BOOK RECOMMENDATION SYSTEM

(PROJECT REPORT – PYTHON FOR DATA SCIENCE)

Kavya Sree. Kaitepalli – [kavyasree.k20@iiits.in](mailto:kavyasree.k20@iiits.in)

Himaja. Anchuri – [himaja.a20@iiits.in](mailto:himaja.a20@iiits.in)

***ABSTRACT*:**

**The book recommendation system is an application which recommends books to the users considering the previous books they have read. The recommendation system works by collecting data on the user’s reading habits such as the genres and authors they enjoy and then prepares a list of books from the database which are nearly related to the books the readers loved.**

**Machine learning models are applied on the data, explicitly clustering techniques so that all the similar books will be brought together which makes it possible for book recommendations. This report provides a clear analysis of the dataset containing details of books and also takes you through the process of suggesting books based upon the results obtained from the clustering techniques applied on the data set. In this project, we have employed the k-means clustering technique as our primary approach for data analysis. K-means clustering is a widely-used unsupervised learning algorithm that allows us to partition a dataset into k distinct clusters based on their similarities. The k-means clustering technique proved to be a highly effective approach to book recommendation, as it was able to identify patterns and similarities among books that could be challenging to detect manually. By using this method, we were able to group books into distinct clusters, and recommend books to users based on their preferred cluster.**

**Keywords:** K-means clustering, unsupervised learning algorithm

**INTRODUCTION:**

Reading books is a widely popular activity enjoyed by a vast majority of people. A book recommendation system is essential for such readers as it helps them discover new books that align with their interests and reading preferences. With millions of books available, it can be challenging for readers to find books that resonate with them, and a book recommendation system can simplify this process. The book recommendation system can help users discover new books of their type which they might not have found otherwise, leading to a more satisfying reading experience. The system is a valuable tool for book lovers, offering a personalized and efficient way to discover new books and authors who write books in the preferred genre the reader. Moreover, a book recommendation system can help users discover books from new genres or authors that they may not have considered otherwise, leading to a more diverse reading experience. This can be particularly beneficial for readers who are looking to expand their reading horizons or explore new topics.

Book recommendation systems are utilized across multiple platforms. Online libraries such as Wattpad, book net, web novel, good reads use this kind of book recommendation system to provide options to their users. Not only the online libraries, but also many e-commerce websites, such as Amazon and Barnes & Noble, use book recommendation systems to suggest books to their customers based on their search and purchase history. Mobile apps such as Scribd and Oyster, social media platforms like Facebook and twitter also use book recommendation systems to suggest books to their readers depending upon the readers interest, reading history likes and shares.

**DATASET DISCRIPTION:**

The books dataset used in the project was obtained through the Goodreads API. The dataset consists of 12 columns and the rows are 11127 in number implying that the dataset is big enough to apply clustering techniques and get a good number of book recommendations.

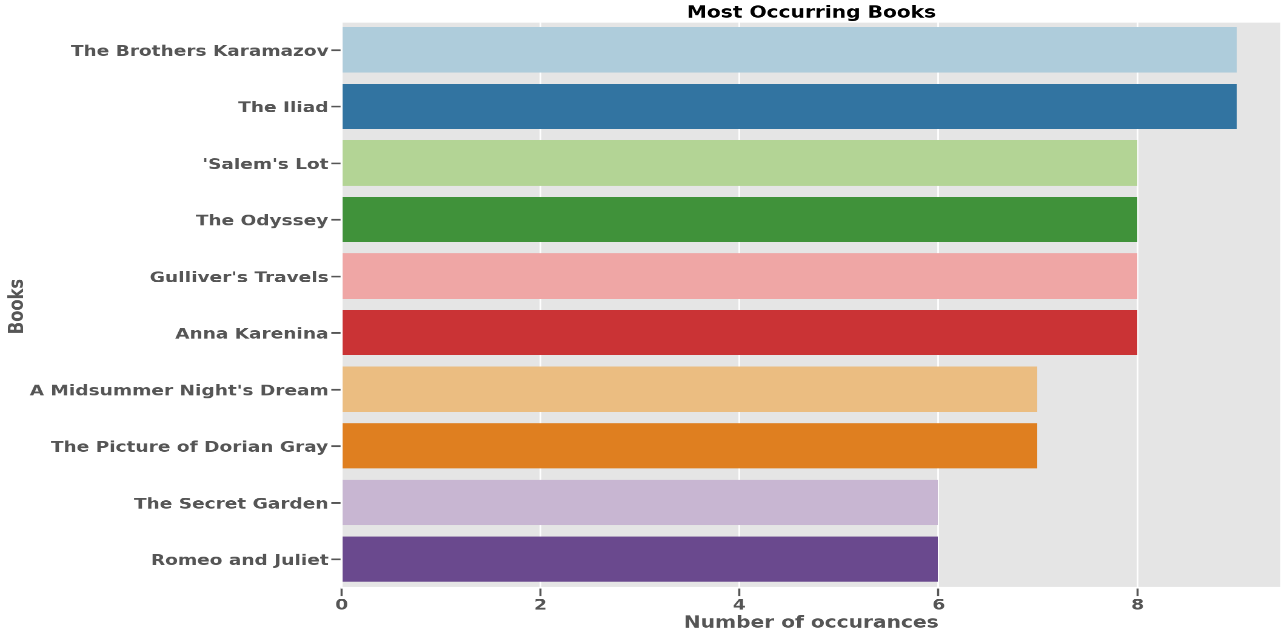
Description of columns in the dataset:

* **bookID:** Contains the unique ID for each book/series.
* **title:** Contains the titles of the books.
* **authors:** Contains the author of the particular book.
* **average\_rating:** the average rating of the books, as decided by the users.
* **ISBN:** ISBN (10) number, tells the information about a book - such as edition and publisher.
* **ISBN 13:** The new format for ISBN, implemented in 2007 (13 digits).
* **language\_code:** Tells the language in which the book is written.
* **Num\_pages:** Contains the number of pages for the book.
* **Ratings\_count:** Contains the number of ratings given for the book.
* **text\_reviews\_count:** Has the count of reviews left by users.

