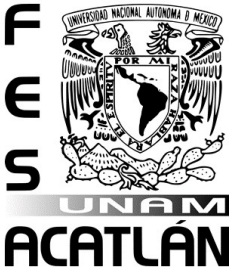



# Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

# Diplomado Ciencia de Datos

Proyecto final

Ambrocio Loreto Luis Manuel

# Índice

# 1. Introducción

Durante las últimas décadas, con el auge de Youtube, Amazon, Netflix y muchos otros servicios web similares, los sistemas de recomendación han ido ocupando cada vez más lugar en nuestras vidas. Desde el comercio electrónico (sugerir a los compradores artículos que podrían interesarles) hasta la publicidad en línea (sugerir a los usuarios los contenidos adecuados que coincidan con sus preferencias), los sistemas de recomendación son hoy inevitables en nuestros viajes diarios en línea.

De manera muy general, los sistemas de recomendación son algoritmos destinados a sugerir elementos relevantes a los usuarios (artículos como películas para ver, textos para leer, productos para comprar o cualquier otra cosa, dependiendo de las industrias).

Los sistemas de recomendación son realmente críticos en algunas industrias, ya que pueden generar una gran cantidad de ingresos cuando son eficientes o también ser una forma de diferenciarse significativamente de la competencia.

# 2. Dataset

El conjunto de datos de Book-Crossing consta de 3 archivos.

**Usuarios**
Contiene los usuarios. Tenga en cuenta que los ID de usuario (*User-ID*) se han anonimizado y se asignan a números enteros. Se proporcionan datos demográficos (*Location, Age*) si están disponibles. De lo contrario, estos campos contienen valores NULL.

**Libros**
Los libros se identifican por su respectivo ISBN. Los ISBN no válidos ya se han eliminado del conjunto de datos. Además, se proporciona cierta información basada en contenido (*Book-Title, Book-Author, Year-Of-Publication, Publisher*), obtenida de Amazon Web Services. Tenga en cuenta que en caso de varios autores, sólo se proporciona el primero. También se proporcionan URL que enlazan con imágenes de portada, que aparecen en tres tipos diferentes (*Image-URL-S, Image-URL-M, Image-URL-L*), es decir, pequeña, mediana y grande. Estas URL apuntan al sitio web de Amazon.

**Calificaciones**
Contiene la información de calificación del libro. Las calificaciones ( Book-Rating) son explícitas, expresadas en una escala del 1 al 10 (los valores más altos indican una mayor apreciación), o implícitas, expresadas en 0.

El conjunto de datos se obtuvo de kaggle: `https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset`

La siguiente gráfica muestra un subconjunto de la tabla de usuarios.

| User-ID | Location | Age |
|---|---|---|
| 1 | nyc, new york, usa | 20.0 |
| 2 | stockton, california, usa | 18.0 |
| 3 | moscow, yukon territory, russia | 12.0 |
| 4 | porto, v.n.gaia, portugal | 17.0 |
| 5 | farnborough, hants, united kingdom | 13.0 |

Cuadro 1: Datos de usuarios

La columna `Location` contenía información de la ciudad, estado y país. Se generaron estas tres variables (`v_ciudad`, `v_estado`, `v_pais`) y se eliminó la columna `Location`. Con este cambio, los datos quedaron de la siguiente forma:

| User-ID | Ciudad | Estado | País | Age |
|---|---|---|---|---|
| 1 | nyc | new york | usa | 20.0 |
| 2 | stockton | california | usa | 18.0 |
| 3 | moscow | yukon territory | russia | 12.0 |
| 4 | porto | v.n.gaia | portugal | 17.0 |
| 5 | farnborough | hants | united kingdom | 13.0 |

Cuadro 2: Datos de usuarios

Los datos de los libros se ven de la siguiente forma

| ISBN | Book-Title | Book - Author | Year | Publisher | Image-URL-S | Image-URL-M | Image-URL-L |
|---|---|---|---|---|---|---|---|
| 0195153448 | Classical Mythology | Mark P. O. Morford | 2002 | Oxford University Press | url | url | url |
| 0002005018 | Clara Callan | Richard Bruce Wright | 2001 | Harper Flamingo Canada | url | url | url |
| 0060973129 | Decision in Normandy | Carlo D'Este | 1991 | Harper Perennial | url | url | url |
| 0374157065 | Flu: The Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It | Gina Bari Kolata | 1999 | Farrar Straus Giroux | url | url | url |
| 0393045218 | The Mummies of Urumchi | E. J. W. Barber | 1999 | W. W. Norton & Company | url | url | url |

Cuadro 3: Datos de libros

En las ultimas van URL's pero no se colocaron por simplicidad

Los datos de rating se ven de la siguiente forma

| User-ID | ISBN | Book-Rating |
|---|---|---|
| 276725 | 034545104X | 0 |
| 276726 | 0155061224 | 5 |
| 276727 | 0446520802 | 0 |
| 276729 | 052165615X | 3 |
| 276729 | 0521795028 | 6 |

Cuadro 4: Datos de libros y calificaciones

Unimos las tres tablas y el total de columnas con su tipo de datos son las siguientes

| Columna | Tipo de Datos |
|---|---|
| User-ID | int64 |
| Age | float64 |
| ISBN | object |
| Book-Rating | int64 |
| Book-Title | object |
| Book-Author | object |
| Year-Of-Publication | object |
| Publisher | object |
| Image-URL-S | object |
| Image-URL-M | object |
| Image-URL-L | object |
| Ciudad | object |
| Estado | object |
| País | object |

Cuadro 5: Datos de tabla concentrada

Los datos se recolectaron en el año 2004, nuestro dataset concentrado se tienen 1031136 registros y 14 columnas.

# 3. Calidad de Datos

## 3.1. Etiquetado de variables

Se renombraron las variables a fin de tener la siguiente nomenclatura

- v_ : variables categóricas
- t_ : variables de texto
- c_ : variables numéricas

Las columnas resultantes son las siguientes

- **id_user**
- **c_age**
- **id_isbn**
- **target_rating**
- **t_book_title**
- **t_book_author**
- **c_year_of_publication**
- **v_publisher**
- **t_image_url_s**
- **t_image_url_m**
- **t_image_url_l**
- **v_ciudad**
- **v_estado**
- **v_pais**

## 3.2. Duplicidad

Nuestra unidad muestral es `id_user` y `id_isbn`. Se revisó que no hubiera duplicados para cada par (`id_user`, `id_isbn`), y no se encontraron datos duplicados.

## 3.3. Precisión /Orden

Se revisó la variable `c_age` y `id_user` y tenían el formato correcto. Posteriormente, se revisó la variable `c_year_of_publication` y se detectó que era de tipo 'Object' y que las columnas se encontraban desfasadas; es decir, el contenido que tenía correspondía a la columna siguiente (`v_publisher`). Se encontraron 4 registros con esta característica, y en los 4 casos, el contenido de `t_book_author` debía ir en `c_year_of_publication`, el valor de `c_year_of_publication` debía ir en `v_publisher`, etc. Además, el dato del autor se encontraba en `t_book_title`, después del nombre del libro. Por ejemplo, el nombre del libro era 'DK Readers: Creating the X-Men, How Comic Books Come to Life (Level 4: Proficient Readers)";James Bu'kley', esto se corrigió y se pasó `c_year_of_publication` a `Int64`.

Para las variables categóricas se hizo una limpieza básica del texto, pasando a minúsculas el texto, quitando espacios a los lados y solo dejando números y letras, eliminando todo tipo de caracteres como guiones, comas, etc.

Para `id_isbn`, se encontraron casos donde el ISBN no era correcto. `id_isbn` contiene principalmente números, pero también puede incluir 'X', y había casos donde había otras letras. Se encontraron 165 casos y estos se eliminaron, ya que no se podían imputar, dado que esto es el identificador principal de cada libro.

Para `v_ciudad`, `v_estado` y `v_pais`, se encontraron registros con el valor 'n/a'. Estos casos se cambiaron a `Null`.

Las variables `t_image_url_s`, `t_image_url_m` y `t_image_url_l`, como ya se mencionó, contenían la URL de la imagen de portada de los libros en formato pequeño, mediano y grande. Esta información no se usará, así que se eliminaron esas columnas. Las demás variables estaban correctas en cuanto a precisión y orden.

## 3.4. Consistencia

### 3.4.1. Variables Numéricas

Primero se revisó la variable de edad, se obtuvieron las siguientes estadísticas.

| Estadística | Valor |
|---|---|
| Recuentos | 753,157 |
| Media | 37.40 |
| Desviación Estándar | 14.10 |
| Mínimo | 0 |
| Percentil 25 % | 28 |
| Mediana (Percentil 50 %) | 35 |
| Percentil 75 % | 45 |
| Máximo | 244 |

Cuadro 6: Estadísticas de la variable c_age

Vemos un valor máximo de 244. Esto no es algo posible, ya que aunque existen personas que superan los 100 años, son muy pocas y es muy poco probable encontrar a personas de más de 100 años calificando libros. Por lo tanto, se consideraron inconsistentes aquellos registros con edades mayores a 100 años y se convirtieron a valores nulos (*Null*). Lo mismo se aplicó para la edad mínima. Aunque no hay una edad fija en la que una persona aprenda a leer, es poco probable que alguien menor de 7 años pueda leer un libro y calificarlo, por lo que estos casos también se convirtieron a *Null*.

