Data Wrangling on Top Selling Books

DSC540 – Data Preparation
Vinay Nagaraj
Data Science, Bellevue University
Nov 20, 2020

Abstract:

Books are a very important aspect in the personal development of a human being as it exposes them to new ideas, perspectives and literary styles. Millions of books have been published over the years and they continue to be an integral aspect of people's lives around the globe.

As part of my project, I have gathered data from three different data sources about the top selling books in the world and performed data cleaning/wrangling steps.

Dataset Source 1: Kaggle

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

Dataset Source 2: Google books API

https://www.googleapis.com/books

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

Dataset Source 3: Wikipedia

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 all the three data sources is based on Book Name & Author.

Kaggle File Cleaning Steps

- The CSV file data was read into Pandas Dataframe which has 12 columns and 11,123 Rows.
- Normalized Author Names for further processing & removed unnecessary columns.
- Removed duplicates in the Dataframe based on the Book Name. It had 775 duplicate rows.
- Checked the % of NaN values in each column and noticed that our dataset was a clean dataset with no missing values.
- Column names were renamed for better understanding.
- Final Dataframe with 11 columns and 10,348 rows was written into a CSV file.

Wikipedia Cleaning Steps

- Information about Top Selling Books were read from Wikipedia using pandas.read_html.
- Validated for any duplicate entries based on Book Name. None found.
- Column names were renamed for better understanding.
- Column values were formatted for better understanding.
- Final DataFrame was loaded into CSV file.

Google API Cleaning Steps

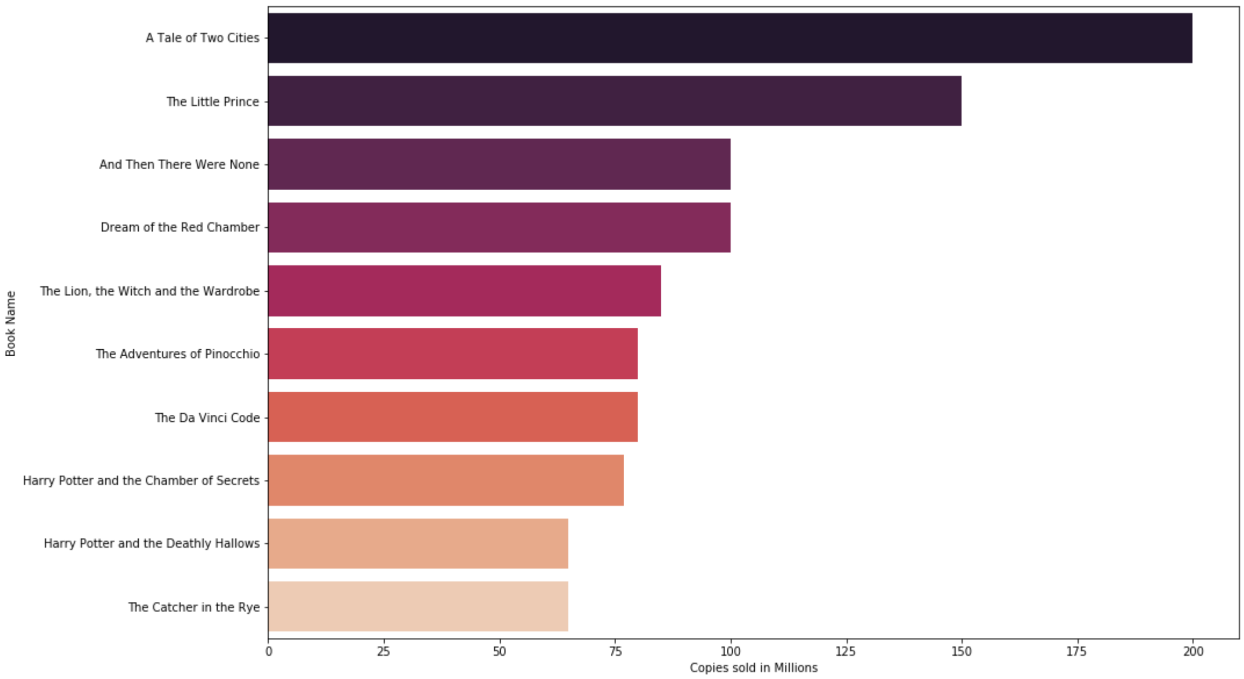
- A successful connection was made using urllib with API and with the list of top selling books from Wikipedia as the search key, the necessary information was retrieved.
- A blank Dictionary was used to store the JSON data retrieved from API. Later was converted to a readable pandas dataframe.
- Total 12 columns and 1,556 rows were retrieved.
- All rows with NaN in Book Name, Author, ISBN & ISBN13 were removed from dataframe.
- Final dataframe was loaded into CSV file.

