Goodreads

Aurora

05 aprile 2019

Goodreads

Il dataset preso in considerazione è tratto da kaggle.com.

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## 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(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
goodreads <- read.csv("good_reads_final.csv") %>%
  mutate(author_genres = stringr::str_replace(author_genres,",","/")) %>%
  separate(author_genres,into=c("first_genre","other_genre"),sep="/") %>%
  mutate(other_genre = ifelse(is.na(other_genre) | other_genre=="", "unknown",other_genre)) %>%
  mutate(other_genre = stringr::str_replace_all(other_genre,",",""))
```

Il dataset è composto da 20 colonne:

- author_average_rating
- author_gender
- author_genres
- author_id
- author_name
- author_page_url
- author_rating_count
- author_review_count
- birthplace
- book_average_rating
- book_fullurl
- book_id
- book_title
- genre1
- genre2
- num_ratings
- num_reviews
- pages
- publish_date
- score

Eliminiamo le colonne inutili

```
goodreads<-goodreads %>%
  select(-author_page_url,-book_fullurl)
```

Sistemare le colonne

```
goodreads <- goodreads %>%
  mutate(author_name = stringr::str_replace_all(author_name, "\n","")) %>%
  mutate(birthplace = stringr::str_replace_all(birthplace, "\n",""))%>%
  mutate(book_title = stringr::str_replace_all(book_title, "\n","")) %>%
  mutate(first_genre = stringr::str_replace_all(first_genre,"-"," "),
         other_genre = stringr::str_replace_all(other_genre,"-"," ")) %>%
  mutate(birthplace = stringr::str_replace_all(birthplace, "  ","")) %>%
  mutate(birthplace = ifelse(birthplace == "","unknown",birthplace))

countries <- unique(goodreads$birthplace) %>%sort()
```

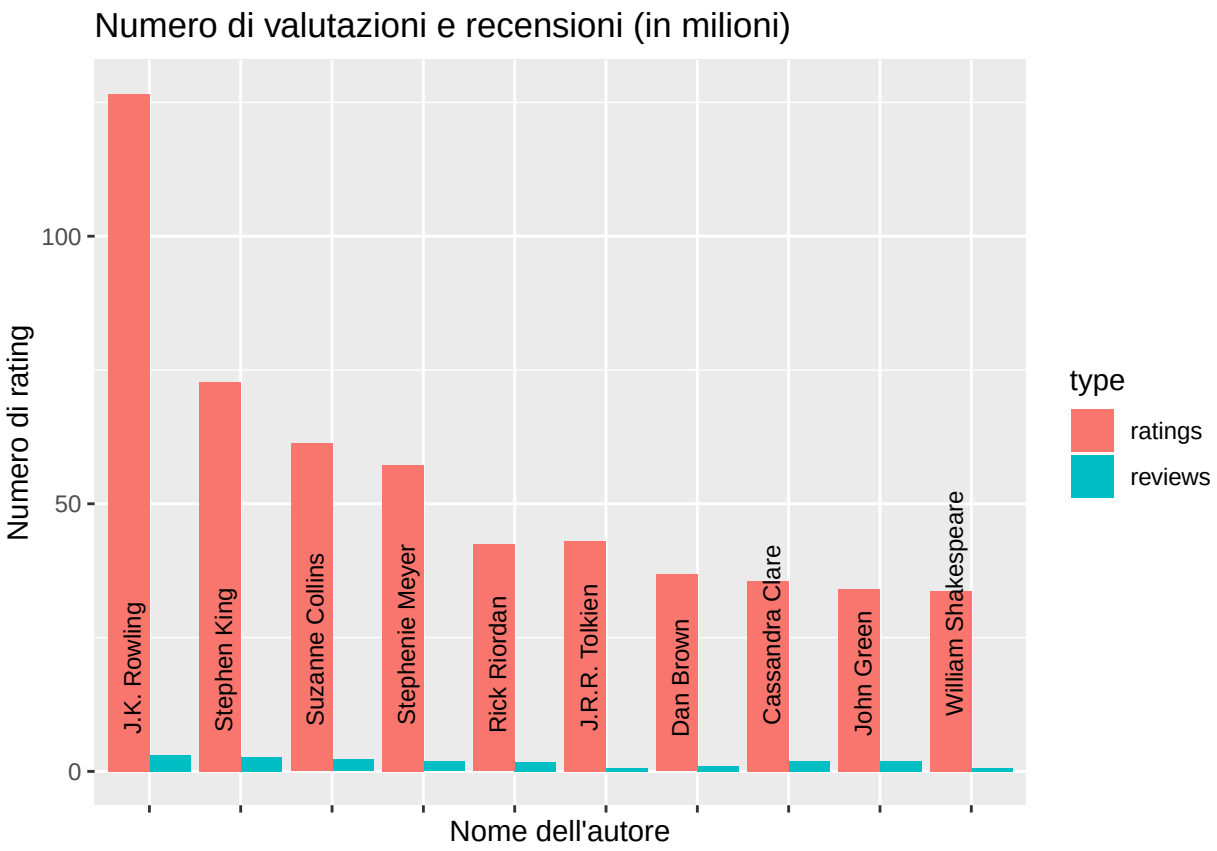
Quali sono gli autori più famosi?

```
authors<-goodreads %>%
  select(author_average_rating:birthplace, book_id) %>%
  select(author_id, author_name,everything()) %>%
  group_by(author_name, author_gender) %>%
  summarise(ratings=sum(author_rating_count), reviews = sum(author_review_count),n_books=n()) %>%
  arrange(-ratings, -reviews)

