DATA VISUALISATION PROJECT ON BOOK REVIEW DATASET

BY: ASIF IKBAL & MAHENDRA NANDI

Visualizing Book Review Dataset

The dataset consists 13 columns. 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.

Dependencies:

- (1) Distribution of average rating
- From the distribution of the average rating in Fig. 1 it is clear that almost all the books have been rated in the range 3.5-4.5 . Out of almost 11000 books 22 books is rated as ‘‘5’’.

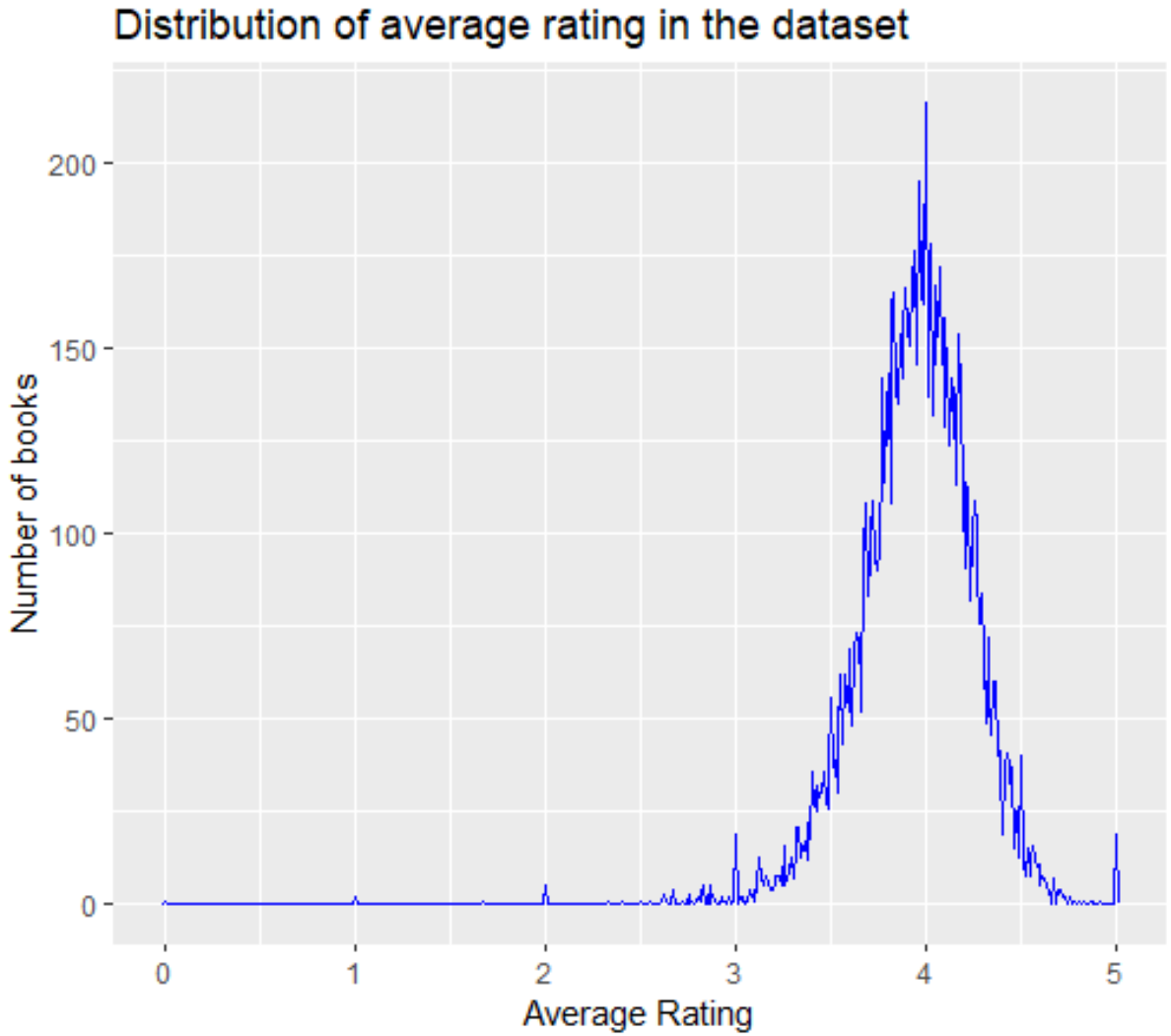


Fig. 1

(2)Dependencies between different review parameters and average ratings

We can see that maximum books got reviews less than 20000 . Fig. 2.a,2.b,2.c and 2.d shows the distribution.

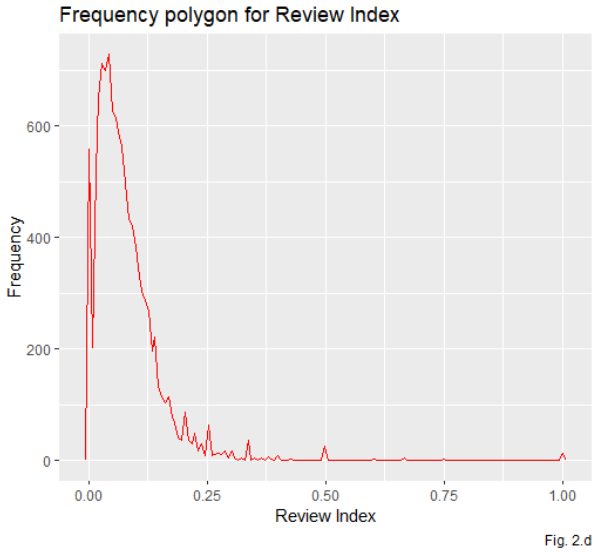


Fig. 2.d

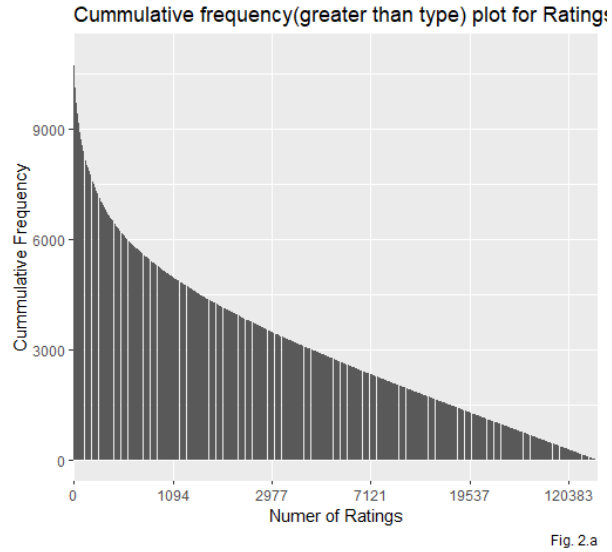


Fig. 2.a

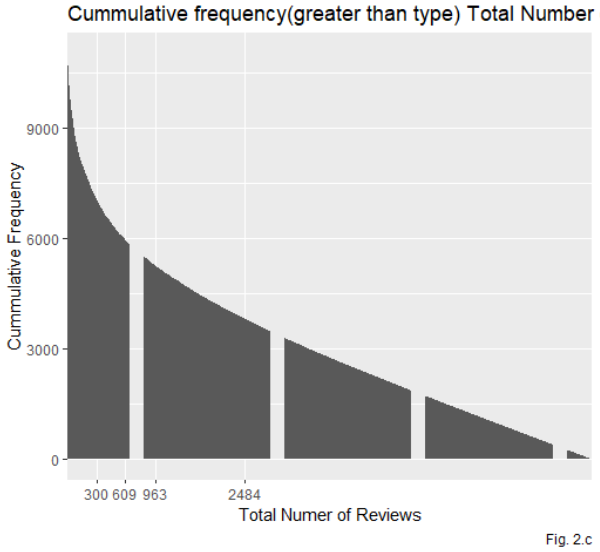


Fig. 2.c

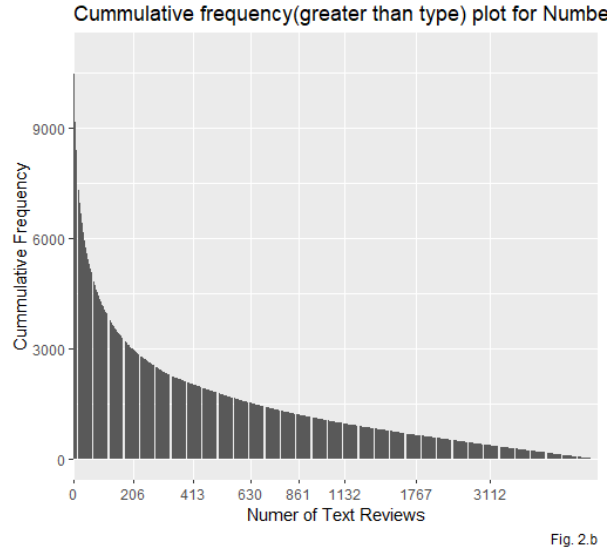


Fig. 2.b

- (3) Number of books in different languages

- Here maximum books are written in “eng” language . 80% book is written in “eng” 13% in “en-US” and the remaining 7% is in other language.

