**NAME : NIVETHA.V**
**DEPARTMENT : DECISION AND COMPUTING SCIENCE**

**Technical Assessment Task 1:**

1. Perform an EDA of this books dataset & share insights
2. Build a system to recommend more books to a reader based on a book already selected

**DATASET**

The dataset consists of

- bookID - containing the unique ID for each book
- 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 - unique 11 digit number to identify the book, International Standard Book Number.
- ISBN 13 - A 13 digit ISBN to identify the book
- language_code - Helps understand what is the primary language of the book. For example, eng is for English.
- Num_pages - Number of pages in the book
- Ratings_count - Total number of ratings the book received.
- text_reviews_count - Total number of written text reviews the book received
- Publication_date - Its tells when the book has been published
- Publisher - By whom it was published

| ookID | title | authors | average_rating | isbn | isbn13 | language_code | num_pages | ratings_count | text_reviews_count | publication_date | pub |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Harry Potter and the Half-Blood Prince (Harry ... | J.K. Rowling/Mary GrandPré | 4.5700 | 0439785960 | 9780439785969 | eng | 652 | 2095690 | 27591 | 9/16/2006 | Sch |
| 2 | Harry Potter and the Order of the Phoenix (Har... | J.K. Rowling/Mary GrandPré | 4.4900 | 0439358078 | 9780439358071 | eng | 870 | 2153167 | 29221 | 9/1/2004 | Sch |
| 4 | Harry Potter and the Chamber of Secrets (Harry... | J.K. Rowling | 4.4200 | 0439554896 | 9780439554893 | eng | 352 | 6333 | 244 | 11/1/2003 | Sch |

DATA PREPROCESSING

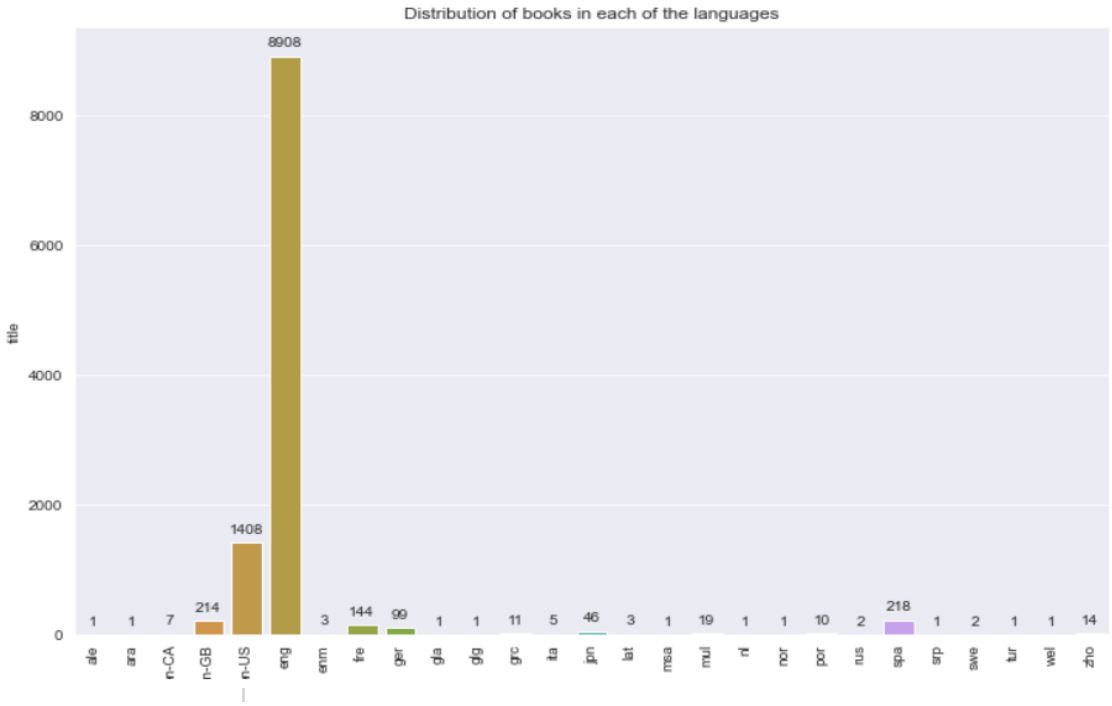
- Finding null values and duplicate values if any
  The dataset does not contain any null and duplicate values
- Counting the number of unique values
  The dataset consists of a number of unique values and it is printed
- Date value for the publication_date is different from one another. Organising the dates into a specified format.
- The dataset contains the value known as 'NOT A BOOK', that indicates that it has not been considered as a book. So those fields have been removed.

