Universidad Peruana de Ciencias Aplicadas

INFORME DEL TRABAJO FINAL

CURSO DE DATA VISUALIZATION

Alumnos:

Francesca Nicole Bances Torres

Marsi Figueroa Larragan

Julio 2025

1. Resumen

En el presente trabajo se analizan los datos de los libros más vendidos en Amazon entre 2009 y 2019, utilizando el dataset *"Amazon Top 50 Bestselling Books"*. El análisis se llevó a cabo en un entorno colaborativo (*Google Colab*), donde se aplicaron procesos de limpieza, transformación y visualización de datos.

Con base en este análisis, se plantearon tres hipótesis que buscaban explorar relaciones entre el precio, la calificación, el género y la permanencia de los libros en el ranking de ventas. A través de diversas visualizaciones, se identificaron patrones que permitieron validar o rechazar dichas hipótesis, obteniendo así una comprensión más profunda del comportamiento de los bestsellers en Amazon.

Este proyecto no solo permitió aplicar herramientas de análisis exploratorio, sino que también fomentó el desarrollo del pensamiento crítico y la interpretación de datos con base en evidencia visual.

TABLA DE CONTENIDOS

2. Objetivo del Estudiante (Student Outcome)

El objetivo del estudiante es desarrollar la capacidad para comprender y brindar soporte en la entrega de sistemas de información dentro de un entorno tecnológico. Este trabajo contribuyó significativamente al fortalecimiento de dicha competencia a través del análisis de los datos proporcionados. Aplicamos técnicas de limpieza y visualización que nos permitieron explorar la información con mayor profundidad y validar diversas hipótesis planteadas.

El proceso inició con la exploración del dataset "bestsellers_with_categories.csv", donde se revisó la estructura de los datos y se identificaron posibles valores nulos o duplicados. Posteriormente, se realizaron representaciones gráficas que facilitaron una mejor comprensión del comportamiento de las variables.

Se evaluaron tres hipótesis relacionadas con la relación entre el precio y la calificación, el género y la calificación, así como la permanencia de los libros en el ranking de los más vendidos. Estas hipótesis fueron analizadas mediante técnicas estadísticas y visualizaciones, promoviendo el desarrollo del pensamiento crítico y la capacidad analítica.

En conclusión, este trabajo nos permitió aplicar de forma práctica nuestras habilidades en la preparación, manipulación y visualización de datos. Estas competencias son fundamentales para la toma de decisiones informadas y constituyen una base sólida para el desarrollo profesional en entornos guiados por los datos.

3. Misión del trabajo

Desarrollar un dashboard visual interactivo, claro y funcional que permita comunicar de manera efectiva los patrones, tendencias y relaciones significativas encontradas en el análisis de los libros más vendidos en Amazon entre los años 2009 y 2019. Para ello, se aplicó el ciclo completo de desarrollo de visualización de datos: desde la comprensión del problema y exploración del dataset, hasta la limpieza, transformación, selección de los gráficos adecuados y validación de los resultados. Este trabajo busca no solo mostrar información descriptiva, sino también

generar conocimiento útil que facilite la toma de decisiones basada en datos, enfocándose en elementos como calificaciones, precios, permanencia en rankings, géneros y autores. Asimismo, se fomenta el pensamiento crítico, el storytelling visual y la capacidad de síntesis, habilidades esenciales en un entorno profesional de análisis de datos.

4. Objetivo de la presentación

Presentar visualizaciones claras, intuitivas y validadas que permitan a la gerencia de Amazon identificar patrones clave y evaluar nuevas oportunidades de negocio relacionadas con el fomento de la lectura digital en mercados estratégicos. El enfoque principal de esta presentación es responder y sustentar, mediante análisis visuales, las siguientes hipótesis:

- H1: Los libros más baratos tienden a recibir mejores calificaciones.

- H2: Existen géneros que consistentemente obtienen buenas calificaciones, destacando especialmente la ficción.

- H3: Una buena calificación sostenida en el tiempo contribuye a que un libro permanezca más años en el ranking de los más vendidos.

Cada visualización ha sido diseñada para apoyar la interpretación de estas hipótesis, permitiendo extraer conclusiones respaldadas por datos históricos del periodo 2009–2019. Este enfoque busca facilitar la toma de decisiones basada en evidencia, optimizando las futuras estrategias de promoción y curaduría de libros en formato digital.