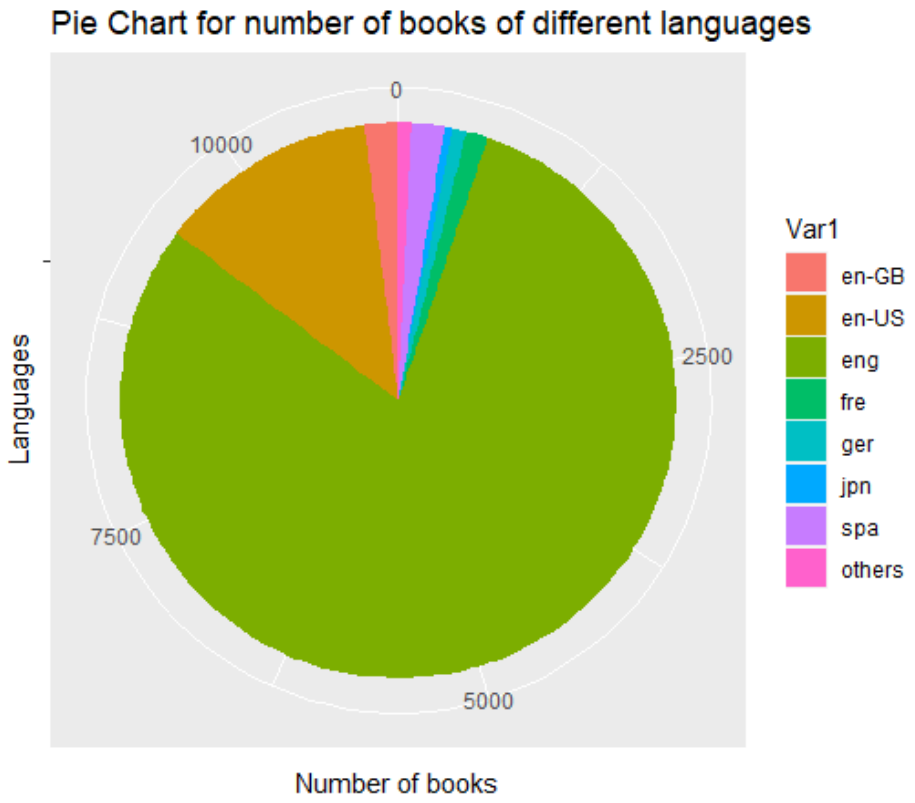


Fig. 3

- (4) Number of books in different page ranges

- Pages of maximum books(around 60%) are in the range of 200-400

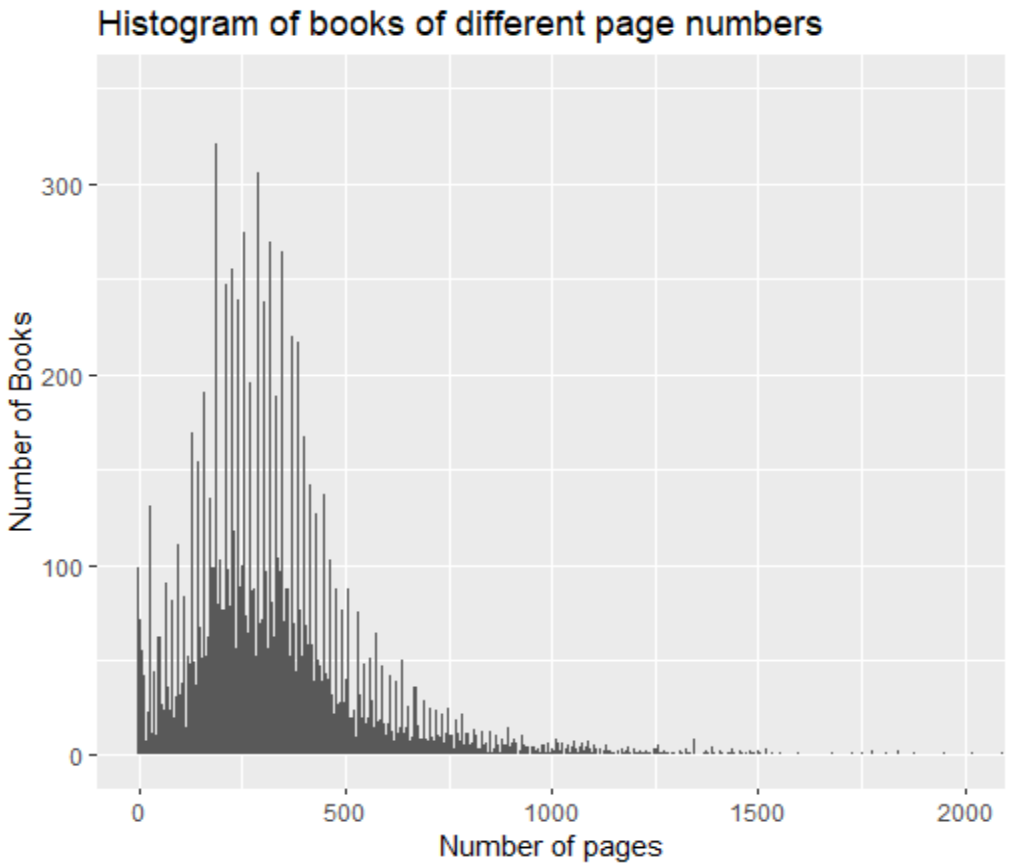


Fig. 4

• (5) Number of authors having exact number of books

- There are more than 6000 authors who wrote only one book . And 1000 authors wrote two books . The number of authors who wrote more than 5 books is very low.

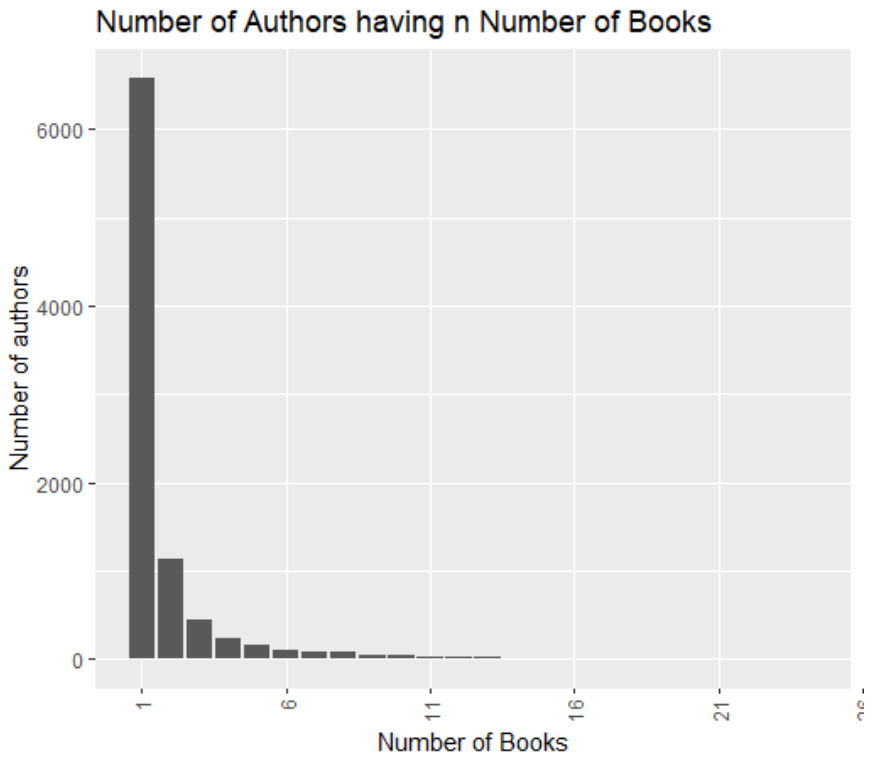


Fig. 5

• (6) Books published in a different years

- It is seen that in every year no of published books is increased largely . But after 2016 it suddenly goes down.

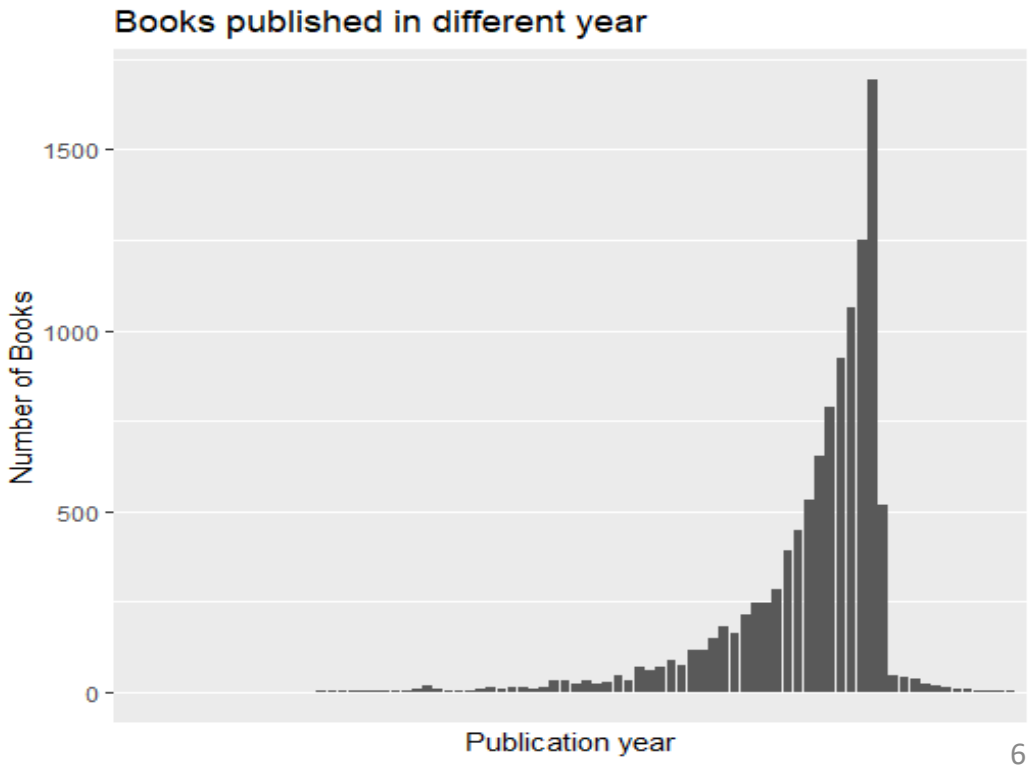


Fig.6

(7)Dependency between Average rating and different languages

- **Though the number of books written in “eng” is much greater than other languages , average rating of books written in 'eng' is ~3.9 . Whereas very few books are written in “wel” but the average rating is 5 .**

