DATA 606 Data Project Proposal

Stephanie Chiang

2024-11-01

```
library(tidyverse)
library(dplyr)
library(ggplot2)
```

Data Preparation

```
raw_books <- read.csv(file = "goodreads_top100.csv")

# select relevant columns
books <- raw_books |>
  select("isbn",
         "title",
         "series_title",
         "series_release_number",
         "rating_score",
         "num_ratings",
         "current_readers",
         "want_to_read")
```

```
# convert blank strings to NAs in text columns
books <- books |>
  mutate(isbn = na_if(isbn, "")) |>
  mutate(series_title = na_if(series_title, "")) |>
  mutate(series_release_number = na_if(series_release_number, ""))

# remove duplicate isbn numbers
books <- books |>
  distinct(isbn, .keep_all = TRUE)

# add column to indicate if in series
books <- books |>
  mutate(serial = !is.na(books$series_title) & !is.na(books$series_release_number))

head(books)
```

```
##           isbn                         title                series_title
## 1 9780689830594                   Summer Story               Brambly Hedge
## 2 9780375704970           The Lake of Darkness                        <NA>
## 3 9780345446671 Beyond the Blue Event Horizon                Heechee Saga
```

```
## 4 9780446403016              St. Peter's Fair  Chronicles of Brother Cadfael
## 5 9780425198773                        Twice Shy                          <NA>
## 6 9780698119604          The Door in the Hedge                          <NA>
##   series_release_number rating_score num_ratings current_readers want_to_read
## 1                     2         4.45        1017               7          512
## 2                  <NA>         3.76        1388              77          623
## 3                     2         3.95       13307             181         3961
## 4                     4         4.12       10493            1298         2502
## 5                  <NA>         3.92        4188             162          642
## 6                  <NA>         3.70        9657             395         6643
##   serial
## 1   TRUE
## 2  FALSE
## 3   TRUE
## 4   TRUE
## 5  FALSE
## 6  FALSE
```

```
# create a 2nd table for any serials-only analysis
serial_books <- filter(books, serial == TRUE)

# add column for if it is first in a series
# may need further cleanup, depends if a prequel counts as first?
serial_books <- serial_books |>
  mutate(first_book = ifelse(grepl("^[01]", serial_books$series_release_number),
                             TRUE,
                             FALSE)) |>
  subset(select = -c(serial))

head(serial_books)
```

```
##            isbn                          title                  series_title
## 1 9780689830594                   Summer Story                 Brambly Hedge
## 2 9780345446671 Beyond the Blue Event Horizon                  Heechee Saga
## 3 9780446403016              St. Peter's Fair  Chronicles of Brother Cadfael
## 4 9780345468642              Pawn of Prophecy                  The Belgariad
## 5 9780553276329               Pacific Vortex!                     Dirk Pitt
## 6 9780786915743   Dragons of Autumn Twilight         Dragonlance: Chronicles
##   series_release_number rating_score num_ratings current_readers want_to_read
## 1                     2         4.45        1017               7          512
## 2                     2         3.95       13307             181         3961
## 3                     4         4.12       10493            1298         2502
## 4                     1         4.16      105412            1777        52200
## 5                     1         3.80       23332             350        11900
## 6                     1         4.01      116639            4499        52800
##   first_book
## 1      FALSE
## 2      FALSE
## 3      FALSE
## 4       TRUE
## 5       TRUE
## 6       TRUE
```

Research question

Are serials (books that are part of multi-volume series) more popular (either by readership, interest or rating) than standalone books? In other words, is there a relationship between being part of a series and popularity for books?

Cases

Each case is a book published between 1980-2023 and ranked in the top 100 for its year, based on Goodreads ratings. There are 4399 cases.

Data collection

This dataset posted on Kaggle was "collected through web scraping techniques" from Goodreads.com

Type of study

This is an observational study.

Data Source

Link to Top Goodreads Books Collection

Describe your variables?

