Final Project Data Checkpoint

---------- Xinyi Zhao

Project code

https://github.com/sarahzhao21/final_project.git

Data sources

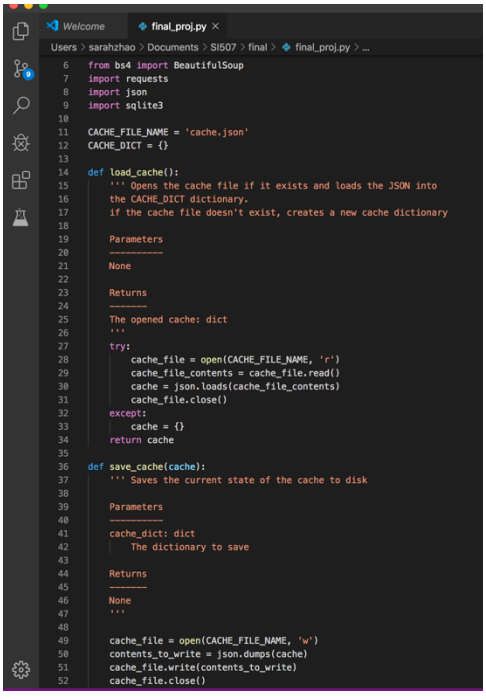
1. The New York Times Best Sellers: https://www.nytimes.com/books/best-sellers/

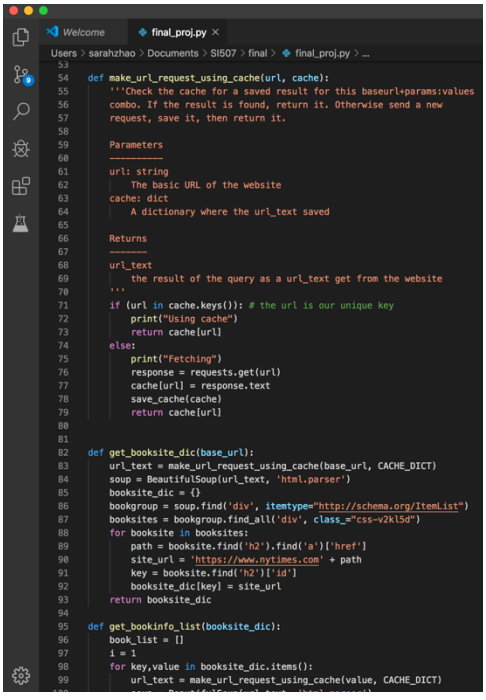
This website provides the lists of the best seller books in different categories in “New York Times”.
Format: HTML, JSON
How: I did web crawling and scraping to get the information of best seller books and get the url of another link in “Apple Books” for each book and do the web crawling to “Apple Books”.
Summary of data:

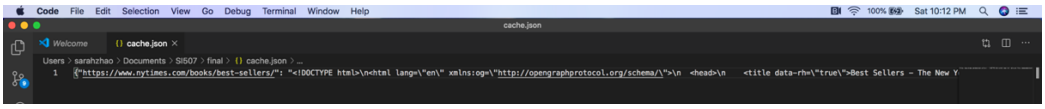
1) There are 140 records available (11 book categories and 10 to 15 best seller books for each categories)
2) 140 records have been retrieved.
3) description of records:

fields	Data type	Description
Id	INTEGER	The Id number of the book
Title	TEXT	The title of the book
Category	TEXT	What kind of book it is (fiction, novel ….)
Author	TEXT	The author who wrote the book
Publisher	TEXT	The publisher of the book
Rank	INTEGER	The rank of this book in it’s category last week
Weeks_on_the_list	INTEGER	How many weeks the book was on the best-seller list
Description	TEXT	A short review about this book
Apple_url	TEXT	The url that connects to the website of ‘Apple Books ‘

4) Evidence of caching







2. Apple Books

   Each book from the New York Time Best Sellers has an 'Apple Books' url, here is one of the example: https://books.apple.com/us/book/american-dirt-oprahs-book-club/id1459876569

   This website provides all kinds of detailed information about the books showed on New York Times.