5. Contexto

Amazon está considerando renovar y reforzar sus estrategias comerciales relacionadas con la venta de libros, particularmente en formatos digitales. Esta decisión responde al crecimiento sostenido del consumo de contenido digital y al interés por entender mejor los hábitos de lectura de sus clientes.

La visualización de datos es el proceso de representar información de manera gráfica para facilitar su comprensión y análisis. Según DataScientest (2023), su objetivo principal es permitir que los usuarios identifiquen patrones, tendencias o anomalías de forma rápida y clara, lo que facilita la toma de decisiones basada en datos.

Con este propósito, se ha solicitado un análisis profundo de un dataset que recopila información sobre los 50 libros más vendidos en cada año entre 2009 y 2019. Esta base de datos, compuesta por 550 registros en total, contiene variables clave como el nombre del libro, autor, calificación promedio de los usuarios, número de reseñas, precio, género y año de aparición en el ranking.

El objetivo del análisis es identificar los factores que contribuyen al éxito de un libro en términos de ventas y recepción por parte del público. Comprender estas dinámicas permitirá a Amazon tomar mejores decisiones, detectar patrones de preferencia y diseñar estrategias más efectivas para promover la lectura digital en mercados clave.

Descripción del dataset:

Nombre: Amazon Top 50 Bestselling Books 2009 - 2019

Origen: El dataset utilizado fue brindado por el enunciado del trabajo y pertenece a Kaggle.

Link del dataset: https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019

Cantidad de registros: 550

Cantidad de variables: 7

A continuación, se detallan las variables incluidas en el dataset:

Variable	Tipo de dato	Descripción	Posibles valores
Name	Categórico	Nombre del libro.	Cadenas de texto

			(ej. "10-Day Green Smoothie Cleanse")
Author	Categórico	Autor del libro.	Cadenas de texto (ej. "JJ Smith")
User Rating	Numérico	Calificación de los usuarios en Amazon	0.0 a 5.0 (valores decimales, ej. 4.7)
Reviews	Numérico	Número de reseñas que los usuarios han escrito sobre el libro.	Números enteros positivos (ej. 17350)
Price	Numérico	Precio del libro.	Números enteros positivos (ej. 0 a 105)
Year	Numérico	Los años en que estuvo en el ranking de los más vendidos.	Años del 2009 al 2019
Genre	Categórico	Género del libro	Non Fiction/ Fiction

