Midterm Project

Atticus Wang

January 17, 2022

In this project we analyze how a book's rating is related to variables such as genre, publisher, country, format, and others.

```
# Load packages

library(tidyverse)
```

We use the following datasets:

- Each observation of `bx_ratings` consists of the ISBN number of the book being reviewed, the book reviewer's ID, and the ratings given (on a scale from 1 to 10).
- Each observation of `bx_users` consists of the person's ID, their location (country, state, city), and their age.
- Each observation of `bx_books` consists of information related to the book: ISBN, title, author(s), year of publication, and publisher.

The three datasets above were collected from the online community Book-Crossing by Cai-Nicolas Ziegler in 2004.

- The file `goodreads_big` contains information about 100000 books collected from Goodreads. Information includes genres and average ratings, among other things. This file was collected by Manav Dhamani.
- The file `goodreads_small` contains more refined information about around 10000 books collected from Goodreads. Information includes language, publication date, and others. This file was collected by Soumik.

All files above are freely available for download on Kaggle.

```
# Read files from folder

bx_ratings <- read_csv2("/Users/atticus_w/Desktop/School/21-22 (Senior)/Data Science/project/BX-Book-Ra

bx_books <- read_csv2("/Users/atticus_w/Desktop/School/21-22 (Senior)/Data Science/project/BX-Books.csv

bx_users <- read_csv2("/Users/atticus_w/Desktop/School/21-22 (Senior)/Data Science/project/BX-Users.csv

goodreads_big <- read_csv("/Users/atticus_w/Desktop/School/21-22 (Senior)/Data Science/project/goodreads

goodreads_small <- read_csv("/Users/atticus_w/Desktop/School/21-22 (Senior)/Data Science/project/goodrea
```

Goodreads data

We first tidy up the `goodreads_big` data: this includes deleting useless information, filtering out null and NA values as well as books that have been rated relatively few times. Then we perform some analyses (see

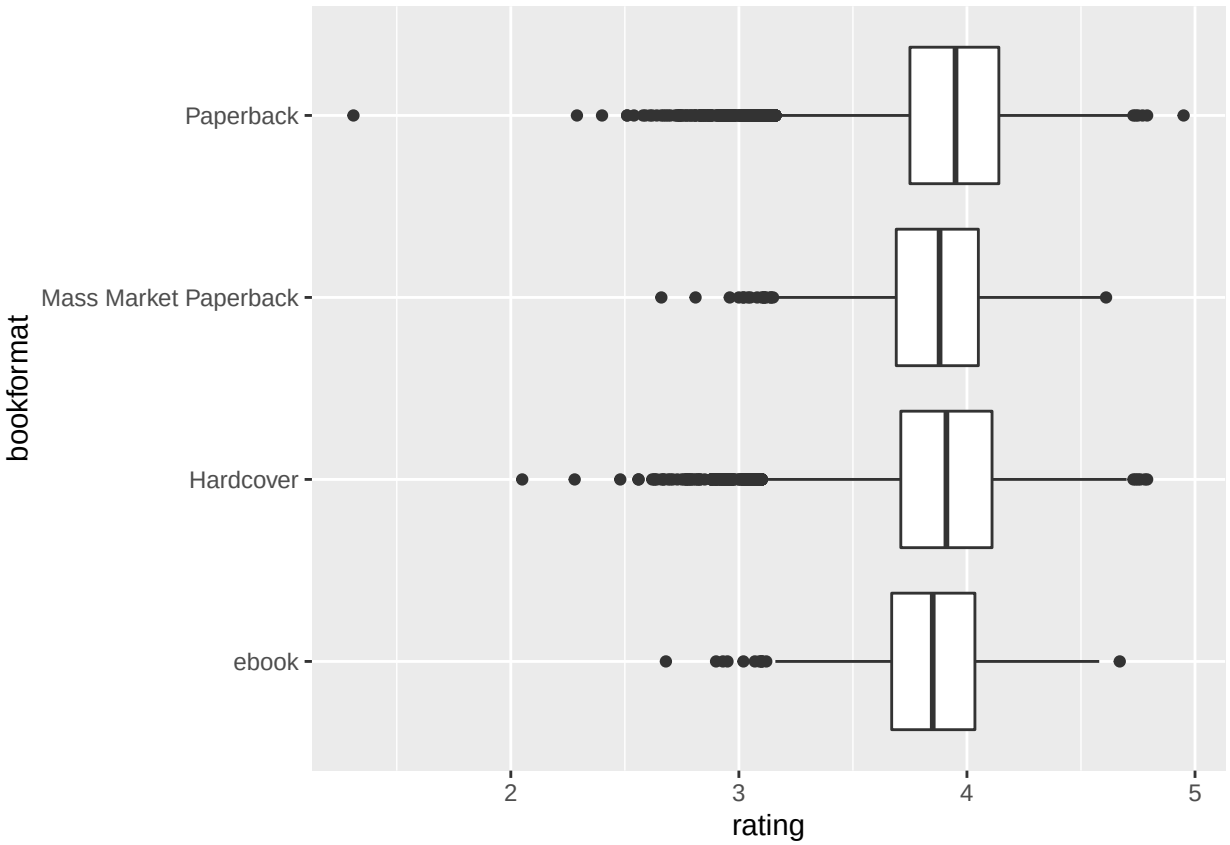
below):

```
goodreads_tidy <- goodreads_big %>%
  select(-img, -link) %>%
  filter(pages != 0 & is.na(isbn13) == FALSE & totalratings >= 100) %>%
  filter(pages <= 2000) %>%
  filter(reviews >= 10) %>%
  mutate(num_authors = 1 + str_count(author, "[,]")) %>%
  filter(num_authors <= 10) %>%
  filter(is.na(isbn) == FALSE)
```

```
# Which book formats are the most common?
goodreads_tidy %>%
  group_by(bookformat) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 75 x 2
##    bookformat            count
##    <chr>                 <int>
##  1 Paperback             24242
##  2 Hardcover             14021
##  3 Mass Market Paperback  2085
##  4 ebook                  1735
##  5 <NA>                    251
##  6 Audio CD                101
##  7 Unknown Binding          81
##  8 Board Book               72
##  9 Trade Paperback          32
## 10 Audiobook                27
## # ... with 65 more rows
```

```
# Is a book's rating related to its format? Not really.
goodreads_tidy %>%
  filter(bookformat %in% c("Paperback","Hardcover","ebook","Mass Market Paperback")) %>%
  ggplot(aes(x = rating, y = bookformat)) +
  geom_boxplot()
```

