A PROJECT ON BOOKS REVIEW

BY THE GUIDANCE OF PROF. SUDEEP MALLICK

[RKMVERI -MSc. BDA]

Abstract

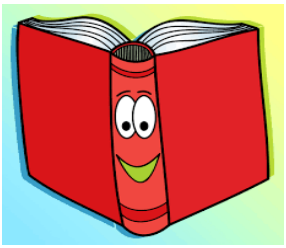
A VISUALISATION OF DIFFERENT FACTORS AFFECTING RATING OF A BOOK

MAHENDRA NANDI

mahendranandi.rkma@gmail.com

Visualizing Book Review Dataset

The dataset consists 13 columns and 11127 rows. Our primary goal is to visualize all the dependencies among them and finding out the key factors for a good review of the books. A good review in the sense that it may have a higher average rating, higher number of reviews.



THE DATASET

IT has 12 columns :

['bookID', 'title', 'authors', 'average rating', 'isbn', 'isbn13','language code', 'no of pages', 'ratings count', 'text reviews count','publication date', 'publisher', 'rating', 'published year']

TARGET COLUMN: 'average rating'

Dependencies :

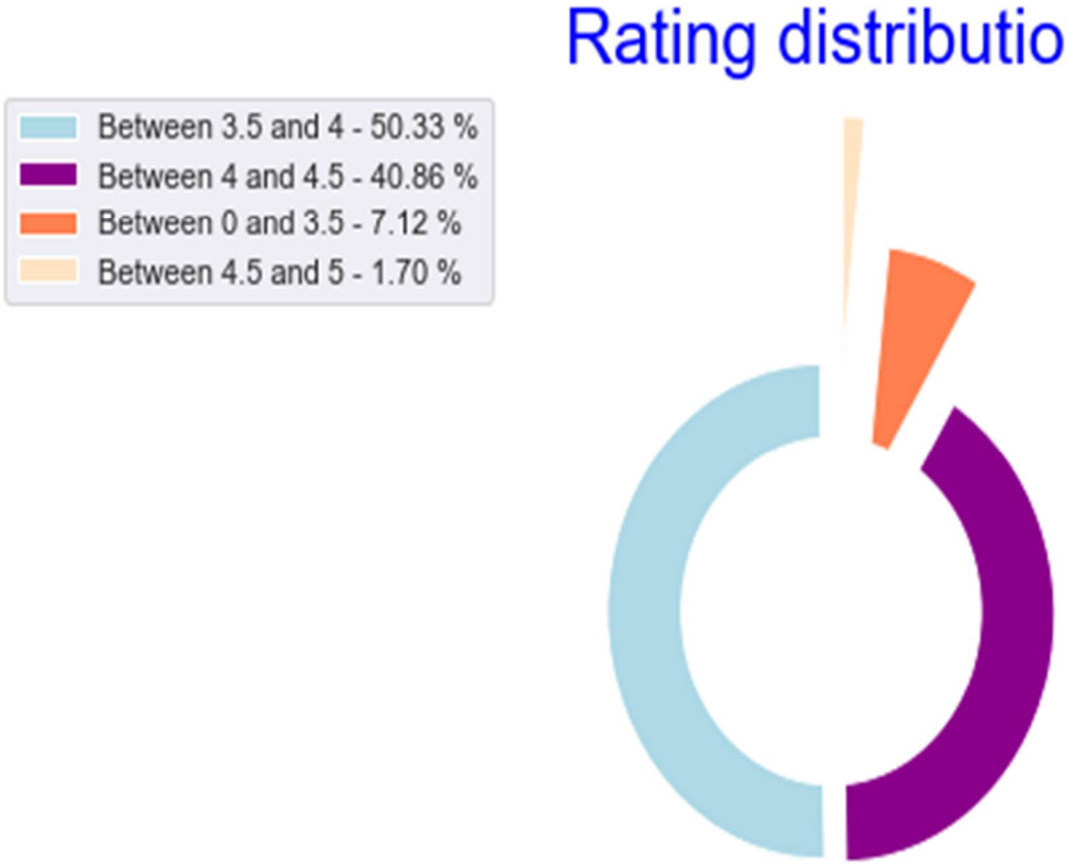
1. 'title' [10352 distinct values]
2. 'authors' [6643 distinct values]
3. 'language code' [27 distinct values] <- .....categorical variable
4. 'no of pages' [997 distinct values] <- .....continuous variable
5. 'ratings count' [5294 distinct values] <- .....continuous variable
6. 'text reviews count'[1822 distinct values] <- .....continuous variable
7. 'publication date' [87 distinct values] <- .....continuous variable
8. 'publisher' [2292 distinct values] <- .....categorical variable
9. 'rating' [4 distinct values] <- .....categorical variable
10. 'published year' [7 distinct values] <- .....categorical variable

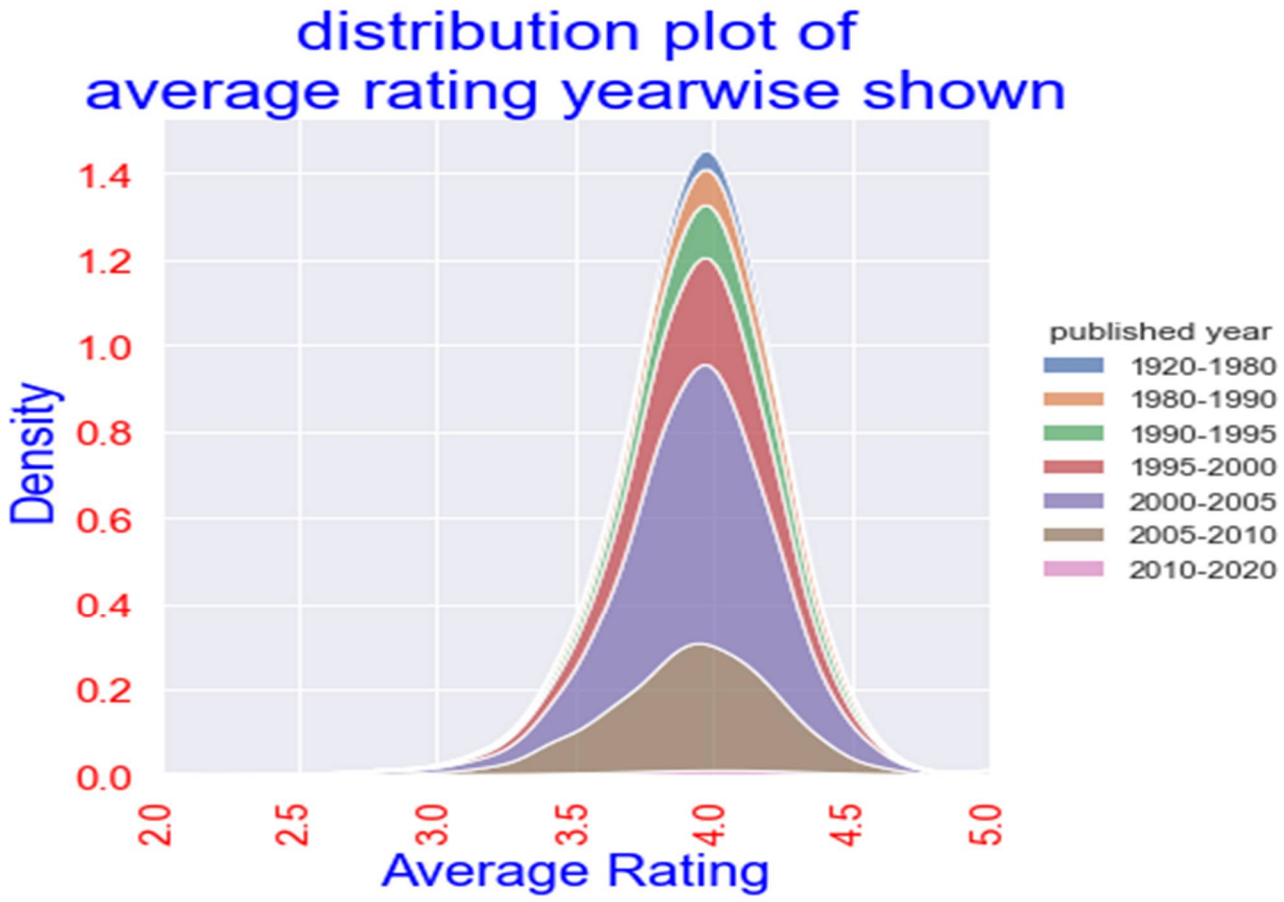
OVERVIEW OF THE IMPORTANT FACTORS:

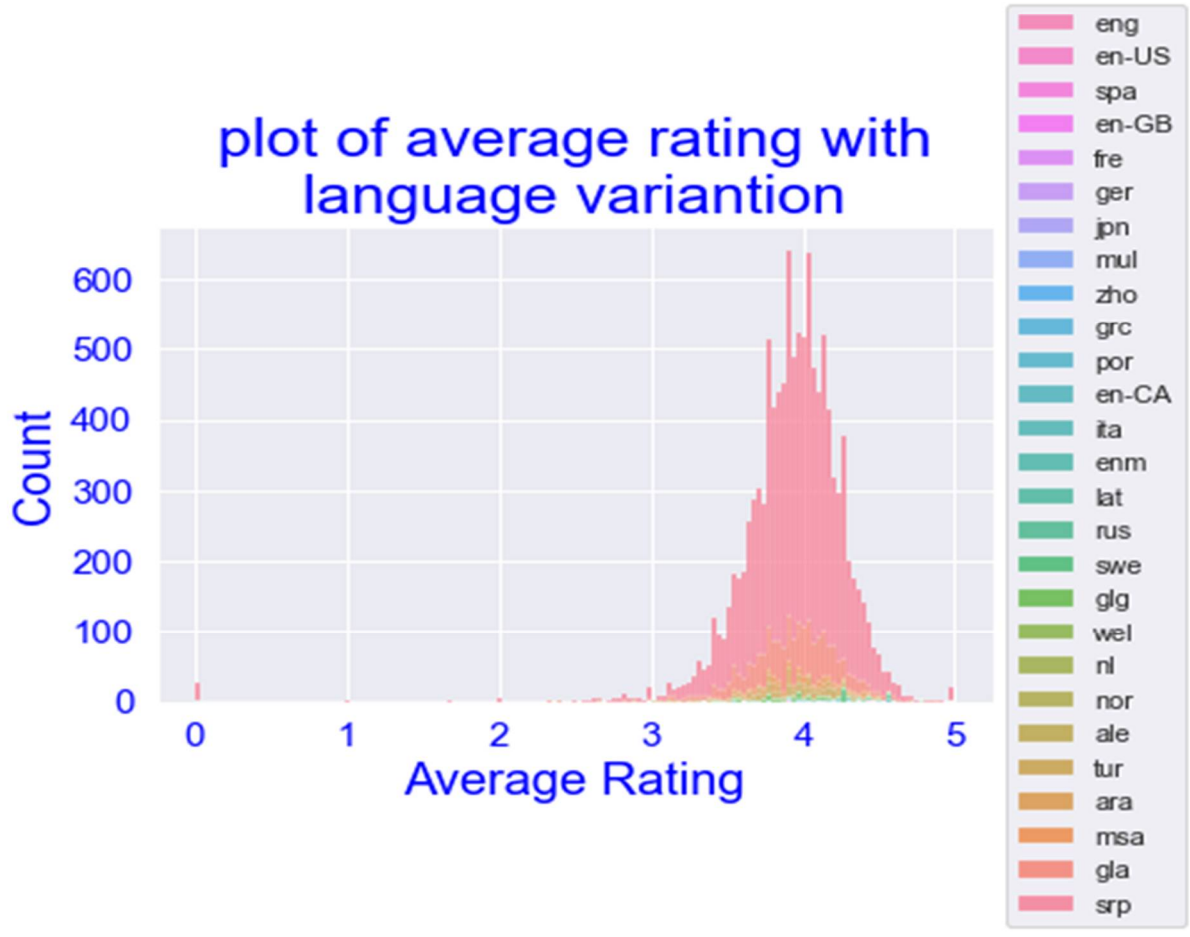
(Visualization of Univariate Data_column)::

[a] univariate continuous:

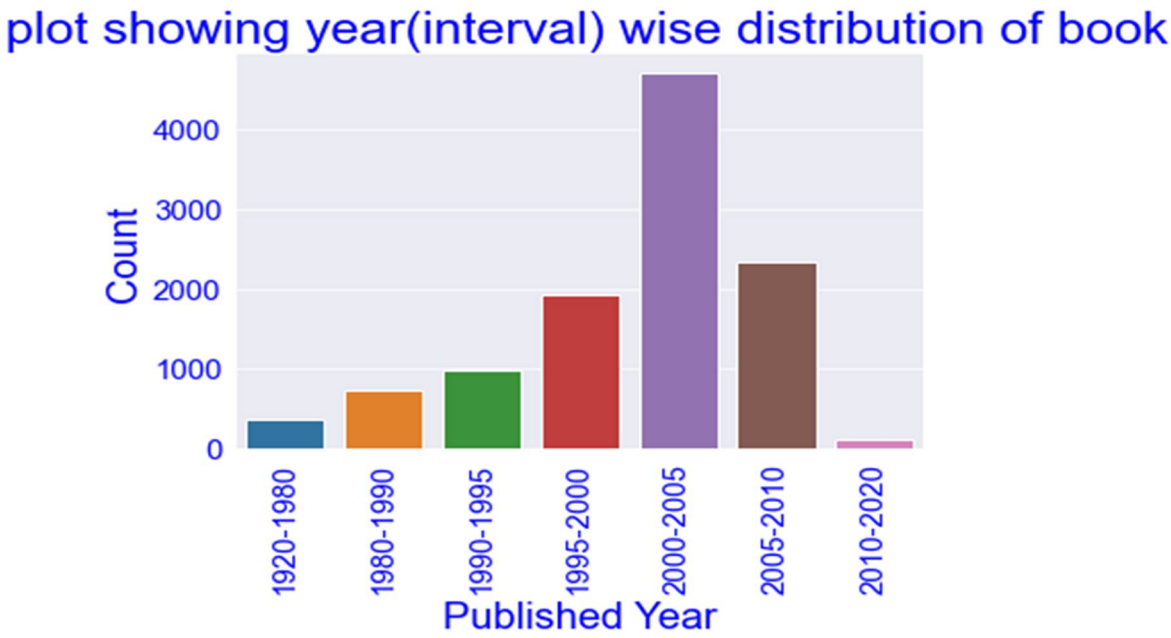
(1) showing the distribution of average rating (the target column),



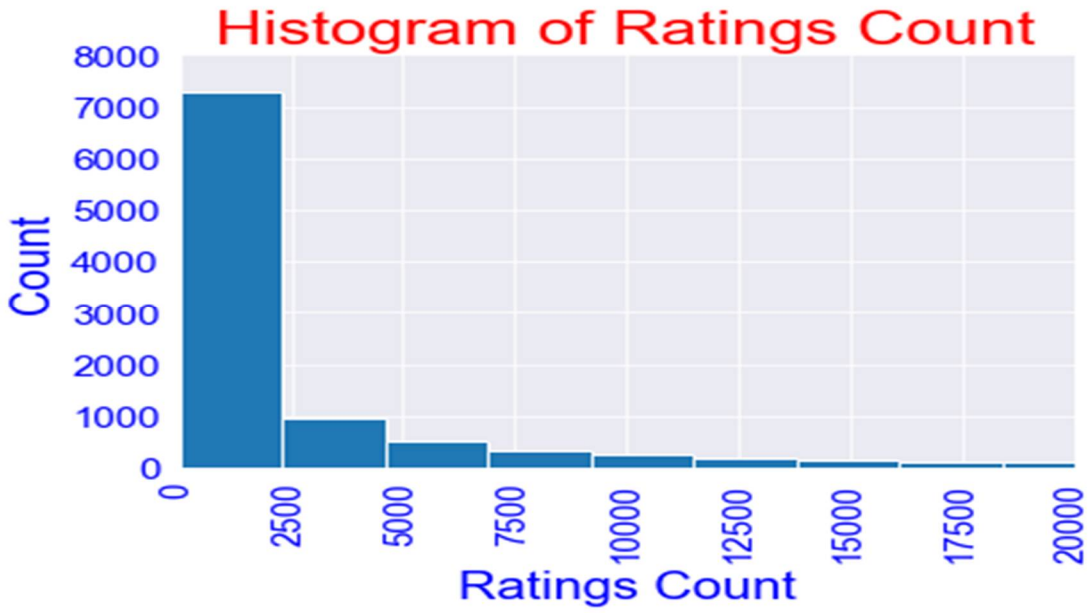

distribution plot of
average rating yearwise shown
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Average Rating
published year
1920-1980
1980-1990
1990-1995
1995-2000
2000-2005
2005-2010
2010-2020


plot of average rating with
language variantion
Count
600
500
400
300
200
100
0
0
1
2
3
4
5
Average Rating
eng
en-US
spa
en-GB
fre
ger
jpn
mul
zho
grc
por
en-CA
ita
enm
lat
rus
swe
glg
wel
nl
nor
ale
tur
ara
msa
gla
srp

(2) showing the distribution of book with time in 7 intervals.

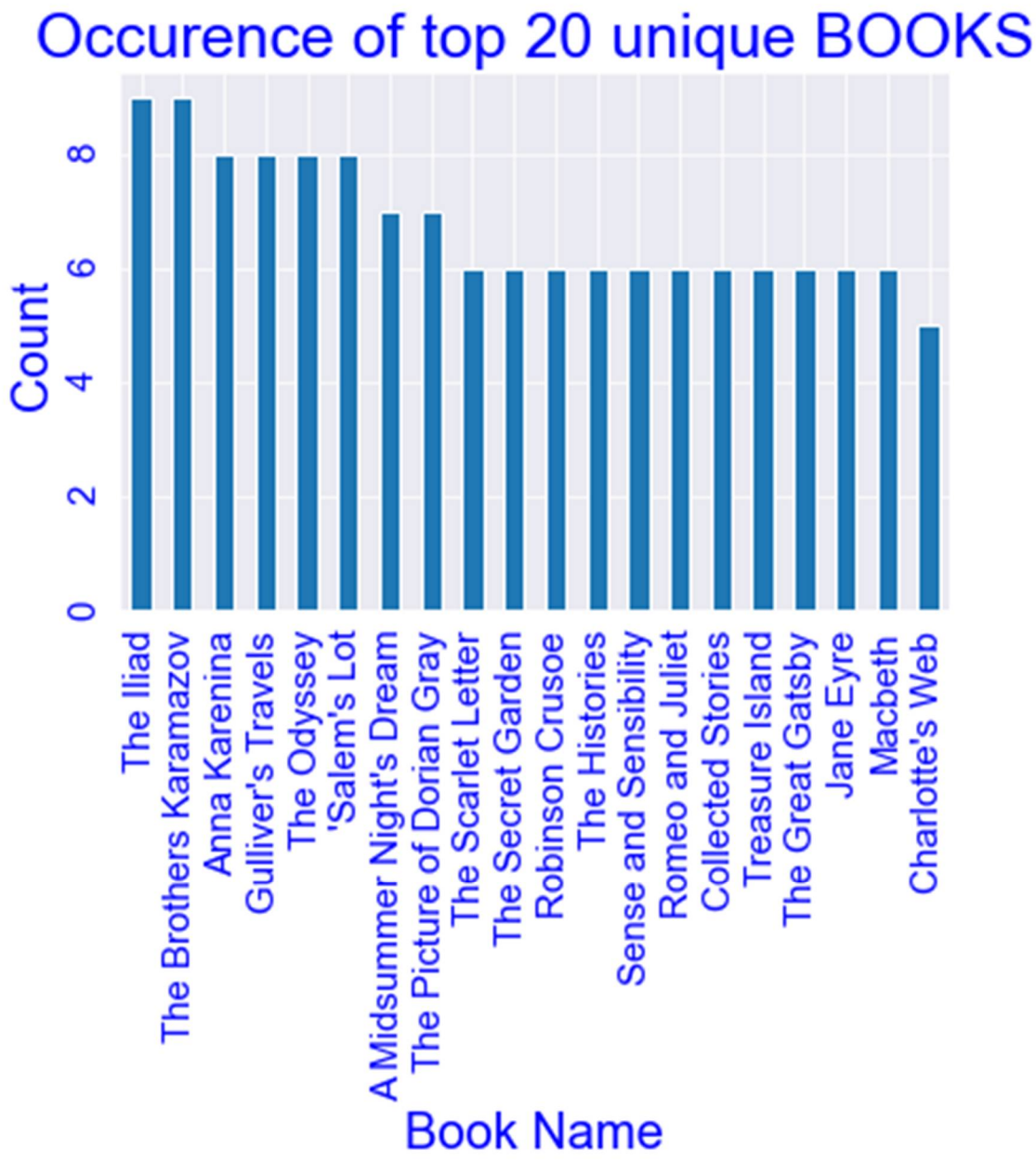


(3) Showing the distribution of Ratings count in histogram plot

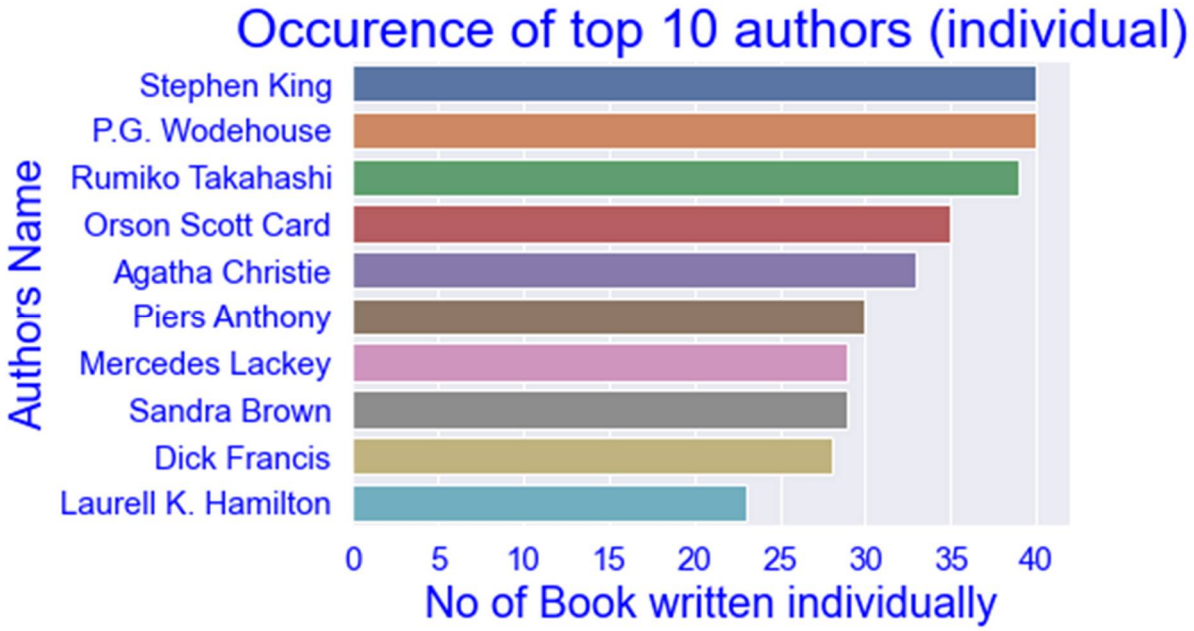


[b] univariate discrete:

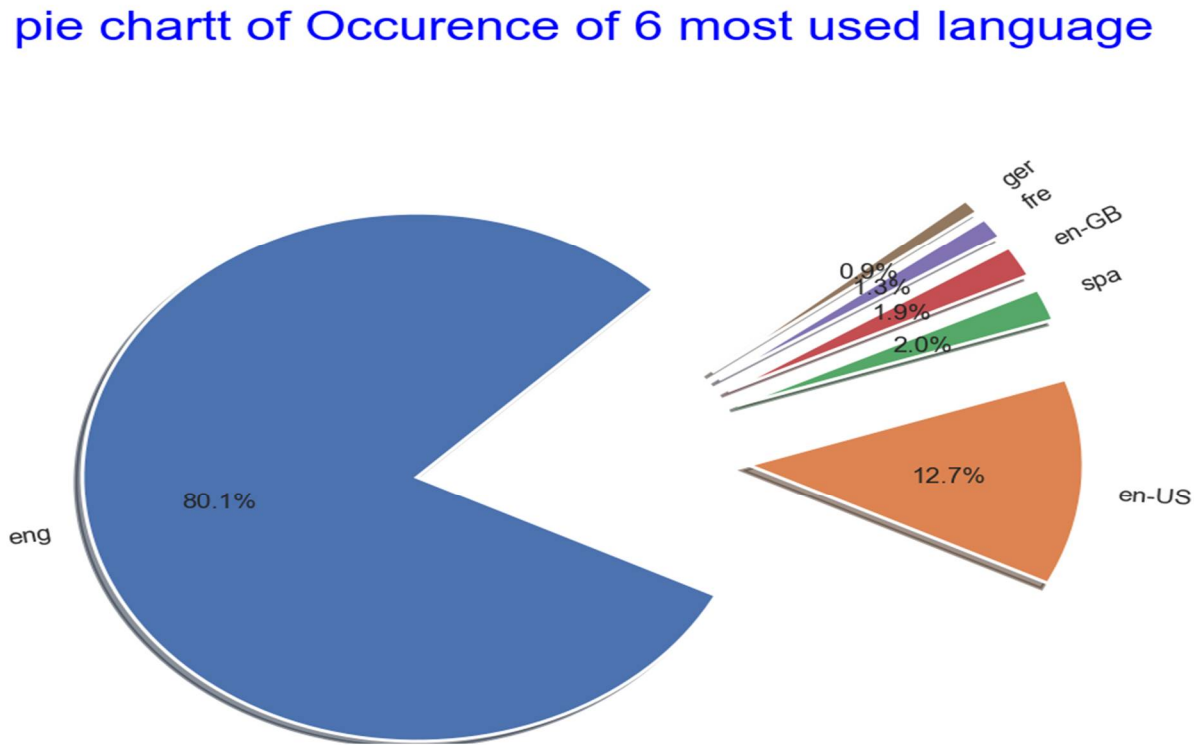
(4) Showing most occurred 20 books which are unique ,



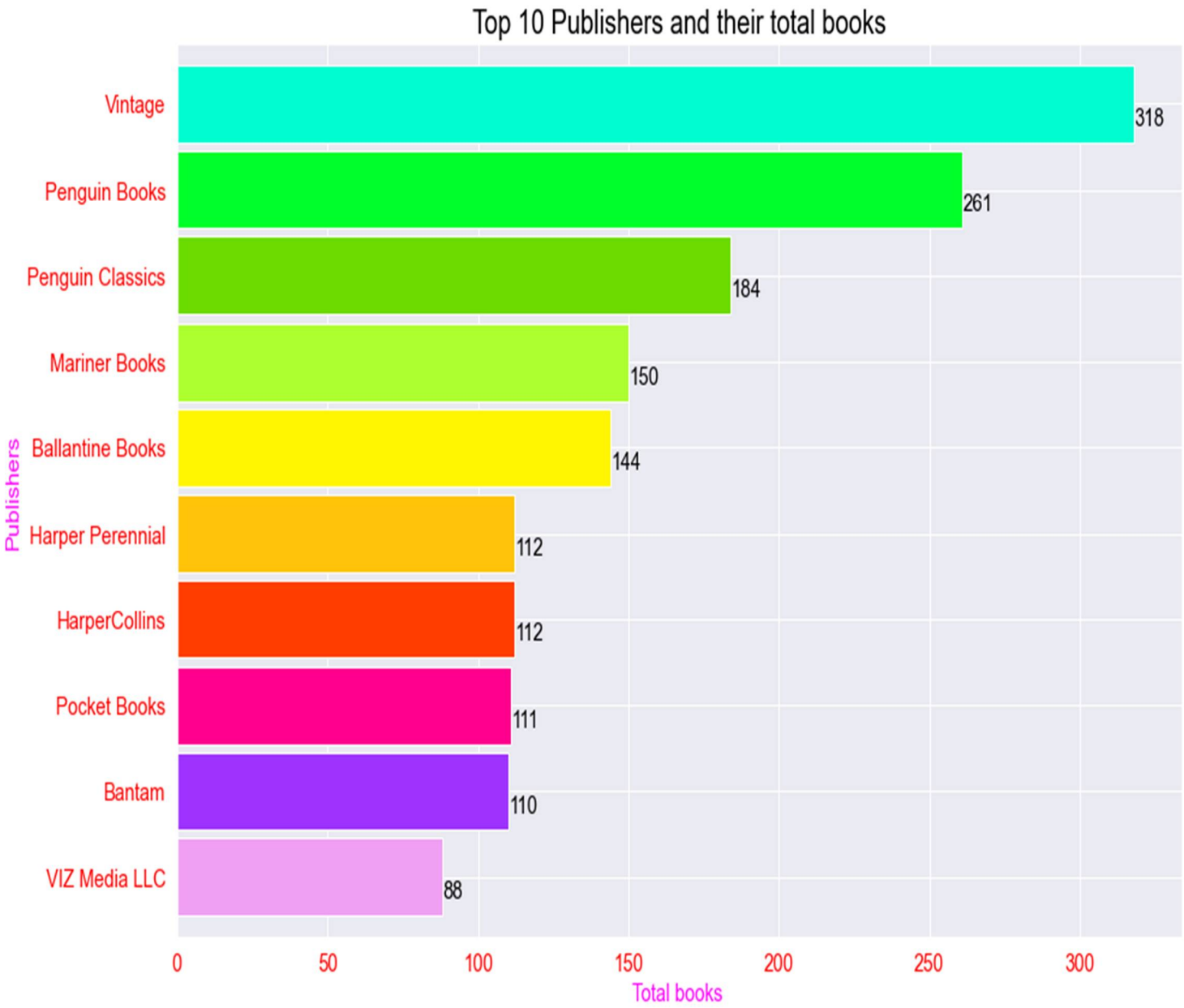
(5) name and no of books written of 10 such authors who have written maximum books by his/her own,



(6) Pie chart of 6 mostly used language, and among then more than 80 percent book is written in “eng”.



(7) Top 10 publishers published most of the books are shown below ,



(Visualization of Bivariate Data_column)::

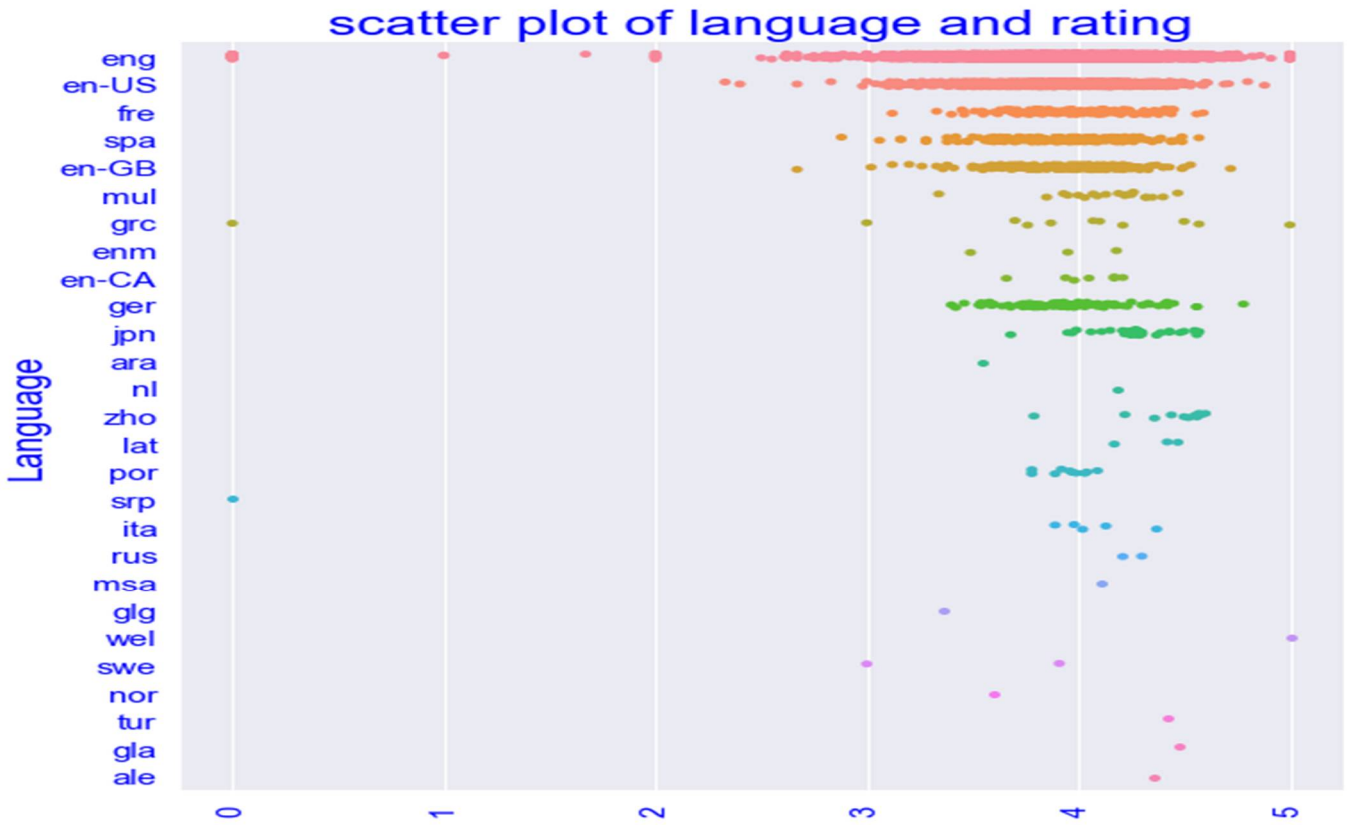
Till now we have seen the distribution of some of the dependent variables . Now we are going to follow the dependency or relation among dependent factors .

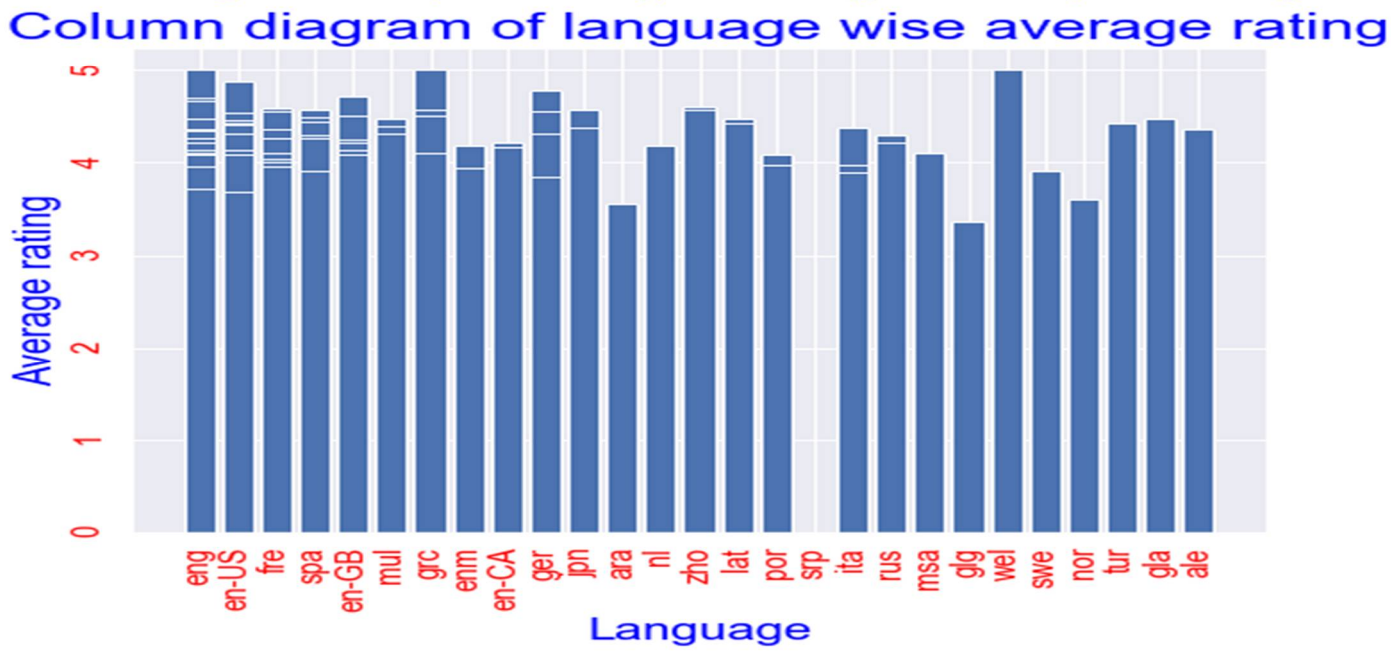
[a]. **bivariate var.(continuous vs continuous)**: the relation between two continuous variables are actually shown latter in “multivariate analysis “

(1) Showing distribution of Total No of Pages of the book having total page less than 10000 and having been written in maximum used 5 language.

