

# ANÁLISIS DE RESEÑAS LITERARIAS

---

Scarramberg, Julián  
Proyecto final  
Septiembre 2023

# Contenido

Descripción de la temática de los datos.....	3
Tabla de versionado .....	4
Hipótesis de trabajo .....	5
Objetivo del proyecto.....	5
Alcance .....	5
Usuario final y nivel de aplicación.....	5
Herramientas tecnológicas .....	5
Diagrama entidad-relación.....	6
Listado de tablas .....	7
Listado de columnas de tablas .....	7
Transformaciones de datos .....	9
Medidas .....	11
Modelo de datos relacional .....	13
Segmentaciones .....	14
Visualización de datos.....	15
Descripción de las solapas .....	15
Futuras líneas .....	19

## Descripción de la temática de los datos

Con los avances tecnológicos de las últimas décadas y la difusión de las herramientas originadas por estos, como el internet y el *smartphone*, la gran mayoría de la población mundial encuentra en la actualidad a disposición todo tipo de contenido y diferentes maneras de acceder al mismo. Uno de los puntos de acceso al contenido más popularizados en el último tiempo ha sido el proporcionado por el gigante estadounidense Amazon.

Amazon es una empresa dedicada principalmente al comercio electrónico originada en los Estados Unidos que ha logrado popularizarse y alcanzar a millones de personas. Fundada en 1994 por el empresario estadounidense Jeff Bezos, fue concebida en un principio como una plataforma que permitiría la venta y compra de libros. Con el tiempo la multinacional se ha expandido y, a la par que encaraba exitosamente una gran variedad de otras actividades, se ha convertido en, tal vez, el medio más importante de comercialización de obras literarias en el mundo.

Con esto en mente es lógico pensar que la mejor y más completa fuente de información para realizar un análisis sobre literatura, y en este caso en particular sobre reseñas literarias, sería la proporcionada por Amazon. Así el proyecto se sostendrá sobre una base de datos provista por dicha empresa.

La base de datos en cuestión proporciona información sobre casi trescientos mil libros y detalles de los mismos, información sobre usuarios y más de un millón de reseñas realizadas por los mismos.

## Tabla de versionado

<b>Versión</b>	<b>Fecha</b>
V1	29/06/23
V2	17/07/23
V3	21/08/23
V4	14/09/23

## Hipótesis de trabajo

Para encarar el proyecto me basaré en la hipótesis de que existe una relación entre ciertos atributos de los usuarios, como su edad y localización, y las reseñas que dan a ciertos libros, géneros y autores.

## Objetivo del proyecto

Como objetivo general se buscará realizar un análisis de las características de los libros y usuarios y también se intentará encontrar tendencias entre las características de los usuarios y las reseñas de los libros.

## Alcance

La base de datos abarca casi trescientos mil libros publicados en los últimos dos siglos y usuarios de todo el mundo y de cualquier edad.