authors
```

```
## # A tibble: 12,156 x 5
## # Groups:   author_name [12,156]
##    author_name         author_gender   ratings reviews n_books
##    <chr>               <fct>             <int>   <int>   <int>
##  1 J.K. Rowling        female        126640336 3098497       6
##  2 Stephen King        male           72795154 2655847       6
##  3 Suzanne Collins     female         61288176 2223663       6
##  4 Stephenie Meyer     male           57284674 1835702       6
##  5 J.R.R. Tolkien      male           43074264  655372       6
##  6 Rick Riordan        male           42373362 1701792       6
##  7 Dan Brown           male           36774808 1009774       6
##  8 Cassandra Clare     female         35621747 1988989       6
##  9 John Green          male           34033693 1938303       6
## 10 William Shakespeare male           33669143  576948       6
## # ... with 12,146 more rows
```

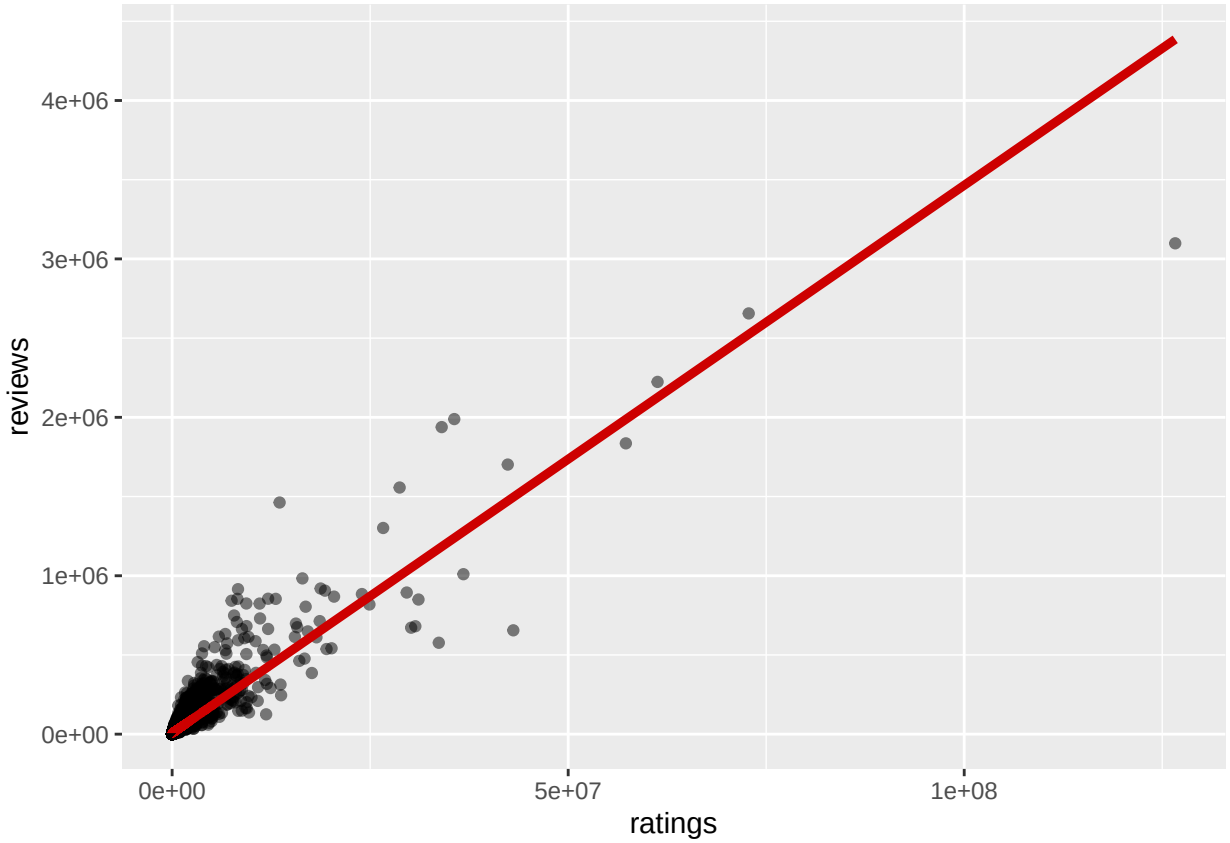
Al crescere dei ratings crescono le recensioni?

```
most_rated_authors <- authors %>%
  arrange(-ratings) %>%
  head(n=10) %>%
  gather("ratings","reviews", key="type",value="number")

ggplot(most_rated_authors, aes(x=reorder(author_name,-number)))+
  geom_bar(stat="identity", position="dodge",aes(y=number/1000000,fill=type)) +
  labs(x = "Nome dell'autore", y="Numero di rating", title="Numero di valutazioni e recensioni (in mili
  geom_text(aes(label=author_name), stat="count", size=3,angle=90, vjust=-0.5, hjust=-0.2)+
  theme(axis.text.x=element_blank())
```



Gli autori con più ratings sono anche i più recensiti?

```
library(modelr)
rew_rat<- lm(reviews~ratings,authors)

authors %>% add_predictions(rew_rat) %>%
  ggplot(aes(ratings))+
    geom_jitter(aes(y=reviews), alpha = 0.5)+
    geom_line(aes(y=pred), color="red3", size=1.6)
```

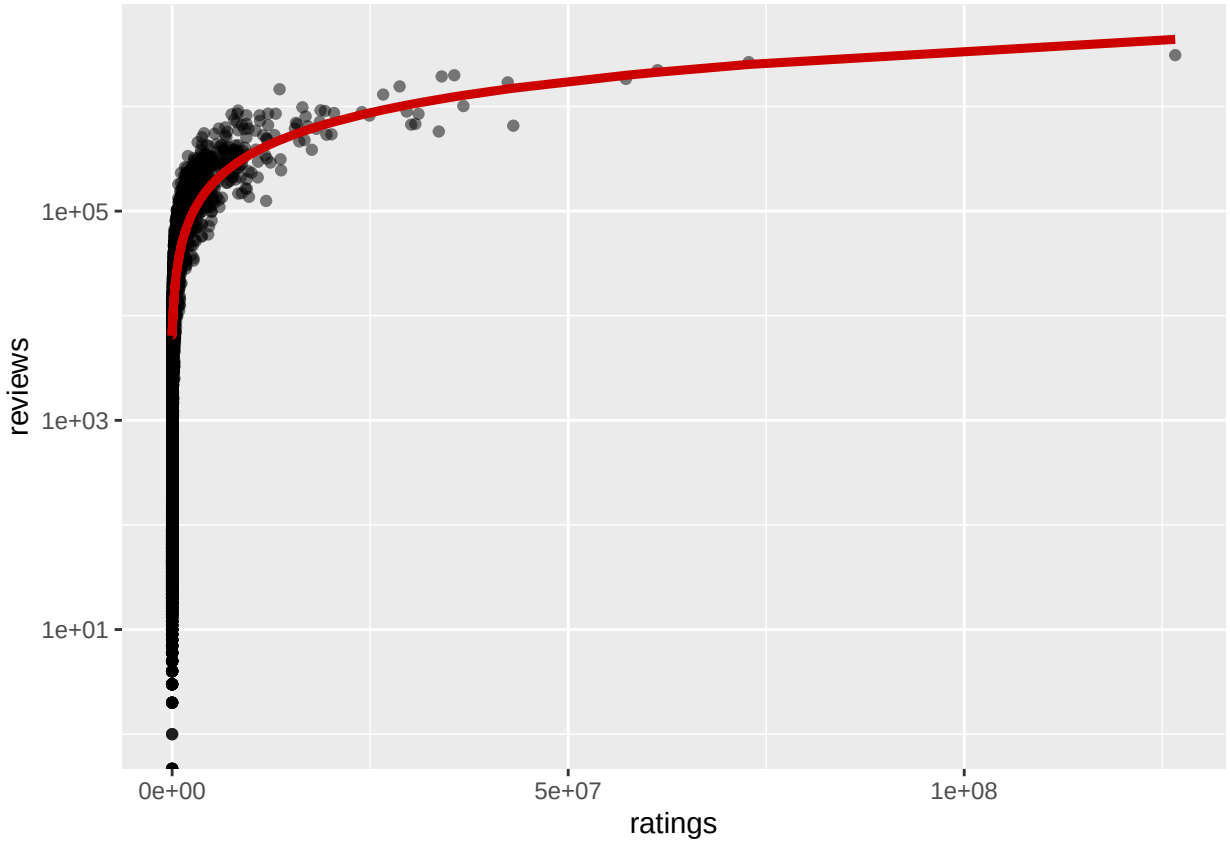


Dall'analisi risultano esserci più recensioni laddove le valutazioni sono in numero minore. Vi sono delle eccezioni, che probabilmente sono quelle precedentemente analizzate. Provando ad applicare una scala logaritmica ai dati, si nota che il modello si adatta meglio rispetto alla relazione lineare precedentemente evidenziata.