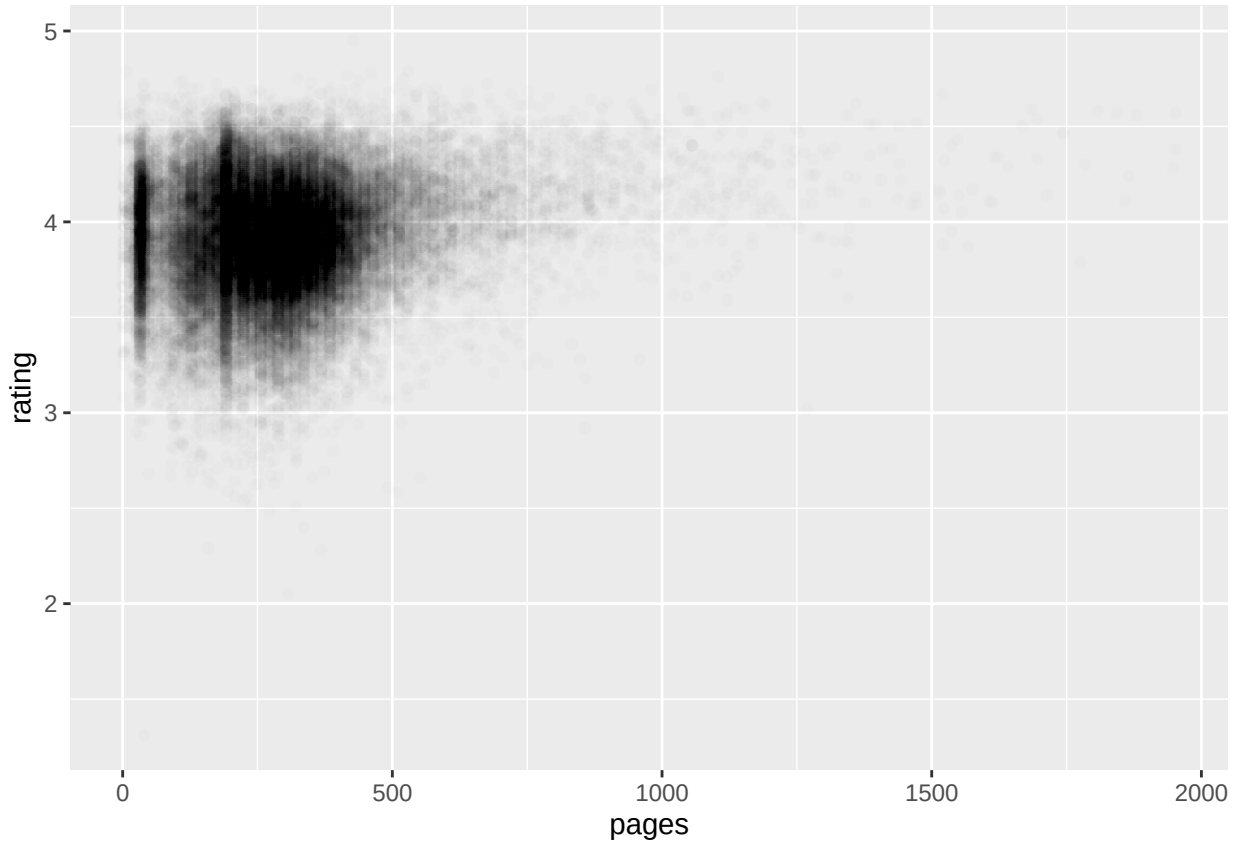



```
# Which books have the highest rating? The answer is a bit unexpected
# (spiritual and religious books, art and photography, and manga series).
goodreads_tidy %>%
  arrange(desc(rating)) %>%
  head(100)
```