Format: HTML, JSON

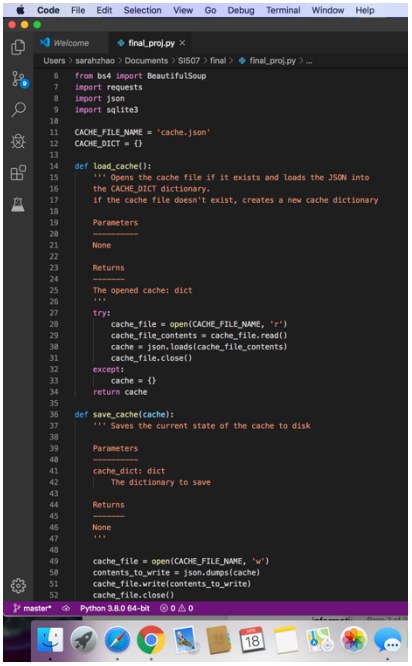
How: By accessing the beautiful soup of the Apple Books url for each book, I did the web scraping to obtain the information of 'title', 'genre', 'price', 'rating', 'released date', 'language', 'length', 'seller' and 'size'.

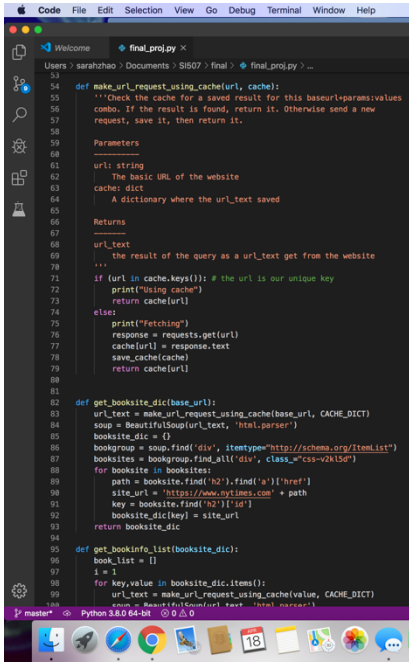
Summary of data:

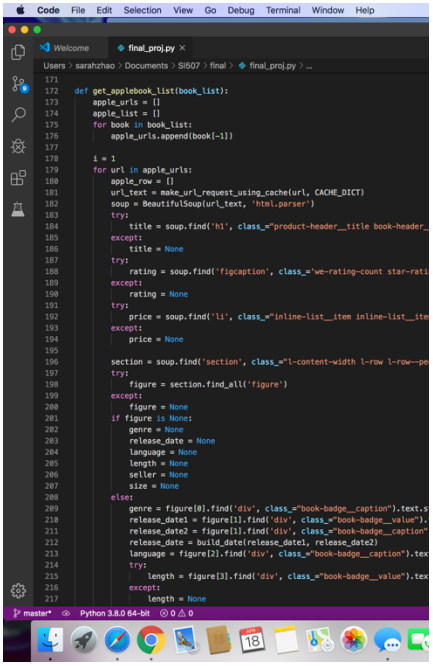
1) There are 140 records available (each record refers to one book on New York Time best seller)
2) 140 records have been retrieved.
3) description of records:

fields	Data type	Description
Id	INTEGER	The Id number of the book
Title	TEXT	The title of the book
Rating	REAL	The rating of the book on "Apple Books"
Price	REAL	The price of the book on "Apple Books"
Genre	TEXT	The genre of the book on "Apple Books"
Released_date	TEXT	The date that the book released on the website
Language	INTEGER	What kind of language the book was written by
Length	INTEGER	The page number of the book
Seller	TEXT	The seller of the book
Size	REAL	How many 'MB' of the book

4) Evidence of caching







Database

A file named ‘best_seller_book.sqlite’ has been created by the program ‘final_proj.py’, when it was opened by DB Browser, there are two tables in this file: ‘Best_seller’ and ‘Apple_book’:

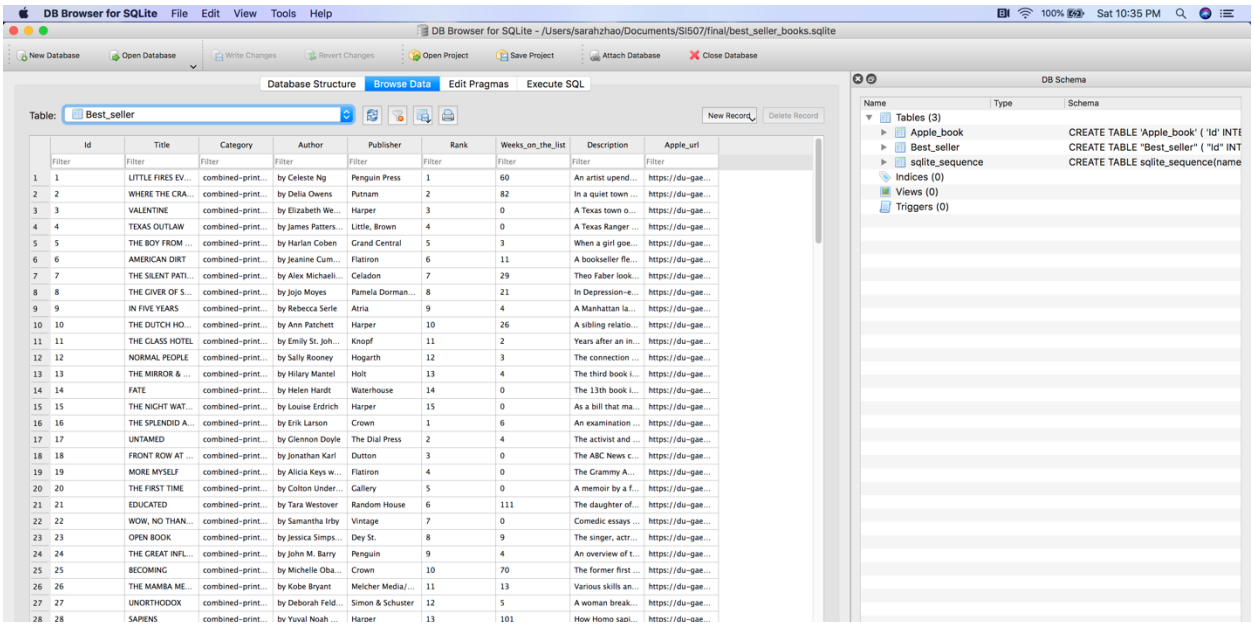
1. ‘Best_seller’ has 9 columns and 140 rows. Each row represents the information of a book, the information of the columns is listed here:

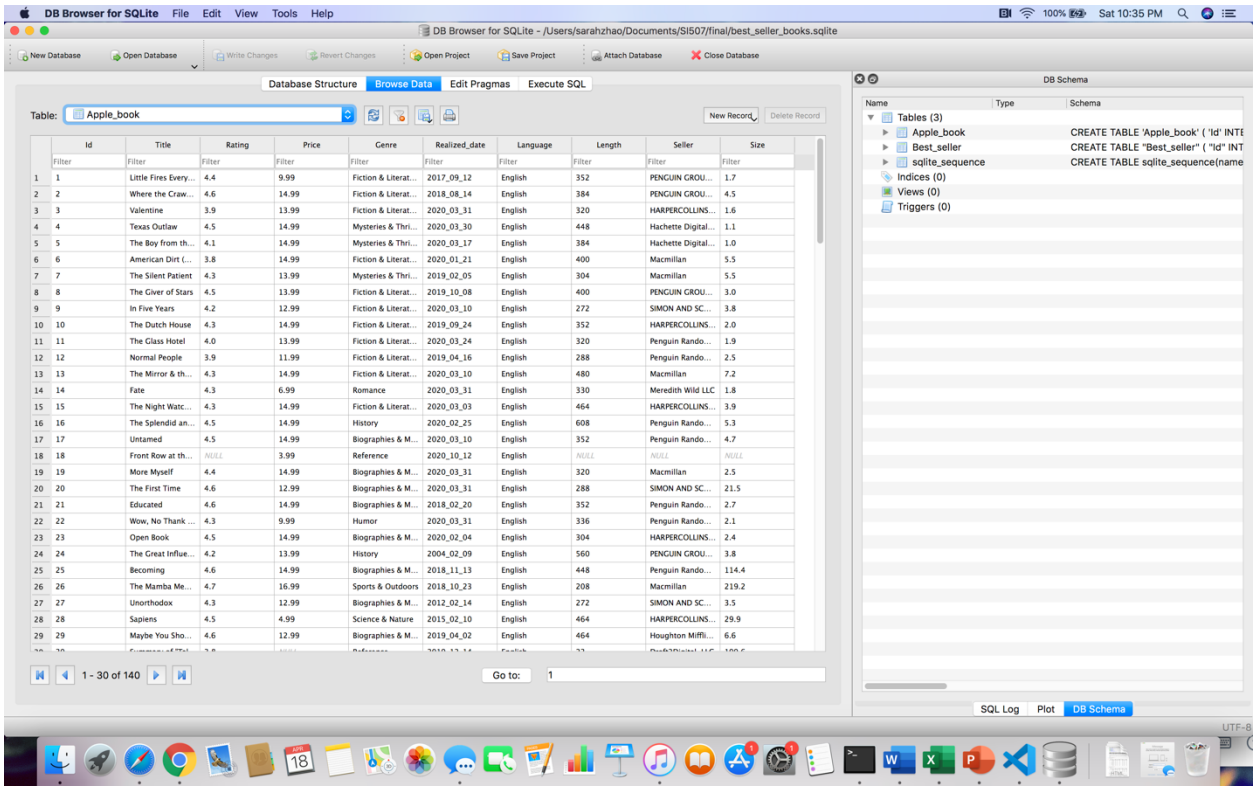
Column name	Data type
Id	INTEGER
Title	Text
Category	TEXT
Author	TEXT
Publisher	TEXT
Rank	INTEGER
Weeks_on_the_list	INTEGER
Description	TEXT
Apple_url	TEXT