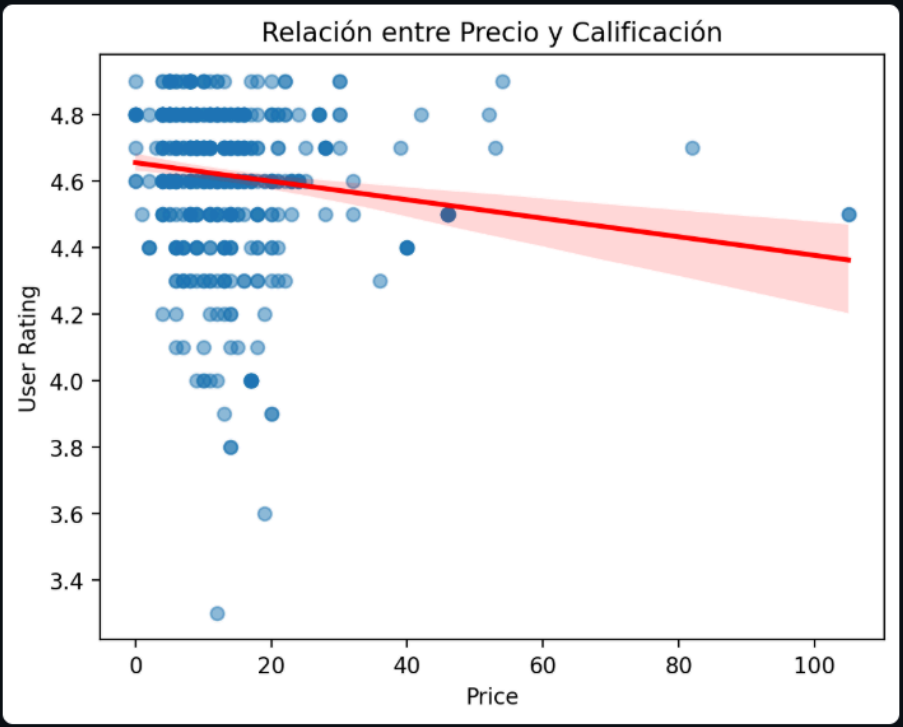
6. Dashboard

Se desarrolló un dashboard interactivo utilizando la biblioteca Streamlit, con el objetivo de visualizar de forma intuitiva y accesible los resultados del modelo predictivo de riesgo cardíaco. Esta interfaz permite al usuario ingresar datos personales relevantes (como edad, salud percibida, actividad física, entre otros) y obtener una predicción inmediata sobre su nivel de riesgo. Además de mostrar la probabilidad calculada por el modelo, el sistema brinda recomendaciones preventivas según el resultado, facilitando su comprensión incluso por parte de usuarios sin conocimientos técnicos. Esta herramienta representa una aplicación práctica y educativa del modelo, con potencial de uso en campañas de salud pública y prevención.

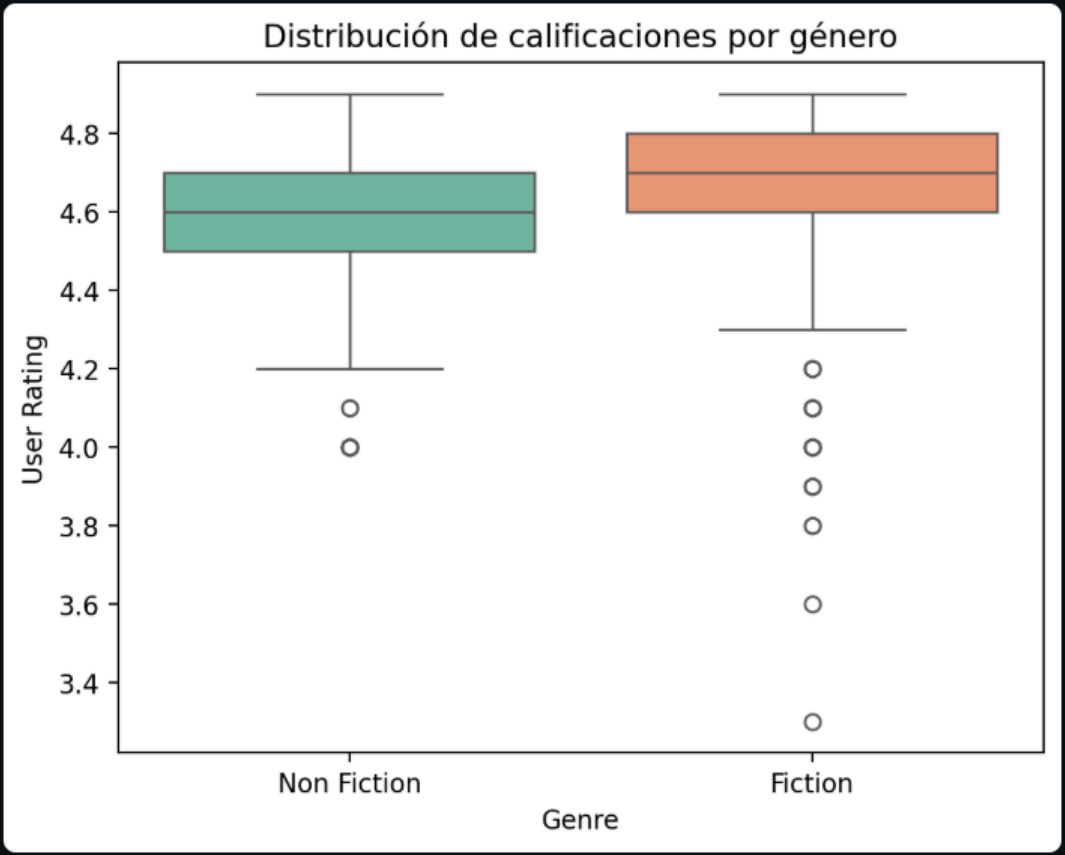
📚 Dashboard – Libros más vendidos en Amazon (2009–2019)