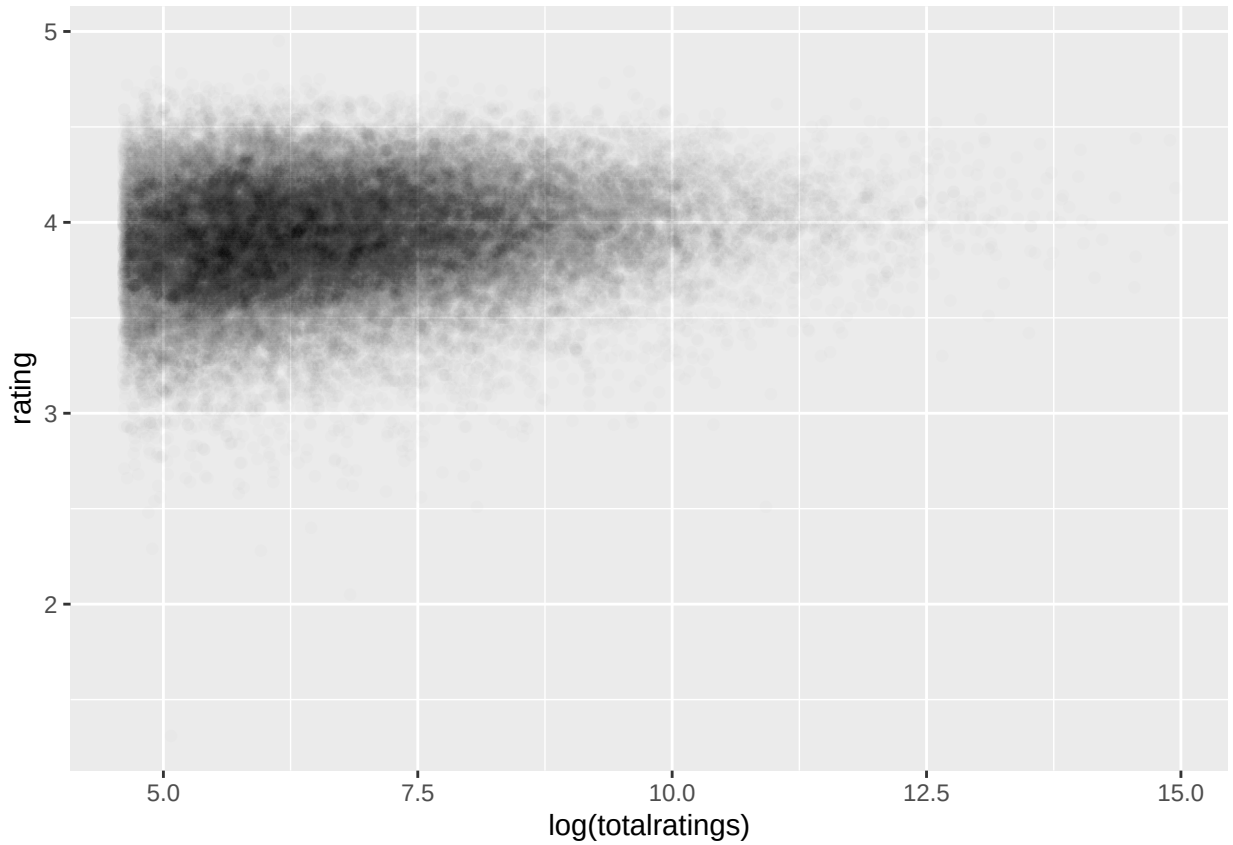
```
## # A tibble: 100 x 12
##    author   bookformat desc    genre   isbn   isbn13 pages rating reviews title
##    <chr>    <chr>      <chr>   <chr>   <chr>   <dbl> <dbl>  <dbl>   <dbl> <chr>
##  1 Paramah~ Paperback  "â€~Th~ Spirit~ 1606~ 9.78e12   427   4.95      74 Livin~
##  2 Brandon~ Paperback  "Accor~ Fantas~ 5751~ 9.78e12   530   4.79     741 The W~
##  3 Francis~ Hardcover  "This ~ Christ~ 9061~ 9.78e12     7   4.79      18 In Co~
##  4 James N~ Hardcover  "A doc~ Art,Ph~ 7148~ 9.78e12   460   4.78      19 Infer~
##  5 Paul Se~ Paperback  "This ~ Spirit~ 3991~ 9.78e12   416   4.77      19 The B~
##  6 Hayao M~ Hardcover  "[ ,\n~ Sequen~ 1421~ 9.78e12  1104   4.76     162 Nausi~
##  7 Abdu'l-~ Paperback  "This ~ Religi~ 8774~ 9.78e12   324   4.75      15 Some ~
##  8 Sebasti~ Hardcover  "â€œIn~ Art,Ph~ 3836~ 9.78e12   520   4.75      36 Genes~
##  9 Deborah~ Hardcover  "This ~ Crafts~ 1603~ 9.78e12   438   4.74      50 The F~
## 10 Jacques~ Paperback  "This ~ Christ~ 8189~ 9.78e12   112   4.74     240 Searc~
## # ... with 90 more rows, and 2 more variables: totalratings <dbl>,
## #   num_authors <dbl>
```

```
# Is a book's rating related to its length? Longer books tend to have a
# higher average rating, and shorter books are more varied. Also, we noticed
# there are not a lot of books with length between roughly 50 to 75 pages
# (there is an obvious gap on the plot).
goodreads_tidy %>%
  ggplot(mapping = aes(x = pages, y = rating)) +
```

