THE POPULARITY OF BOOKS ON GOODREADS

DATA 606 Final Project

Stephanie Chiang

Fall 2024

Abstract

This is an analysis of some of the factors that may contribute to the popularity of books on Goodreads. The observational dataset from Kaggle was originally sourced from Goodreads' Top 100 lists of the most popular books for each year from 1980 to 2023. In particular, the focus will be to answer the question:

Are serials, or books that are part of multi-volume series, associated with significantly different average user ratings and readership numbers than standalone books? In other words, is there a relationship between the independent variables (standalone books vs serials, and first in series vs sequel) and the dependent variables (mean user ratings, number of current readers and potential readership counts)?

Using summary statistics, visualizations and regression modeling, the results indicate that serials do not show statistically significant differences in mean user ratings; and sequels also do not show notable difference in average ratings than a first published installment. The differences in numbers of current and potential Goodreads users for standalone books vs series is also statistically insignificant. This analysis could provide potentially valuable insights for publishers, authors, and marketers for data-driven decision-making in their industry.

Data Preparation

The explanatory variables are created from text fields in the raw data. Each of the 4399 rows is given a value in a new categorical column with two levels (TRUE or FALSE) called `serial` based on whether there are non-empty strings under `series_title` and `series_release_number`.

Then, in a second data frame for only those books with `serial` set to TRUE, each observation is marked in a new `first_book` column with a TRUE if it is the first book published in its series or FALSE for sequels and prequels. (The determination here is that since prequels are published *after* the initial volume, they should not be considered the first in a series.)

The response variables are numerical: `rating_score` (out of 5), `num_ratings`, `current_readers`, `want_to_read`.

```r
library(tidyverse)

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 / repeated books
books <- books |>
  distinct(isbn, .keep_all = TRUE)

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

knitr::kable(head(books[, 2:5]))
```

title	series_title	series_release_number	rating_score
Summer Story	Brambly Hedge	2	4.45
The Lake of Darkness	NA	NA	3.76
Beyond the Blue Event Horizon	Heechee Saga	2	3.95
St. Peter's Fair	Chronicles of Brother Cadfael	4	4.12
Twice Shy	NA	NA	3.92
The Door in the Hedge	NA	NA	3.70

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

# add column for if it is the first release of its series
series <- series |>
  mutate(first_book = ifelse(grepl("^1(?!\\d)", series$series_release_number, perl = TRUE),
                             TRUE,
                             FALSE)) |>
  subset(select = -c(serial))

knitr::kable(head(series[, 2:5]))
```

title	series_title	series_release_number	rating_score
Summer Story	Brambly Hedge	2	4.45
Beyond the Blue Event Horizon	Heechee Saga	2	3.95
St. Peter's Fair	Chronicles of Brother Cadfael	4	4.12
Pawn of Prophecy	The Belgariad	1	4.16
Pacific Vortex!	Dirk Pitt	1	3.80

title	series_title	series_release_number	rating_score
Dragons of Autumn Twilight	Dragonlance: Chronicles	1	4.01

Summary Statistics & Data Visualizations

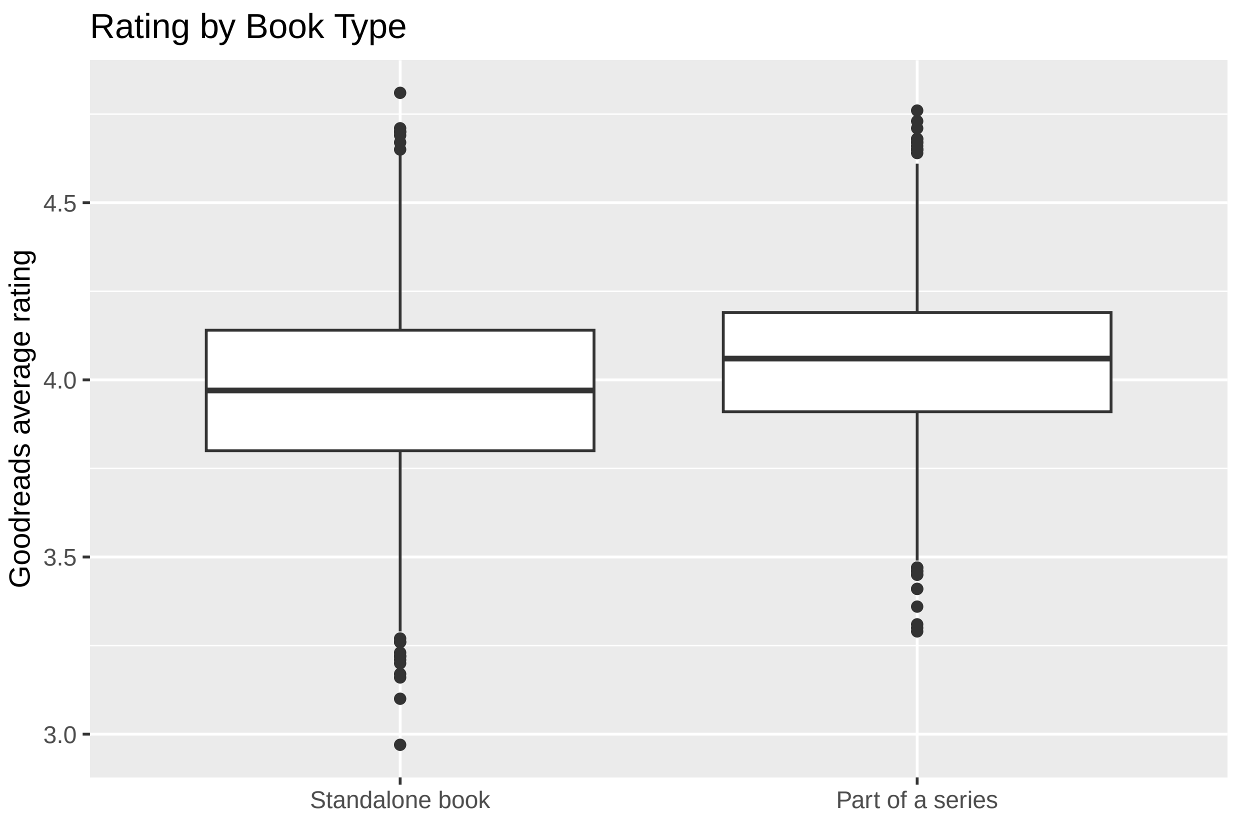
Mean Ratings: Series vs Standalone Books

When comparing the average user ratings, there is an observed preference by Goodreads users for books in serials standalone books.