Para la variable `c_year_of_publication`, obtuvimos las siguientes estadísticas.

| Estadística | Valor |
|---|---|
| Recuentos | 1,030,971 |
| Media | 1968.33 |
| Desviación Estándar | 230.55 |
| Mínimo | 0 |
| Percentil 25 % | 1992 |
| Mediana (Percentil 50 %) | 1997 |
| Percentil 75 % | 2001 |
| Máximo | 2050 |

Cuadro 7: Estadísticas de la variable c_year_of_publication

Se especifica que el año de recopilación de datos fue en 2004, por lo que no es consistente tener años de publicación mayores a 2004. También encontramos años con valor 0. En estos casos, los registros se cambiaron a valores nulos (*Null*). Las demás variables numéricas se encontraron correctas.

### 3.4.2. Variables Categóricas

Se calcularon las estadísticas para la variable `v_publisher`, y se encontró lo siguiente.

| Estadística | Valor |
|---|---|
| Recuento | 1,030,969 |
| Únicos | 16,120 |
| Valor más común | ballantine books |
| Frecuencia del valor más común | 34,724 |

Cuadro 8: Estadísticas de la variable v_publisher

Tenemos 16,120 editoriales diferentes y se encontraron 6,843 casos donde la editorial solo aparece una vez. Esto podría ser debido a errores de escritura. Para detectar errores de escritura, se revisó si había casos en los que existiera poca diferencia de texto entre las distintas editoriales. Para ello, se utilizó la distancia de Levenshtein. Esta distancia nos proporciona el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Por ejemplo, la distancia de Levenshtein entre çasaz çalle.$^{es}$ de 3, porque se necesitan al menos tres ediciones elementales para cambiar una en la otra. Aquellos casos con poca distancia implican que son muy similares y solo tienen algunos caracteres diferentes. Esta información se colocó en una tabla como la siguiente.

| v_publisher1 | v_publisher2 | distancia |
|---|---|---|
| harperperennial | harper perennial | 1 |
| putnam pub group | putnam pub group j | 2 |
| berkley publishing group | berkeley publishing group | 1 |
| random house | randon house | 1 |
| scribner | scribners | 1 |
| wonder books | wooden books | 2 |
| sams pub | haus pub | 2 |
| trollcommunications llc | troll communications llc | 1 |
| landoll s | landoll | 2 |
| lone star books | lodestar books | 2 |

Cuadro 9: Distancia Levenshtein entre las editoriales

Como se puede ver, hay casos en los que las editoriales son las mismas, pero se identifican como diferentes debido a errores de escritura. Para solucionar esto, sustituimos la editorial con más frecuencia en la editorial con menos frecuencia para aquellos casos en los que la distancia de palabras fuera menor o igual a 3. Al realizar este cambio, terminamos con 13,243 editoriales diferentes.

Para `v_ciudad`, encontramos 14,648 ciudades diferentes, y las que aparecían solo una vez se revisaron y se confirmó que existen como ciudades válidas, por lo que no fueron mal escritas o algo por el estilo. Sin embargo, para evitar tener información tan granular que no aporte mucho, consideramos que una opción adecuada sería normalizar por estado. Dado que esta columna ya existe, decidimos eliminar la columna `v_ciudad`.

En el caso de `v_estado`, se encontraron 1,953 estados diferentes. Se revisaron las distancias de palabras y, al igual que con `v_publisher`, se encontraron estados mal escritos. A continuación, se muestra un subconjunto de la distancia de palabras para `v_estado`.

| v_estado1 | v_estado2 | distancia |
|---|---|---|
| british columbia | brithish columbia | 1 |
| alberta | aberta | 1 |
| tennessee | tenessee | 1 |
| odense | orense | 1 |
| graubunden | graubuenden | 1 |

Cuadro 10: Distancia Levenshtein entre las editoriales entre los estados

Se realizó un proceso similar al de `v_publisher` para las variables `v_estado` y `v_pais`. En el caso de `v_estado`, sustituimos el estado menos frecuente por el más frecuente, pero solo se consideraron casos donde la distancia de palabras era 1. Al final, nos quedamos con 1,598 estados diferentes. Para `v_pais`, se siguió el mismo enfoque y también sustituimos los valores para reducir la variabilidad. Al hacerlo, pasamos de tener 373 a 101 países diferentes.

### 3.4.3. Variables de texto

Para `t_book_title` se revisó que el texto estuviera en ingles, nos interesan solo libros en ingles ya que nuestros modelos van a estar limitados a ingles, si hay libros en otro idioma los modelos no funcionarían bien, no se puede imputar ya que se requiere que el nombre del libro sea preciso, por lo tanto se opto por eliminar aquellos libros con otro idioma que no sea ingles, se eliminaron 281514 registros, que es poco más del 27 % del total de nuestros registros originales. Tanto para `t_book_title` como para `t_book_author` se aplicó lo siguiente:

1. Se eliminaron stopwords, estas son palabras muy comunes pero que no aportan mucho información del contexto del texto, por ejemplo 'el,ella,y,como'

2. Se eliminaron HAPAXE, estas son palabras que aparecen solo una vez en todo el texto, de igual forma no aportarían información valiosa.

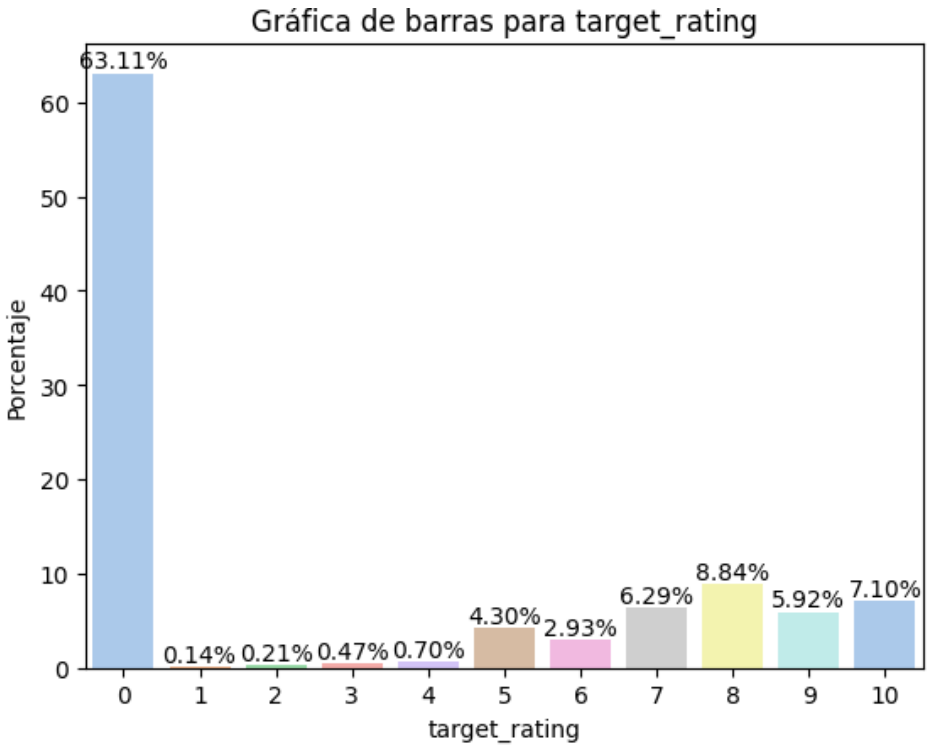
3. Se pasaron a null aquellos cambios vacíos