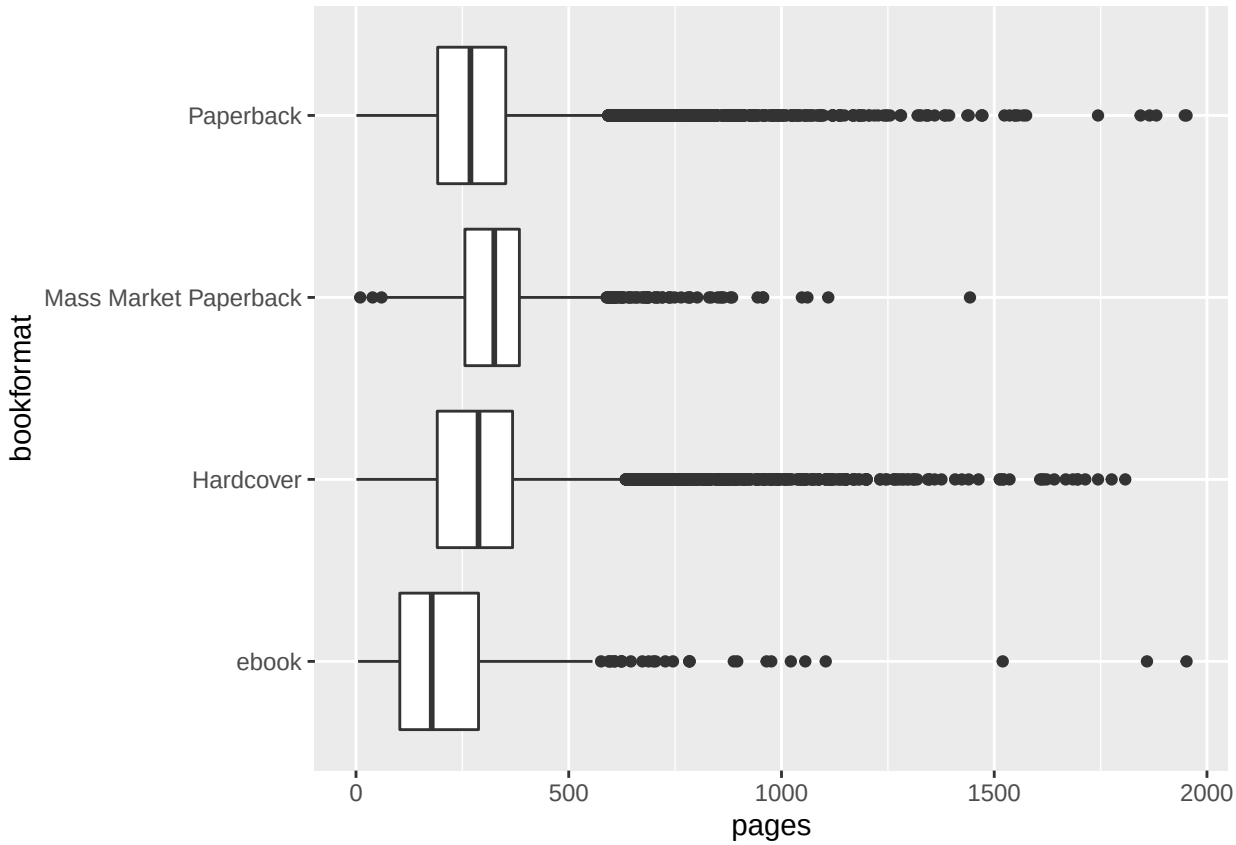
```
  geom_point(alpha = 0.01)
```




```
# Is a book's rating related to the number of its posted ratings? We found
# that books with more ratings have a slightly higher average rating.
goodreads_tidy %>%
  ggplot(mapping = aes(x = log(totalratings), y = rating)) +
  geom_point(alpha = 0.01)
```

