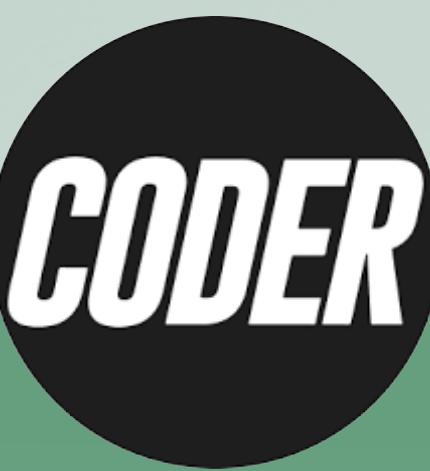




Comisión 61175



Data Science II: Machine Learning para la Ciencia de Datos

Primera Entrega

Alumno:

Diego Lopez Castan

Objetivo del Proyecto



El objetivo es poder lograr una comprensión del dataset de libros. Se analizará el dataset para comprender las diferentes características de los libros y cuales de ellos son los más importantes. Los temas principales a los cuales voy a analizar son:

- **Analizar las relaciones entre el rating de un libro y sus géneros:** Identificar si ciertos géneros están asociados con mejores ratings en general.
- **Estudio de popularidad por series literarias:** Evaluar el impacto de pertenecer a una serie en el éxito de los libros (por ejemplo, comparar el rating de los libros de una serie con otros libros del mismo autor que no pertenecen a una serie).
- **Segmentación por idioma y género:** Estudiar la distribución de géneros según los diferentes idiomas de los libros y su relación con los ratings.
- **Relación entre la longitud del libro y el éxito de un libro:** Investigar si los libros más cortos o más largos tienen alguna ventaja en términos de popularidad (rating).
- **Estudio de la relación entre el número de géneros y el rating de un libro:** Analizar si los libros que pertenecen a un genero específico recibe mejor rating.

Datos



Este conjunto de datos está formada por **52478 libros** que han sido incluidos en la lista de los mejores libros de la historia del sitio GoodReads.com.

El conjunto de datos presenta información diversa sobre cada libro, desde las puntuaciones hasta los premios y los géneros que lo componen .

Alcance del Proyecto



El alcance del proyecto sobre los libros incluye:

1. **Análisis de los libros:** Evaluación de los libros más importantes según el sitio web GoodReads.com.
2. **Estudio de la relación libros-rating:** identificar las tendencias de los libros .
3. **Estudio de la relación libros-precio:** identificar relaciones entre características de los libros y el precio.
4. **Limitaciones:** El proyecto solo incluye los 52478 libros.
5. **Datos:** Los datos que se van a analizar corresponden hasta noviembre de 2020.

Hipótesis



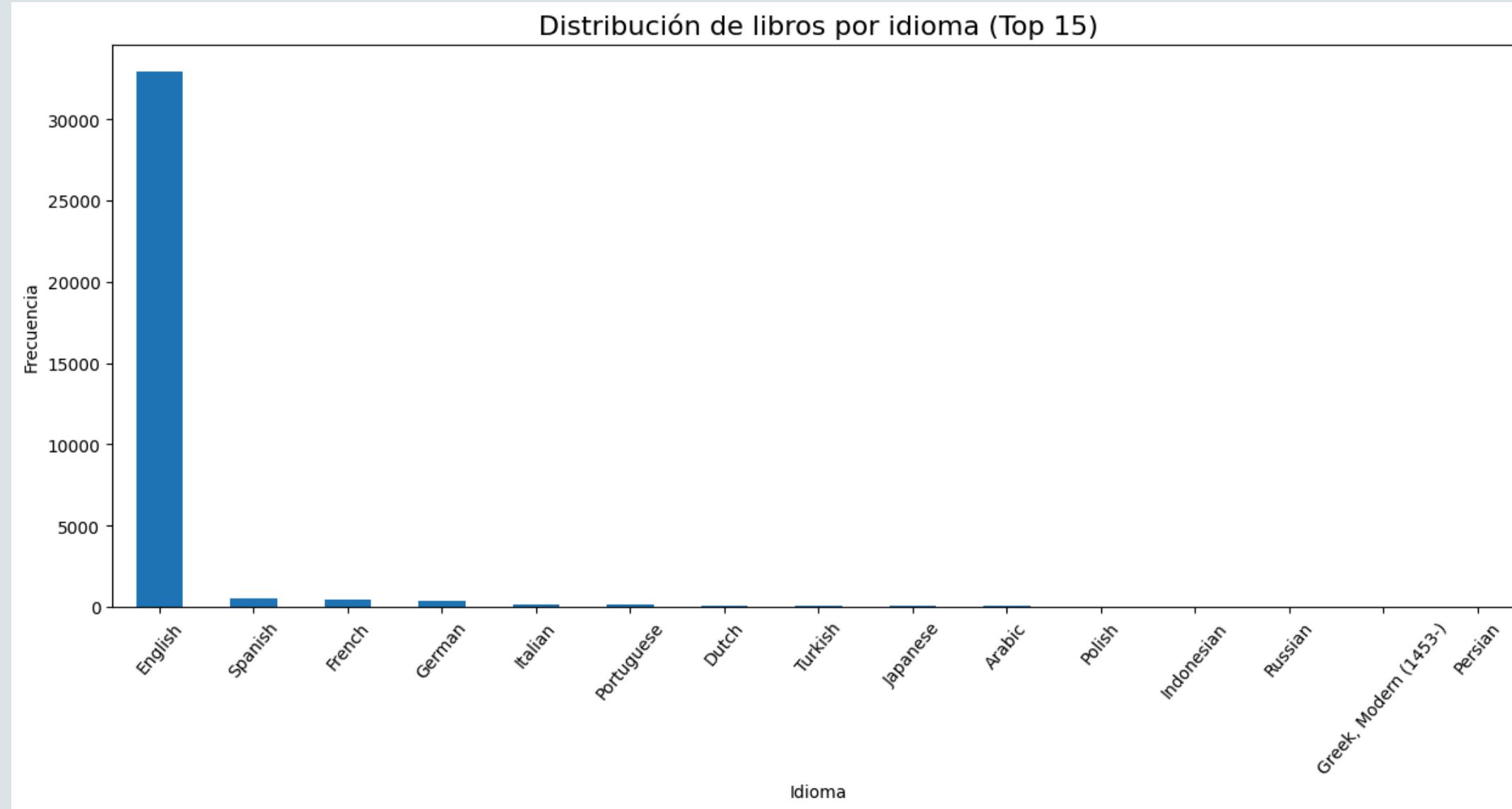
Ciertos géneros están más asociados a libros con ratings altos: Géneros como "Fantasía" y "Ciencia ficción" tienen mejores ratings promedio que géneros como "Romance" o "Ficción histórica".

