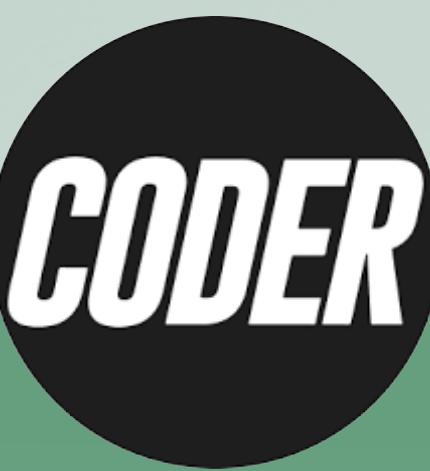




Comisión 61175



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



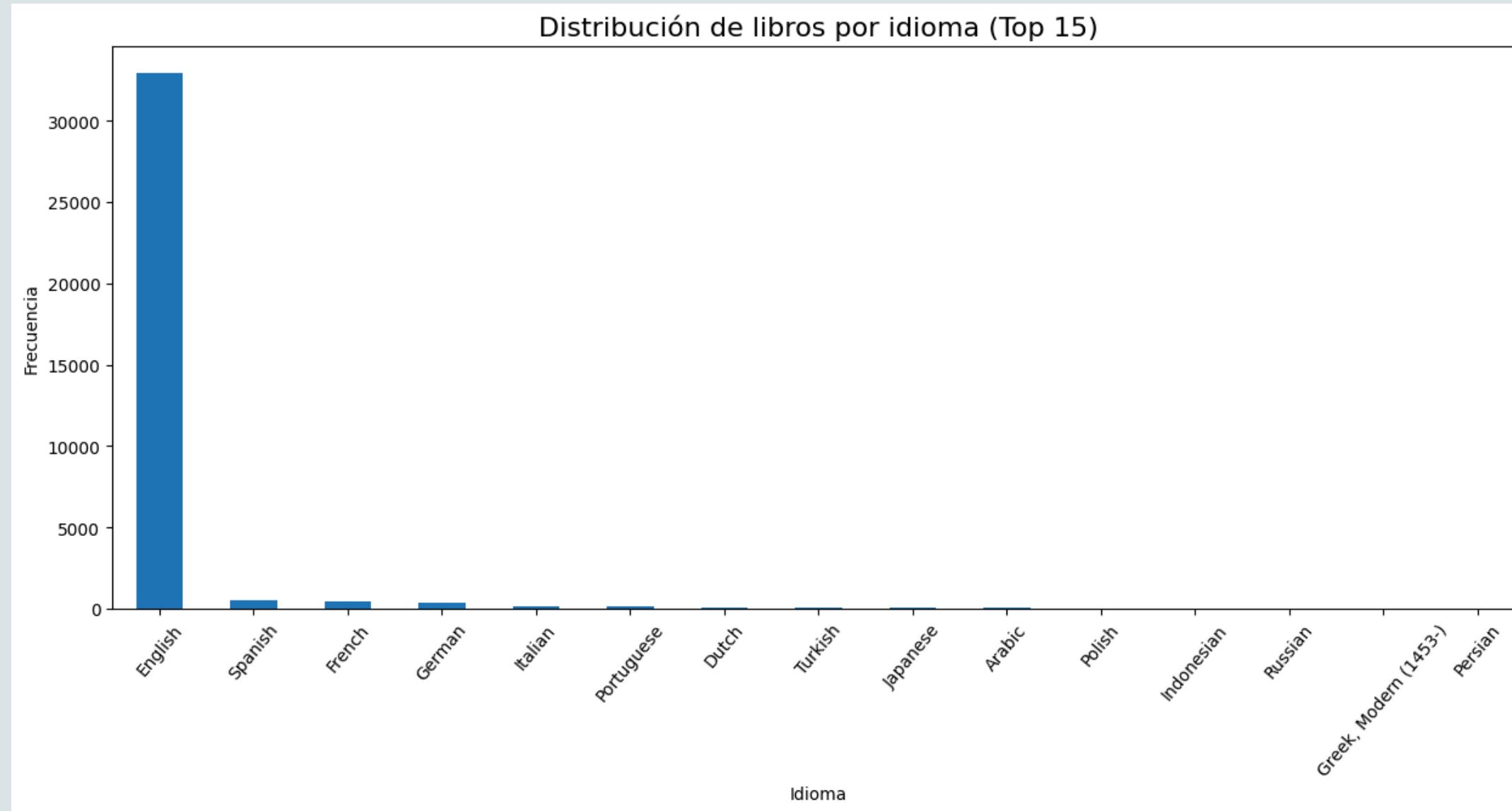
Ciertos géneros están más asociados a libros con ratings altos: Géneros como "Fantasía" y "Ciencia ficción" tienen mejores ratings promedio que géneros como "Romance" o "Ficción histórica".

