Mini_Poster

Kapil Sahu

10/9/2021

Title:

Analyzing Goodreads Dataset to gain interesting insights about factors impacting the popularity of Books

Dataset Description:

The data set used for analysis is collected from the Goodreads database(Goodreads API: https://goodreads.com/api) (https://www.kaggle.com/hoshi7/goodreads-analysis-and-recommending-books/data) and is available on Kaggle. It contains book title along with authors, language, ratings, reviews, size of book and ISBN. A brief description of data available in it is below:

Columns Description:

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 for the books **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

```
dfb <- read.csv("C:\\Users\\kapil\\OneDrive\\Desktop\\IDMP\\books\\books.csv")

head(dfb,20) # Printing top 20 rows of dataset
```

```
##    bookID
## 1       1
## 2       2
## 3       3
## 4       4
## 5       5
## 6       8
## 7       9
## 8      10
## 9      12
## 10     13
## 11     14
## 12     16
## 13     18
## 14     21
## 15     22
## 16     23
## 17     24
## 18     25
## 19     26
## 20     27
##
## 1                                                      Harry Potter and the Half-Blood Prince (Harry P
## 2                                                   Harry Potter and the Order of the Phoenix (Harry P
## 3                                                       Harry Potter and the Sorcerer's Stone (Harry P
## 4                                                     Harry Potter and the Chamber of Secrets (Harry P
## 5                                                    Harry Potter and the Prisoner of Azkaban (Harry P
## 6                                                         Harry Potter Boxed Set  Books 1-5 (Harry Pot
## 7                            Unauthorized Harry Potter Book Seven News: Half-Blood Prince Analysis and S
## 8                                                                   Harry Potter Collection (Harry Pot
## 9  The Ultimate Hitchhiker's Guide: Five Complete Novels and One Story (Hitchhiker's Guide to the Gal
## 10                                                                 The Ultimate Hitchhiker's Guide to
## 11                                   The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide to the G
## 12                                   The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide to the G
## 13                                       The Ultimate Hitchhiker's Guide (Hitchhiker's Guide to the Ga
## 14                                                                          A Short History of Nearly
## 15                                                                                   Bill Bryson's Afr
## 16                                Bryson's Dictionary of Troublesome Words: A Writer's Guide to Gettin
## 17                                                                                        In a Sunburn
## 18                             I'm a Stranger Here Myself: Notes on Returning to America After Twenty
## 19                                                             The Lost Continent: Travels in Small To
## 20                                                                     Neither Here nor There: Travels
##                         authors average_rating       isbn        isbn13
## 1  J.K. Rowling-Mary GrandPrÃ©           4.56 0439785960 9780439785969
## 2  J.K. Rowling-Mary GrandPrÃ©           4.49 0439358078 9780439358071
## 3  J.K. Rowling-Mary GrandPrÃ©           4.47 0439554934 9780439554930
## 4                  J.K. Rowling           4.41 0439554896 9780439554893
## 5  J.K. Rowling-Mary GrandPrÃ©           4.55 043965548X 9780439655484
## 6  J.K. Rowling-Mary GrandPrÃ©           4.78 0439682584 9780439682589
## 7        W. Frederick Zimmerman           3.69 0976540606 9780976540601
## 8                  J.K. Rowling           4.73 0439827604 9780439827607
## 9                 Douglas Adams           4.38 0517226952 9780517226957
## 10                Douglas Adams           4.38 0345453743 9780345453747
## 11                Douglas Adams           4.22 1400052920 9781400052929
## 12    Douglas Adams-Stephen Fry           4.22 0739322206 9780739322208
## 13                Douglas Adams           4.38 0517149257 9780517149256
## 14 Bill Bryson-William Roberts           4.20 076790818X 9780767908184
## 15                  Bill Bryson           3.43 0767915062 9780767915069
## 16                  Bill Bryson           3.88 0767910435 9780767910439
## 17                  Bill Bryson           4.07 0767903862 9780767903868
## 18                  Bill Bryson           3.90 076790382X 9780767903820
## 19                  Bill Bryson           3.83 0060920084 9780060920081
## 20                  Bill Bryson           3.87 0380713802 9780380713806
##    language_code X..num_pages ratings_count text_reviews_count
## 1            eng          652       1944099              26249
## 2            eng          870       1996446              27613
## 3            eng          320       5629932              70390
## 4            eng          352          6267                272
## 5            eng          435       2149872              33964
## 6            eng         2690         38872                154
## 7          en-US          152            18                  1
## 8            eng         3342         27410                820
## 9            eng          815          3602                258
## 10           eng          815        240189               3954
## 11           eng          215          4416                408
## 12           eng            6          1222                253
## 13         en-US          815          2801                192
## 14           eng          544        228522               8840
## 15           eng           55          6993                470
## 16           eng          256          2020                124
## 17           eng          335         68213               4077
## 18           eng          304         47490               2153
## 19         en-US          299         43779               2146
## 20           eng          254         46397               2127
```

```r
sapply(dfb, function(x) sum(is.na(x))) #To check Missing value count
```

```
##             bookID              title            authors     average_rating 
##                  0                  0                  0                  0 
##               isbn             isbn13      language_code       X..num_pages 
##                  0                  0                  0                  0 
##      ratings_count text_reviews_count 
##                  5                  5
```

```r
dups <-dfb[duplicated(dfb$bookID)|duplicated(dfb$bookID, fromLast=TRUE),]
dups # finding duplicate values in bookID
```

```
##      bookID                                                           title
## 4         4 Harry Potter and the Chamber of Secrets (Harry Potter  #2)
## 5689      4                                                                
## 7058      4                                                                
##           authors average_rating       isbn        isbn13 language_code
## 4    J.K. Rowling           4.41 0439554896 9780439554893           eng
## 5689                                                                   
## 7058                                                                   
##      X..num_pages ratings_count text_reviews_count
## 4             352          6267                272
## 5689                         NA                 NA
## 7058                         NA                 NA
```

```r
dfb1<-dfb

dfb2<-dfb1[dfb1$title != "",] # Removed all books which have no title

dups <-dfb2[duplicated(dfb2$bookID)|duplicated(dfb2$bookID, fromLast=TRUE),]
dups # finding duplicate values of bookID in updated dfb2
```

```
##  [1] bookID             title              authors            average_rating    
##  [5] isbn               isbn13             language_code      X..num_pages      
##  [9] ratings_count      text_reviews_count
## <0 rows> (or 0-length row.names)
```

```
sapply(dfb2, function(x) sum(is.na(x))) #No NA values now
```

```
##             bookID              title            authors     average_rating
##                  0                  0                  0                  0
##               isbn             isbn13      language_code       X..num_pages
##                  0                  0                  0                  0
##      ratings_count text_reviews_count
##                  0                  0
```

```
dfb2$authors <- gsub("-.*","",dfb2$authors) #Kept only primary author

dfb2[, c(4,8)] <- sapply(dfb2[, c(4,8)], as.numeric)
```

```
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
```

```
a<- table(dfb2$authors)
```