## 3.5. Normalización

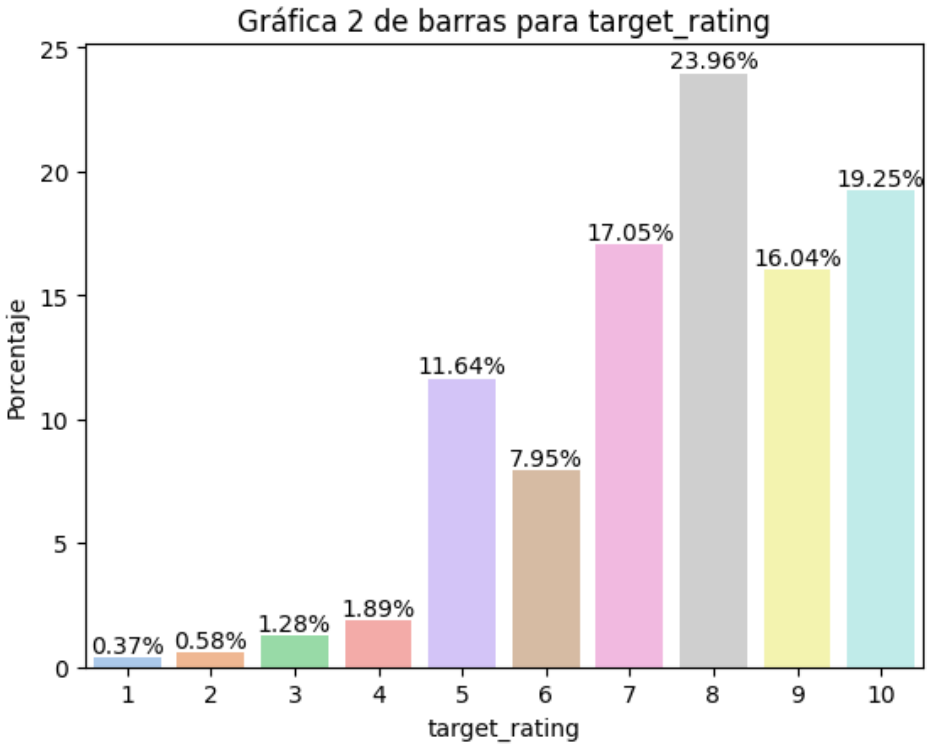
Se empezó revisando la variable `v_publisher`, teniendo en cuenta que una editorial publica varios libros, y que la mayoría de los libros en este dataset son calificados más de una vez, sería bastante raro encontrar editoriales con menos de 10 apariciones, todos estos casos se agruparon en una categoría 'Otros', se encontraron 5152 editoriales que se normalizaron a la categoría 'Otros', al final nos quedamos con 1997 diferentes editoriales. Para `v_estado` y `v_pais` aplicamos la misma lógica, para estado normalizamos a 'Otros' aquellos estados donde aparecían menos de 20 veces en todo el dataset y para país se normalizaron a 'Otros' aquellos países donde aparecían menos de 60 veces, al final nos quedamos con 341 estados diferentes y 52 paises.

# 4. Análisis Exploratorio

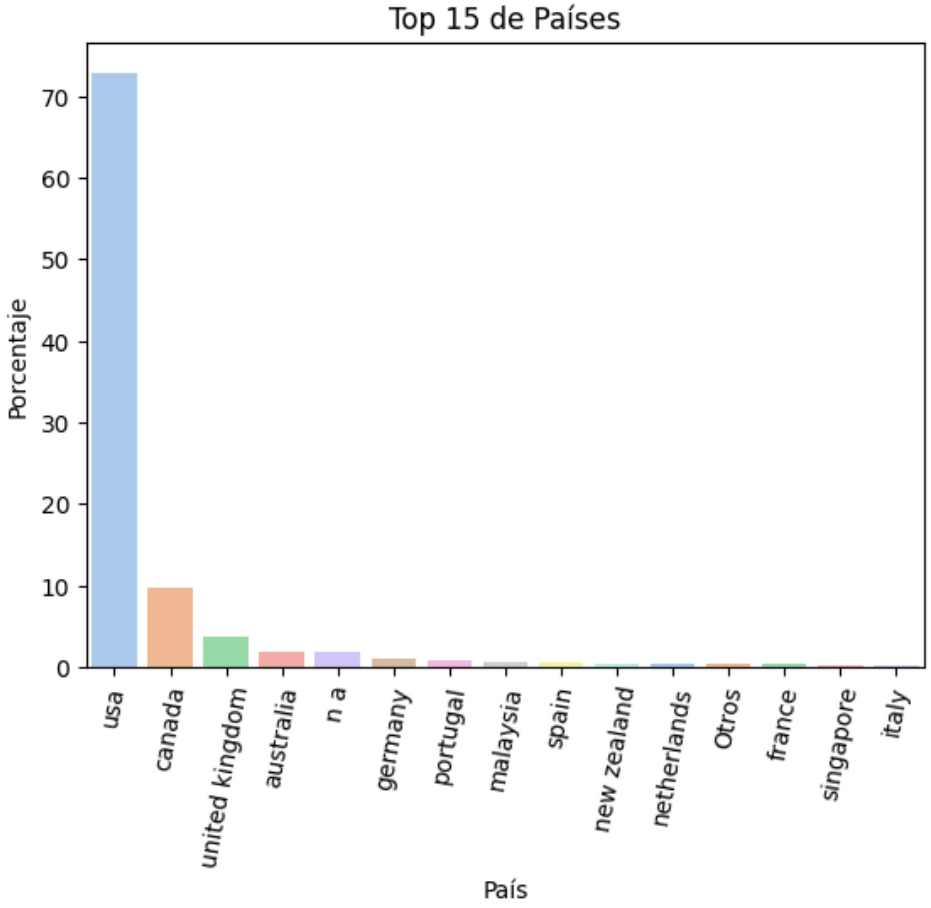
El siguiente gráfico representa un diagrama de barras que visualiza la variable `target_rating`. Los valores de 0 en `target_rating` indican que el libro no ha recibido una calificación. Observamos que el 63 % de los libros no ha sido calificado. La presencia de un gran número de registros sin calificación podría sesgar los resultados, por lo tanto, sería más apropiado eliminar estos registros, incluso si esto resultara en una cantidad menor de registros totales. En este escenario, nos quedaríamos con más de 200,000 registros, lo cual sigue siendo una cantidad considerable y robusta para nuestro análisis. Esta acción contribuiría a obtener resultados más precisos y significativos.



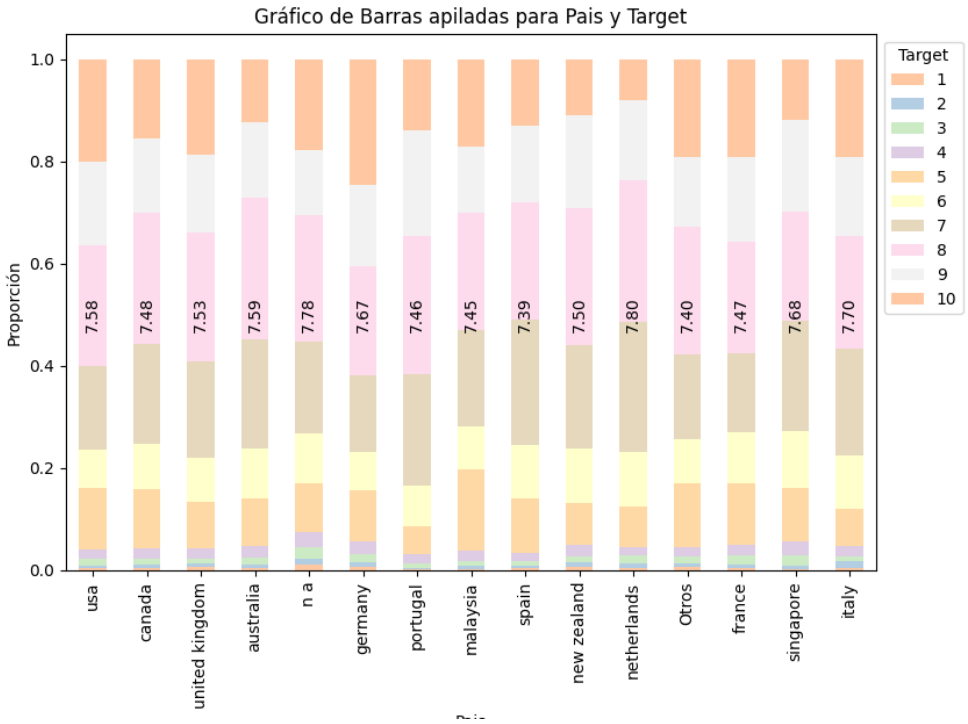

En relación al conjunto de datos, se procedió a eliminar el 63 % de las observaciones que presentaban el mismo valor. Con los datos actualizados, se observa claramente que la calificación más común es de 8. En general, entre las personas que proporcionan calificaciones, son pocas las que otorgan una valoración baja.



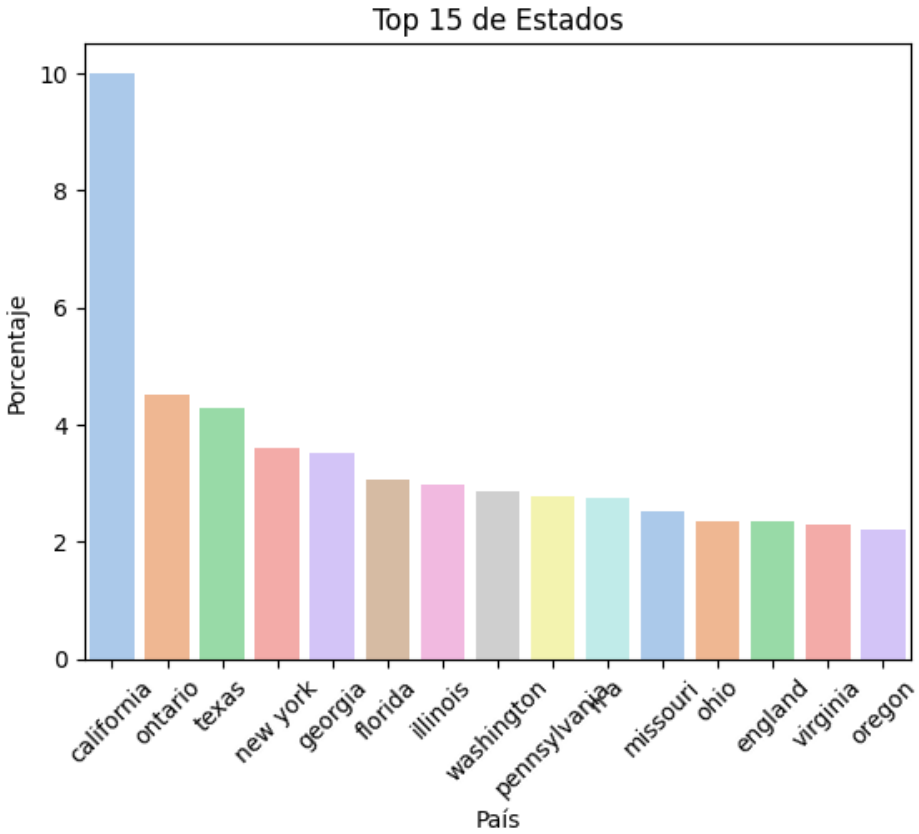

La gráfica siguiente representa el top 5 de países que aparecen con mayor frecuencia en nuestro conjunto de datos. Es evidente que la mayor cantidad de información disponible proviene de los Estados Unidos.