**EXPLORATORY DATA ANALYSIS:**

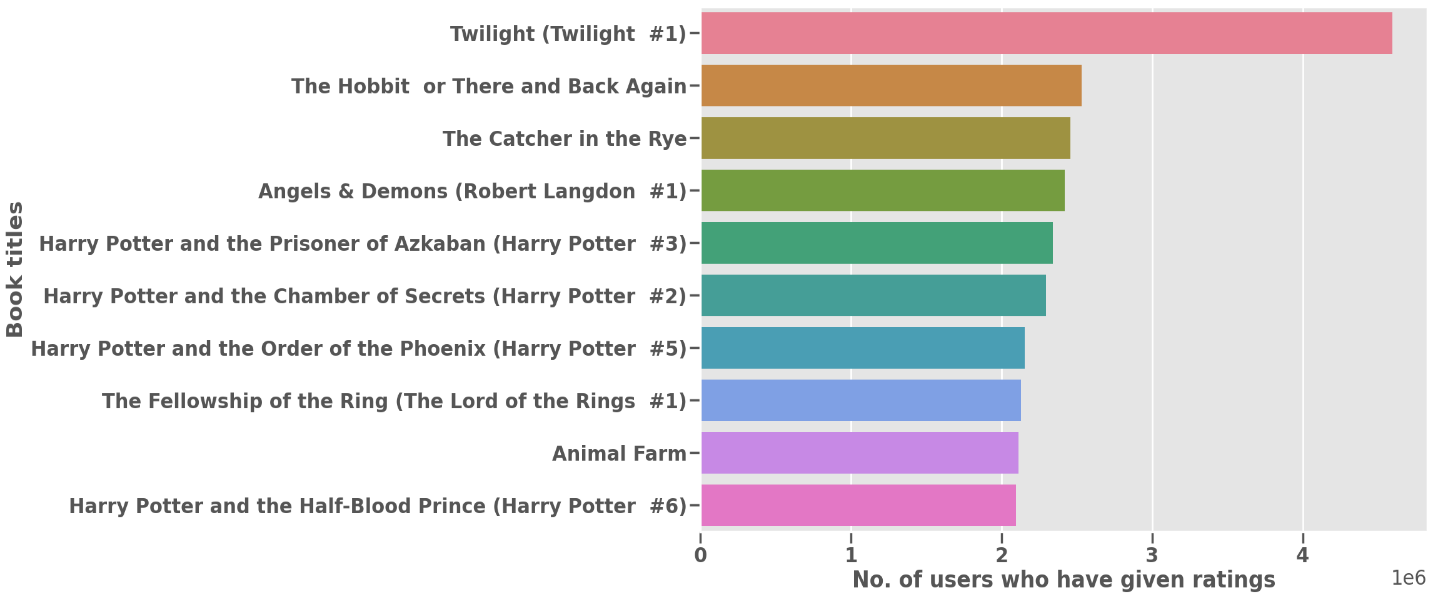
* **Books vs No. of occurrences:**

The following section presents an analysis of the publication frequency of various books. Specifically, the number of times each book has been published will be examined**.**



The analysis of the dataset reveals that 'The Brothers Karamazov' and 'The Iliad' have appeared the most frequently, with various publication editions bearing the same name. While this suggests that there is sustained interest in these books over time, it is important to note that the number of publications alone does not necessarily indicate a book's success. Other factors, such as marketing, distribution, and historical context, could also contribute to a book's popularity. Therefore, while the high frequency of appearances of these books in the dataset could suggest that they have been well-received by readers, further investigation is necessary to determine the full extent of their impact and success.

* **Finding the top 10 most rated books:**

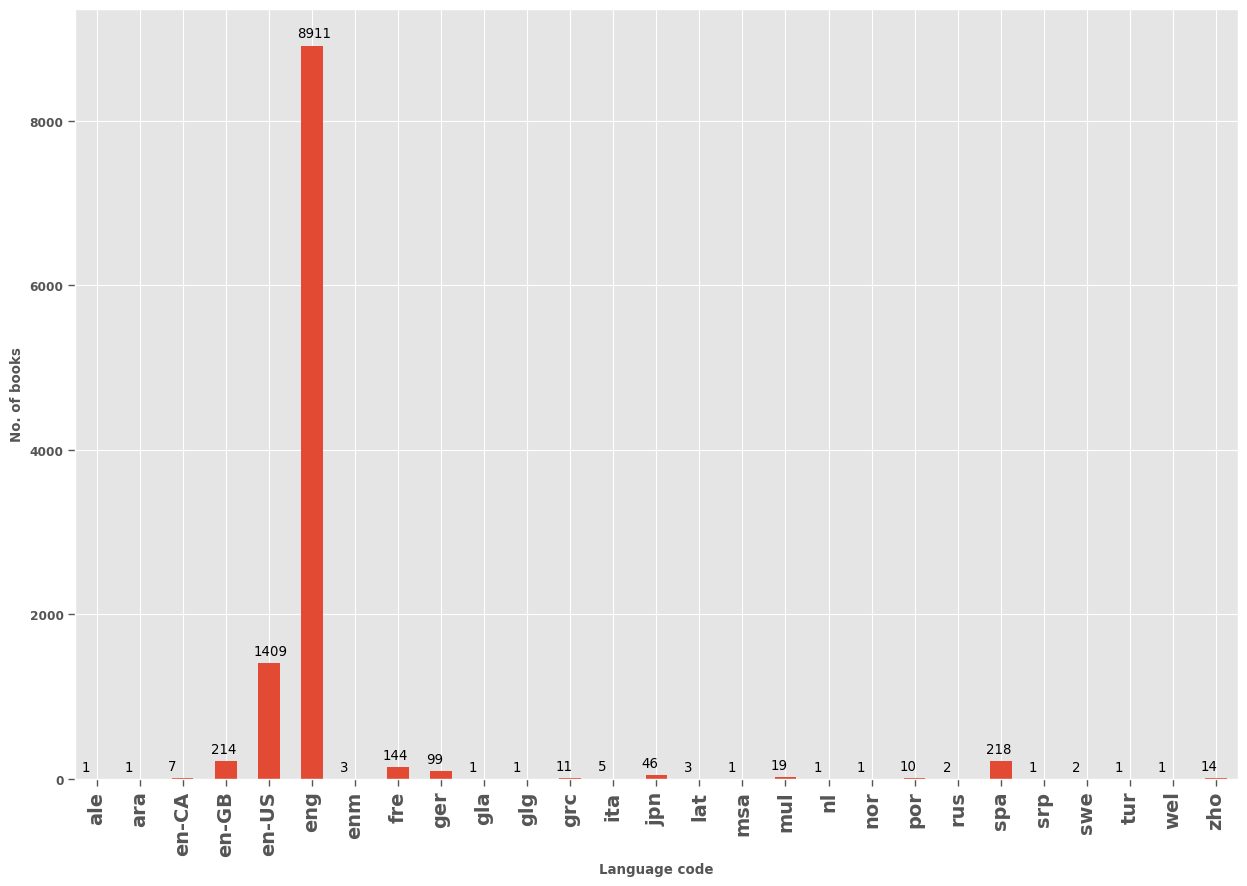
****

Upon analyzing the dataset, it becomes apparent that a significant number of readers have provided ratings for the books titled "Twilight" and "The Hobbit or There and Back Again". These books appear to have generated a notable level of reader engagement and interest, as evidenced by the number of ratings submitted by readers. This finding underscores the enduring popularity and cultural significance of these books, as well as the influence they continue to wield in contemporary literature.