DATA VISUALIZATION

Description of the data tells us the summary statistics of the books_data.

| | bookID | average_rating | isbn13 | num_pages | ratings_count | text_reviews_count |
|---|---|---|---|---|---|---|
| **count** | 11,123.0000 | 11,123.0000 | 11,123.0000 | 11,123.0000 | 11,123.0000 | 11,123.0000 |
| **mean** | 21,310.8570 | 3.9341 | 9,759,880,247,639.1816 | 336.4056 | 17,942.8481 | 542.0481 |
| **std** | 13,094.7273 | 0.3505 | 442,975,846,058.3530 | 241.1526 | 112,499.1535 | 2,576.6196 |
| **min** | 1.0000 | 0.0000 | 8,987,059,752.0000 | 0.0000 | 0.0000 | 0.0000 |
| **25%** | 10,277.5000 | 3.7700 | 9,780,345,453,803.5000 | 192.0000 | 104.0000 | 9.0000 |
| **50%** | 20,287.0000 | 3.9600 | 9,780,582,461,536.0000 | 299.0000 | 745.0000 | 47.0000 |
| **75%** | 32,104.5000 | 4.1400 | 9,780,872,208,045.5000 | 416.0000 | 5,000.5000 | 238.0000 |
| **max** | 45,641.0000 | 5.0000 | 9,790,007,672,386.0000 | 6,576.0000 | 4,597,666.0000 | 94,265.0000 |
| **Variance** | 171,471,881.8133 | 0.1228 | 196,227,600,191,113,673,048,064.0000 | 58,154.5892 | 12,656,059,531.6634 | 6,638,968.5087 |