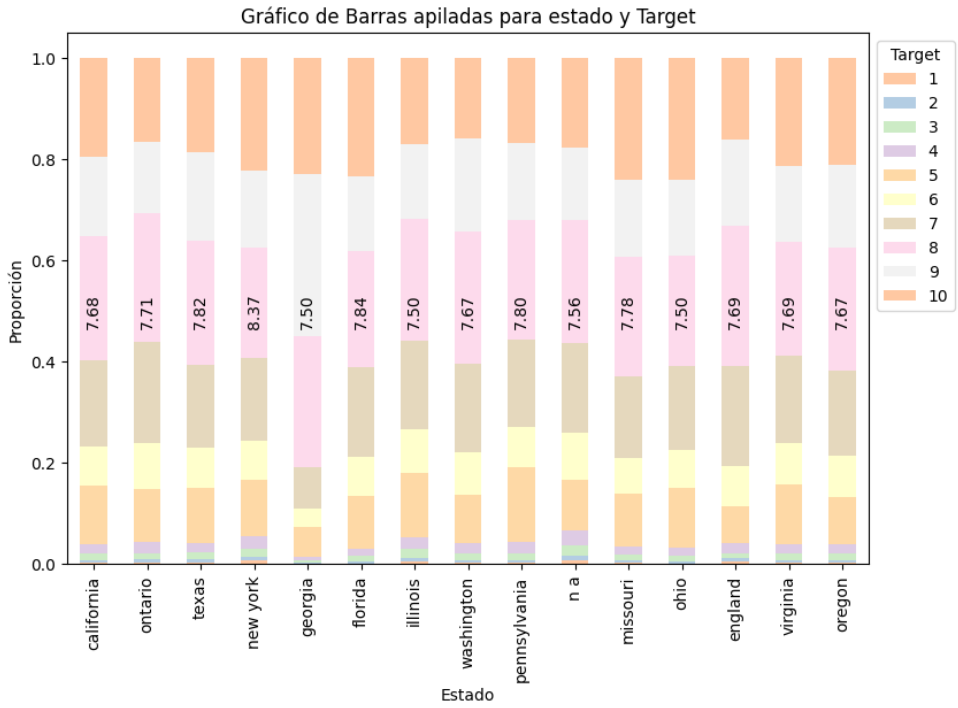

La siguiente gráfica ilustra la proporción de cada calificación en los quince países principales de nuestro conjunto de datos. Aunque no se aprecia un patrón claramente definido, es notable que las calificaciones varían entre cada país.



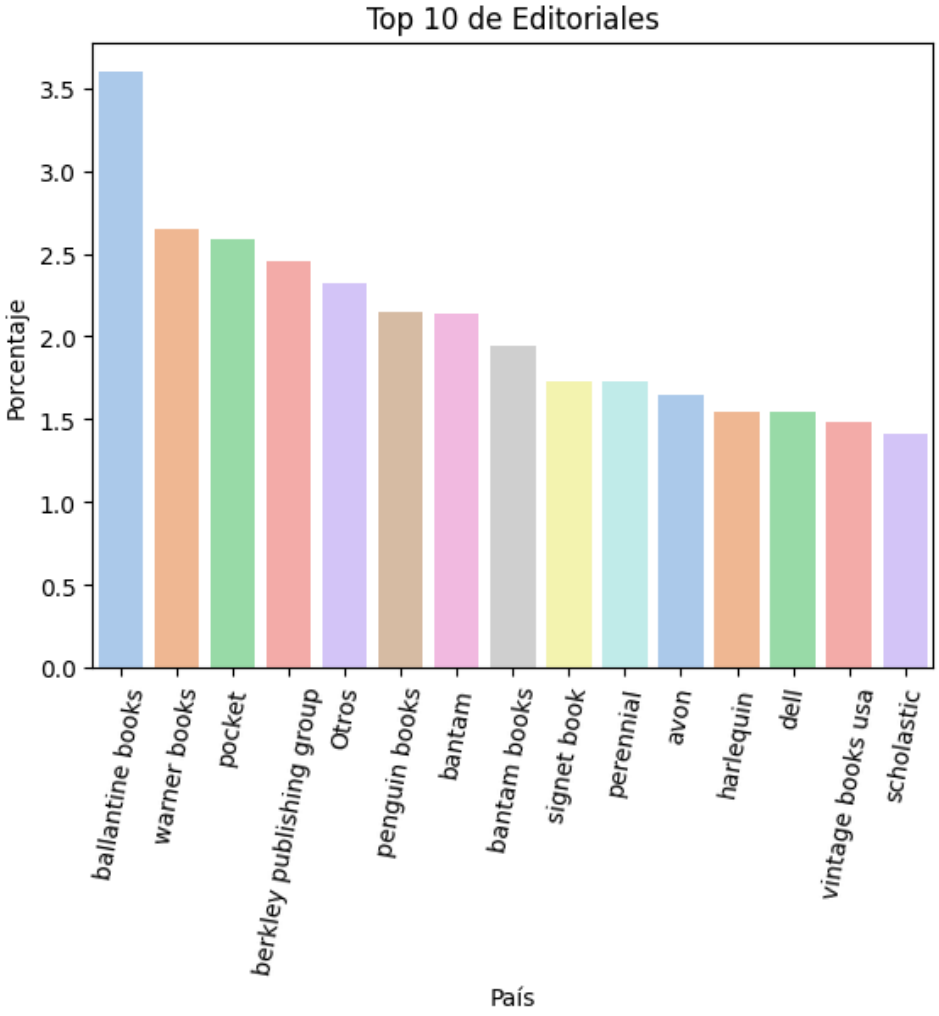
En el caso de los estados, se puede observar que los quince estados más frecuentes en nuestro conjunto de datos están principalmente ubicados en Estados Unidos.



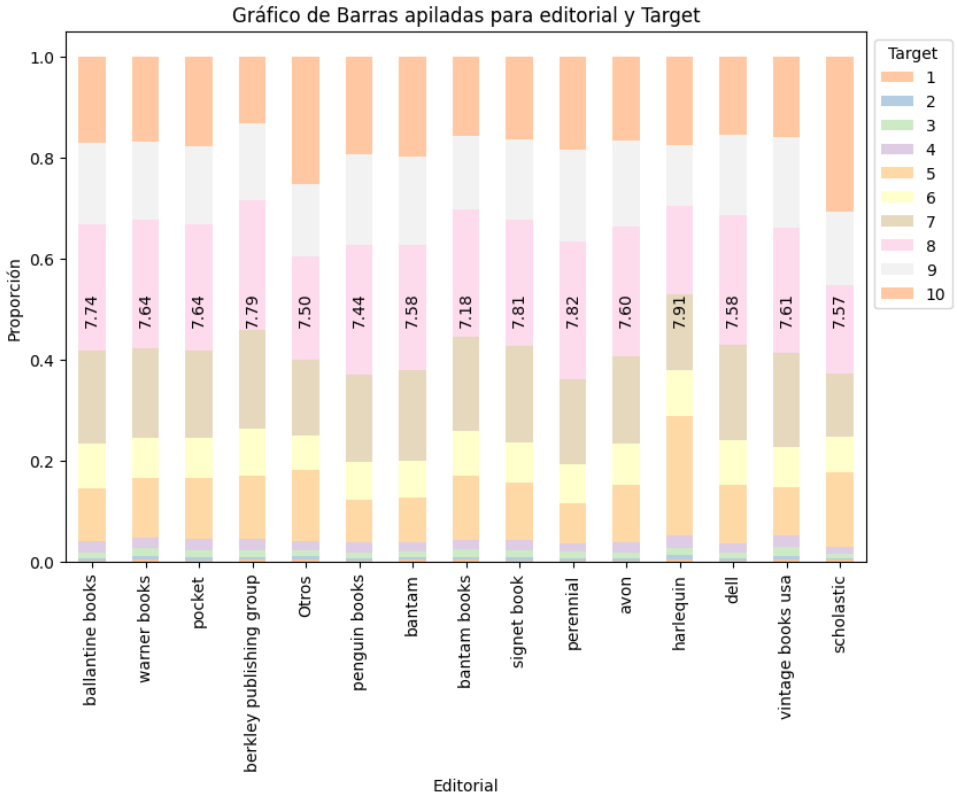

En el gráfico de barras apiladas que compara el estado y la variable objetivo, se confirma que, al igual que con los países, existen diferencias notables entre cada estado en términos de la variable objetivo.