Printed first 20 lines of dataset above.

Tidied Data:

As mentioned along with code in the comments I have tidiesd the dataset by doing below tasks: 1. Checked missing values 2. Removed duplicate values 3. Removed all books which have no title 4. Kept only primary author to simplify the dataset analysis

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

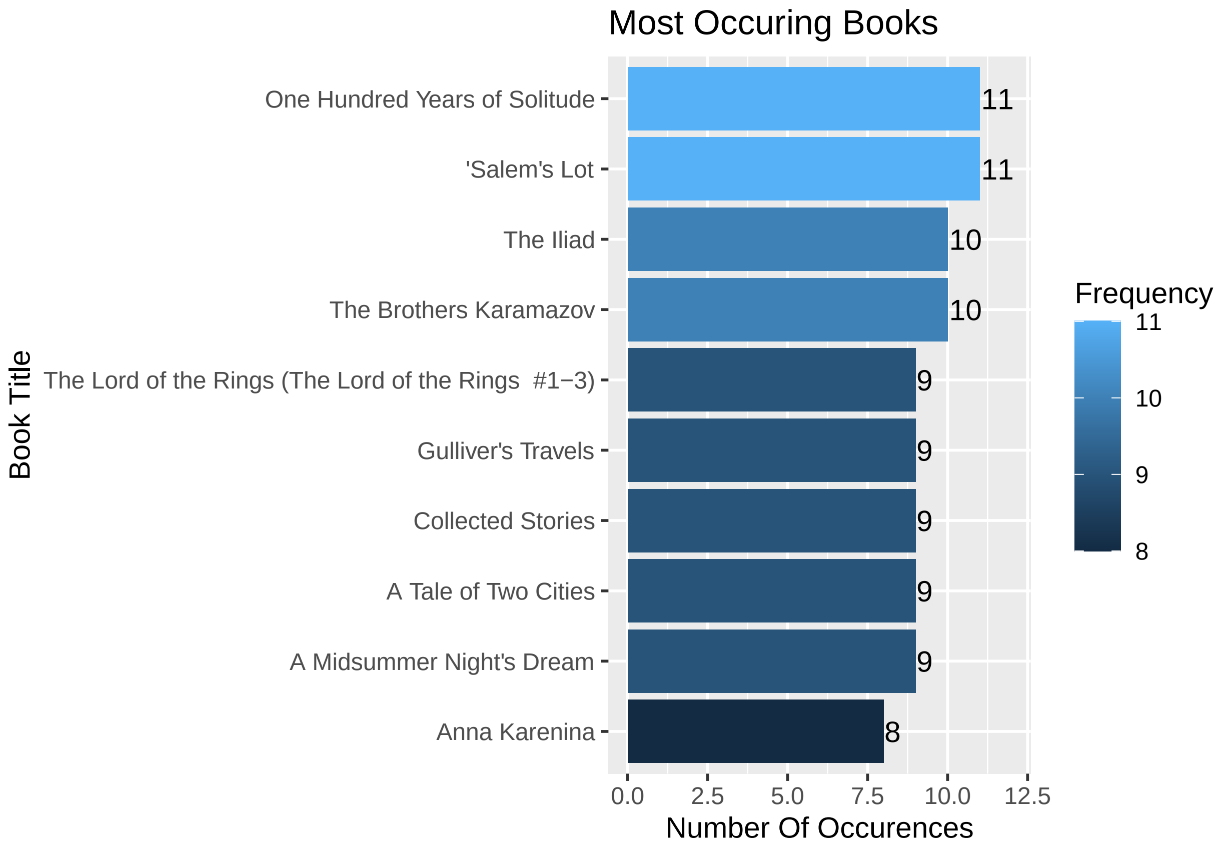
```
library(scales)

# Titles of top 10 most occuring books:
df_title_top <- dfb2 %>%
  group_by(dfb2$title)  %>%
  summarise(count=n()) %>% arrange(desc(count)) %>% head(10)

names(df_title_top)[1] <- 'ttl'
names(df_title_top)[2] <- 'frq'

df_title_top$frq = as.numeric(df_title_top$frq)

#Barplot showing most occuring books
df_title_top %>%
  ggplot(aes(x = frq, y = reorder(ttl,frq), fill = frq))+
  geom_col()+
  geom_text(aes(label = frq), hjust = -0.05)+
  ggtitle("Most Occuring Books")+
  xlab("Number Of Occurences")+
  ylab("Book Title")+
  labs(fill = "Frequency")+
  coord_cartesian(xlim = c(0, 12))
```