The response variables are numerical: rating_score, num_ratings, current_readers, want_to_read. The explanatory variables are text fields, which will be converted into boolean / categorical: series_title (exists yes/no), series_release_number (first in series or not)

Relevant summary statistics

```
# summary statistics for rating_score, serials vs standalone
rating_summary <- books |>
  group_by(serial) |>
  reframe(
    count = n(),
    mean = mean(rating_score),
    sd = sd(rating_score),
    median = median(rating_score),
    min = min(rating_score),
    max = max(rating_score),
  )

rating_summary
```

```
## # A tibble: 2 x 7
##   serial count  mean    sd median   min   max
##   <lgl>  <int> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 FALSE   1813  3.96 0.256   3.97  2.97  4.81
## 2 TRUE    1805  4.06 0.216   4.06  3.29  4.76
```

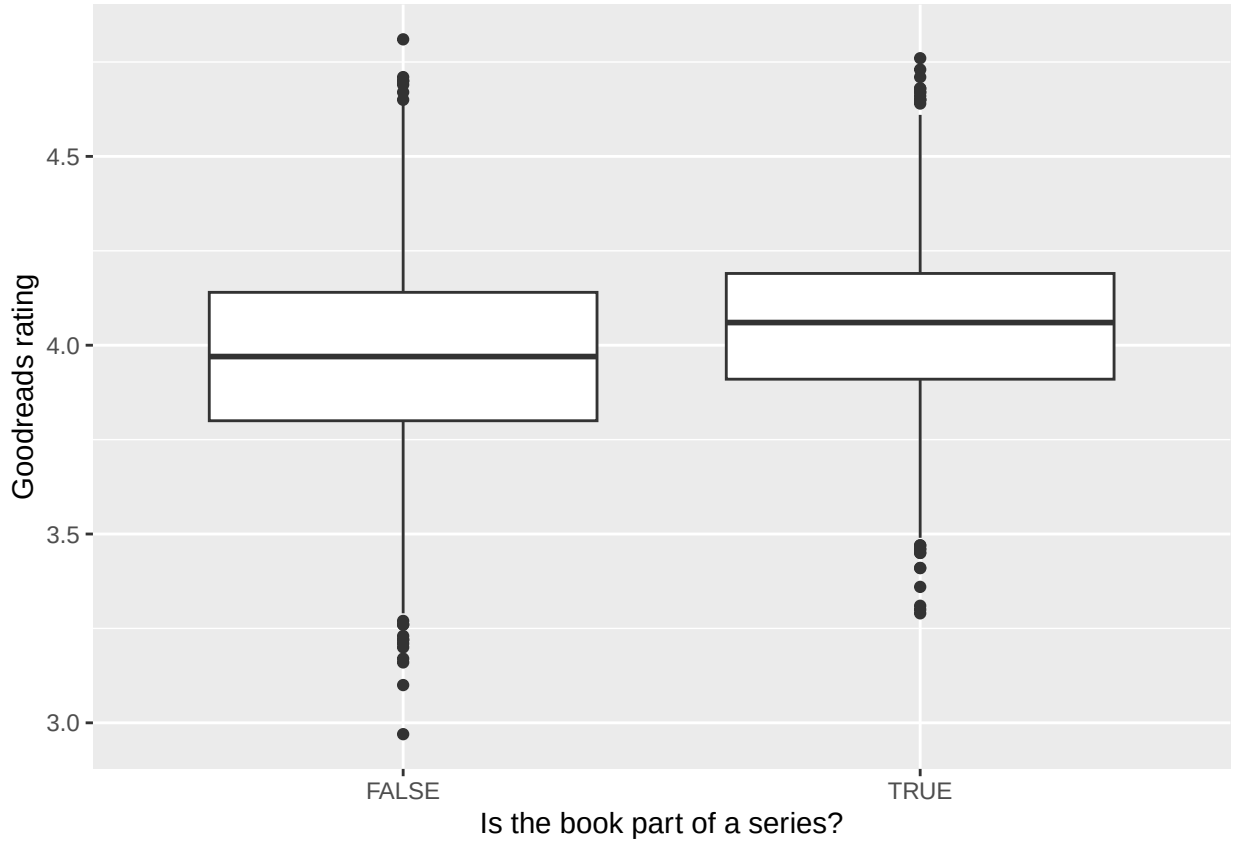
```r
# summary statistics for rating_score, first in a series vs sequel
rating_summary_sequels <- serial_books |>
  group_by(first_book) |>
  reframe(
    count = n(),
    mean = mean(rating_score),
    sd = sd(rating_score),
    median = median(rating_score),
    min = min(rating_score),
    max = max(rating_score),
  )

rating_summary_sequels
```

```
## # A tibble: 2 x 7
##   first_book count  mean    sd median   min   max
##   <lgl>      <int> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 FALSE       1006  4.10 0.206   4.1   3.41  4.76
## 2 TRUE         799  4.00 0.215   4.01  3.29  4.73
```