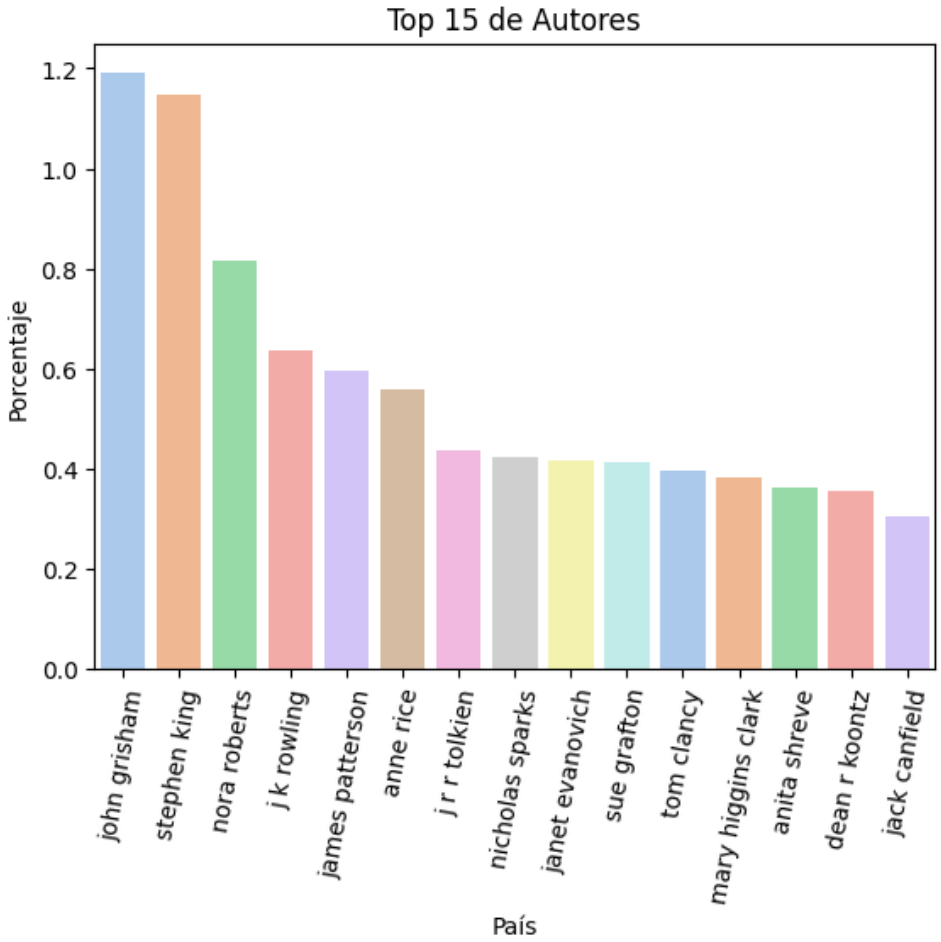

La gráfica siguiente representa el top 15 de las editoriales que más apariciones tienen en nuestro conjunto de datos.



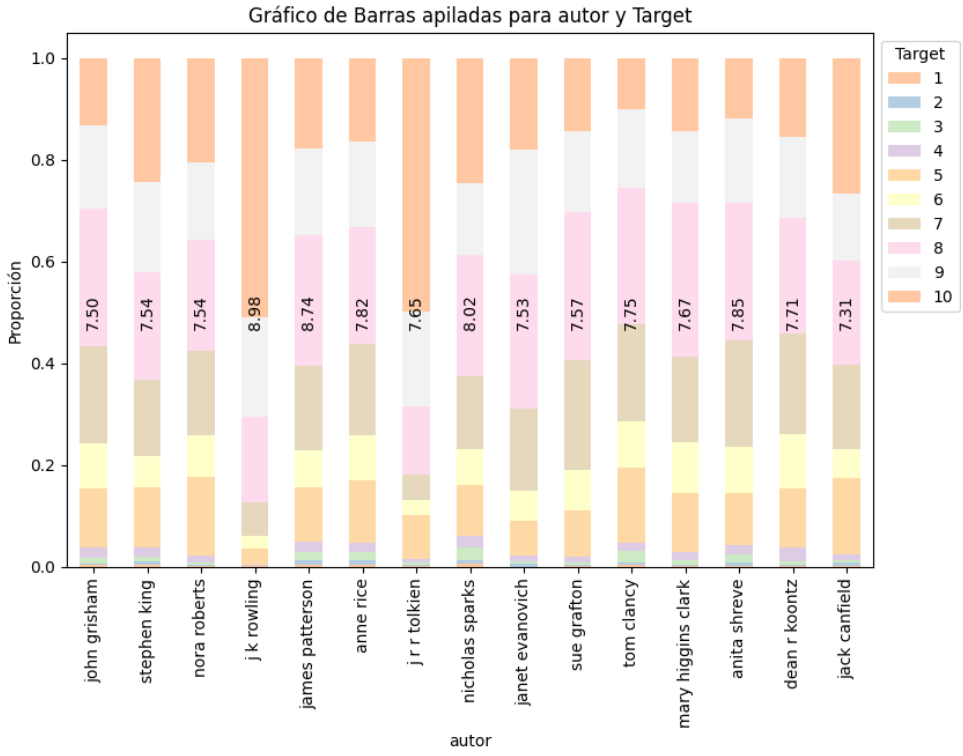

En el gráfico de barras apiladas que compara las editoriales con la variable objetivo, se evidencian diferencias notables en cómo se distribuye la variable objetivo entre las quince editoriales principales.



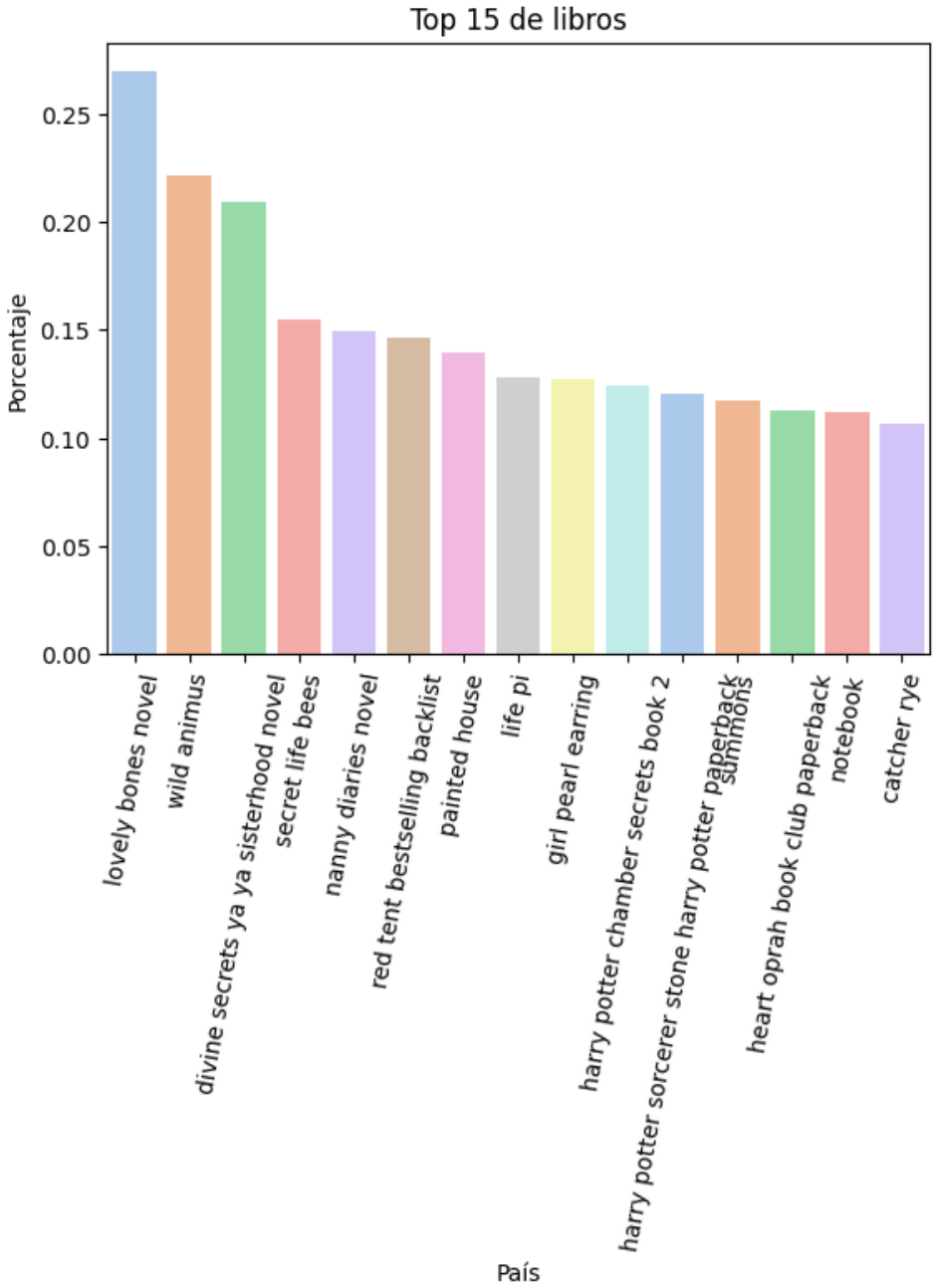

La siguiente gráfica presenta el top 15 de autores que aparecen con mayor frecuencia en nuestro conjunto de datos.



