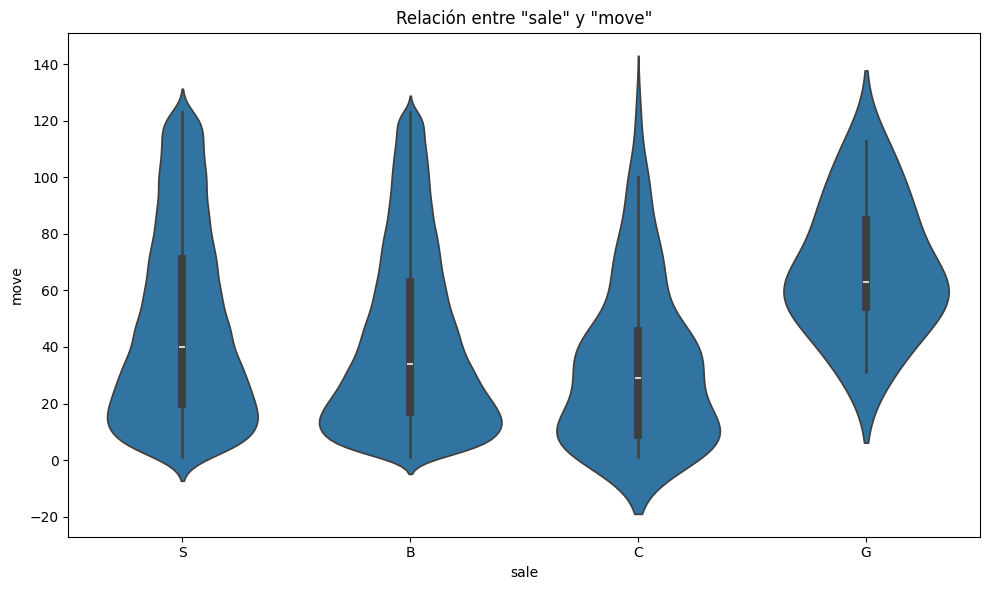
* Manejo de Valores Faltantes. Se tuvo 2 casos: para las ventas que registraban un valor cero y precio cero se determinó eliminar estos datos y para la variable sale donde el campo era vacío se trataron como una categoría separada.
* Síntesis de Datos. Para aumentar el conjunto de datos y abordar las limitaciones relacionadas con su alcance temporal, se generaron datos sintéticos utilizando una metodología descrita en la comprensión de los datos. Este proceso implicó el muestreo de las distribuciones de los datos originales para crear nuevos puntos de datos que preservaran las propiedades estadísticas del conjunto de datos original al tiempo que extendían su cobertura temporal.

Ilustración 5 - relación entre variable categórica sale(tipo de promoción) y move (cantidad de ventas)

* + La estategia promocional G tiene el mayor número de movimientos(ventas) y también una mediana más alta.
  + Las estategias promocionales S y B presentan distribuciones similares: más concentradas entre 10 y 60 movimientos(ventas), aunque con algunos valores altos.
  + La estategia promocional tiene una distribución más ancha en la parte inferior, lo que indica más valores bajos de "move" (ventas).Gráfico

    Descripción generada automáticamente

Ilustración 6 - relación entre variable categórica sale(tipo de promoción) y profit (ganancias)

* + Las estrategias promocionales S y B muestran una dispersión mayor y presentan valores negativos, lo cual indica que hay pérdidas asociadas a estas ventas.
  + C y G tienen distribuciones más concentradas, especialmente C, lo que sugiere ganancias más estables.
  + La estrategia promocional G presenta los valores más altos de "profit", lo cual puede indicar que es la venta más rentable.

Gráfico

Descripción generada automáticamente

Ilustración 7 - relación entre variable categórica sale(tipo de promoción) y precio.

* + Las estrategias promocionales S y B presentan una distribución más amplia en precios, con algunos valores extremos (precios altos).
  + La estrategia C parece tener precios más bajos y estables, concentrados en torno a 1.
  + G también tiene precios bajos y poco dispersos, pero levemente superiores a C.

También se realizaron análisis de las tendencias para asegurarse que los datos sintéticos siguieran estas mismas tendencias de las variables numéricas.