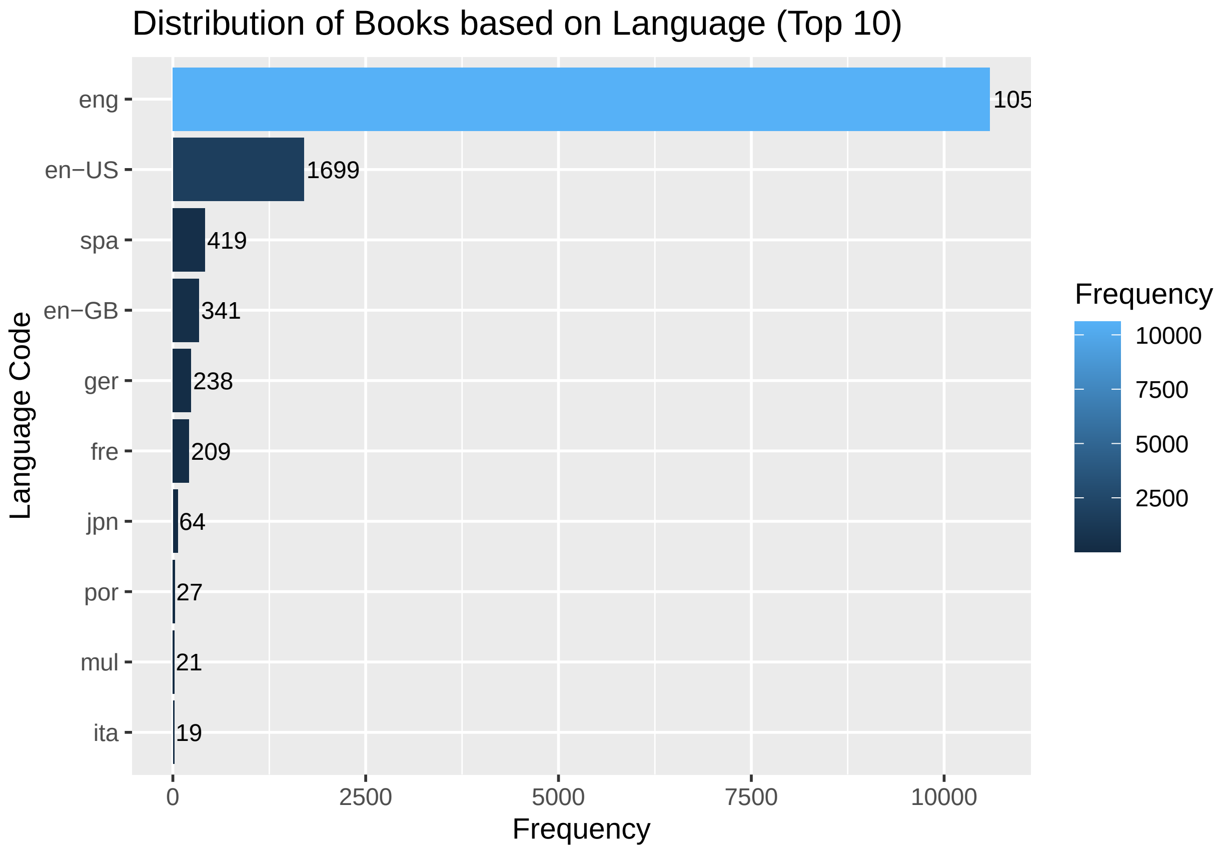
Observation:

One Hundred Years Of Solitude’ and ‘Salem’s Lot’ have the most number of occurrences with the same name in the data.These books have come up in this database over and over again, with various publication editions.

```
#Top 10 languages in which books are published
df_lang <- dfb2 %>% group_by(dfb2$language_code) %>%
  summarise(count=n()) %>% arrange(desc(count)) %>% head(10)
```

```r
names(df_lang)[1] <- 'lang'
names(df_lang)[2] <- 'frq'


#barplot showing distribution of books for all languages
df_lang %>%
  ggplot(aes(x = frq, y =reorder(lang,frq), fill = frq ))+
  geom_col() +
  geom_text(aes(label = frq), size = 3, hjust =-0.05)+
  ggtitle("Distribution of Books based on Language (Top 10)")+
  xlab("Frequency")+
  ylab("Language Code")+
  labs(fill = "Frequency")
```



Observation:

Majority of the books are in english languages, with some further categorized into English-US, english-UK and english-CA. Then some are in Spanish, German and French.

```r
pdf("Output.pdf") # For miniposter

#Top 10 most rated books
df_rating <- dfb2 %>% arrange(desc(ratings_count)) %>% head(10)
```

```
#Plotting top 10 most rated books
df_rating %>%
  ggplot(aes(x = ratings_count,
             y = reorder(title,ratings_count), fill = ratings_count))+
  geom_col()+
  geom_text(aes(label = ratings_count), size = 3, hjust =-0.05)+
  ggtitle("Most Rated Books (Top 10)")+
  ylab("Title")+
  xlab("Rating Count")+
  labs(fill = "Rating Count")+
  coord_cartesian(xlim = c(0, 7000000))

dev.off() # for miniposter
```

```
## pdf
##   2
```

Analysis:

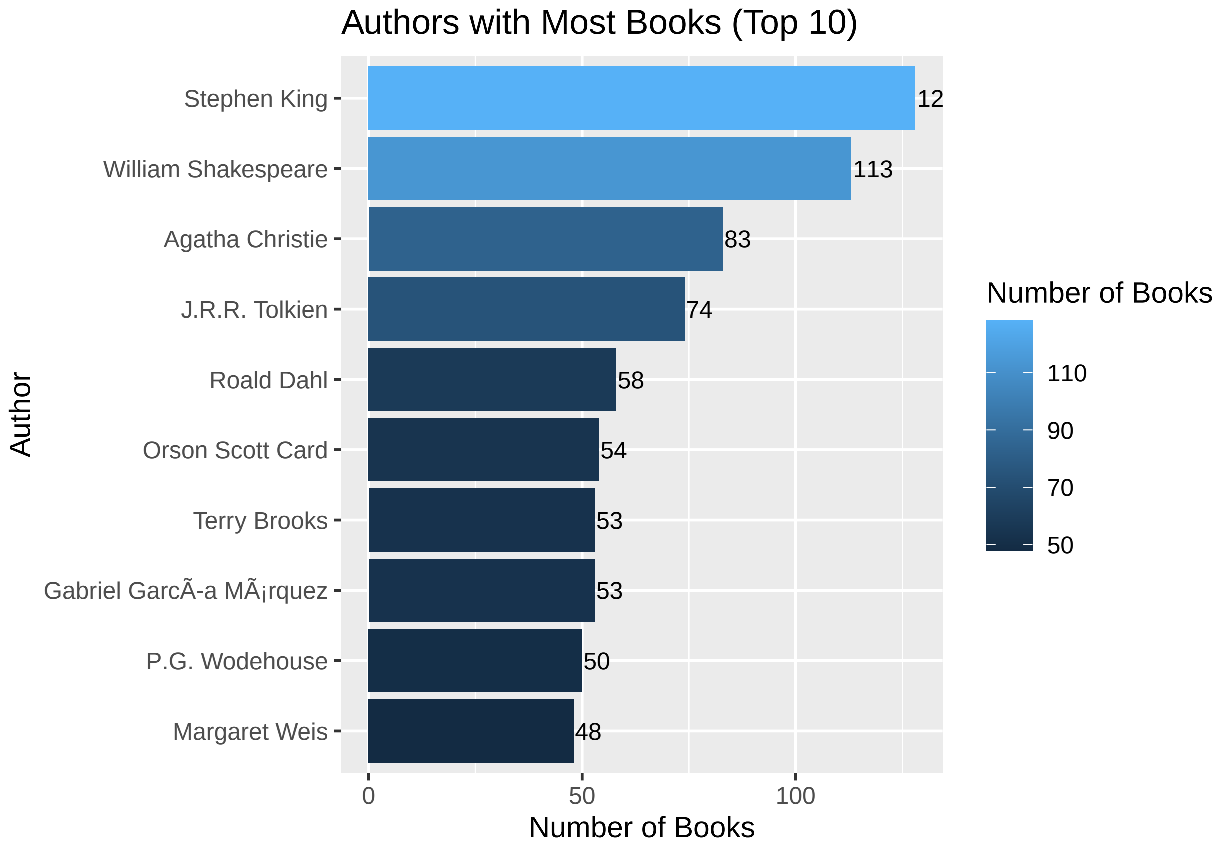
Here we are looking at the top 10 books with the most reviews. The first book of any series usually has most of the ratings, i.e, Harry Potter and the Sorcerer's Stone, Twilight #1, The Hobbit, Angels and Demons #1. However, A huge gap in ratings(approx. 50%) of Harry Potter(#1) and Harry Potter(#2) indicates that fiction enthusiasts did not pick up the sequel in the series as much as they liked the first one. This is quite interesting as generally, people eagerly wait for a sequel of a book if the first part is a big hit and authors expect a positive response mostly. But this data tells otherwise.

```
#Authors with most books

df_auth <- dfb2 %>% group_by(dfb2$authors) %>%
  summarise(count=n()) %>% arrange(desc(count)) %>% head(10)

names(df_auth)[1] <- 'author'
names(df_auth)[2] <- 'no_of_books'

df_auth %>%
  ggplot(aes(x= no_of_books, y = reorder(author,no_of_books),
             fill = no_of_books)) +
  geom_col()+
  geom_text(aes(label = no_of_books), size = 3, hjust =-0.05)+
  ggtitle("Authors with Most Books (Top 10)")+
  xlab("Number of Books")+
  ylab("Author")+
  labs(fill = "Number of Books")
```