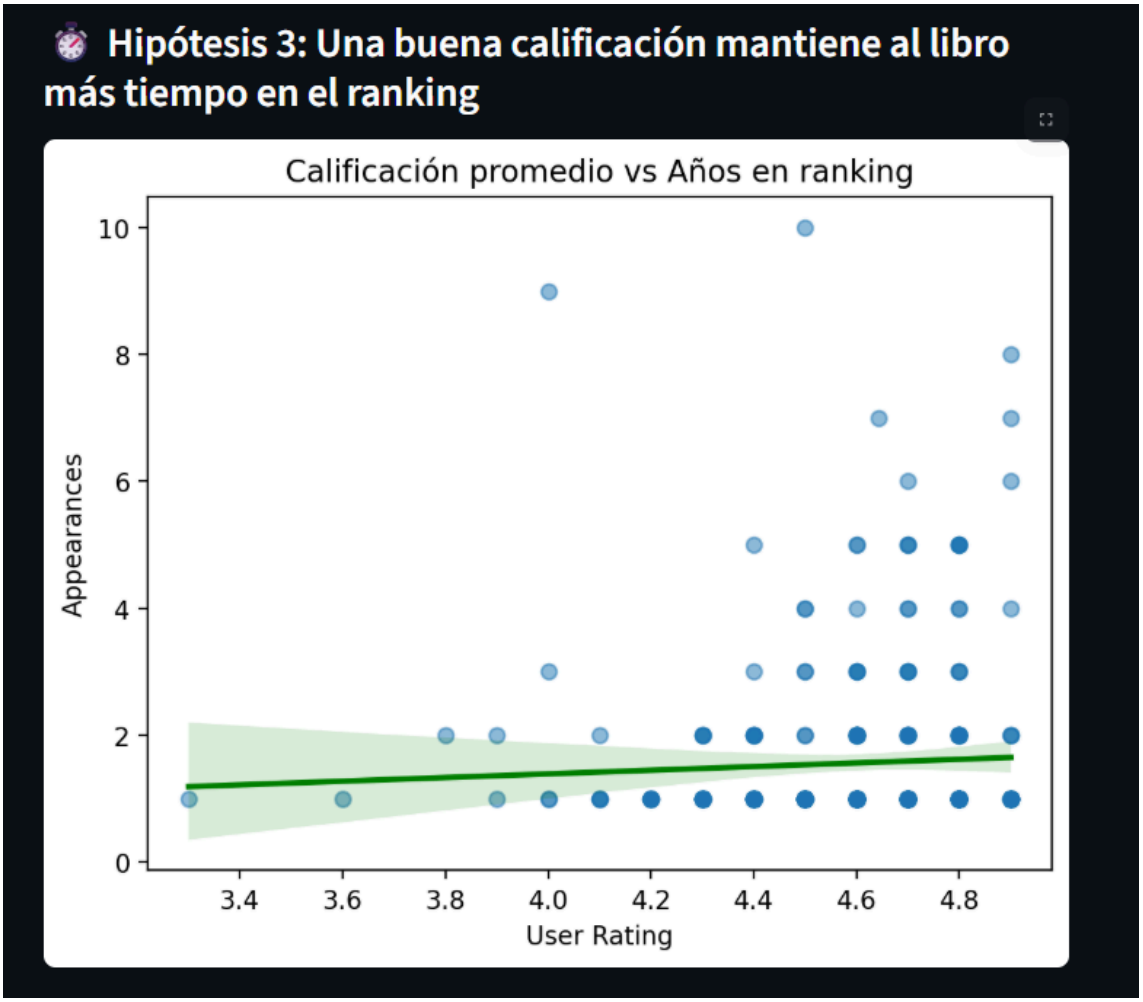
Visualización de datos basada en 3 hipótesis extraídas del análisis histórico del top 50 de Amazon.