```r
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
ggplot(books, aes(x = serial, y = rating_score)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("FALSE" = "Standalone book", "TRUE" = "Part of a series")) +
  labs(y = "Goodreads average rating", x= "", title = "Rating by Book Type")
```



Since each book is an independent observation and the sample sizes for each group are comfortably large, the conditions for inference are satisfied; a hypothesis test for the difference of the two means can determine if the association is noteworthy.

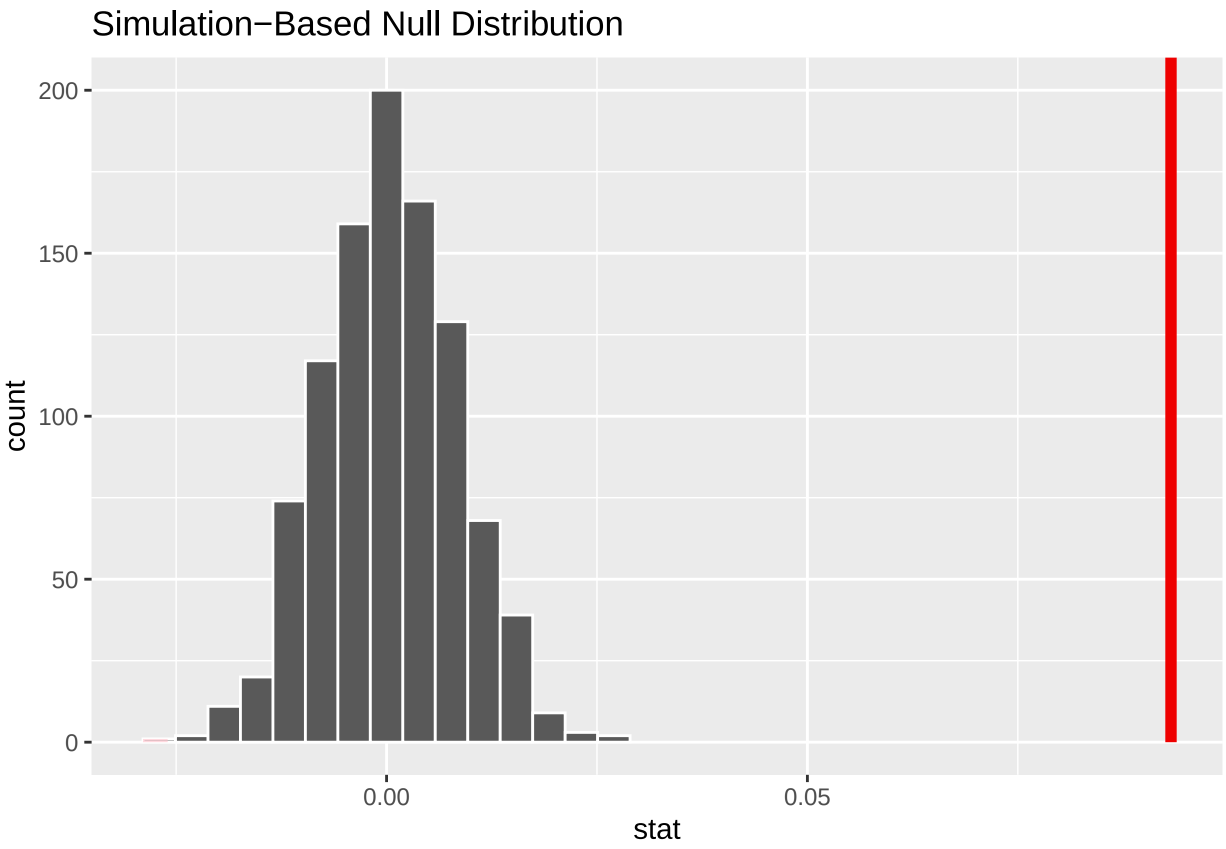
- The null hypothesis H0: There is no relationship between being part of a series and average rating.
- The alternative hypothesis H1: The average ratings are significantly different for serial books.

Below, the difference in means is calculated in the order of TRUE - FALSE != 0. Then, a test is then simulated; the null distribution is plotted for demonstration.

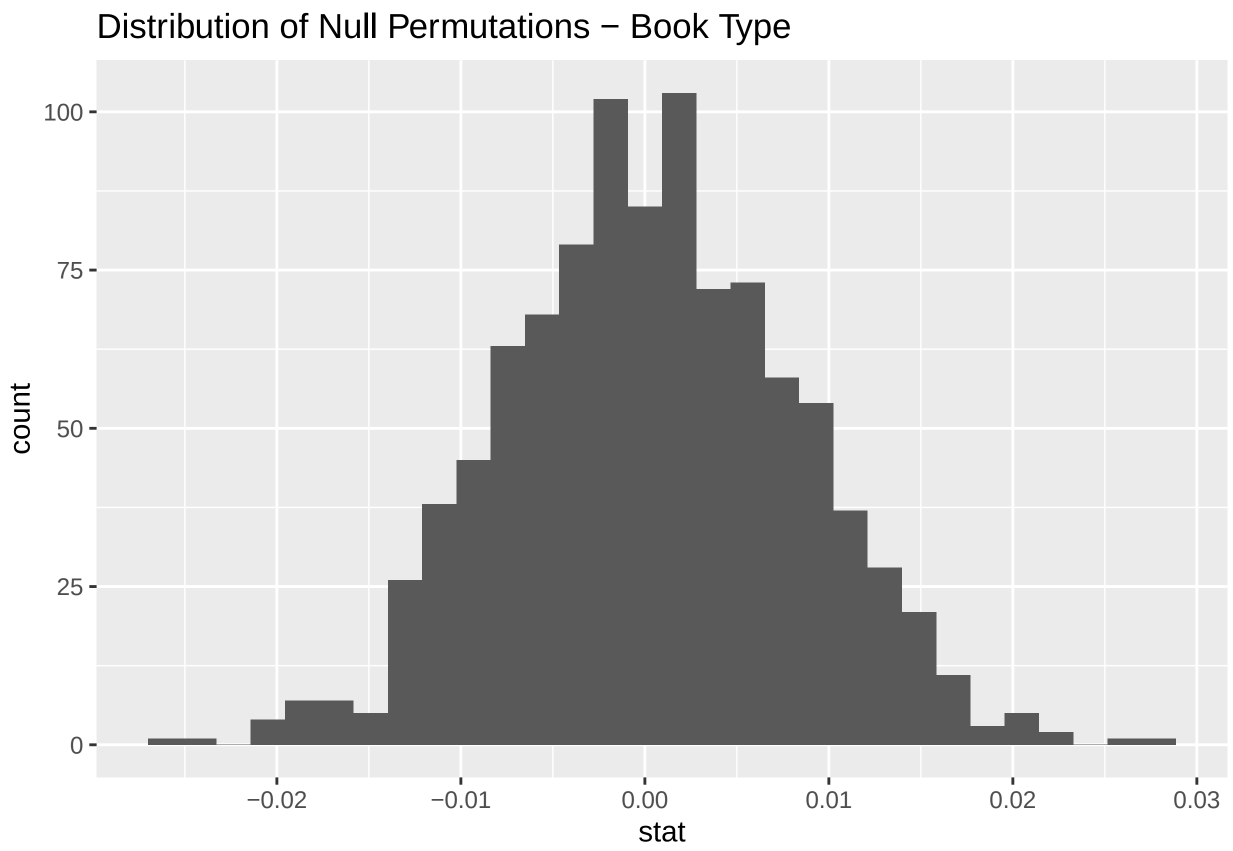
```
library(infer)
set.seed(99)