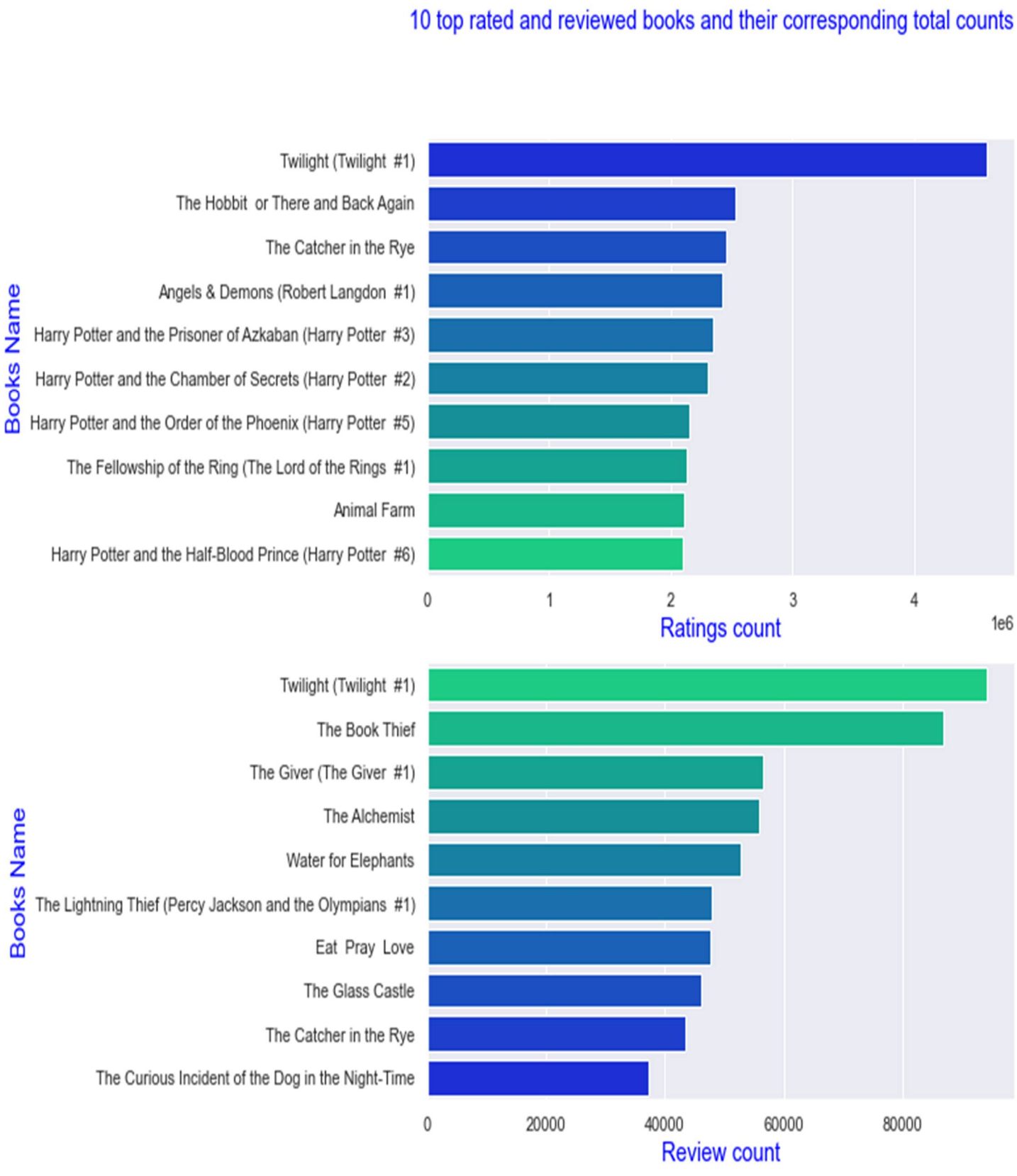
[b] bivariate var. (continuous vs discrete):

(2) showing scatter plot of average rating and the languages used to write the books, and then showing lang uage wise average rating.

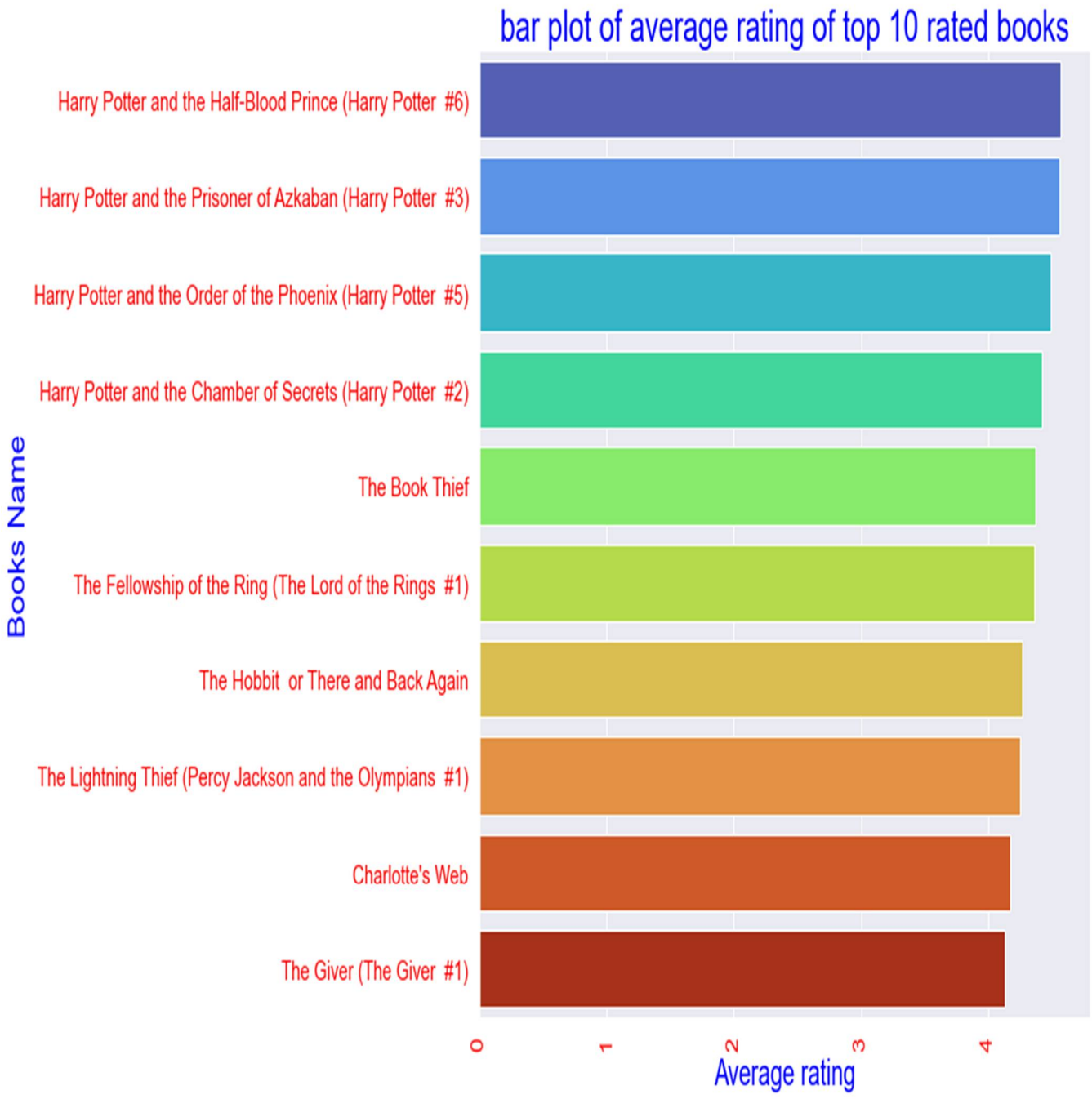




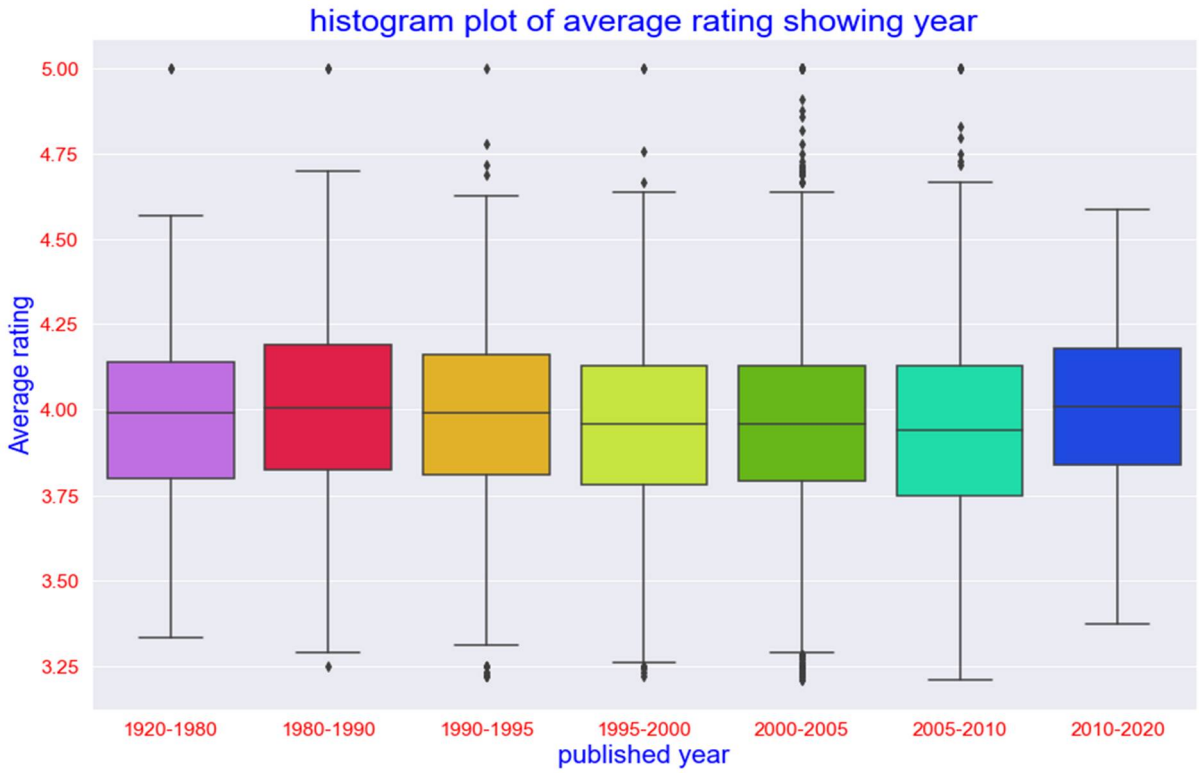
(3) Books with maximum count of rating and review and corresponding count is shown here.