Observation:

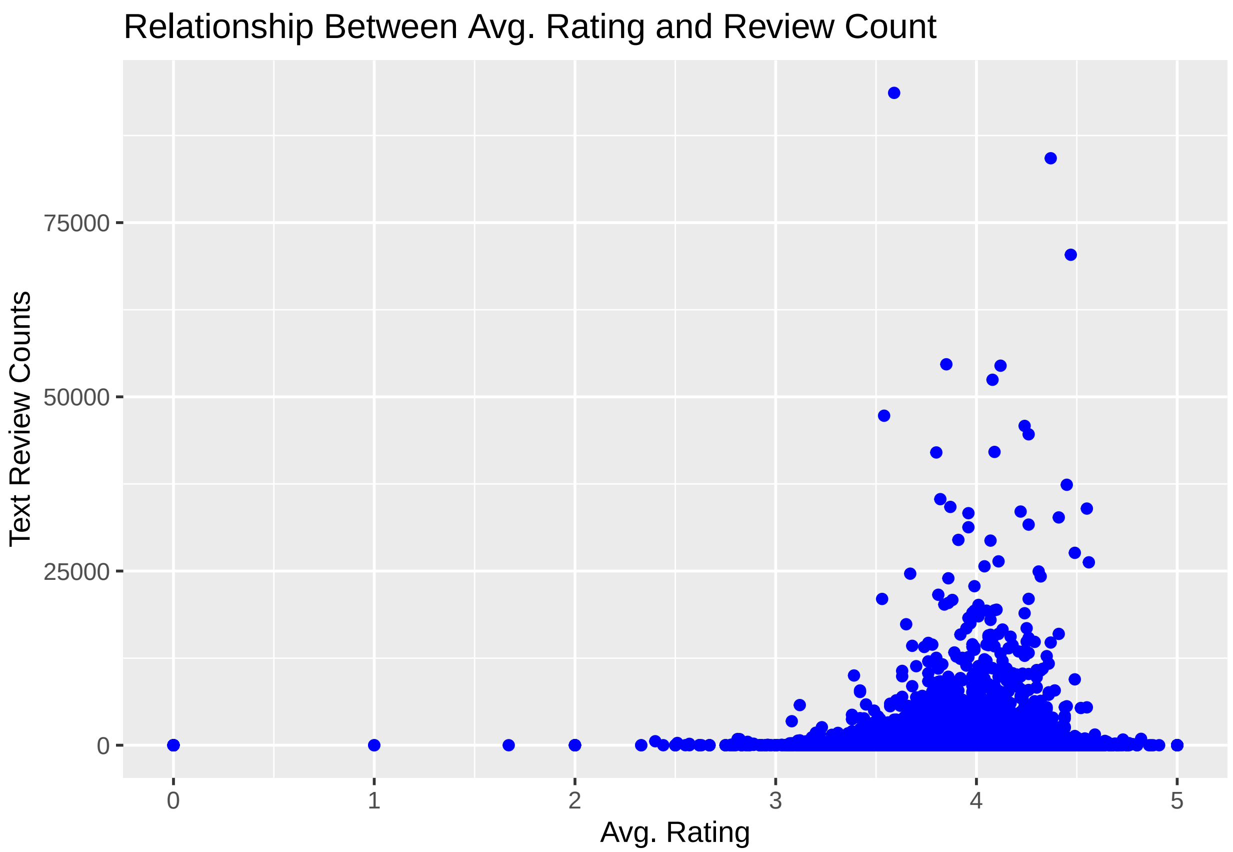
Stephen King is the author with with most books followed by William Shakespeare and Agatha Christie. Most of the authors have either been writing for decades, churning numerous books from time to time, or are authors who are known as the 'classics' in our history.

```
#For Miniposter
#Relationship between avg rating and review count:

df_avg_interval <- dfb2
df_avg_interval$average_rating = as.numeric(df_avg_interval$average_rating)

df_avg_interval %>%
  ggplot(aes(x= average_rating, y = text_reviews_count)) +
