Analysis on Books data from different platforms to increase Library Usage

1st Abhishek Sunil Padalkar
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19221576@student.ncirl.ie

2nd Silky Jain
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19213590@student.ncirl.ie

3rd Sai Srimaha Vishnu Valluri
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19208758@student.ncirl.ie

4th Namita Mohan
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x19212500@student.ncirl.ie

***Abstract*—Library usage has decreased drastically in current time. Due to the presence of online resources, users prefer to learn online instead of reading books from libraries. Analysis performed on the San Francisco library usage data, in this work, shows a clear trend of decreased library usage, majorly among the youths. Though libraries are good information resource with collection of books, journals, and research papers, it is slowly getting replaced by other sources. An extensive analysis is performed on the bestselling and trending books based on rankings, pricing, reviews, rating, genres from multiple source such as Kindle, Goodreads, New York Times such that if these books are made available in libraries, it might increase library usage and registrations significantly.**

***Index Terms*—Library Usage, Kindle, Goodreads, New York Times, bestselling, trending**

I. Introduction

The main objective is to find the attributes affecting the library usage and to increase it further by analyzing the books data available at multiple platforms such as Kindle, Goodreads, and New York Times. The declining library usability is also caused because of the increase in online platforms providing e-books services, so to overcome this problem analysis is done on the trends of bestselling books and publications based on ratings and reviews provided by the people on online platforms. Using these analysis, the popular books can be added into the library sections to increase library users.

Library Usage data is extracted from San Francisco's Library Integrated system which includes the library patron records with circulation data having records from year 2003 to 2016 which contains data like age range, total renewals, total checkouts, Home library, circulation active month and year, year patron registered to the library. This data can be utilized to visualize the age range with respect to total number of checkouts and renewals per year showing the usage of libraries is gradually decreasing over the years [1].

Today, New York Times Bestselling ranking is considered as the top correct ranking for all the books in the market. The process for their ranking for hardcovers and e-books is based on confidential information on units sold given to the New York Times Department by all the vendors [2]. New York Times has a developers page which made the bestselling books data publicly accessible using API. This data is used to know top grossing bestselling books, authors and book categories for overall project goal.

Goodreads is an excellent social cataloguing website to browse through books, quotes and reviews. Users can add books they have read or want to read to their bookshelf. Also, it allows you to check out other users' bookshelves and what they are reading. It's the feature of reviews that makes Goodreads the best place to discover new books and share books that one loves. Goodreads provides APIs and a python wrapper for developers to access their books database. This report presents the analyses like what are the most rated books, who are the most popular authors and the top publishers on Goodreads. Also, how the book's ratings are distributed.

Amazon Kindle is one of the most prominent platforms used to access e-book. Roughly 80 million kindle products have been sold worldwide. This makes it a very important contributor towards the sale of books. Primarily dealing with ebooks leads to the availability of having a large amount of data regarding customer response to the available content. Availability of this information helps us understand the trends in what sort of content is preferred by people more and also helps in understanding what books perform well in a sales perspective.

II. Research Questions

For the purpose of the main research aim, that is to increase library users, this project is divided into 4 sub-research questions. They are as follows:

1) What are the attributes affecting the library usage the most like Renewals, checkouts, age range, type of patrons?

2) NewYork Times Bestselling books analysis to find top grossing bestselling books, authors and book categories from 2018-2020.
3) Analysis to find top ranking books, authors and publishers using ratings and reviews of books on Goodreads.
4) Analysing trends in pricing and customer response to books and authors on the Kindle platform.

III. Related Work

Cynthia H. Kumah conducted a study for the comparison of Internet and Library use, assuming that the students prefer the internet moreover library. After the analysis, the results suggested that both library and internet are used but still there is less usage of library observed during these years because of recent advancements in technology [3].

Chamani Gunasekera performed a study by survey undertaking the satisfaction of library services, in which it was found that library services are not well utilized. This led to the recommendation of a comprehensive information literacy program for undergraduates and also the library has to evolve as an electronic source of information for general and reference material to fulfil the user's requirements [4].

Antonios Makris, et al. compared No-SQL MongoDB with the relational database PostgreSQL by applying the spatio-temporal queries on both the database. Performance of PostgreSQL was better in all the cased from MongoDB in terms of average response time which is quite less in PostgreSQL. This doesn't mean that MongoDB is any less than PostgreSQL, MongoDB has vast usage but in terms of efficiency PostgreSQL outperforms MongoDB [5]. Thus, both these efficient databases are used for the purpose of this project.

