Zeru-Zhou-project07

October 18, 2021

1 Project 7 – Zeru Zhou

TA Help: NA

Collaboration: NA

- Get help from Dr.Ward's video

1.1 Question 1

[1]:
```
library(data.table)
```

[2]:
```
books <- fread("/depot/datamine/data/goodreads/csv/goodreads_books.csv")
```

[7]:
```
head(books)
```

A data.table: 6 x 26

isbn <chr>	author_id <int>	text_reviews_count <int>	country_code <chr>	language_code <chr>	asin <chr>
0312853122	604031	1	US		
0743509986	626222	6	US		
	10333	7	US	eng	B00071I
0743294297	9212	3282	US	eng	
0850308712	149918	5	US		
1599150603	3041852	7	US		

[8]:
```
sort(tapply(books$average_rating, books$publication_month, mean, na.rm=T),␣
↪decreasing=T)
```

11 3.89601935646053 **10** 3.89468731483113 **12** 3.88409020780097 **9** 3.87714531536457 **7** 3.87555508189365 **3** 3.87409088400331 **6** 3.87360441984598 **8** 3.87124292174196 **5** 3.86889534991733 **4** 3.86650617328914 **2** 3.86366476090338 **1** 3.84461825591151 **25** 0

As a result, average rating is highest in November and lowest in January, according the table listed above. I suggest Dr. Ward publish his work in November because the rating (3.896) is relative higher than the other months across the whole year.

1.2 Question 2

[4]:
```
summary(books$num_pages)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
    0.0    147.0    245.0    264.3    344.0 945077.0   323929
```

[5]:
```
books$book_size_cat <- cut(books$num_pages, breaks= c(0, 100, 400, Inf),␣
↪include.lowest=T, useNA="always", labels= c("small","medium","large"))
```

[6]:
```
table(books$book_size_cat, useNA= "always")
```

```
 small medium  large   <NA>
110311 466681  99079 323929
```

In above, I use Dr.Ward's break period to justify my method is correct. Then I'm gonna use my break period.

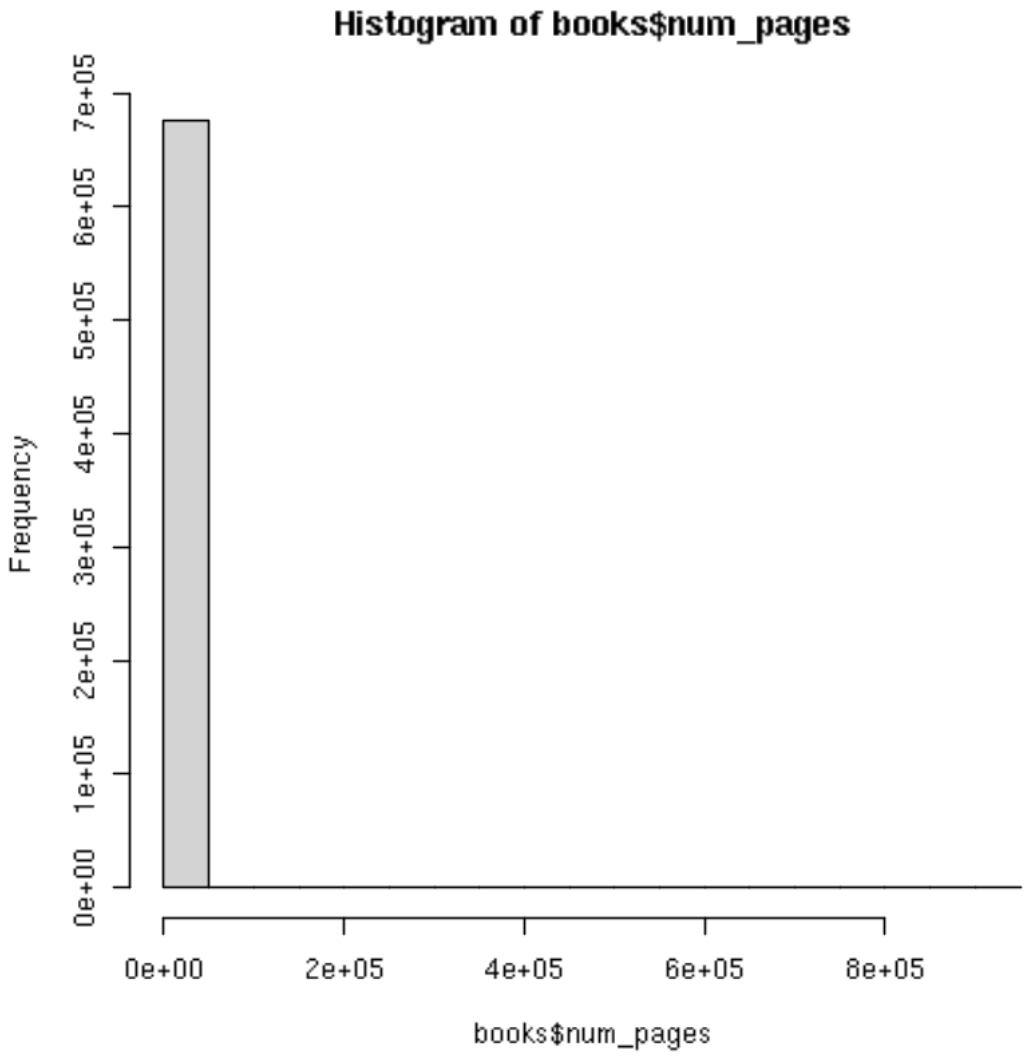
[4]:
```
books$book_size_cat <- cut(books$num_pages, breaks= c(0, 145, 345, Inf),␣
↪include.lowest=T, useNA="always", labels= c("small","medium","large"))
```