  geom_point(color = "blue") +
  ggtitle("Relationship Between Avg. Rating and Review Count")+
  xlab("Avg. Rating")+
  ylab("Text Review Counts")+
  scale_x_continuous(breaks=c(0,1,2,3,4,5))
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



Analysis Plot 1:

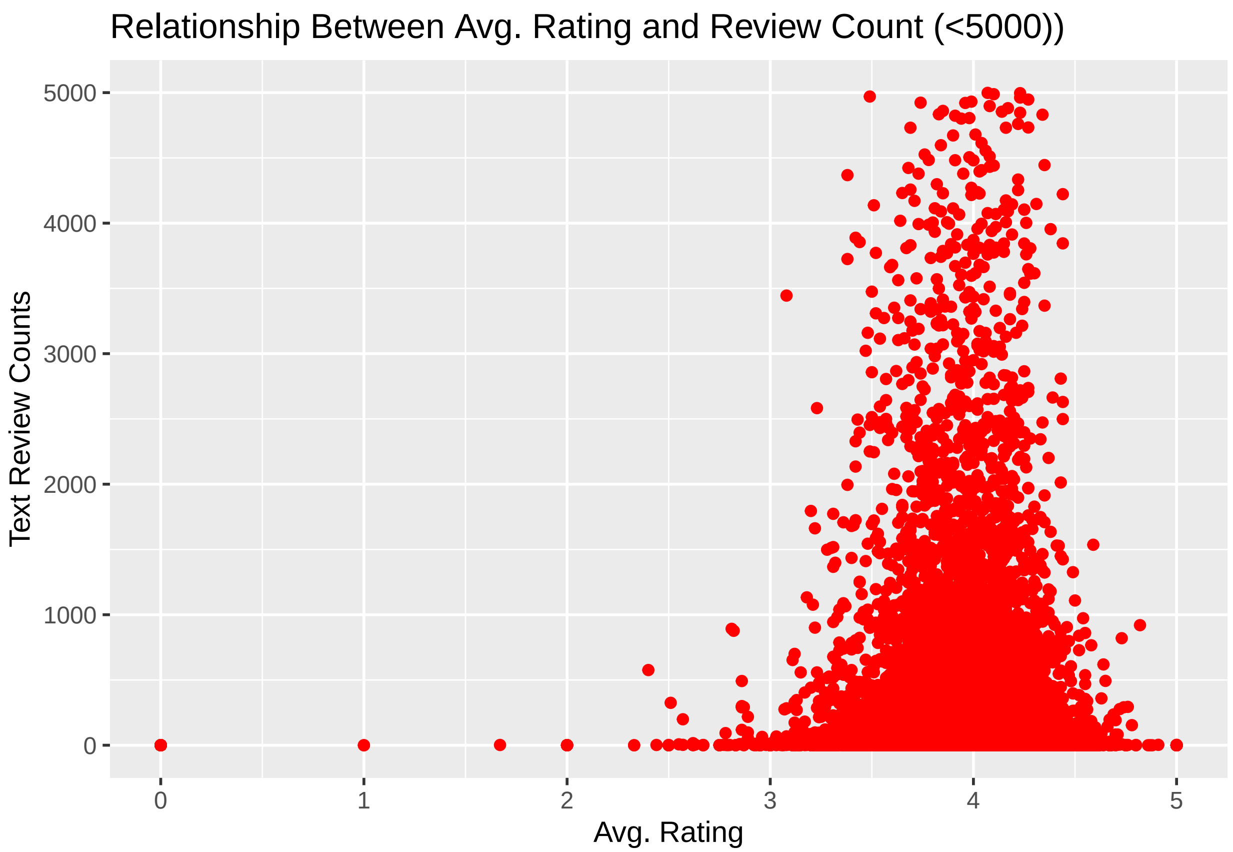
It is a scatter plot from which we can infer that most of the ratings for the books seem to lie near 3-4, with a heavy number of reviews lying mostly near 5000, approx. Hence, we need to look closely at the area where the review count is below 5000 to get some insights.