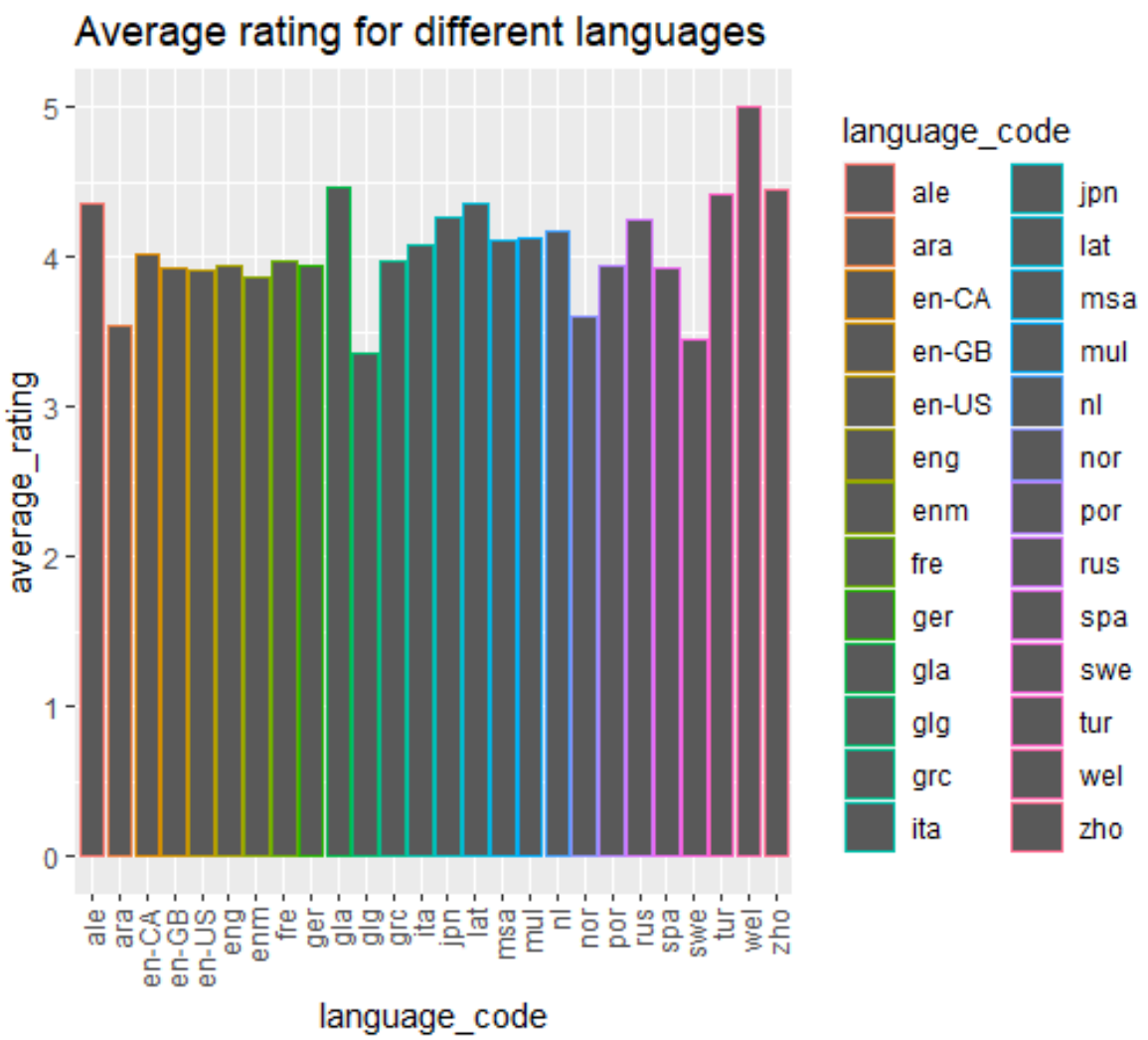


Fig. 7

- (8) Relation between Average Rating and Number of pages

- We can find here that rating is almost varies as the num of pages of the book . Book with no of pages 2600-2700 is of high rated . Though there are few books in this range , It actually ensures that if a book is interesting to the readers, total number pages of book does not matter.

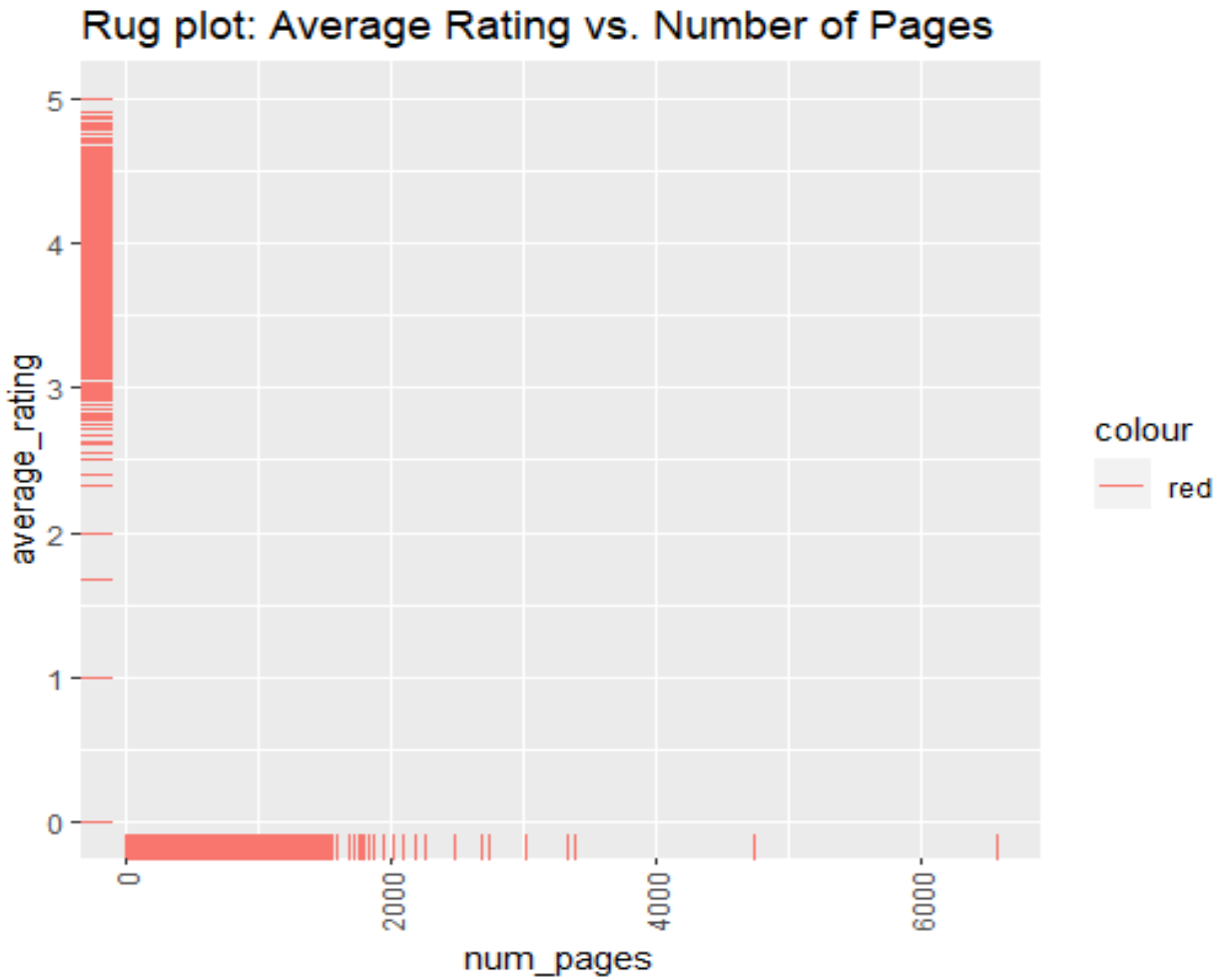


Fig. 8

- .(9)Dependencies of the various review parameters with average ratings

- As we earlier saw that there are very few books with very high number of reviews . The books with very high reviews also fall under the rating range 3.5-4.5 . So, we can say maximum people, who gives a review , responds to a book with an average rating of 3.5-4.5 ,unless they get a good satisfaction with the book or become totally hopeless .Which is intuitively acceptable.