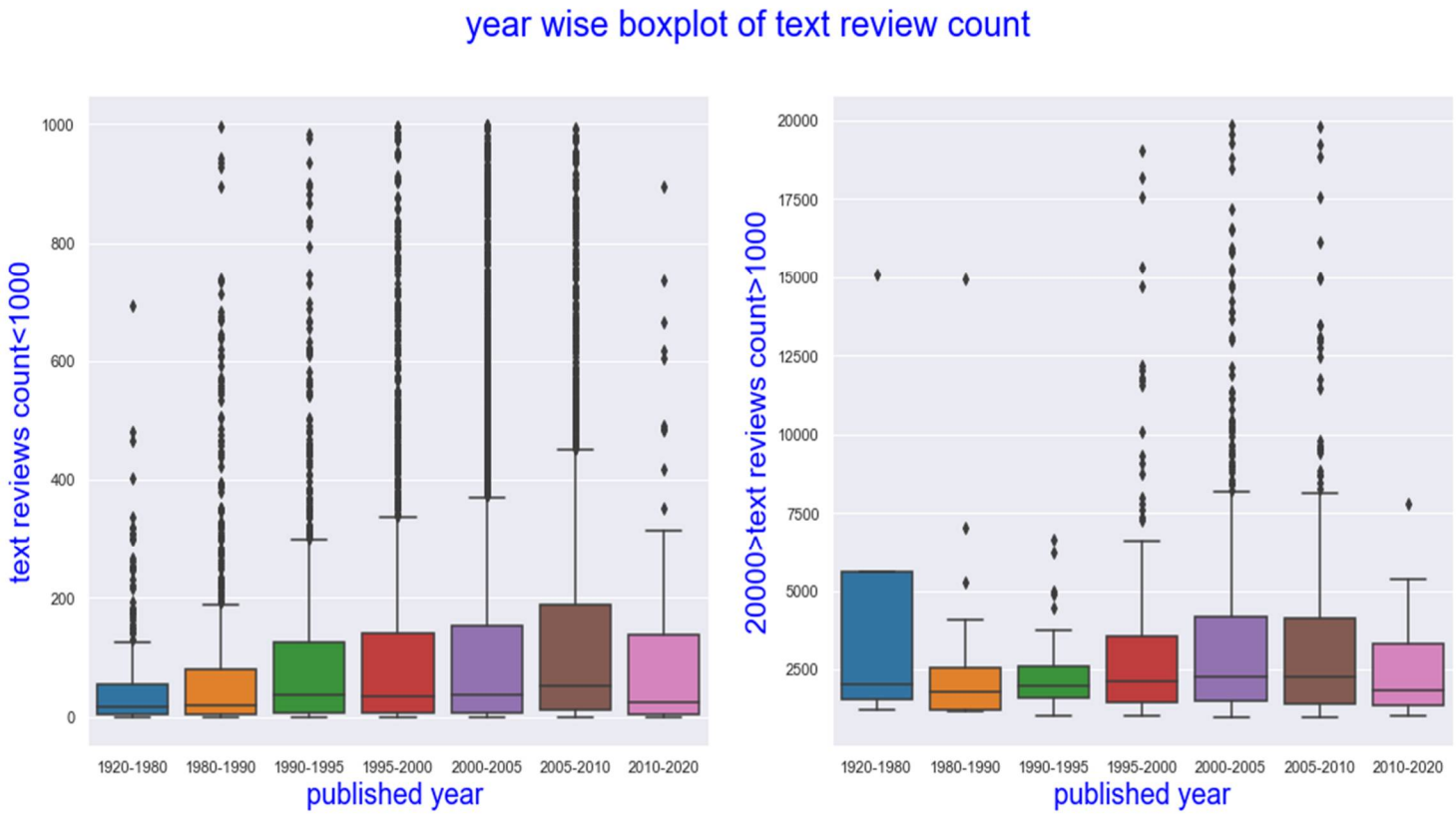


(4) books which have rated in top 10 are selected and a plot with corresponding average rating is shown. Here we can follow that average rating and total rating count is different



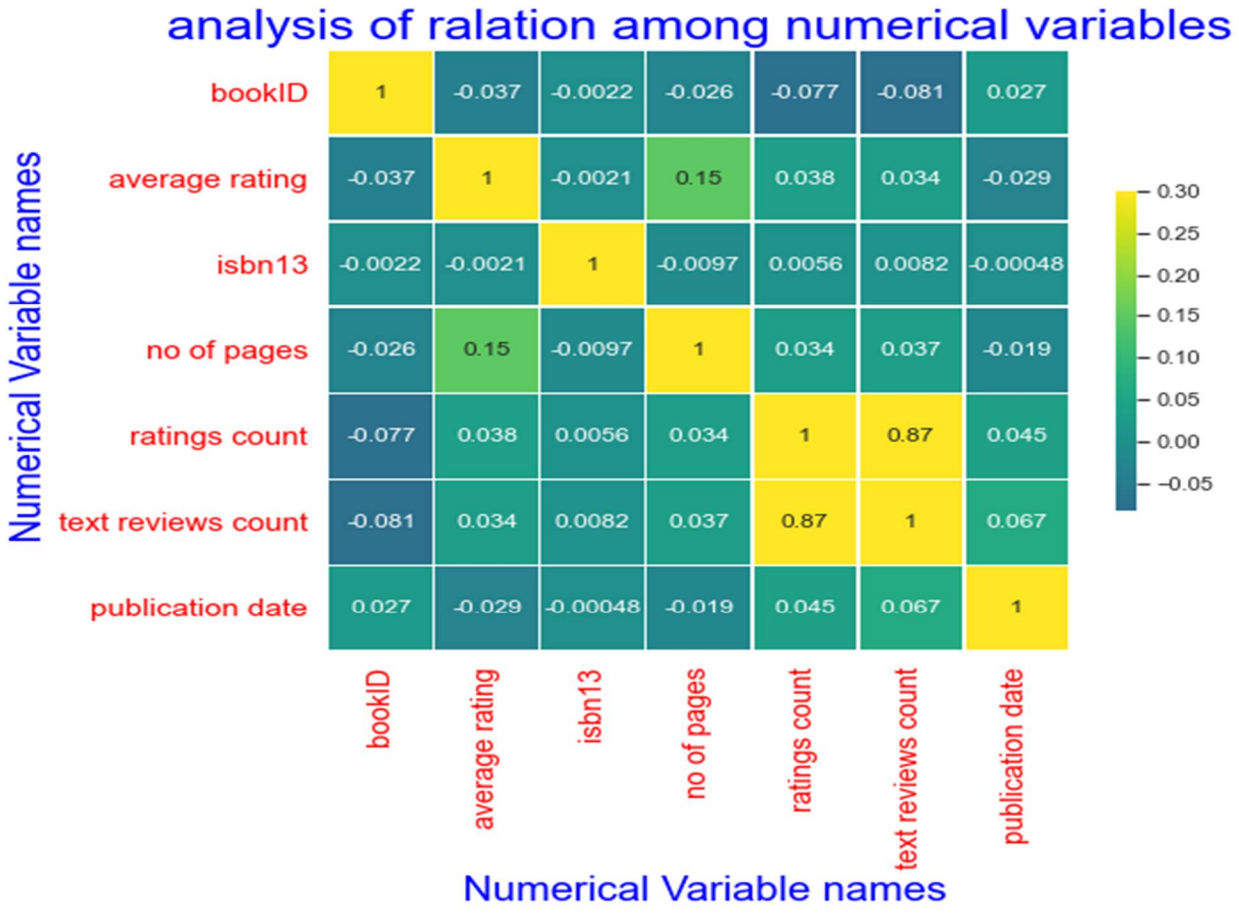
(5) we can now infer some insights one by one , here we have seen that a slight decreasing pattern is observed with time but after 2010 it is opposite to that .





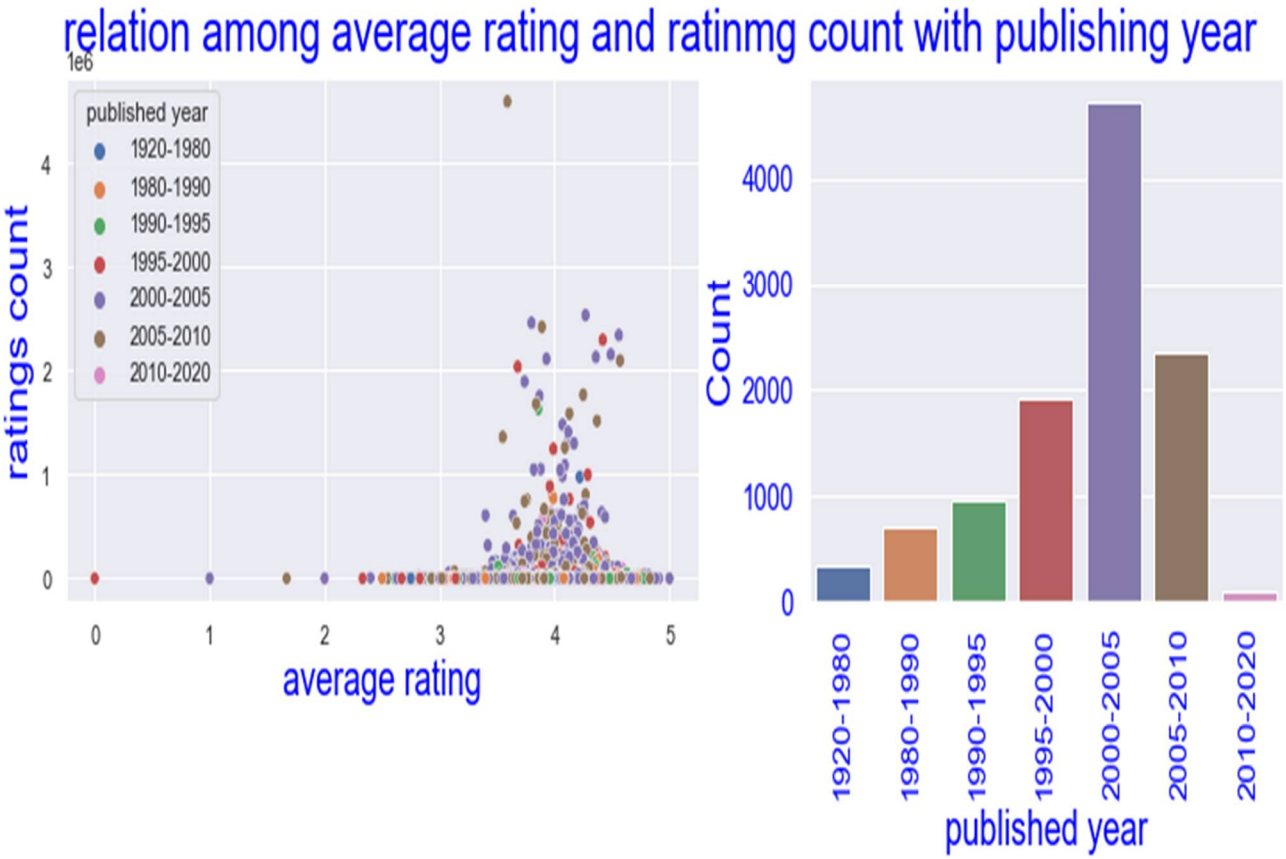
(6) Similarly in the plot above we can see text reviews has a increasing pattern with time but after 2010 it is being opposite

(7) And above all we have the correlation plot . we can see frem here that only rating count and reviw count have a strong correlation .



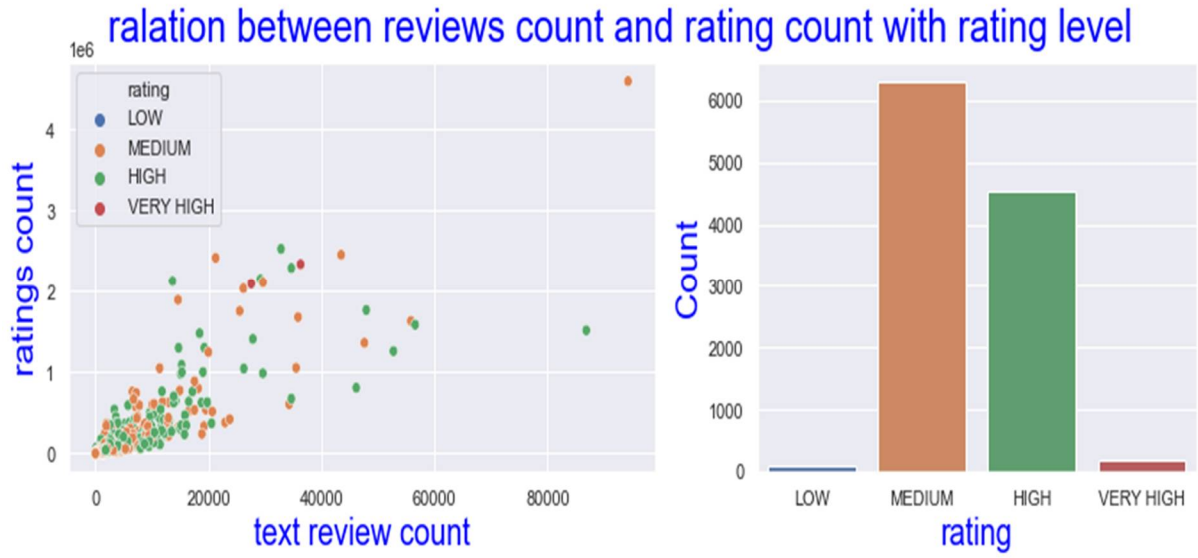
(Visualization of Multivariate relations):