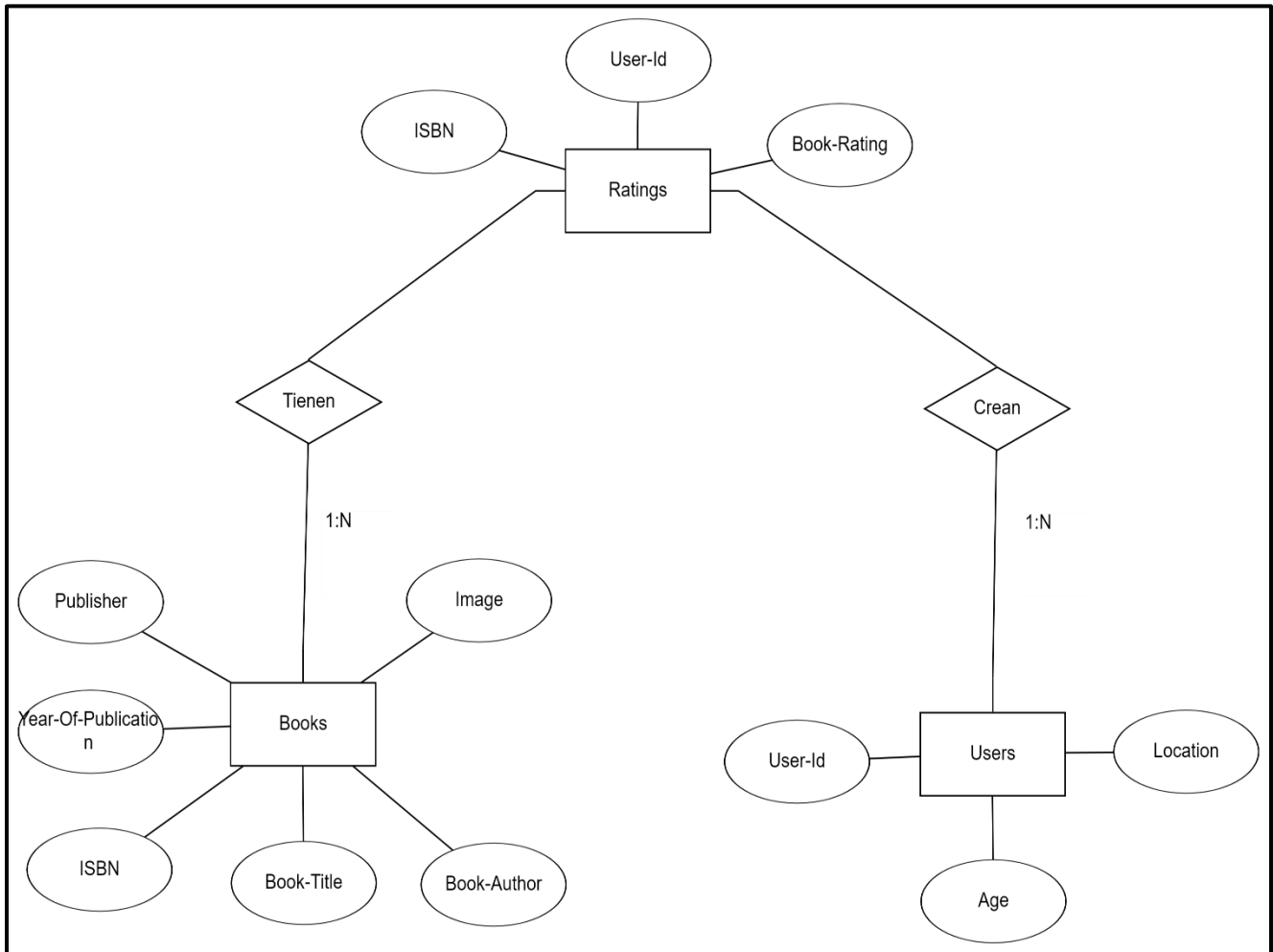
## Usuario final y nivel de aplicación

Los datos obtenidos podrán ser utilizados para, luego de identificar tendencias en las reseñas, dar recomendaciones más acertadas a los usuarios de Amazon

## Herramientas tecnológicas

Las herramientas utilizadas para la realización del proyecto fueron SQL para el modelado de la base de datos y Power BI para la visualización y análisis de los datos.

## Diagrama entidad-relación



## Listado de tablas

❖ Books

❖ Users

❖ Ratings

## Listado de columnas de tablas

Tabla 1. Books:

Esta tabla contiene para cada libro su título, su autor, su editorial, el año en el que fue publicado, su código de identificación (ISBN) y un hipervínculo a una imagen del mismo.

Tipo de clave	Campo	Tipo de campo	Detalle
PK	ISBN	VARCHAR (100)	Código de identificación único del libro
	Book-Title	VARCHAR (100)	Título del libro
	Book-Author	VARCHAR (100)	Autor del libro
	Year-Of-Publication	INT	Año en el que se publicó el libro
	Publisher	VARCHAR (100)	Editorial que publicó el libro
	Image-URL-L	VARCHAR (100)	Imagen de la portada del libro

Tabla 2. Users:

Esta tabla contiene para cada usuario su código de identificación, su localización y su edad.

Tipo de clave	Campo	Tipo de campo	Detalle
PK	User-ID	VARCHAR (100)	Código de identificación único del usuario
	Location	VARCHAR (100)	Localización geográfica del usuario
	Age	VARCHAR (100)	Edad del usuario

Tabla 3. Ratings:

Esta tabla contiene puntuaciones definidas por los usuarios para los libros.

Tipo de clave	Campo	Tipo de campo	Detalle
PK	Rating-ID	VARCHAR (100)	Código de identificación único del rating
FK	ISBN	VARCHAR (100)	Código de identificación único del libro
FK	User-ID	VARCHAR (100)	Código de identificación único del usuario
	Book-Rating	VARCHAR (100)	Puntuación que el usuario dió al libro



# Transformaciones de datos

## Tabla books:

- Se eliminó la columna "Image-URL-S".
- Se eliminó la columna "Image-URL-M".
- Se eliminaron las filas con entradas duplicadas en la columna "ISBN".
- Se eliminaron las filas con entradas con errores en la columna "Year-Of-Publication".
- Se cambió a minúsculas y se capitalizó todas las entradas de la columna "Book-Author".
- Se removieron las filas con entradas vacías en la columna "Book-Title".