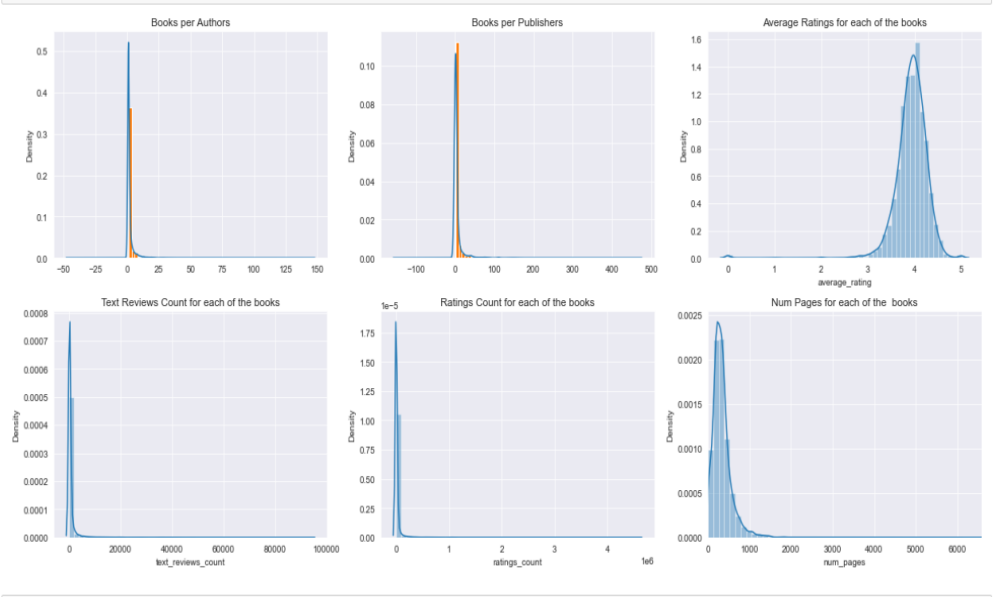
DISTRIBUTION OF BOOKS IN EACH LANGUAGE



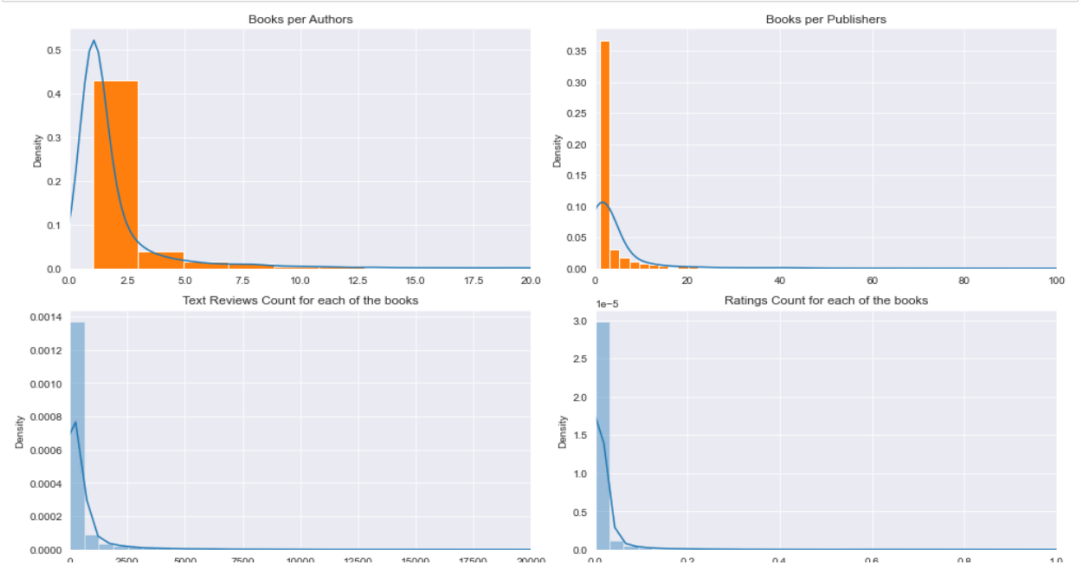

In the above graph, 90% of books were in english language and other forms of english like en-US, en-GB and en-CA.

UNIVARIATE ANALYSIS

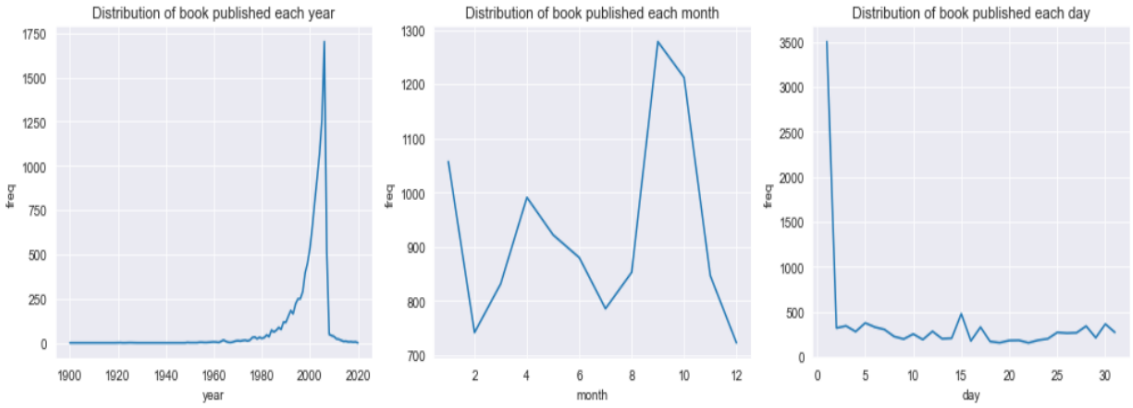
Here we analyse the distribution of each and every feature.