📉 Hipótesis 1: Los libros más baratos tienden a recibir mejores calificaciones



📊 Hipótesis 2: Existen géneros que siempre serán bien calificados

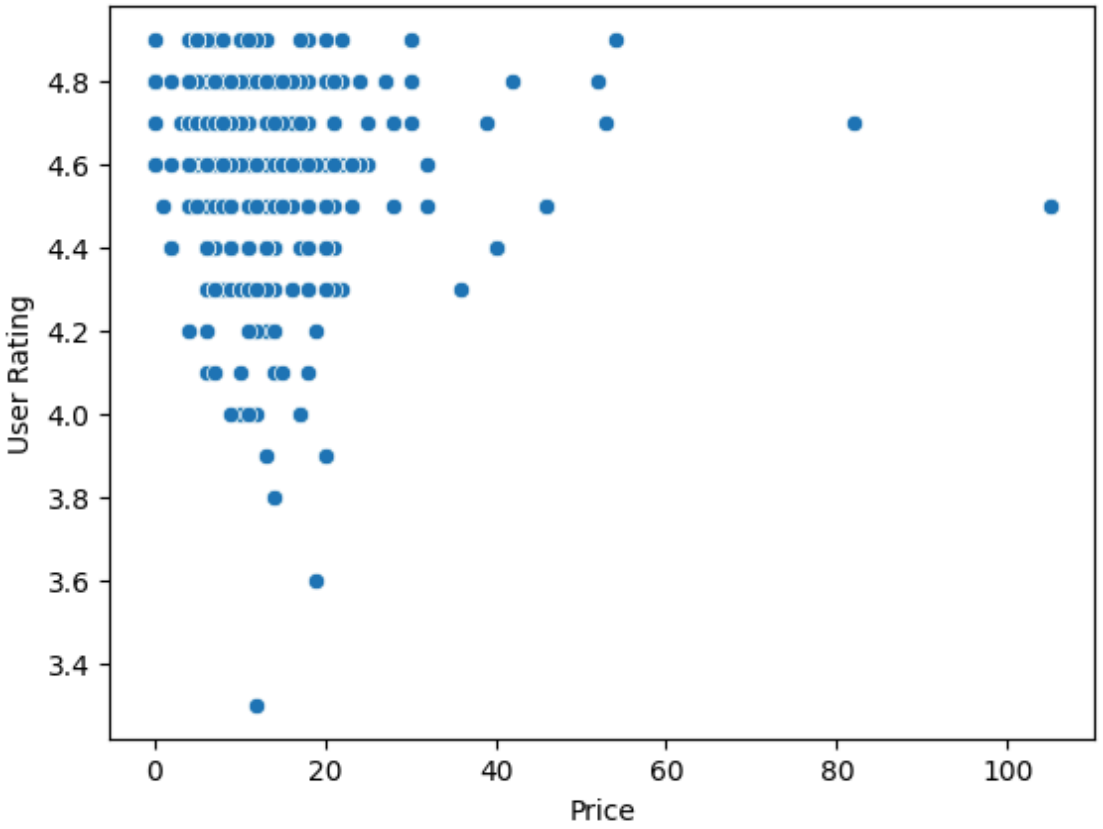




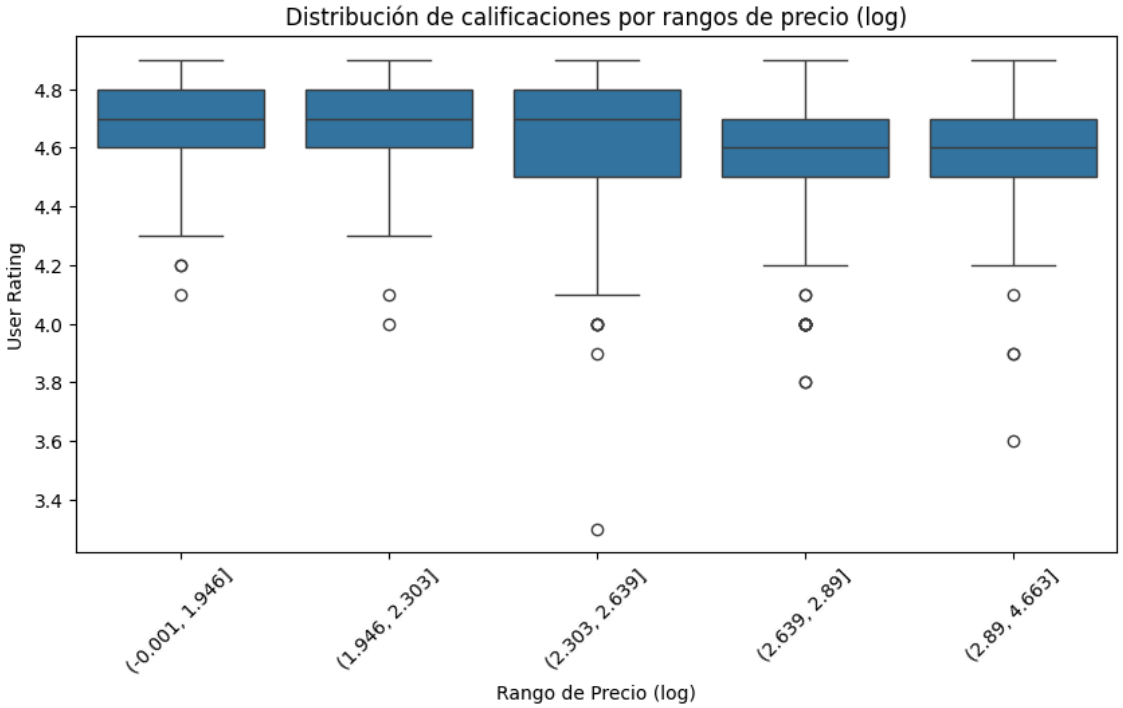
Link al dashboard: https://tf-data-visualization.streamlit.app/