## Tabla users:

- Se eliminaron las filas con entradas con errores en la columna "User-ID".
- Se cambió a mayúsculas todas las entradas de la columna "Location".
- Utilizando la herramienta "Split Column" se separó la columna "Location" en tres:
  - o La columna "City", que contiene la ciudad de los usuarios.
  - o La columna "Province/State" que contiene el distrito de los usuarios.
  - o La columna "Country" que contiene el país de los usuarios.
- Se reemplazó todas las entradas con valor NULL en la columna "Age" por entradas en blanco.
- Se removieron las filas con entradas en blanco en la columna "Age".
- Se removieron las filas con entradas en blanco en la columna "Country".

## Tabla ratings:

- Se removieron las filas con entradas con errores en la columna "User-ID".

- Se removieron las filas con entradas con errores en la columna "ISBN".
- Se removieron las filas con entradas en blanco en la columna "ISBN"

#### Tabla calendar:

- Se creó la tabla calendar a partir de la columna "Year-Of-Publication" de la tabla "books".

# Medidas

## Cantidad Autores:

Cantidad\_Autores = `DISTINCTCOUNT`(books[Book-Author])

Esta medida retorna la cantidad de entradas diferentes que existen en la columna "Book-Author" de la tabla "books". Se utiliza como dato.

## Cantidad Books:

Cantidad\_Books = `COUNT`(books[ISBN])

Esta medida retorna la cantidad de entradas que existen en la columna "ISBN" de la tabla "books". Se utiliza como dato.

## Cantidad Editoriales:

Cantidad\_Editoriales = `DISTINCTCOUNT`(books[Publisher])

Esta medida retorna la cantidad de entradas diferentes que existen en la columna "Publisher" de la tabla "books". Se utiliza como dato.

## Cantidad Ratings:

Cantidad\_Ratings = `COUNT`(ratings[ISBN])

Esta medida retorna la cantidad de entradas que existen en la columna "ISBN" de la tabla "ratings". Se utiliza como dato y como filtro.

## Cantidad Users:

Cantidad\_Users = `COUNT`(users[User-ID])

Esta medida retorna la cantidad de entradas que existen en la columna "User-ID" de la tabla "users". Se utiliza como dato y como filtro.

## Promedio Edad:

Promedio\_Edad = `AVERAGE`(users[Age])

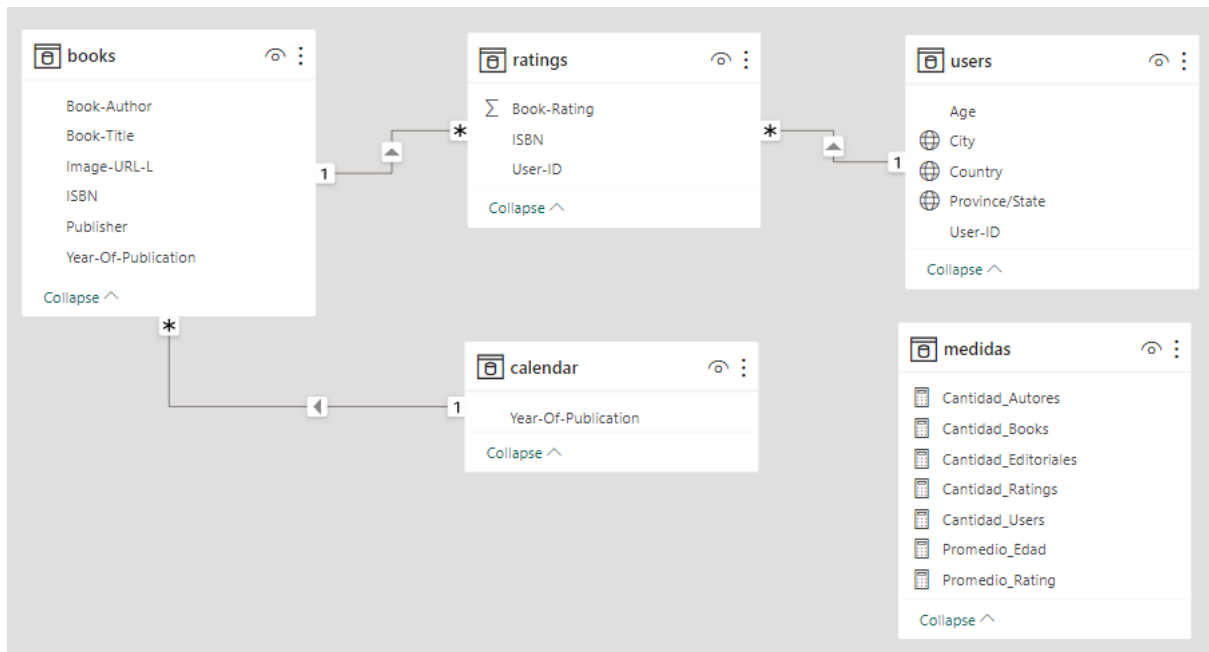
Esta medida retorna el valor promedio de la columna “Age” de la tabla “users”. Se utiliza como dato.

