Informe EDA

1. Objetivo del Análisis

El objetivo principal de este análisis es explorar la base de datos de libros para identificar patrones en las calificaciones de los usuarios, determinar cuáles son las categorías mejor valoradas y analizar la distribución de datos relevantes. Se busca detectar tendencias en la percepción de los libros según su género, año de publicación y otros factores clave que puedan influir en su recepción por parte de los lectores.

2. Metodología

Para llevar a cabo este análisis, se siguieron los siguientes pasos:

- **Limpieza de datos**: Se identificaron y trataron valores nulos en columnas como "subtitle", "authors" y "categories".
- **Normalización de datos**: Se aplicó el escalado MinMax para columnas numéricas como "average_rating", "num_pages" y "ratings_count" con el fin de hacerlas comparables.
- **Análisis de distribución**: Se examinaron los ratings promedio por categoría mediante diagramas de caja (boxplots) para visualizar la dispersión y distribución de las calificaciones en cada género literario.
- **Cálculo de estadísticas descriptivas**: Se obtuvieron medidas de tendencia central como la media, mediana, desviación estándar y percentiles para comprender mejor la variabilidad de los datos.

3. Principales Hallazgos

3.1. Distribución de Ratings por Categoría

El análisis de los ratings promedio por categoría reveló las siguientes observaciones clave:

- **Categorías mejor valoradas**: Se identificaron "Philosophy", "Poetry" y "Religion" como las categorías con mayor promedio de calificación, lo que sugiere que los lectores suelen valorar positivamente los libros dentro de estos géneros.
- **Variabilidad en las calificaciones**: Categorías como "Travel" y "Social Science" mostraron una mayor dispersión en las calificaciones, lo que podría indicar que la percepción de los libros dentro de estas áreas varía significativamente entre los lectores.
- **Categoría más representada**: "Fiction" es una de las categorías con mayor cantidad de libros en la base de datos, presentando una distribución de calificaciones más amplia, aunque con una mediana en torno a los 4 puntos.

3.2. Manejo de Datos Nulos

- Se detectaron valores nulos en "subtitle", "authors", “thumbnail” y "description". Se tomó la decisión de rellenar las filas indicando que no tienen para evitar una pérdida excesiva de datos.
- Para otras variables, como "categories" y “published_year”, se optó por eliminarlas ya que era una cantidad menor que no afectaba a nuestro análisis.
- Para "average_rating", "num_pages" y "ratings_count" se decidió implementar la mediana para obtener un equilibrio.

3.3. Distribución Temporal de los Libros

- La mayoría de los libros analizados fueron publicados entre **1990 y 2019**, con un notable aumento en el número de publicaciones durante los años 2000.
- Se observó que los libros publicados antes de 1980 tienden a recibir menos calificaciones en comparación con los más recientes, lo que puede estar relacionado con la digitalización y el acceso a plataformas de reseñas en línea.

4. Conclusiones y Recomendaciones

A partir del análisis realizado, se pueden extraer las siguientes conclusiones y sugerencias:

- Las categorías de libros con mejor calificación pueden indicar una **mayor aceptación por parte de los lectores**, lo que puede ser relevante para editoriales y librerías al momento de definir estrategias de mercado.
- Se recomienda un estudio más profundo sobre la relación entre la fecha de publicación y el número de reseñas recibidas, lo que podría ayudar a entender mejor la evolución de la popularidad de ciertos libros con el tiempo.
- Para mejorar la calidad del dataset, podría completarse la información faltante en "subtitle" y "description" mediante técnicas de web scraping o mediante el uso de bases de datos complementarias.
- Se podrían desarrollar modelos de predicción para estimar el rating de un libro en función de su categoría y año de publicación, lo que podría ser útil para autores y editores al evaluar el potencial éxito de nuevos lanzamientos.

El presente informe ofrece un panorama general sobre la distribución y tendencias en las calificaciones de libros, proporcionando una base para futuras investigaciones y estrategias en el ámbito editorial y bibliográfico. Para complementar el análisis, se sugiere incluir visualizaciones gráficas como histogramas y heatmaps que faciliten la interpretación de los datos y sus relaciones.