

Informe Predictivo de Precios en el Mercado Editorial Argentino









Caso de Estudio: Catálogo Yenny–El Ateneo

Autor: Marcos Praga

Área: *Ciencia de Datos*

Fecha: Octubre de 2025

Contenido :

 Objetivo y Alcance del Proyecto.....	3
 Usuario Final y Nivel de Aplicación.....	4
 Objetivos del Análisis Exploratorio.....	5
 Preguntas / Hipótesis que Queremos Responder.....	6
 Resumen de Metadata.....	7
Estructura general del dataset:.....	7
Resumen técnico:.....	8
 Insights más relevantes.....	8
Análisis Univariado.....	8
Análisis Bivariado.....	12
Análisis Multivariado.....	14

Objetivo y Alcance del Proyecto

El objetivo principal de este proyecto es analizar los factores que influyen en la formación de precios dentro del mercado editorial argentino y desarrollar un modelo predictivo capaz de estimar el precio óptimo de un libro en función de sus características editoriales (como número de páginas, tipo de encuadernación, género literario, y sello editorial). Este análisis busca aportar información estratégica para la toma de decisiones en pricing dinámico, permitiendo a editoriales y cadenas de librerías optimizar sus márgenes de rentabilidad, definir precios competitivos y mejorar la gestión de inventarios.

El alcance del proyecto abarca la exploración, limpieza y análisis descriptivo del catálogo público de Yenny–El Ateneo, identificando patrones, correlaciones y segmentos de mercado que expliquen la variabilidad de precios. A partir de estos hallazgos, se establecen las bases para un modelo de machine learning orientado a predecir el precio promedio de nuevos títulos, incluso cuando sólo se dispone de información parcial sobre sus características. En conjunto, los resultados ofrecen una herramienta de apoyo a la estrategia comercial del sector editorial, promoviendo una gestión más eficiente y basada en datos.

Usuario Final y Nivel de Aplicación

El usuario final de este proyecto está compuesto por gerentes comerciales y de producto de editoriales, así como por responsables de pricing, planeamiento y gestión de inventario en librerías y cadenas de retail. Estos perfiles requieren herramientas que faciliten la toma de decisiones estratégicas basadas en datos, especialmente en lo que respecta a la definición de precios, posicionamiento de catálogo y planificación de stock.

El nivel de aplicación del análisis se ubica en el ámbito estratégico y operativo:

- A nivel estratégico, los resultados del modelo permiten comprender las dinámicas de precios en el mercado editorial argentino, optimizar políticas de precios y detectar oportunidades de posicionamiento competitivo entre sellos o categorías.
- A nivel operativo, el análisis ofrece una base para implementar modelos predictivos de pricing dinámico, capaces de estimar precios óptimos de nuevos títulos o ajustar los existentes según las características del producto, la demanda y la competencia.

En síntesis, el proyecto brinda un apoyo analítico integral que puede incorporarse tanto en las decisiones comerciales de alto nivel (como la planificación de lanzamientos y estrategias editoriales), como en las operaciones cotidianas de retail y distribución, potenciando la eficiencia, la rentabilidad y la coherencia del pricing dentro del ecosistema editorial.



Objetivos del Análisis Exploratorio

Durante la etapa de análisis exploratorio se busca comprender la estructura y comportamiento del catálogo editorial a fin de identificar patrones, tendencias y relaciones que permitan explicar la formación de precios en el mercado.

Los objetivos específicos del EDA son:

1. **Analizar la distribución general de precios** en el catálogo para identificar rangos predominantes, dispersiones y valores atípicos que puedan influir en las estrategias de pricing.
2. **Examinar la composición del catálogo por categorías y géneros literarios**, determinando cuáles concentran mayor volumen de títulos y cómo se comportan en términos de precios promedio.
3. **Evaluar la relación entre el número de páginas y el precio de venta**, explorando si existe una correlación significativa entre ambas variables.
4. **Investigar el impacto del tipo de encuadernación** (rústica, tapa dura, etc.) sobre el precio final, considerando su efecto en la percepción de valor y costos de producción.
5. **Detectar posibles combinaciones de variables** (por ejemplo, género + encuadernación + editorial) que expliquen variaciones significativas en el precio.
6. **Identificar segmentos de mercado y patrones comunes** que puedan servir como base para el desarrollo del modelo predictivo de precios.