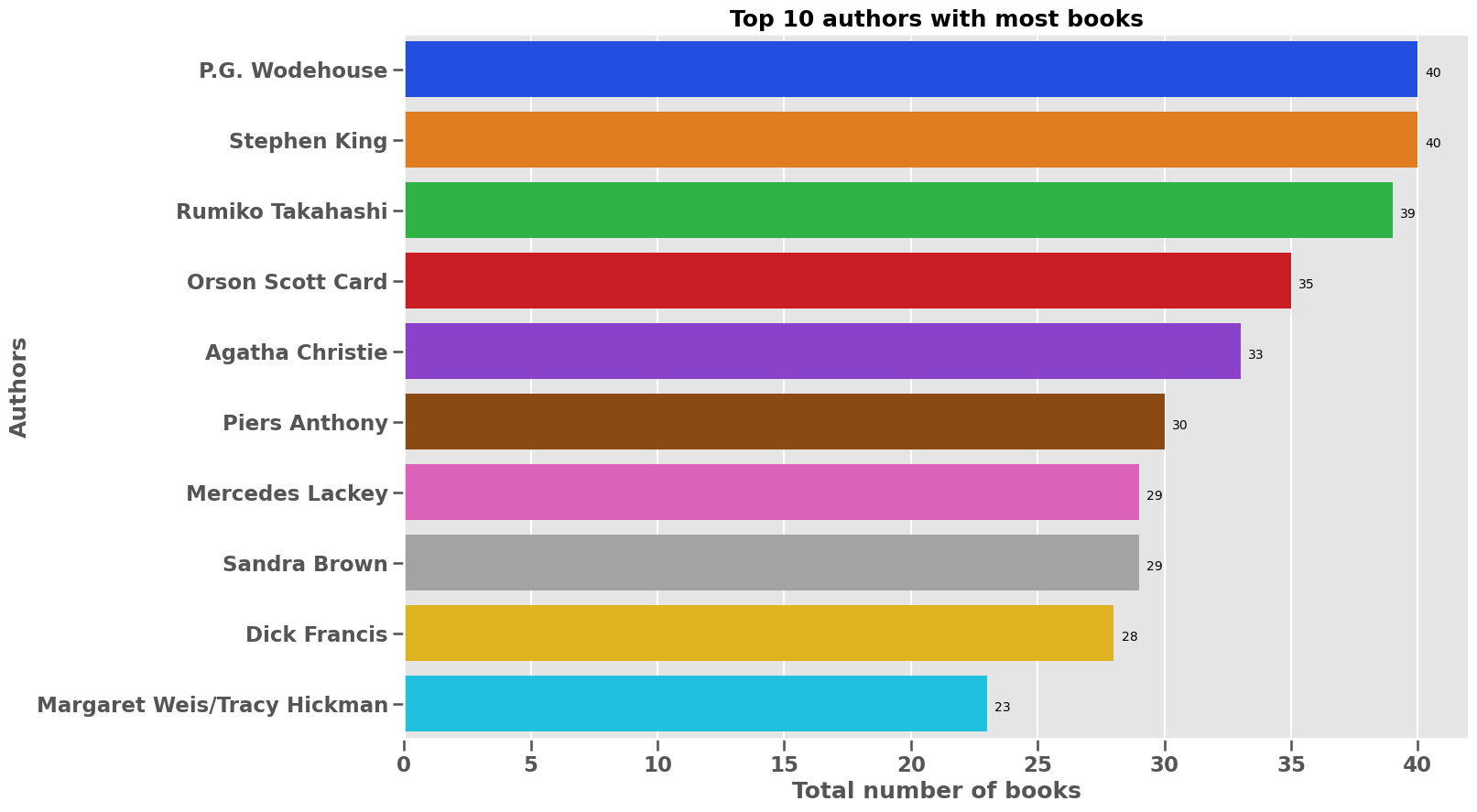
* **Distribution of books across languages:**

Upon analyzing the dataset, it is apparent that the majority of the books included are written in the English language. This observation suggests that English is the dominant language in the publishing industry, at least within the scope of the dataset analyzed. This finding underscores the importance of the English language in the global dissemination of ideas and knowledge through literature.