### Promedio Rating:

Promedio\_Rating = `AVERAGE`(ratings[Book-Rating])

Esta medida retorna el valor promedio de la columna “Book-Rating” de la tabla “ratings”. Se utiliza como dato y como filtro.

# Modelo de datos relacional



## Segmentaciones

- Desplegable para la elección de un año específico
- Desplegable para la elección de un país específico
- Desplegable para la elección de un autor específico
- Desplegable para la elección de un libro específico
- Desplegable para la elección de una editorial específica

## Visualización de datos

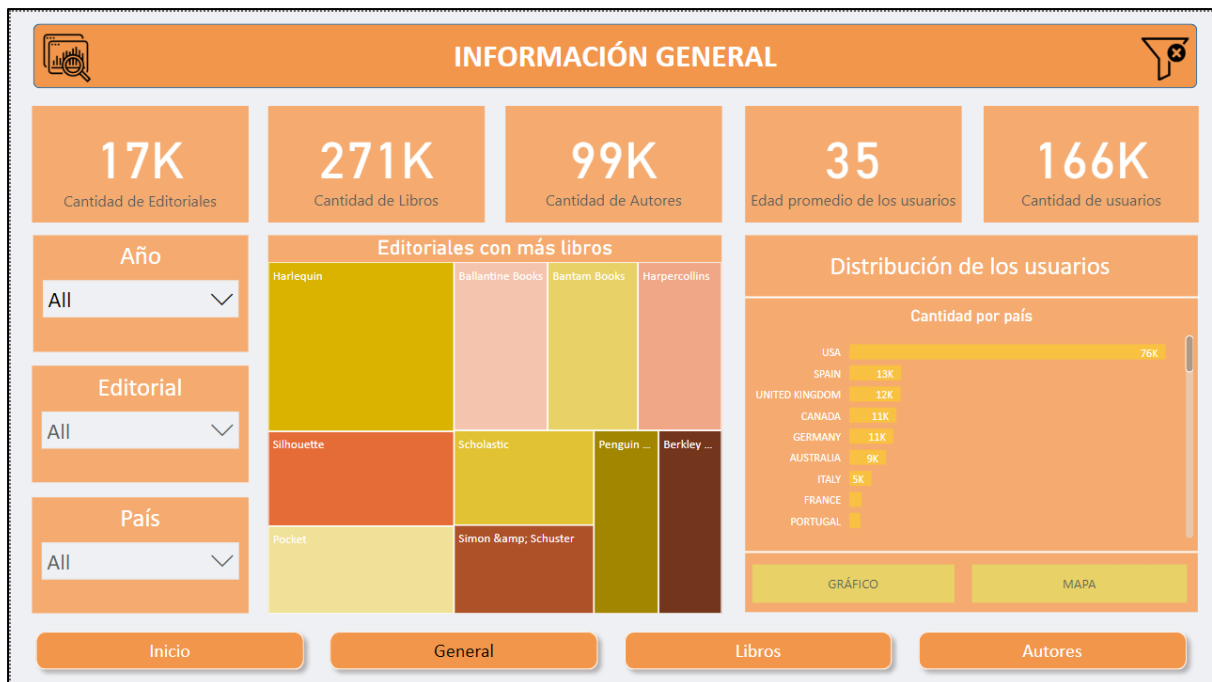
### Descripción de las solapas

#### Solapa 1: “Portada”



En esta solapa se introduce a la temática con un título y una imagen descriptivos de la misma, el logo de *Amazon*, puesto que el dataset provino de dicha empresa; la última actualización de los datos analizados y botones para acceder a las otras solapas.

