Capstone Project - 1

Goodreads Book Recommendation System

Submitted by : Anshul Vyas Mentor: Kevin Glynn

======================================================================

1. Proposal with a problem statement
2. Data collection and wrangling summary
3. Exploratory data analysis summary (visualization and inferential statistics)
4. Results and In-depth analysis using machine learning

====================================================================

1. Problem statement:

1. Data collection and wrangling summary:

Data has been gathered from a Kaggle competition page. Following is the link

For the same:

<https://www.kaggle.com/jealousleopard/goodreadsbooks/kernels>

A lot of Data cleaning and Wrangling techniques were applied to the data sets.

There was a total of 5 different CSV files available with various information.

Like books, id, title, author, cover links, ratings, tags, tag names etc.

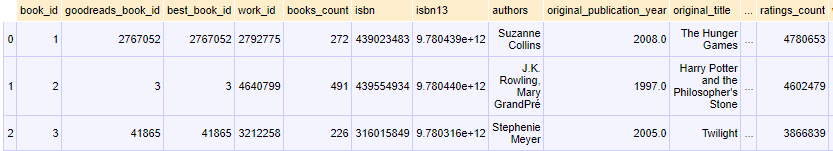
Many unwanted columns were removed like the link for the book cover.

For the id, there were 3 various types of columns available. So two of them were removed as there was no link between those columns and other datasets.

Various datasets were merged to gain the full picture of the available data.

Book\_tag and Tag datasets were merged so that we can all information related to tags at on place.

* Here are the few data frames:



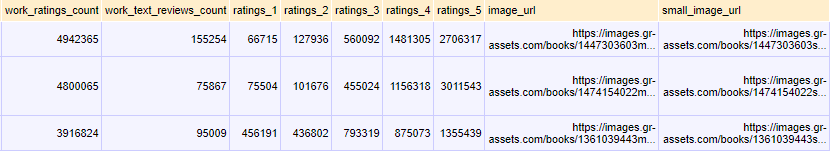


Fig 1: Raw dataset for books

* Book\_tag information looks like this:

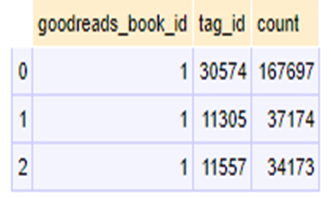


Fig 2: Book\_tags

* Tag information dataset:

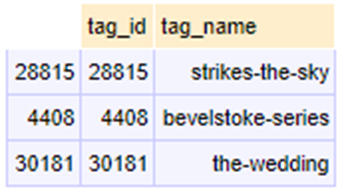


Fig 3 : Tags

* Rating information:

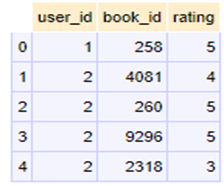


Fig 4: Ratings

1. Exploratory data analysis summary (visualization and inferential statistics):

From the book data set, the list of top 10 books was inferred on the basis of the total number of ratings given to a book. Though, this data is not going to help us to find any concrete information.

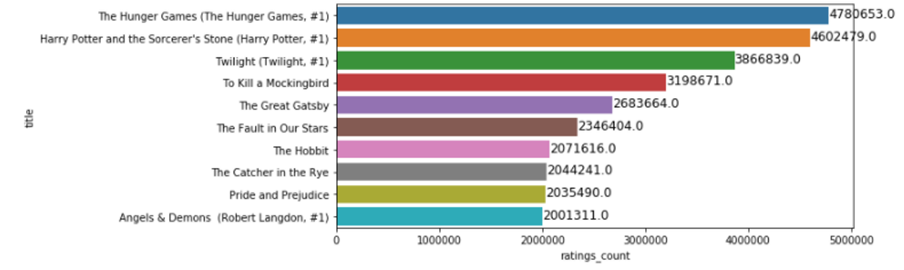


Fig 5: Top 10 rated books on the basis of frequency of votes

There were approximately 98% of books written in the English language. Following is the visual pie chart for the same:

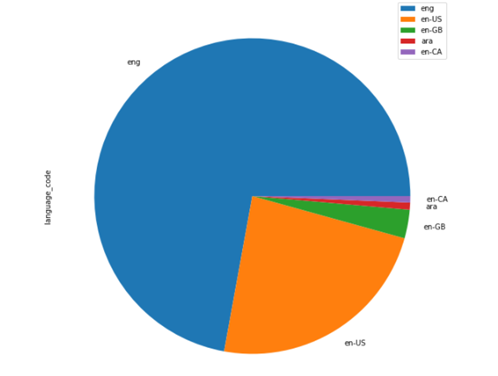


Fig 6: Language distribution of all books

Insights from Dataframes :

1) 'book\_id' : This column is present in df\_rating also. We need to merge these both data frames to get the result

2) 'goodreads\_book\_id': These are unique book ids which can be mapped with df\_book\_tag

3) 'best\_book\_id','work\_id','books\_count','isbn','isbn13': All these columns are of no use in further analysis and can be dropped.

4) 'authors','title' this column will be useful for further analysis.

5) 'Original\_title','small\_image\_url' these columns can be dropped.

6) There are total of six columns related to rating on the scale of 1-5 and then average rating. But the data is not mapped with ratings given in ratings.csv file. so discarding these columns.

Relation Between rating and frequency:

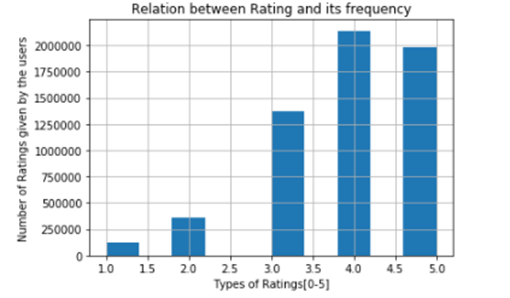


Fig 7: Rating /Frequency plot

Relationship between Books and the number of users:

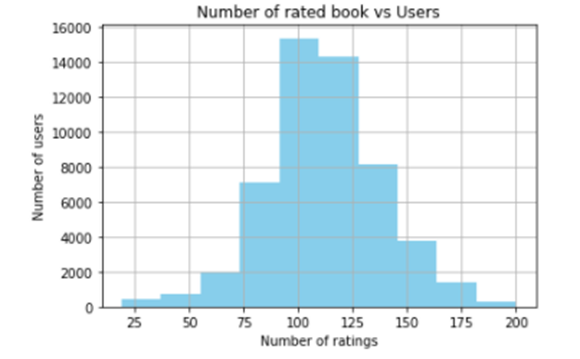


Fig 8: Book/User plot

1. Results