Los libros de ficción tienen ratings más elevados que los libros de no ficción: Existe alguna diferencia entre los libros de ficción y no ficción.

Los libros de ficción distópica tienen mejores ratings que los de ficción general: Los libros categorizados como "distopía" tienen mejores críticas debido a la popularidad de este subgénero en los últimos años.

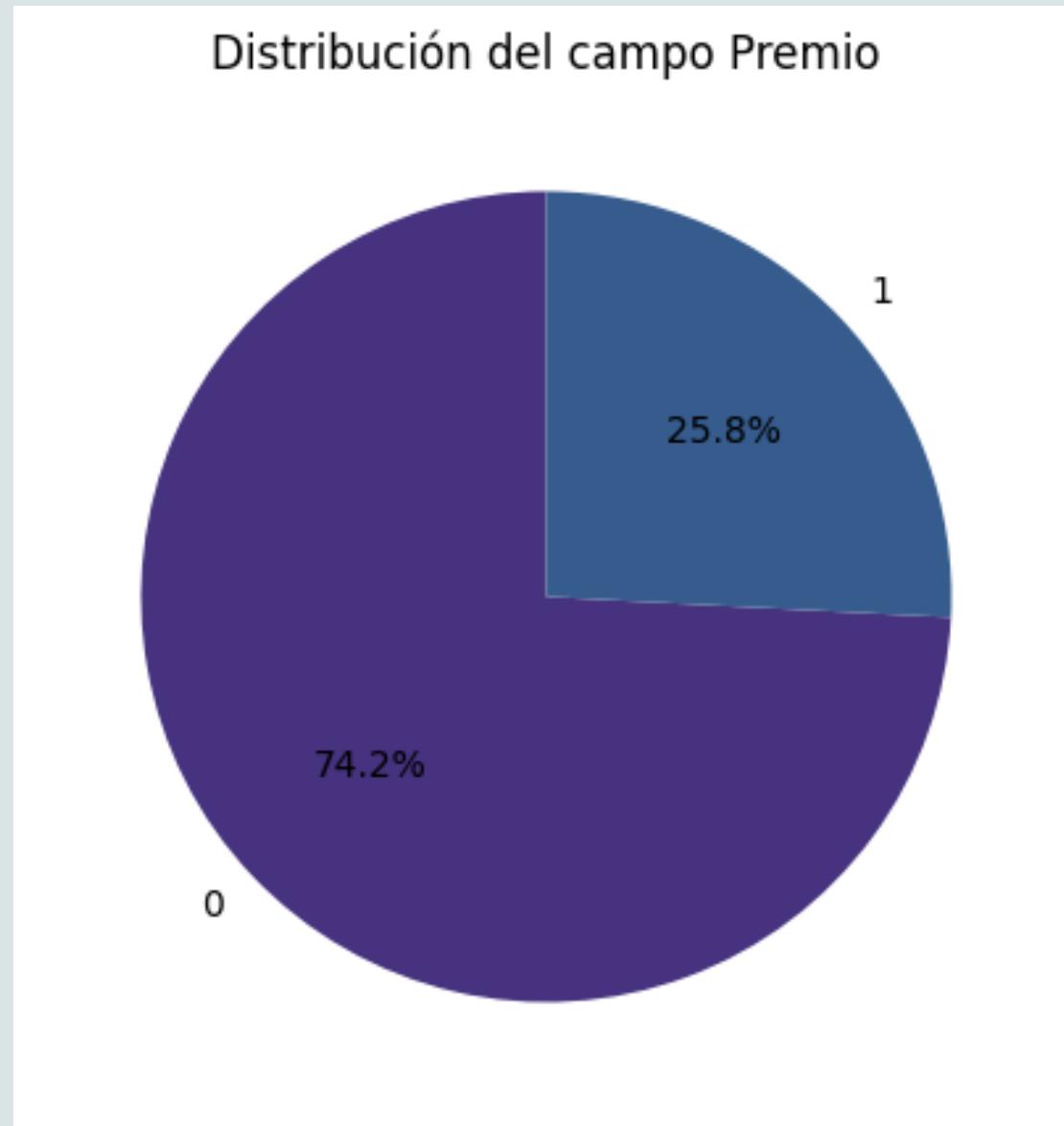
Análisis Exploratorio

Análisis Exploratorio



La mayoría de los libros que se encuentran en el dataset son en el idioma Ingles.

