Sprawozdanie

Alicja Myśliwiec, Natalia Lach

2022-12-03

1. Wstęp

eee

2. Opis danych

Zbiór danych przedstawia cechy najlepszych 5281 książek kryminalnych i zagadkowych. (jak przetłumaczyć crime and mystery to ja nie wiem xd)

Wczytanie:

```
plik <- read.csv('best_crime_and_mystery_books.csv', na.strings=c("", "NA"), header = TRUE)
plik[5,]
```

```
##   book_rank     id         title   book_author publication_year publisher
## 5         5 168642 In Cold Blood Truman Capote             1994   Vintage
##   language_code num_pages average_rating ratings_count
## 5           eng       343           4.07        463437
```

```
plik$publication_year[plik$publication_year == '6'] <- 2006
plik$publication_year[plik$publication_year == '17'] <- 2017
```

opis zmiennych - mozna sprobowac zrobic tabele z tymi zmiennymi z kolumnami: zmienna, rodzaj (kateoryczna/ciągła), typ danych (int/numeric/chr), min i max wartosci, ilość braków, krótki opis co przedstawia (zamiast takiego wypisywania)

book rank

id

title

book_author

publication_year

publisher

language_code

num_pages

average_rating

ratings_count

3. pytania badawcze, cel analizy

Cel analizy - odpowiedz na pytanie jakie cechy mają książki, którymi interesuje się najwięcej ludzi? analiza popularności książek kryminalnych w zależności od ich aspektów, takich jak ilość stron, wydawnictwo, autor.

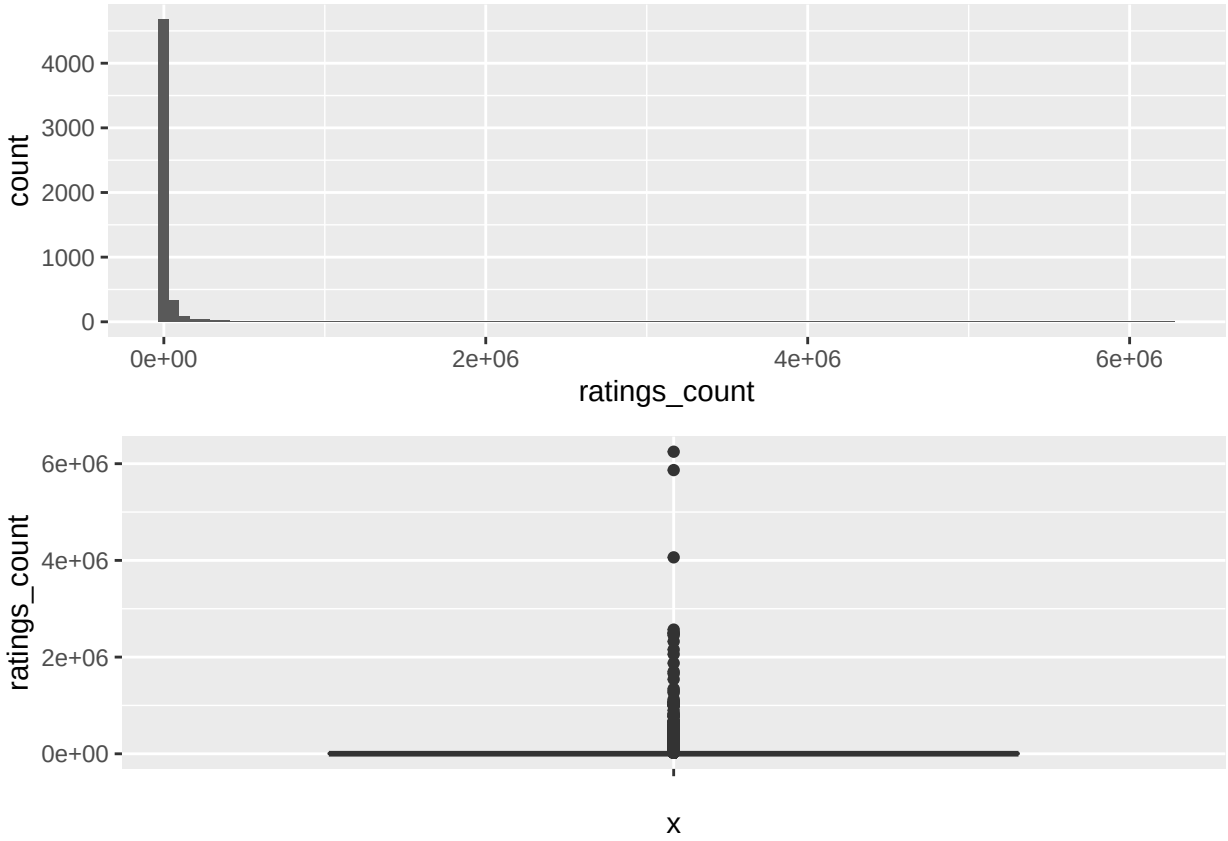
3.1 quick analiza popularnosci = zainteresowania = ratings_count

najwazniejsza zmienna dla nas, przedstawia ilosc osob zainteresowanych tą książką na tyle, żeby dać jej ocene, narysowac hist

```
p1 <- ggplot(plik, aes( x = ratings_count)) + geom_histogram(bins=100)

p2 <- ggplot(plik, aes(x = "", y = ratings_count)) +
  geom_boxplot()

grid.arrange(p1, p2, nrow=2)
```



Quick analiza, ze srednia ilosc glosow oddanych na ksiazke wynosi tyle 3.0241333×10^4 , mediana tyle 1234, wiec rozklad taki (. . . -skosny), duzo odstajacych itd.