(1)Till now we have seen the relations among different factors affecting the book rating , now we will gradually try to make some insight . Firstle, we can see that most of the data is concentrated in the rating level (3.5-4.5)

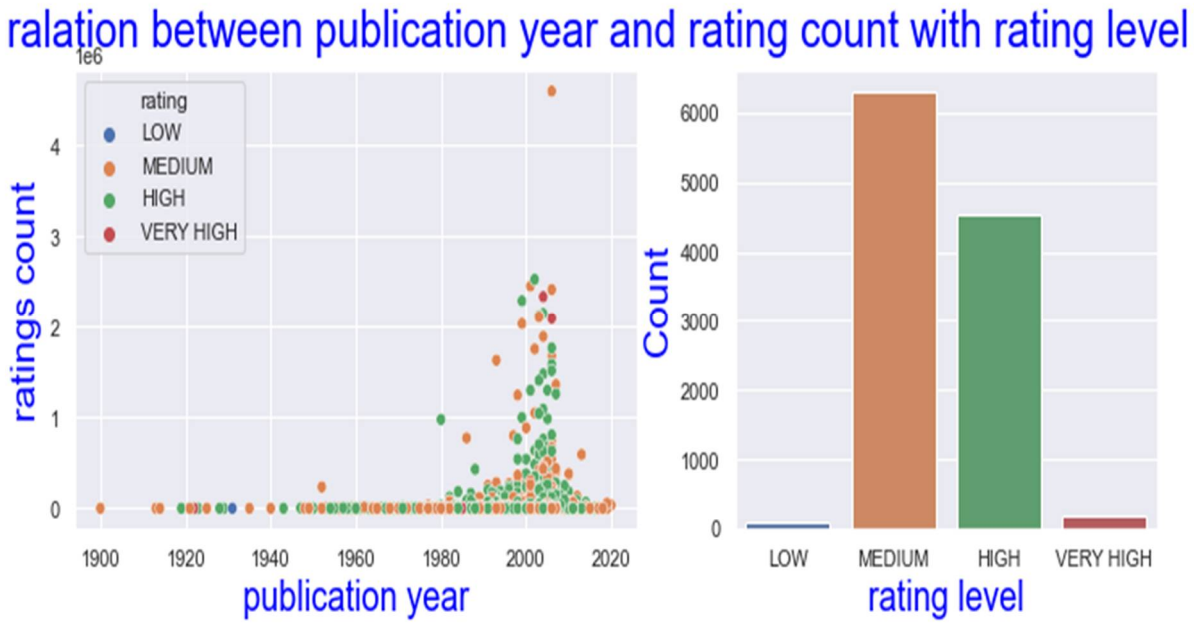


And in case of total pages, total rating count, total review counts the data is mostly concentrated at very low value.

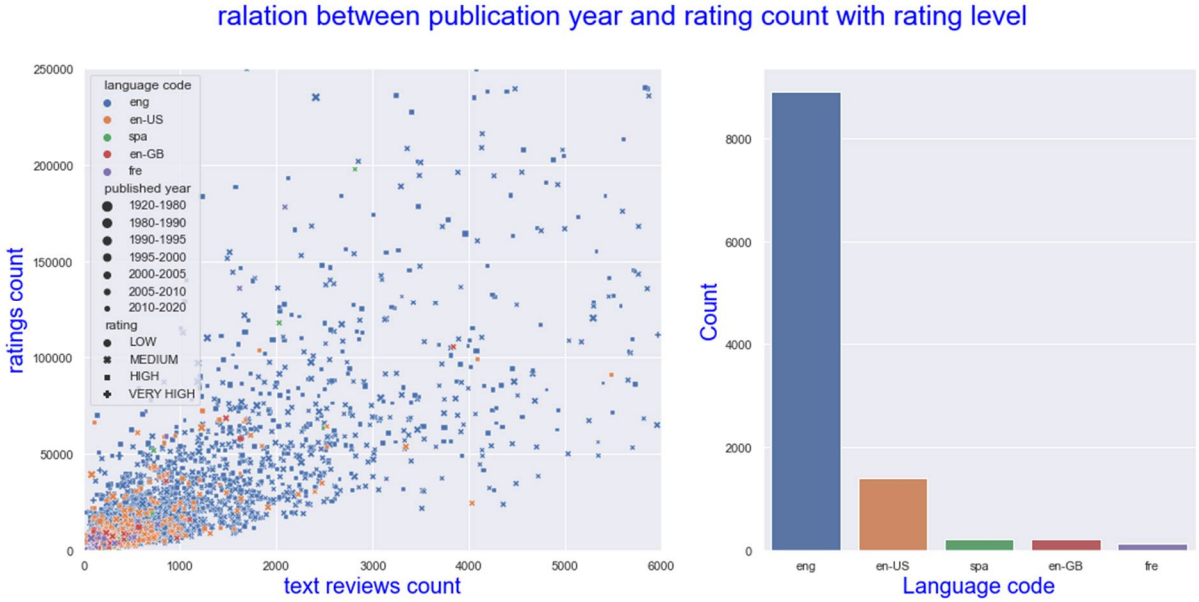
(2)only there is a positive and high correlation among total rating and review as shown below.



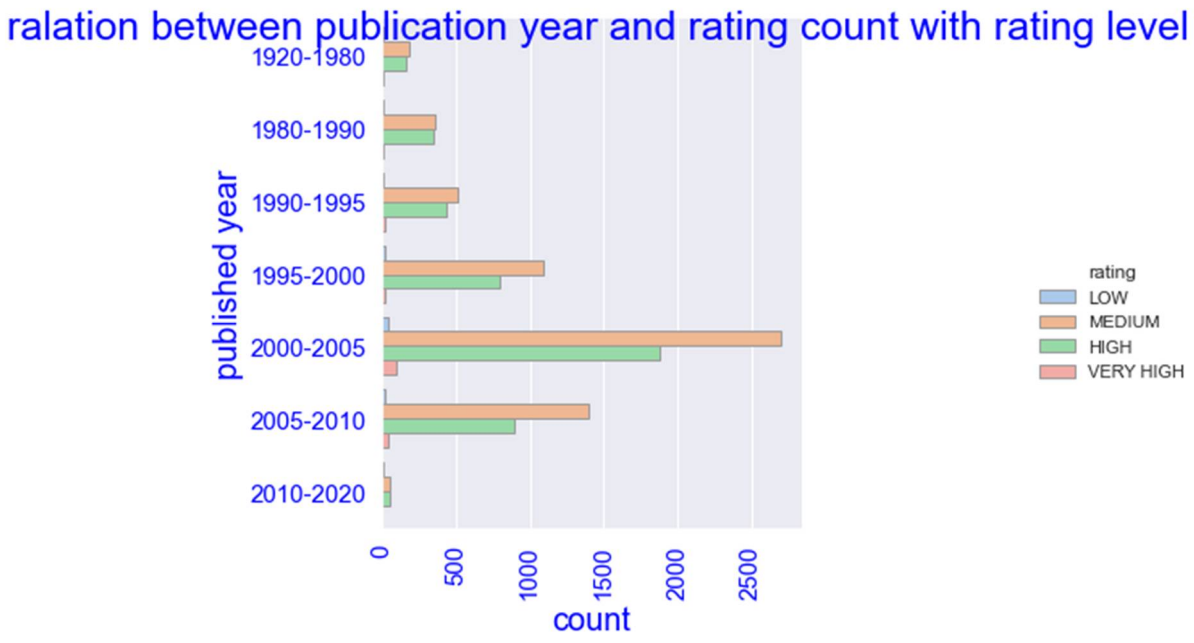
(3) we can see most of the books of the data frame has been published in between 2000 and 2010 , and also in this interval total rating is also high.



(4)As , mostly “emg” language is present here the interesting distribution is also in this language as shown

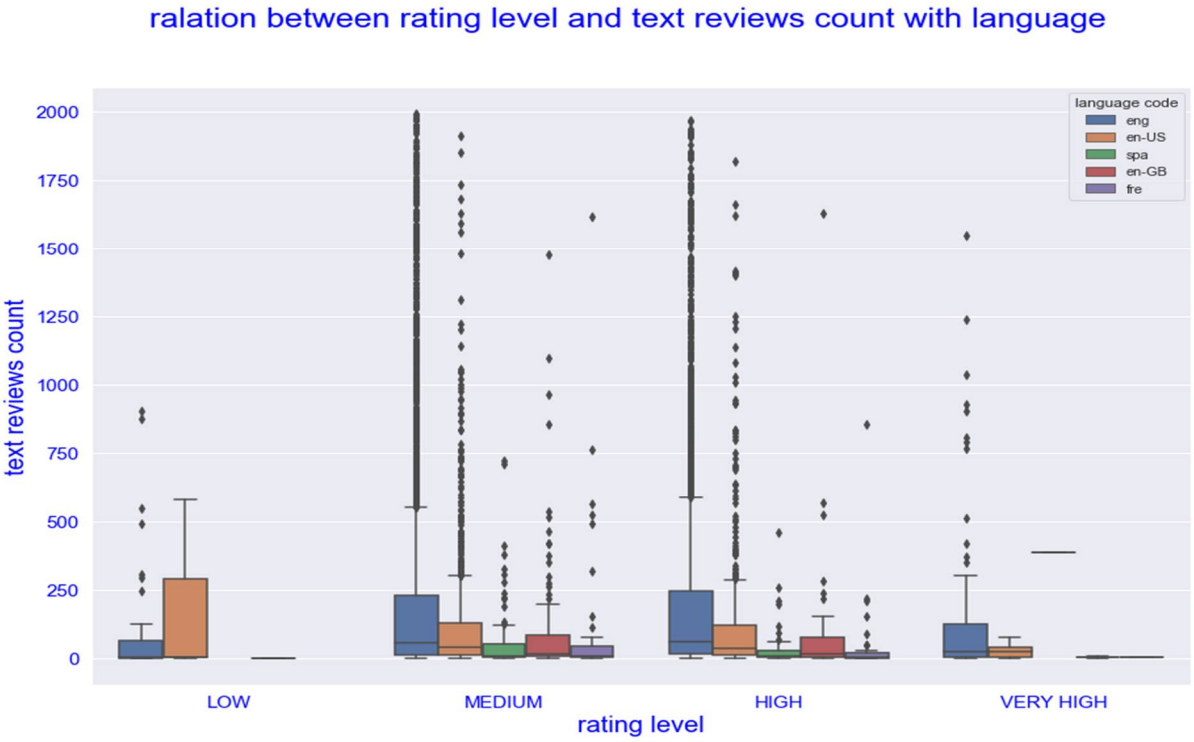


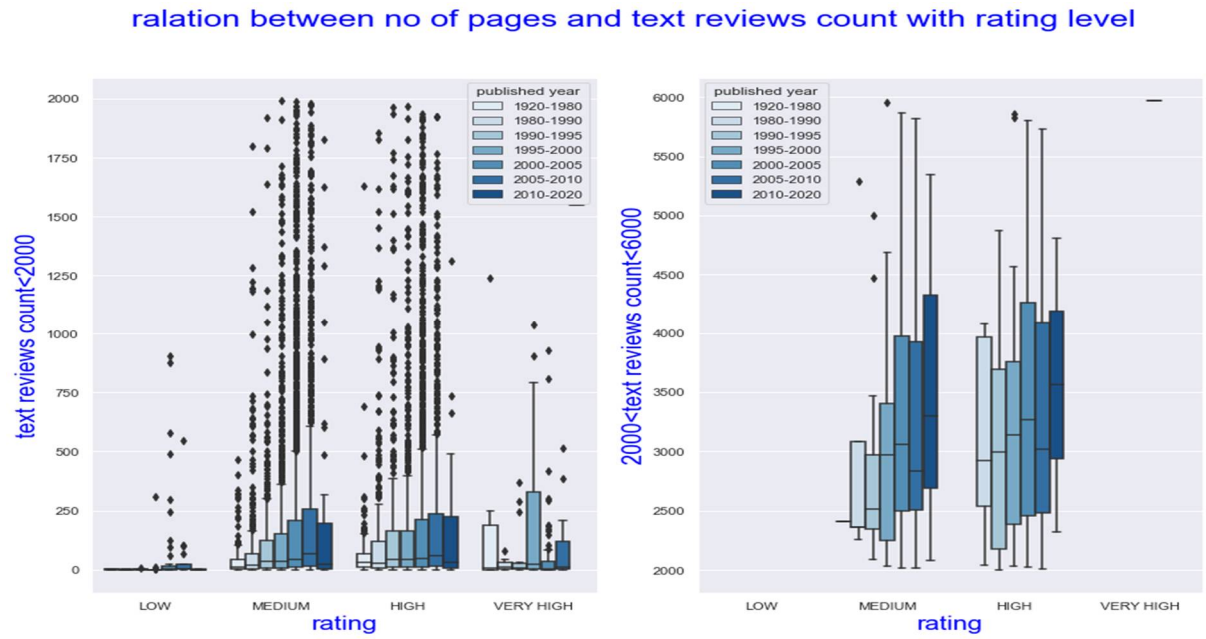
(5)We can see here the rating level named “ medium” is present in maximum through out the time period of publication .