Observation on this graph concludes that most of the books have average ratings that lie between 3.7 to 4.3 and which are quite smaller in number of pages.



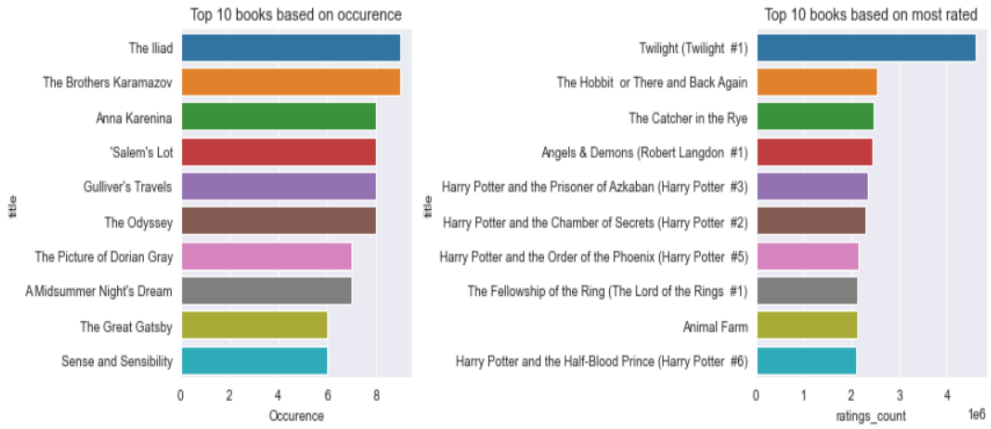

From the above graph, I observe that most of the authors write only one book. There is a large group of publishers who publishes 10 or 20 books. Most Reviews of a book has less than 1000 words and most books has only less than 100000 ratings.