Najpopularniejsze książki to:

```
plik[order(-plik$ratings_count),][1:5,c(3,10)]
```

```
##                                                          title ratings_count
## 3386 Harry Potter and the Sorcerer's Stone (Harry Potter, #1)       6247740
## 1279                  The Hunger Games (The Hunger Games, #1)       5867734
## 113                                      To Kill a Mockingbird       4063329
## 970                       The Hobbit, or There and Back Again       2568612
## 3883                                The Diary of a Young Girl       2503131
```

Faktycznie są to bardzo znane książki, z dużą iloscia ocen, są to te wartosci odstające, bo count jest way bigger niz srednia albo mediana, no i jest pare ksiazek z zerowymi głosami. (jak one trafiły do tego zestawienia top ksiazek to idk)

3.2 Czy czas miał wpływ na popularność? Książki z którego roku cieszą się największymz zainteresowaniem?

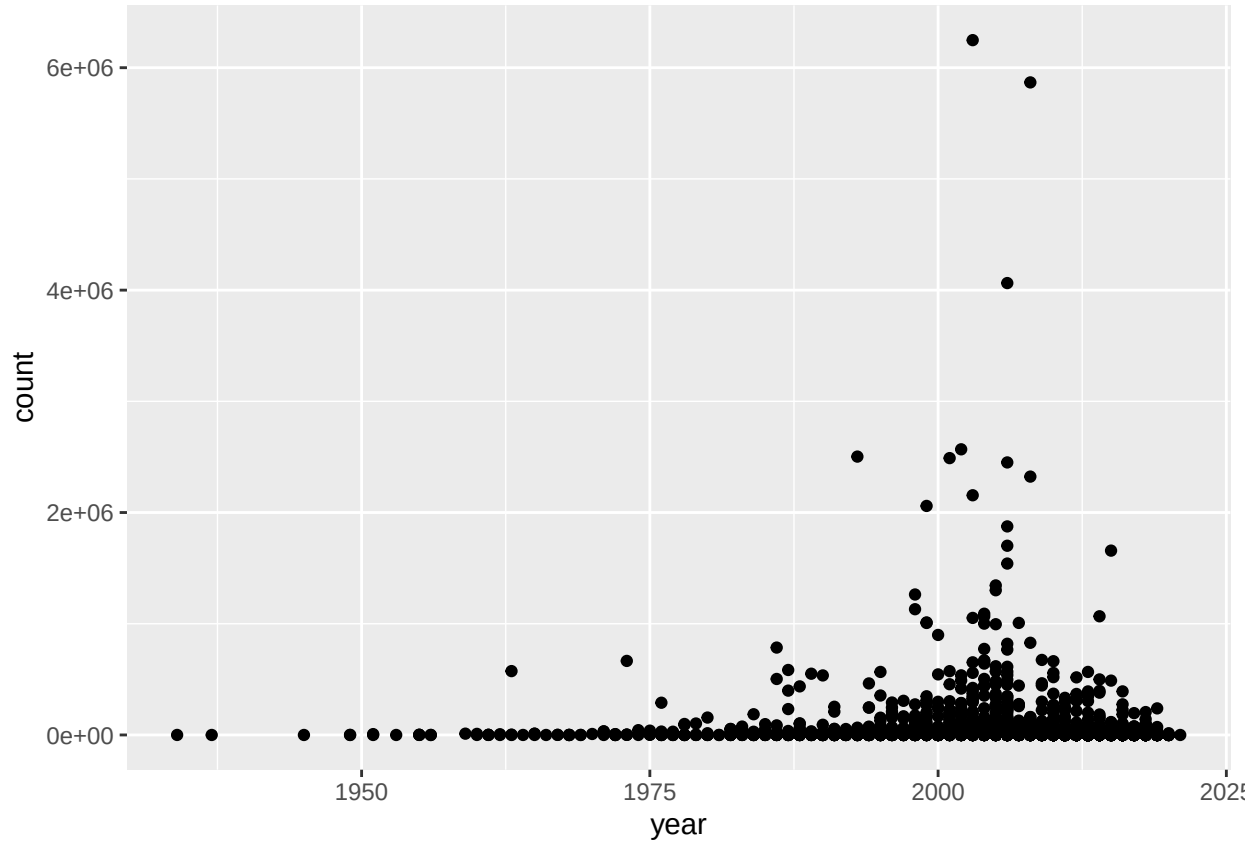
wczytanie danych:

```
##wczytanie danych - z pominieciem brakujących wartości
year_vs_count <- na.omit(data.frame( year = plik$publication_year, count = plik$ratings_count))
head(year_vs_count)
```

```
##   year   count
## 1 2008 2323151
```

```
## 2 2004  642138
## 3 2006 2450604
## 4 2013  200400
## 5 1994  463437
## 6 2002  287416
```

```
p1 <- ggplot(year_vs_count, aes(x=year, y=count)) + geom_point()
p1
```



rok a ilosc ksiazek a laczna ilosc ocen a srednia ilosc ocen na ksiazke

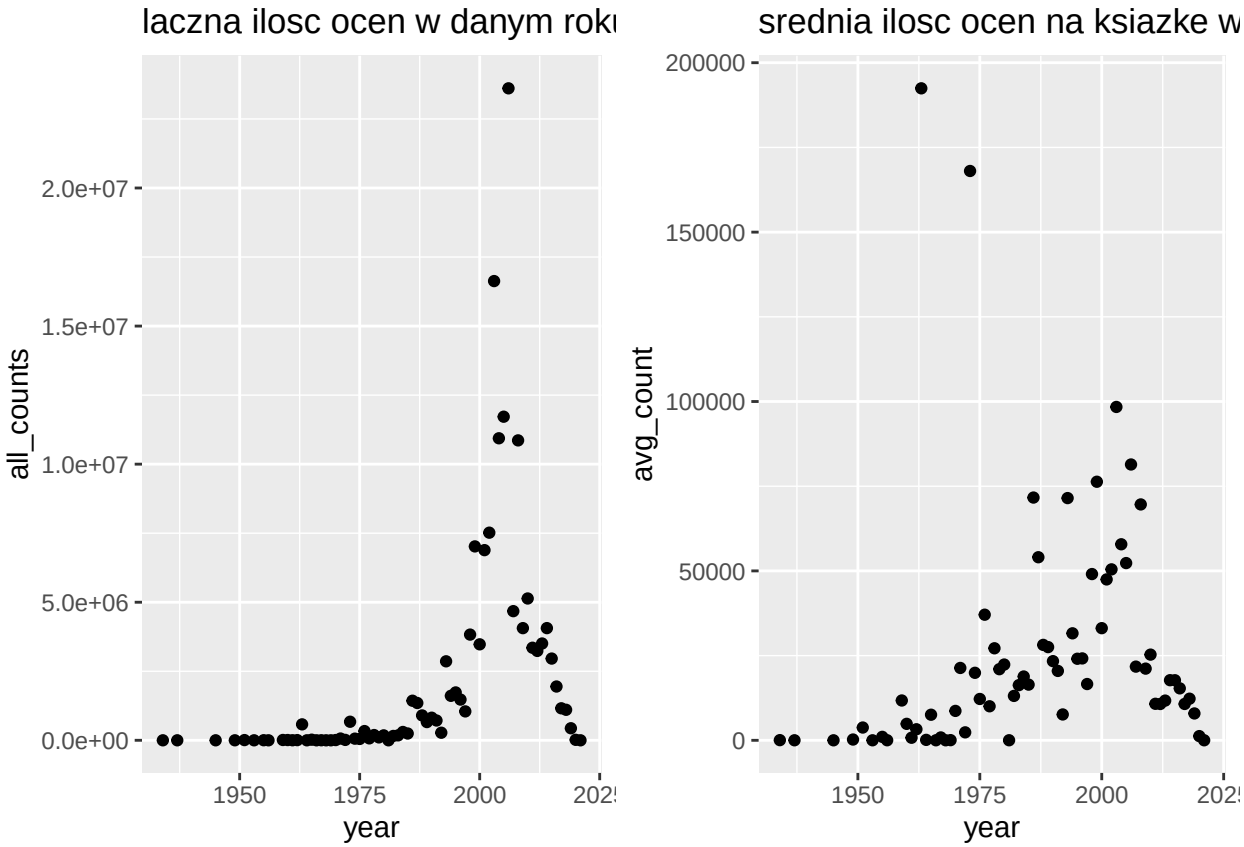
```
df <- as.data.frame(table(year_vs_count$year))
colnames(df) <- c('year', 'bpy')
df$all_counts <- aggregate(year_vs_count$count, by=list(Category=year_vs_count$year), FUN=sum)[,2]
df$avg_count <- df$all_counts/df$bpy
df$year <- as.numeric(as.character(df$year))
```