```
#For Miniposter
#To get better resolution for rating < 5000

df_avg_interval2 <- subset(df_avg_interval,
                           df_avg_interval$text_reviews_count < 5000)

df_avg_interval2 %>%
  ggplot(aes(x= average_rating, y = text_reviews_count )) +
  geom_point(color = "red") +
  ggtitle("Relationship Between Avg. Rating and Review Count (<5000))")+
  xlab("Avg. Rating")+
  ylab("Text Review Counts")+
  scale_x_continuous(breaks=c(0,1,2,3,4,5))
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



Analysis Plot 2:

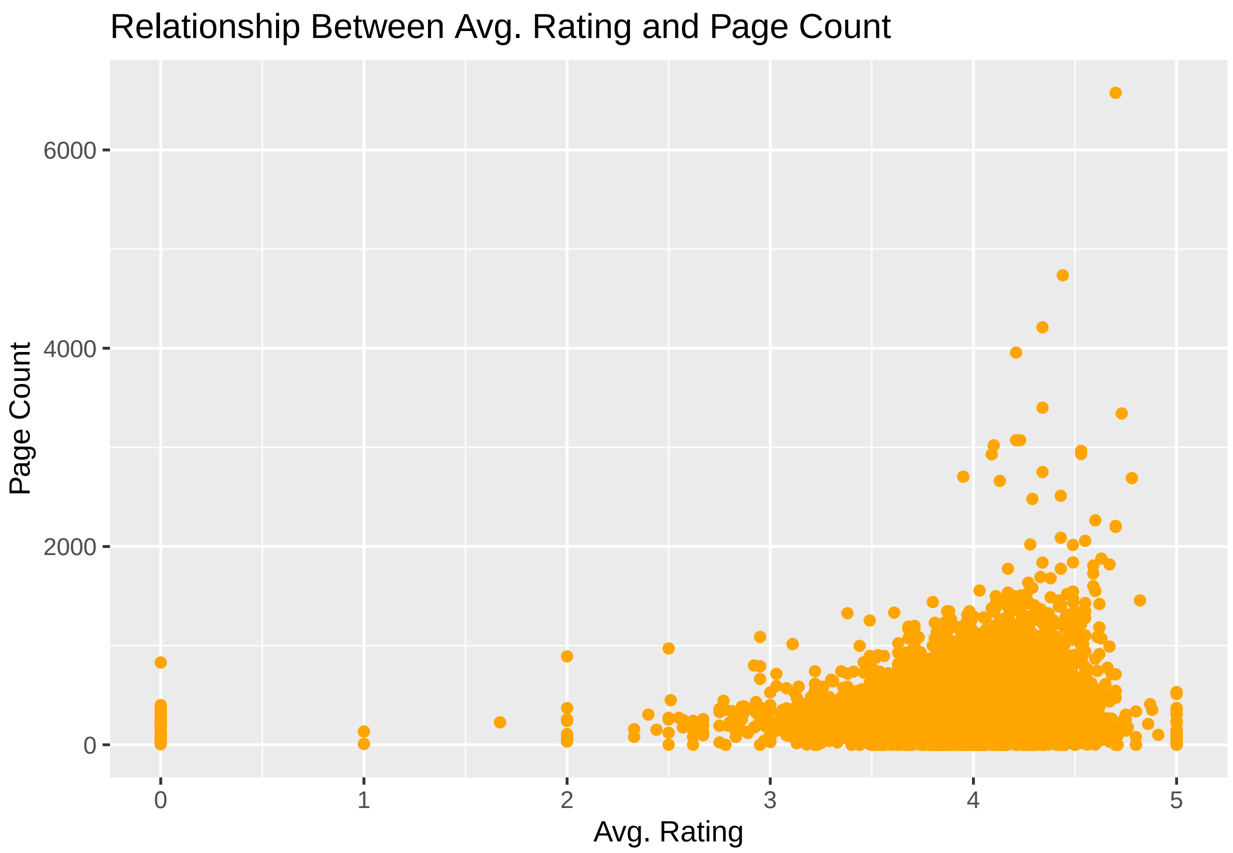
This scatter plot is a kind of magnified version of plot 1 where we are looking at observations where the review count is less than 5000. Even after looking at a smaller scale, most text reviews for books still lie under 1000, making data results inconclusive. The reviews seem to be predominant amongst books with good ratings. Maybe this is pointing towards a possibility that these are all fake reviews. Fake reviews are usually posted to promote a product in the market and hence are not from genuine readers which is somewhat self-explanatory since most of the time, bibliophiles don't just rate a book, they also share their honest reviews no matter positive or negative to help the community grow.

```
#Relationship between avg rating and number of pages:

df_avg_rat_pages <- dfb2