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.

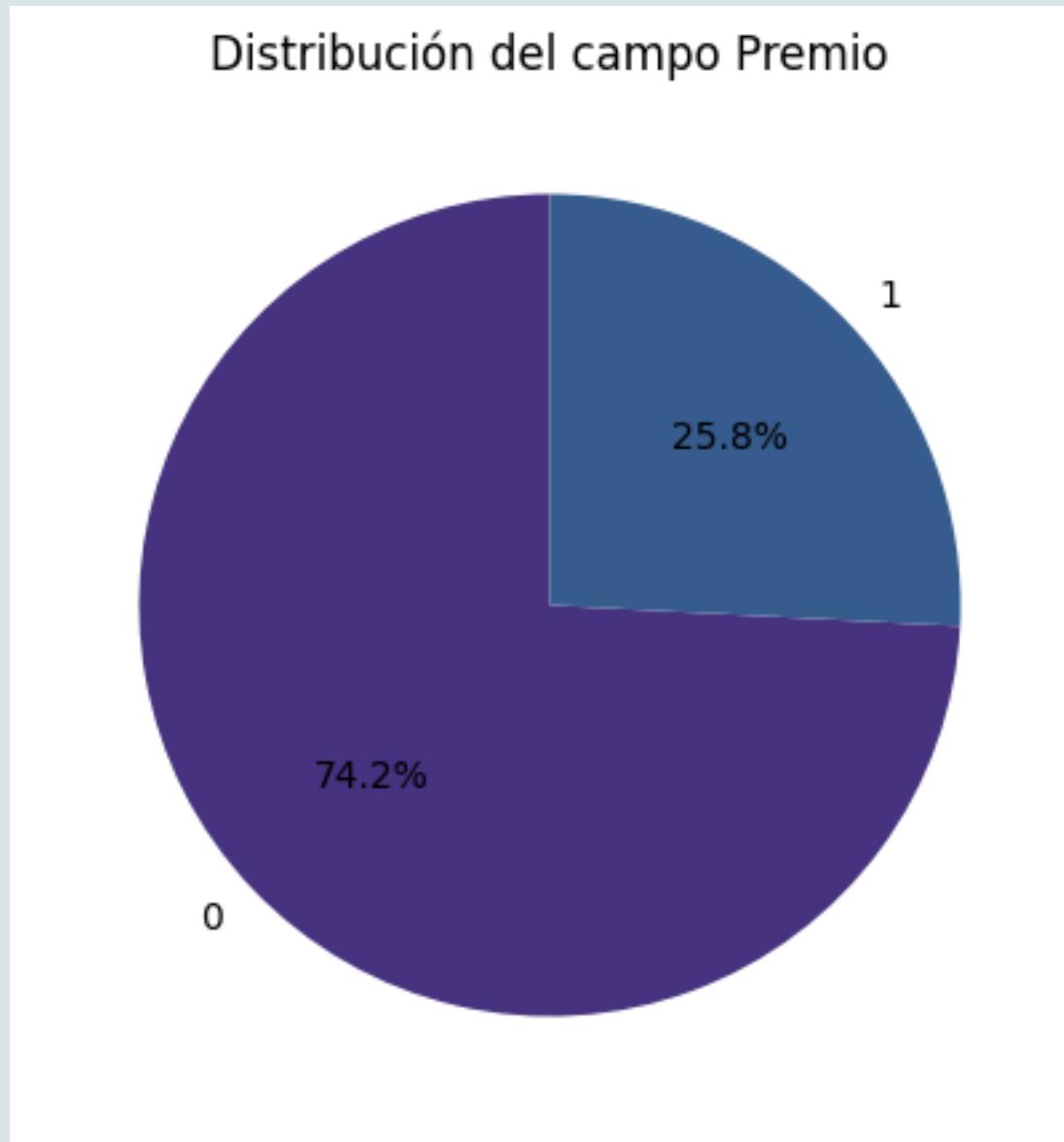
Análisis Exploratorio

