Data Narrative – Assignment 1

Parag Sarvoday Sahu
Electrical Engineering
(Roll No. - 22110179)
IIT Gandhinagar
Gandhinagar, Gujarat
parag.sahu@iitgn.ac.in

***Abstract*—To mine the given Dataset containing book descriptions from the website goodreads.com and report the observations.**

Keywords—python, pandas, data-analysis, goodreads

I. Overview of the Dataset

The given Dataset has description of 10,000 books contained in 5 CSV files. A short description of each CSV file is as follows:

1. books.csv – It contains information like various books and their authors, year of first publication, average rating and number of ratings gained by a book.
2. ratings.csv – It contains information about number of stars given as rating to various books by the users.
3. to_read.csv – It contains information about various books contained in a user's to-read list.
4. book_tags.csv – It contains information about number of books contained in various tags.
5. tags.csv – It contains information about the names of various tags along with their tag IDs.

II. Procedure employed for analysis

The dataset was first cleaned by removing rows which contained empty cells and repetitive rows or rows containing same information but with different names were merged. Thereafter, the dataset was analyzed by asking questions from the user's point of view to retrieve meaning information.

III. Scientific Questions/Hypotheses

A. What is the distribution of the books present in the datadet amongst the various languages?

B. Who are the most popular authors amongst the masses?

C. What are the most read books? Are only recent books popular or there are all time favourites too?

D. Hypotheses : Avid readers are generous while rating books and they do not give very low rating to any book.

IV. Details of Libraries and Functions

The various libraries used in this analysis are as follows:

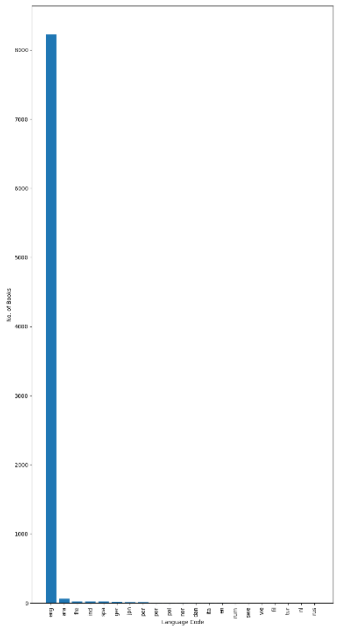
1. Pandas – This library is used to process and analyse large datasets.
2. Matplotlib – This library is used to get various types of plots using python language.
3. Numpy – This library is used to perform quick and efficient calculations using arrays, in python.
4. Seaborn – This library is used to get more detailed and modern looking plots than Matplotlib using python language.

V. Answers to the Questions

A. Language distribution of books

Books in English language heavily outnumber books written in other languages. The languages that follow English are Arabic, French, Indonesian, Spanish, German respectively. This clearly shows that goodreads.com being an American origin website has majority of its readers from English speaking regions.

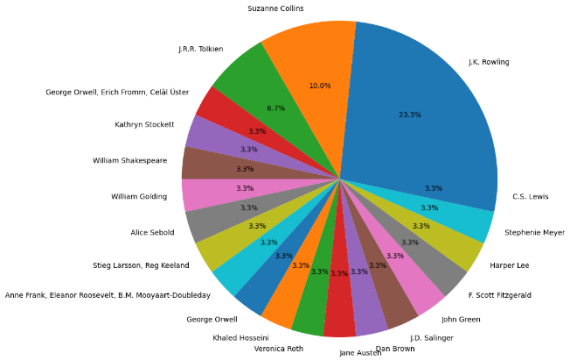
Assumption – All the various types of English such as American, British English are considered under the same language code 'eng' to simplify the analysis.