wykresiki:

```
p1 <- ggplot(df, aes(x=year, y=all_counts)) + geom_point() + ggtitle('łączna ilość ocen w danym roku')

p2 <- ggplot(df, aes(x=year, y=avg_count)) + geom_point() + ggtitle('średnia ilość ocen na książkę w da

grid.arrange(p1, p2, ncol=2)
```



Jakas analiza brak widocznej zaleznosci liniowej, ale rangi spearmana nawet wysokie

```
cor(df$all_counts, df$year, method = "pearson")
```

```
## [1] 0.4564137
```

```
cor(df$all_counts, df$year, method = "kendall")
```

```
## [1] 0.5983903
```

```
cor(df$all_counts, df$year, method = "spearman")
```

```
## [1] 0.7614353
```

```
cor(df$avg_count, df$year)
```

```
## [1] 0.1691126
```

top rok pod wzgledem ocen:

```
df[df$all_counts == max(df$all_counts),] ##łącznie
```

```
##    year bpy all_counts avg_count
## 56 2006 290   23609016   81410.4
```

```
df[df$avg_count == max(df$avg_count),] ##średnio
```

```
##    year bpy all_counts avg_count
## 13 1963   3     577323    192441
```

wnioski:

3.3 Czyli czy ilość stron ma wpływ na zainteresowanie książką

```
num_vs_count <- na.omit(data.frame(num_pages = plik$num_pages, ratings_count = plik$ratings_count))
```

```
cor(num_vs_count$num_pages, num_vs_count$ratings_count, method = "pearson")
```

```
## [1] 0.02211679
```

```
cor(num_vs_count$num_pages, num_vs_count$ratings_count, method = "spearman")
```

```
## [1] 0.2876353
```

```
cor(num_vs_count$num_pages, num_vs_count$ratings_count, method = "kendall")
```

```
## [1] 0.1986348
```

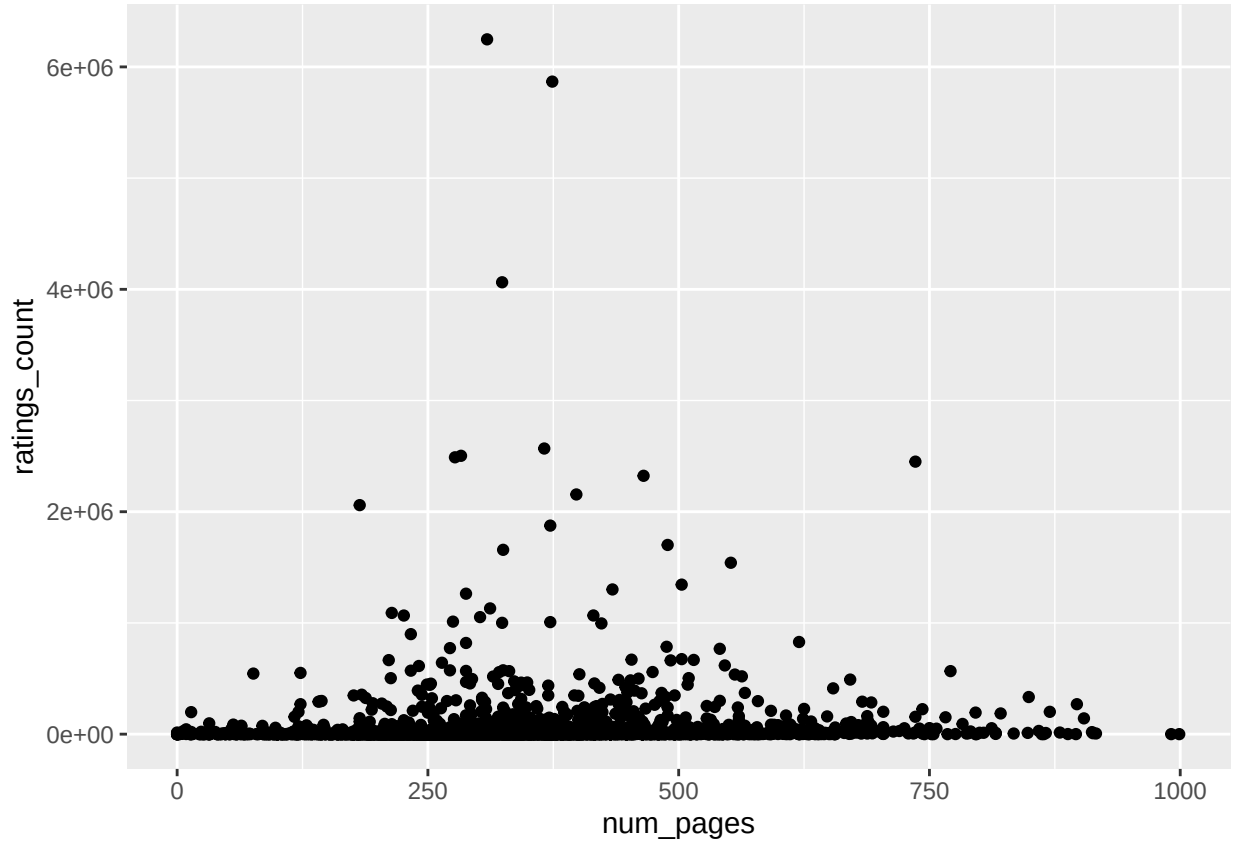
```
plik[plik$ratings_count == max(plik$ratings_count),]
```

```
##      book_rank id                                                    title
## 3386      3107  3 Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
##       book_author publication_year      publisher language_code num_pages
## 3386 J.K. Rowling             2003 Scholastic Inc           eng       309
##      average_rating ratings_count
## 3386           4.47       6247740
```

```
p <- ggplot(num_vs_count, aes(x=num_pages, y=ratings_count)) + geom_point() + scale_x_continuous(limits=

p
```

```
## Warning: Removed 26 rows containing missing values (geom_point).
```




```
n_p <- avg_p <- num_books <- c()
i <- 1
ns <- 50

for (n in seq(ns, max(num_vs_count$num_pages), ns)) {
    all <- sum(num_vs_count[num_vs_count$num_pages <=n & num_vs_count$num_pages > n - ns ,2])
    how_many <- sum(num_vs_count$num_pages <=n& num_vs_count$num_pages >= n - ns)
    n_p[i] <- all
    num_books[i] <- how_many
    avg_p[i] <- all/how_many
    i <- i+1
}
count_per_pages <- na.omit(data.frame(less_than_pages = seq(ns, max(num_vs_count$num_pages), ns), overa
```

