**Project Name**

Good Books Analysis

**Group Members**

Ryan Gonzalez, Benjamin Kramskoi, Saroja Shreenivasan, Celeste Muniz-Lithgow

**Project Description**

By analyzing the books data, we will be creating readily available data of the good books based on review, price, etc.

**Why your final database will be useful to a hypothetical organization:**

An interactive database that a client can use to sort through a collection of books by rating, author, popularity(to\_read list), genre, price, number of pages, language and best selling books by genre.

**Data Sources:**

1. <https://www.kaggle.com/jealousleopard/goodreadsbooks>
2. <https://en.m.wikipedia.org/wiki/Lists_of_The_New_York_Times_Fiction_Best_Sellers>
3. <https://en.m.wikipedia.org/wiki/Oprah%27s_Book_Club>

**A brief summary of the three ETL steps you will take to create this database:**

* Extract - using python from kaggle and web scraping from  barnes and noble
* Transform - using pandas, python, sql and transform the data that is extracted
* Load - make a connection between pandas and postgres; load the data into the respective tables

**A description of what each team member will be responsible for:**

1. Pre-work: (assess the sources and find the best data which will suit the DB) - Ryan, Benjamin Kramskoi, Celeste, Saroja
2. Extraction - Ryan
3. Transforming - Celeste, Saroja
4. ERD Diagram - Benjamin Kramskoi
5. Load - Saroja
6. Github Repository - Benjamin Kramskoi

**Information to Extract:**

Kaggle - GoodReads

ISBN

Title

Author

Average Rating

ISBN

Language

Num Pages

Table (wikipedia all 3? - Bestsellers)

Title

Was best seller?

Year was best seller

**Tables:**

Book\_desc

ISBN - VARCHAR PK

Title (VARchar 50)

Author (Varchar 50)

No\_of\_pages (INT)

Publication\_date (DATE)

Language (VARCHAR 50)

Avg\_rating (float)

New York Times

ISBN - (INT) FK

YEar (INT)

Oprah’s

ISBN - (INT) FK

Bookclub\_Month (VARCHAR)

Boolclub\_year (INT)