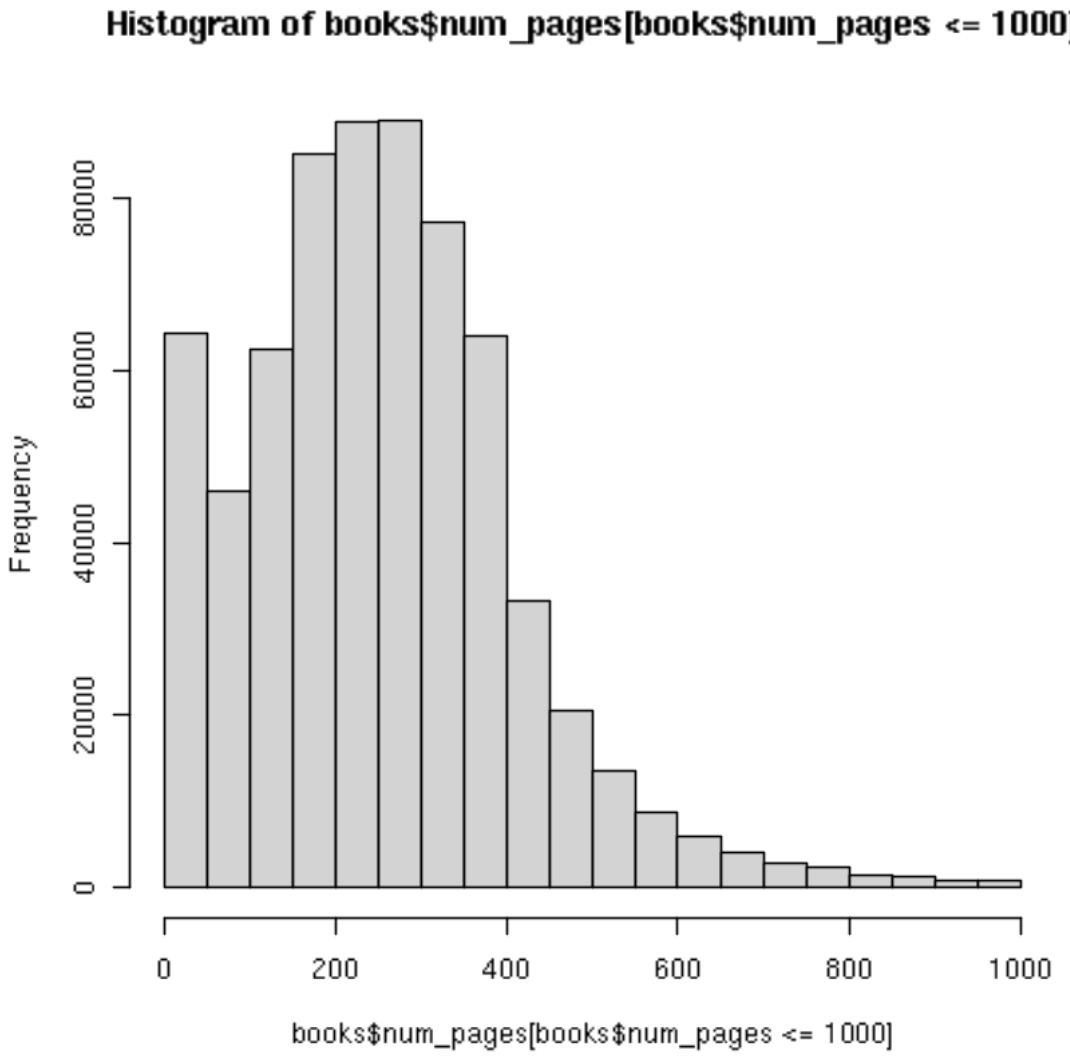
[9]:
```
table(books$book_size_cat,useNA="always")
```