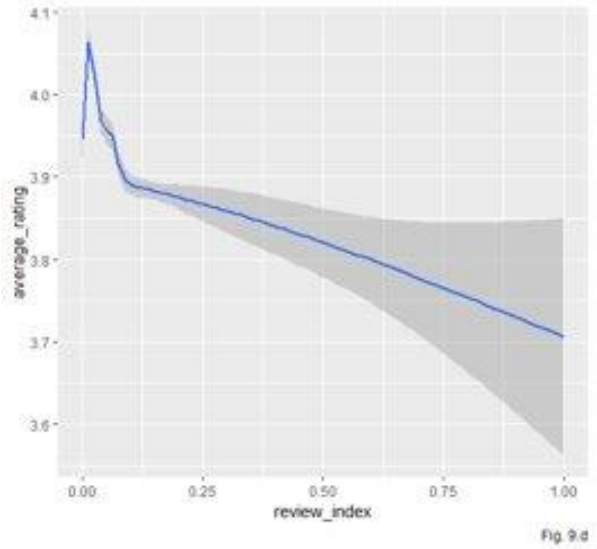


Fig. 9.d

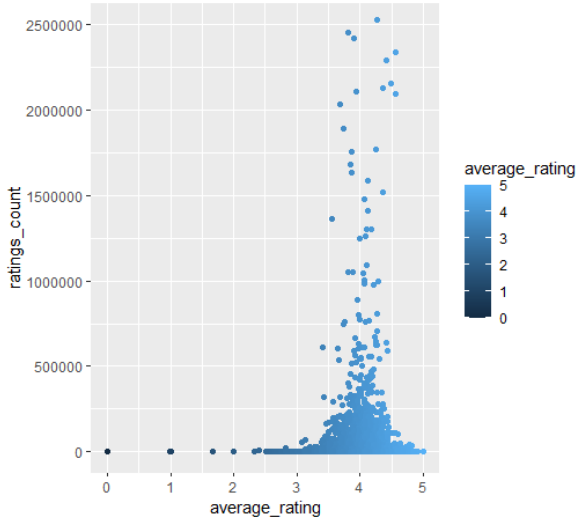


Fig. 9.a

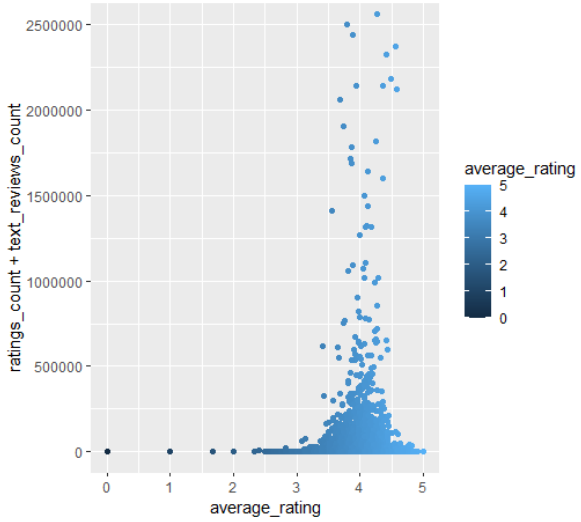


Fig. 9.c

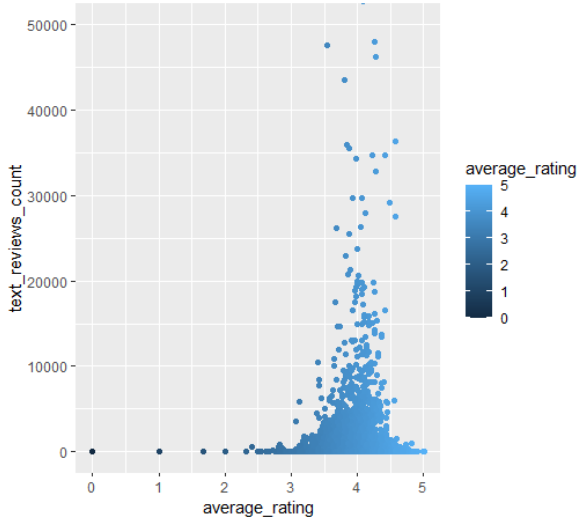


Fig. 9.b

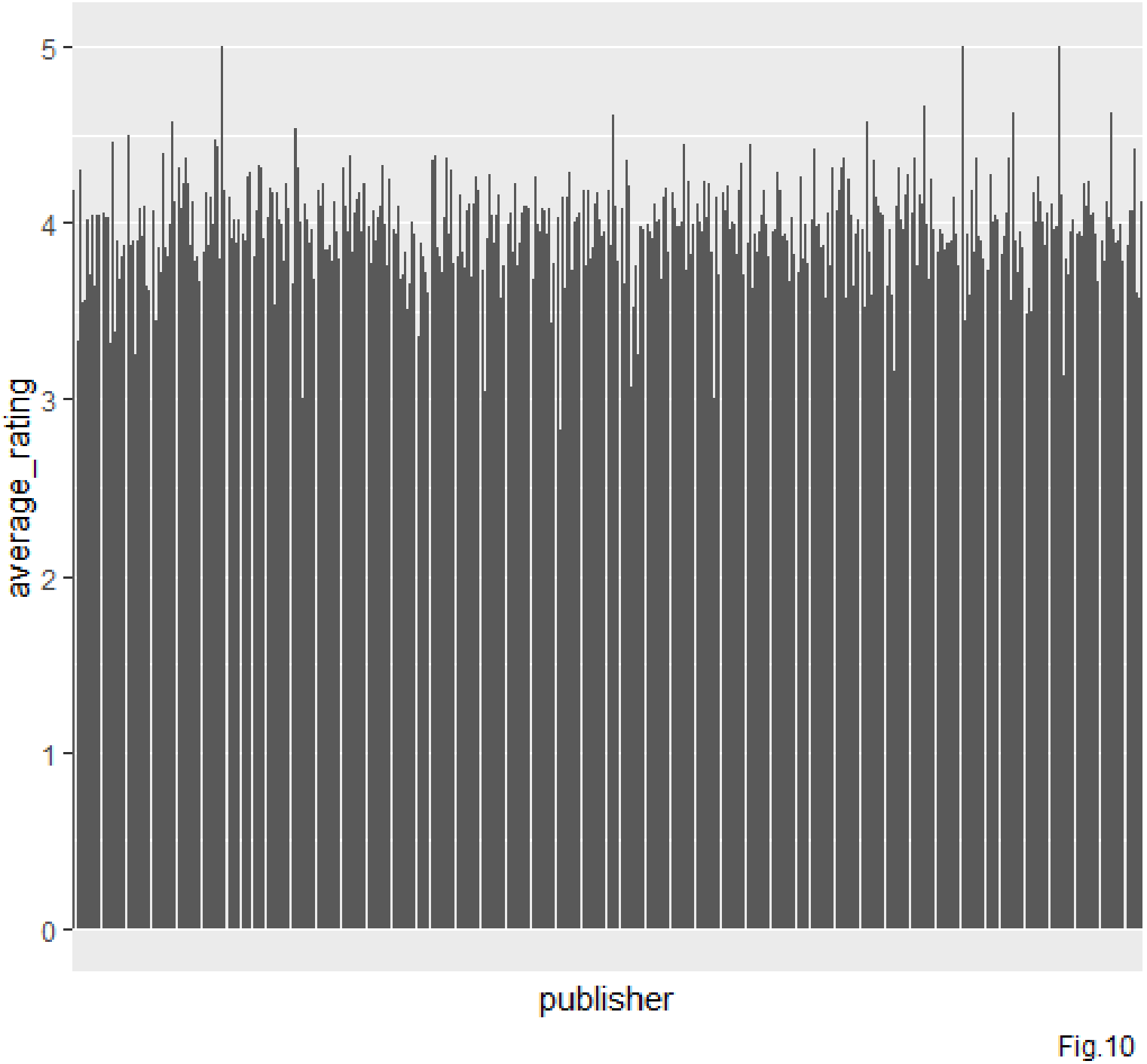


Fig.10

(10) How average rating is differed for different publishers

There are 2290 publishers published the books . Out of 2290 publishers 47 publishers got average rating greater than 4.5 , 2190 got in the range [3.5,4.5] , and the remaining 134 publishers got rating less than 3.5.

(11) Average number of pages for different languages

- Among the languages, the most used language is ‘eng ’ . And we can see that books with “eng” language have 300 pages in average . Whereas books written in “enm” has maximum no of average pages (around 1100) . But it is also a point to note that books of most average-rated language “wel” are of total average pages 150 (which is minimum among the used languages)

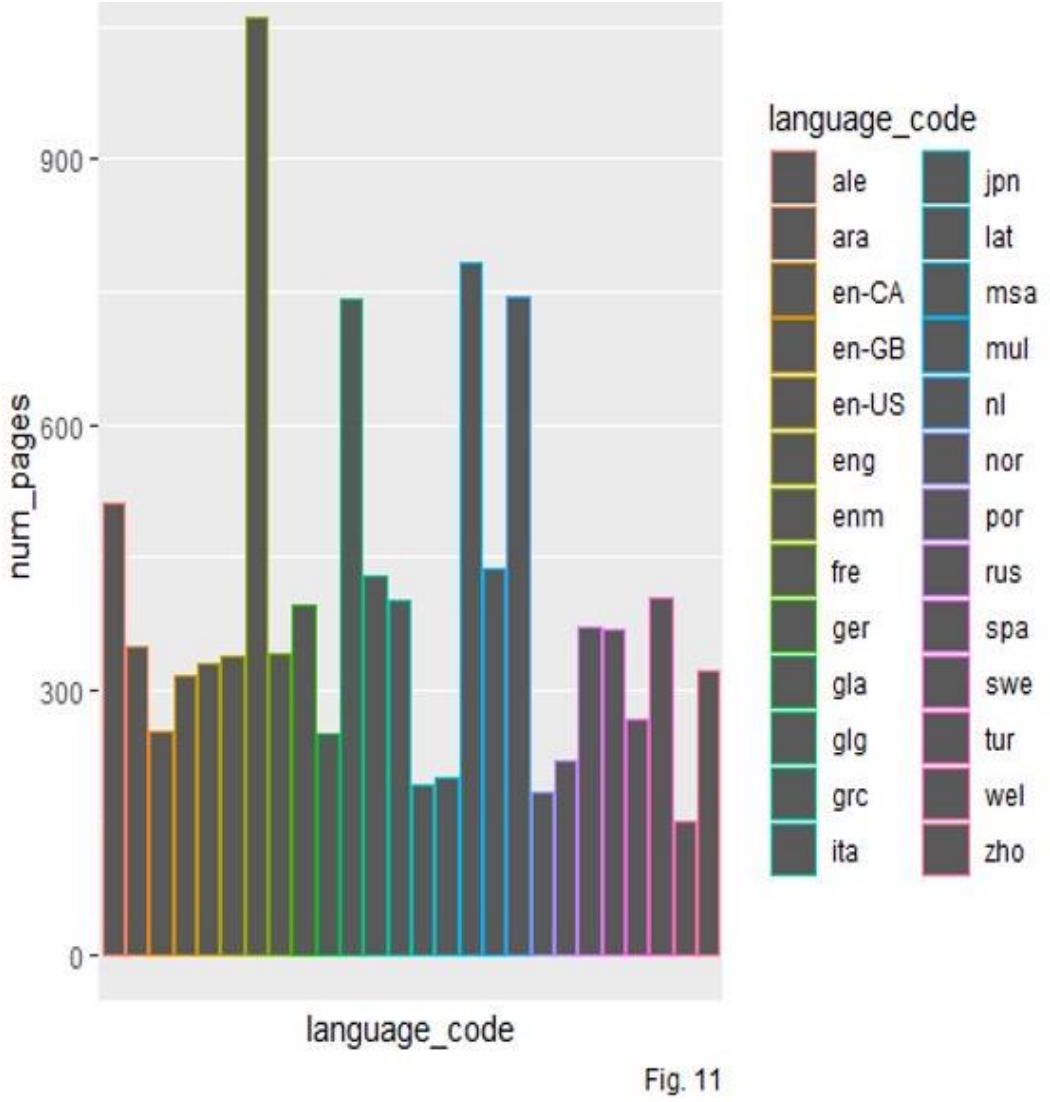


Fig. 11

(12) Number of reviews for different languages

- As the most used language to write a book is “eng”, it has highest total reviews . But also it has a highest number of reviews which is far more from others . So we can conclude that most of the people who read books of ‘eng’ language used to give reviews after reading it .

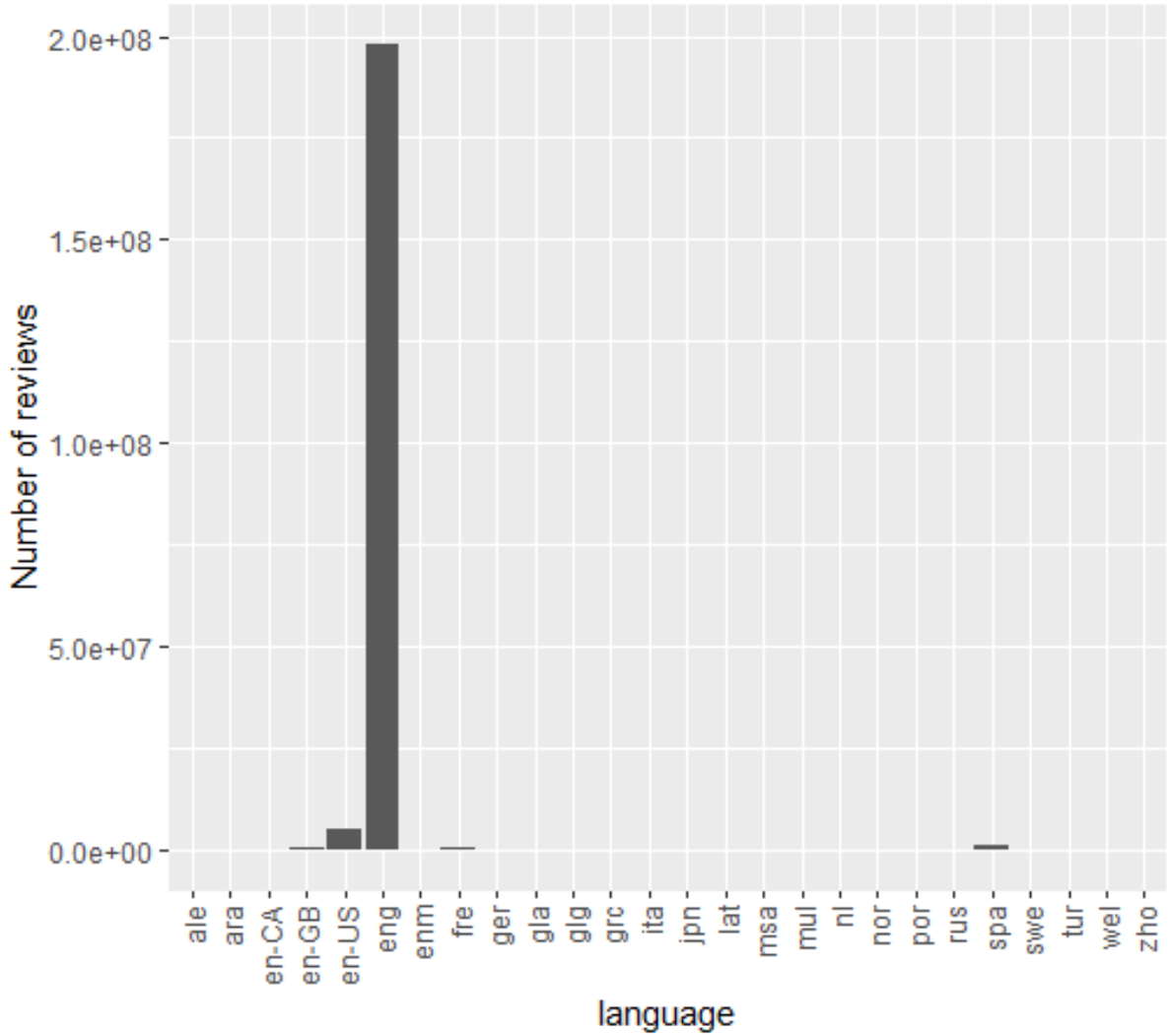


Fig. 12

(13) Dependency between Number of pages and Review parameters

- Books are reviewed most for the books having pages less than 1000. Number of reviews for books having more number of pages are very low.