series_obs_diff <- books |>
  specify(rating_score ~ serial) |>
  calculate(stat = "diff in means", order = c(TRUE, FALSE))

series_obs_diff
```

```
## Response: rating_score (numeric)
## Explanatory: serial (factor)
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 0.0932
```

```r
series_null_dist <- books |>
  specify(rating_score ~ serial) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in means", order = c(TRUE, FALSE))

series_null_dist |>
  visualize() +
  shade_p_value(obs_stat = series_obs_diff, direction = "two-sided")
```



```r
ggplot(data = series_null_dist, aes(x = stat)) +
  geom_histogram() +
  labs(title = "Distribution of Null Permutations - Book Type")
```



The plots show that the null permutations fall entirely below the observed difference between serial and standalone ratings.

```
series_null_dist |>
  get_p_value(obs_stat = series_obs_diff, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
series_diff_ci <- series_null_dist |> get_ci(level = 0.95)

series_diff_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1  -0.0139   0.0154
```

At a confidence level of 95%, the difference in mean ratings between series and standalone books should fall between -0.014 to 0.015. Since this contains 0, there is a failure to reject the null hypothesis; there is no significant difference in average rating for serials.

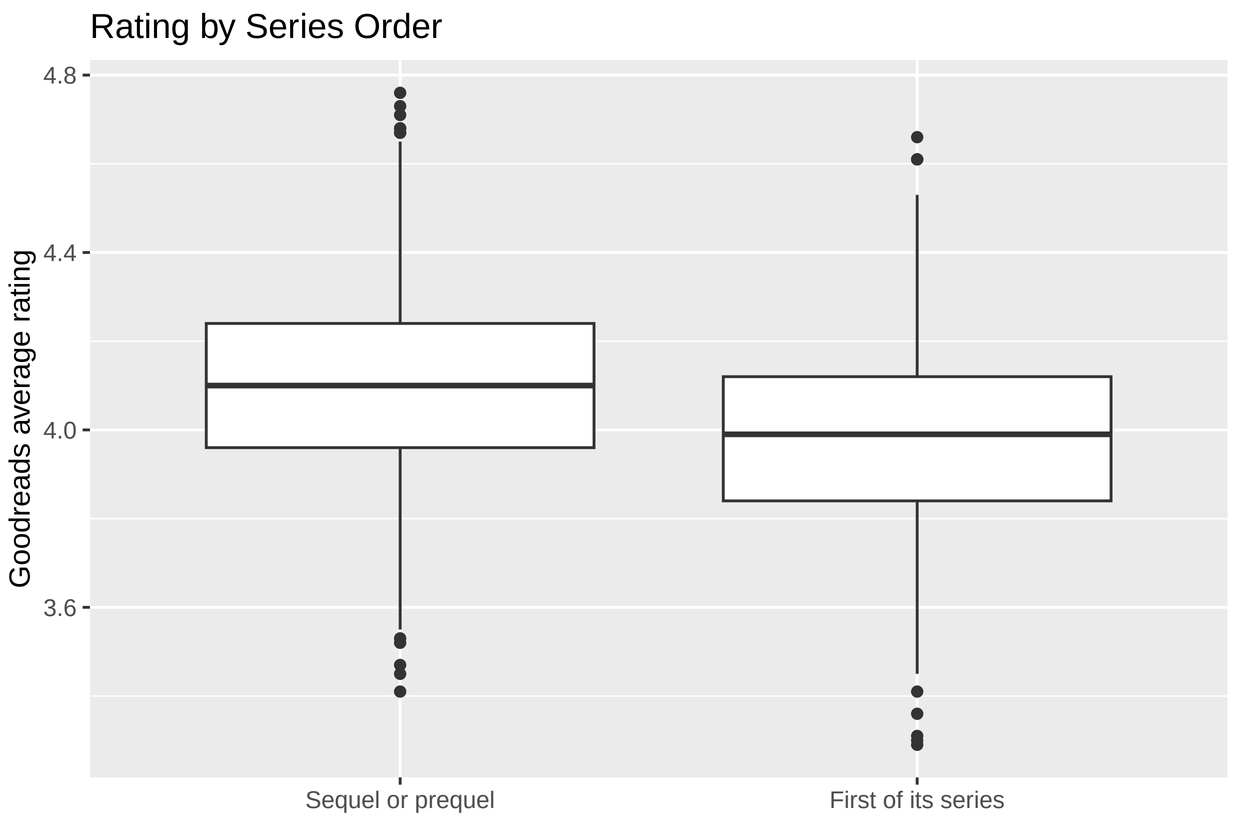
Average Ratings: Firsts vs Sequels

Within book series, the observed difference in ratings for sequels vs a first book in a series is even larger.