We can clearly observe it in the graph below…

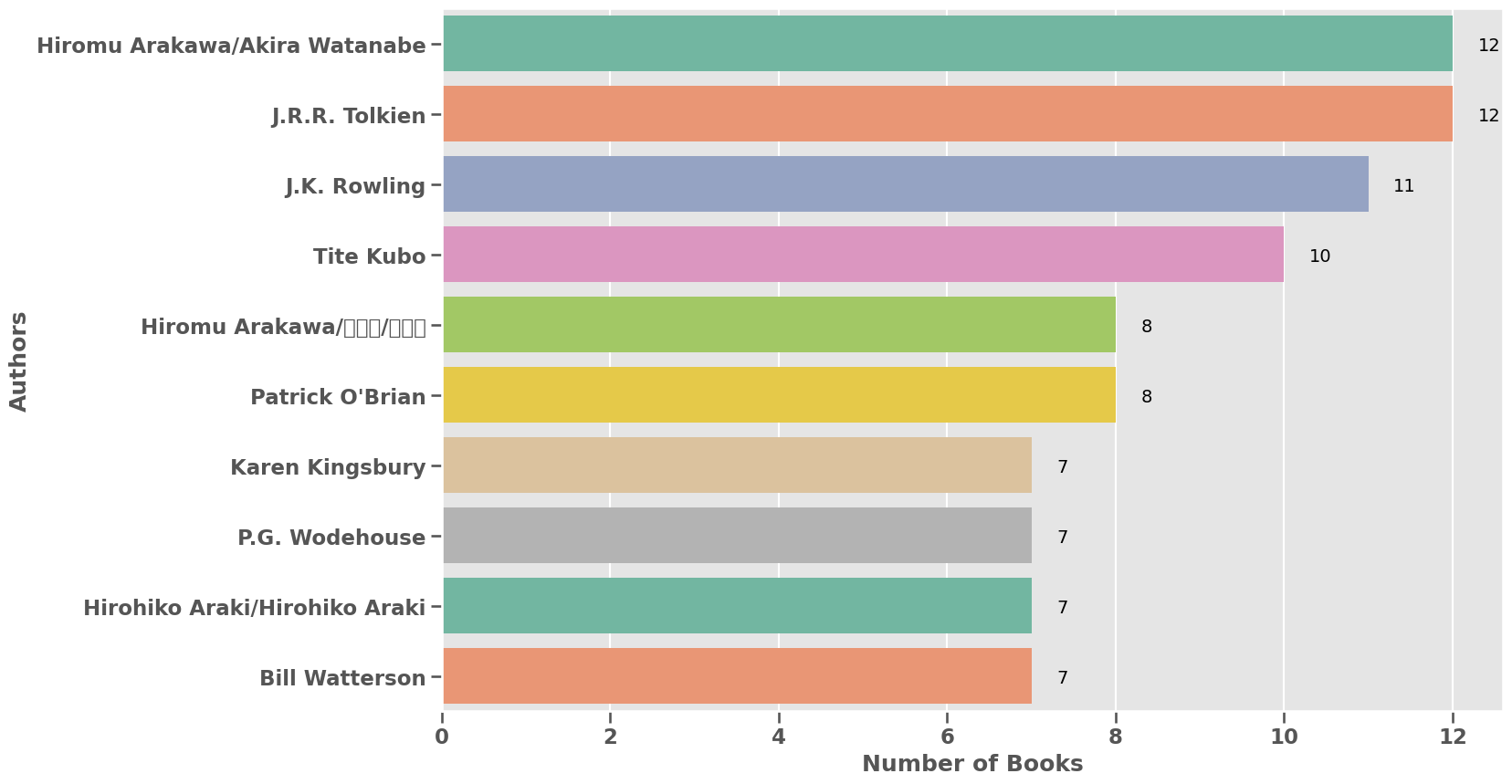


* **Finding the authors who wrote most books:**

****

Based on the findings in the preceding analysis, P.G. Wodehouse and Stephen King has written 40 books each out of which “The Most of P.G. Wodehouse” (4.48), “Life With Jeeves (Jeeves #6 2 & 4)” (4.39) of P.G. Wodehouse and “Carrie / 'Salem's Lot / The Shining” (4.54), “The Green Mile” (4.44) of Stephen King are the high rated books.

**Finding the top-rated authors:**

****

As per the analysis, 12 books from Hiromu Arakawa/Akira Watanabe and J.R.R. Tolkien are rated highly, which showcases their exceptional writing skills and popularity among readers.

Hiromu Arakawa is a renowned Japanese manga artist and writer, while Akira Watanabe is a Japanese light novel writer.

J.R.R. Tolkien was an English writer, poet, and professor who is famous for his epic high fantasy novels.

Top 5 highly rated books of Hiromu Arakawa/Akira Watanabe are from the series

“Fullmetal Alchemist”. The top 5 ratings of the volumes are as follows:

Fullmetal Alchemist Vol. 10 4.60

Fullmetal Alchemist Vol. 12 4.60

Fullmetal Alchemist Vol. 11 4.59

Fullmetal Alchemist Vol. 14 4.59

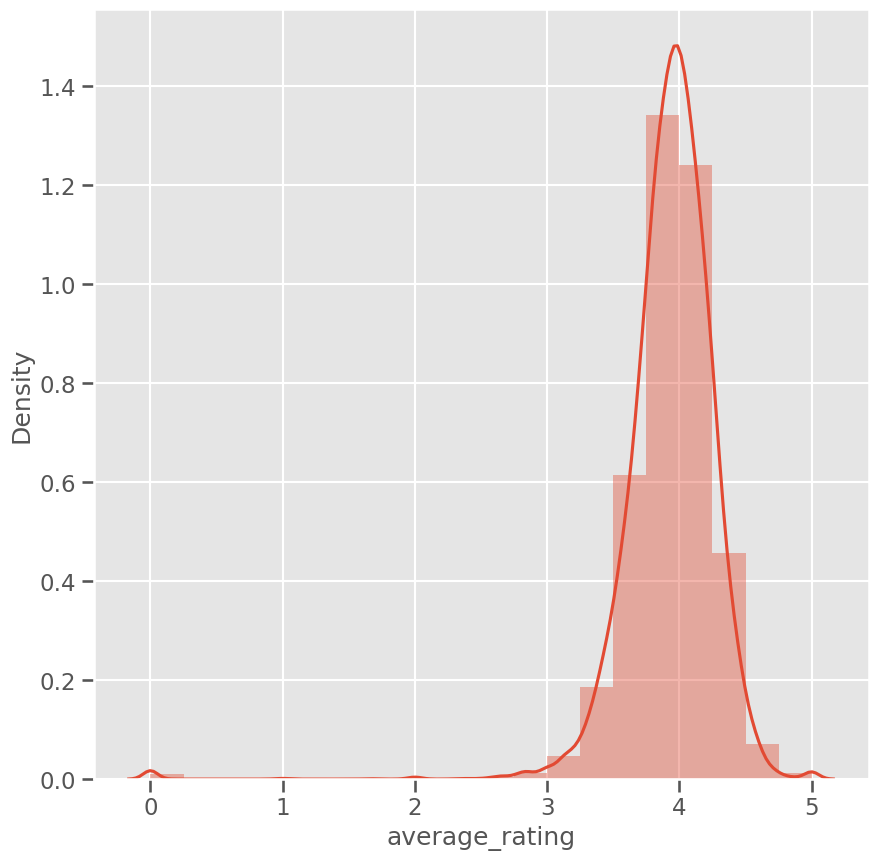
Fullmetal Alchemist Vol. 6 4.58

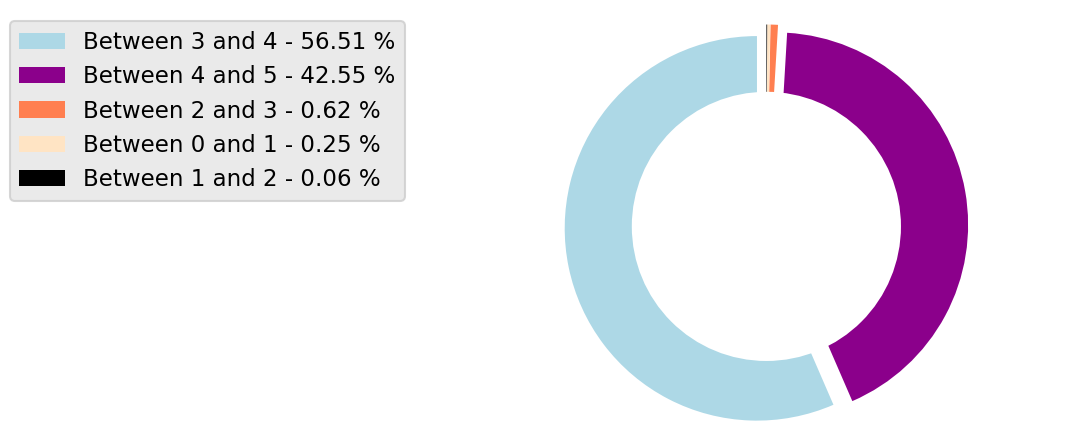
The top highly rated books from J.R.R. Tolkien are as follows:

The Hobbit and the Lord of Rings 4.59

The Return of the King (The Lord of the Rings #3) 4.53

**Rating Distribution of the books:**





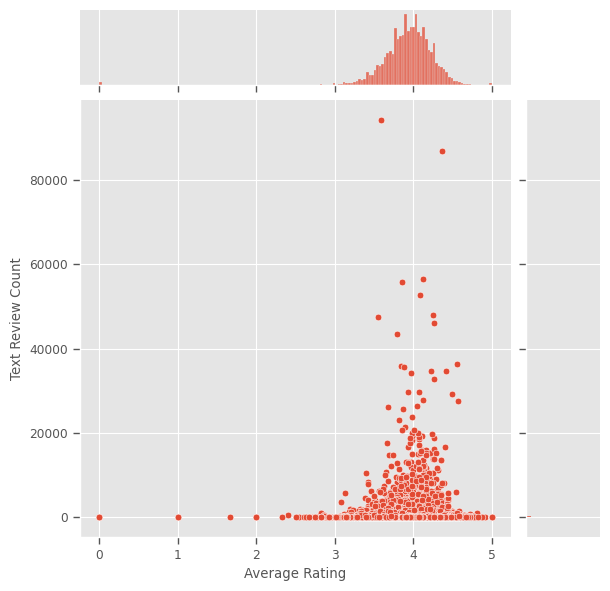
Based on our analysis of book ratings, it appears that most books tend to be rated around 4, with a relatively small number of books receiving low ratings. Additionally, we observed that the number of books decreases as we move further away from the 4 rating, towards both lower and higher ratings.

This suggests that, overall, readers tend to rate books positively, with a significant number of books falling within the 3.5 to 4.5 range. However, there are still some books that receive lower ratings, which may indicate that these books are not well-received by readers, or may have some weaknesses in their writing or storytelling.

It is worth noting that the exact distribution of book ratings can vary depending on the specific genre or category of books being considered. Some genres may have a higher proportion of highly-rated books, while others may have a more even distribution of ratings. Nevertheless, our analysis suggests that most books are generally well-liked by readers, with a smaller number of books receiving low ratings.

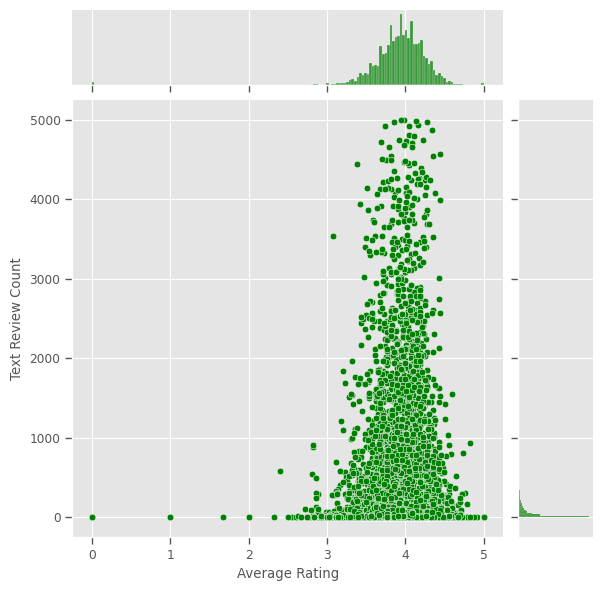
**Checking relation between rating and reviews count:**

Checking the relationship between ratings and review count of books can help to determine the credibility of the ratings and provide insights into how readers are responding to a particular book

****

Our analysis of book ratings revealed that the majority of books tend to receive ratings between 3-4, with the number of books decreasing as we move towards both lower and higher ratings. We also found that there is a significant number of books with reviews close to 5000.

However, to investigate the relationship between ratings and reviews more accurately, we chose to focus on books with less than 5000 reviews. By doing so, we aimed to reduce the impact of popular books with a large number of reviews, which may skew the relationship between ratings and reviews. After plotting the ratings and reviews for these books, we were able to observe a clearer pattern in the relationship between the two variables.

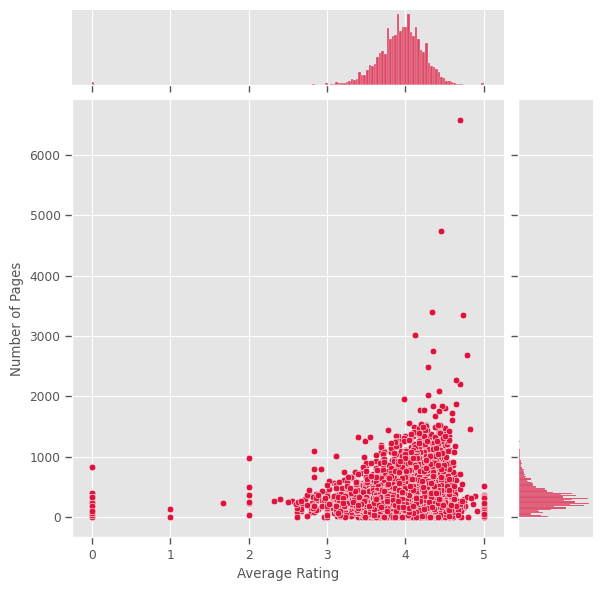


Our analysis of book ratings and reviews count revealed that the majority of reviews for books tend to fall within the rating range of 3-4, with the number of reviews decreasing as we move towards both lower and higher ratings. Additionally, we found that most of these reviews counted for books have a count of less than 1000.

However, we also observed that books with text reviews between 4000-5000 tend to have an average rating of 3-4, although the number of books in this category is relatively low. This suggests that a high number of reviews does not necessarily guarantee a higher rating, and that there may be other factors at play that influence a book's overall rating..