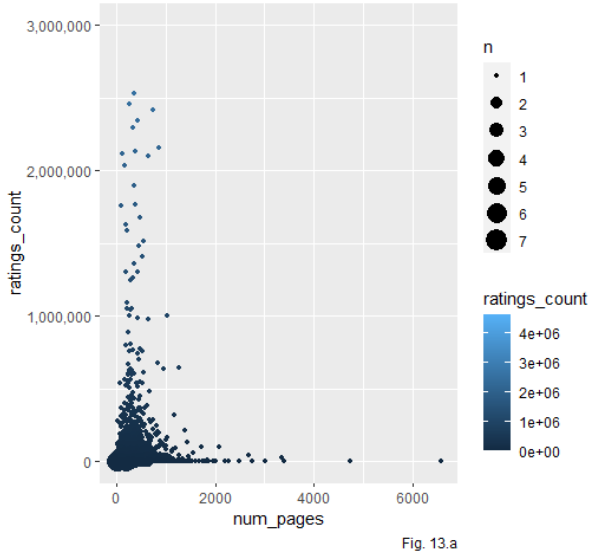


Fig. 13.a

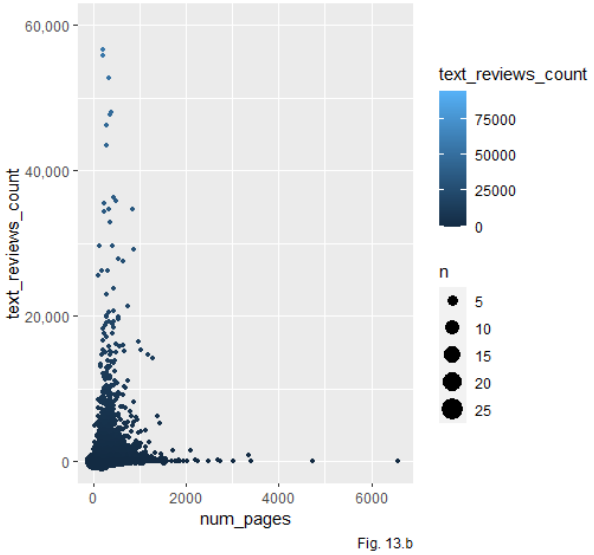


Fig. 13.b

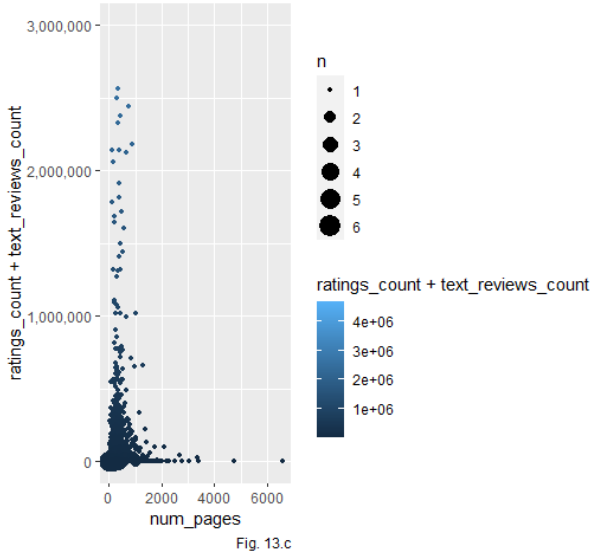


Fig. 13.c

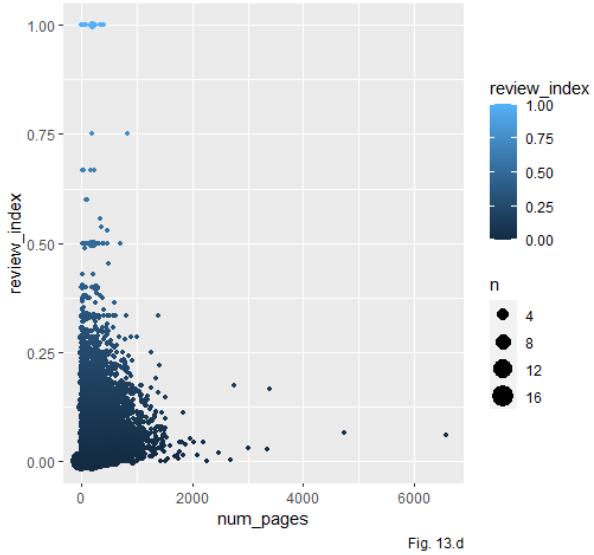
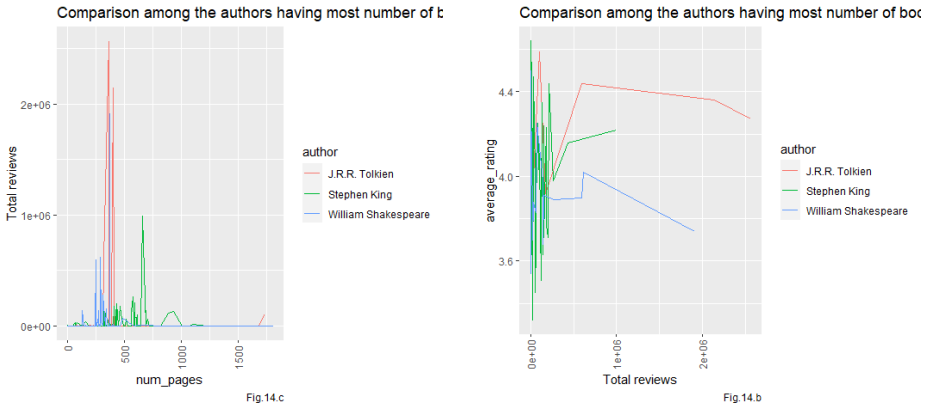


Fig. 13.d

(14) Comparison between top three authors having most number of books

- Initially averae rating decreases fore more number of pages ,but later on it increases for all 3 author.
- For low number of reviews all the authors has a mixed average rating. But when total number of reviews increases, average rating decreased for J.R.R. Tolkein and William Shakespeare but increases for Stephan King.
- All three author has more reviews in pages between 250 to 750.