7. Hipótesis

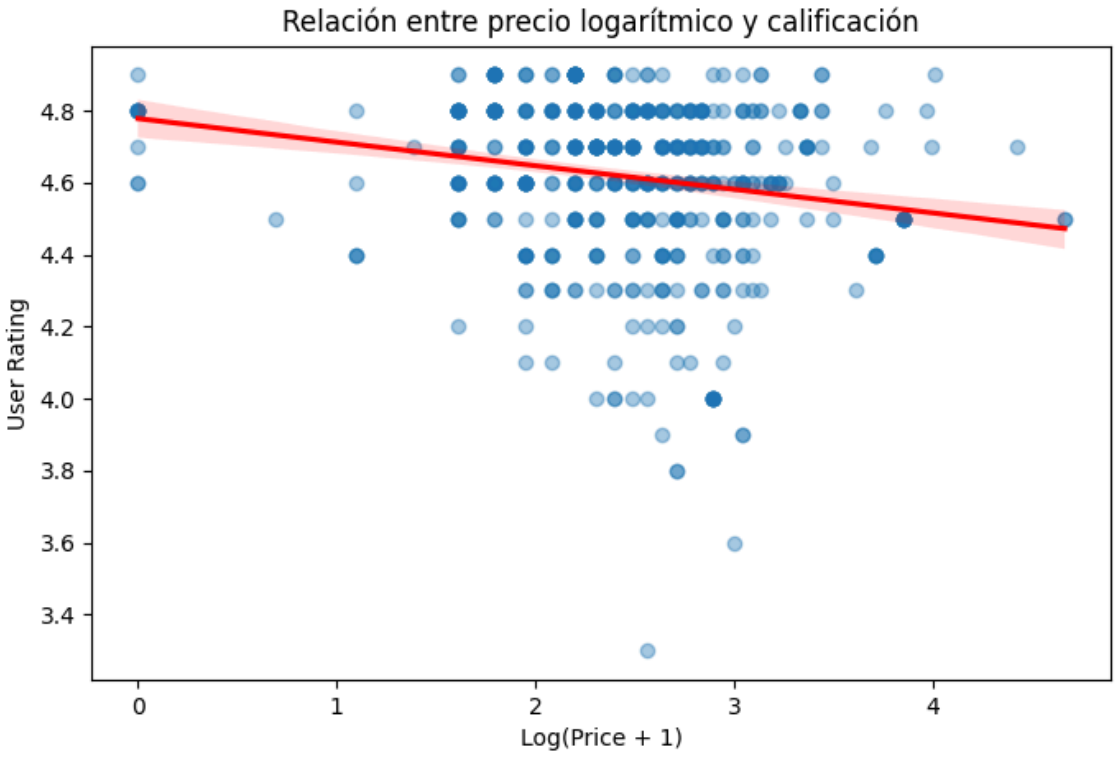
- **H1: Los libros baratos son mejor calificados**



- Alta concentración en el rango de precio de 0 a 20 unidades con calificaciones altas entre 4.2 y 4.8.
- A medida que el precio aumenta, los puntos se dispersan y las calificaciones son más variadas.
- Los libros caros tienen buenas calificaciones, pero son menos frecuentes.