Análisis Exploratorio



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

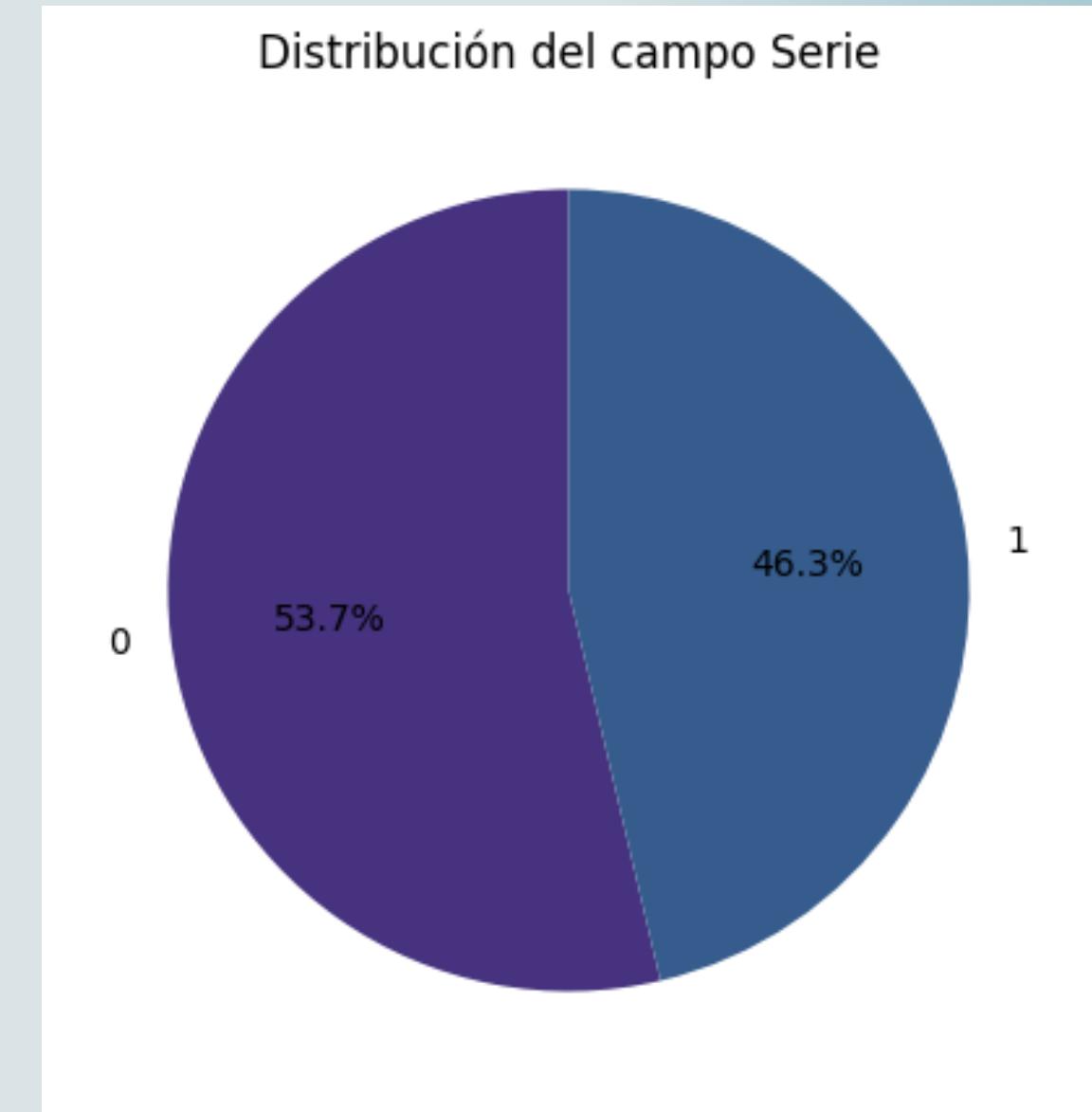
Análisis Exploratorio