```
authors %>% add_predictions(rew_rat) %>%
  ggplot(aes(ratings))+
    geom_jitter(aes(y=reviews), alpha=0.5)+
    geom_line(aes(y=pred), color="red3", size=1.6)+
    scale_y_log10(limits=c(1,NA))
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



Alcune recensioni vanno in negativo perché molti libri hanno meno di 10 recensioni. Proviamo a indagare sulla distribuzione dei dati delle reviews:

```
summary(authors$reviews)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     303    1291   17644    6427 3098497
```

I dati sono concentrati tra 303 e 17644, mentre il massimo supera i 3 milioni. Ci sono però anche libri che non hanno recensioni.

In che modo il numero di valutazioni influisce sul punteggio medio?

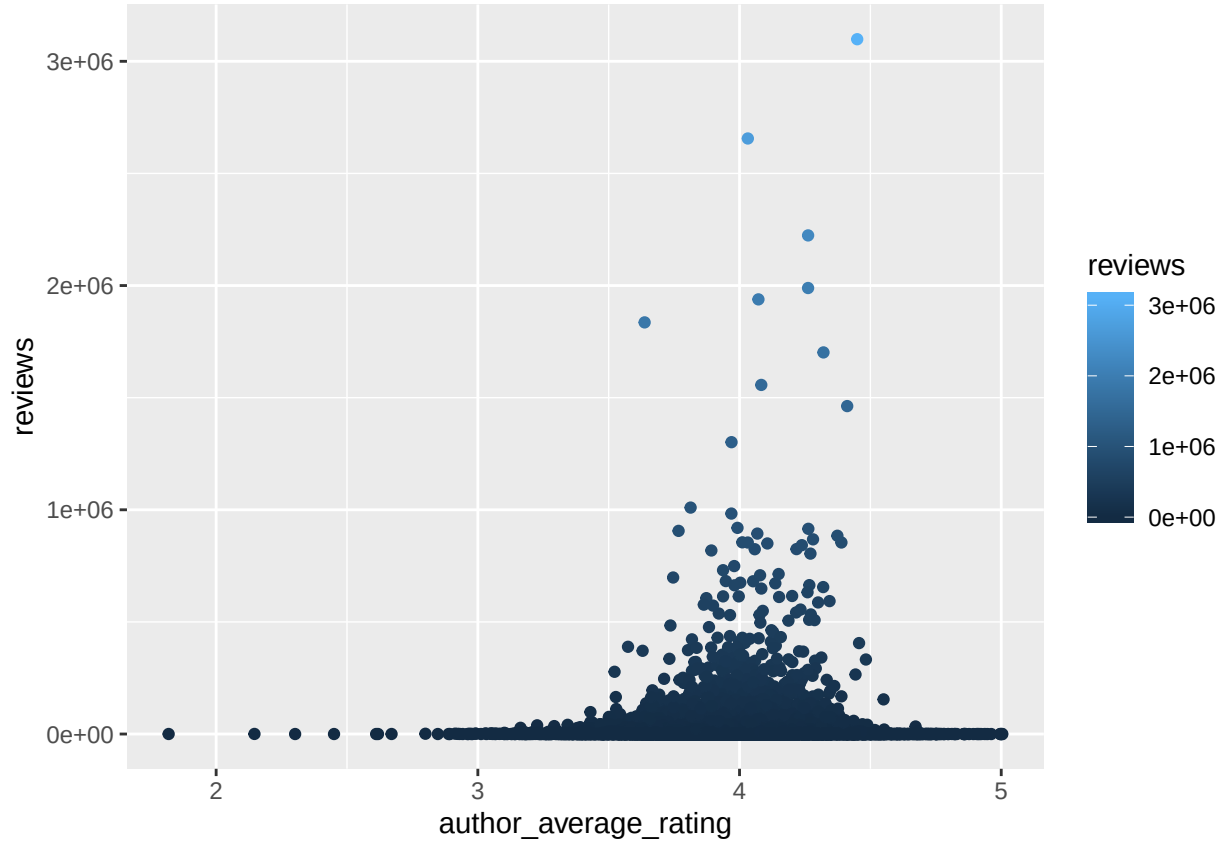
```
rating<- authors %>%
  select(author_name, reviews) %>%
  left_join(select(goodreads, author_name, author_average_rating)) %>%
  group_by(author_name,reviews) %>%
  summarise(author_average_rating = round(mean(author_average_rating), digits = 2))
```

```
## Joining, by = "author_name"
```

```
rating
```

```
## # A tibble: 12,156 x 3
## # Groups:   author_name [12,156]
##    author_name         reviews author_average_rating
##    <chr>                 <int>                 <dbl>
##  1 19                      126                  3.92
##  2 Ã-mer Seyfettin          93                  3.56
##  3 A.  Kirk               1505                  4.2
##  4 A. Digger Stolz          58                  4.18
##  5 A. Lee Martinez        5458                  3.85
##  6 A. Lynden Rolland        81                  3.94
##  7 A. Manette Ansay       3154                  3.39
##  8 A. Meredith Walters   42278                  4
##  9 A. Payne                354                  4.18
## 10 A. Wilding Wells        665                  4.24
## # ... with 12,146 more rows
```

```
ggplot(rating, aes(y=reviews,x=author_average_rating ))+
  geom_jitter(aes(color=reviews))
```