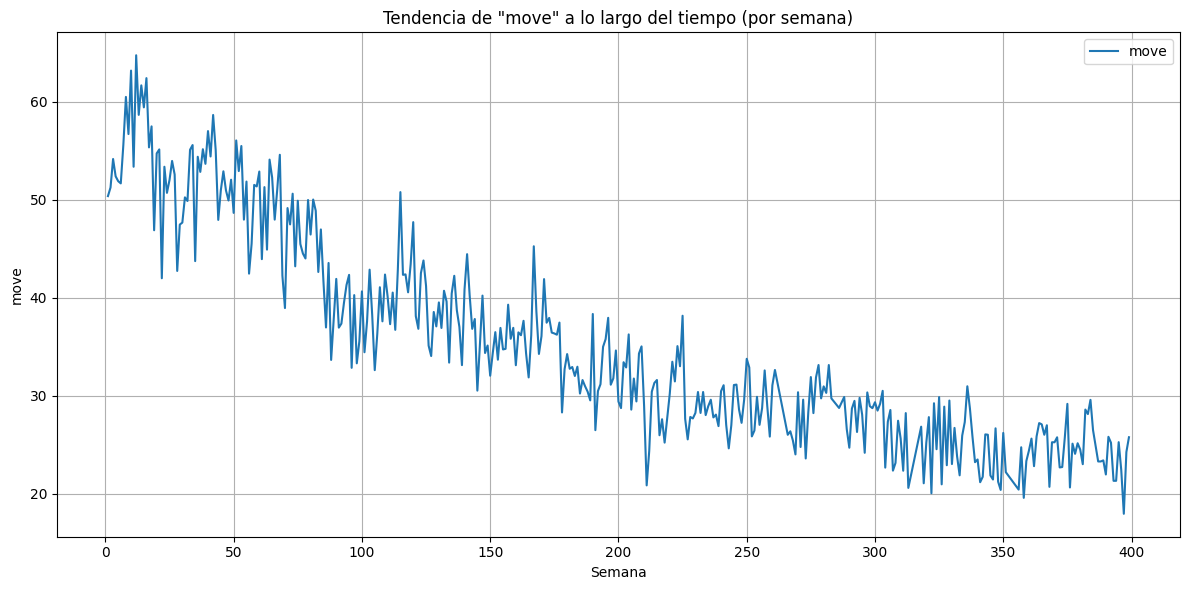


Ilustración 8 - Tendencia de las ventas a lo largo del tiempo

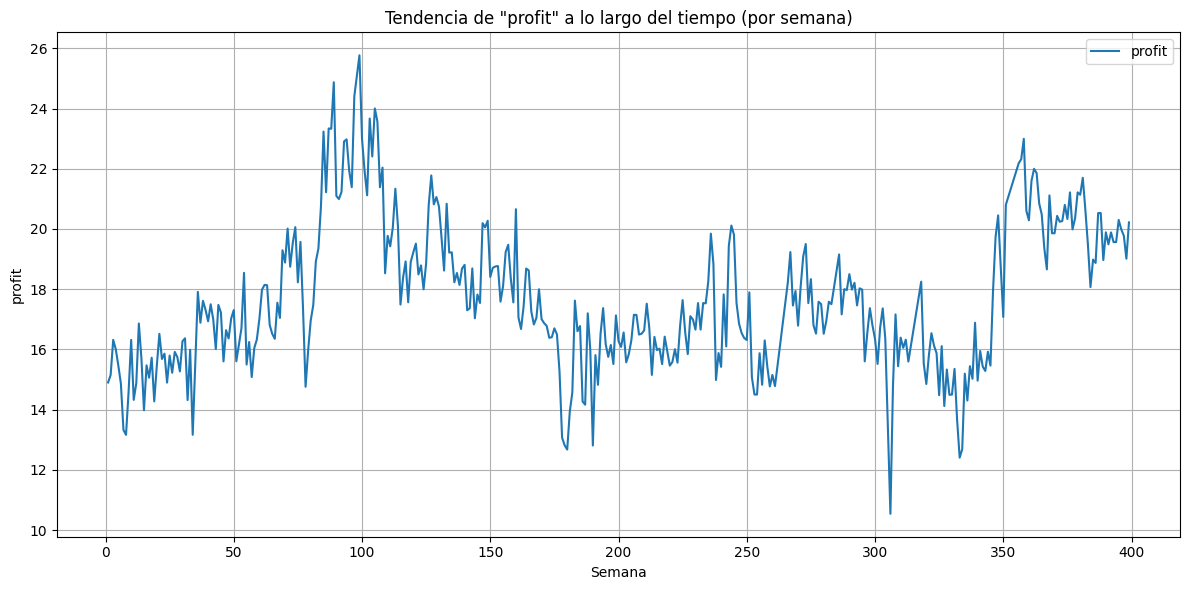


Ilustración 9 - Tendencia de las ganancias a lo largo del tiempo

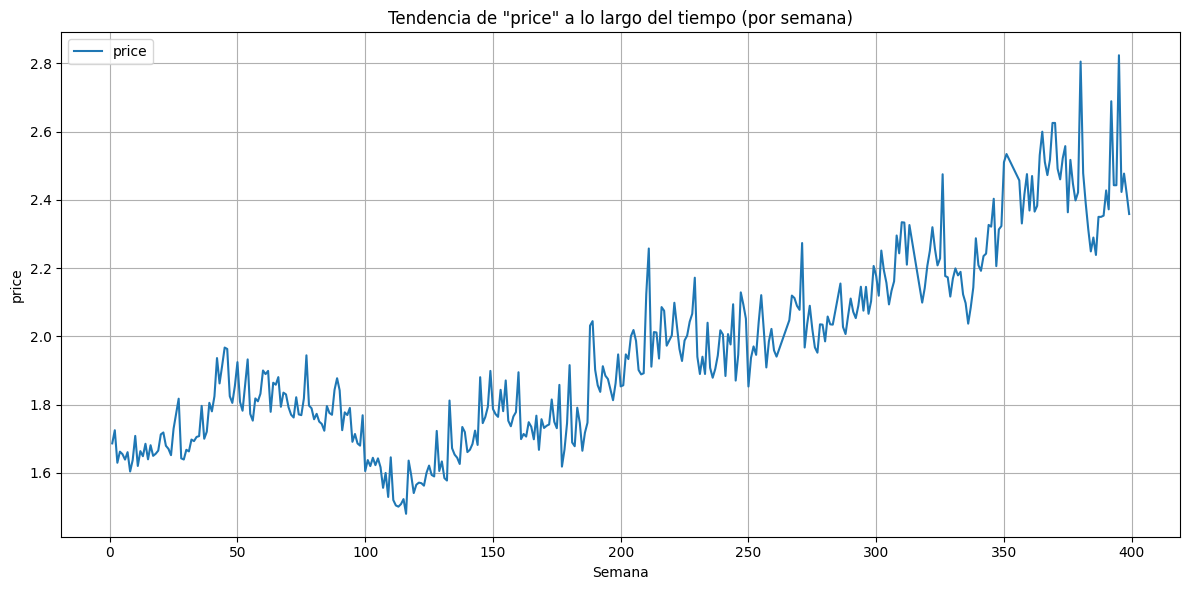
* + 

Ilustración 10 - Tendencia del precio a lo largo del tiempo

* Análisis de Valores Atípicos. Se utilizaron diagramas de caja para identificar visualmente posibles valores atípicos en las variables numéricas (*move, price, qty, profit*). Si bien se observaron valores atípicos, se tomó la decisión de retenerlos en el conjunto de datos, considerando su potencial relevancia para capturar eventos extremos o variabilidad en los datos.
* Transformación de Variables. Se calculó una nueva variable, value, que representa el valor total de las ventas, multiplicando *price, move y profit*. Las variables numéricas se estandarizaron utilizando el *StandardScaler* de la biblioteca *scikit-learn*. Esta transformación escaló las variables para tener una media de 0 y una desviación estándar de 1, facilitando la comparación y mejorando el rendimiento de ciertas técnicas analíticas.