En conjunto, estos objetivos buscan construir una visión integral y fundamentada del mercado editorial, que permita transformar los datos en información útil para la toma de decisiones comerciales y la futura implementación de un modelo de predicción de precios eficiente y escalable.

? Preguntas / Hipótesis que Queremos Responder

El análisis parte de un conjunto de hipótesis orientadas a comprender qué factores explican las variaciones en los precios de los libros dentro del catálogo editorial. Estas preguntas guían las etapas del EDA (Exploratory Data Analysis) y preparan el terreno para el desarrollo del modelo predictivo.

A continuación, se detallan las principales preguntas de análisis, junto con el tipo de enfoque aplicado en cada caso:

1. Análisis Univariado:

- ¿Cuál es la distribución de precios en el catálogo?
- ¿Qué categorías y géneros literarios dominan el mercado?
- ¿Cómo se distribuye el número de páginas de los libros?
- ¿Cuál es la distribución de tipos de encuadernación?

2. Análisis Bivariado:

- ¿Existe correlación entre el número de páginas y el precio?
- ¿Qué categorías literarias tienen los precios más altos en promedio?

3. Análisis Multivariado:

- ¿Existen interacciones entre variables que impacten en el precio?
- ¿Cuál es el índice de correlación entre las variables?
- ¿Se pueden identificar segmentos de mercado con características similares?

Resumen de Metadata

El conjunto de datos utilizado en este análisis proviene del **catálogo público de Yenny–El Ateneo**, correspondiente a libros disponibles en su tienda online. La información fue obtenida de una fuente abierta y estructurada en formato tabular, con el objetivo de explorar los factores que influyen en la formación de precios dentro del mercado editorial argentino.

El dataset está compuesto por **15.642 registros** (libros únicos) y **13 variables** que describen distintas características de cada título, abarcando aspectos editoriales, físicos y económicos.

Estructura general del dataset:

Tipo de dato	Variables incluidas	Ejemplos / Descripción
Numéricas	nro. de páginas, precio	Extensión total del libro y valor de venta en pesos argentinos.
Categorías	editorial, idioma, encuadernación, categoría, género, subgénero, título, autor	Información descriptiva que caracteriza el libro y su posición en el catálogo, identificación de la obra y su autoría.
Fechas	fecha_publicacion	Año de publicación del título.
Identificadores y URLs	código isbn, url	Información única para cada ejemplar y enlace al sitio de venta.

Resumen técnico:

- **Cantidad de filas:** 15.642
- **Cantidad de columnas:** 13
- **Tipos de datos:** `object` (10), `int64` (2), `float64` (1)
- **Campos con valores faltantes:** la variable `nro_paginas` presenta algunos registros nulos (~9 % del total).



Insights más relevantes

Análisis Univariado

Para responder a la primera pregunta realizamos en la notebook algunos cálculos para entender cómo se comporta la variable target:

- **¿Cuál es la distribución de precios en el catálogo?**

Media: \$25.548,27

Mediana: \$23.900

Moda: \$7.500

Desv. Estándar: \$14.561,17

Mínimo: \$1.050

Máximo: \$367.080

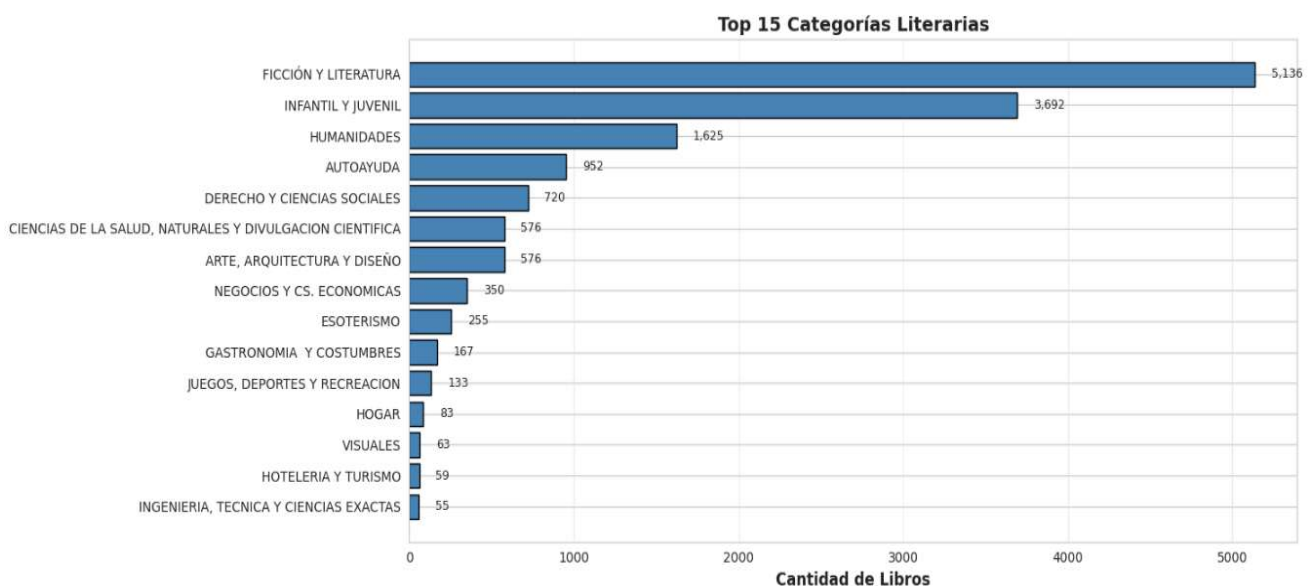
Rango Intercuartil (IQR): \$15.600

Esto nos daría a entender inicialmente que hay mucha amplitud entre los valores extremos de los precios pero a su vez la media y la mediana generaron una tendencia hacia la derecha en la campana que comprobamos graficando:



Con respecto a la segunda pregunta decidimos optar por un gráfico de barras horizontales teniendo en cuenta que los datos a comparar son categóricos:

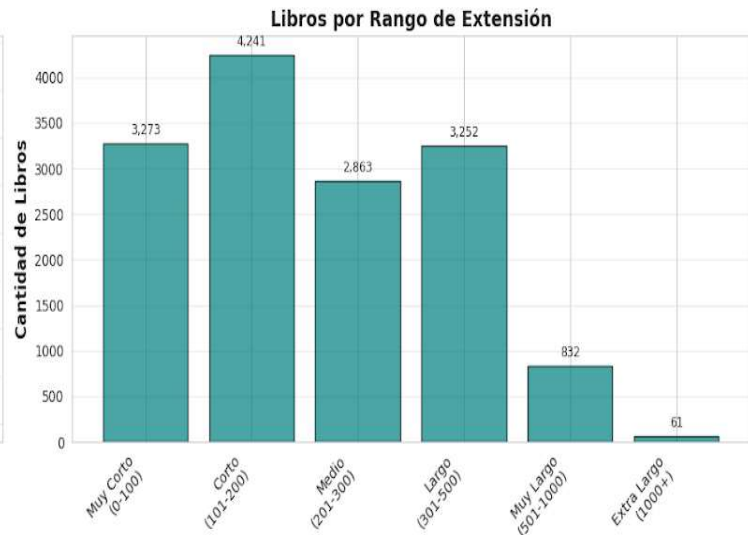
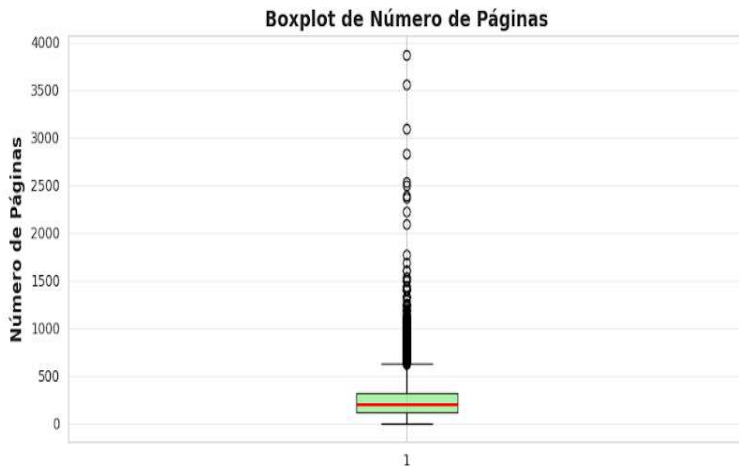
- **¿Qué categorías y géneros literarios dominan el mercado?**