Bar graph showing the distribution of various languages

B. *Most popular authors amongst the masses*

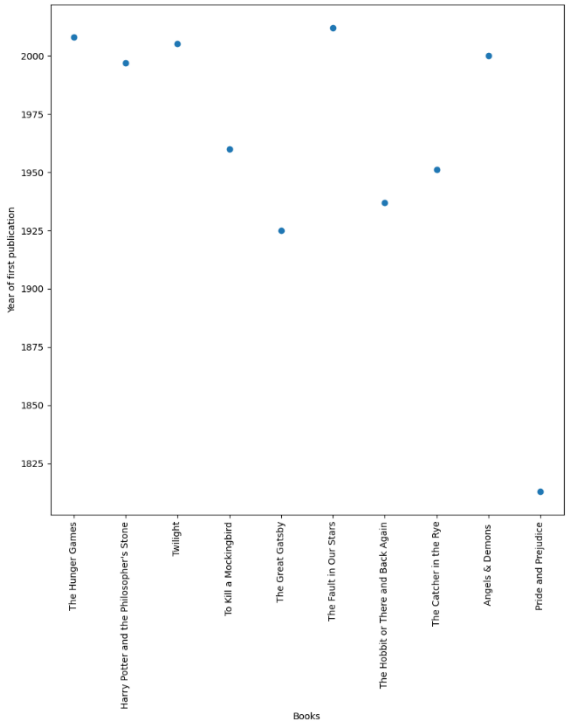
The popularity of books is gauged from the number of ratings received by each of them. Thirty books in the dataset have more than 15 lakh ratings. Analysis shows that J.K. Rowling is the most popular author followed by Suzanne Collins and J.R.R Tolkien respectively.



Pie-chart depicting the author-wise distribution of books which received more than 15 lakh ratings.

C. *Most read books*

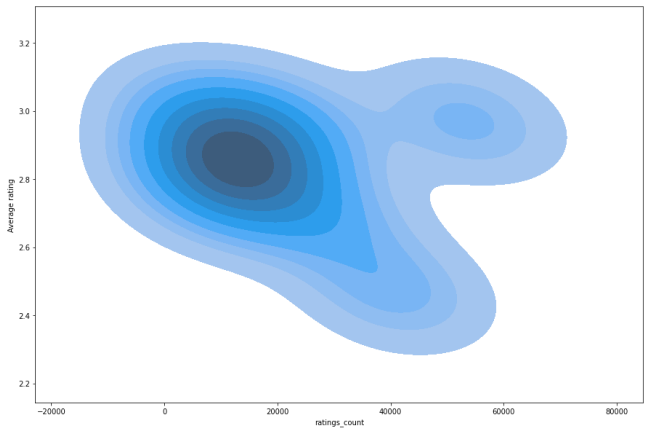
For this analysis, books having more than 20 lakh ratings have been considered. There are 10 books which confirm to this criterion. Most books in this list are from the 20th century or later except for 'Pride and Prejudice' which was published in 1813. This means that the general population prefers reading the recently released books but some all-time favourites also exist.



A scatter plot depicting the most popular books and their year of first publication.

D. *Readers do not give very low rating to any book*

The books with low average ratings i.e., below 3 have ratings around are around 20,000. The following plot shows that people are generous and they do not give very low ratings like 1 or 2 to even the books they do not like.



Kernel Density Estimation (KDE) plot of books having average ratings less than

VI. Summary of the Observations

The books that are already popular tend to get more popular over time. Hence, there are some books like 'The Hunger Games' and 'Harry Potter and the Philosopher's Stone' which have exceptionally high number of ratings of the order of 45 lakhs. There exist some all-time favourites which are being read by people despite being old and hence, they will continue to be read by the future generations.

Acknowledgment

I would like thank Prof. Shanmuganathan Raman for providing us with this opportunity of experiencing Data Analysis in our 1st year of BTech program. I also thank the creators of all the free resources available on the internet.

References

- Goodreads dataset : https://github.com/zygmuntz/goodbooks-10k
- Pandas user guide : https://pandas.pydata.org/docs/user_guide/index.html
- Matplotlib user guide : https://matplotlib.org/stable/users/index.html
- Seaborn user guide : https://seaborn.pydata.org/tutorial.html
- Data Analysis with Python Course : https://jovian.com/learn/data-analysis-with-python-zero-to-pandas