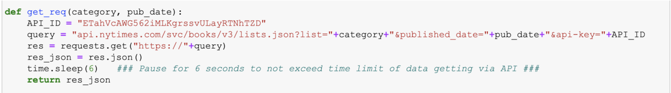
Wang et al., they used Goodreads data to study the use of online review for book impact assessment. The study involves understanding the reviewer's sentiment, identity and motivation using review, rating and role on Goodreads. They concluded that the Goodreads reviews are suitable for content-based recommendation and analysis [6].

T Bao and TS Chang performed a study to see the influence of traditional media and social media on market sales to understand the effect of both combined. The traditional media is reporting by journalists on New York Times BestSeller where as social media is based on reviews and comments of public. It is found that they influence each other, and increase sales together [7]. This confirms, considering New York Times BestSelling Data, being a traditional media, is a better to perform analysis along with Goodreads and Kindle data which have social media features.

IV. Datasets and Sources

A. New York Times Bestselling Dataset

To perform analysis on bestselling books and authors, data was fetched using NewYork Times API. NewYork Times has a developer website which allows public to access their database. This acted as the data source for the visualization and analysis. To make requests using API, a prerequisite is to open an developers account on the website: https://developer.nytimes.com/apis. It consists of various different APIs and for the purpose of this project, Books API is selected and the API key is generated for later getting the data programmatically. API documentation: https://developer.nytimes.com/docs/books-product/1/overview. Fig. 1, shows the API call and get request for fetching the data.



```
def get_req(category, pub_date):
    API_ID = "ETahVcAWG562iMLKgrssvULayRTNhTZD"
    query = "api.nytimes.com/svc/books/v3/lists.json?list="+category+"&published_date="+pub_date+"&api-key="+API_ID
    res = requests.get("https://"+query)
    res_json = res.json()
    time.sleep(6)   ### Pause for 6 seconds to not exceed time limit of data getting via API ###
    return res_json
```

Fig. 1: API call to fetch New York Times Bestselling data

The Dataset consists of attributes such as "list_name", "author", "title", "weeks_on_list", "rank", "bestseller_date", "published_date", and more 19 attributes. These are main attributes used for the aim of the project purpose. Attribute "list_name" is the book category name, "weeks_on_list" is how many weeks the book was on the bestselling list, "rank" is the rank of the book in the bestselling week, "published_date" is the date when the bestseller list was published.

B. Library Usage Dataset

Library Usage Data is derived from Dan Francisco's Integrated Library System which is composed of the records including the data on patron records and circulation data. This data includes the patron registration year and since that time how heavily the library is being utilized by the patrons based on the checkout and renewal activities. This dataset consists of 4,23,448 records with 15 attributes such as Patron_Type_Definition, Total_Checkouts, Total_Renewals, Age_Range,Home_Library_Definition,Year_Patron_Registered and 9 more attributes. These attributes become the factors for the analysis of library usage in the past years. Total Renewals and checkouts describe how heavily the library is being utilized which can be further categorized using age range. Registration analysis is based on Patron Registration year and patron type definition.

C. Goodreads Dataset

The dataset used for the analysis of books and authors is from Goodreads [8]. It contains 19 attributes and 30006 observations for each book. The dataset includes information such as title, author, rating, voters count, publication and description. These features can be used to perform exploratory data analysis and predicting bestselling books and features contributing to that

D. Kindle Dataset

The dataset used here is the kindle dataset [9]. This dataset contains information from the kindle platform regarding e-books, their prices, authors, publishers, customer reviews and information regarding availability of Kindle features. This data can be used to perform exploratory data analysis and observe trends in what content customers are more inclined towards in the e-book market.

V. Methodology

As discussed in the previous section, this project consists of 4 sub-research questions which cumulatively is used for solving the main research topic of the project. For all the individual sub-research tasks, the method followed to perform the final visualization is same. For Datasets Library Usage, Goodreads, and Kindle, the data is extracted from Kaggle website in a semi-structured format where-as for New York Times Bestselling books, the data is fetched from the New York Times Bestselling Developers API as discussed earlier in the previous section. The method followed is explained in step-wise manner as follows:

1) For individual sub-research tasks, a semi-structured data is downloaded in json format, from either a data portal or using API.
2) The downloaded data is stored in MongoDB database, as it is suitable for NoSQL data format objects.
3) The unstructured data is structurised in a tabular format and cleaned for the main visualization and research findings.
4) The structured data is stored in the PostGreSQL database.
5) The visualization and research is performed by retrieving the structured data from PostGreSQL database.