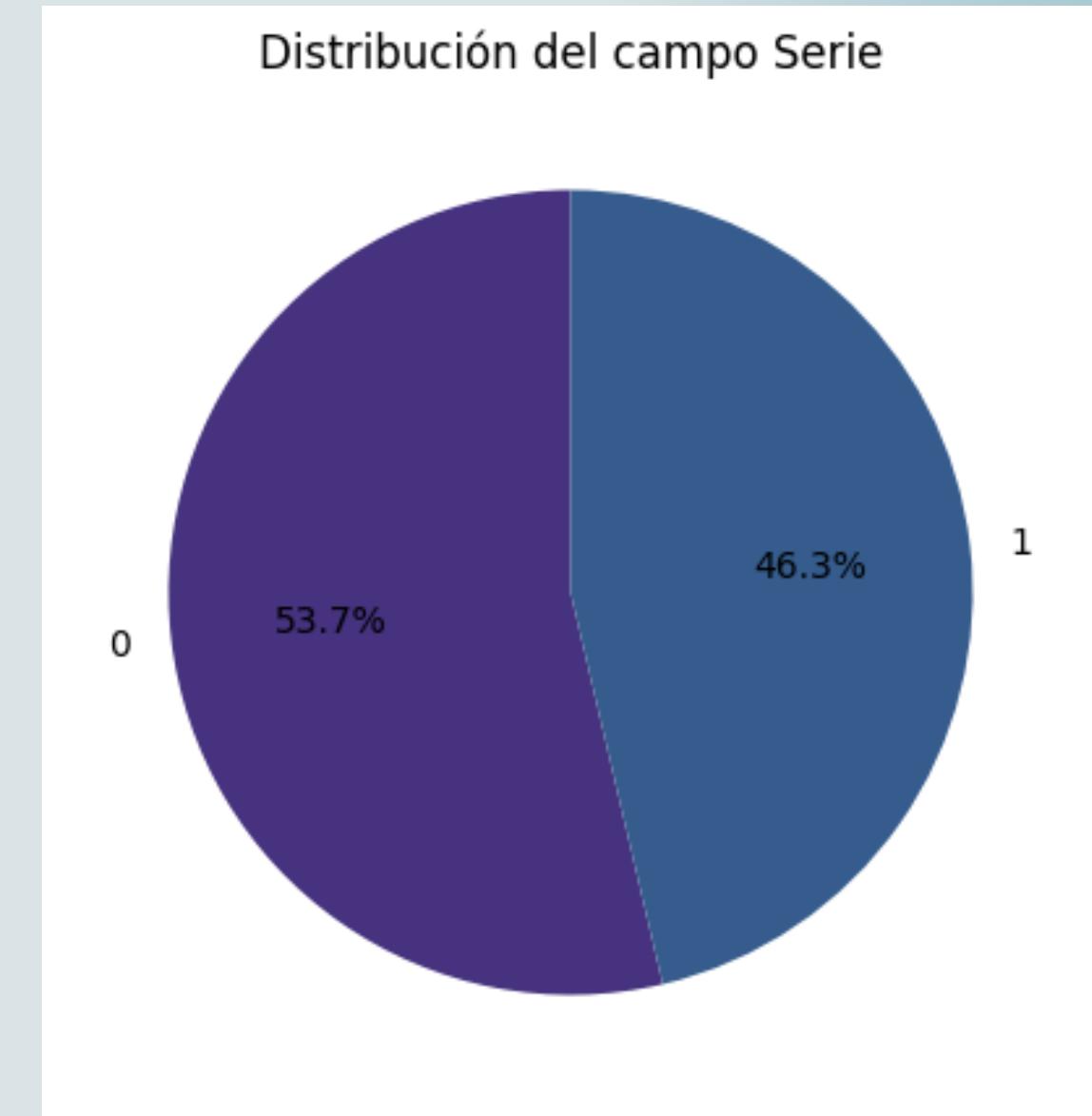
Análisis Exploratorio



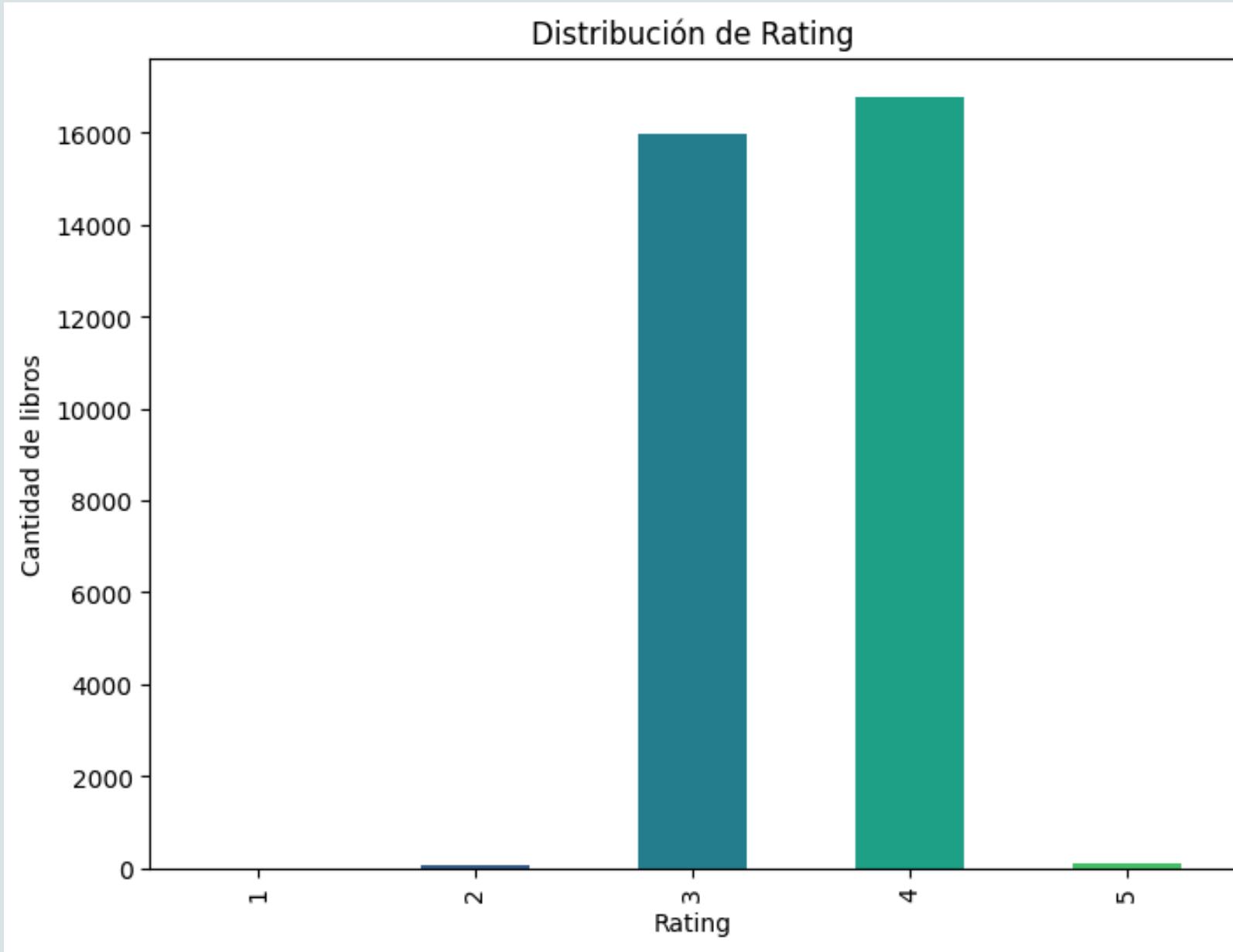
Más del 25% de los libros que se encuentran en el dataset tienen al menos un premio.

Análisis Exploratorio

Existen muchos libros que forman parte de una serie de libros. Son casi 46,3%.