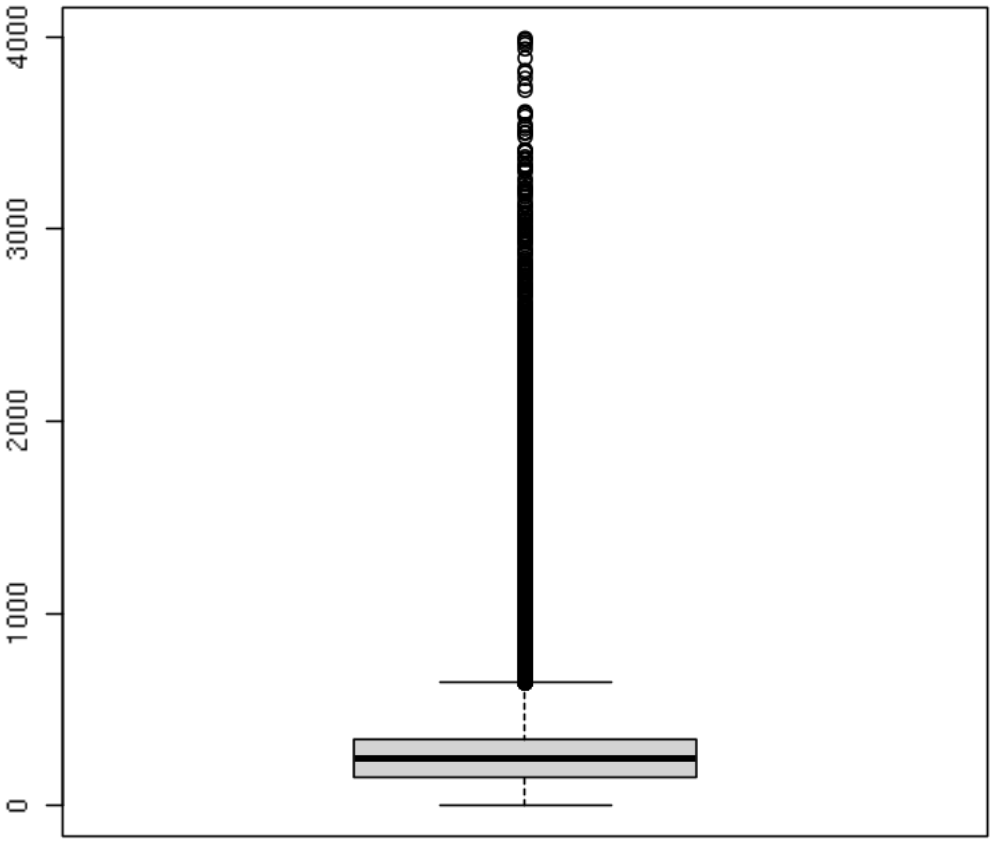
```
 small medium  large   <NA>
167855 340389 167827 323929
```

[17]:
```
summary(books$num_pages)
hist(books$num_pages)
hist(books$num_pages[books$num_pages <= 1000])
boxplot(books$num_pages[books$num_pages < 4000])
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
    0.0    147.0    245.0    264.3    344.0 945077.0   323929
```


Histogram of books$num_pages
Frequency
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0e+00
2e+05
4e+05
6e+05
8e+05
books$num_pages


Histogram of books$num_pages[books$num_pages <= 1000]
Frequency
0
20000
40000
60000
80000
0
200
400
600
800
1000
books$num_pages[books$num_pages <= 1000]



I pick [0,145] as small since the first quarter is 147, so I just find a number 145 near it to define as "small". Then, I pick [145,345] as medium since the third quarter is 344, I find a number 345 near it to divide the boundry between medium and large. Then, if the number id greater than 345, it is defined as large because it is larger than 3/4 of the data. The result of running "table(books$book_size_cat)" is listed above.

