**Summary of the Books Project**

* The purpose of my project is to consolidate all the book’s data into records which I can store in the database to serve the customers effectively as all the data will be in one place.
* First, I took the books.csv file downloaded from the internet and applied transformations. I cleaned the data as necessary and saved it in a CSV file.

Link to books.csv: [*7K books | Kaggle*](https://www.kaggle.com/code/aiswaryarana/7k-books/input?select=books.csv)

* Second, I scraped the website for book data and applied transformations to clean the data and format it as necessary.

Link to website: [*Catalog of books - wpDataTables - Tables and Charts WordPress Plugin*](https://wpdatatables.com/documentation/table-examples/catalog-of-books/)

* Third, I called APIs to get the Books data based on the ISBN value, applied some transformations to it, and saved it to a CSV file.

Link to the website to call the API: [*https://openlibrary.org/dev/docs/api/books*](https://openlibrary.org/dev/docs/api/books)

* Fourth, I read all the above three CSV files into individual data frames, connected to the SQLite database, and saved them into three tables.
* Fifth, I combined the three tables into a single table based on the common ISBN, created a data frame from it, and applied visualizations.

**Ethical Implications of the books data**

* The original book.csv, that I downloaded from the website had large data. So, I truncated the data to get only 6810 rows to be easy for analysis. Also, I deleted many rows with NA. I would have missed many important books (for analysis). So, this book's info may be biased.
* The data on the website has only 98 rows which has only 98 records of book data whose ratings are more than 3. So, the website does not contain data for low-rating or average-rating books. So, the data seems to be biased since it shows only books with higher ratings. Many other books' data is not available on this website which may be relevant for analysis. Customers may have different tastes in reading books despite low ratings. So, here we are missing the opportunity to provide these books' data.
* This open Library does not have the endpoint to get all the book info at once. There is only data of one book returned based on the book ISBN given as the input parameter. So, needed to make an API call for each book. So, I ended up making around 2000 API calls. So, I have taken the ISBN info from the previous 2 milestones and made API calls. This way, I may not have made API calls and could not have pulled other important book info which may lead to bias.
* Also, I had to make around 2000 API calls which may not be a good option. Rather, I should have taken an API to return all book data simultaneously. But I chose this option because this way, I called the API with the ISBNs I already have from Excel data and website data. So, merging the three datasets got easier, and getting all the data for some ISBNs made sense. The disadvantage with this is that I would have left many other books data that customers may like.