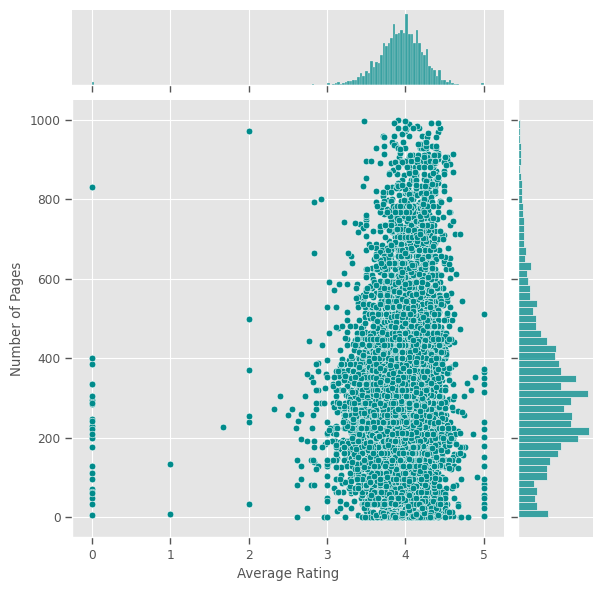
**Checking the relation between number of pages and ratings:**

Checking the relationship between the number of pages and the ratings of books can provide insights into reader preferences and how book length might affect their perception of the book's quality

****

Our analysis of book ratings and number of pages revealed that most books tend to have a relatively low number of pages, with the majority falling below 1000 pages. Furthermore, we found that these books tend to have an average rating between 3-5, with a relatively small number of books with more than 1000 pages.

To investigate this relationship further, we plotted the average rating against the number of pages for books with less than 1000 pages.



Our analysis of book ratings and books with number of pages less than 1000 revealed that books with a moderate number of pages tend to receive the highest ratings. Specifically, we found that the highest ratings were given to books with a page range of 200-400, with a peak near 250 pages.

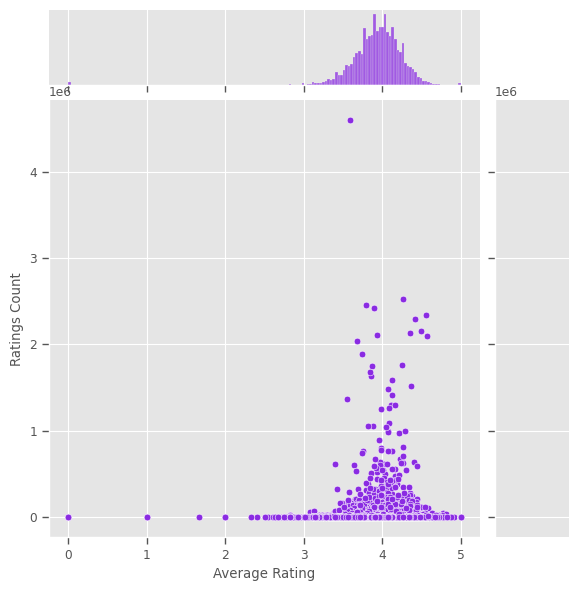
This suggests that most people prefer books that are not too long or too short, but fall within a moderate range of pages. Additionally, our analysis suggests that thicker books may be intimidating to some readers, leading them to choose books with a lower page count.

However, it is important to note that while the number of pages can influence a book's rating, other factors such as the quality of writing, the plot, and the characters can also have a significant impact on a book's overall popularity and success. Therefore, it is important to consider multiple factors when evaluating a book's quality and appeal.

Overall, our analysis highlights the importance of considering the number of pages when selecting a book to read or evaluating its popularity, but also underscores the importance of other factors such as writing quality and plot.

**Checking the relationship between ratings and ratings count:**

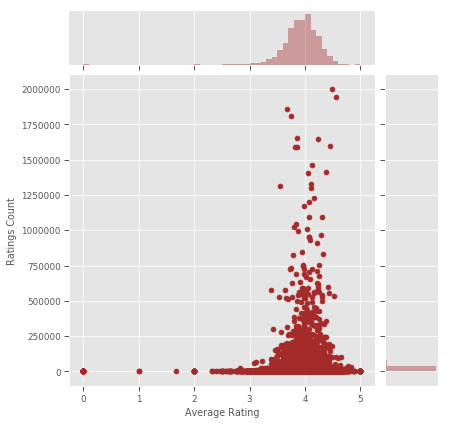
We need to find the relationship between average rating and ratings count of books because it can provide insights into the popularity and overall reception of a book. Books with a high average rating and a large number of ratings may indicate that the book is well-received by a large audience, while books with a low average rating and a small number of ratings may indicate that the book is less popular or less well-received. Additionally, analyzing the relationship between average rating and ratings count can help identify trends and patterns in readers' preferences, which can be useful for authors and publishers when deciding which books to publish or promote. Overall, understanding the relationship between average rating and ratings count can provide valuable insights into the success and popularity of books**.**

****

The graph above shows that the number of ratings for books increase with the average rating of the book. This means that books with a high average rating tend to also have a high number of ratings.

This can be explained by the fact that books that are highly rated are more likely to be read and recommended by others, which in turn leads to more people rating and reviewing the book. Additionally, popular books tend to receive more exposure and marketing, which can further increase their rating count.

In other words, books that are well-received by readers tend to attract more attention and positive reviews, which can contribute to their overall popularity and success. Therefore, understanding the relationship between average rating and rating count can be useful for both authors and readers in evaluating the quality and appeal of a book.



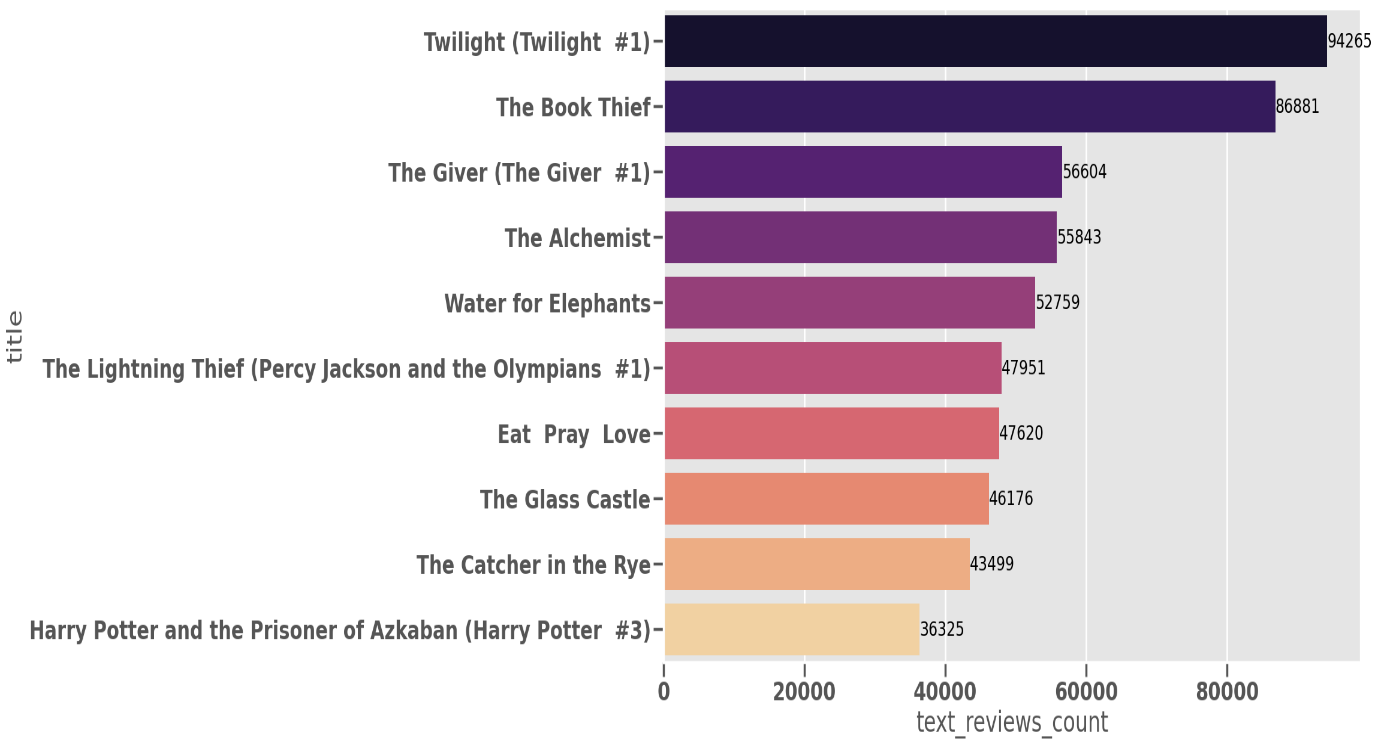
Based on our analysis of the graph, we can see that there may be a relationship between the average rating and ratings count of books. Specifically, as the number of ratings for a book increases, the average rating for that book tends to converge towards 4, with the ratings becoming sparser.