All the sub-parts of the project is performed using Python programming language. Fig. 2, illustrates the project architecture.

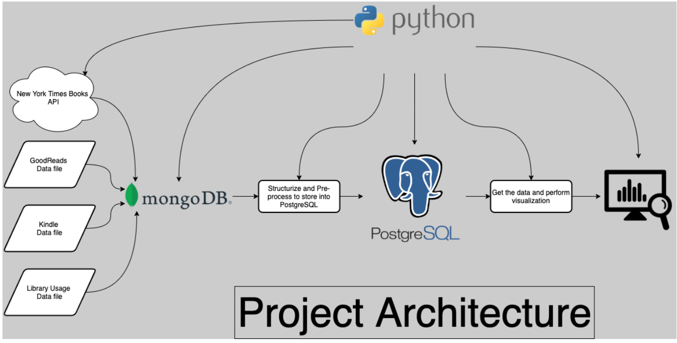


Fig. 2: Architecture

For the New York Times Bestselling Analysis, data is fetched using New York Times Developers API. The NewYork Times Books API allows maximum request limit per hour, thus, data was retrieved by pausing the process for 6 seconds per request made. For the purpose of this project, 3 years of data is downloaded for 4 different categories only. Only 4 categories are chosen as time taken for downloading best-selling books for all the categories would have taken many days and NewYork Times API has a limit to download. After retrieving the data, a copy is stored on the local machine. After downloading, the above process is followed to create the required visualization.

VI. Results and Discussion

This section focuses on the analysis performed on the cleaned structured data and findings for each dataset research question.

A. *New York Times Bestselling Data*

To perform BestSelling analysis, books, authors and categories of books which were in bestselling list is considered for our project aim. First visualization is performed to know which book categories dominated the bestselling list every year. From the fig. 3 , it is seen that NewYork Times database has an equal number of bestselling books per category stored except for year 2020. Number of Best Selling Books for all 4 categories in our data for years 2018, and 2019 is 795 and for year 2020, it is 735.

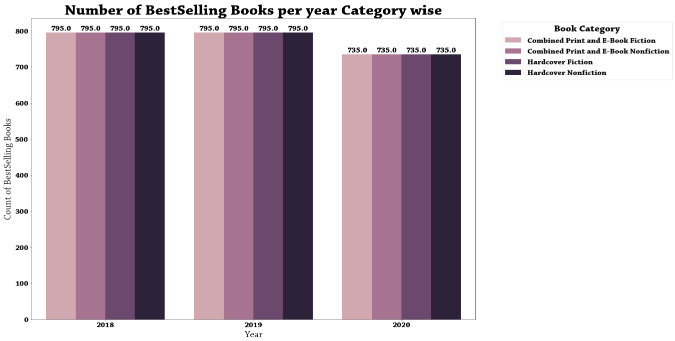


Fig. 3: Dominating Book Categories in BestSelling List every Year

Second visualization is performed to know top 5 authors having highest bestselling books on the list. Line chart is used to perform this visualization. For Year 2018, from fig. 4, we see that ”Tara Westover” with around 88 was the top bestselling Author. In Year 2019, again ”Tara Westover” and a new author in the list ”Delia Owens” with around 107 times were on the bestselling list for all their books. In Year 2020, ”Delia Owens” maintained her top position from 2019 with a score of around 95. A continuous yearly increase is seen for author ”John Grisham” in no. of times being on the bestselling list. The lines denote those authors whose books were consecutively in the top bestselling list over 2-3 years.