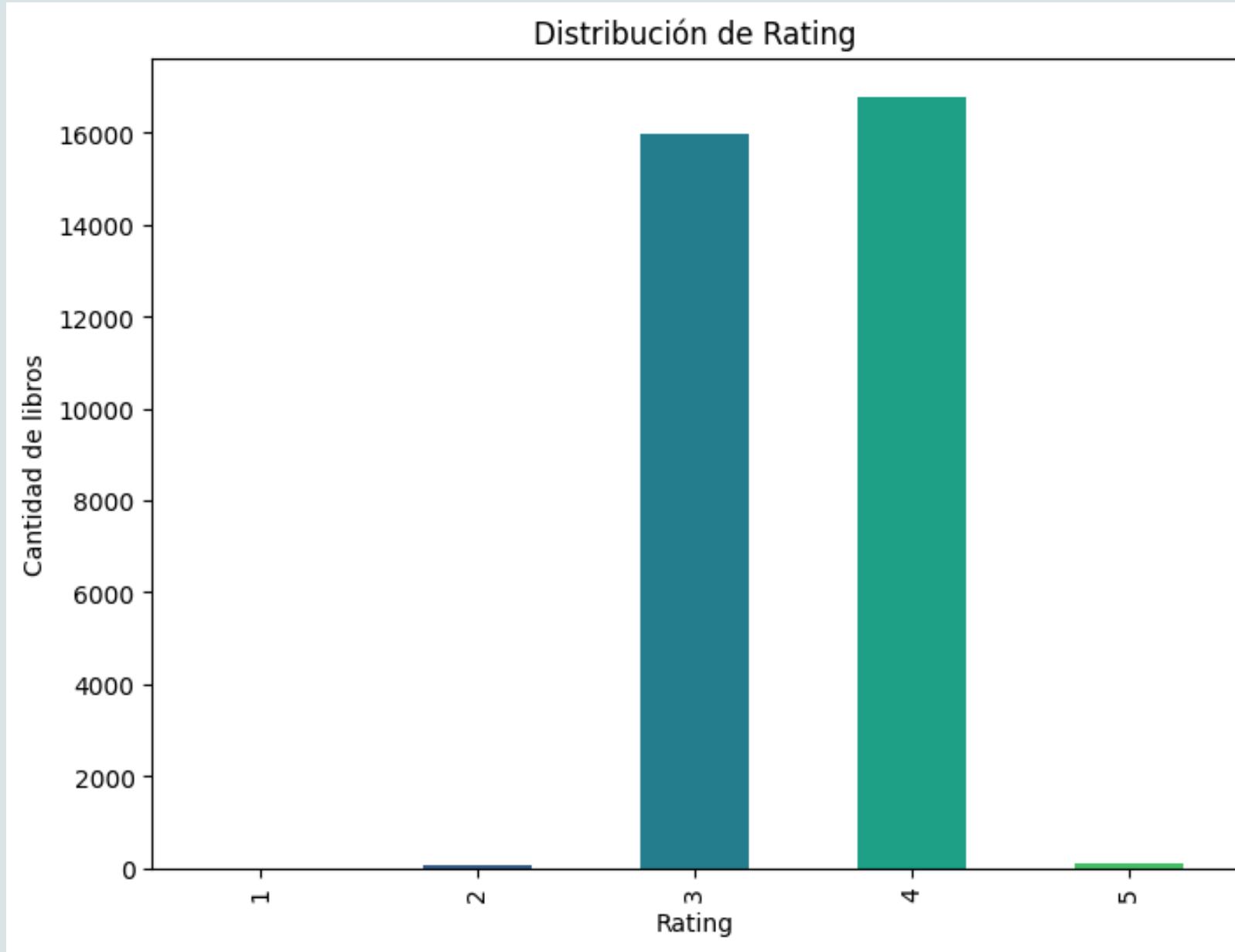
Más del 25% de los libros que se encuentran en el dataset tienen al menos un premio.