This suggests that books that are highly rated initially tend to attract a larger number of readers and ratings, which can contribute to their overall success and popularity. However, as the number of ratings increases, the rating for the book may become more moderate, as readers with varying opinions and preferences begin to leave their own reviews.

Additionally, the fact that the average rating becomes sparse as the number of ratings decreases may indicate that books with a lower rating count may be less popular or less widely-read, and may therefore attract fewer reviews overall.

Overall, this analysis highlights the complex relationship between average rating and ratings count, and suggests that a book's overall popularity and success may be influenced by a range of factors, including its initial reception, its marketing and exposure, and the opinions and preferences of individual readers.

**No. of books vs text review count:**



Based on our analysis of the data, we can conclude that while reviews are an important indicator of a book's success and popularity, there is no clear and consistent relationship between the number of reviews and a book's ranking.

**K-MEANS CLUSTERING**

Now that we have finished analysing the data, we can move on to applying clustering techniques to further explore patterns and relationships within the dataset.

Clustering is a popular unsupervised learning technique used in machine learning to group similar data points together based on their features or attributes. By grouping books together based on their shared characteristics, clustering can help us identify common themes, genres, or trends within the dataset that may not be immediately apparent from a visual or statistical analysis.

Here, we use k-means clustering technique to get the similar books together which makes it easy to recommend books to the readers.

We perform k-means clustering two times:

* k-means clustering with outliers
* k-means clustering without outliers

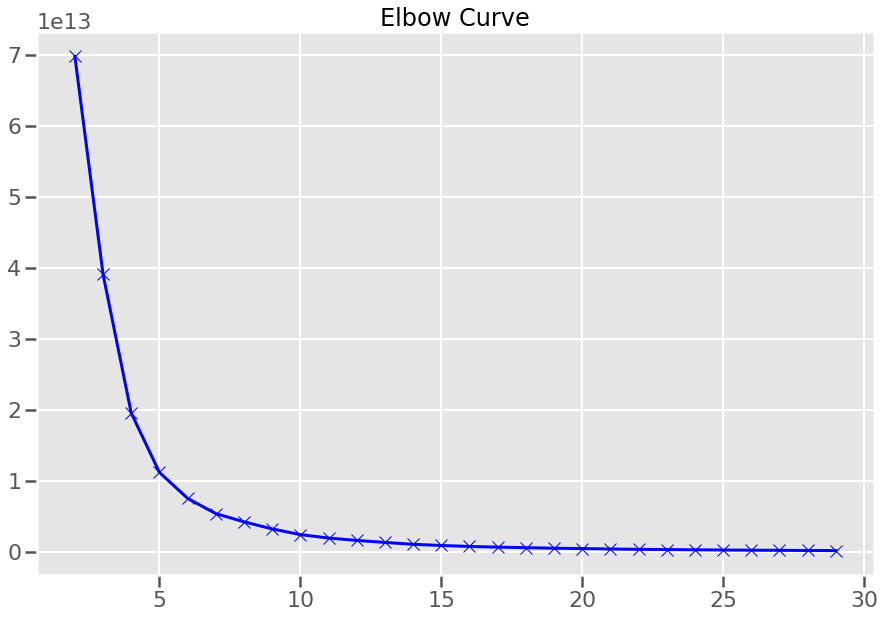
**Finding number of clusters:**

In this project, we employed the elbow method to determine the optimal number of clusters to use in our clustering analysis.

The elbow curve is a graphical method used to determine the optimal number of clusters for a k-means clustering algorithm. In k-means clustering, the objective is to partition a dataset into a given number of clusters, with the goal of minimizing the sum of squared distances between the data points and their assigned cluster centres.