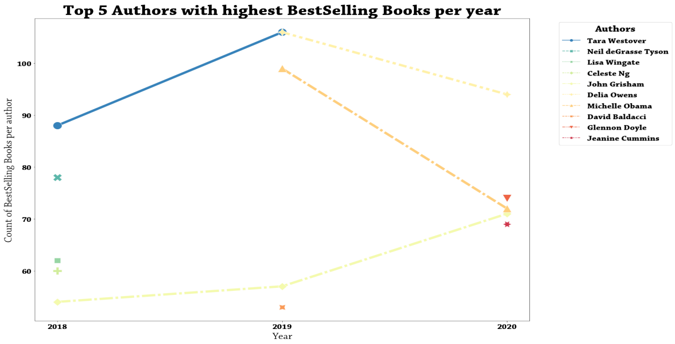


Fig. 4: Top 5 Author having more BestSelling Books per Year

Next visual focuses on which top 20 authors were on the bestselling list for the longest period of time in the span of 3 years. Bar chart is used to carry out the above findings. From fig. 5, we see in the time period 2018-2020, ”Tara Westover” was the top author ranking first for being on the Best Selling List for 132 weeks in total. Second and Third position to secure are ”John Grisham” and ”Ta-Nehisi Coates” being for 126 weeks.

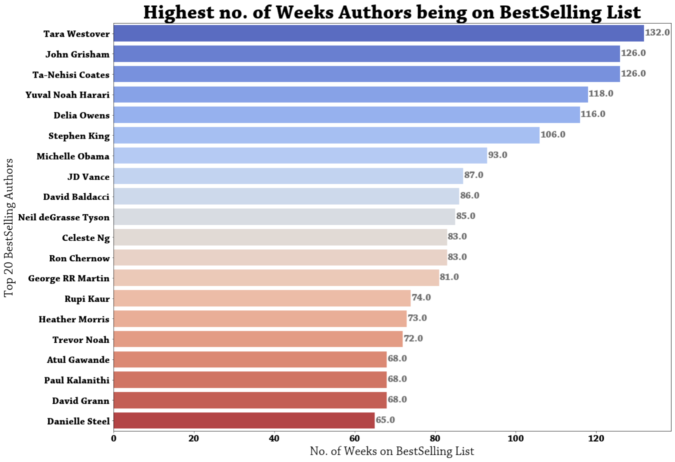


Fig. 5: Top 20 Author’s books(all) on best selling week for longest period

The fourth graph, fig. 6, depicts the top 15 books which showed up for the longest period on the bestselling list in 3 years. For all these 3 years, Book named ”Educated” is seen for the highest of 132 no. of weeks times on the BestSelling List. Second Book seen most of the time, 116 weeks, on the list is ”Where The Crawdads Sing”.

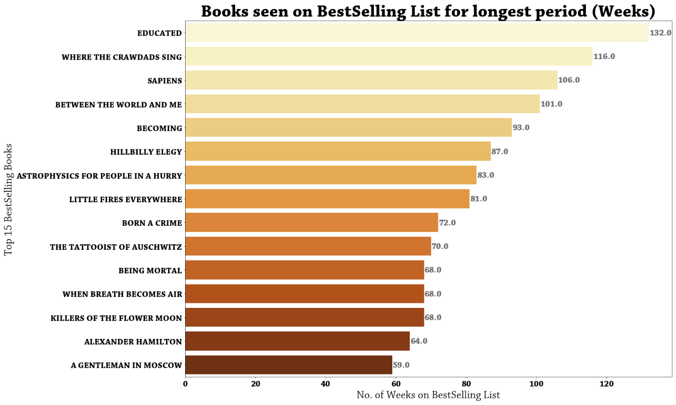


Fig. 6: Which books showed up longest on the bestselling list (Top 15)

B. Library Usage Data

Fig 7, shows the plot for the count of the users registered in the library based on the age range, it can be observed that most of the users are of age 25 to 34 years, followed by 35 to 44 years people age group. Least library members are from 60 to 64 and 75 and over age groups.

Fig 8, bar graph shows how the registration were highest in the year 2000, close to 2.5k which drastically decreased to less than 500 in 2014 and then there is a gradual increase in the registrations over the year increasing up to 2014 and again a sudden decrease is observed post 2014.

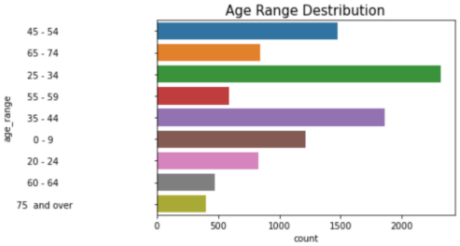


Fig. 7: Plot showing the count of Age Range Distribution of the library users

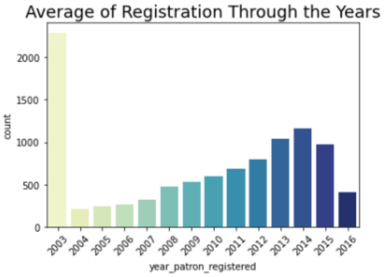


Fig. 8: Average number of the Registration through the years

Fig 9, shows the decrease in the trend of checkouts and renewals activities from the library over the years which clearly shows the decrease in library usage over the years.