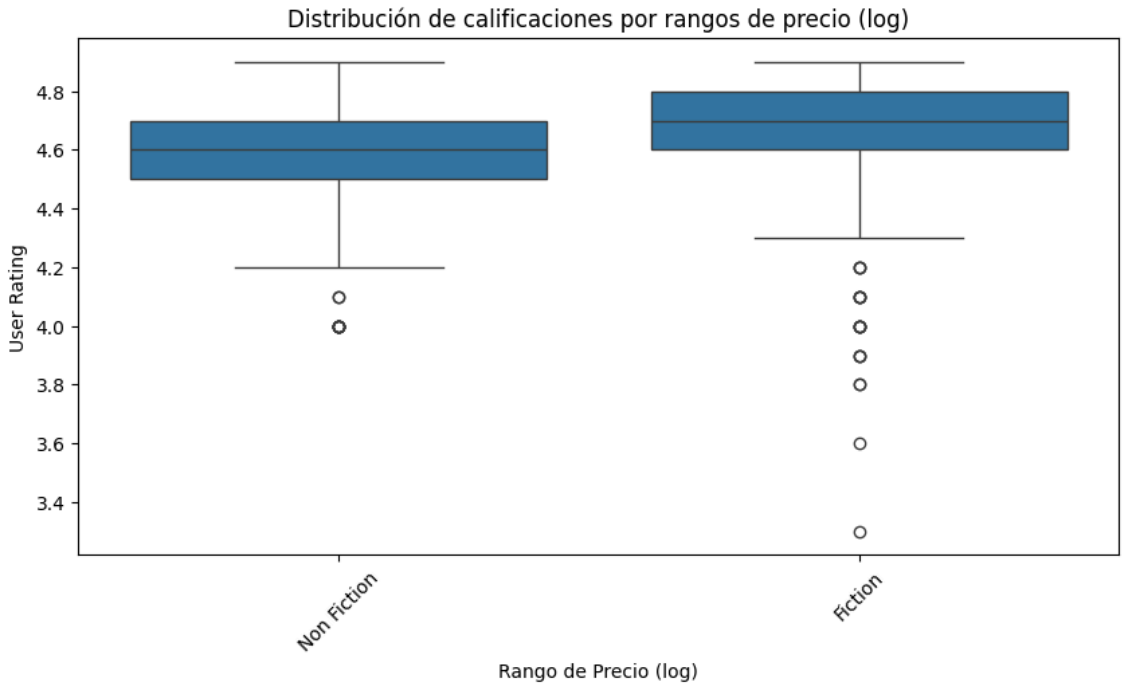
Los rangos de precio más bajos, especialmente los dos primeros, muestran medianas de calificación más altas y menor dispersión. A medida que se avanza hacia rangos de precio más altos, la mediana de las calificaciones disminuye la dispersión aumenta.



La línea de tendencia muestra una pendiente descendente, lo que indica que a medida que el precio (logarítmico) aumenta, la calificación promedio tiende a disminuir. Esta relación negativa respalda directamente la hipótesis de que los libros más baratos tienden a recibir mejores calificaciones.

En conclusión, se acepta la hipótesis H1. Los libros más baratos tienden a recibir mejores calificaciones.

- **H2: Existen géneros que siempre serán bien calificados**



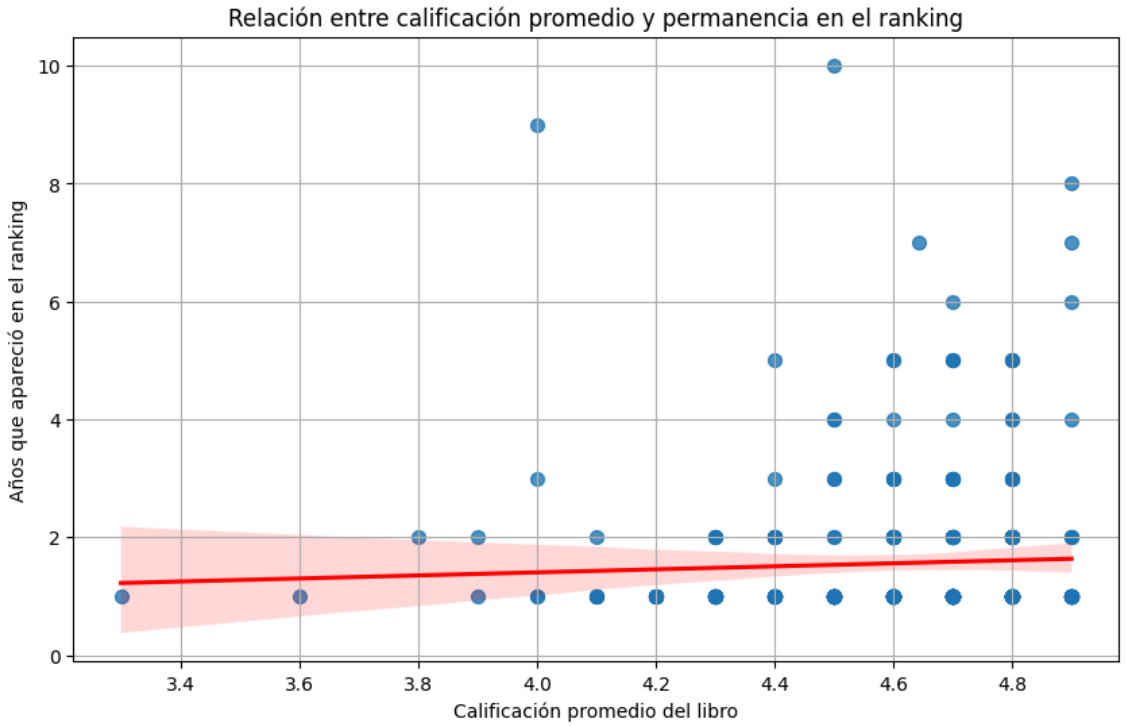
- Ambos géneros tienen calificaciones altas (concentradas entre 4.5 y 4.8), sin embargo la mediana de Fiction tiene un valor mayor (4.7) que Non Fiction (4.6). Esto significa que aunque ambos géneros son bien calificados, los libros de ficción tienden a tener calificaciones más altas.
- Las calificaciones de Non Fiction tienden a ser menos variables lo que sugiere que las calificaciones de los usuarios son más consistentes en este género.
- El hecho que Fiction tenga mayor cantidad de outliers indica que aunque en promedio tiene mejores calificaciones, también presenta bastante variabilidad hacia calificaciones más bajas.

