**Introduction:**

As part of the term project for this semester, I will be performing data preparation on the list of best-selling books. I could find some key fields related to this list across different sources. I will bring all of them into one table and perform my analysis.

**Data Sources**

**Flat File:** **books.csv** downloaded from <https://www.kaggle.com/jealousleopard/goodreadsbooks>

The file contains details about 11,128 books with its book name, author, isbn, isbn13, publisher, publication date and other related information.

**API**: <https://developers.google.com/books>

Google books API link where I can find details about the cost of the book, the link to buy the book along with book name, author, isbn, isbn13, publisher, publication date.

**Website**: <https://en.wikipedia.org/wiki/List_of_best-selling_books>

This Wikipedia link has details about the best-selling books. Best-selling refers to the number of copies sold. It has details regarding the book name, author, approximate number of copies sold & genre.

**Relationship between sources:**

Key relationship between Flat file & API will be the International Standard Book Number (isbn) field. The Website does not contain the isbn information, so the relation here needs to be built around the book name & author. This might be a challenge since book name is string field and not a unique numerical field like isbn.