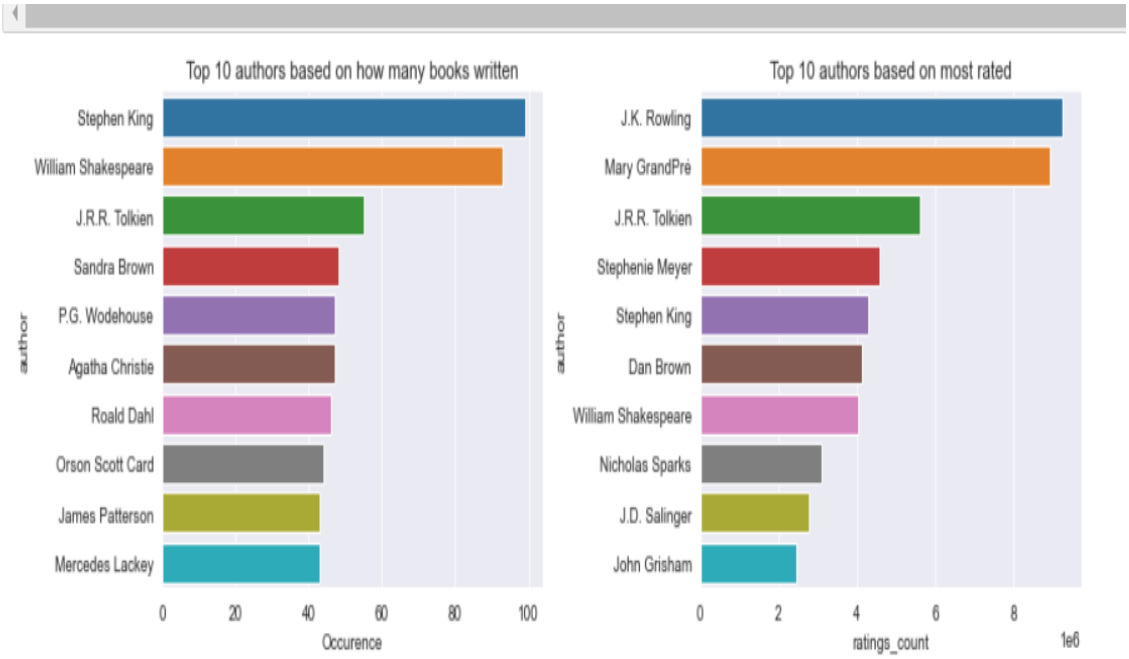

Most books in this goodreads review were published around 2000-2010. Most books are also published in the last quarter of the month which is September followed by April and January while books are less published in December.

Last, most books are published in the beginning of the month and followed by the middle of the month as the second most.

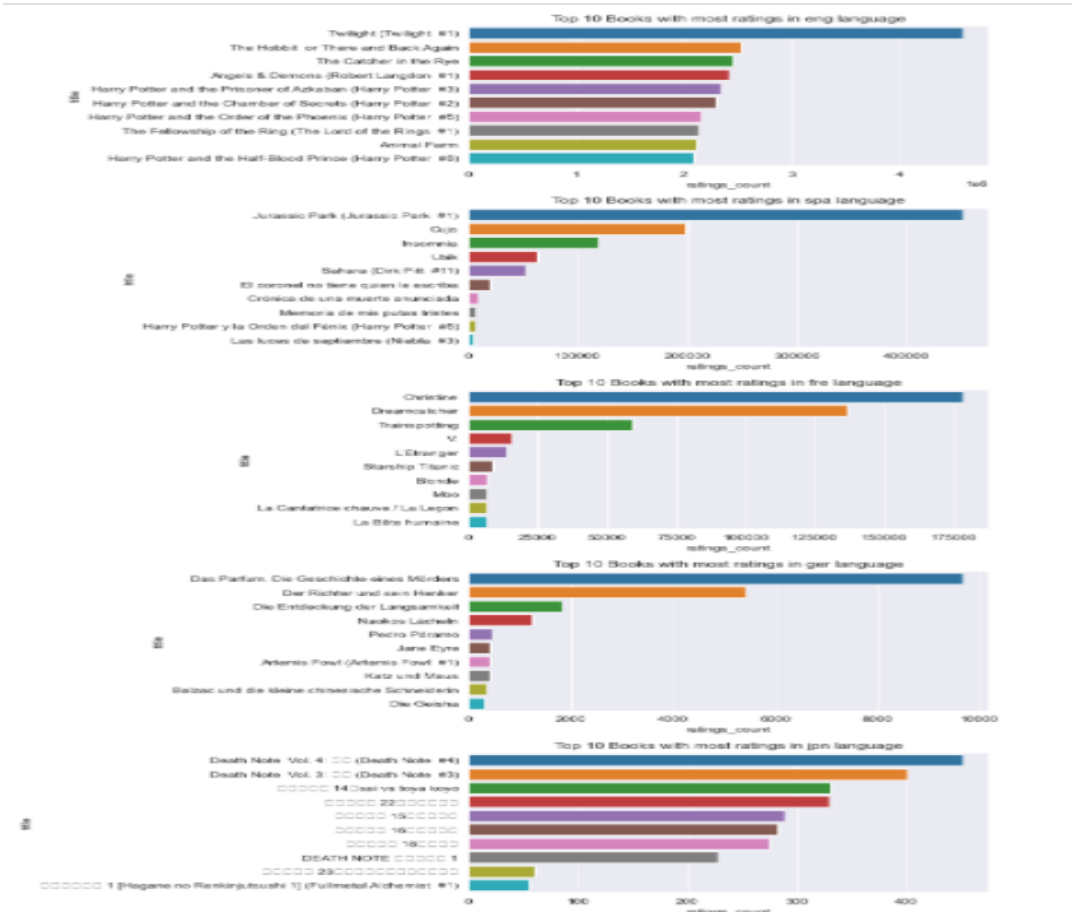
## TOP 10 BOOKS BASED ON RATINGS AND OCCURRENCE