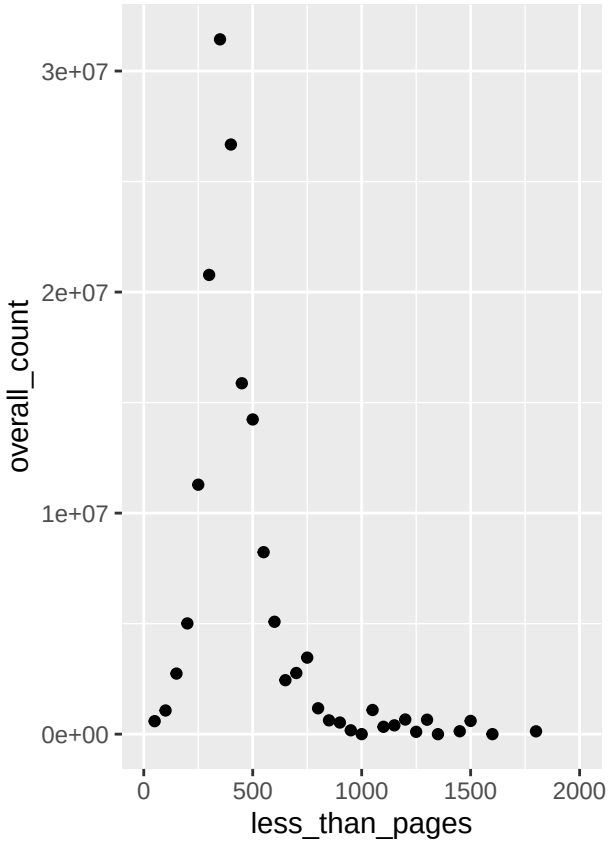
wykr

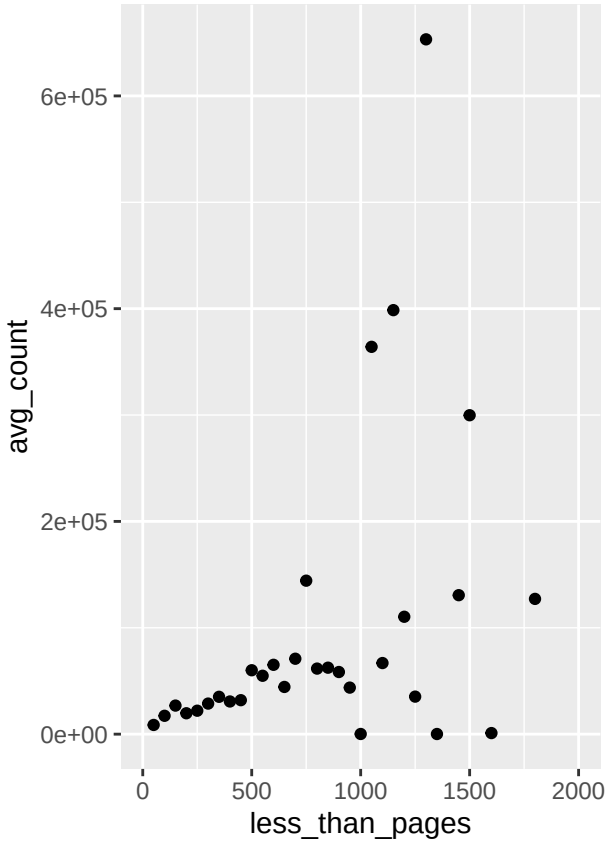
```
p1 <- ggplot(count_per_pages, aes(x=less_than_pages, y=overall_count)) + geom_point() + scale_x_continu


p2 <- ggplot(count_per_pages, aes(x=less_than_pages, y=avg_count)) + geom_point() + scale_x_continuous(


grid.arrange(p1, p2, ncol=2)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
## Removed 2 rows containing missing values (geom_point).
```





top wyniki

```
count_per_pages[count_per_pages$overall_count == max(count_per_pages$overall_count),]
```

```
##   less_than_pages overall_count avg_count num_books
## 7             350      31433726  35121.48       895
```