Análisis Exploratorio

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

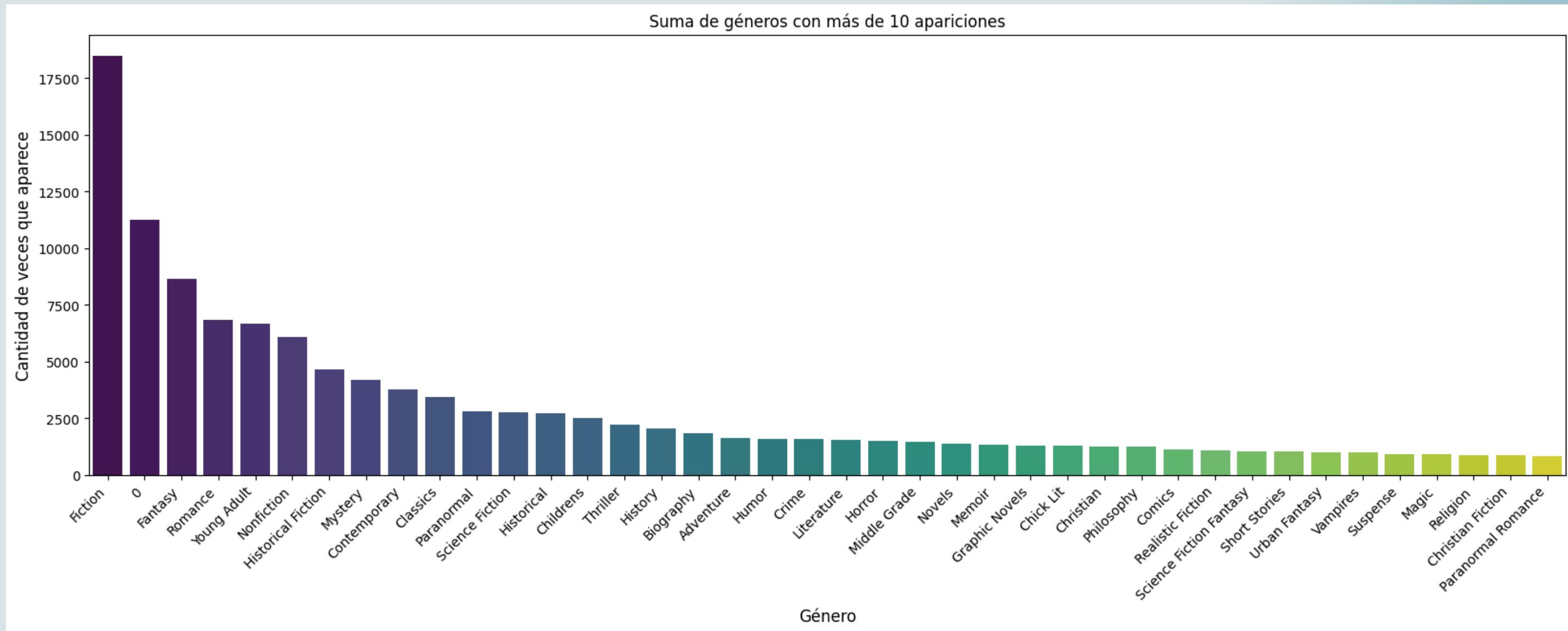
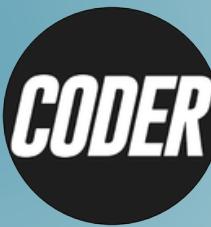