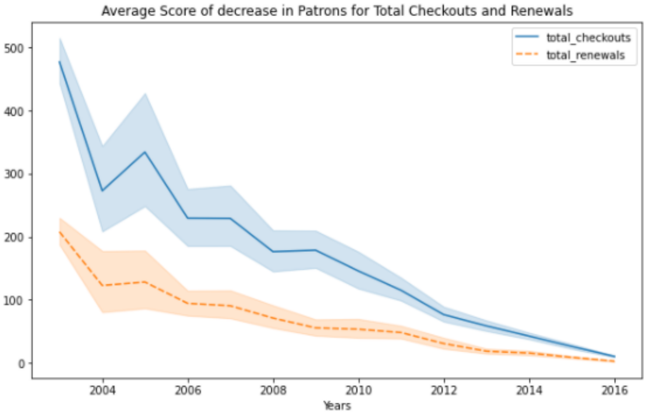


Fig. 9: Average number of renewals and checkout counts over the years

Fig 10, graphs show that even though the number of registrations is less for the age range 60 to 64 years, it is most utilized by these users as it can be observed that the number of renewals and checkouts is high for this age range people. On the other hand, users with the age range 25 to 34 years least utilize the library even though the most of users from this category.

C. Goodreads Dataset

Fig. 11, Histogram representing how often different ratings are occurring in the dataset. We can observe a normal distribution of the rating with a peak of 3600 at rating 4. Also, there is a small peak at rating 5 and very few books with rating 0.

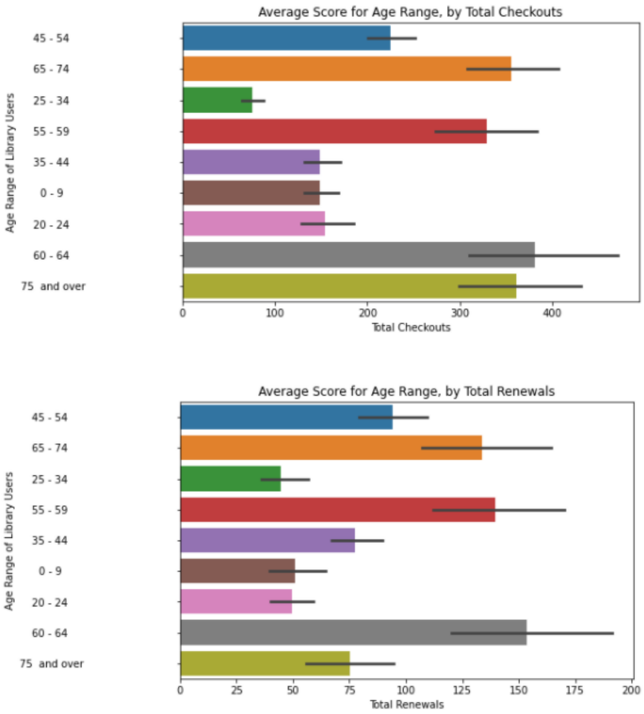


Fig. 10: Average Score for Age Range by total checkouts and renewals

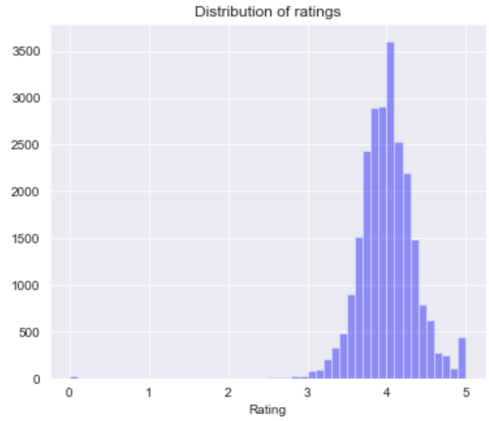


Fig. 11: Distribution of Ratings

Fig. 12, Bar graph showing top 10 books on the basis of rating and voters count on Goodreads. We can observe that "The Keystroke Killer: Transcendence" is the highest 5 rated book with 38 voters followed by "Broken Promises: The Suspenseful Sequel To The Novel, I, Beauty" with 28 voters.