Merging and Database Load

- The 3 CSV files from the different sources were read and was merged using merge() and ‘inner’ join.
- Fuzzy matching technique was used to merge the Wikipedia data and the Google API data. Book Name & Author was used as a key and a threshold of 95. This merged dataframe was then loaded into the SQLite database.
- Kaggle data was loaded to the database and then the above merged data was then merged with the Kaggle data using ‘Inner’ join.
- The data was then read from the database to create Visualizations.

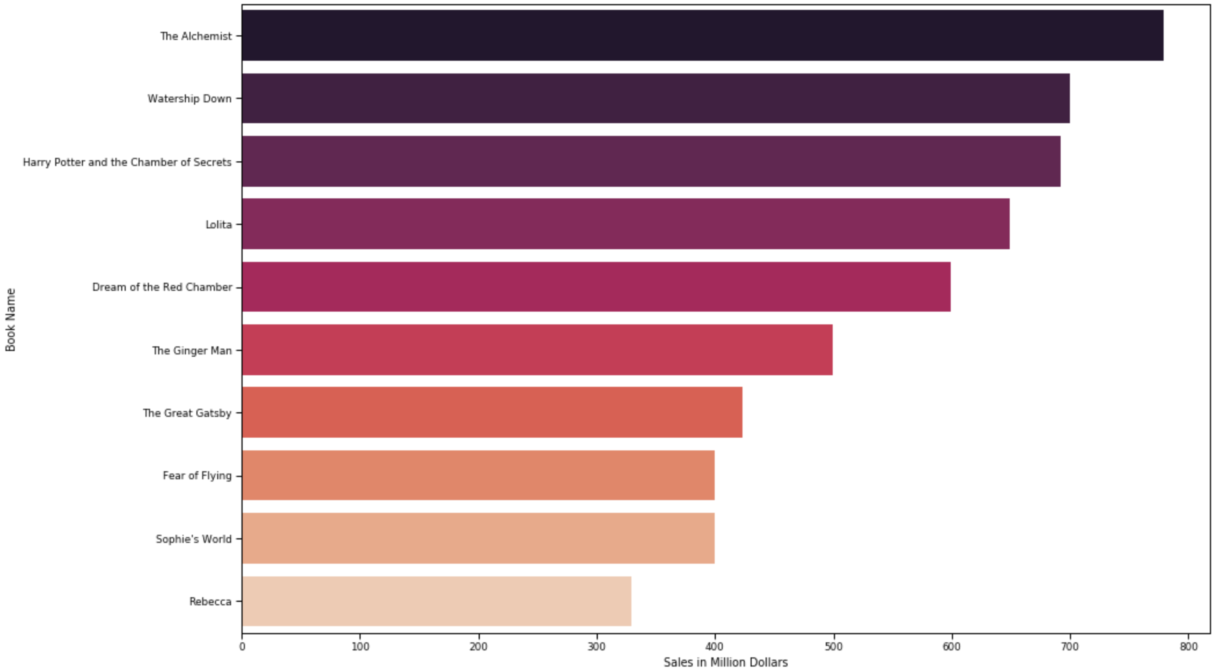
Visualization

Top 10 books sold



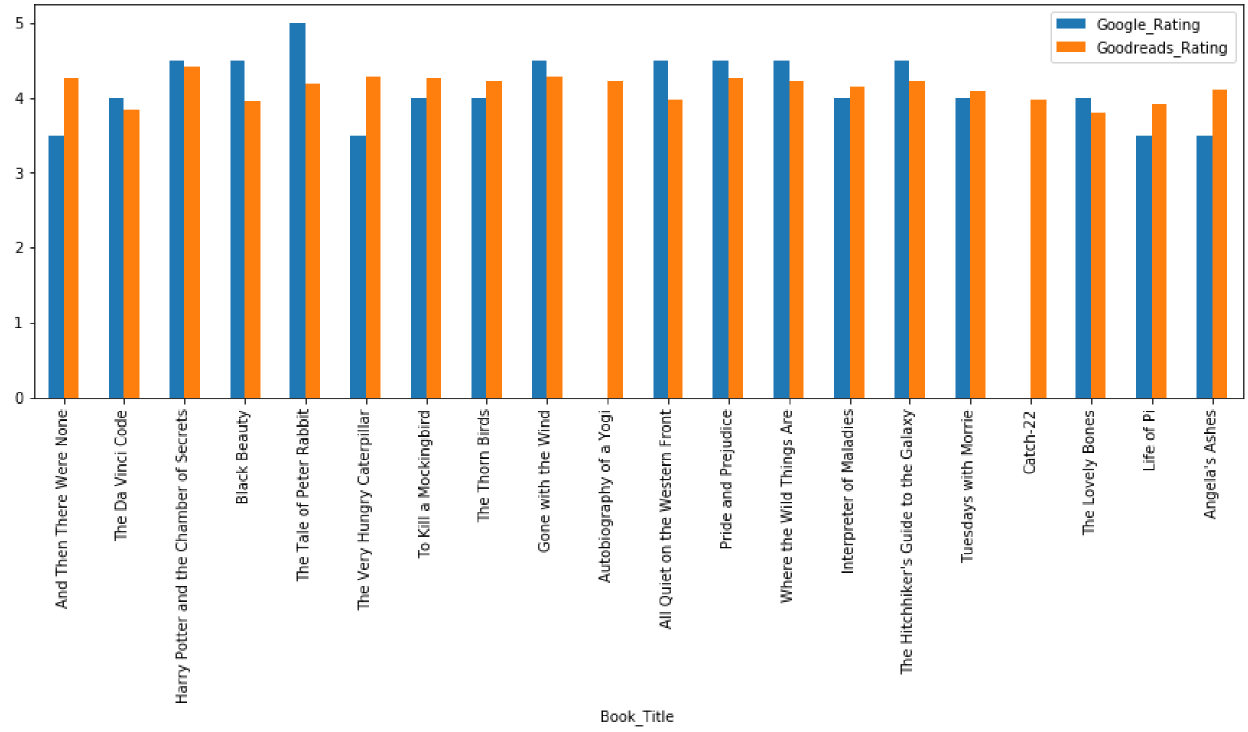
A Tale of Two Cities was the Top selling books which sold 200 Million copies.