(6) Though all the categories have same distribution (almost normal distribution) we can see the average rating is uniformly distributed.

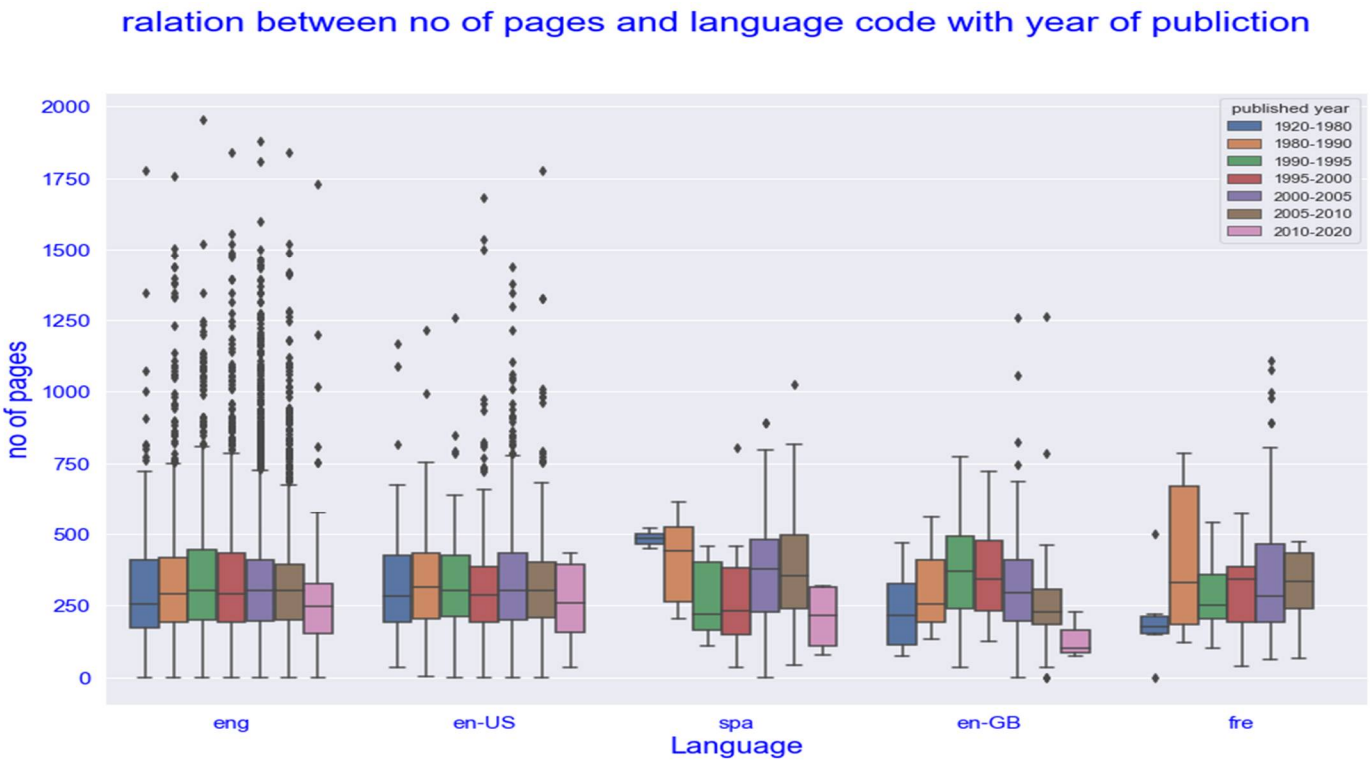
(7)There is a general trend of decreasing total review count as shown in figure with language of the book.



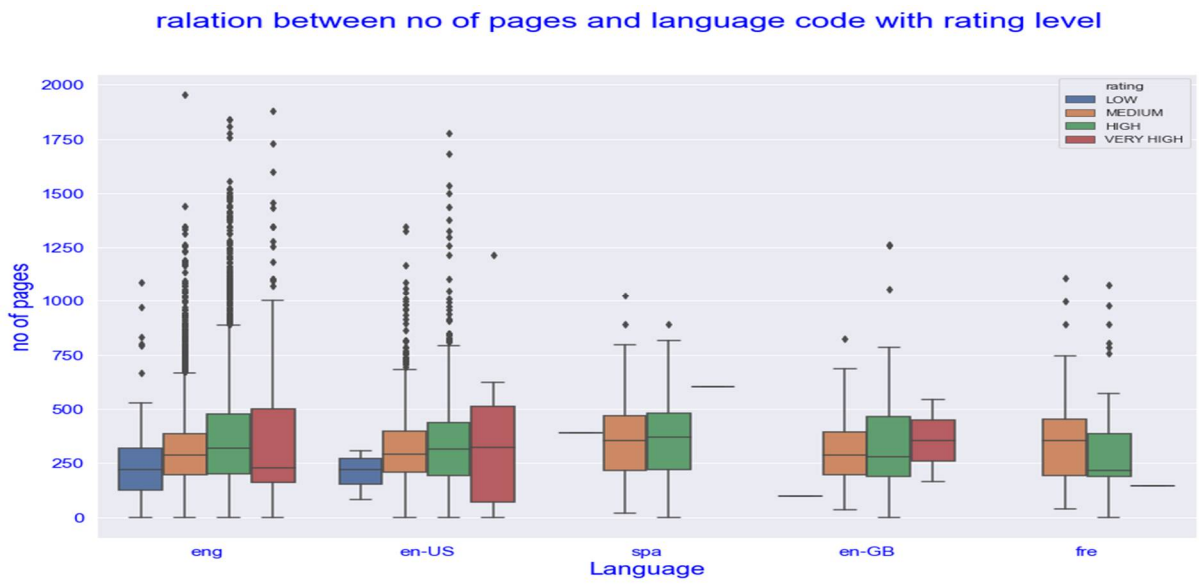


(8) But with time reviews is increasing shown in the above plot in all type of rating categories.

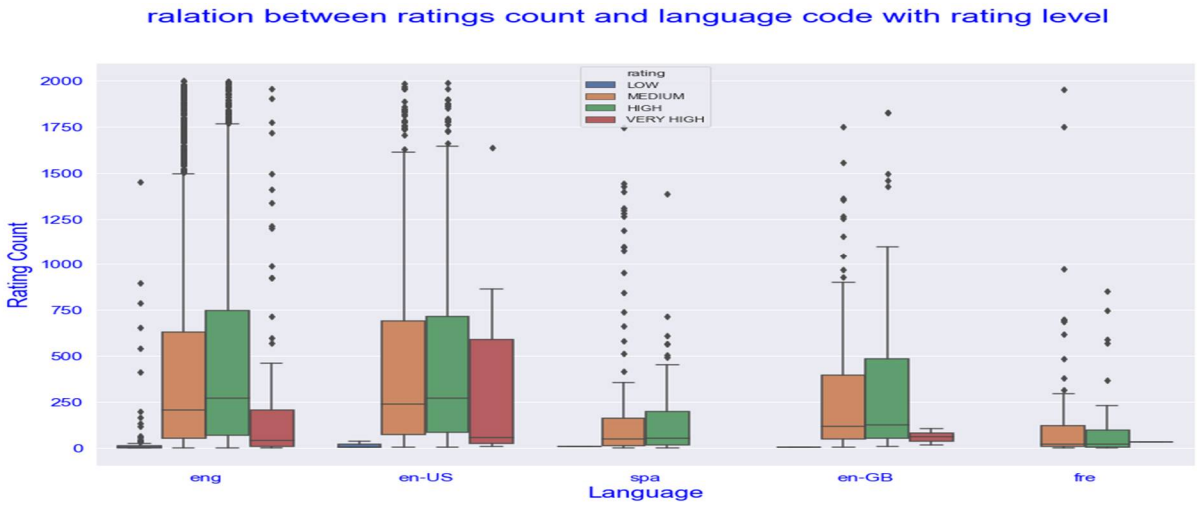
(9) Here is an interesting fact that with time the total no of pages of a book is going to increase till 2010 in 4 mostly used language , but after that year interval i.e, 2010 to 2021 it has a sudden decrease . and it is very obvious because of the modern technologies of using e-book and not liking to read more .



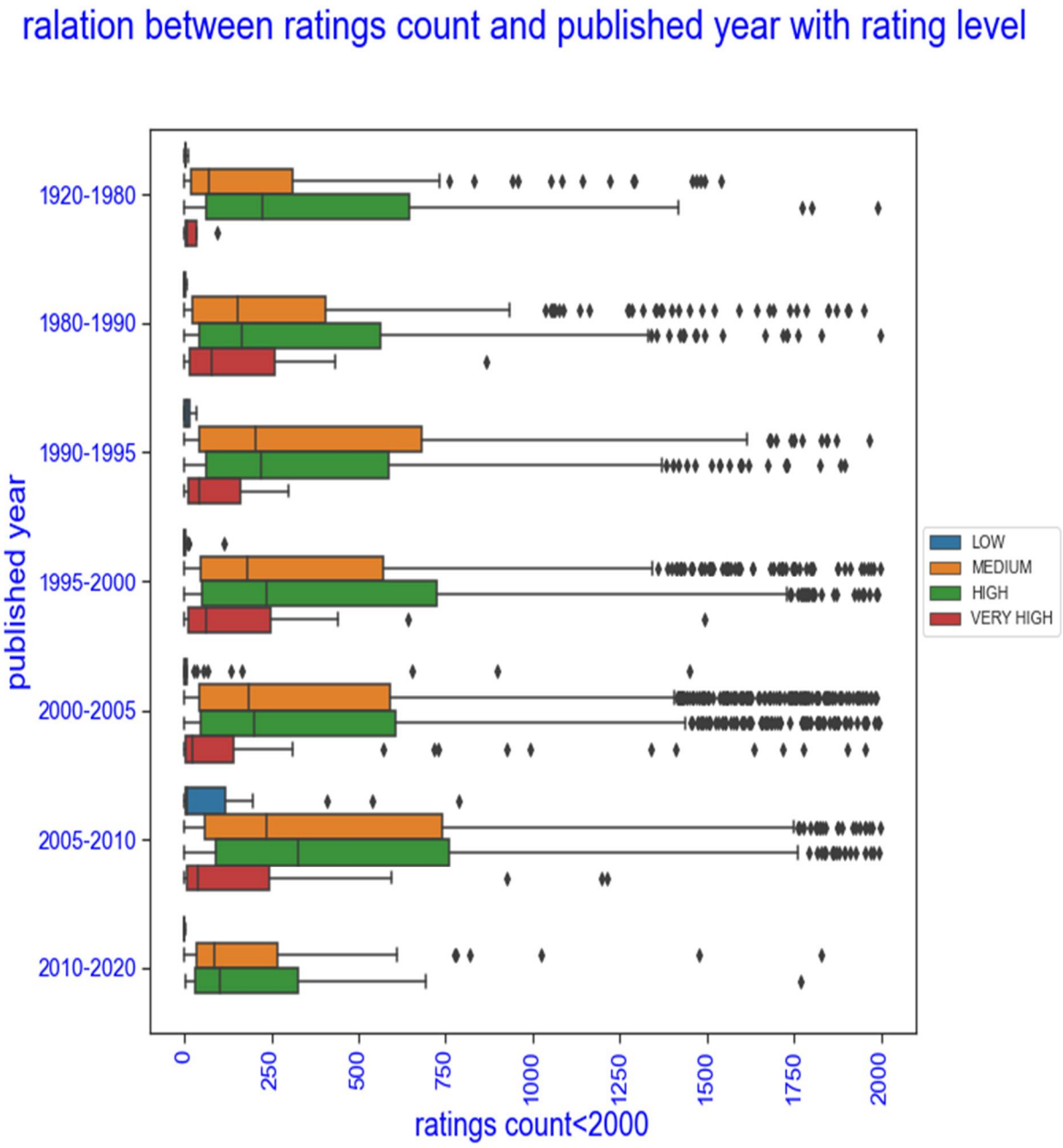
(10) Here also same conclusion with no of pages shown in different language.