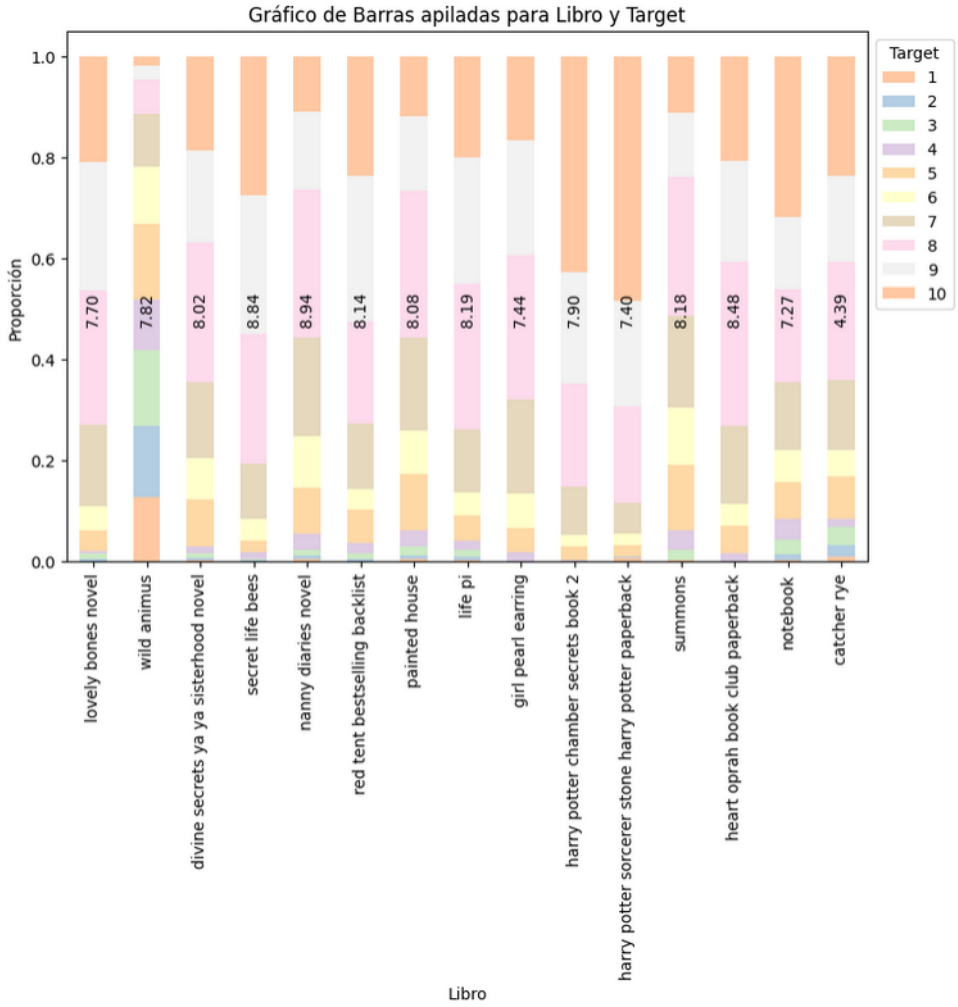

En el gráfico de barras apiladas que relaciona los autores con la variable objetivo, se aprecian diferencias más marcadas. Por ejemplo, en el caso de "J. R. R. Tolkien", se observa que casi el 50 % de las calificaciones son de 10.



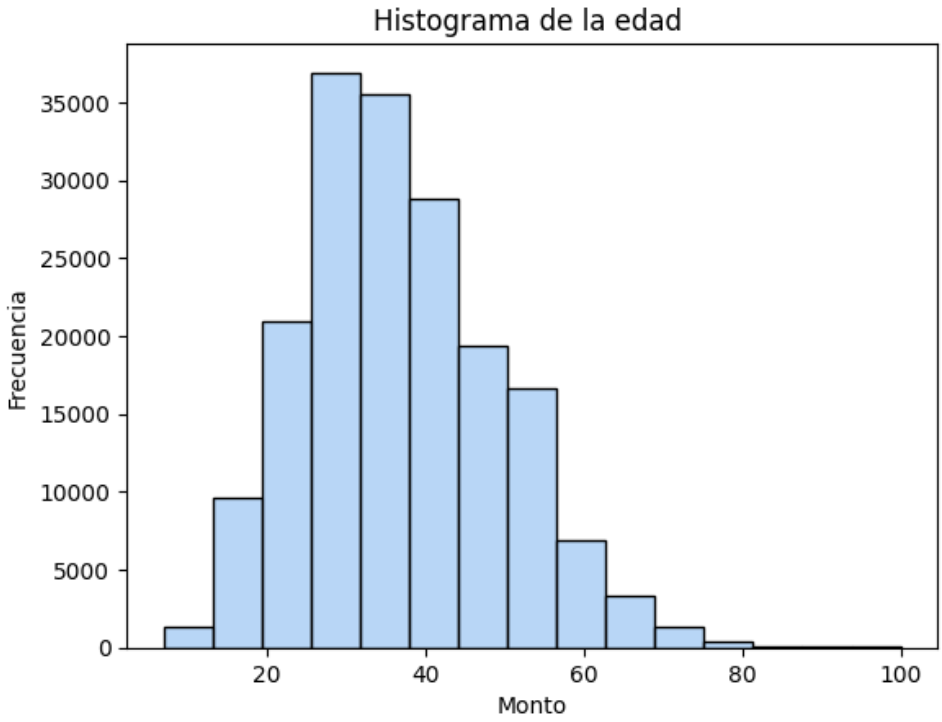

La siguiente gráfica presenta el top 15 de los libros que tienen más apariciones en nuestro conjunto de datos.



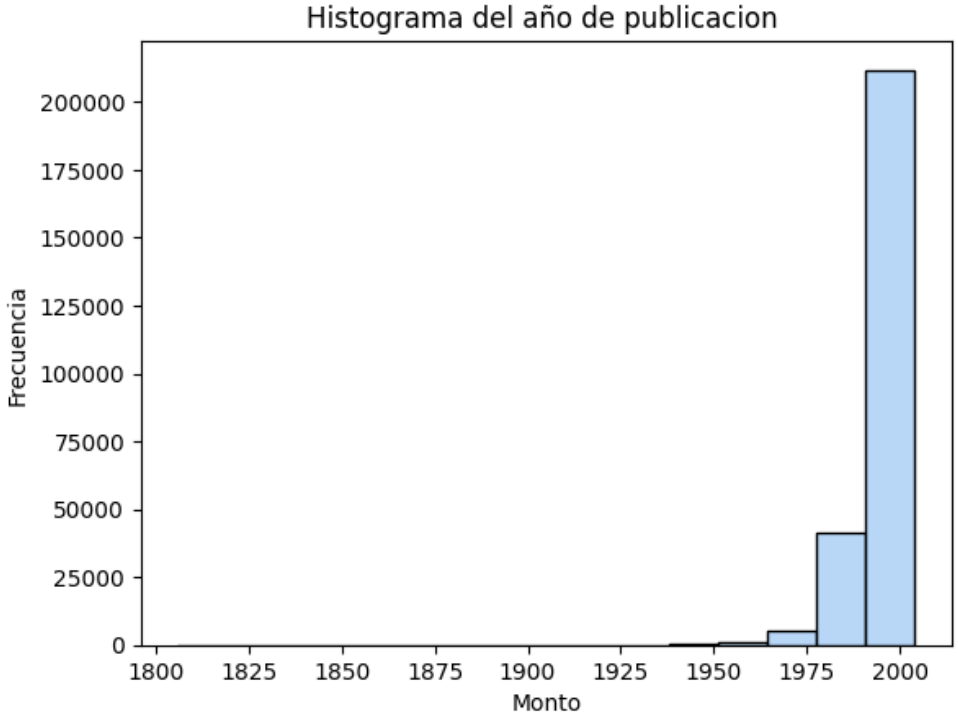

En el caso de los títulos de los libros, las diferencias son aún más notables. Por ejemplo, "Wild Animus", a pesar de ser uno de los más leídos, presenta calificaciones bajas en general.