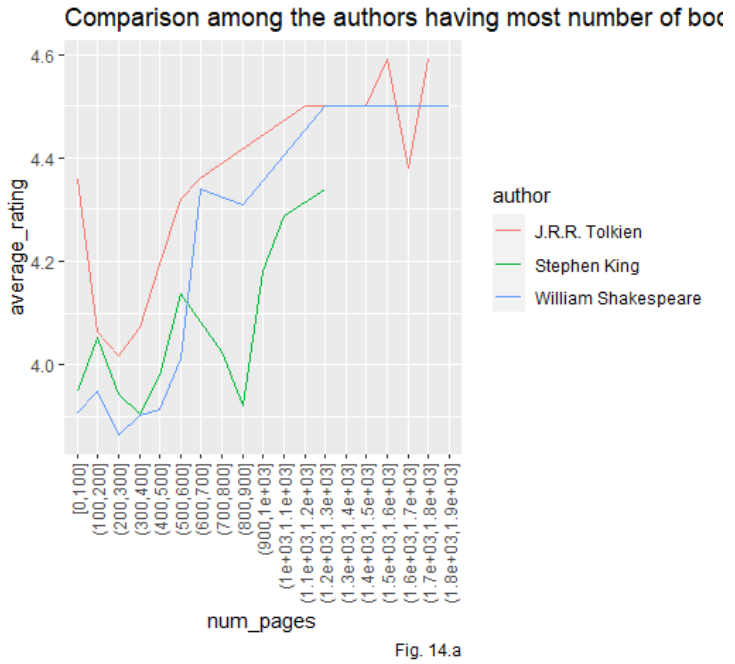


Fig. 14.a

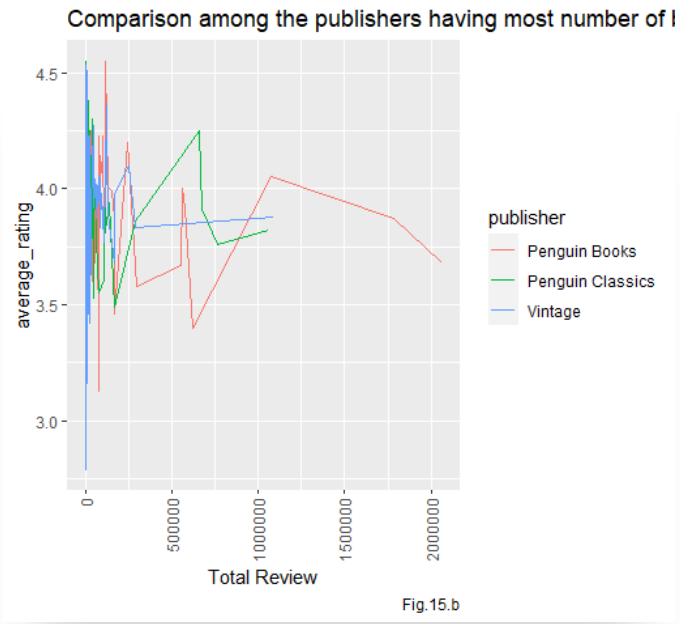


Fig.15.b

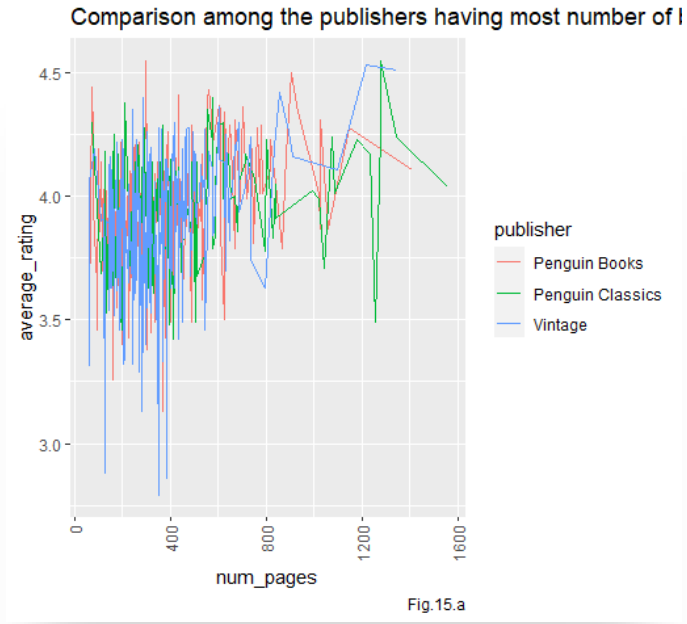


Fig.15.a

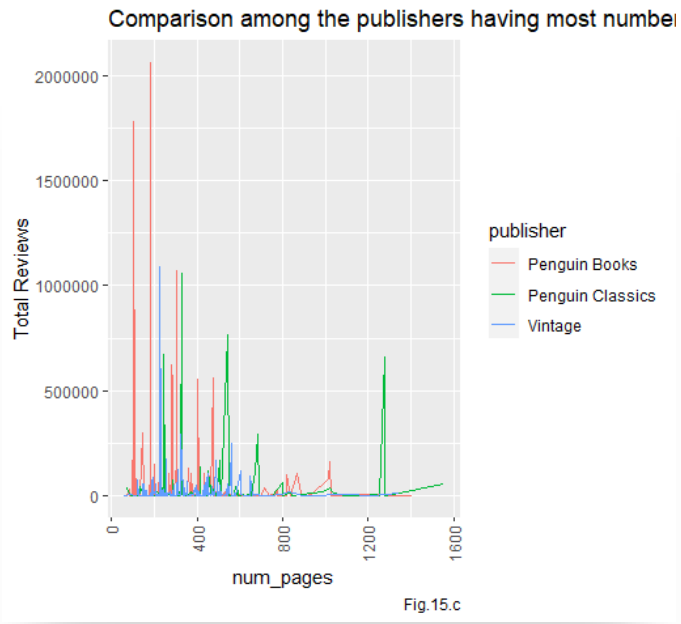
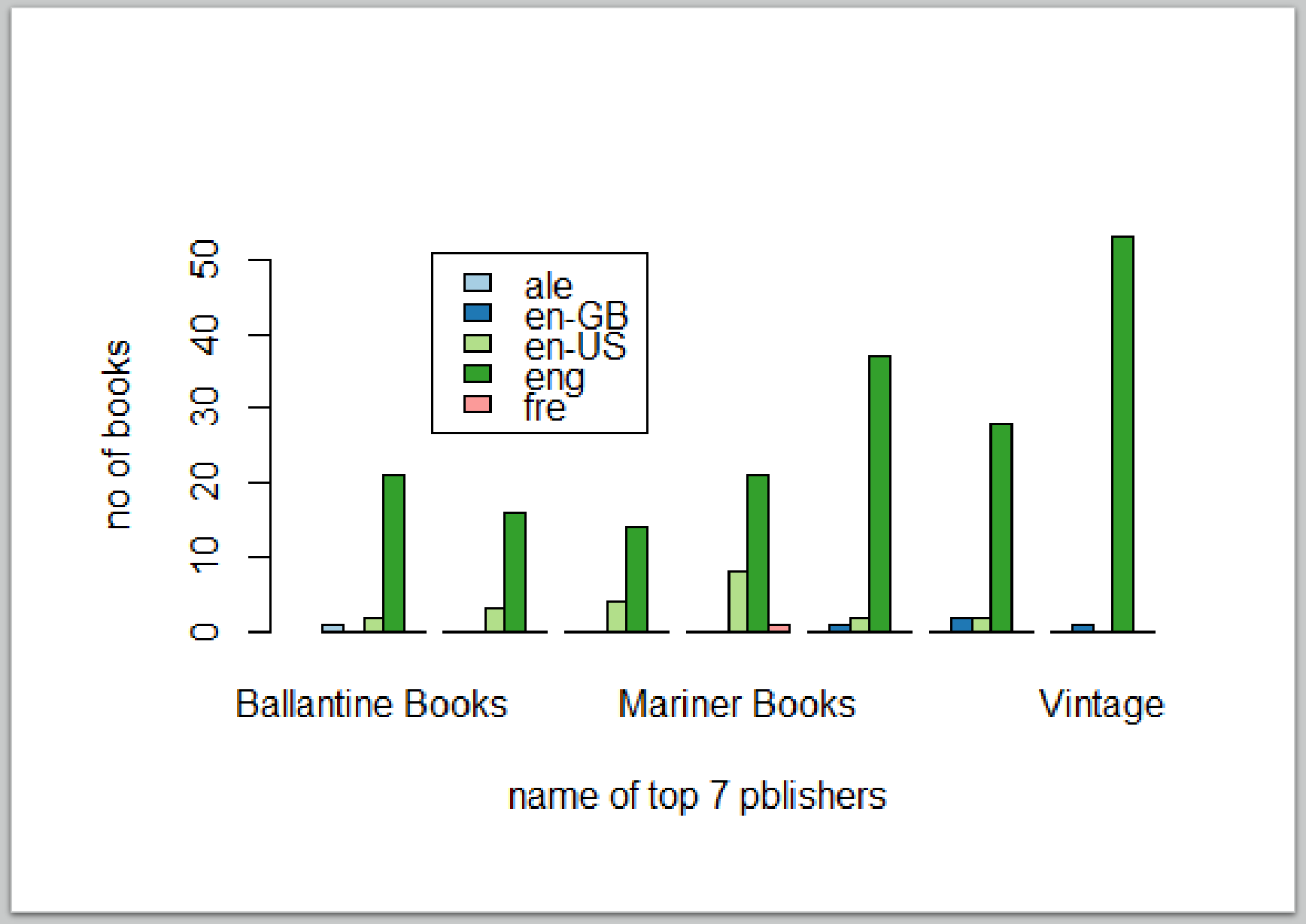


Fig.15.c

(15) Comparison between top three publishers

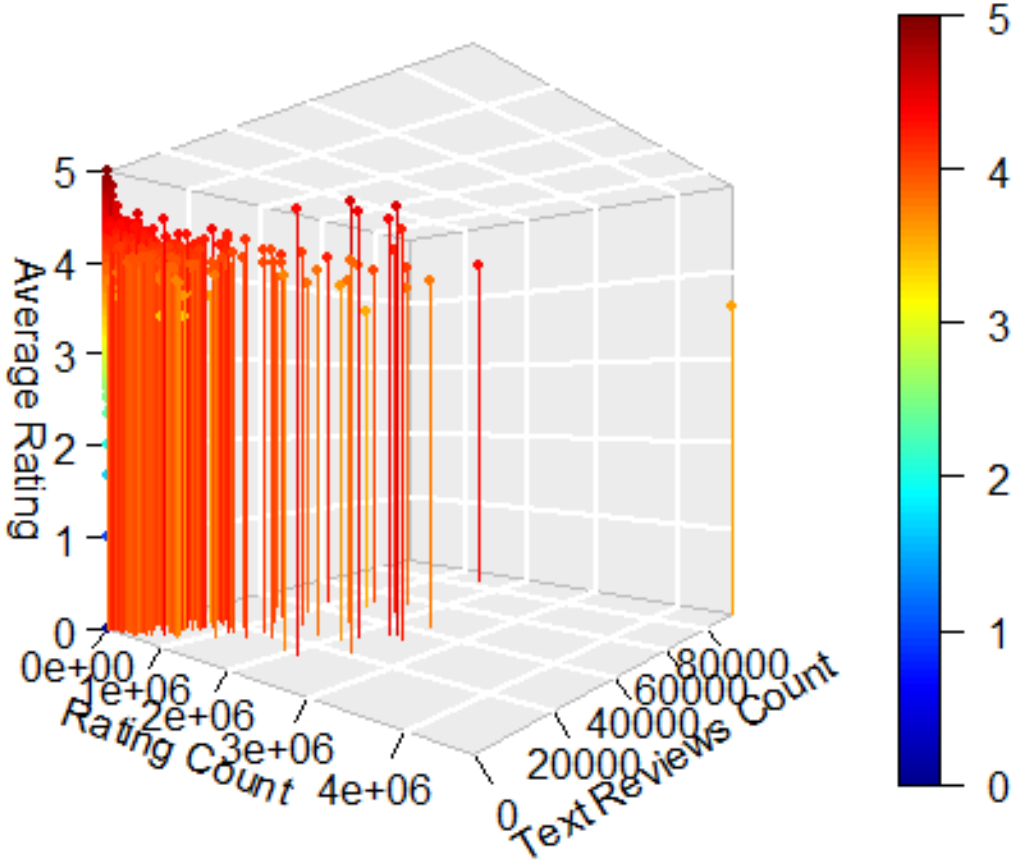
- For higher number of reviews, average ratings are below 4 for these publishers
- Average rating has not such pattern for different number of pages
- For hiher number of paes , total reviews decreases for all publishers

(16) Books published by top publishers in different languages



Average Rating for different Rating count and Text reviews count

- Average ratings are mostly between 4 and 4.5
- Less number of points for large values of Ratings count and text review count



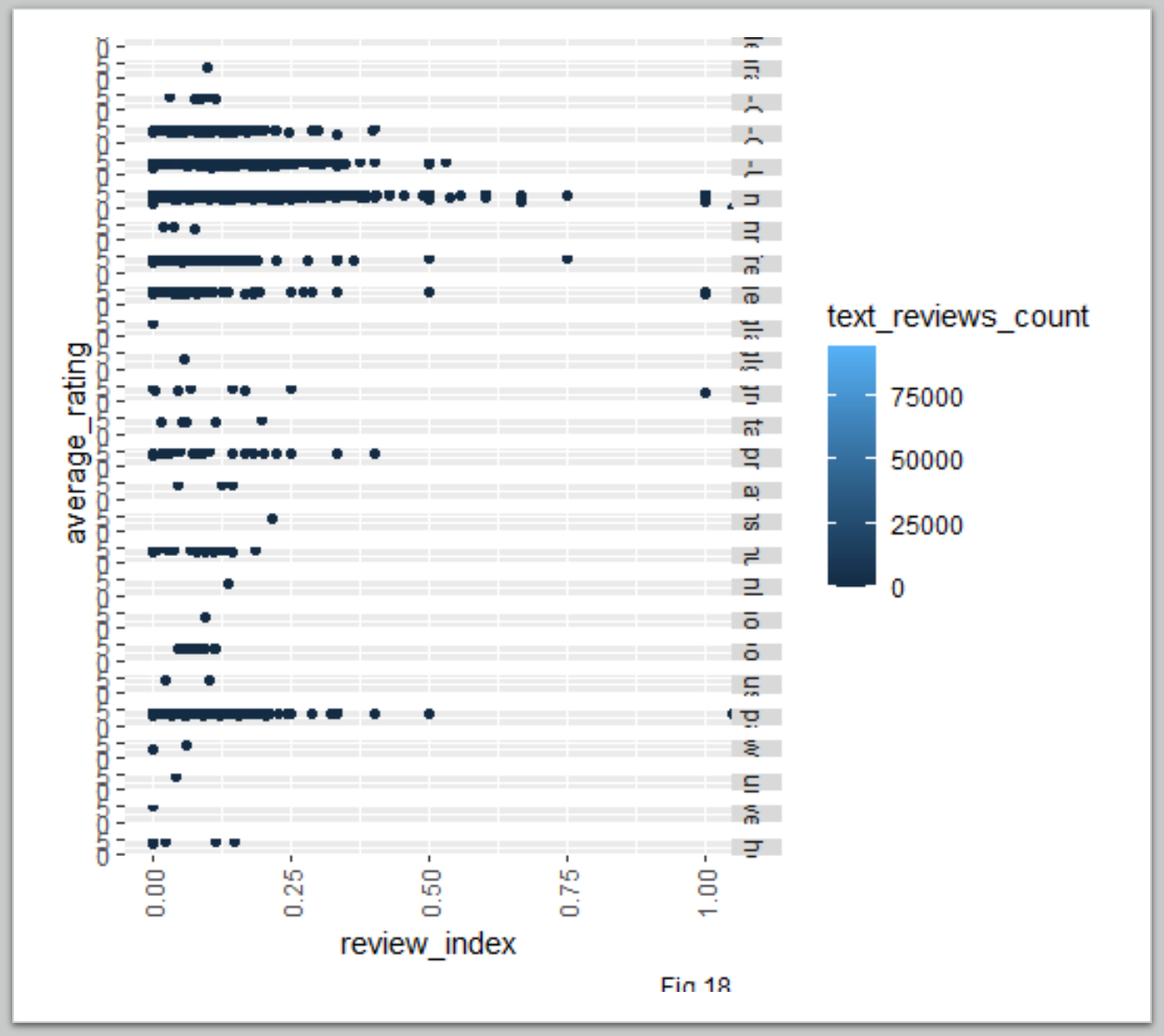


Fig 18

Dependencies of review index and average rating for different languages and text review count

Review index mostly lie between 0-0.25 for most of the languages. For few, it goes above 0.5.