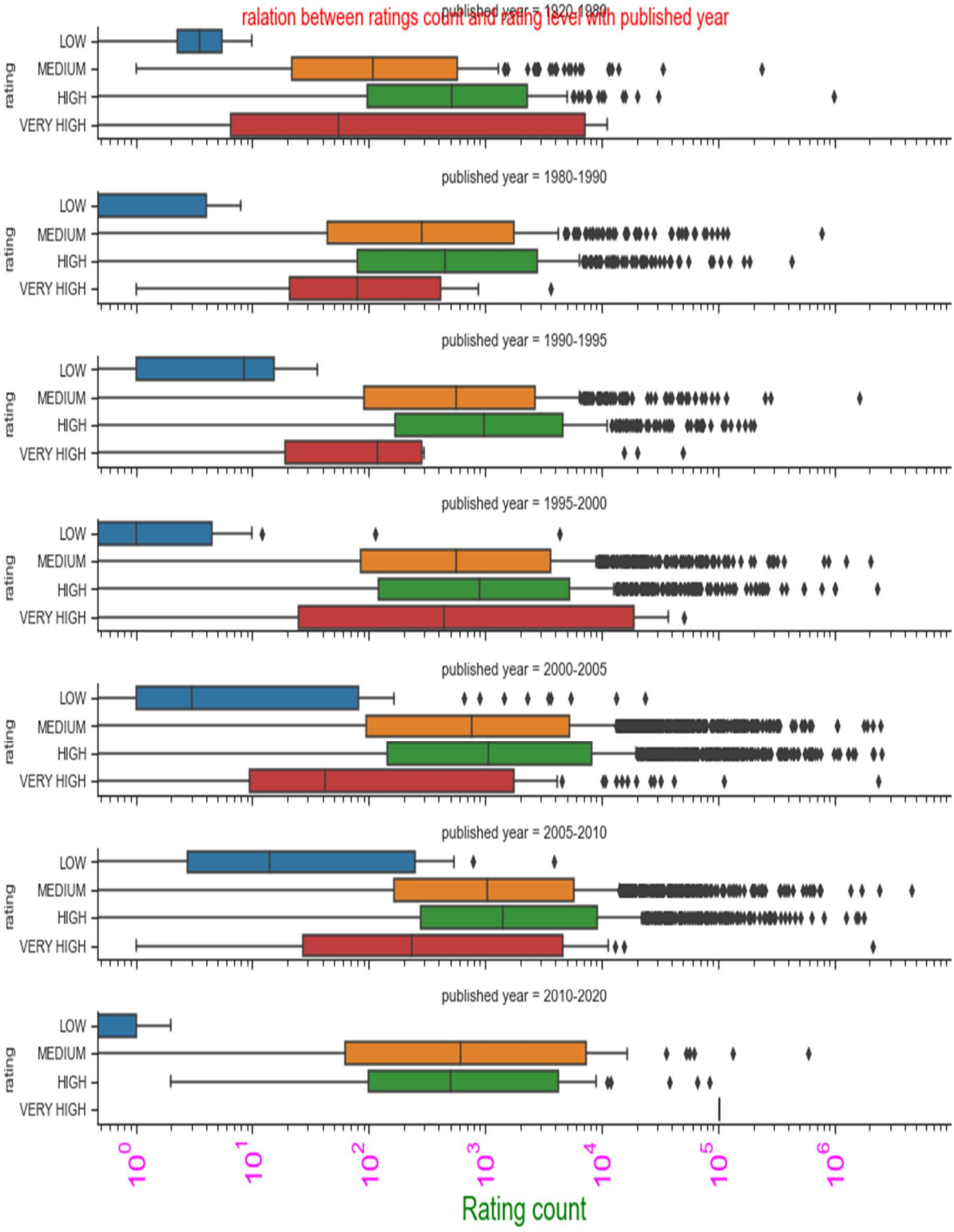
(11) Another interesting fact that the total rating count also have a general trend to increase in all language but in high rated book it is opposite. That means a high average rated book doesn't mean a high total rated or responded ones.



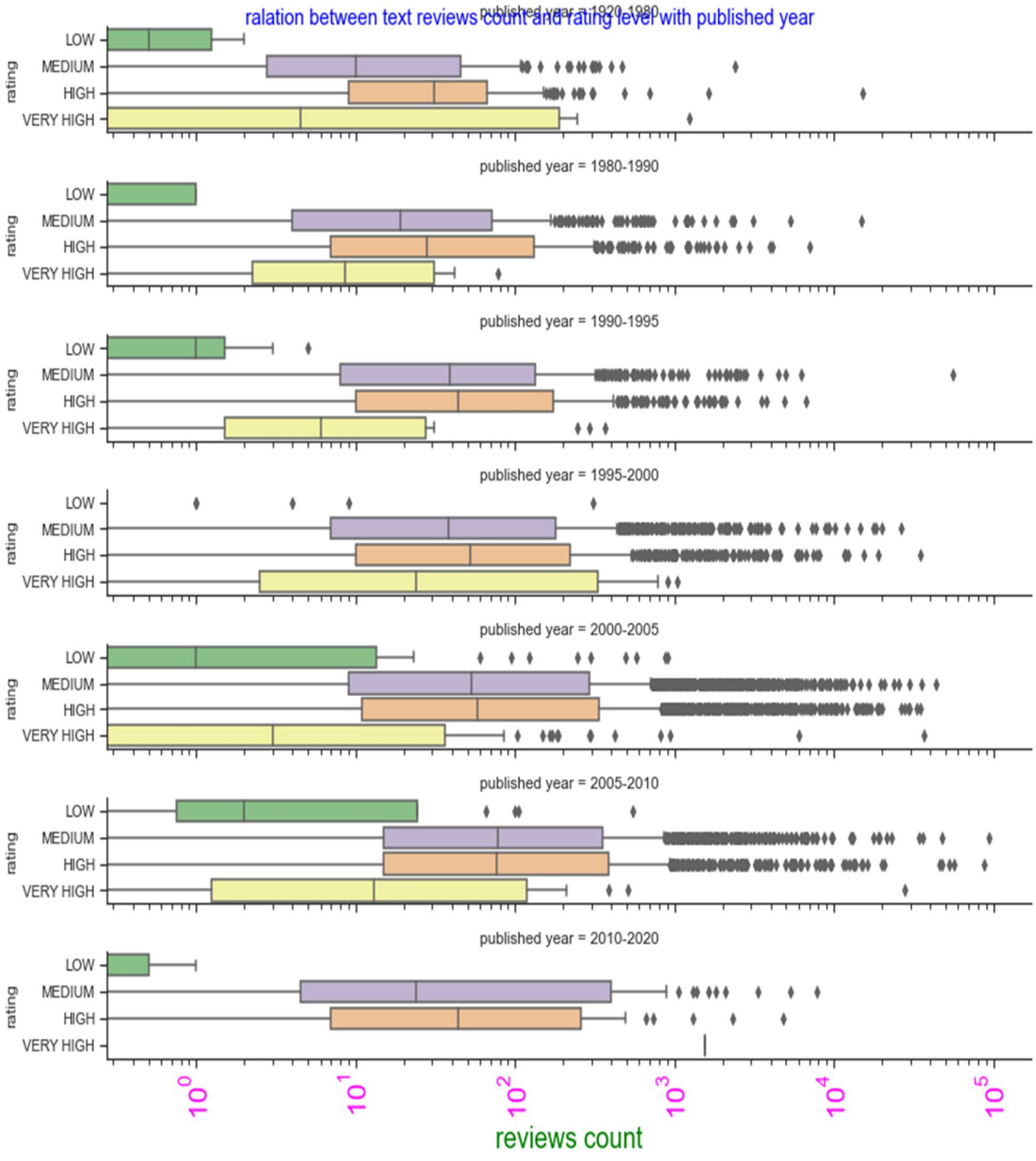
(12) Here relation between ratings count and publishing year is shown with different rating level, the distribution vividly says the above said fact of high rated books.



(13) Here also distribution of rating is shown .



(14) Here relation between reviews count and publishing year is shown with different rating level, the distribution vividly says the above said fact of high rated books.



(15) The three figures below are representing the variation of different factors in case of 3 top authors and publishers.

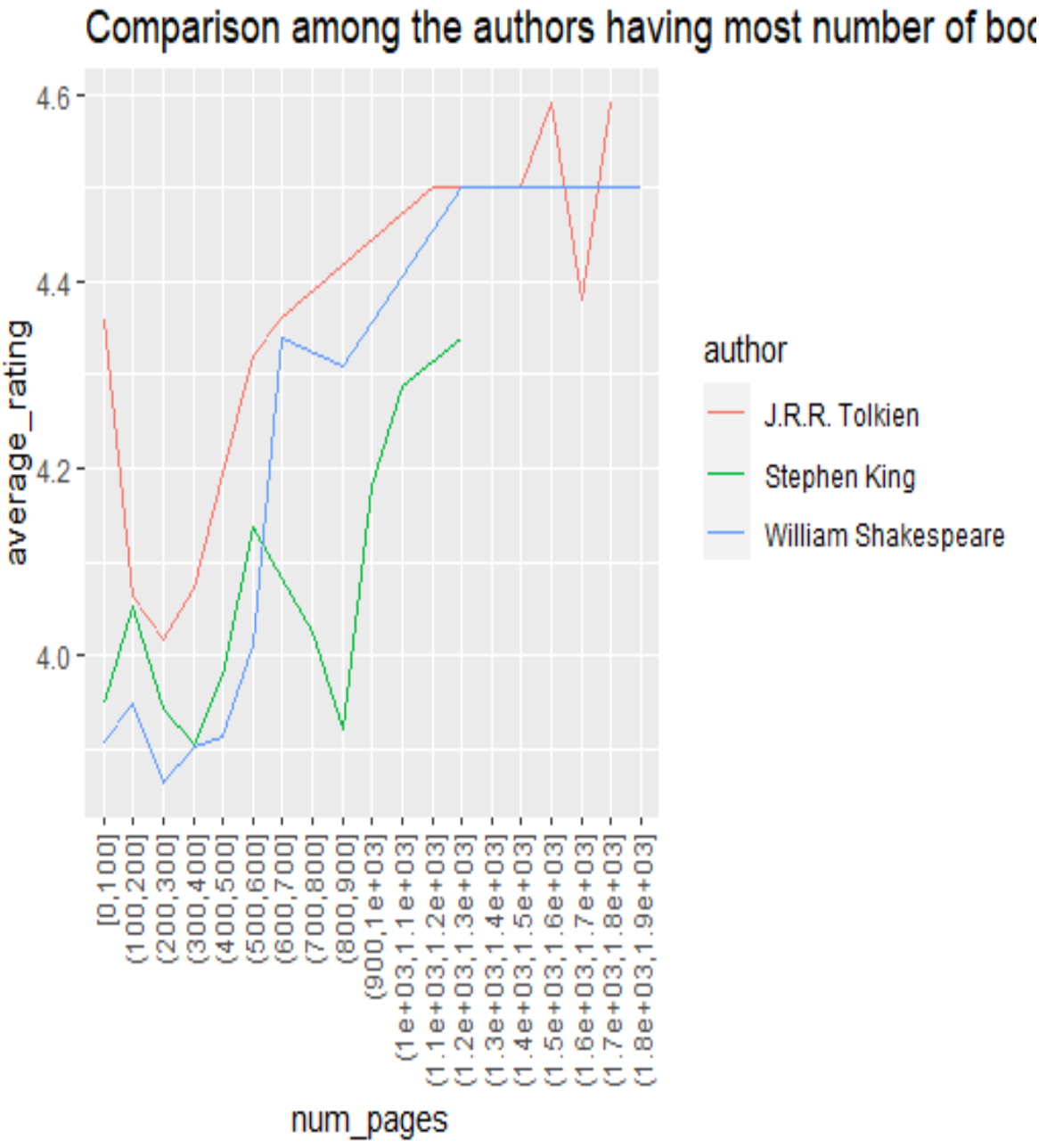


Fig. 14.a

Fig.14.b

Fig.15.a

**********----------THE END----------**********