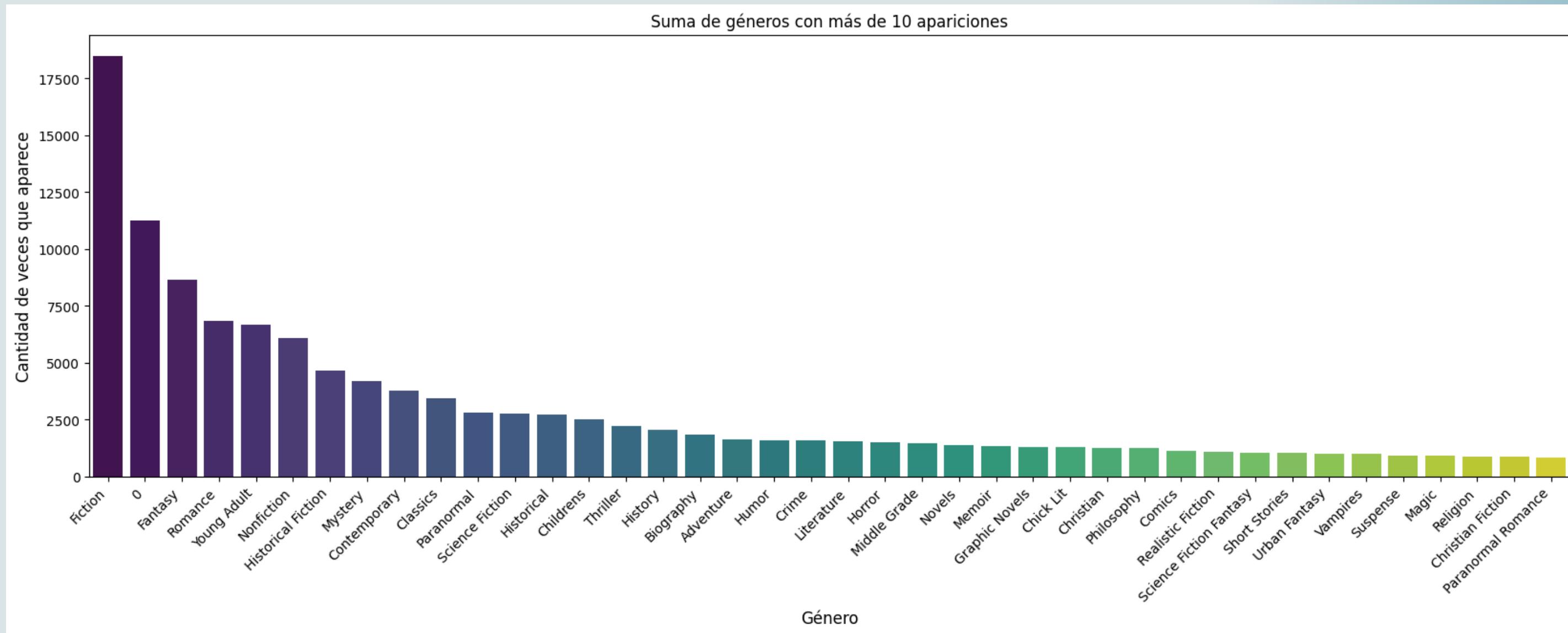
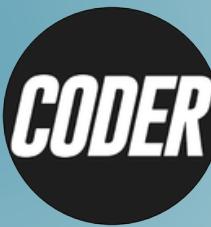


Análisis Exploratorio



La mayoría de los libros se encuentran entre los 3 y 4 puntos del rating.