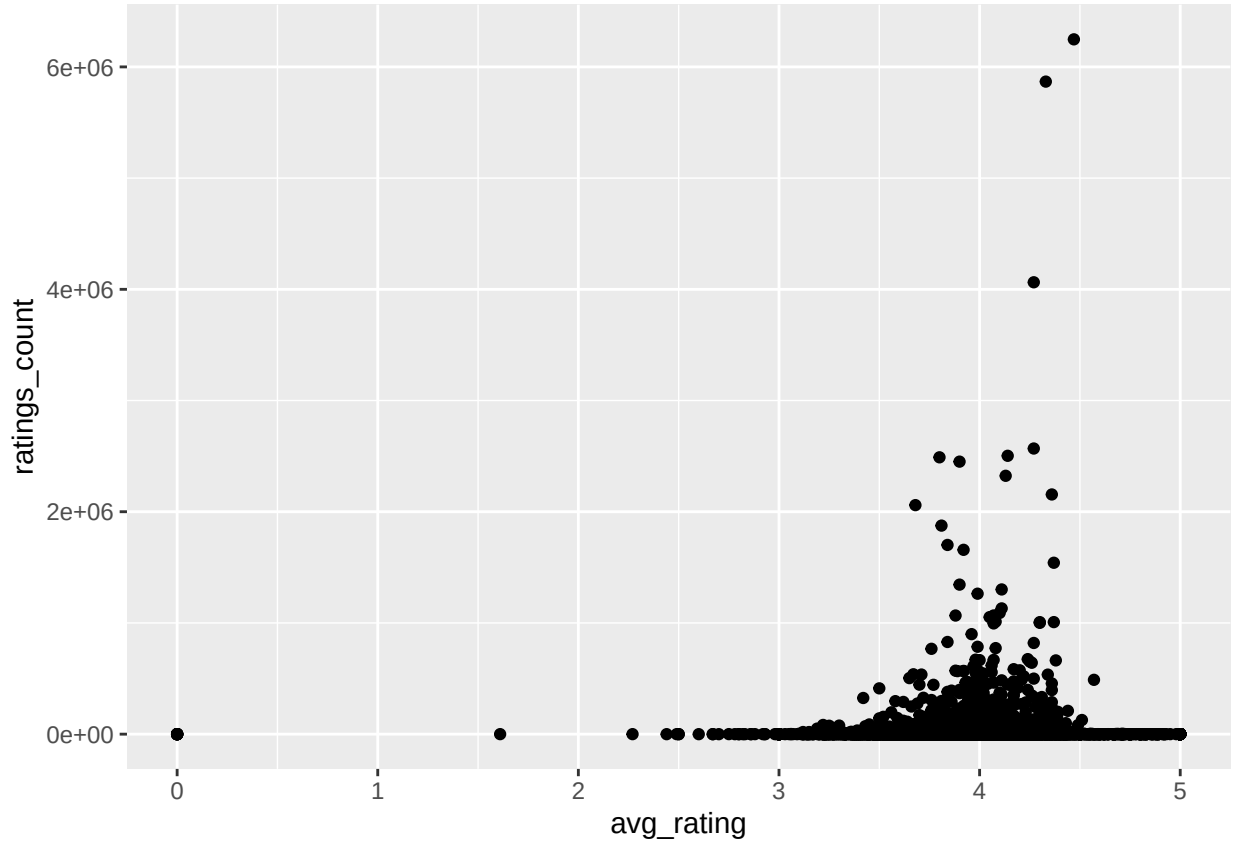
```
count_per_pages[count_per_pages$avg_count == max(count_per_pages$avg_count),]
```

```
##    less_than_pages overall_count avg_count num_books
## 26            1300        653184    653184         1
```

mozna tez te boxploty tu dac

3.4 How are average rating and popularity related? AVG VS COUNT

```
avg_vs_count <- na.omit(data.frame(avg_rating = plik$average_rating, ratings_count = plik$ratings_count
p <- ggplot(avg_vs_count, aes(x=avg_rating, y=ratings_count)) + geom_point()
p
```




```
cor(avg_vs_count$avg_rating, avg_vs_count$ratings_count, method = "pearson")
```

```
## [1] 0.04642757
```

```
cor(avg_vs_count$avg_rating, avg_vs_count$ratings_count, method = "kendall")
```

```
## [1] -0.009206791
```

```
cor(avg_vs_count$avg_rating, avg_vs_count$ratings_count, method = "spearman")
```

```
## [1] -0.02883265
```

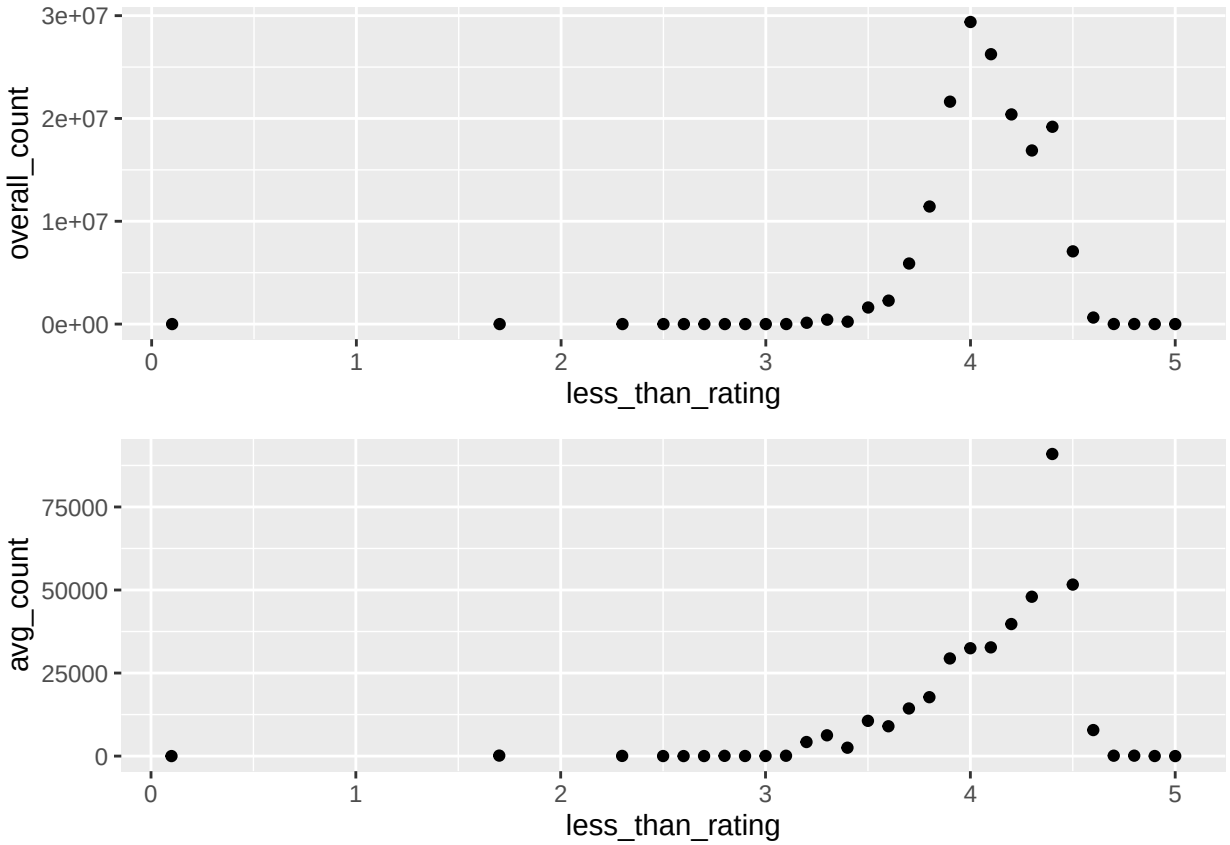
```
n_r <- avg_p <- num_books <- c()
i <- 1
ns <- 0.1

for (n in seq(ns, max(avg_vs_count$avg_rating), ns)) {
    all <- sum(avg_vs_count[avg_vs_count$avg_rating < n & avg_vs_count$avg_rating >= n - ns ,2])
    how_many <- sum(avg_vs_count$avg_rating  <=n& avg_vs_count$avg_rating  >= n - ns)
    n_r[i] <- all
    num_books[i] <- how_many
    avg_p[i] <- all/how_many
    i <- i+1
}
avg_per_count <- na.omit(data.frame(less_than_rating = seq(ns, max(avg_vs_count$avg_rating), ns), overa
```