df_avg_rat_pages$average_rating = as.numeric(df_avg_rat_pages$average_rating)
df_avg_rat_pages$X..num_pages = as.numeric(df_avg_rat_pages$X..num_pages)

df_avg_rat_pages %>%
  ggplot(aes(x= average_rating, y = X..num_pages)) +
  geom_point(color= "orange") +
  ggtitle("Relationship Between Avg. Rating and Page Count ")+
  xlab("Avg. Rating")+
  ylab("Page Count")+
  scale_x_continuous(breaks=c(0,1,2,3,4,5))
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

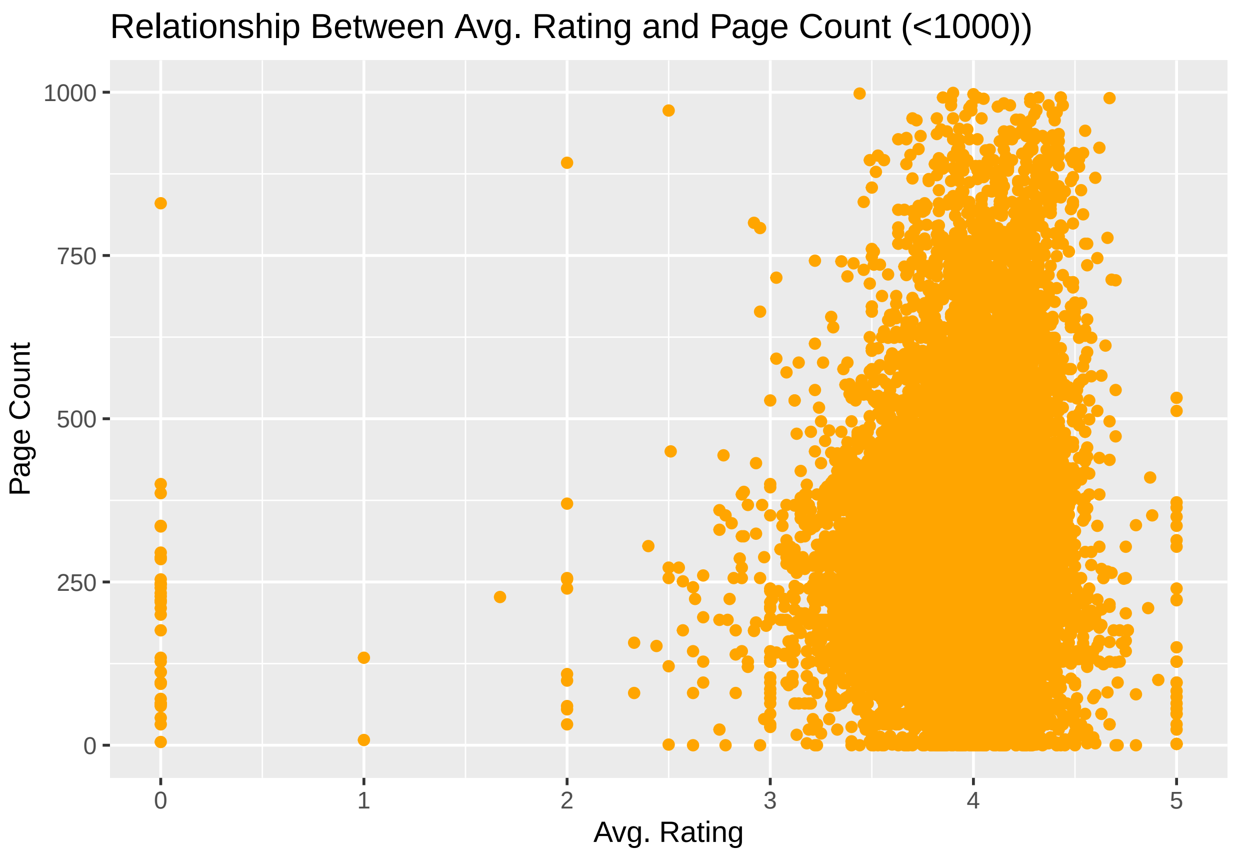


Analysis Plot 1:

This plot doesn't give that much of an accurate inference due to the massive presence of outliers for books above 1000 pages, for the maximum density is between 0-1000 pages.

```
df_avg_rat_pages2 <- df_avg_rat_pages <- subset(df_avg_rat_pages,
                                          df_avg_rat_pages$X..num_pages < 1000)

df_avg_rat_pages2 %>%
  ggplot(aes(x= average_rating, y = X..num_pages)) +
  geom_point(color= "orange") +
  ggtitle("Relationship Between Avg. Rating and Page Count (<1000))")+
  xlab("Avg. Rating")+
  ylab("Page Count")+
  scale_x_continuous(breaks=c(0,1,2,3,4,5))
```



Analysis Plot 2:

From the given plot, we can infer that the highest ratings ever given, usually are for books with the page range of 200-400, peaking near 250. Which could mean that most of the people seem to prefer books with a moderate amount of pages, and people are not that fond of thicker books.