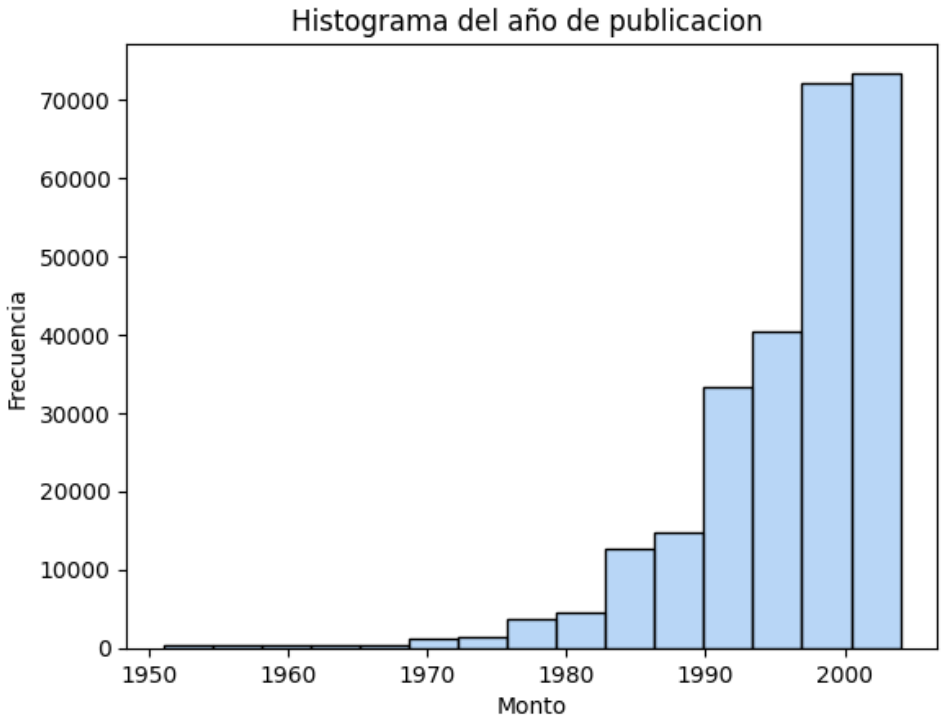

El histograma que muestra la distribución de edades revela que la mayoría de las personas que respondieron se encuentran en el rango de edades comprendido entre los 20 y los 60 años. De hecho, el pico más pronunciado de la distribución se ubica aproximadamente alrededor de los 35 años. Además, se observa una ligera simetría hacia la derecha en la distribución.




El histograma que representa el año de publicación muestra claramente que la mayor concentración de registros se sitúa en torno a la década de los 2000. Este patrón indica que los lectores tienden a preferir libros más recientes en su elección de lectura.

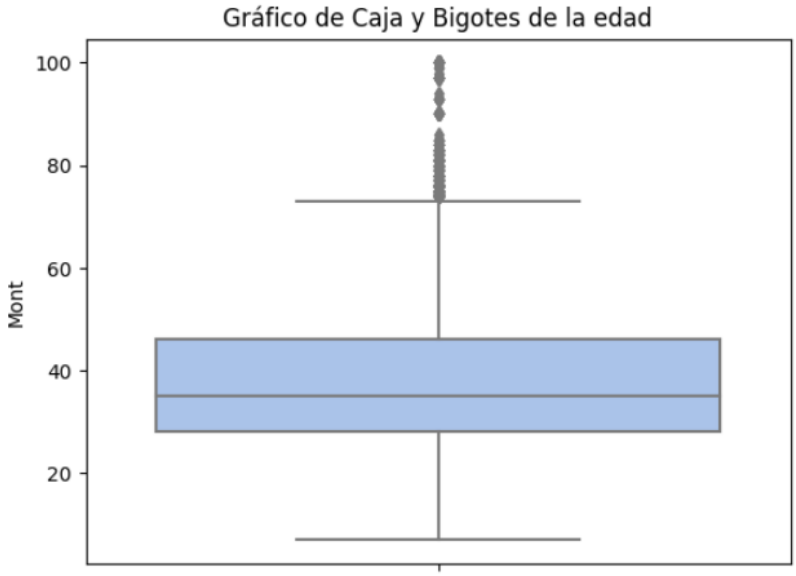


El histograma que muestra el año de publicación ofrece una visión más clara cuando se eliminan los posibles datos atípicos.

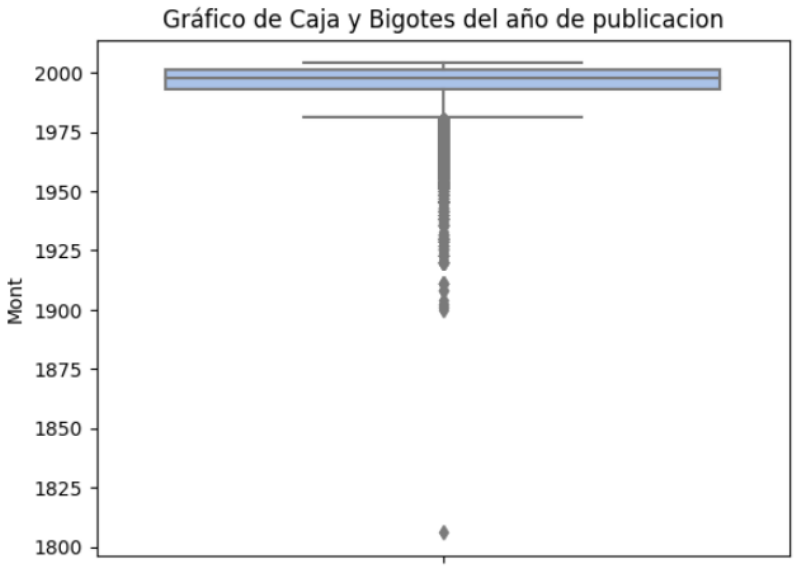


# 5. Datos atípicos

Para revisar datos atípicos empezamos revisando las gráficas de caja y bigotes de `c_age` y `c_year_of_publication`, primero para `c_age` observamos muy pocos registros por arriba de 60 años, una edad de 60-100 años es una edad correcta, sin embargo, hay muy pocos registros de ese rango de edad y esto podría sesgar los resultados




En `c_year_of_publication` sin más evidentes los datos atípicos, la mayoría de registros estan entre 1975 y 2004, hay muy pocos registros por debajo de 1975, de nuevo aunque la fecha sea correcta el mantener estos registros podría sesgar los resultados




Para determinar cuales son datos atípicos se aplicaron 3 métodos:

- IQR
- Método de percentiles
- Mean change

Cada método dio resultados diferentes sobre cuales registros son atípicos, por lo que se considero como atípicos aquellos registros que más de un método clasificara como atípico, en las siguiente tablas resumimos la información y observamos cuantos atípicos detecto cada método para cada variable, así como el porcentaje que esto representa del total de datos, se muestra también los registros que si se consideran como atípicos y el porcentaje de atípicos finales para cada variable.

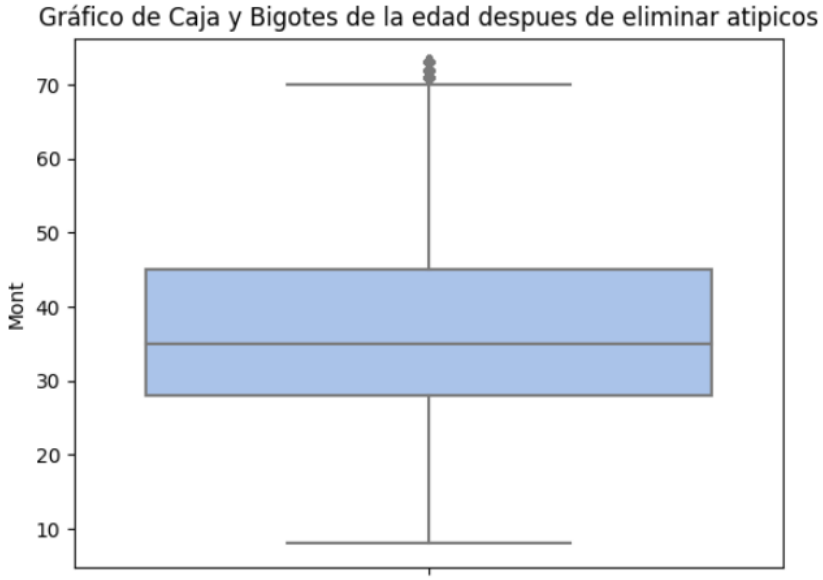
Cuadro 11: Número de valores atípicos en características

| **Variables** | **n_out_IQR** | **n_out_Percentil** | **n_out_Mean_Change** | **total_out** |
|---|---|---|---|---|
| c_year_of_publication | 9561 | 19974 | 1 | 9561 |
| c_age | 898 | 15424 | 68 | 830 |

Cuadro 12: Porcentaje de valores atípicos en características

| **Variables** | **n_out_IQR %** | **n_out_Percentil %** | **n_out_Mean_Change %** | **%out** |
|---|---|---|---|---|
| c_year_of_publication | 3.64 % | 7.61 % | 0.00 % | 3.64 % |
| c_age | 0.30 % | 5.88 % | 0.03 % | 0.32 % |