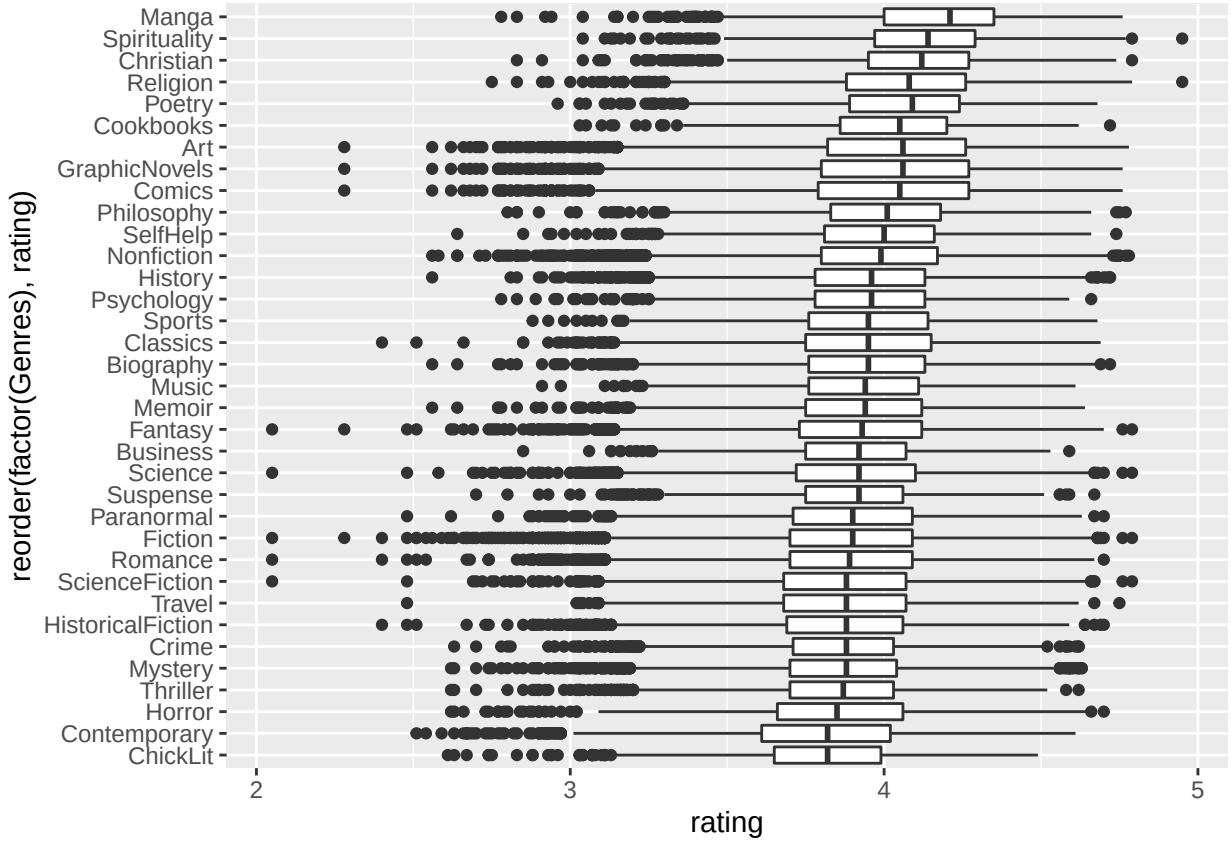



```
# Is a book's format related to its length? I expected paperback and mass
# market paperback books to be shorter, but it turns out to be false.
goodreads_tidy %>%
  filter(bookformat %in% c("Paperback","Hardcover","ebook","Mass Market Paperback")) %>%
  ggplot(mapping = aes(x = pages, y = bookformat)) +
  geom_boxplot()
```




```
# Now we take into account the genres. I created a column of logical
# variables for each major listed genre:
goodreads_genres <- goodreads_tidy %>%
  mutate(Art = str_detect(genre, "Art"),
         Biography = str_detect(genre, "Biography"),
         Business = str_detect(genre, "Business"),
         ChickLit = str_detect(genre, "Chick Lit"),
         Childrens = str_detect(genre, "Children's"),
         Christian = str_detect(genre, "Christian"),
         Classics = str_detect(genre, "Classics"),
         Comics = str_detect(genre, "Comics"),
         Contemporary = str_detect(genre, "Contemporary"),
         Cookbooks = str_detect(genre, "Cookbooks"),
         Crime = str_detect(genre, "Crime"),
         Ebooks = str_detect(genre, "Ebooks"),
         Fantasy = str_detect(genre, "Fantasy"),
         Fiction = str_detect(genre, "Fiction"),
         GayAndLesbian = str_detect(genre, "Gay and Lesbian"),
         GraphicNovels = str_detect(genre, "Graphic Novels"),
         HistoricalFiction = str_detect(genre, "Historical Fiction"),
         History = str_detect(genre, "History"),
         Horror = str_detect(genre, "Horror"),
         HumorAndComedy = str_detect(genre, "Humor and Comedy"),
         Manga = str_detect(genre, "Manga"),
         Memoir = str_detect(genre, "Memoir"),
         Music = str_detect(genre, "Music"),
         Mystery = str_detect(genre, "Mystery"),
```

```
         Nonfiction = str_detect(genre, "Nonfiction"),
         Paranormal = str_detect(genre, "Paranormal"),
         Philosophy = str_detect(genre, "Philosophy"),
         Poetry = str_detect(genre, "Poetry"),
         Psychology = str_detect(genre, "Psychology"),
         Religion = str_detect(genre, "Religion"),
         Romance = str_detect(genre, "Romance"),
         Science = str_detect(genre, "Science"),
         ScienceFiction = str_detect(genre, "Science Fiction"),
         SelfHelp = str_detect(genre, "Self Help"),
         Suspense = str_detect(genre, "Suspense"),
         Spirituality = str_detect(genre, "Spirituality"),
         Sports = str_detect(genre, "Sports"),
         Thriller = str_detect(genre, "Thriller"),
         Travel = str_detect(genre, "Travel"),
         YoungAdult = str_detect(genre, "YoungAdult"))

# Which genre has the highest median rating? Contrary to our expectation
# (which are fantasy novels), the highest ranking ones are Manga,
# Spirituality, Christian, Religion, Poetry, Cookbooks, and Art.
goodreads_genres %>%
  select(-author,-bookformat,-desc,-genre,-isbn,-isbn13,-pages,-reviews,-title,-totalratings,-num_autho
  pivot_longer(c(Art, Biography, Business, ChickLit, Childrens, Christian, Classics, Comics, Contempora
  filter(Value == TRUE) %>%
  ggplot(aes(x = rating, y = reorder(factor(Genres), rating))) +
  geom_boxplot()
```