Análisis Exploratorio

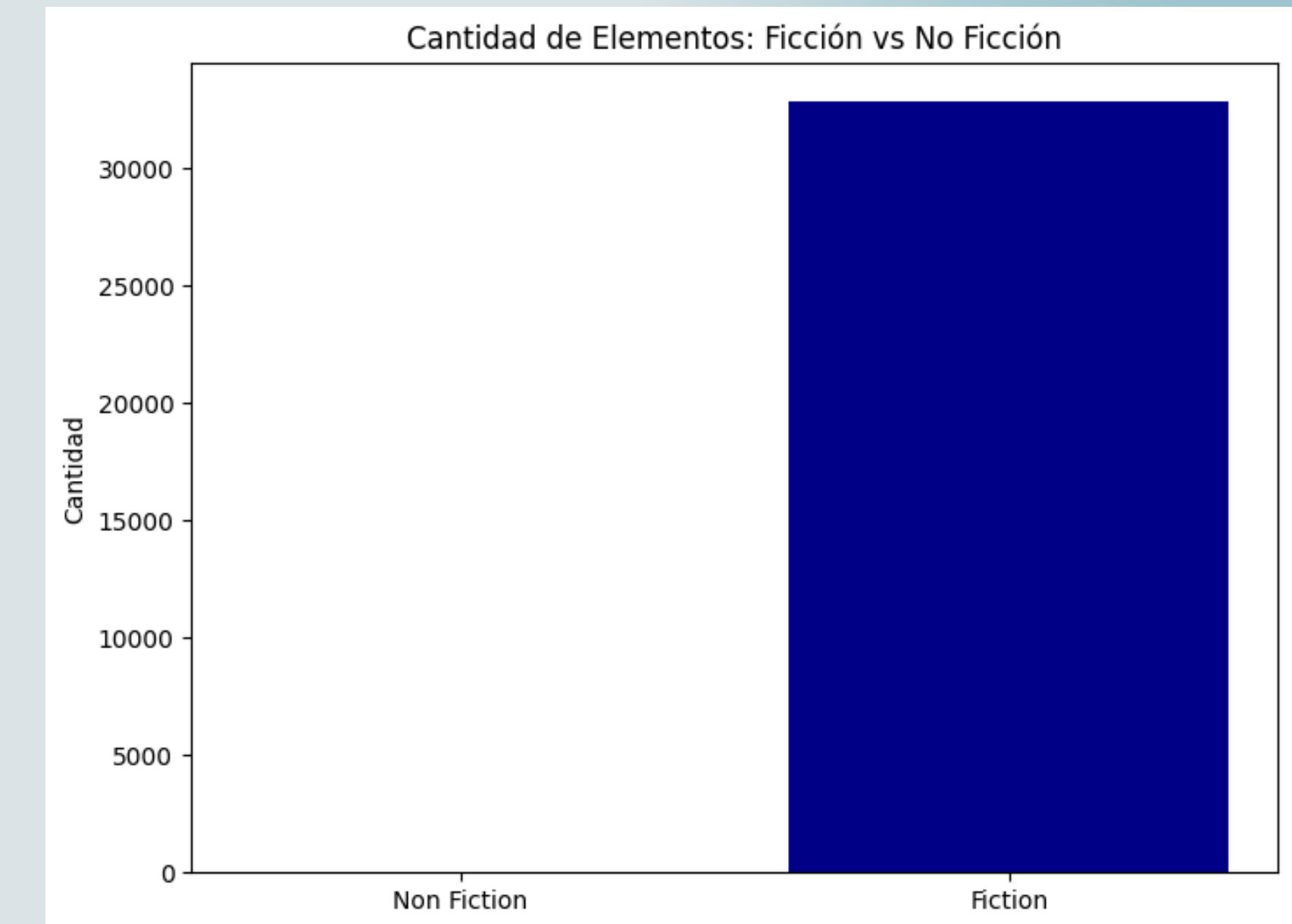


En este gráfico se pueden ver las principales generos de los libros que se encuentran dentro del dataset.