Highest Grossing Books in USA

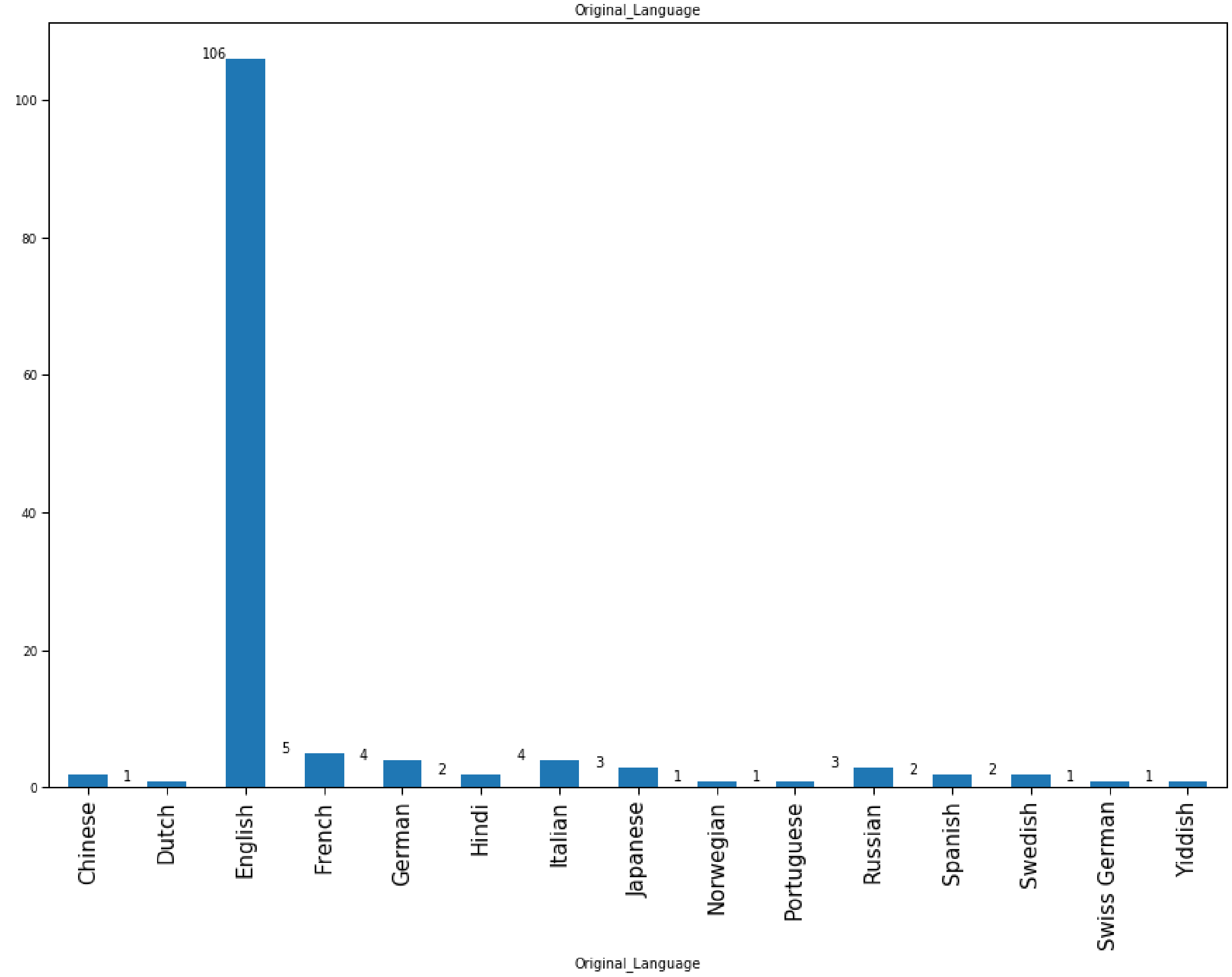


The Alchemist was the book with the top dollar value sales of 779.35 Million Dollars.

Compare Google Ratings & Goodreads Ratings

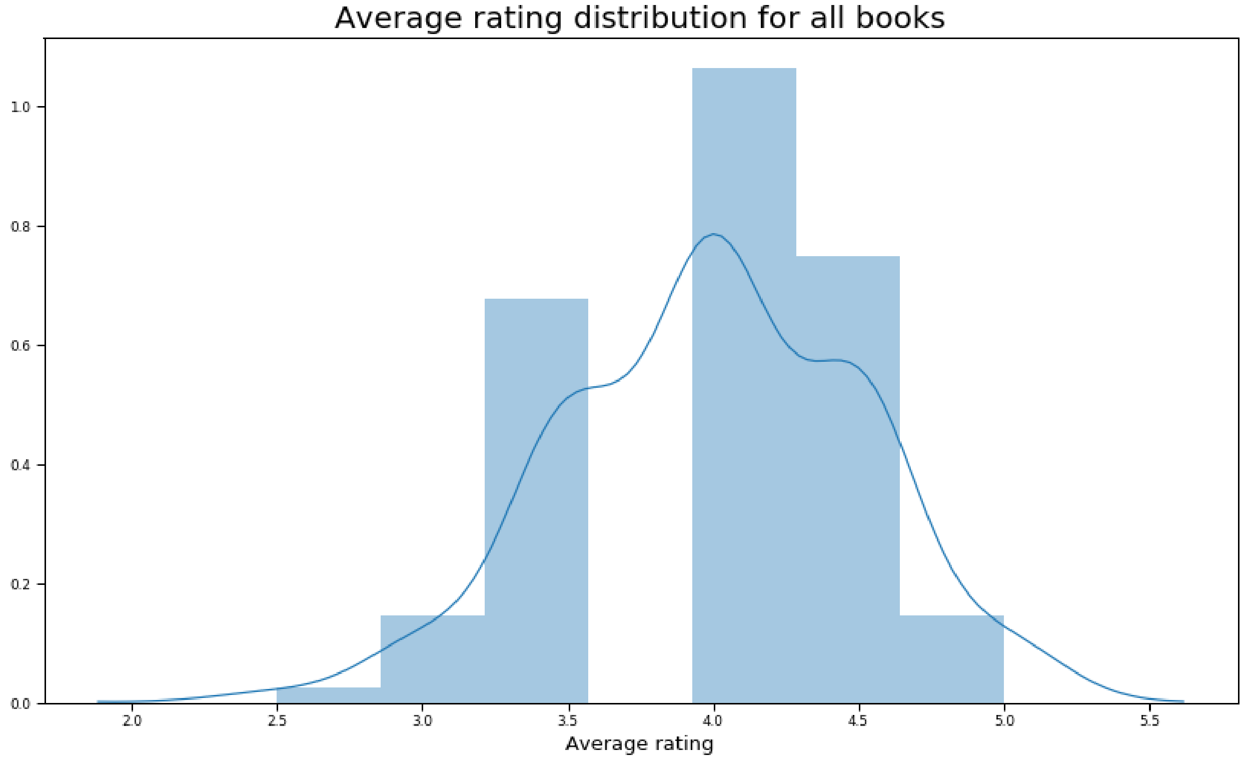


Distribution of books for all languages



Majority of the Top Selling booksare written in English.

Average Google rating distribution for all books



The overall rating of most of the top selling books are around 4.0.