Next, we analyze the smaller but more detailed dataset (`goodreads_small`). Against, we first tidy up the data, then perform some analyses (see below).

```
# In particular, we create a column (author1) containing the primary authors.
goodreads_small_tidy <- goodreads_small %>%
  filter(ratings_count >= 1000) %>%
  filter(is.na(isbn) == FALSE & is.na(isbn13) == FALSE) %>%
  separate(publication_date, into = c("publication_month", "publication_date", "publication_year")) %>%
  separate(authors, into = c("author1", "author2", "author3"), sep = "/")
```

```
# Which authors who wrote at least 3 books have the highest rating? The
# top 10 are Bill Watterson, Hiromu Arakawa, Hayao Miyazaki, J.K. Rowling,
# James Herriot, Karen Kingsbury, Arthur Conan Doyle. Anton Chekhov.
# Edgar Allan Poe, and Viktor E. Frankl.
goodreads_small_tidy %>%
  group_by(author1) %>%
  summarize(author_avg_rating = mean(average_rating), count = n()) %>%
  filter(count >= 3) %>%
  arrange(desc(author_avg_rating))
```

```
## # A tibble: 536 x 3
##    author1            author_avg_rating count
##    <chr>                          <dbl> <int>
##  1 Bill Watterson                  4.72     6
##  2 Hiromu Arakawa                  4.57    12
##  3 Hayao Miyazaki                  4.56     6
##  4 J.K. Rowling                    4.54    14
##  5 James Herriot                   4.42     4
##  6 Karen Kingsbury                 4.41    12
##  7 Arthur Conan Doyle              4.38     8
##  8 Anton Chekhov                   4.38     4
##  9 Edgar Allan Poe                 4.36     4
## 10 Viktor E. Frankl                4.36     3
## # ... with 526 more rows
```

```
# Books from which year have the highest average rating? Turns out to be
# years 1984, 86 and 88.
goodreads_small_tidy %>%
  group_by(publication_year) %>%
  summarize(year_avg_rating = mean(average_rating), count = n()) %>%
  filter(count >= 9) %>%
  arrange(desc(year_avg_rating))
```

```
## # A tibble: 32 x 3
##    publication_year year_avg_rating count
##    <chr>                      <dbl> <int>
##  1 1984                        4.12    21
##  2 1986                        4.07    27
##  3 1988                        4.04    28
##  4 1979                        4.03    12
##  5 1989                        4.02    46
##  6 1990                        4.02    50
##  7 1992                        4.02    80
##  8 1985                        4.02    22
##  9 1995                        4.01   106
## 10 2012                        4.00    12
```

```
## # ... with 22 more rows
```

```
# Books from which publisher have the highest average rating? The top three
# are VIZ Media LLC (which mainly publishes mangas), Tyndale House Publishers
# (which mainly publishes spiritual and religious books), and Vertigo (a
# part of DC Comics).
goodreads_small_tidy %>%
  group_by(publisher) %>%
  summarize(publisher_avg_rating = mean(average_rating), count = n()) %>%
  filter(count >= 10) %>%
  arrange(desc(publisher_avg_rating))
```

```
## # A tibble: 124 x 3
##    publisher                 publisher_avg_rating count
##    <chr>                                    <dbl> <int>
##  1 VIZ Media LLC                             4.34    53
##  2 Tyndale House Publishers                  4.27    19
##  3 Vertigo                                   4.24    37
##  4 Houghton Mifflin Harcourt                 4.19    20
##  5 Everyman's Library                        4.16    14
##  6 Scholastic                                4.14    17
##  7 Touchstone                                4.10    10
##  8 Farrar  Straus and Giroux                 4.09    24
##  9 Greenwillow Books                         4.09    13
## 10 Thomas Nelson                             4.09    11
## # ... with 114 more rows
```