Análisis Exploratorio



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

Análisis Exploratorio

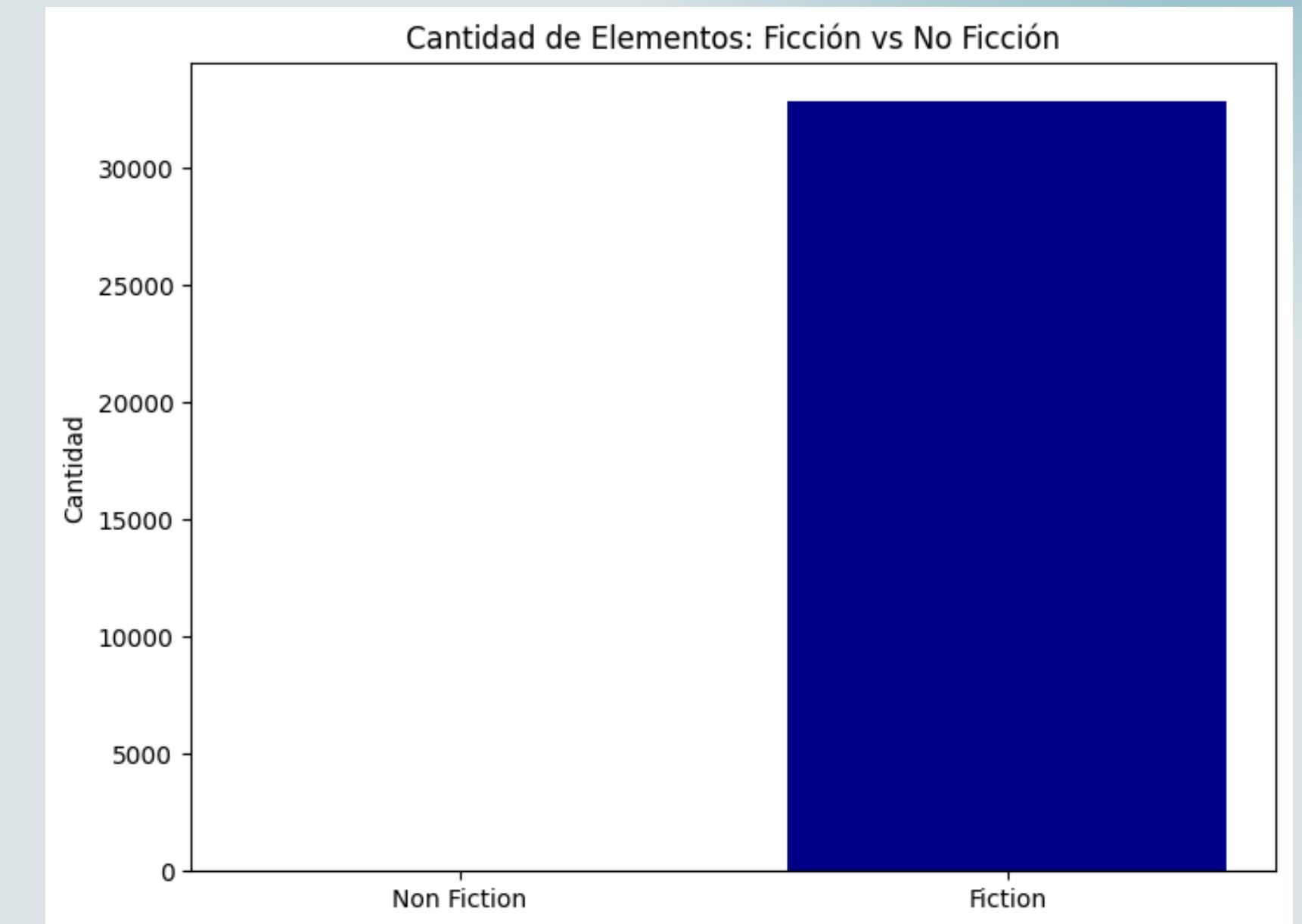


En este gráfico se pueden ver las principales generos de los libros que se encuentran dentro del dataset.