1.3 Question 3

[10]:
```
tapply(books$text_reviews_count, books$book_size_cat, mean)
```

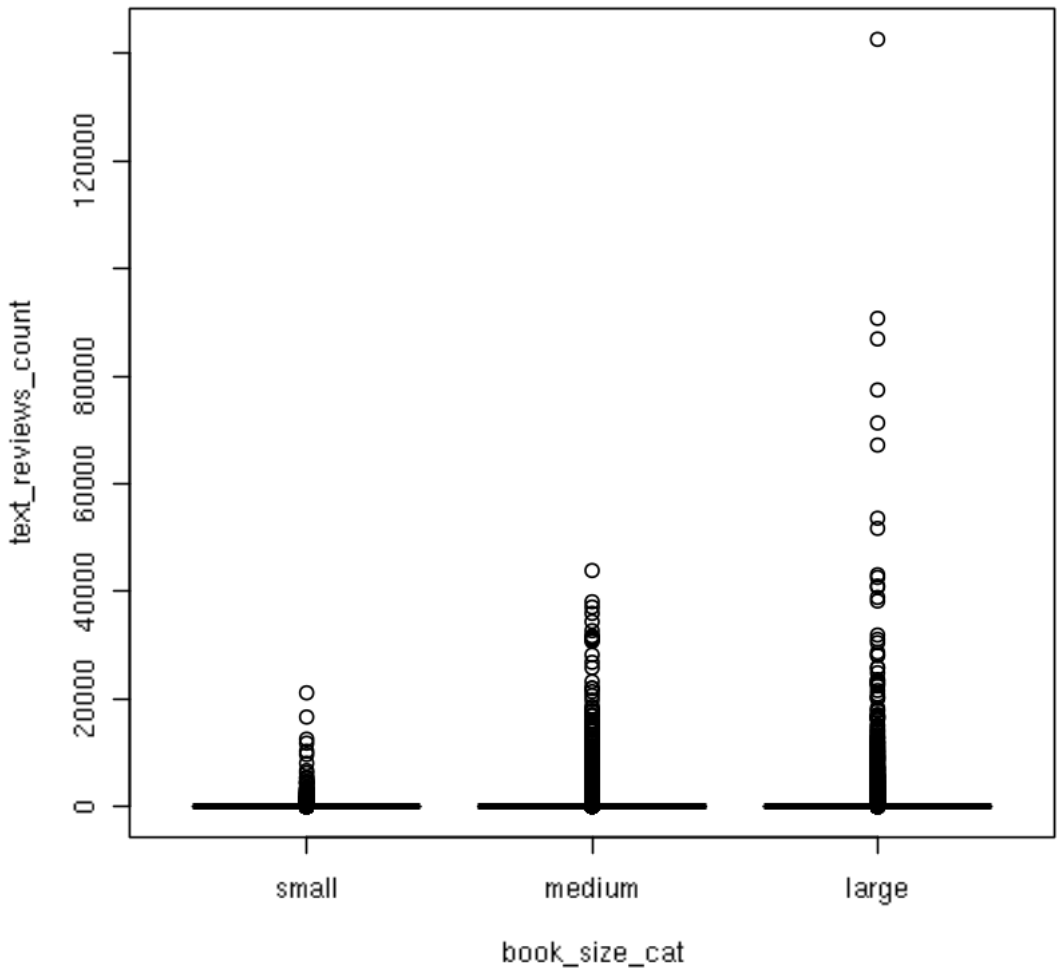
small 15.2653242381818 **medium** 34.1107438842031 **large** 59.5537905104661

As the table above, we see that text reviews are far more in category "large" than in "medium" or "small". Therefore, as a firm believer in feedback, Dr. Ward should consider about large size as

book size, which is more than 345 pages in my division in last problem.