Entendemos a partir de esto que el top 3 de este resultado lo lideran los libros de ficción y fantasía, los infanto juveniles, y los libros de humanidades mientras que los que tienen menor cantidad en el catálogo son los de ingeniería, técnica y ciencias exactas, hotelería y turismo y los visuales.

- ¿Cómo se distribuye el número de páginas de los libros?

Para responder esto decidimos hacer dos gráficos:

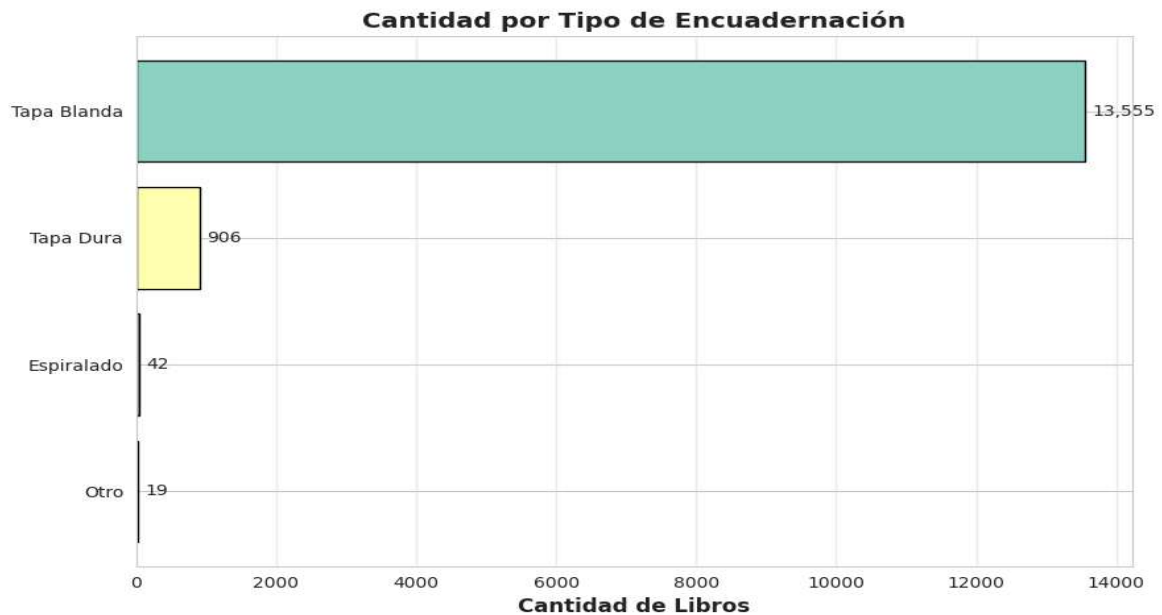


Podemos deducir que hay una **distribución asimétrica** de los datos, la mayoría de los libros tienen un número de páginas relativamente bajo, concentrándose por debajo de 500 páginas, mientras que hay unos pocos libros con un número de páginas extremadamente alto, lo que genera **outliers**. Un comportamiento muy similar al del precio.

Presencia de outliers: Existen varios libros con más de 1000 páginas, algunos incluso cercanos a 4000 páginas pero son pocos. Esto indica que hay títulos excepcionalmente largos que no representan el comportamiento general.