Fig. 13, Bar graph showing top 10 publishers on the basis of total rating on Goodreads. Here we can see a significant difference between the rating of "CreateSpace" and other publications. CreateSpace is the most popular self-publishing service owned by Amazon where authors can publish their books without Editing and verification.

Fig. 14, Represents top 10 authors on the basis of ratings. We can see that "Erin Hunter" is the highest rated author followed by "Stephen King" and "James Patterson".

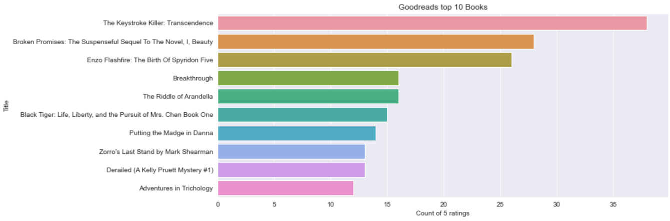


Fig. 12: Top 10 Books on Goodreads

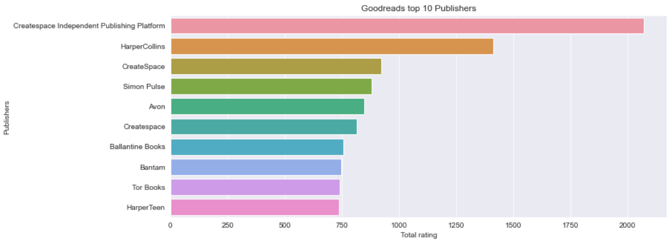


Fig. 13: Top 10 Publishers on Goodreads

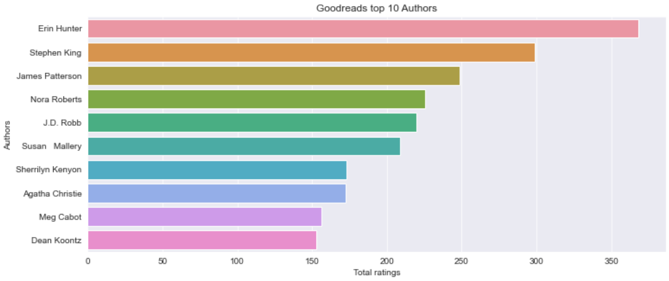


Fig. 14: Top 10 Authors on Goodreads

D. *Kindle Dataset*

Fig. 15, the first visualization, is on the e-books with reference to their customer rating. Here we get to see what books have been customer favorites over the time that this data has been collected. For the purpose of this visual we have represented the top 15 books with the highest customer ratings. From the bar plot, we see that the highest customer rated e-book is "The Hunger Games" and it has 19723 positive customer ratings. Followed by "Gone Girl: A novel" at 15,866 positive

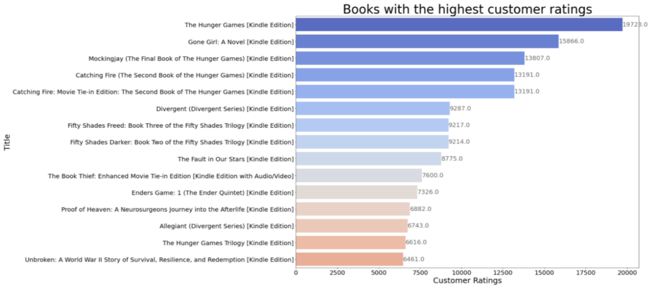


Fig. 15: Customer ratings

In the next graph, fig. 16, we find information regarding the 15 highest rated authors on the platform according to customers. We see that Gillian Flynn has the highest number of positive ratings at 15866 positive reviews.

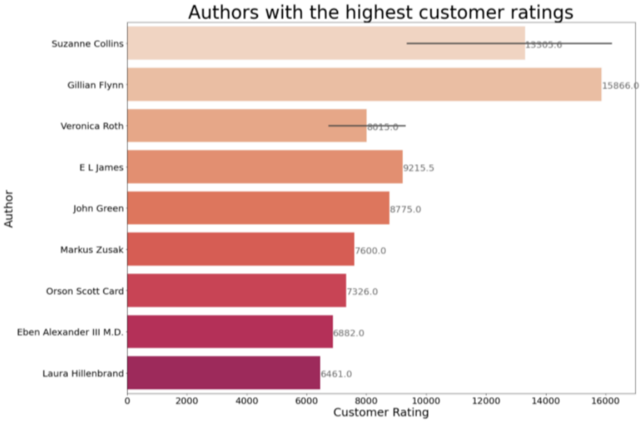


Fig. 16: Authors with the highest customer ratings

In next visual, fig. 17, we look at the 10 costliest books from the entirety of the data. From the graph, we see that "Fiscal Administration" is the book with the highest price at $186.49. Followed by, "Microeconomics" at $124.02 and "Construction Accounting & Financial Management" at $100.46.