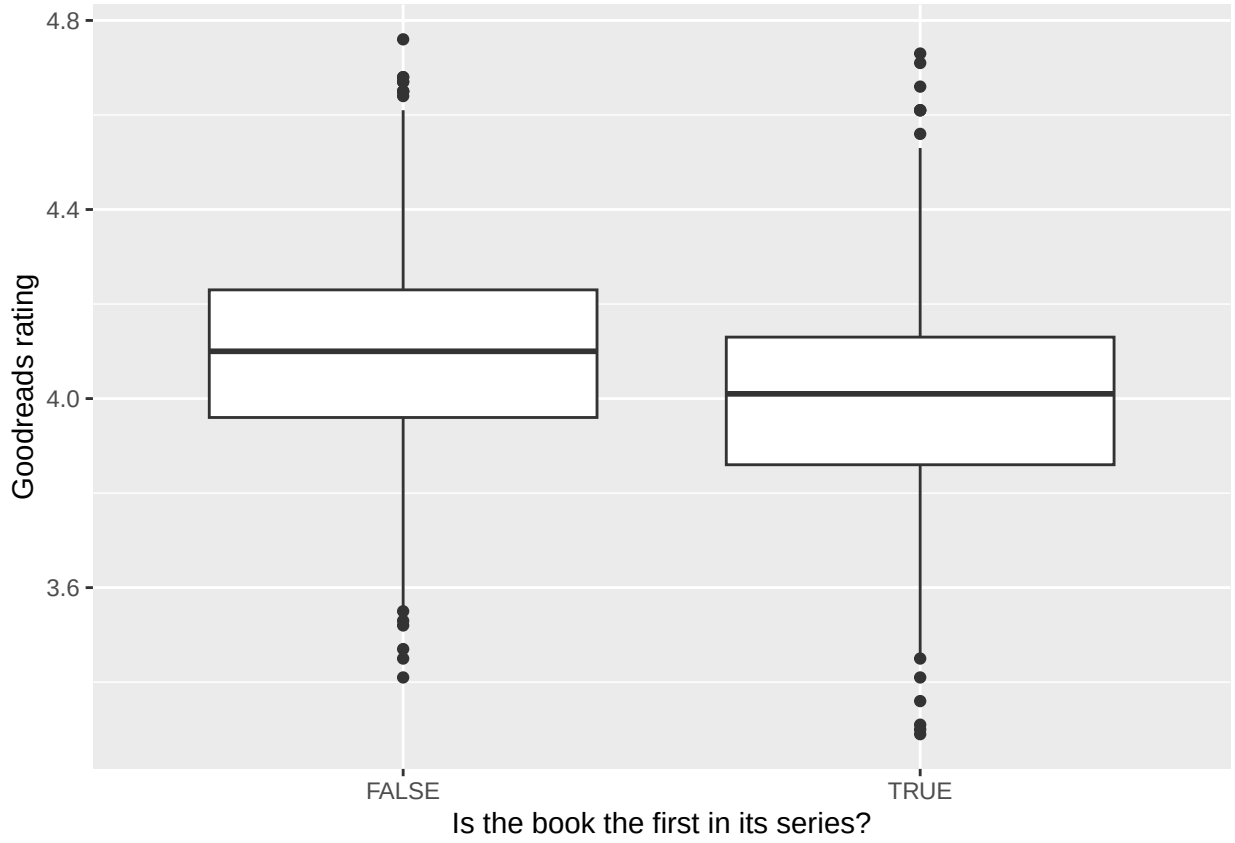
```r
# comparing mean of the numbers of current readers vs interested/potential readers
books_readership <- books |>
  pivot_longer(cols = c("current_readers", "want_to_read"),
               names_to = "reader_type",
               values_to = "readership")

readership_summary <- books_readership |>
  group_by(serial, reader_type) |>
  summarize(mean_readership = mean(readership, na.rm = TRUE))

readership_summary
```

```
## # A tibble: 4 x 3
## # Groups:   serial [2]
##   serial reader_type     mean_readership
##   <lgl>  <chr>                     <dbl>
## 1 FALSE  current_readers           5013.
## 2 FALSE  want_to_read             88904.
## 3 TRUE   current_readers           3874.
## 4 TRUE   want_to_read             62413.
```