- **¿Cuál es la distribución de tipos de encuadernación?**

Este fue un insight con el resultado menos útil porque fue muy fácil de predecir pero no por ello menos importante porque ayuda a confirmar una información que todos suponemos desde el principio:

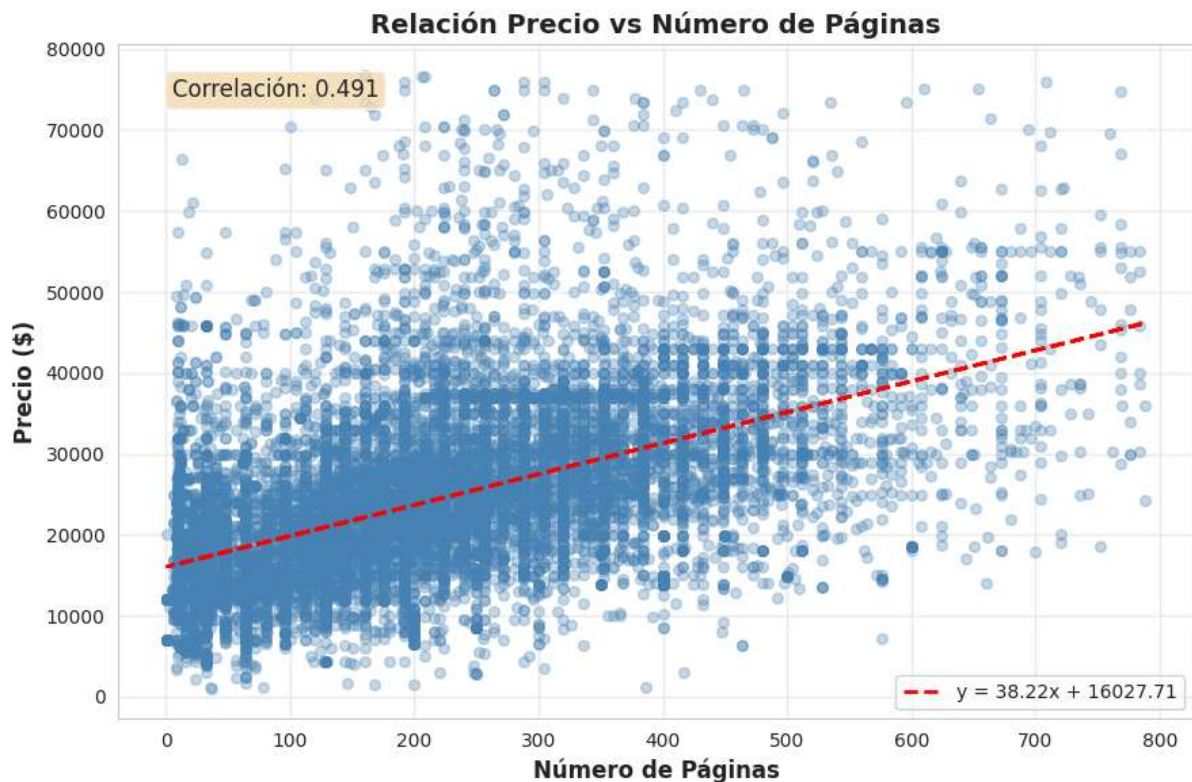


Los libros de tapa blanda son los más masivos y comunes de todos.

Análisis Bivariado

- ¿Existe correlación entre el número de páginas y el precio?

Respondemos entonces de manera definitiva a la duda que nos llevó a eliminar los datos nulos del dataset inicial:

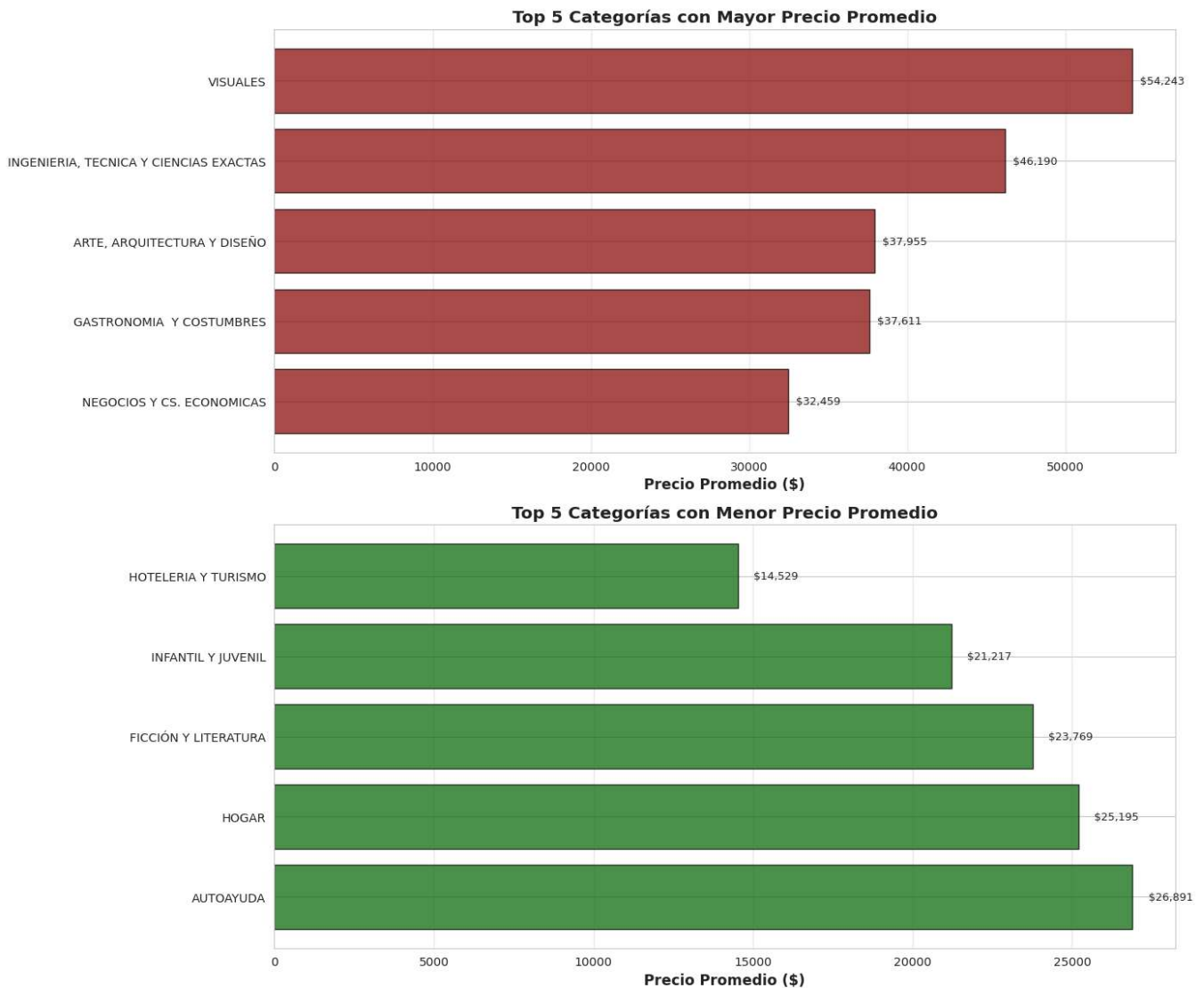


La relación es directamente proporcional con una correlación positiva de casi 0.5. A mayor número de páginas, mayor es el precio, pero no son variables completamente dependientes.

Todavía hay un gran margen de diferencia en el precio que no puede ser explicado solo con la cantidad de páginas que tiene.

- ¿Qué categorías literarias tienen los precios más altos en promedio?