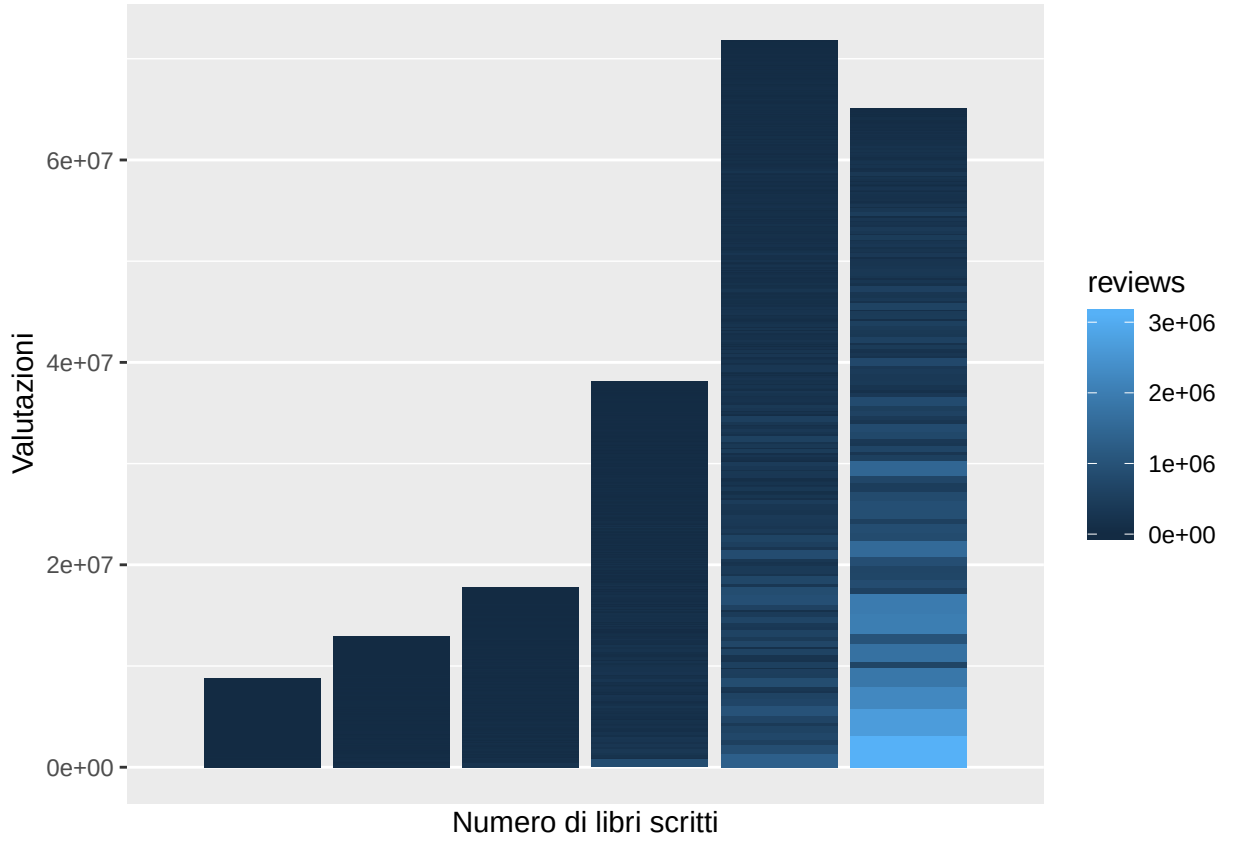

```
summary(rating$author_average_rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.820   3.790   3.950   3.947   4.110   5.000
```

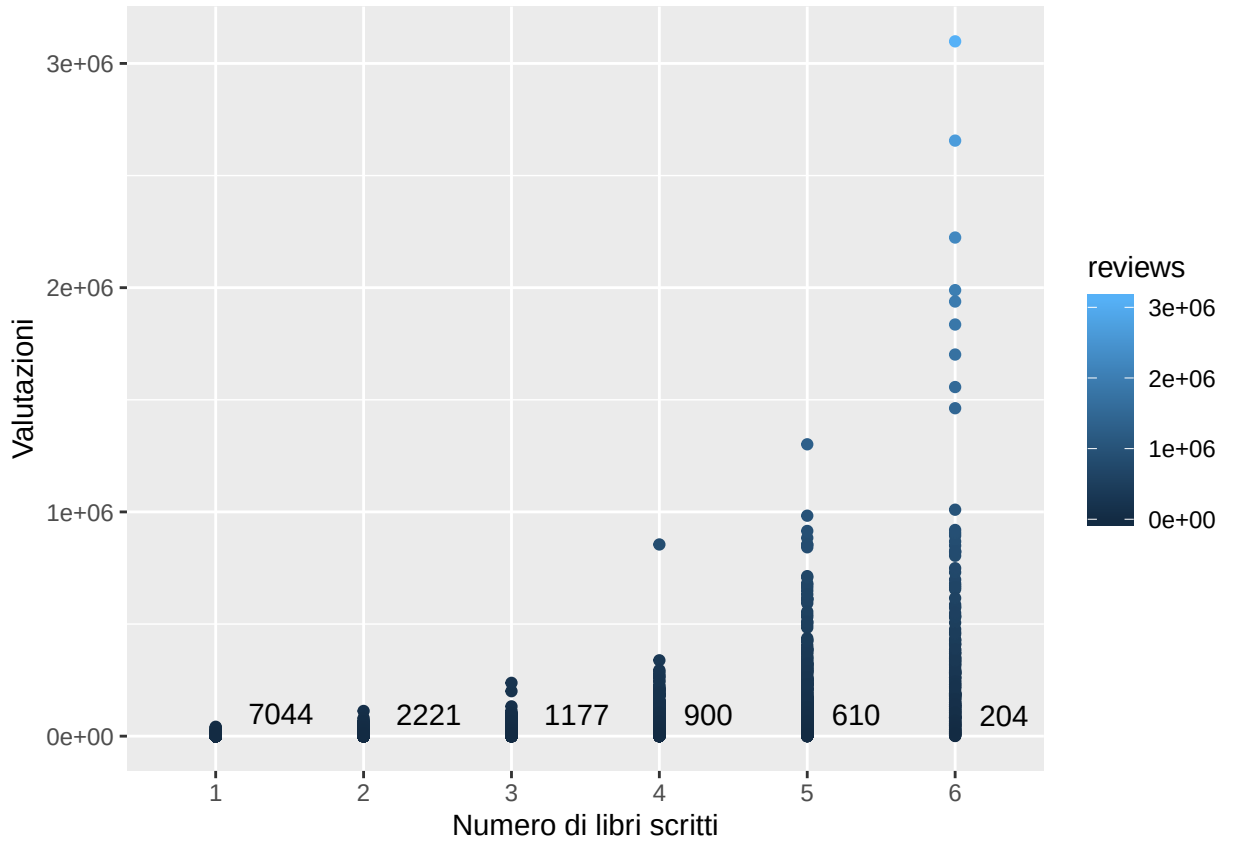
Come si nota dal grafico, le valutazioni sono molto alte, infatti la media è su 4. C'è da notare però che nonostante le valutazioni siano alte, il numero di review è molto basso, quindi chi tende a valutare il libro generalmente non lo recensisce.