En conclusión, se acepta la hipótesis H2. Existen géneros que siempre serán bien

calificados

Link al notebook de Google collab: https://colab.research.google.com/drive/1rHUlmRvz6f-RBIBF-3c-uvXZb8wBBEUZ?usp=sharing

- **H3: Una buena calificación de manera continua mantiene de manera continua en el ranking a un libro.**



- La mayoría de libros que aparecen muchos años tienen calificaciones altas (5 o más), generalmente mayores a 4.5.
- Casi ningún libro con calificaciones por debajo de 4.3 aparece consistentemente en muchos años.
- La línea de tendencia asciende ligeramente sugiere que existe una relación positiva débil entre la calificación promedio de un libro y su permanencia en el ranking. Los libros con mejores calificaciones tienden

a estar más años en el ranking pero la tendencia nos indica que hay una correlación baja entre ambas variables.

En conclusión, se cumple la hipótesis H3. Una buena calificación de manera continua mantiene de manera continua en el ranking a un libro.

8. Idiomas gráficos seleccionados

Las visualizaciones fueron generadas mayormente utilizando la librería Seaborn (Waskom, 2021), debido a su capacidad para producir gráficos estadísticos claros, personalizables y adaptables al análisis exploratorio de datos.

- Heatmap: Utilizado para verificar la presencia de valores nulos o faltantes (NA) en los datos.
- Countplot o gráfico de barras de frecuencia: Empleado para visualizar la frecuencia de calificaciones (ratings) y comparar la cantidad de libros por género.
- Boxplot o diagrama de cajas: Aplicado para analizar la distribución de precios y reseñas, así como para comparar los rangos de calificaciones y precios por género.
- Barplot: Utilizado para mostrar los 10 autores más populares y los 10 libros más repetidos en el Top 50.
- Lineplot o gráfico de líneas: Implementado para analizar la tendencia del promedio de reseñas por año, y para comparar la evolución del promedio de calificaciones de usuarios y precios a lo largo del tiempo.
- Scatterplot o gráfico de dispersión: Sirve para visualizar la relación entre el precio y la calificación de los usuarios.
- Regplot: Utilizado para observar la relación entre el precio y la calificación incluyendo una línea de tendencia calculada algorítmicamente.

9. Recomendaciones

- Una mejora importante sería la creación de un dashboard dinámico que permita a los usuarios interactuar con los datos filtrando por año, género o autor. Esto facilitaría la toma de decisiones basada en datos.

- Aunque nuestras hipótesis fueron evaluadas con base en tendencias visuales, reconocemos que incorporar pruebas estadísticas más formales podría fortalecer la validez de los hallazgos y dar mayor soporte cuantitativo a las conclusiones.
- El preprocesamiento realizado fue adecuado para nuestros objetivos, pero identificamos oportunidades para automatizar y documentar mejor cada paso. Esto permitiría replicar el análisis de forma más ordenada y profesional en el futuro.

10. Bibliografía

DataScientest. (2023, agosto 7). *DataViz: definición, objetivos y usos*. https://datascientest.com/es/dataviz-definicion-objetivos-y-usos

Kaggle. (2020, October 13). Amazon Top 50 Bestselling Books 2009 - 2019. Kaggle. Retrieved July 13, 2025, from https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019

Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.