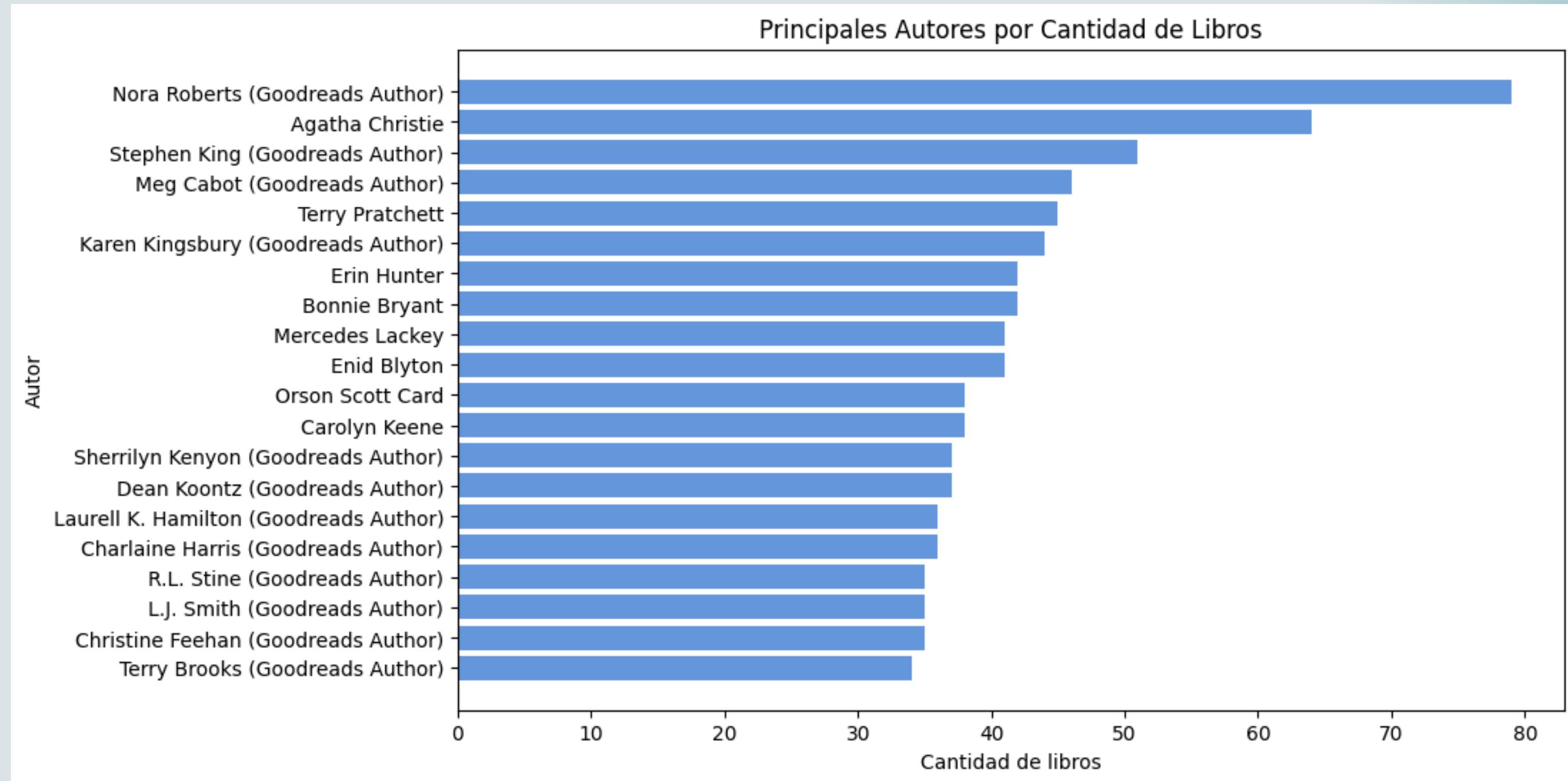
Análisis Exploratorio



Solo 25 libros que se encuentran en el dataset son de No Ficción.



Análisis Exploratorio



En este gráfico se puede visualizar los autores con más libros editados.

Nuevos Modelos

Busco el mejor modelo



Model		MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
xgboost	Extreme Gradient Boosting	0.0998	0.0324	0.1799	0.8764	0.0402	0.0291	0.3610
lightgbm	Light Gradient Boosting Machine	0.0988	0.0335	0.1829	0.8722	0.0411	0.0290	1.3180
rf	Random Forest Regressor	0.0816	0.0342	0.1850	0.8694	0.0414	0.0240	12.0440
et	Extra Trees Regressor	0.1165	0.0495	0.2223	0.8112	0.0496	0.0340	2.9750
dt	Decision Tree Regressor	0.0631	0.0641	0.2529	0.7556	0.0565	0.0186	0.2320
gbr	Gradient Boosting Regressor	0.1762	0.0658	0.2564	0.7489	0.0573	0.0514	4.3550
knn	K Neighbors Regressor	0.1338	0.0763	0.2762	0.7087	0.0617	0.0395	0.2050
ridge	Ridge Regression	0.3433	0.1554	0.3942	0.4069	0.0900	0.1017	0.1100
br	Bayesian Ridge	0.3434	0.1554	0.3942	0.4069	0.0900	0.1018	0.0450
lr	Linear Regression	0.3433	0.1554	0.3942	0.4069	0.0900	0.1017	1.1770
lar	Least Angle Regression	0.3430	0.1574	0.3966	0.3993	0.0903	0.1017	0.0400
en	Elastic Net	0.3790	0.1678	0.4095	0.3598	0.0914	0.1109	0.4560
huber	Huber Regressor	0.3840	0.1735	0.4165	0.3377	0.0928	0.1123	0.3670
lasso	Lasso Regression	0.4230	0.1955	0.4421	0.2540	0.0982	0.1235	0.6590
llar	Lasso Least Angle Regression	0.4230	0.1955	0.4421	0.2540	0.0982	0.1235	0.0390
ada	AdaBoost Regressor	0.4815	0.2357	0.4855	0.1002	0.1082	0.1400	1.7280
omp	Orthogonal Matching Pursuit	0.5047	0.2615	0.5113	0.0020	0.1142	0.1476	0.0390
dummy	Dummy Regressor	0.5058	0.2621	0.5120	-0.0005	0.1144	0.1480	0.0310
par	Passive Aggressive Regressor	1.9105	131.6896	10.2773	-495.2799	0.4188	0.5328	0.0530

Para mejorar las métricas he utilizado la librería **pycaret**. Según este modelo los mejores modelos son xgboost y lightgbm.

Comparo modelos



	modelo	mse	r2
0	LinearRegression	0.148332	0.424452
1	XGBRegressor	0.064698	0.748963
2	LGBMRegressor	0.031400	0.878162

Al comparar los modelos se puede apreciar que los mejores resultados se han obtenido con el modelo LGBMRegressor.

Según los datos obtenidos el mejor modelo fue el LGBMRegressor. Esto lo podemos verificar ya que el menor error cuadrático medio (MSE) es de 0.031400, el más bajo entre los tres modelos. También el r2 es el más alto llegando a 0.878162.