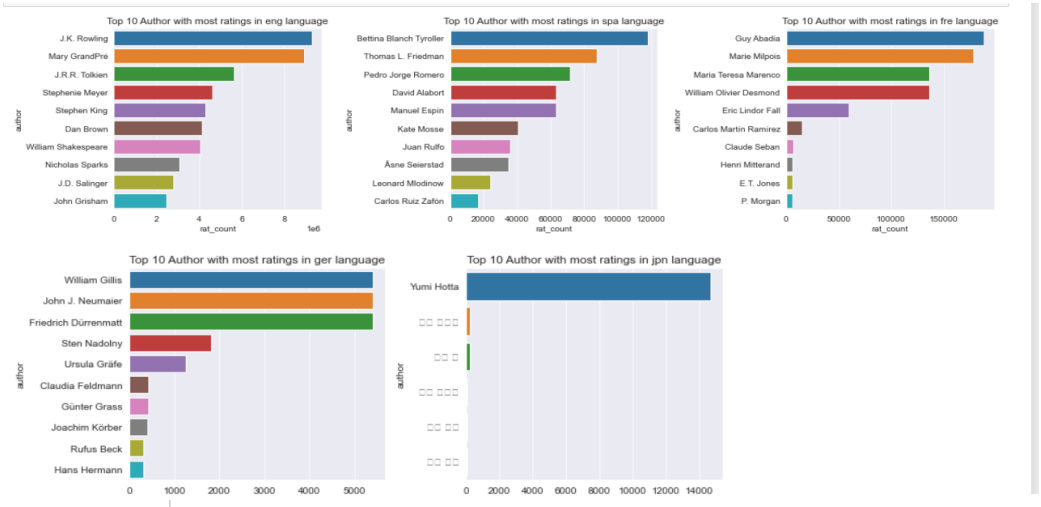

## TOP 10 AUTHORS BASED ON OCCURENCE AND RATINGS



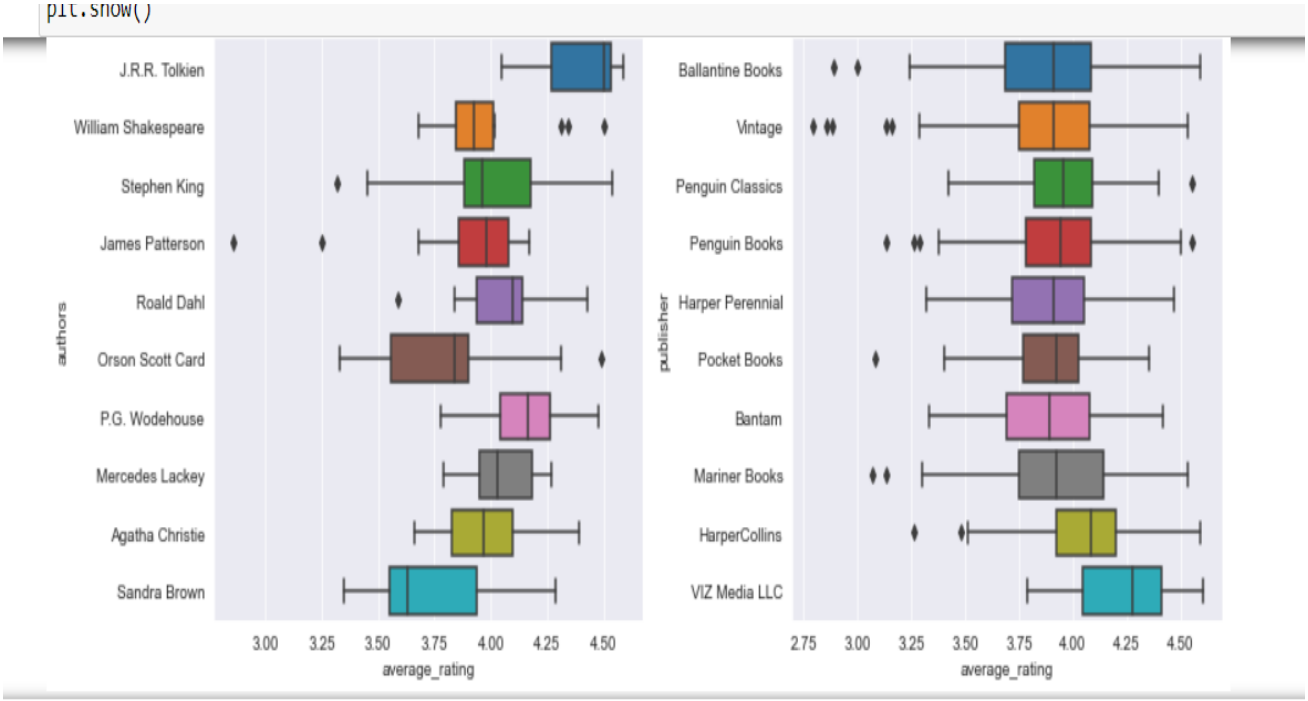
## TOP 10 TITLED BOOKS OF EACH LANGUAGE



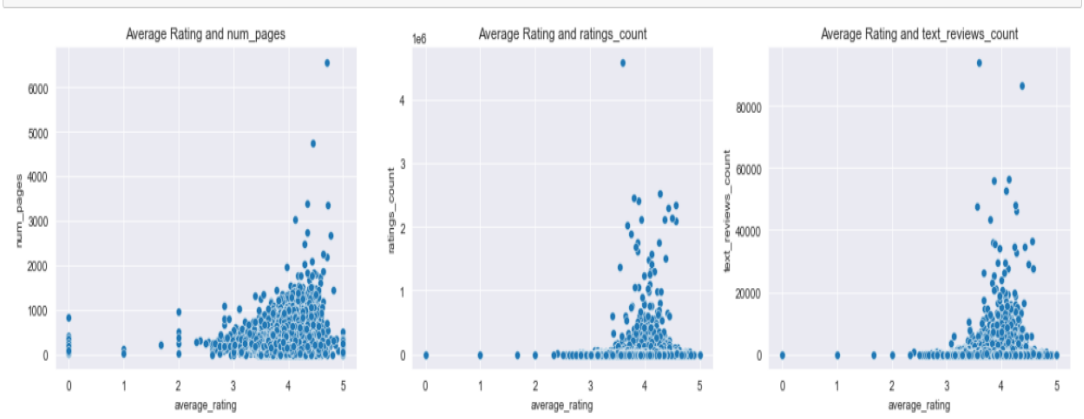

## TOP 10 AUTHORS BASED ON RATINGS IN TOP 5 LANGUAGES



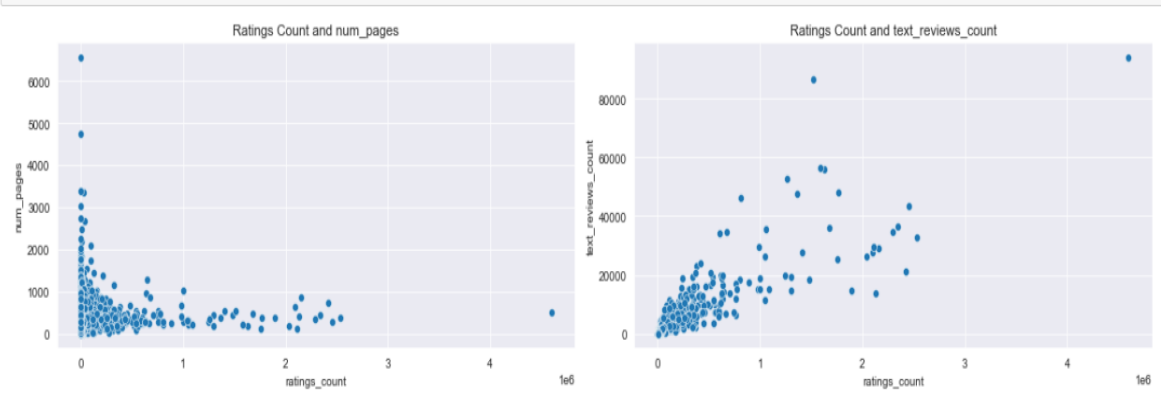
## AVERAGE RATING DISTRIBUTION OF TOP 10 AUTHORS AND PUBLISHERS




## CORRELATION ANALYSIS




Seems like there is no correlation between average_rating and num_pages, ratings_count and text_reviews_count




There is a positive correlation between text_reviews count and ratings_count but it's hard to tell about num_pages and ratings_count