C'è una relazione tra il numero di libri scritti e il successo dell'autore?

```
authors %>%
  ggplot(aes(x = n_books, y=reviews))+
    geom_bar(stat="identity", aes(fill=reviews))+
    scale_x_discrete(breaks = c(seq(1,6)))+
    labs(x="Numero di libri scritti", y="Valutazioni")
```

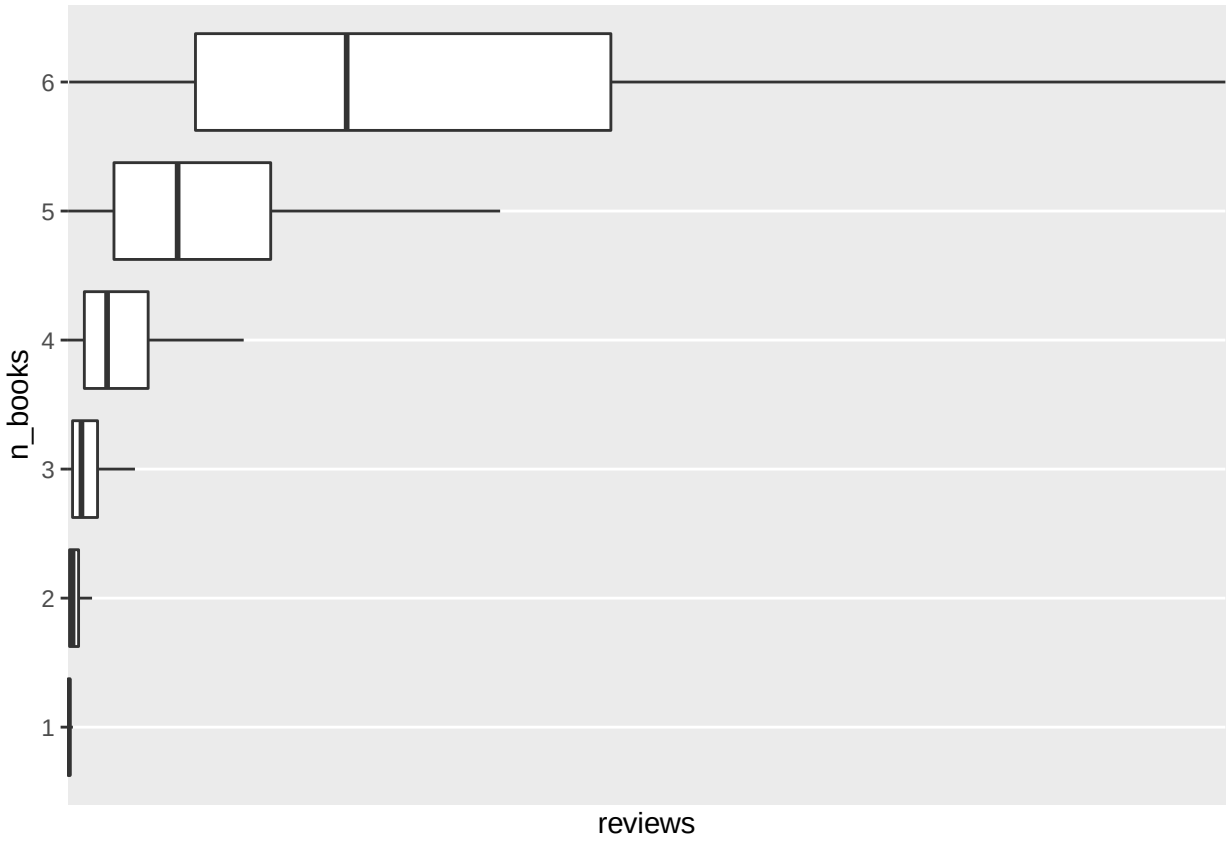



```
authors$n_books = as.factor(authors$n_books)
authors %>%
  ggplot(aes(x = n_books))+
    geom_point(aes(y = reviews,color = reviews))+
    scale_x_discrete(breaks = c(seq(1,6)))+
    labs(x="Numero di libri scritti", y="Valutazioni")+
    geom_text(aes(label=..count..), stat="count", hjust=-0.5, vjust=-0.5)
```



Sembra che all'aumentare dei libri scritti aumentino anche le valutazioni. Probabilmente perchè gli scrittori sono incentivati a scrivere se ricevono feedback positivi dei precedenti capitoli. Questo ovviamente solo nel caso in cui siano famosi. Come emerge dal grafico ci sono comunque molti autori che pur avendo scritto molti libri non sono valutati molto rispetto ad altri.

```
ggplot(authors, aes(n_books,reviews))+
  geom_boxplot(outlier.color = "blue", outlier.shape = NA)+
  coord_flip()+
  scale_y_discrete(breaks=seq(1,10,by=2))
```



Dal boxplot emerge comunque che la mediana aumenta a ogni libro in più scritto. In particolar modo la distribuzione dei dati con 6 libri è più variabile e contiene svariati outlier.