BookCrossing data

As usual, we tidy up the data, remove duplicates, and do some analysis. Here we also use relational data, combining data from all three BX datasets.

```
bx_books <- bx_books %>%
  select(-`Image-URL-L`,-`Image-URL-M`,-`Image-URL-S`)

bx_books %>%
  count(ISBN) %>%
  filter(n > 1)
```

```
## # A tibble: 1 x 2
##   ISBN           n
##   <chr>      <int>
## 1 0486404242     2
```

```
bx_books <- bx_books %>%
  filter(ISBN != 0486404242)

bx_ratings <- bx_ratings %>%
  filter(`Book-Rating` != 0)

# Join data into one dataframe
bx_all <- bx_ratings %>%
  left_join(bx_books, by = "ISBN") %>%
  left_join(bx_users, by = "User-ID") %>%
  filter(is.na(`Book-Title`) == FALSE) %>%
```

```
  filter(`Year-Of-Publication` != 0)

# Books from which year have the highest average rating? It's 1970, 71 and 76.
bx_all %>%
  group_by(`Year-Of-Publication`) %>%
  summarize(avg_rating = mean(`Book-Rating`), count = n()) %>%
  filter(count >= 100) %>%
  arrange(desc(avg_rating))
```

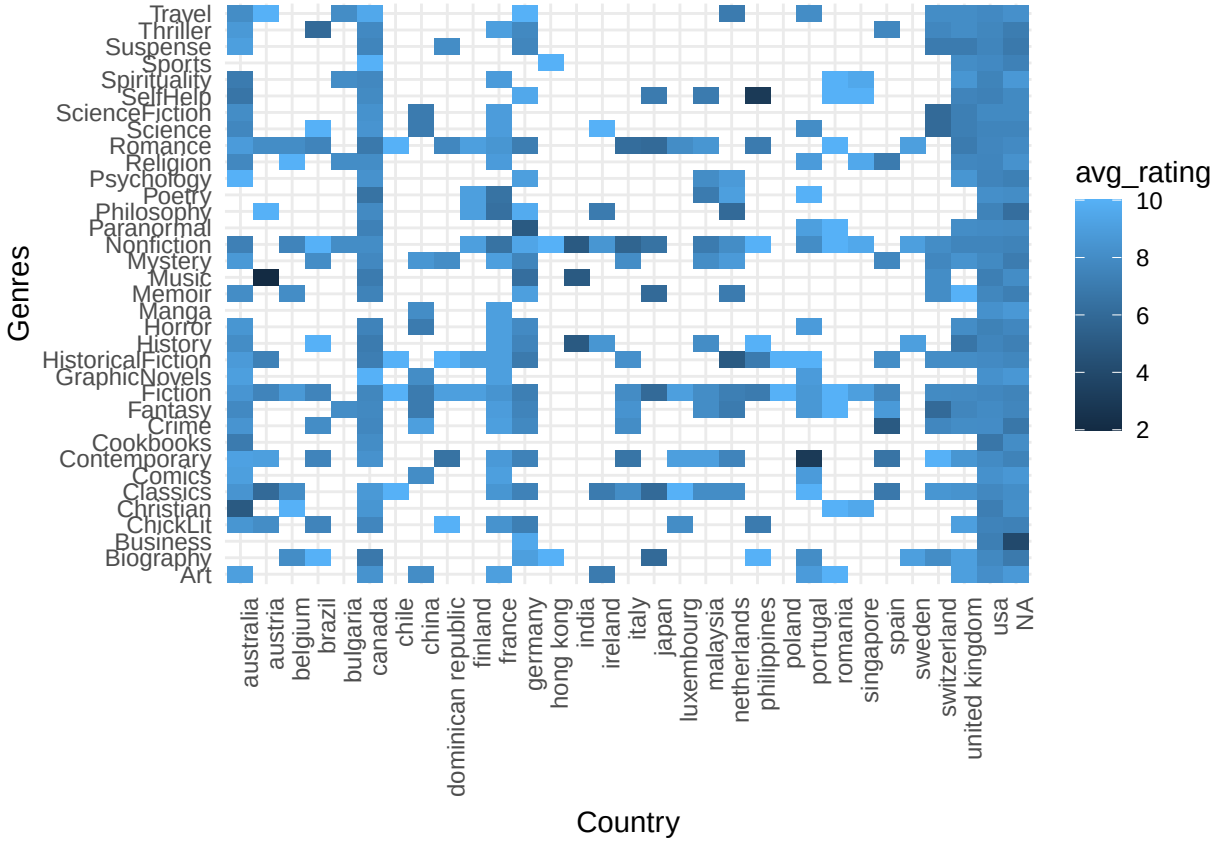
```
## # A tibble: 36 x 3
##    `Year-Of-Publication` avg_rating count
##                    <dbl>      <dbl> <int>
##  1                  1970       7.96   186
##  2                  1971       7.83   178
##  3                  1976       7.80   566
##  4                  1980       7.71   738
##  5                  1974       7.71   327
##  6                  1979       7.71   562
##  7                  1984       7.70  1890
##  8                  1986       7.68  2368
##  9                  2003       7.68 11410
## 10                  1983       7.65  1617
## # ... with 26 more rows
```

```
# We also join data from the goodreads dataset to get information about
# genres.
goodreads_reduced <- goodreads_genres %>%
  mutate(ISBN = isbn) %>%
  select(ISBN, rating, Art, Biography, Business, ChickLit, Childrens, Christian, Classics, Comics, Cont

# Isolate the book reviewer's country for the next step
bx_countries <- bx_all %>%
  dplyr::mutate(Country = str_extract(Location, "[^,]+$")) %>%
  left_join(goodreads_reduced, by = "ISBN") %>%
  filter(is.na(rating) == FALSE)

bx_countries
```

```