Mejores Parámetros



- **subsample=0.6:**
 - Define la fracción de datos que se usa para entrenar cada árbol. Por ejemplo, con subsample=0.6, se usará el 60% de los datos en cada iteración. Esto ayuda a prevenir el sobreajuste (overfitting) y puede mejorar la generalización.
- **num_leaves=100:**
 - Especifica el número máximo de hojas (o nodos terminales) en cada árbol. Un mayor número de hojas permite capturar relaciones más complejas en los datos, pero aumenta el riesgo de sobreajuste.
- **n_estimators=250:**
 - Representa el número total de árboles que se construirán en el modelo. Más árboles generalmente mejoran la precisión del modelo hasta cierto punto, pero también incrementan el tiempo de entrenamiento y el riesgo de sobreajuste.
- **max_depth=9:**
 - Indica la profundidad máxima de los árboles. Limitar la profundidad ayuda a controlar el tamaño del modelo y a prevenir el sobreajuste. Una profundidad más alta permite capturar relaciones más complejas, pero también aumenta el riesgo de sobreajuste.
- **learning_rate=0.1:**
 - Es un factor de reducción que controla cuánto contribuye cada árbol al modelo final. Un valor más bajo generalmente lleva a mejores resultados, pero necesita más iteraciones (más árboles) para converger. Un valor más alto puede acelerar el entrenamiento, pero con riesgo de no capturar bien las relaciones en los datos.
- **colsample_bytree=1.0:**
 - Define la proporción de características (o columnas) que se usan para entrenar cada árbol. Reducir este valor puede hacer el modelo más robusto contra el sobreajuste y puede acelerar el entrenamiento.

Aplicaciones

App para Predecir Rating

Predicción de Rating

Introduce las características del libro para predecir su rating.

Páginas: 0

Porcentaje de Likes: 0

¿Es parte de una serie? (1 = Sí, 0 = No): 0

¿Tiene premio? (1 = Sí, 0 = No): 0

Cantidad de estrellas 5: 0

Cantidad de estrellas 4: 0

Cantidad de estrellas 3: 0

Cantidad de estrellas 2: 0

Cantidad de estrellas 1: 0

¿Es ficción? (1 = Sí, 0 = No): 0

Precio: 0

Submit

Use via API 🔍 · Construido con Gradio 🎵

Al seleccionar los diferentes parámetros la aplicación va a indicar cual es la predicción del rating.

App para Predicir Precio

Predicción de Precio

Introduce las características del libro para predecir su precio.

Páginas

Porcentaje de Likes

¿Es parte de una serie? (1 = Sí, 0 = No)

¿Tiene premio? (1 = Sí, 0 = No)

Cantidad de estrellas 5

Cantidad de estrellas 4

Cantidad de estrellas 3

Cantidad de estrellas 2

Cantidad de estrellas 1

¿Es ficción? (1 = Sí, 0 = No)

Clear **Submit**

Use via API  · Construido con Gradio 

Al seleccionar los diferentes parámetros la aplicación va a indicar cual es la predicción del precio.

Diego Lopez Castan

App para Recomendar libros

Recomendador de Libros

book_title

output

Clear Submit Flag



Al seleccionar los diferentes parámetros la aplicación va a indicar cual es la predicción del precio.

App Streamlit para predecir rating

The screenshot shows a Streamlit application running at `localhost:8501`. The title of the app is "CODERHOUSE". The main section is titled "Predicción de Rating" and contains the following text: "Introduce las características del libro para predecir su rating." Below this, there are several input fields for different book features:

- Páginas: 100
- Porcentaje de Likes: 50.00
- ¿Es parte de una serie?: 0
- ¿Tiene premio?: 0
- Cantidad de estrellas 5: 0
- Cantidad de estrellas 4: 0
- Cantidad de estrellas 3: 0
- Cantidad de estrellas 2: 0
- Cantidad de estrellas 1: 0
- ¿Es ficción?: 0
- Precio: 10.00

At the bottom of the form is a button labeled "Predecir Rating".

Diego Lopez Castan

App Streamlit para predecir precios

The screenshot shows a Streamlit application running at `localhost:8501`. The interface features a large **CODERHOUSE** logo at the top. Below it, the title **Predicción de Precio** is displayed. A sub-instruction reads: "Introduce las características del libro para predecir su precio." The form contains the following input fields:

- Páginas: 100
- Porcentaje de Likes: 50.00
- ¿Es parte de una serie?: 0
- ¿Tiene premio?: 0
- Cantidad de estrellas 5: 0
- Cantidad de estrellas 4: 0
- Cantidad de estrellas 3: 0
- Cantidad de estrellas 2: 0
- Cantidad de estrellas 1: 0
- ¿Es ficción?: 0

A "Predecir Precio" button is located at the bottom of the form.

Diego Lopez Castan

App Streamlit para recomendar libros

A screenshot of a Streamlit application running locally at port 8501. The page has a header with the CoderHouse logo and the title "Recomendador de Libros". It includes a text input field for entering a book title, a dropdown menu showing "The Book Thief" as the selected item, and a red button labeled "Recomendar Libros". Below the button, a message says "Si te gustó 'The Book Thief', también te podrían gustar:" followed by a list of five recommended books: 1. I Am the Messenger, 2. The Thief, 3. The Thief of Always, 4. Before We Were Yours, and 5. Moloka'i.

localhost:8501

CODERHOUSE

Recomendador de Libros

Introduce el título de un libro y te recomendaremos otros similares.

Selecciona un libro

The Book Thief

Recomendar Libros

Si te gustó 'The Book Thief', también te podrían gustar:

1. I Am the Messenger
2. The Thief
3. The Thief of Always
4. Before We Were Yours
5. Moloka'i