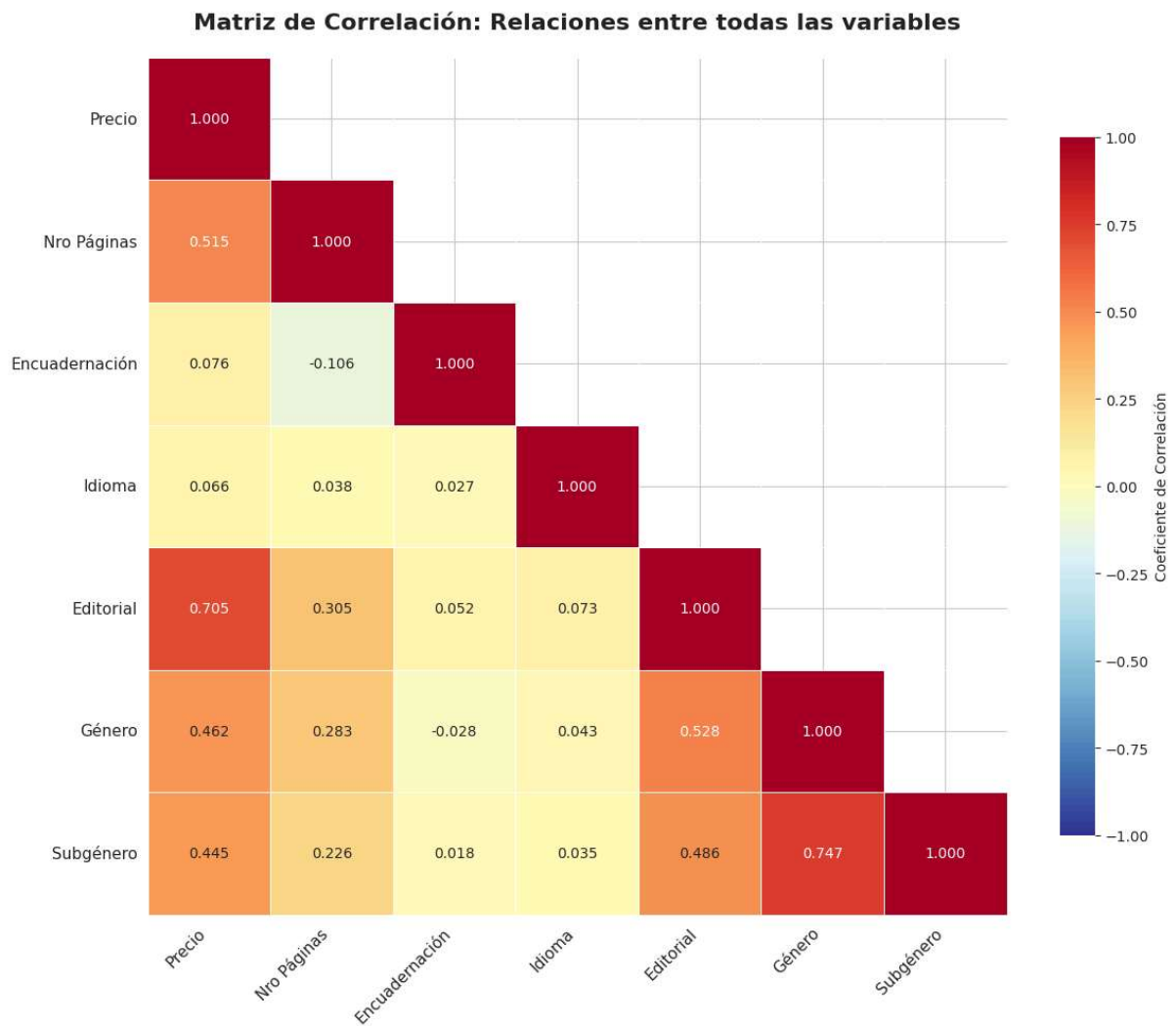
Para responder esto hicimos dos rankings expresados en gráficos de barras:



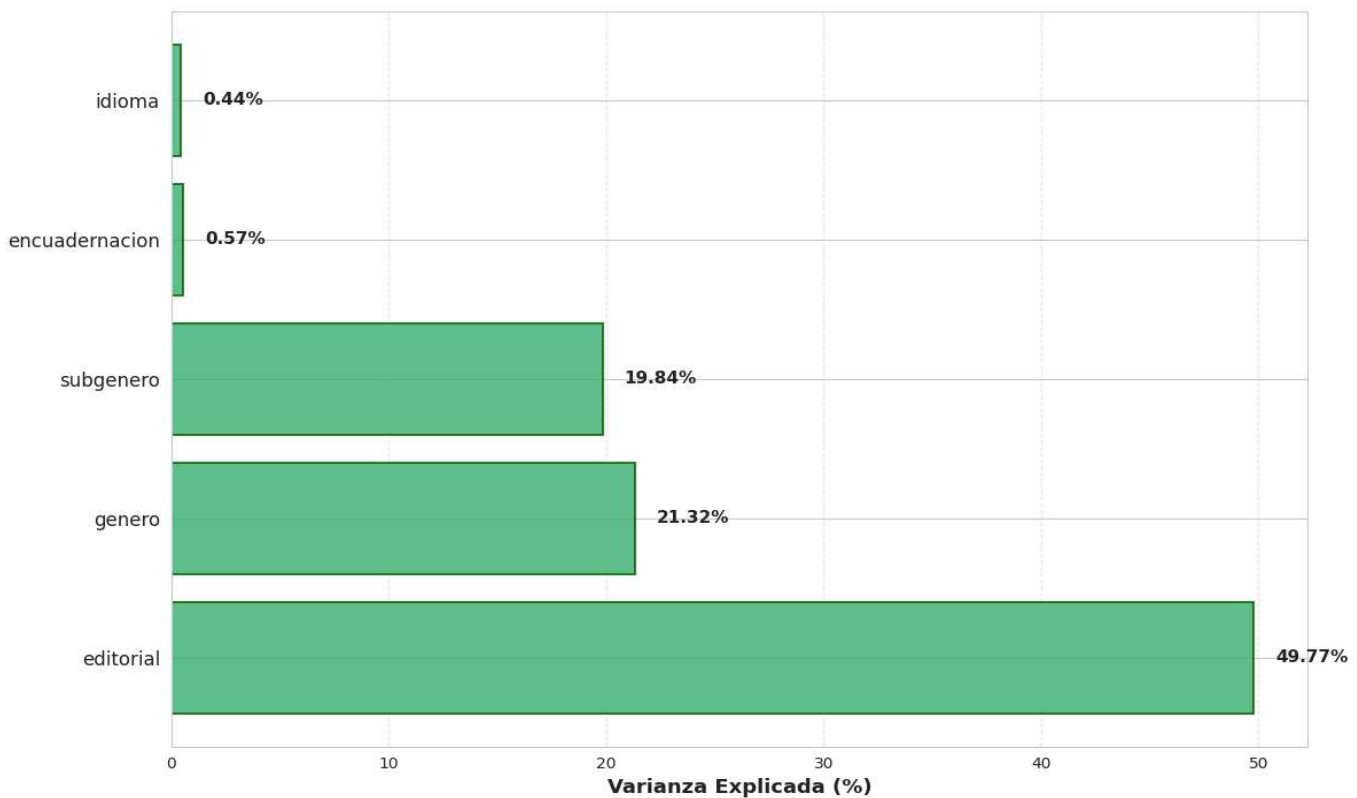
En este gráfico en si no vemos nada muy relevante pero si volvemos al gráfico de categorías con más cantidad de libros podemos entender que hoteleria y turismo a pesar de ser el más barato no hay mucha cantidad, lo cual puede decirnos que hay un nicho de marketing. Y por otro lado los de ficción y los infanto juveniles son muy baratos y hay gran cantidad, al contrario de los más visuales que son más caros y a su vez hay menos cantidad, lo que quiere decir que puede haber otras variables como el tipo de papel (los visuales y de arte suelen ser más plastificados) o la cantidad de imágenes que no estamos pudiendo relacionar y que podrían estar afectando a la variable target.

Análisis Multivariado

Para el análisis multivariado decidimos mostrar cual de todas las variables tiene una relación más directa con la definición del precio en el dataset y a su vez cuales se relacionan mas entre si. Para ello graficamos en base a la varianza y al precio promedio:



Poder Explicativo de cada Variable sobre el Precio



En estos gráficos podemos observar en primera instancia el top 3 de variables relacionadas con el precio:

- **Editorial:** 0.705 [FUERTE positiva]
- **Nro Páginas:** 0.515 [FUERTE positiva]
- **Género:** 0.462 [MODERADA positiva]

Y también el top 3 de relaciones entre las demás variables:

- Subgénero ↔ Género: 0.747 (**positiva**)
- Precio ↔ Editorial: 0.705 (**positiva**)
- Editorial ↔ Género: 0.528 (**positiva**)