**ETL Project**

Kenny Dao

Sean Pei

[Mickey Sandhu]

Our task was to source, clean up and load the data to a production database. We used the Pandas library for our data manipulation. We used Jupyter notebook to pull API from selected data source.

Also, we created DataFrames using CSV files about books data from Kaggle and Goodreads API.

Data source:

1. Kaggle: <https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m?select=book900k-1000k.csv> (dataset is being updated every 2 days)
2. API: <https://www.goodreads.com/api>
3. <https://www.kaggle.com/pelinsoylu/amazon-the-most-read-books-of-the-2019-dataset> (bestsellers with categories.csv)

EXTRACT

We used the books data (book900k-1000k.csv) from Kaggle. It’s a csv file with over 40k rows and 20 columns that content lot of information and a good starting point for us.

We also used the bestsellers with categories data to create the API request (from Goodreads API) about information about authors.

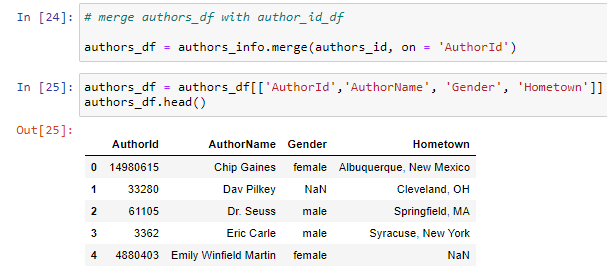
TRANSFORM

Having found the books dataset on Kaggle, we have chosen to make more information relating to books available for inquiries, such as finding the most popular book titles (bestsellers title by year), the highest-rated books, recently released titles, etc. For books table, 11 columns were removed and cleaned up/convert to suitable forms. The focused columns are BookId, Title, Rating, ISBN,

PublishY, PublishM, PublishD and Language.

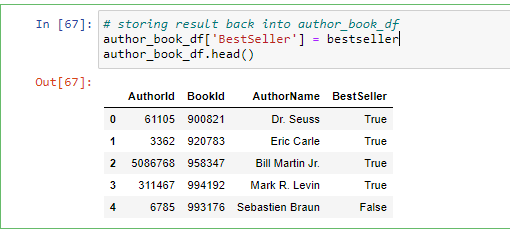
We have pulled API about author information from Goodreads API. The data is in XML file so we converted into dataframe using Jupyter notebook and created an authors table that includes AuthorId, AuthorName, Gender, and Hometown. At present, the authors table has 244 records.





We then processed and merged the data from books data, bestseller data, authors table (created above) to created an author\_book dataframe to make correlation between author and book adding BestSeller feature (to determine whether a book title is one of the bestseller books.





LOAD

We loaded our CSV into Postgres as our database and make it available on-line using google cloud platform. After that, we have tested out with some queries as below.

In summary, based on our observation and finding, there are multiple books dataset available. However, the data format is vastly different and still much of missing data (null value), which we can further update a long the way. Applying Postgresql to store database about books would be an alternative to present data in a better data structure for inquiries.