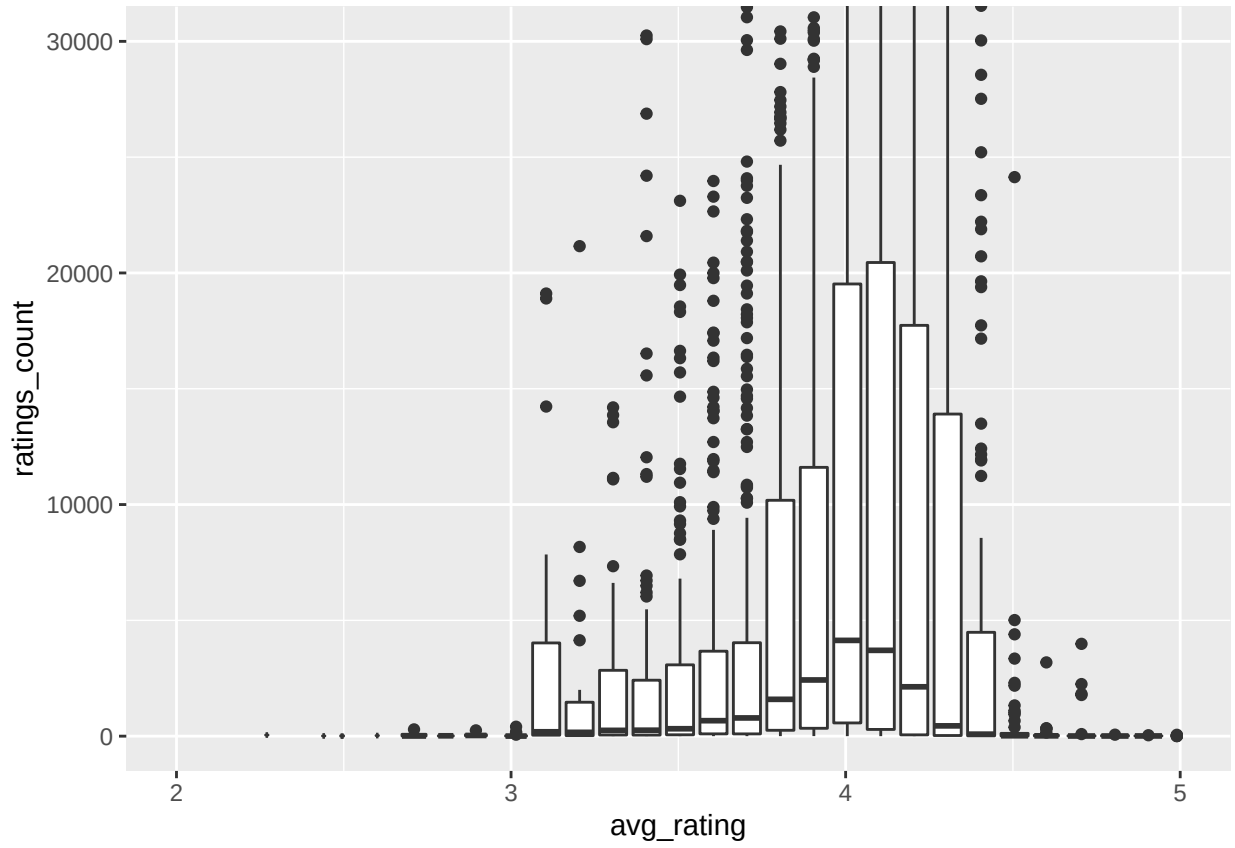
```
p1 <-  ggplot(avg_per_count, aes(x=less_than_rating, y=overall_count)) + geom_point()
```

```
p2 <-  ggplot(avg_per_count, aes(x=less_than_rating, y=avg_count)) + geom_point()

grid.arrange(p1, p2, nrow=2)
```




```
p1 <-ggplot(avg_vs_count, aes(x=avg_rating, y=ratings_count)) +
geom_boxplot(aes(group = cut_width(avg_rating, 0.1))) +coord_cartesian(xlim =c(2,5), ylim = c(0, 30000))
p1
```