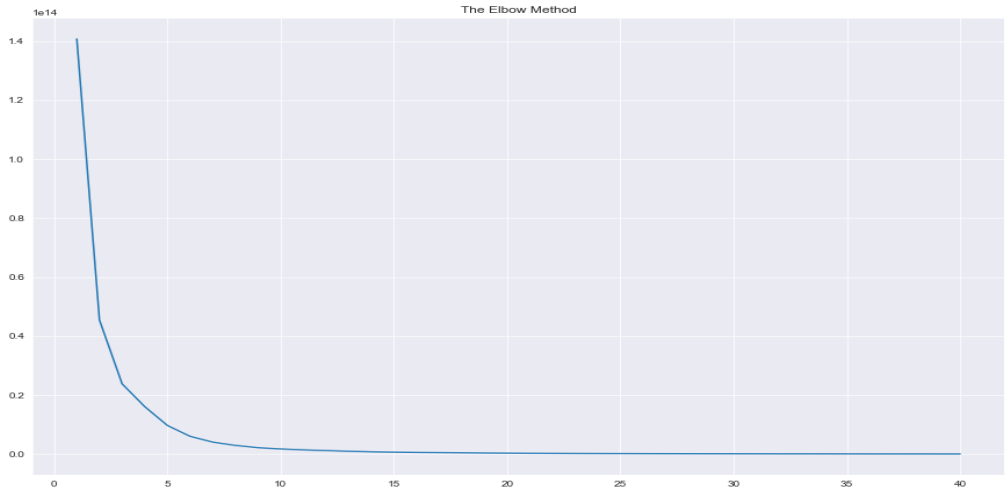
## WORD CLOUD


## RECOMMENDATION SYSTEM USING COSINE SIMILARITY

```
In [47]: recommend(2)

Out[47]: 4415     Harry Potter and the Chamber of Secrets (Harry...
         6              Harry Potter Collection (Harry Potter  #1-6)
         10675    Harry Potter and the Goblet of Fire (Harry Pot...
         1        Harry Potter and the Order of the Phoenix (Har...
         10674    Harry Potter and the Philosopher's Stone (Harr...
         Name: title, dtype: object
```

## ELBOW CURVE FOR K-MEANS CLUSTERING




## RECOMMENDATION SYSTEM USING K-MEANS CLUSTERING

In [66]:
```
def book_recommendation_engine(book_name):
    book_list_name = []
    book_id = book_data[book_data['title'] == book_name].index
    book_id = book_id[0]
#     print('book_id', book_id)
    for newid in idlist[book_id]:
#         print(newid)
        book_list_name.append(book_data.loc[newid].title)
#         print(new_data.loc[newid].title)
    return book_list_name
```

In [67]:
```
book_list_name = book_recommendation_engine('The Da Vinci Code (Robert Langdon  #2)')
book_list_name
```

Out[67]:
```
['The Da Vinci Code (Robert Langdon  #2)',
 'The Alchemist',
 'Of Mice and Men',
 'Romeo and Juliet',
 'Lord of the Flies',
 'Eat  Pray  Love']
```