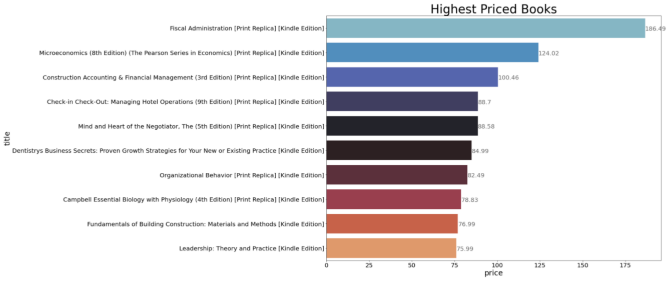


Fig. 17: Highest Priced books

VII. Conclusion and Future Work

In the analysis of Library Usage dataset, it is found that with time there is a clear downward trend in library usage. Although there are a large number of registrations happening from youth but by the renewals and checkouts activities in libraries, it is observed that the youth are not engaging in utilizing the library services. To increase the usability of library services to best of its use, analysis is performed on the online platforms providing e-book services such as Kindle and Goodreads. Also, BestSelling books overall on all platforms are analysed from New York Times Bestselling Data. Using these datasets, useful information on the current trending books, authors, publishers, genres based on the public rating and reviews given to these books is obtained. There are some books which are on top of best-Selling list for the longest period and some of the authors are very popular among people for their contents.

Currently, library access is available for a limited number of hours which also adds on the decrease in library usage, so if the information on the public interests for the best author, books, publishers, genres is used efficiently then it can help in optimizing the future services of Library and ensuring digital availability of the library services. The future scope of this project is to perform deeper analysis from many books sources and online platforms and avail these analyses to build a system for libraries to increase library usage by enhancing the services.

References

[1] Kaggle data Repository, San Francisco Library Usage, Anonymized library usage data by over 420,000 patrons. Available From: https://www.kaggle.com/datasf/sf-library-usage-data.

[2] "About the Best Sellers". Accessed on: Jan. 5, 2021. [Online]. Available: https://www.nytimes.com/books/best-sellers/methodology/

[3] Kumah, Cynthia H., "A Comparative Study of use of the Library and the Internet as Sources of Information by Graduate Students in the University Of Ghana" (2015). Library Philosophy and Practice (e-journal). 1298. http://digitalcommons.unl.edu/libphilprac/1298.

[4] Chamani Gunasekera, "Students Usage of an Academic Library: a user survey conducted at the Main Library University of Peradeniya" Journal of the University Librarians Association of Sri Lanka 14(1) DOI: 10.4038/jula.v14i1.2687.

[5] Antonios Makris,Konstantinos Tserpes, Giannis Spiliopoulos, Dimitrios Zissis, Dimosthenis Anagnostopoulos in "MongoDB Vs PostgreSQL: A comparative study on performance aspects" Received: 7 August 2019 / Revised: 10 March 2020 / Accepted: 13 April 2020 /©The Author(s) 2020, corrected publication 2020 DOI: 10.1007/s10707-020-00407-w.

[6] K. Wang, X. Liu, and Y. Han, "Exploring Goodreads reviews for book impact assessment," J. Informetr., vol. 13, no. 3, pp. 874–886, Aug. 2019, doi: 10.1016/j.joi.2019.07.003.

[7] Tong Bao, Tung-lung Steven Chang, "Why Amazon uses both the New York Times Best Seller List and customer reviews: An empirical study of multiplier effects on product sales from multiple earned media," Decision Support Systems, Volume 67, 2014, Pages 1-8, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2014.07.004.

[8] "Books descriptions & movie adpt or not more." https://kaggle.com/davidpitts2bds/30k-books-tagged-if-made-to-movie-or-not-and-more (accessed Jan. 07, 2021).

[9] Kaggle data Repository, Kindle Books Dataset. Available From: https://www.kaggle.com/snathjr/kindle-books-dataset.