2. ‘Apple_book’ has 10 columns and 140 rows. Each row represents the information of a book on “Apple Books” and refer to the same book on the table ‘Best_seller’ with the same order. The information of the columns is listed here:

Column name	Data type
Id	INTEGER
Title	TEXT
Rating	REAL
Price	REAL
Genre	TEXT
Released_date	TEXT
Language	INTEGER
Length	INTEGER
Seller	TEXT
Size	REAL

3. The primary key is the ‘Id’ on ‘Best_seller’ table and the foreign key is the ‘Id’ on ‘Apple_book’ table. The sequences of these two sets of ‘Id’s are the same.
4. Screen shot:





Interaction and Presentation Plans

Approach: Flask, Plotly
I plan to use Flask to design a submission form for the users to select what kinds of books they what to know. The form for users should look like this:

Welcome to the New York Time Best Seller Book Brower!

Select Category:

- Combined Print & E-Book Fiction
- Combined Print & E-Book Nonfiction
- Hardcover Fiction
- Hardcover Nonfiction
- Paperback Trade Fiction
- Paperback Trade Nonfiction
- Advice, How-To & Miscellaneous
- Children's Middle Grade Hardcover
- Children's Picture Books
- Children's Series
- Young Adult Hardcover

Sort by:

- Rank

- Rating on Apple Books
- Price on Apple Books
- Length on Apple Books

Sort direction:
- High to low
- Low to high

Plot results?

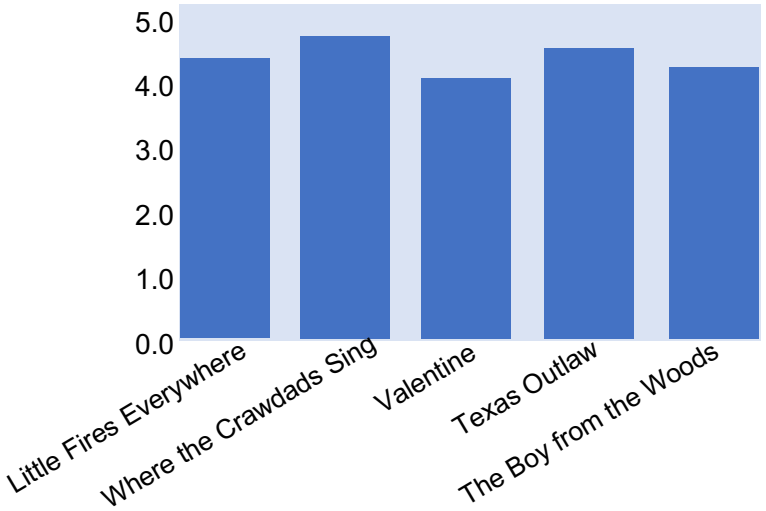
The presentation to users should look like this:

Here are your results:

Title	Genre	Author	Rank	Rating	Price	Length	Seller
Little Fires Everywhere	Fiction & Literature	Celeste Ng	1	4.4	9.99	352	PENGUIN GROUP
Where the Crawdads Sing	Fiction & Literature	Delia Owens	2	4.6	14.99	384	PENGUIN GROUP
Valentine	Fiction & Literature	Elizabeth Wetmore	3	3.9	13.99	320	HARPERCOLLINS
Texas Outlaw	Mysteries & Thrillers	James Patterson	4	4.5	14.99	448	Hachette Digital, Inc
The Boy from the Woods.	Mysteries & Thrillers	Harlan Coben	5	4.1	14.99	384	Hachette Digital, Inc

Or this:

Here is your graph:



The values that the graph will show depend on the ‘Sort by’ selected by the users. For example, if the user selects ‘Sort by’ ‘Rating on Apple Books’, the bar chart will show the ratings of the books labeled by the title of each book.