1.4 Question 4

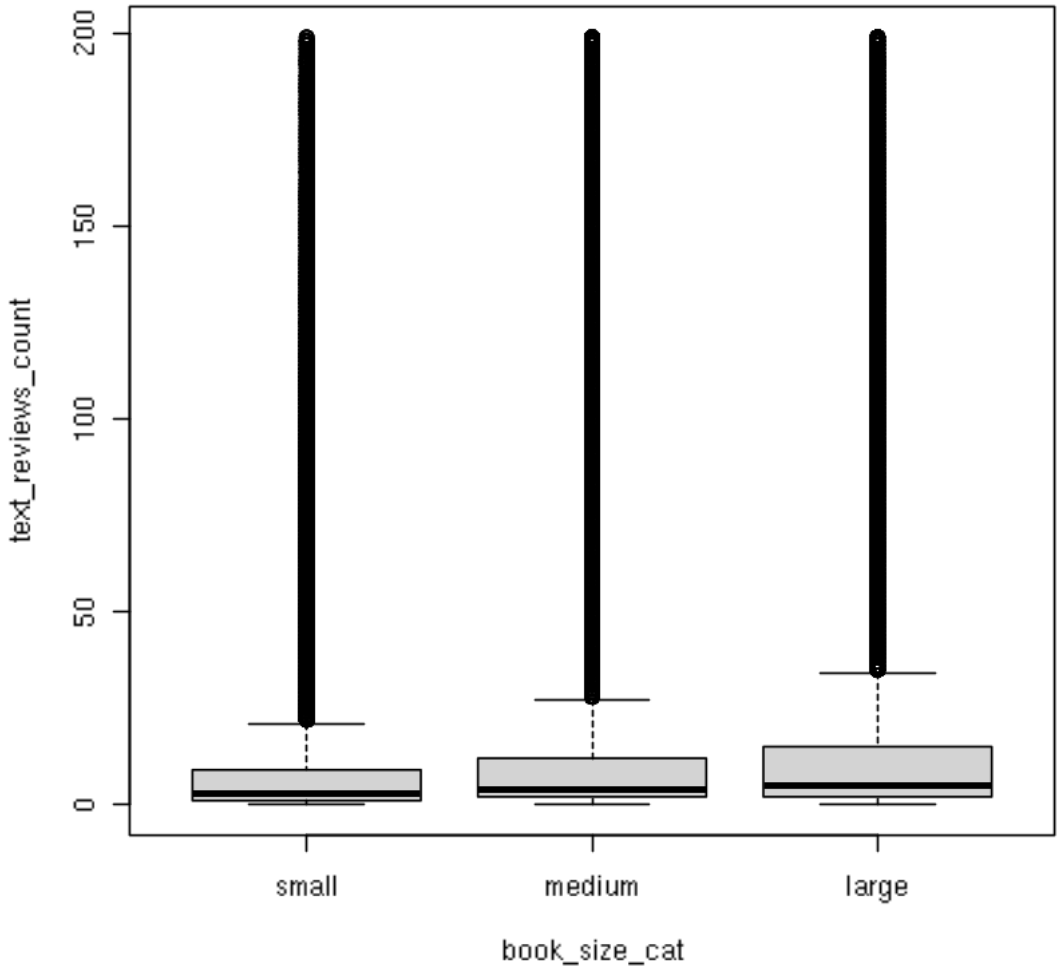
[12]:
```
boxplot(text_reviews_count~book_size_cat, data= books)
```



My answer won't change based on the boxplot because "large" is still category that owns more reviews, and "small" has obviously less reviews than the other categorys. So, I would still recommand "large" book size. To be honest, the box is hard to read because the value beyond the box is so large that I can hardly see the box.

1.5 Question 5

[13]:
```
boxplot(text_reviews_count~book_size_cat, data= subset(books,␣
↪text_reviews_count<200))
```



The box is shown up since we use a smaller data set to concentrate on smaller values. It is easier to read than before, and from the plot, we can see that large box has obviously larger number of reviews than medium and small.