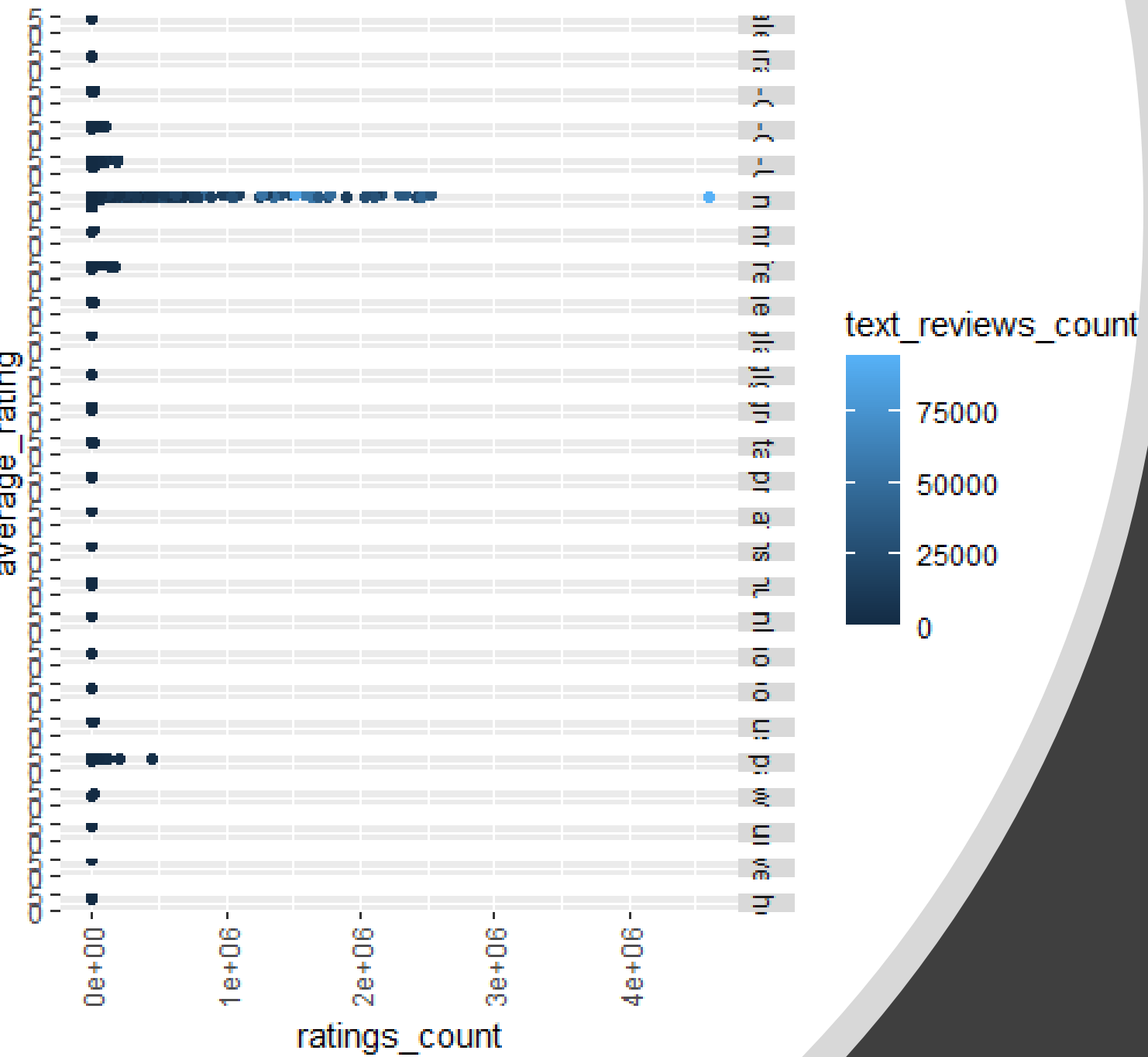


Fig.19

Dependencies of ratings count and average rating for different languages and text review count

Conclusion

Good average ratings are for the books having page ranges 250-750

Number of reviews are mostly for the books written in english. Probably because of the reader of the language are more than others.Also most books in the dataset are written in this language.

Review index (ratio between text reviews and ratings count) are mostly below 0.25. So only a quarter of the reviewer gives a text review.

For higher review index, average rating decreases.