Para `c_year_of_publication` se detectaron 9561 registros atípicos lo que corresponde a un 3.64 % y para `c_age` se detectaron 830 que corresponde a 0.32 % de los datos, estos registros se elimina por lo cual se estaría eliminando un 3.96 % de los datos, después realizamos de nuevo las gráficas de caja y bigotes de `c_age` y `c_year_of_publication`, observamos en ambas gráficas que ya no hay datos atípicos



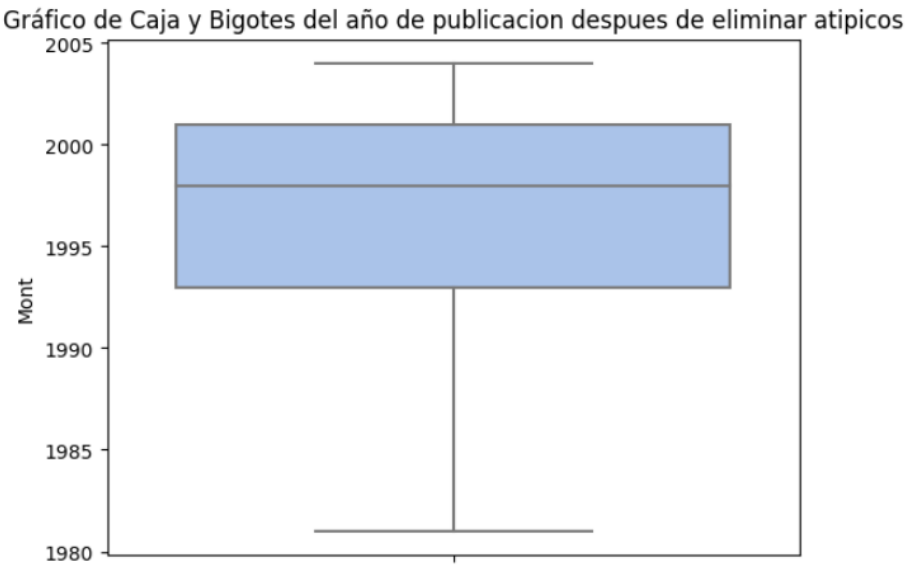


Gráfico de Caja y Bigotes del año de publicacion despues de eliminar atipicos
Mont
2005
2000
1995
1990
1985
1980

# 6. Datos Null

A partir de este momento se hizo un partición entre conjunto de entrenamiento y conjunto de prueba de los datos, los conjuntos se generaron mediante una división estratificada a partir de la target (es decir las proporciones en la target se mantuvieron en cada conjunto) y el conjunto de entrenamiento se quedo con el 70 % mientras que el conjunto de prueba se quedo con el 30 %, hasta este punto teníamos 252046 registros y el conjunto de entrenamiento se quedo con 176432 registros y el conjunto de prueba con 75614 registros. A continuación se muestra el total de registros con datos nulos por variable en el conjunto de entrenamiento y el conjunto de prueba.

Cuadro 13: Resumen de datos faltantes

| **Columna** | **nulos_train** | **nulos_test** | **pct_nulos_train** | **pct_nulos_test** |
|---|---|---|---|---|
| id_user | 0 | 0 | 0.000000 | 0.000000 |
| c_age | 9561 | 9358 | 0.311752 | 0.309546 |
| id_isbn | 0 | 0 | 0.000000 | 0.000000 |
| target_rating | 0 | 0 | 0.000000 | 0.000000 |
| t_book_title | 22 | 23 | 0.000935 | 0.001018 |
| t_book_author | 0 | 0 | 0.000006 | 0.000000 |
| c_year_of_publication | 1233 | 1236 | 0.011336 | 0.011149 |
| v_publisher | 0 | 0 | 0.000000 | 0.000000 |
| v_estado | 2444 | 2471 | 0.023624 | 0.024202 |
| v_pais | 2896 | 2888 | 0.028169 | 0.028077 |

Observamos que la variable c_age tiene más del 30 % de datos faltantes, también se observa que las otra variables con datos faltantes son categóricas o de texto, consideramos que para este problema aporta más información crear una categoría llamada 'sin informacion' que imputar con la moda (que es el método más común para variables categóricas), para las variables de texto, de igual forma, aporta más información especificar que no se tiene información a tratar de imputar, en el caso de la variable de edad se reconstruyó la variable a categórica y después se aplico lo mismo que para las otras categóricas, para reconstruir la variable se generaron las siguientes 4 categorías.

- <20
- 20-40
- 40-60
- >60

Después se cambio el nombre a v_age y se creo la categoría 'sin informacion' para los datos nulos. Para la variable c_year_of_publication se hizo una transformación, considerando que los datos se obtuvieron en el 2004, se la diferencia entre la recopilación de datos y el año de publicación, a un modelo ML no es buena idea pasarle una variable de años, ya que en un futuro no sabrá que hacer con los años nuevos (por ejemplo hoy 2023, el modelo nunca recibió ese dato), la antigüedad del libro en cambio, si es un valor que se puede generalizar, se creo la variable v_antiguedad_book y después se crearon las siguientes categorías

- <5
- 5-10
- 10-15
- 15-20
- >20

Después se agrego la categoría 'sin informacion' para los datos nulos y se elimino c_year_of_publication.

# 7. Ingeniería de variables

Hasta este punto ya hicimos varias cosas correspondientes a ingeniería de variables, como lo es calcular la antigüedad del libro o convertir variables numéricas a categóricas, no se puso en esta sección ya que era necesario hacerlo antes. Primero se manipulo la variable t_book_title, queríamos tener una representación numérica a 3 di menciones de los datos, la idea que se tuvo fue calcular los embedding de cada palabra de cada nombre de libro, para eso se utilizo un modelo pre entrenado que se descargo en la siguiente liga `https://fasttext.cc/docs/en/english-vectors.html`, al aplicar el modelo nos regresa un vector de 300 dimensiones por cada palabra, si el libro se llama 'neuromancer remembering tomorrow' obtendriamos una matriz de $3 \times 100$, con esta representación lo que se hizo después fue promediar los 3 vectores a fin de tener solo un vector de 300-D, a partir de esta información para obtener una dimensión 3D se aplico PCA (es un método de reducción de dimensiones) manteniendo solo 3 componentes principales, las variables creadas a partir de este método son:

- c_embedding_book1
- c_embedding_book2
- c_embedding_book3

Para t_book_author también queríamos una representación 3D, lo que hicimos fue promediar los embedding de todos los libros que ha escrito cada autor, por ejemplo si william gibson ha escrito 1000 libros, obteníamos la representación 3D del titulo de los 1000 libros, luego eso lo promediamos a fin de tener una representación numérica de cada autor, no se utilizo el mismo metodo que en t_book_title porque los embedding no trabajan bien en nombres, las variables creadas a partir de este método son:

- c_embedding_autor1
- c_embedding_autor2
- c_embedding_autor3

Para c_estado y c_pais promediamos el valor de target_rating y creamos 2 nuevas variables con esa información, las variables creadas a partir de este método son:

- c_mean_rate_estado
- c_mean_rate_pais

Para v_antiguedad_book y v_age se aplico one-hot encoding, es un método para representar numéricamente a variables categóricas, se crea una columna por categoría y si por cada registro se coloca 1 en la categoría que originalmente tenia y 0 en las demás, funciona como un indicador para mostrar a que categoría pertenece asignando 1 a dicha categoría, las variables creadas a partir de este método son:

- v_age_40-60
- v_age_¡20
- v_age_¿60

- v_age_sin informacion
- v_antiguedad_book_15-20
- v_antiguedad_book_5-10
- v_antiguedad_book_¡5
- v_antiguedad_book_¿20
- v_antiguedad_book_sin informacion

Finalmente las variables id_user y id_isbn se establecieron como los índices de nuestro conjunto de datos.

# 8. Datos Finales

Cuadro 14: Resumen de datos finales

| Conjunto | Registros | Columnas |
|---|---|---|
| Conjunto de entrenamiento | 176432 | 18 |
| Conjunto de prueba | 70633 | 18 |

Cuadro 15: Descripción de las columnas

| **Columna** | **Tipo de dato** |
|---|---|
| target_rating | int64 |
| c_embedding_book1 | float64 |
| c_embedding_book2 | float64 |
| c_embedding_book3 | float64 |
| c_embedding_autor1 | float64 |
| c_embedding_autor2 | float64 |
| c_embedding_autor3 | float64 |
| c_mean_rate_estado | float64 |
| c_mean_rate_pais | float64 |
| v_age_40-60 | float64 |
| v_age_¡20 | float64 |
| v_age_¿60 | float64 |
| v_age_sin_informacion | float64 |
| v_antiguedad_book_15-20 | float64 |
| v_antiguedad_book_5-10 | float64 |
| v_antiguedad_book_¡5 | float64 |
| v_antiguedad_book_¿20 | float64 |
| v_antiguedad_book_sin_informacion | float64 |