## Solapa 2: "General"



En esta solapa se presenta información general sobre las distintas variables que se analizarán. Se busca particularmente demostrar el alcance del dataset y a quien representan los datos. Además, se da la posibilidad de filtrar por año, para saber cuantos libros, editoriales y autores en ese año; por país, para conocer la información sobre los usuarios en el país elegido; y por editorial, para saber cuantos libros y autores tiene esa editorial.

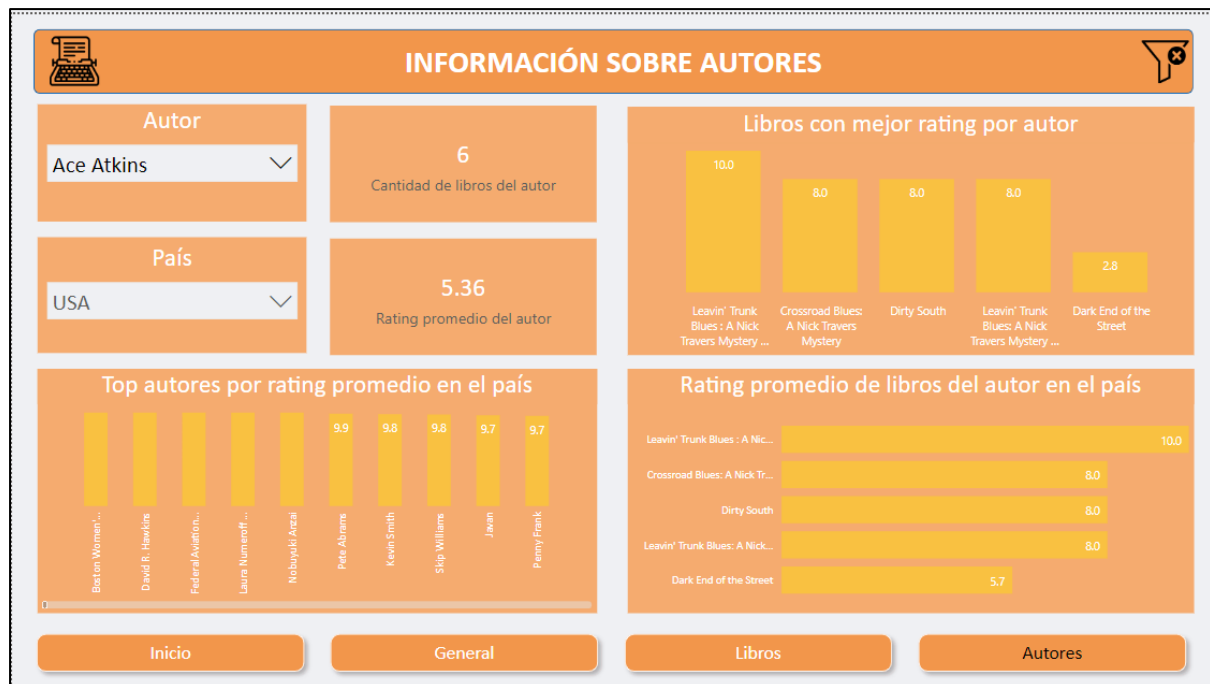


### Solapa 3: "Libros"



En esta solapa se busca principalmente analizar información sobre un libro específico, utilizando un filtro para seleccionar el mismo. Además, se busca, de manera general o filtrando por año, obtener información sobre las editoriales y libros con las valoraciones más altas.

## Solapa 4: "Autores"



En esta solapa se busca principalmente proveer información sobre un autor específico, utilizando un filtro para elegir el mismo. Además, se da la posibilidad de observar, de manera general o filtrando por país, los autores con valoraciones más altas. Por último se puede observar la performance de los libros de un autor específico en el país elegido.

## Futuras líneas

A futuro se podría poner como objetivo para el análisis de la temática la expansión del dataset, incluyendo nuevas columnas como género de los libros, país y año de nacimiento de los autores, y otras que incluyeran información cuantitativa de los libros y autores, como, por ejemplo: descripciones o reseñas de los mismos. Esta ampliación de los datos permitiría enriquecer y profundizar el análisis realizado.

Agregar a cada libro su género correspondiente permitiría observar los libros con mayor valoración dentro de los géneros, cuales son los géneros con mejor valoración y cuales con peor, que géneros se prefieren por edad y cuales por país, entre otros análisis que se podrían realizar.

Sumar información cuantitativa tanto a los autores como a los libros permitiría contextualizar a los mismos dando a los usuarios la posibilidad de entender, por ejemplo, porque un autor escribe cierto género o porque ciertos libros son más populares que otros.