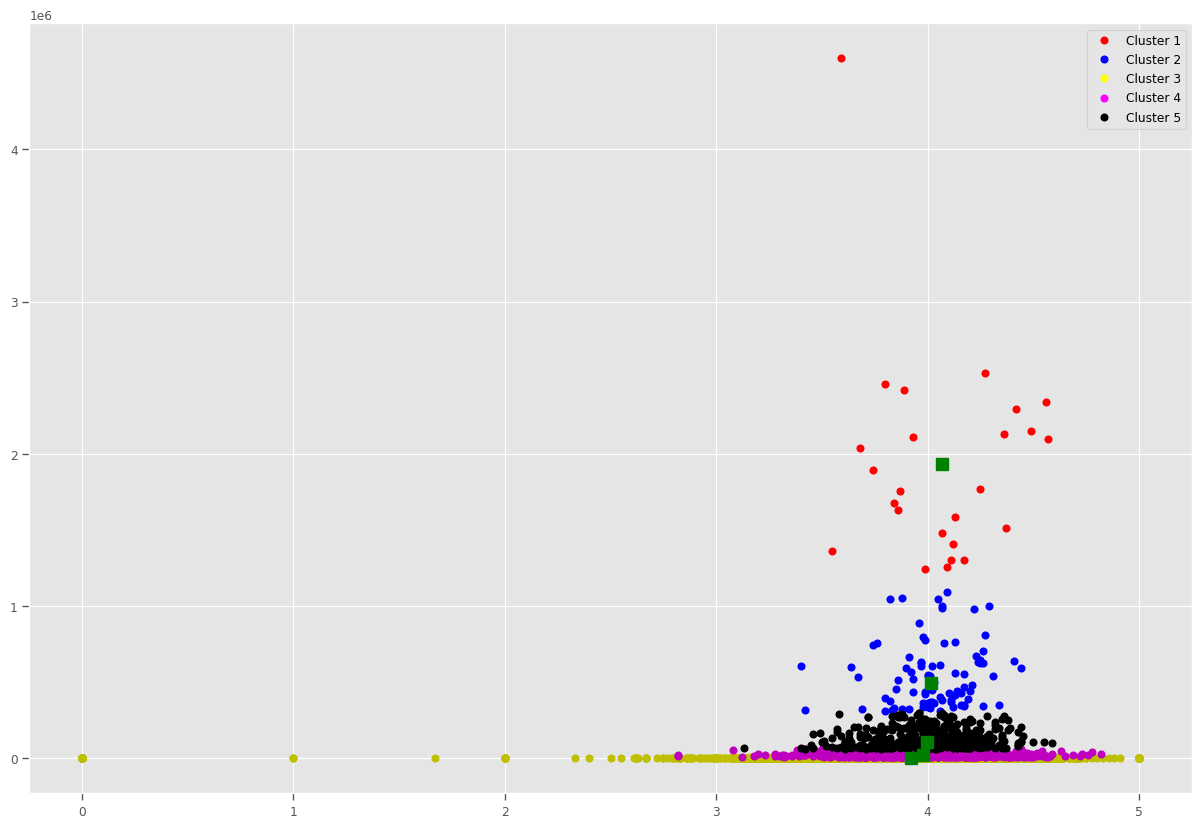
The elbow curve works by plotting the sum of squared distances against the number of clusters. The curve typically resembles an arm, with the "elbow" point representing the optimal number of clusters. This is the point at which the addition of more clusters does not significantly improve the sum of squared distances.

This point gives us the optimal number of clusters that can group the data properly.

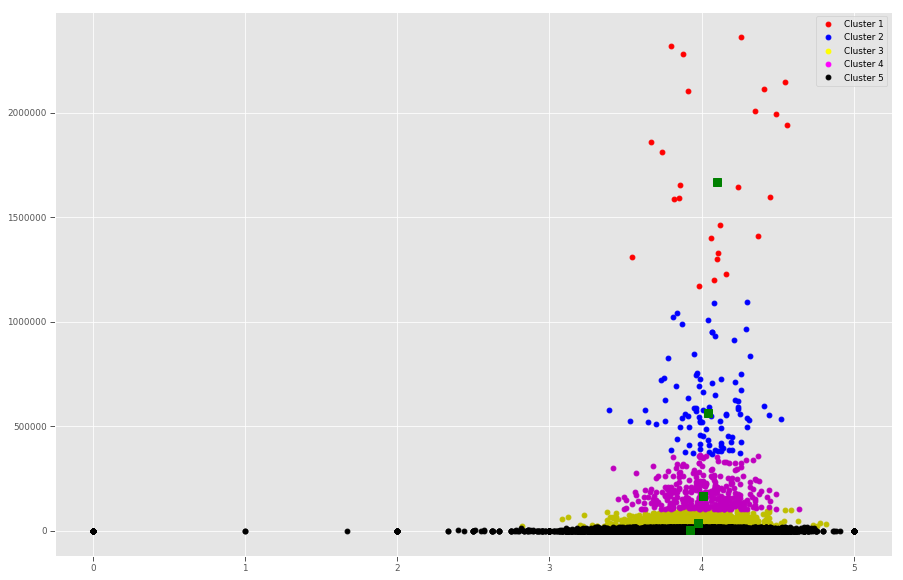


Based on the analysis of the elbow curve, it can be observed that the sum of squared distances decreases at a rapid pace up to the point where the number of clusters is 5. Beyond this threshold, the decrease in the sum of squared distances is not as significant. Hence, it can be inferred that 5 clusters would be an appropriate choice for applying k-means clustering in this context.

After clustering the data set, we got the 5 clusters and centroids:



We can see from the above plot, that because of two outliers, the whole clustering algorithm is skewed, that is, the centroids are getting displaced. Let's remove them and form inferences.



Now, we can see that the centroids are in correct place and the datapoints are properly grouped into different clusters.

As the rating count seems to decrease, the average rating seems to become sparser, with higher volatility and less accuracy.

**RECOMMENDATION RESULTS**

The results of book recommendations are as follows:

Book recommendations for “Warrior of the Light”:

* The Cat Who Walks Through Walls (The World As Myth)
* All Through the Night
* Question Quest (Xanth #14)
* Tolkien and the Great War: The Threshold of Middle-earth
* The Wanting Seed

Book recommendations for “Imaginary Friends”:

* The Fashion System
* Dragon's Treasure
* Kevin Trudeau's Mega Memory: How to Release Your Superpower Memory in 30 Minutes or Less a Day
* The Good Liar
* Exile's Children (Exiles #1)

Getting books by giving partial names of book (for "Harry Potter and the "):

* Harry Potter and the Half-Blood Prince (Harry Potter #6) 0
* Harry Potter and the Order of the Phoenix (Harry Potter #5) 1
* Harry Potter and the Chamber of Secrets (Harry Potter #2) 2
* Harry Potter and the Prisoner of Azkaban (Harry Potter #3) 3
* Harry Potter and the Half-Blood Prince (Harry Potter #6) 0
* Harry Potter and the Prisoner of Azkaban (Harry Potter #3) 3
* Harry Potter and the Chamber of Secrets (Harry Potter #2) 2
* Harry Potter and the Sorcerer's Stone (Harry Potter #1) 8876
* Harry Potter and the Philosopher's Stone (Harry Potter #1) 10678
* Harry Potter and the Goblet of Fire (Harry Potter #4) 10679

Getting similar books by giving book IDs as input:

* The Spider's House
* Siddhartha
* On Being Ill
* 'Salem's Lot
* Revolutionary Girl Utena Vol. 5: To Blossom

**CONCLUSION**

Our book recommendation system successfully identified books that are likely to be of interest to readers based on the features of language, average ratings, and ratings count.

By considering the average ratings and ratings count, we were able to identify books that have a high level of popularity and positive reception among readers. This allowed us to recommend books that have been well-received using K-Means Clustering and enjoyed by a wide audience.

Overall, our recommendation system has demonstrated its effectiveness in suggesting books that are likely to match the preferences and interests of readers.