Appendix-1

R codes

```
#Visualizing different factors for having a good review of books

#

#_____________________________________________________________________________###


library(ggplot2)

library(dplyr)


#Read data from file

books<-read.csv("C:\\Users\\kabir\\OneDrive\\Desktop\\pds_visualization\\books.csv")

books<-na.omit(books)


#datatype conversion

books$language_code<-as.factor(books$language_code)

books$average_rating<-as.numeric(books$average_rating)

books$num_pages<-as.integer(books$num_pages)
```

```
books$ratings_count<-as.integer(books$ratings_count)

books$text_reviews_count<-as.integer(books$text_reviews_count)


#Adding derived column

books<-cbind(books,review_index=books$text_reviews_count/books$ratings_count)


books<-na.omit(books)


#(1) average rating vs. number of books of that rating

ggplot(books, aes(average_rating)) + geom_freqpoly(binwidth=0.01,color = "blue")+

  labs( x = "Average Rating", y = "Number of books",title ="Distribution of average rating in the dataset",caption = "Fig. 1" )


#(2) Number of reviews

#(a) ratings_count distribution

ratingcount.df<-data.frame(table(books$ratings_count))

names(ratingcount.df)<-c("ratings_count","cum_freq")

ratingcount.df$cum_freq<-rev(cumsum(rev(ratingcount.df$cum_freq)))

ggplot(ratingcount.df, aes(x=ratings_count, y=cum_freq)) +  geom_col()+

  labs(x="Numer of Ratings",y="Cummulative Frequency",title="Cummulative frequency(greater than type) plot for Ratings count",caption="Fig. 2.a")+

  scale_x_discrete(breaks = levels(ratingcount.df$ratings_count)[c(T,rep(F,999))])
```

```
#(b) text reviews distribution

treviewcount.df<-data.frame(table(books$text_reviews_count))

names(treviewcount.df)<-c("text_reviews_count","cum_freq")

treviewcount.df$cum_freq<-rev(cumsum(rev(treviewcount.df$cum_freq)))

ggplot(treviewcount.df, aes(x=text_reviews_count, y=cum_freq)) +  geom_col()+

  labs(x="Numer of Text Reviews",y="Cummulative Frequency",title="Cummulative
frequency(greater than type) plot for Number of Text Reviews",caption = "Fig. 2.b")+

  scale_x_discrete(breaks = levels(ratingcount.df$ratings_count)[c(T,rep(F,205))])




#(c) total reviews distribution

reviews.df<-data.frame(table(books$ratings_count+books$text_reviews_count))

names(reviews.df)<-c("reviews_count","cum_freq")

reviews.df$cum_freq<-rev(cumsum(rev(reviews.df$cum_freq)))

ggplot(reviews.df, aes(x=reviews_count, y=cum_freq)) +  geom_col()+

  labs(x="Total Numer of Reviews",y="Cummulative Frequency",title="Cummulative
frequency(greater than type) Total Number of Reviews",caption = "Fig. 2.c")+

  scale_x_discrete(breaks = levels(ratingcount.df$ratings_count)[c(T,rep(F,299))])



#(d) review index distribution

ggplot(books,aes(review_index))+
```

```
  geom_freqpoly(binwidth=0.007,colour="red")+

  labs(x="Review Index",y="Frequency",title="Frequency polygon for Review Index",caption =
"Fig. 2.d")



#(3) Number of books for different languages

lang<-data.frame(table(books$language_code))

lang<-lang[order(lang$Freq,decreasing=T),]

levels(lang$Var1)<-c(levels(lang$Var1),"others")

lang<-rbind(lang %>% top_n(7,lang$Freq),c("others",sum(lang$Freq[8:31],na.rm=T)))

ggplot(lang, aes(x="", y=as.integer(Freq), fill=Var1))+

  geom_bar(width = 1, stat = "identity")+

  coord_polar("y", start=0)+

  labs(x="Languages",y="Number of books",title="Pie Chart for number of books of different
languages",caption = "Fig. 3")



#(4) Number of books of different pages

ggplot(books,aes(num_pages))+

  geom_histogram(binwidth = 5)+

  labs(x="Number of pages",y="Number of Books",title="Histogram of books of different page
numbers",caption = "Fig. 4")+

  coord_cartesian(xlim = c(0,2000),ylim = c(0,350))
```

```
#(5) number of authers having exactly certain number of books barplot


# creates a dataframe with number of authors having n number of books

author<-data.frame(table(table(unlist(strsplit(books$authors,split = "/")))))

names(author)<-c("no_of_books","no_of_authors")

ggplot(author, aes(no_of_books, no_of_authors)) +

  geom_col() +

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  scale_x_discrete(breaks=levels(author$no_of_books)[c(T,rep(F,4))])+

  labs(x="Number of Books",y="Number of authors",title="Number of Authors having n Number
of Books",caption = "Fig. 5")+

  coord_cartesian(xlim = c(0,25))+

  scale_fill_brewer(palette = "Blues")



#(6) Book published in different years

pubdate<-substr(books$publication_date, nchar(books$publication_date)-4+1,
nchar(books$publication_date))

pubdate<-as.integer(pubdate)

pubdate<-pubdate[pubdate>=1900]

pubdate<-data.frame(table(pubdate))

ggplot(data=pubdate, aes(x=pubdate, y=Freq)) +

  geom_bar(stat = 'identity') +
```

```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  scale_x_discrete(breaks=levels(pubdate)[T,rep(F,9)])+

  scale_fill_brewer(palette = "Blues")+

  labs(x="Publication year",y="Number of Books",title = "Books published in different
year",caption = "Fig.6")
```

```
# (7) Language vs average rating


avg.lang<-aggregate(average_rating~language_code, data=books, FUN = mean)

ggplot(avg.lang, aes(language_code, average_rating)) +

  geom_col(aes(colour=language_code))+

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  labs(title="Average rating for different languages",caption="Fig. 7")
```

```
#(8) number of pages vs average ratings
```

```
ggplot(books,aes(num_pages,average_rating))+
  geom_rug(aes(colour="red"))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(title = "Rug plot: Average Rating vs. Number of Pages",caption = "Fig. 8")



#(9) Reviews vs average rating
#(a) rating count vs average rating
ggplot(books,aes(average_rating,ratings_count))+
  geom_jitter(aes(colour=average_rating))+
  labs(caption="Fig. 9.a")+
  coord_cartesian(ylim = c(0,2500000))



#(b) text reviews vs average rating
ggplot(books,aes(average_rating,text_reviews_count))+
  geom_jitter(aes(colour=average_rating))+
  labs(caption="Fig. 9.b")+
  coord_cartesian(ylim = c(0,50000))
```

```
#(c) Total reviews vs average rating
ggplot(books,aes(average_rating,ratings_count+text_reviews_count))+
  geom_jitter(aes(colour=average_rating))+
  labs(caption="Fig. 9.c")+
  coord_cartesian(ylim = c(0,2500000))



#(d) review_index vs average rating
ggplot(books,aes(review_index,average_rating))+
  geom_smooth(aes(colour=review_index))+
  labs(caption="Fig. 9.d")

#(10) average ratings for different publishers
avg.pub<-aggregate(average_rating~publisher, data=books, FUN = mean)
ggplot(avg.pub, aes(publisher,average_rating)) +
  geom_col()+
  scale_x_discrete(breaks=NULL)+
  labs(caption = "Fig.10 ")
```

```
#(11) pages per book for different languages

pagesperbook<-aggregate(num_pages~language_code,data=books, FUN=mean)

ggplot(pagesperbook,aes(language_code,num_pages))+

  geom_col(aes(colour=language_code))+

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  scale_x_discrete(breaks=NULL)+

  labs(caption="Fig. 11")
```

```
#(12) total reviews for different languages

reviews.lang<-aggregate(ratings_count+text_reviews_count~language_code,data=books,
FUN=sum)

names(reviews.lang)<-c("language","reviews")

ggplot(reviews.lang,aes(x=language,y=as.integer(reviews)))+

  geom_bar(stat = "identity")+

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  labs(y="Number of reviews",caption = "Fig. 12")
```

```
#(13) number of pages vs  reviews

#(a) pages vs rating count

ggplot(books,aes(num_pages,ratings_count))+

  geom_count(aes(colour=ratings_count))+

  scale_y_continuous( labels = scales::comma)+

  coord_cartesian(ylim = c(0,3000000))+

  labs(caption="Fig. 13.a")


#(b) pages vs text reviews count

ggplot(books,aes(num_pages,text_reviews_count))+

  geom_count(aes(colour=text_reviews_count))+

  scale_y_continuous( labels = scales::comma)+

  coord_cartesian(ylim = c(0,60000))+

  labs(caption="Fig. 13.b")



#(c) total reviews vs number of pages

ggplot(books,aes(num_pages,ratings_count+text_reviews_count))+

  geom_count(aes(colour=ratings_count+text_reviews_count))+

  scale_y_continuous( labels = scales::comma)+

  coord_cartesian(ylim = c(0,3000000))+
```

```
  labs(caption="Fig. 13.c")



#(d) review_index vs number of pages

ggplot(books,aes(num_pages,review_index))+

  geom_count(aes(colour=review_index))+

  scale_y_continuous( labels = scales::comma)+

  labs(caption="Fig. 13.d")




#(14) 3 authors having most number of books

book.author<-data.frame(table(unlist(strsplit(books$authors,split = "/"))))

book.author<-book.author%>% slice_max(Freq,n=3)


  book.author.df<-books[grep(as.character(book.author$Var1[1]),books$authors),]

  book.author.df$num_pages<-cut_width(book.author.df$num_pages,100,boundary=0)

  book.author.df<-aggregate(average_rating~num_pages,data = book.author.df,mean)

  author<-rep(book.author$Var1[1],nrow(book.author.df))

  book.author.df1<-cbind(book.author.df,author)
```

```
 book.author.df<-books[grep(as.character(book.author$Var1[2]),books$authors),]
book.author.df$num_pages<-cut_width(book.author.df$num_pages,100,boundary=0)
book.author.df<-aggregate(average_rating~num_pages,data = book.author.df,mean)
author<-rep(book.author$Var1[2],nrow(book.author.df))
book.author.df2<-cbind(book.author.df,author)


book.author.df<-books[grep(as.character(book.author$Var1[3]),books$authors),]
book.author.df$num_pages<-cut_width(book.author.df$num_pages,100,boundary=0)
book.author.df<-aggregate(average_rating~num_pages,data = book.author.df,mean)
author<-rep(book.author$Var1[3],nrow(book.author.df))
book.author.df3<-cbind(book.author.df,author)

book.author.df<-rbind(rbind(book.author.df1,book.author.df2),book.author.df3)
#(a) number of pages vs average ratings
ggplot(book.author.df,aes(num_pages,average_rating))+
  geom_line(aes(group = author,colour=author))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
```

```
   labs(title = "Comparison among the authors having most number of books",caption = "Fig. 14.a")



  #(b) total reviews vs average rating for top 3 author


# book.author.review<-
books[c(grep(as.character(book.author$Var1[1]),books$authors),grep(as.character(book.author$Var1[2]),books$authors),grep(as.character(book.author$Var1[3]),books$authors)),]

  book.author.review.df1<-books[grep(as.character(book.author$Var1[1]),books$authors),]

  book.author.review.df2<-books[grep(as.character(book.author$Var1[2]),books$authors),]

  book.author.review.df3<-books[grep(as.character(book.author$Var1[3]),books$authors),]

  author1<-rep(book.author$Var1[1],nrow(book.author.review.df1))

  author2<-rep(book.author$Var1[2],nrow(book.author.review.df2))

  author3<-rep(book.author$Var1[3],nrow(book.author.review.df3))


  book.1<-cbind(book.author.review.df1,author=author1)

  book.2<-cbind(book.author.review.df2,author=author2)

  book.3<-cbind(book.author.review.df3,author=author3)


  book.author.review<-rbind(rbind(book.1,book.2),book.3)
```

```
  ggplot(book.author.review,aes(ratings_count+text_reviews_count,average_rating))+

    geom_line(aes(group = author,colour=author))+

    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

    labs(x="Total reviews",caption = "Fig.14.b",title ="Comparison among the authors having most number of books" )

#(c) total reviews vs number of pages

  ggplot(book.author.review,aes(num_pages,ratings_count+text_reviews_count))+

    geom_line(aes(group = author,colour=author))+

    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

    labs(y="Total reviews",caption = "Fig.14.c",title ="Comparison among the authors having most number of books" )




#(15) For top 3 publishers

  publishers.top<-data.frame(table(books$publisher))

publishers.top<-publishers.top %>% slice_max(Freq,n=3)

pub.top<-as.character(publishers.top$Var1)

publishers.top.df1<-books[books$publisher==pub.top[1] | books$publisher==pub.top[2] | books$publisher==pub.top[3] ,]



# (a) number of pages vs average ratings

ggplot(publishers.top.df1,aes(num_pages,average_rating))+
```

```
  geom_line(aes(group = publisher,colour=publisher))+

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  labs(caption = "Fig.15.a",title ="Comparison among the publishers having most number of
books" )



# (b) total reviews vs average ratings

ggplot(publishers.top.df1,aes(ratings_count+text_reviews_count,average_rating))+

  geom_line(aes(group = publisher,colour=publisher))+

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  labs(x="Total Review",caption = "Fig.15.b",title ="Comparison among the publishers having
most number of books" )



#(c) number of pages vs total number of reviews

ggplot(publishers.top.df1,aes(num_pages,ratings_count+text_reviews_count))+

  geom_line(aes(group = publisher,colour=publisher))+

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+

  labs(y="Total Reviews",caption = "Fig.15.c",title ="Comparison among the publishers having
most number of books" )



#(16)

pub<-data.frame(table(books$publisher))

pub<-pub[order(pub$Freq,decreasing=T),]

pub7 <-pub %>% top_n(7,pub$Freq)
```

```
top7publisher  <-as.vector(pub7$Var1)

lang.pub <-table(books[books$publisher==top7publisher,c(7,12)])

barplot(lang.pub,1,  beside = T,legend.text= rownames(lang.pub),col =blues9,args.legend  =
list(x=ncol(lang.pub)+350,y=50))

library(RColorBrewer)

barplot(lang.pub,beside=T,xlim=
c(0,ncol(lang.pub)+300),col=brewer.pal(nrow(lang.pub),"Paired"),ylab="no of books",xlab=
"name of top 7 pblishers",legend.text= T,args.legend= list(x=ncol(lang.pub)+370))



#(17)

library(plot3D)

rating_count<-books$ratings_count

text_reviews_count<-books$text_reviews_count

average_rating<-books$average_rating

scatter3D(rating_count,text_reviews_count,average_rating,xlab="Rating Count",ylab="Text
Reviews Count",zlab="Average Rating",pch = 19,  bty = "g",  type = "h", phi = 0,ticktype =
"detailed",cex=0.5)




#(18)


ggplot(books,aes(review_index,average_rating,colour=text_reviews_count))+

  geom_jitter()+

  facet_grid(vars(language_code))+
```

```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(caption = "Fig.18" )+
  scale_y_continuous(n.breaks = 2)



#(19)
ggplot(books,aes(ratings_count,average_rating,colour=text_reviews_count))+
  geom_point()+
  facet_grid(vars(language_code))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(caption = "Fig.19" )+
  scale_y_continuous(n.breaks = 2)


################################################################################
##############################################

################################################################################
##############################################
```

Appendix-2

Different type of plots

A.one variable continuous : Fig. 1, Fig. 2.a, Fig. 2.b , Fig 2.c, Fig 2.d , Fig. 4

B. One variable discrete (categorical data) : Fig. 3 , Fig 5, Fig. 6

C. Two variables (both continuous) : Fig. 8, Fig. 9.d

D. Two variables (one discrete and one continuous) : Fig. 7, Fig. 9.a,Fig. 9.b, Fig. 9.c,Fig 10,Fig. 13.d

E. Two variables (both discrete) : Fig. 11, Fig. 12,Fig. 13.a,Fig. 13.b,Fig. 13.c,Page-16

F. Three variable : Fig. 14.a,Fig. 14.b, Fig. 14.c,Fig. 14.d,Fig.15.a,Fig. 15.b,Fig., 15.c,Page-17

G. Four Variable : Page-18,Page-19