## # A tibble: 1,029 x 51
##    `User-ID` ISBN       `Book-Rating` `Book-Title` `Book-Author` `Year-Of-Public~
##        <dbl> <chr>              <dbl> <chr>        <chr>                    <dbl>
##  1    276872 8571648972            10 "Estação Ca~ Drauzio Vare~             1999
##  2    276875 8838910987             8 "Il birraio~ Andrea Camil~             1995
##  3    276925 002542730X            10 "Politicall~ James Finn G~             1994
##  4    276925 8423976645             8 "Olvidado R~ Ana Maria Ma~             1996
##  5    276939 2253144452             9 "La R\xc3?\~ Bernard Werb~             1998
##  6    276939 226612269X             9 "Ou Es-Tu?"  Marc Levy                  2003
##  7    277359 1591820537             4 "Paradise K~ Ai Yazawa                 2002
##  8    277427 002542730X            10 "Politicall~ James Finn G~             1994
##  9    277427 1572304510            10 "Without Co~ Robert D. Ha~             1999
## 10    277435 031215125X             8 "The Far Pa~ M. M. Kaye                 1997
## # ... with 1,019 more rows, and 45 more variables: Publisher <chr>,
## #   Location <chr>, Age <chr>, Country <chr>, rating <dbl>, Art <lgl>,
## #   Biography <lgl>, Business <lgl>, ChickLit <lgl>, Childrens <lgl>,
```

```
## #   Christian <lgl>, Classics <lgl>, Comics <lgl>, Contemporary <lgl>,
## #   Cookbooks <lgl>, Crime <lgl>, Ebooks <lgl>, Fantasy <lgl>, Fiction <lgl>,
## #   GayAndLesbian <lgl>, GraphicNovels <lgl>, HistoricalFiction <lgl>,
## #   History <lgl>, Horror <lgl>, HumorAndComedy <lgl>, Manga <lgl>, ...
```

```r
# Is there a relationship between a country and its preferences in genres?
bx_countries %>%
  select(-`User-ID`, -Location) %>%
  pivot_longer(c(Art, Biography, Business, ChickLit, Childrens, Christian, Classics, Comics, Contempora
  filter(Value == TRUE) %>%
  group_by(Country, Genres) %>%
  summarise(avg_rating = mean(`Book-Rating`)) %>%
  ggplot(aes(x = Country, y = Genres)) +
  geom_tile(mapping = aes(fill = avg_rating)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Conclusion

The conclusion is that there are some correlations between book ratings and other variables, especially genres and country. It was surprising that religious books had some of the highest ratings, but overall, it's safe to say that nothing affects a book's rating as much as the quality of the book itself.