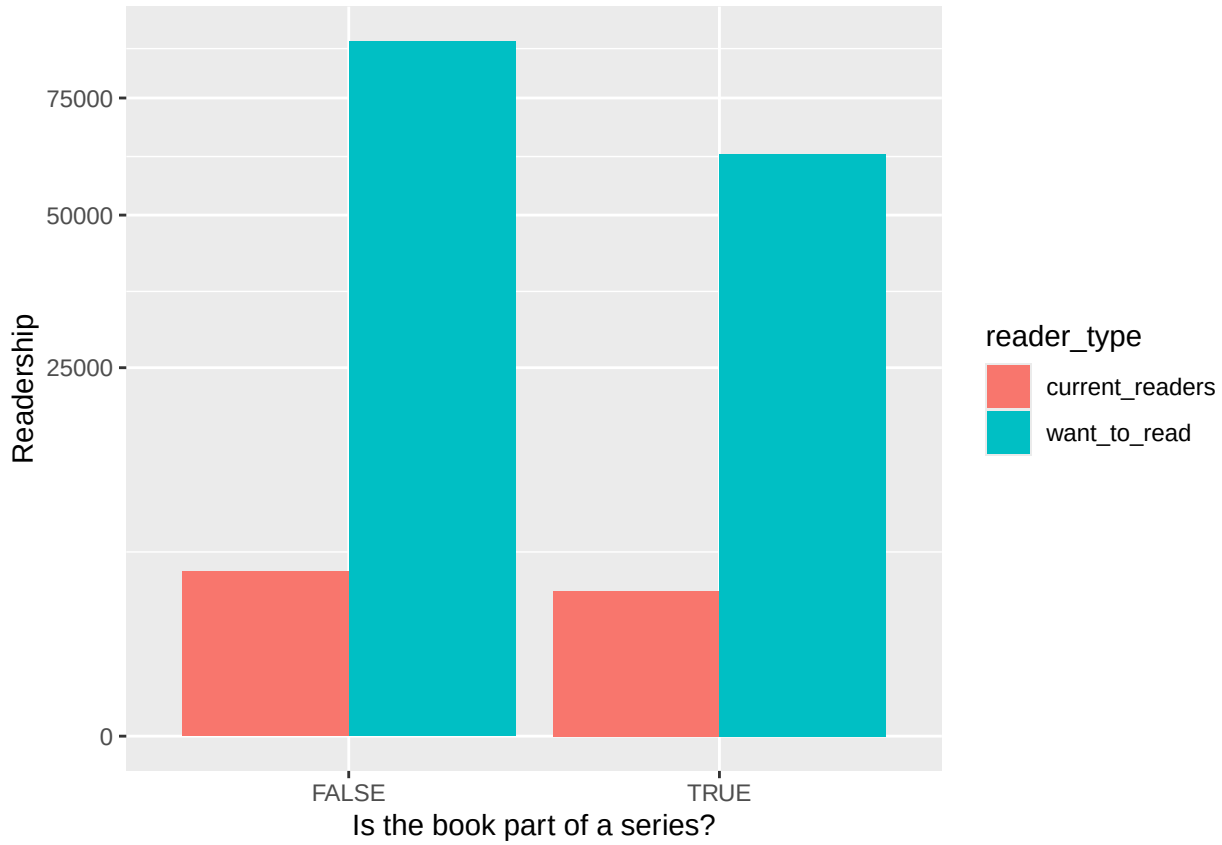
```r
ggplot(books, aes(x = serial, y = rating_score)) +
  geom_boxplot() +
  labs(x = "Is the book part of a series?", y = "Goodreads rating")
```




```
ggplot(serial_books, aes(x = first_book, y = rating_score)) +
  geom_boxplot() +
  labs(x = "Is the book the first in its series?", y = "Goodreads rating")
```




```
ggplot(readership_summary, aes(fill = reader_type, x = serial, y = mean_readership)) +
  geom_bar(position = "dodge", stat = "identity") +
  scale_y_sqrt() +
  labs(x = "Is the book part of a series?", y = "Readership")
```


75000
50000
25000
0
Readership
FALSE
TRUE
Is the book part of a series?
reader_type
current_readers
want_to_read