1.6 Question 6

[3]:
```
authors <- fread("/depot/datamine/data/goodreads/csv/goodreads_book_authors.
↪csv")
```

[15]:
```
dim(authors)
```

1. 829529 2. 5

[4]:
```
names(authors)
```

1. 'average_rating' 2. 'author_id' 3. 'text_reviews_count' 4. 'name' 5. 'ratings_count'

[5]:
```
names(authors) %in% names(books)
```

1. TRUE 2. TRUE 3. TRUE 4. FALSE 5. TRUE

[17]:
```
books_authors <- merge(books,authors, by.x="author_id", by.y="author_id")
```

[18]:
```
dim(books_authors)
```

1. 1000000 2. 31

[19]:
```
Sub <- subset(books_authors, name %in% c("Douglas Adams","Lloyd␣
↪Alexander","William Shakespeare","John Donne","John Keats"))
```

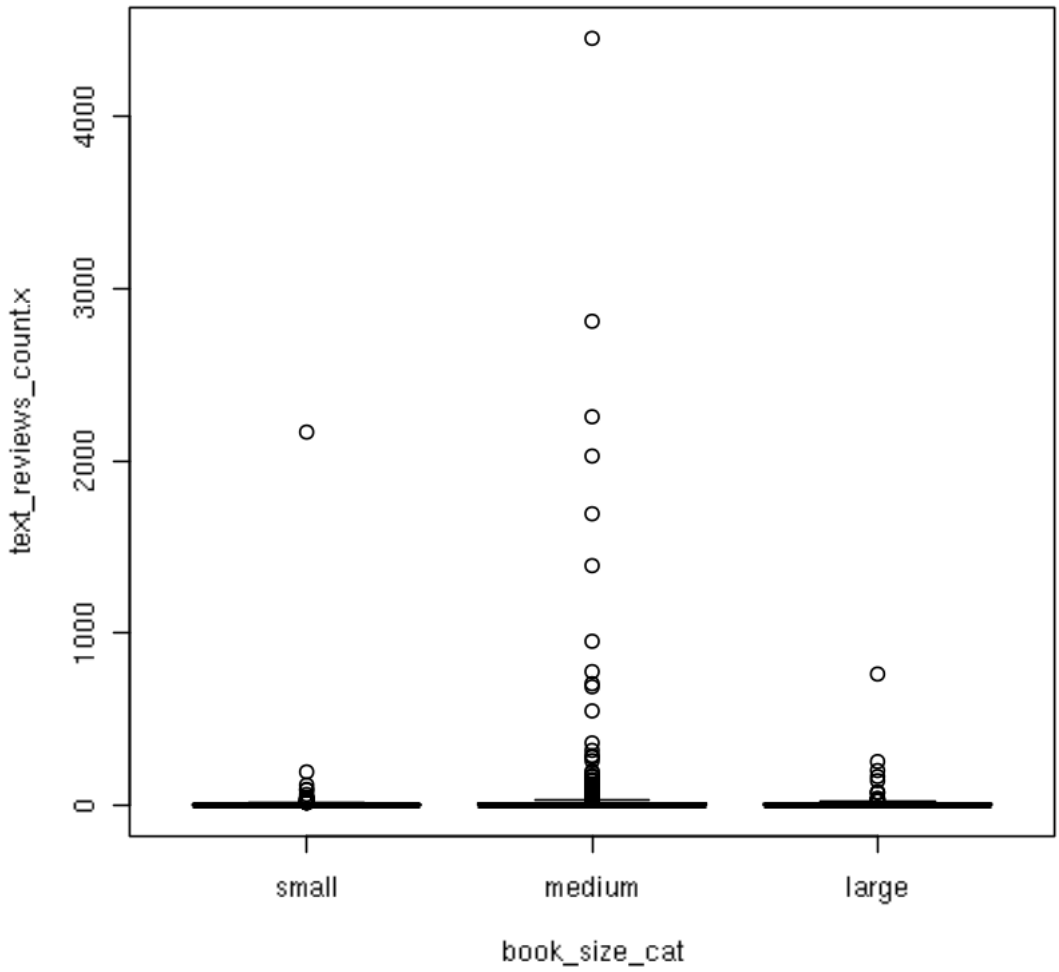
[20]:
```
dim(Sub)
```

1. 1394 2. 31

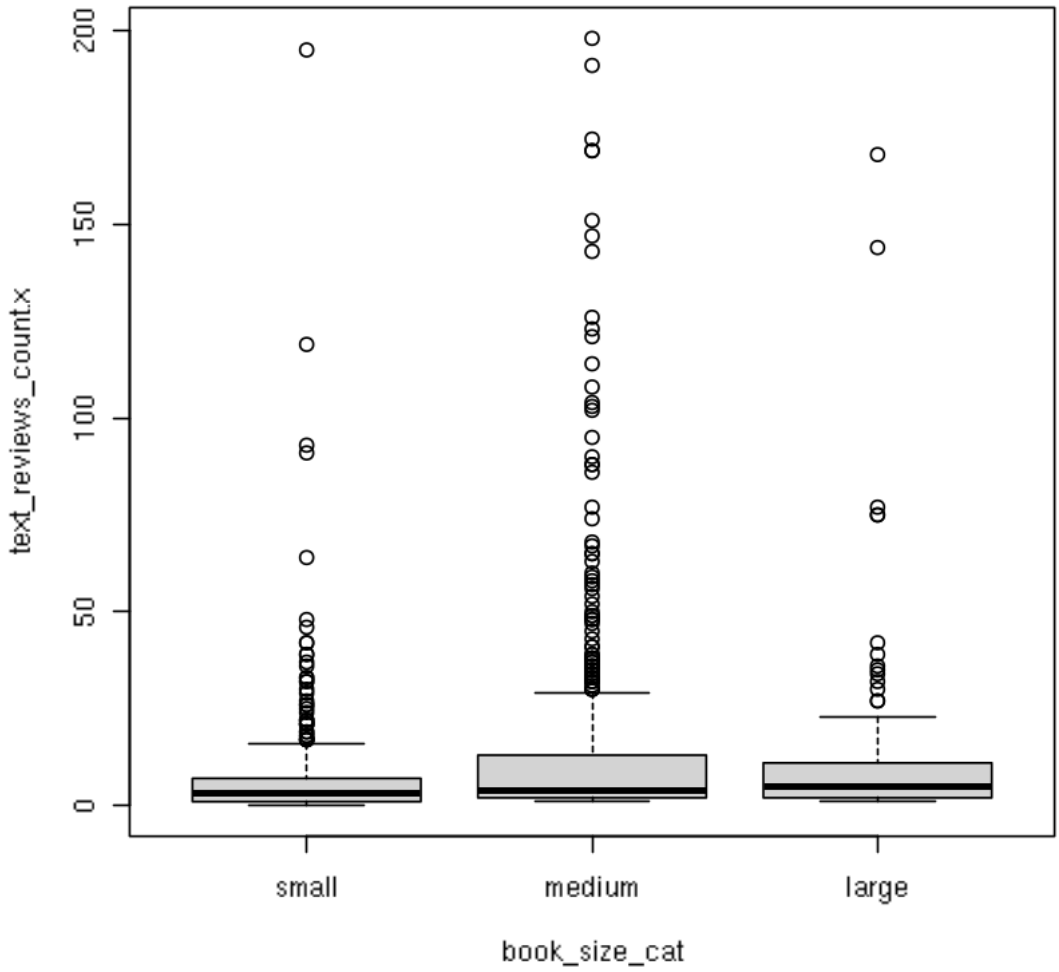
[24]:
```
names(Sub)
```

1. 'author_id' 2. 'isbn' 3. 'text_reviews_count.x' 4. 'country_code' 5. 'language_code' 6. 'asin' 7. 'is_ebook' 8. 'average_rating.x' 9. 'kindle_asin' 10. 'description' 11. 'format' 12. 'link' 13. 'publisher' 14. 'num_pages' 15. 'publication_day' 16. 'isbn13' 17. 'publication_month' 18. 'edition_information' 19. 'publication_year' 20. 'url' 21. 'image_url' 22. 'book_id' 23. 'ratings_count.x' 24. 'work_id' 25. 'title' 26. 'title_without_series' 27. 'book_size_cat' 28. 'average_rating.y' 29. 'text_reviews_count.y' 30. 'name' 31. 'ratings_count.y'

[27]:
```
boxplot(text_reviews_count.x ~ book_size_cat, data= Sub)
```



[28]:
```
boxplot(text_reviews_count.x ~ book_size_cat, data= subset(Sub,␣
↪text_reviews_count.x <200 ))
```



As we can see, recommandation should change to "medium" book size when we concentrate on these 5 authors: Douglas Adams, Lloyd Alexander, William Shakespeare, John Donne, and John Keats. In both original and restricted boxplot, "medium" book size seems to have greater number of reviews than "large" and "small". Thus, I would rather recommand "medium" book size here.

1.7 Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together – We are Purdue.