Analisi sui generi

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

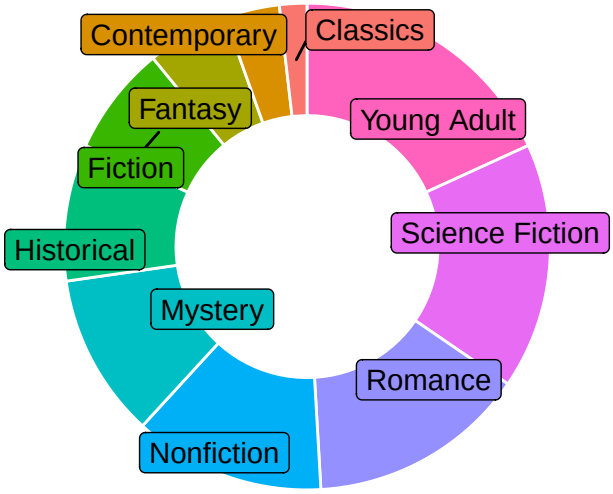
```
par(mfrow = c(2,1))

genres1 <- goodreads %>%
  count(genre_1) %>%
  rename(genre = genre_1)

genres2 <- goodreads %>%
```

```
  count(genre_2) %>%
  rename(genre = genre_2)

genres <- full_join(genres1,genres2, by="genre") %>%
  mutate(n.x = ifelse(is.na(n.x),0,n.x),
         n.y = ifelse(is.na(n.y),0,n.y))%>%
  group_by(genre) %>%
  summarise(n = n.x+n.y)
```

```
## Warning: Column `genre` joining factors with different levels, coercing to
## character vector
```

```
loved_genres<- genres %>%
  arrange(-n) %>%
  head(10) %>%
ggplot(aes(x = 2,y=genre, fill=genre))+
  geom_bar(stat="identity", color="white", show.legend = FALSE)+
  coord_polar(theta="y", start=0)+
  ggrepel::geom_label_repel(aes(label=genre),
                            position=position_stack(vjust=0.5),
                            show.legend = FALSE)+
  theme(legend.position="none")+
  theme_void()+
  xlim(0.5,2.5)+
  labs(title="I generi più amati")

hollow_genres <- genres %>%
  arrange(n) %>%
  head(10) %>%
  ggplot(aes(x = 2,y=genre, fill=genre))+
  geom_bar(stat="identity", color="white", show.legend = FALSE)+
  coord_polar(theta="y", start=0)+
  ggrepel::geom_label_repel(aes(label=genre),
                            position=position_stack(vjust=0.5),
                            show.legend = FALSE)+
  theme(legend.position="none")+
  theme_void()+
  xlim(0.5,2.5)+
  labs(title="I generi di nicchia")

grid.arrange(loved_genres,hollow_genres,ncol=2)
```

I generi più amati



I generi di nicchia

Dai due grafici a ciambella emergono i generi più letti e quelli più di nicchia, ovvero i cui testi sono rari.

Gender gap?

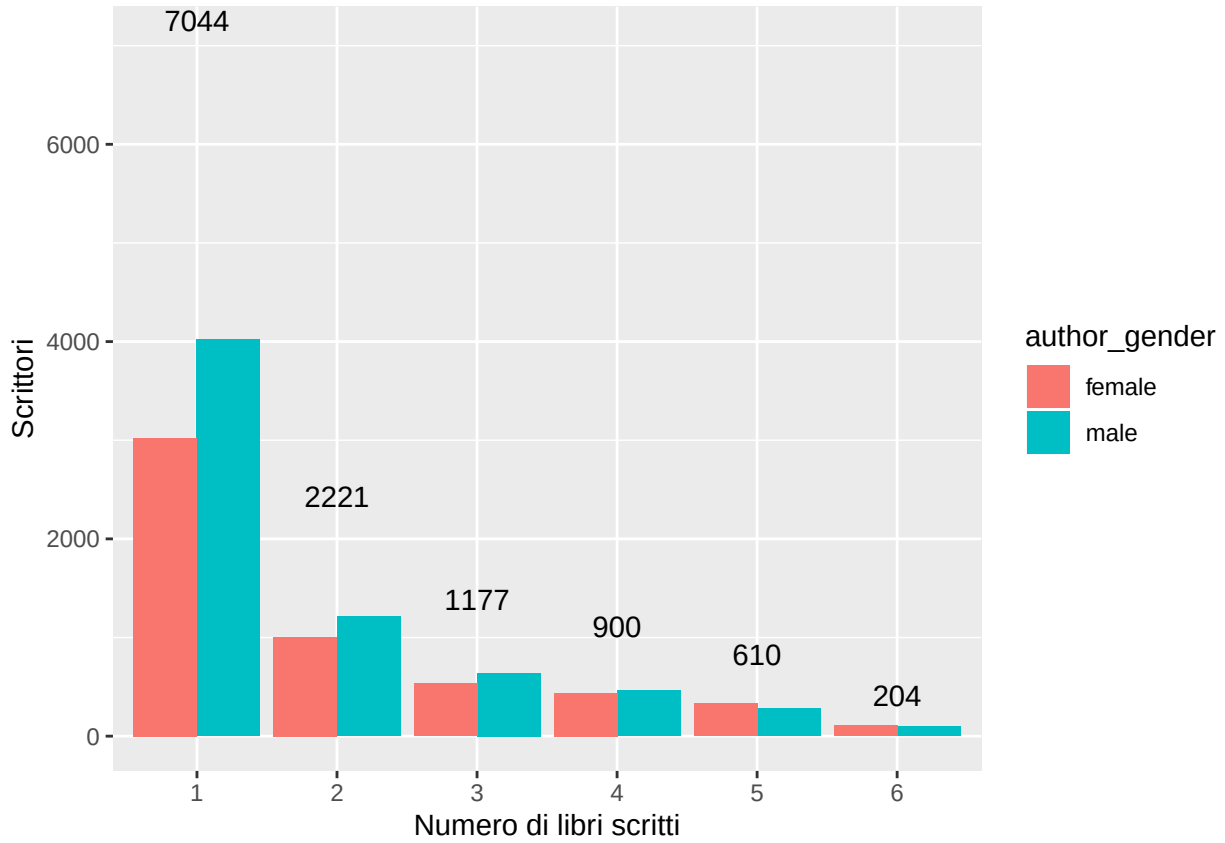
Analizziamo ora il sesso degli scrittori, per analizzare se vi sono più scrittrici o scrittori e se c'è una relazione tra genere e successo.

```
authors %>%
  group_by(author_gender) %>%
  summarise(n=n())
```

```
## # A tibble: 2 x 2
##   author_gender     n
##   <fct>         <int>
## 1 female         5439
## 2 male           6717
```

Da una prima analisi emerge che vi sono più scrittori maschi che femmine. Analizziamo ora la relazione tra numero di libri scritti, reviews e genere dello scrittore.

```
authors %>%
  ggplot(aes(x = n_books))+
    geom_bar(stat="count",position="dodge", aes(fill=author_gender))+
    labs(x="Numero di libri scritti", y="Scrittori")+
    geom_text(aes(label=..count..), stat="count", vjust=-0.5)
```