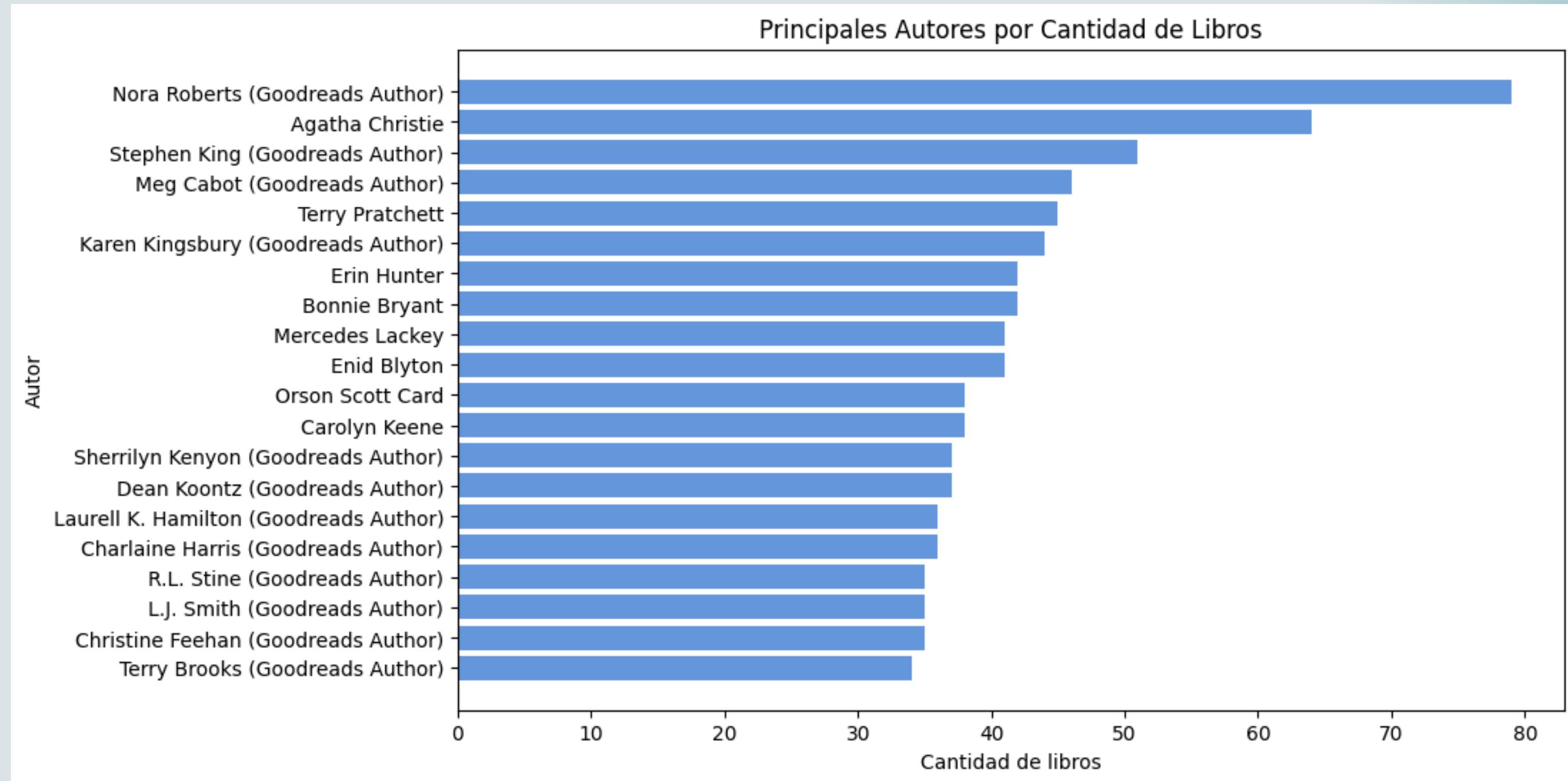
Análisis Exploratorio



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.

Próximos Pasos

- **Explorar otros modelos:** intentar usar otros algoritmos, como árboles de decisión, random forests.
- **Revisar los datos:** verificar que no haya outliers significativos y considera escalar los datos.
- **Hacer ingeniería de características:** añadir o transformar variables para ver si puedes mejorar el ajuste del modelo.
- **Validación cruzada:** verificar si el modelo se comporta de manera consistente con diferentes particiones de los datos.