boxplot zgadza sie z pierwszym wykresem

3.5 publisher VS count

```
pub_vs_count <- na.omit(data.frame(publisher = plik$publisher, count = plik$ratings_count))

df <- as.data.frame(table(pub_vs_count$publisher))
colnames(df) <- c('publisher', 'bpp')
df$all_counts <- aggregate(pub_vs_count$count, by=list(Category=pub_vs_count$publisher), FUN=sum)[,2]
df$avg_count <- df$all_counts/df$bpp
df$publisher <- as.character(df$pub)
##top 10 publisherów pod wzgledem ilosci wydanych ksiazek
head(df[order(-df$bpp),])
```

```
##                      publisher bpp all_counts avg_count
## 813              Minotaur Books  83     473157  5700.687
## 104                     Bantam  81    5587398 68980.222
## 103           Ballantine Books  79    1532055 19393.101
## 499    Grand Central Publishing  73    2859470 39170.822
## 475          G.P. Putnam's Sons  69     512806  7431.971
## 1320 Vintage Crime/Black Lizard  69     361169  5234.333
```

```
##top 10 publisherów pod wzgledem lacznej ilosci ocen
head(df[order(-df$all_counts),])
```

```
##                         publisher bpp all_counts  avg_count
## 931                  Penguin Books  59    6777997  114881.31
```

```
## 1096                   Scholastic Inc   1    6247740 6247740.00
## 1098                 Scholastic Press   3    6087228 2029076.00
## 104                           Bantam  81    5587398   68980.22
## 732         Little, Brown and Company  38    5432257  142954.13
## 548  Harper Perennial Modern Classics   4    5207640 1301910.00
```

```
##top 10 publisherów pod wzgledem sredniej ilosci ocen
head(df[order(-df$avg_count),])
```

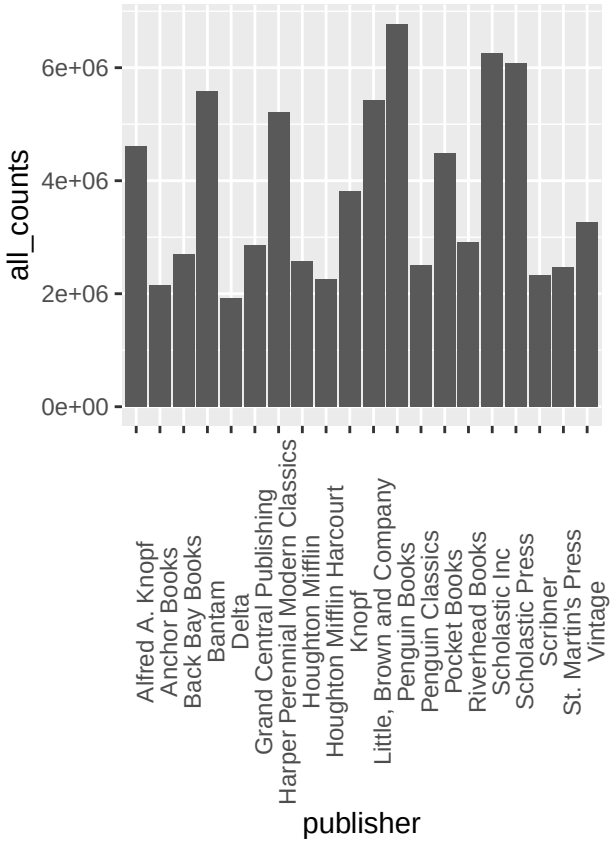
```
##                               publisher bpp all_counts avg_count
## 1096                    Scholastic Inc   1    6247740   6247740
## 1098                  Scholastic Press   3    6087228   2029076
## 548      Harper Perennial Modern Classics   4    5207640   1301910
## 1021  Random House Large Print Publishing   1    1301127   1301127
## 603                     Houghton Mifflin   2    2568974   1284487
## 565  HarperPerennial / Perennial Classics   1    1263125   1263125
```

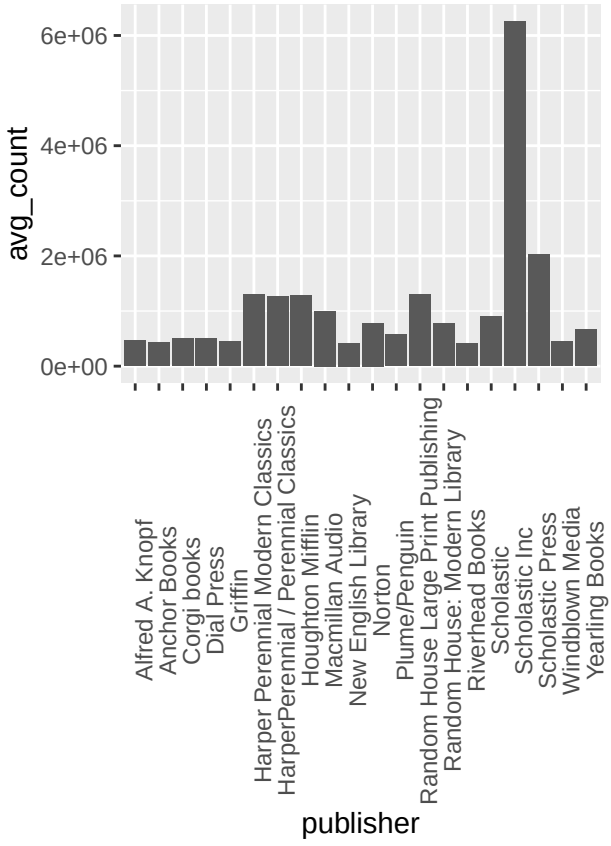
jakas analiza znowu i wnioski

```
##chwalimy sie umiejetnoscia robienia barplotow
top20 <- df[order(-df$all_counts),][1:20,]
p1 <- ggplot(top20, aes(x=publisher, y=all_counts)) + geom_bar(stat="identity") +
theme(axis.text.x = element_text(angle = 90))

top20_2 <-  df[order(-df$avg_count),][1:20,]
p2 <- ggplot(top20_2, aes(x=publisher, y=avg_count)) + geom_bar(stat="identity") +
theme(axis.text.x = element_text(angle = 90))

grid.arrange(p1, p2, ncol=2)
```