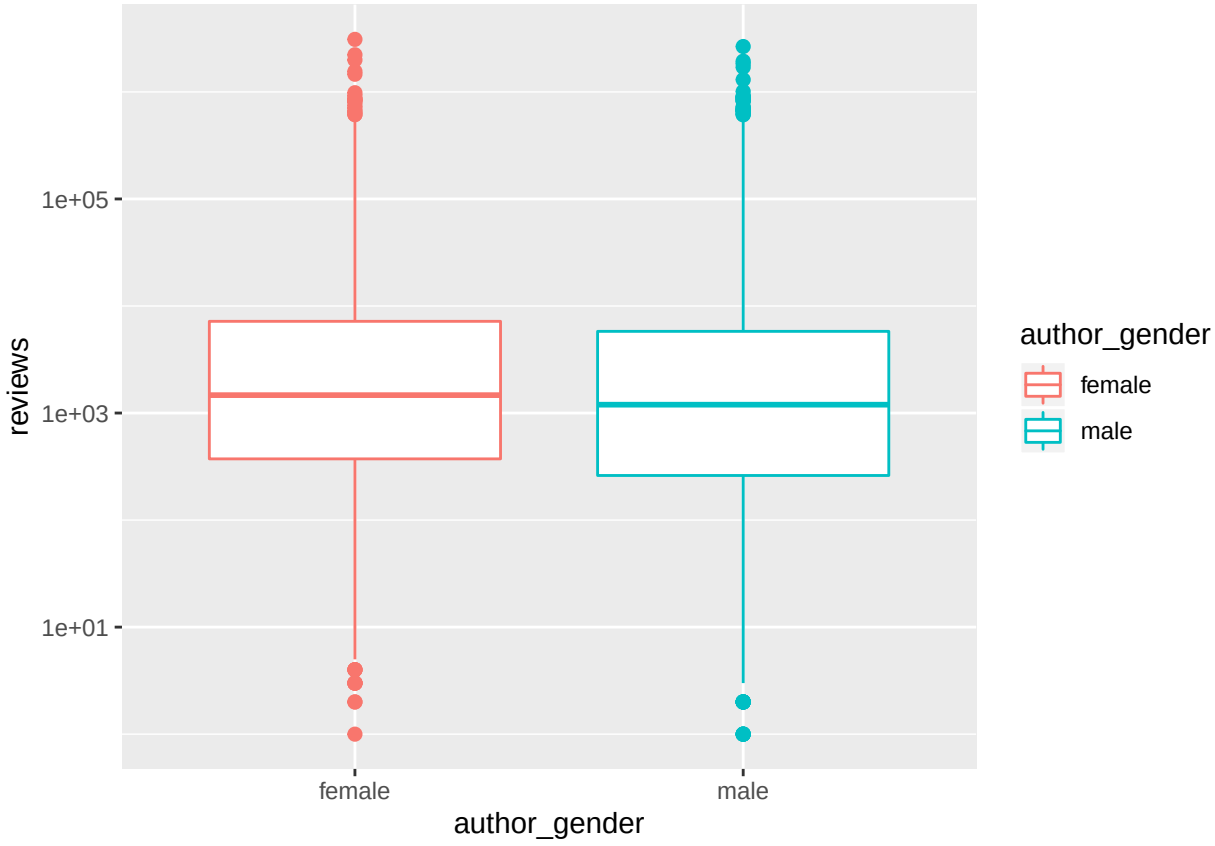


Non sembra esserci una differenza sostanziale, ma proviamo a esplorare la relazione tra numero di reviews e sesso.

```
ggplot(authors, aes(x=author_gender, y=reviews, color=author_gender))+
  geom_boxplot(outlier.size=2)+
  scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```




```
summary(authors$reviews)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     303    1291   17644    6427 3098497
```

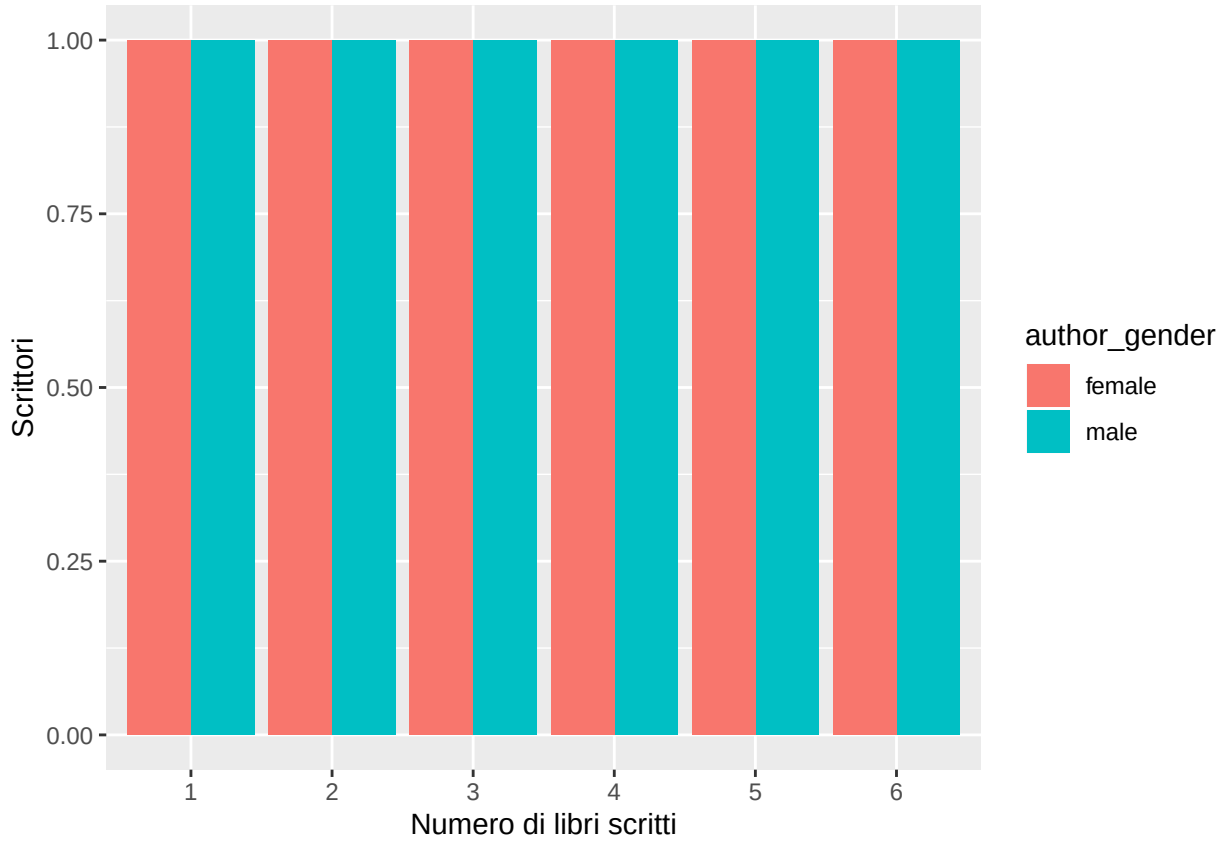
La distribuzione dei dati delle scrittrici sembra essere leggermente superiore rispetto a quella maschile, anche se la mediana sembra coincidere. Gli outlier superiori, quindi gli scrittori di successo, non sembrano differire di numero, mentre quelli inferiori sembrano essere il doppio rispetto agli scrittori maschi.

```
#Sopra il quantile 75%
authors[authors$reviews > quantile(authors$reviews,0.75) | authors$reviews < quantile(authors$reviews,0
  transform(quantile = ifelse(reviews > quantile(authors$reviews,0.75),0.75,0.25)) %>%
  group_by(author_gender, quantile) %>%
  summarise(n=n())
```

```
## # A tibble: 4 x 3
## # Groups:   author_gender [2]
##   author_gender quantile     n
##   <fct>            <dbl> <int>
## 1 female            0.25  1222
## 2 female            0.75  1447
## 3 male              0.25  1814
## 4 male              0.75  1592
```

La maggioranza dei valori al di fuori del box per i maschi si trova al di sotto del quantile 0.25, mentre per le femmine si trova sopra il quantile 0.75. Proviamo ad analizzare le frequenze relative

```
n_writers <- nrow(authors)

authors %>%
  ggplot(aes(x = n_books))+
    geom_bar(position="dodge", aes(y=..prop..,fill=author_gender))+
    labs(x="Numero di libri scritti", y="Scrittori")
```



In termini di frequenze relative si può concludere che il numero di valutazioni

Nazionalità degli scrittori

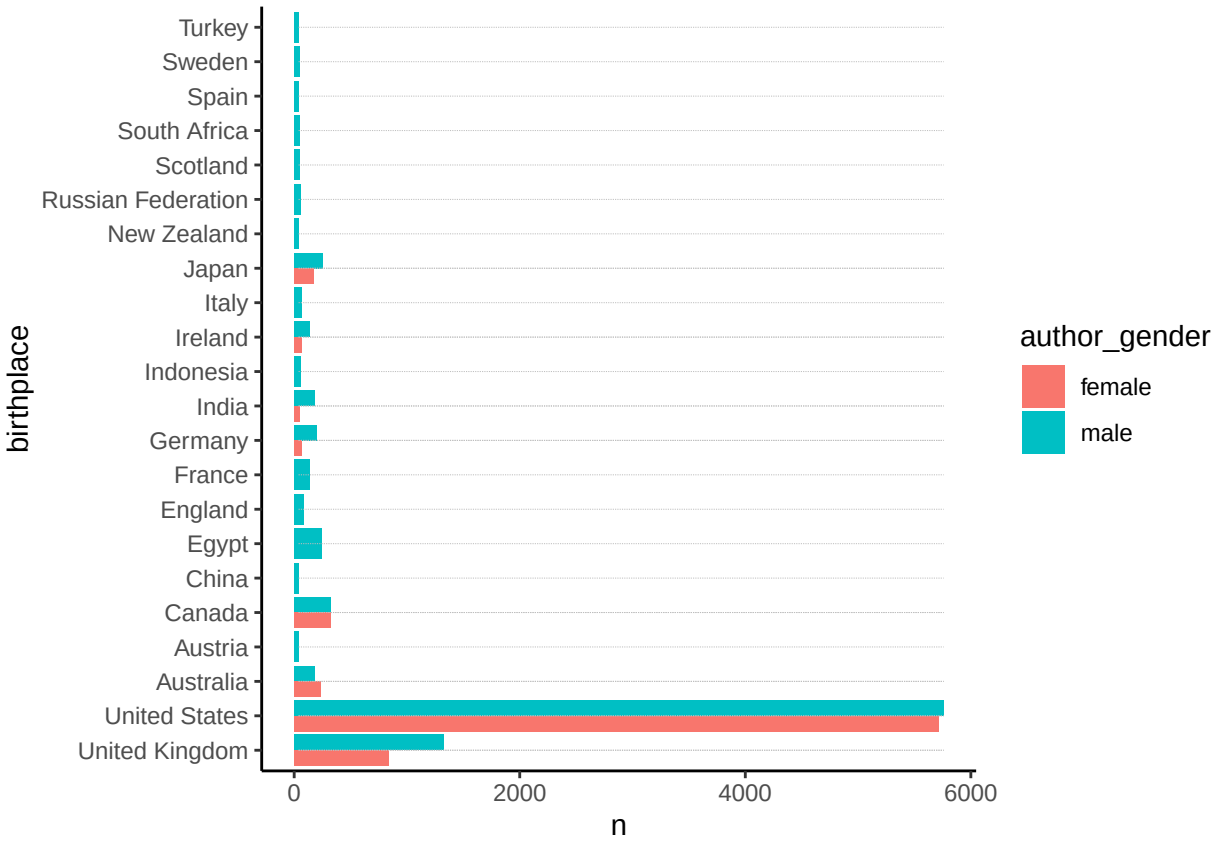
```
goodreads %>%
  select(birthplace) %>%
  count(birthplace) %>%
  arrange(-n) %>%
  head(30)
```

```
## # A tibble: 30 x 2
##    birthplace            n
##    <chr>             <int>
##  1 " United States"  11471
##  2 unknown            4414
##  3 " United Kingdom"  2164
##  4 Canada              646
```

```
##  5 Japan               428
##  6 Australia           420
##  7 Germany             263
##  8 Egypt               250
##  9 India               230
## 10 Ireland             207
## # ... with 20 more rows
```

Di molti scrittori non è nota la nazionalità, ma la maggioranza proviene dagli stati uniti. Sembra comunque prevalere la lingua inglese sulle altre, probabilmente perchè l'applicazione è ideata da uno statunitense e pensata per essere monolingua. Tralasciando le nazionalità sconosciute

```
top30countries <- goodreads %>%
  select(birthplace, author_gender) %>%
  count(birthplace, author_gender) %>%
  filter(birthplace !="unknown") %>%
  arrange(-n) %>%
  head(30)


ggplot(top30countries,aes(birthplace,n))+
  geom_bar(stat="identity", position="dodge", aes(fill=author_gender))+
  geom_segment(aes(x=birthplace,
                   xend=birthplace,
                   y=min(n),
                   yend=max(n)),
               linetype="dashed",
               size=0.05,
               color="grey")+
  coord_flip()+
  theme_classic()
```


Turkey
Sweden
Spain
South Africa
Scotland
Russian Federation
New Zealand
Japan
Italy
Ireland
Indonesia
India
Germany
France
England
Egypt
China
Canada
Austria
Australia
United States
United Kingdom
birthplace
0
2000
4000
6000
n
author_gender
female
male