wnioskiii

3.6 top 10 authors - authors vs count

```
aut_vs_count <- na.omit(data.frame(author = plik$book_author, count = plik$ratings_count))

df <- as.data.frame(table(aut_vs_count$author))
colnames(df) <- c('author', 'bpa')
df$all_counts <- aggregate(aut_vs_count$count, by=list(Category=aut_vs_count$author), FUN=sum)[,2]
df$avg_count <- df$all_counts/df$bpa
df$publisher <- as.character(df$author)
##top 10 autorow pod wzgledem ilosci wydanych ksiazek
head(df[order(-df$bpa),])
```

```
##                 author bpa all_counts avg_count          publisher
## 32      Agatha Christie  93    2412726 25943.290    Agatha Christie
## 1077    James Patterson  36    1908945 53026.250    James Patterson
## 201  Arthur Conan Doyle  31    1101980 35547.742 Arthur Conan Doyle
## 2189       Ruth Rendell  29      51173  1764.586       Ruth Rendell
## 902        Harlan Coben  26     652441 25093.885       Harlan Coben
## 978        Isaac Asimov  24    1138158 47423.250       Isaac Asimov
```

```
##top 10 publisherów pod wzgledem lacznej ilosci ocen
head(df[order(-df$all_counts),])
```

```
##               author bpa all_counts avg_count       publisher
## 1009    J.K. Rowling   1    6247740 6247740.0    J.K. Rowling
```

```
## 517        Dan Brown  11    6033868  548533.5       Dan Brown
## 2396 Suzanne Collins   1    5867734 5867734.0 Suzanne Collins
## 1016  J.R.R. Tolkien   2    4724085 2362042.5  J.R.R. Tolkien
## 903       Harper Lee   1    4063329 4063329.0      Harper Lee
## 2338    Stephen King  18    3872836  215157.6    Stephen King
```

```
##top 10 publisherów pod wzgledem sredniej ilosci ocen
head(df[order(-df$avg_count),])
```

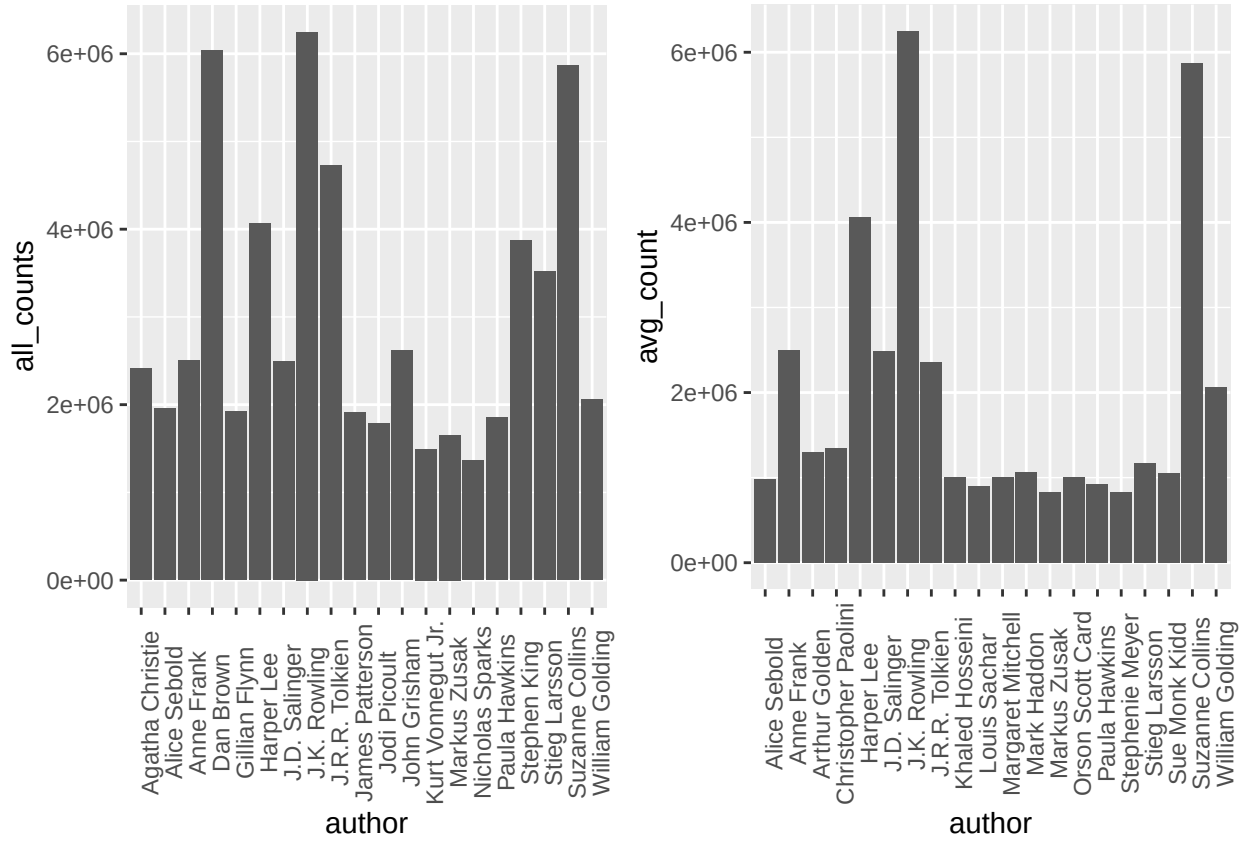
```
##               author bpa all_counts avg_count       publisher
## 1009    J.K. Rowling   1    6247740   6247740    J.K. Rowling
## 2396 Suzanne Collins   1    5867734   5867734 Suzanne Collins
## 903       Harper Lee   1    4063329   4063329      Harper Lee
## 168       Anne Frank   1    2503131   2503131      Anne Frank
## 1002   J.D. Salinger   1    2489479   2489479   J.D. Salinger
## 1016  J.R.R. Tolkien   2    4724085   2362043  J.R.R. Tolkien
```

bla bla

```
##chwalimy sie umiejetnoscia robienia barplotow
top20 <- df[order(-df$all_counts),][1:20,]
p1 <- ggplot(top20, aes(x=author, y=all_counts)) + geom_bar(stat="identity") +
theme(axis.text.x = element_text(angle = 90))

top20_2 <-  df[order(-df$avg_count),][1:20,]
p2 <- ggplot(top20_2, aes(x=author, y=avg_count)) + geom_bar(stat="identity") +
theme(axis.text.x = element_text(angle = 90))

grid.arrange(p1, p2, ncol=2)
```



wnioskiii

4. podumowanie i wnioski

jestesmy super