```
rating_summary_sequels <- series |>
  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       1141  4.10 0.206   4.1   3.41  4.76
## 2 TRUE         664  3.98 0.209   3.99  3.29  4.66
```

```
ggplot(series, aes(x = first_book, y = rating_score)) +
  geom_boxplot() +
  scale_x_discrete(labels = c("FALSE" = "Sequel or prequel",
                              "TRUE" = "First of its series")) +
  labs(y = "Goodreads average rating", x = "", title = "Rating by Series Order")
```



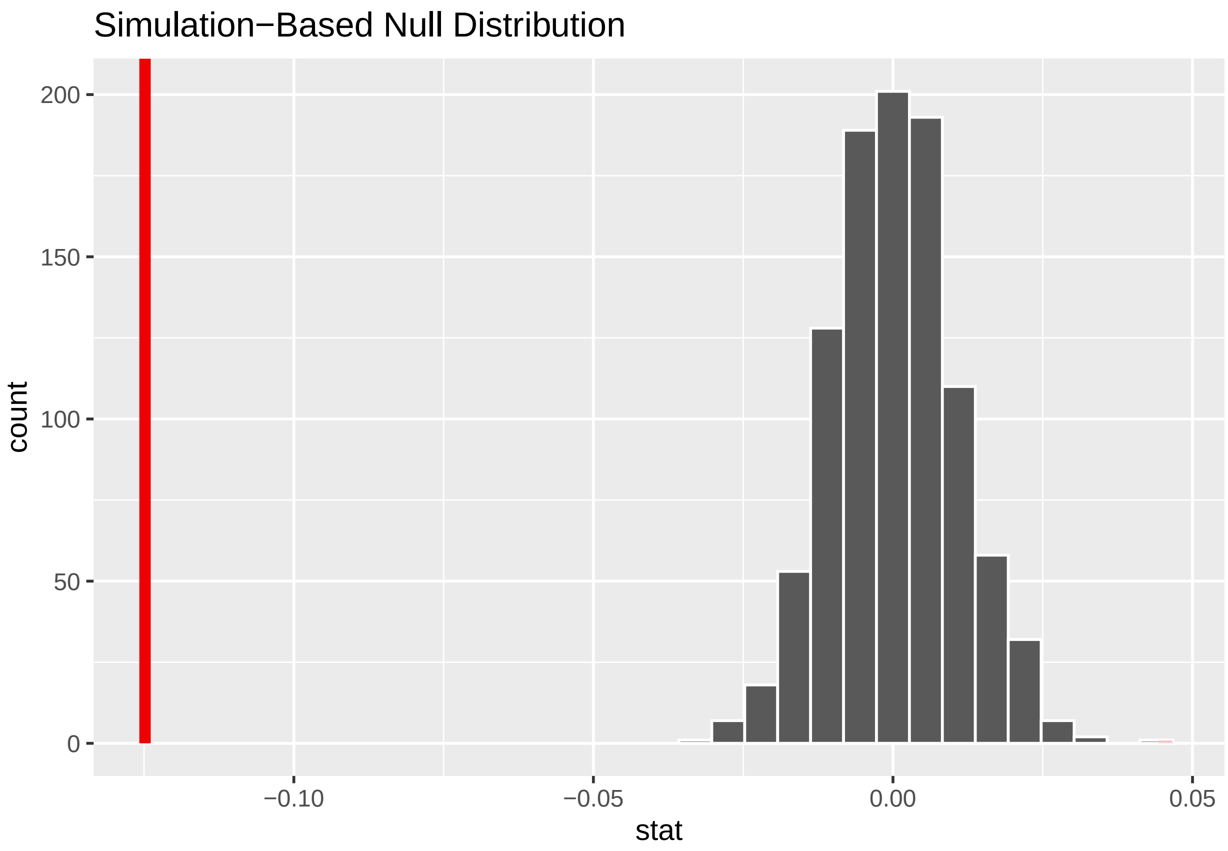
The hypothesis test for serials is as follows:

- The null hypothesis H0: There is no relationship between being the first in a series and average rating.
- The alternative hypothesis H1: The average ratings are significantly different for sequels than first books.

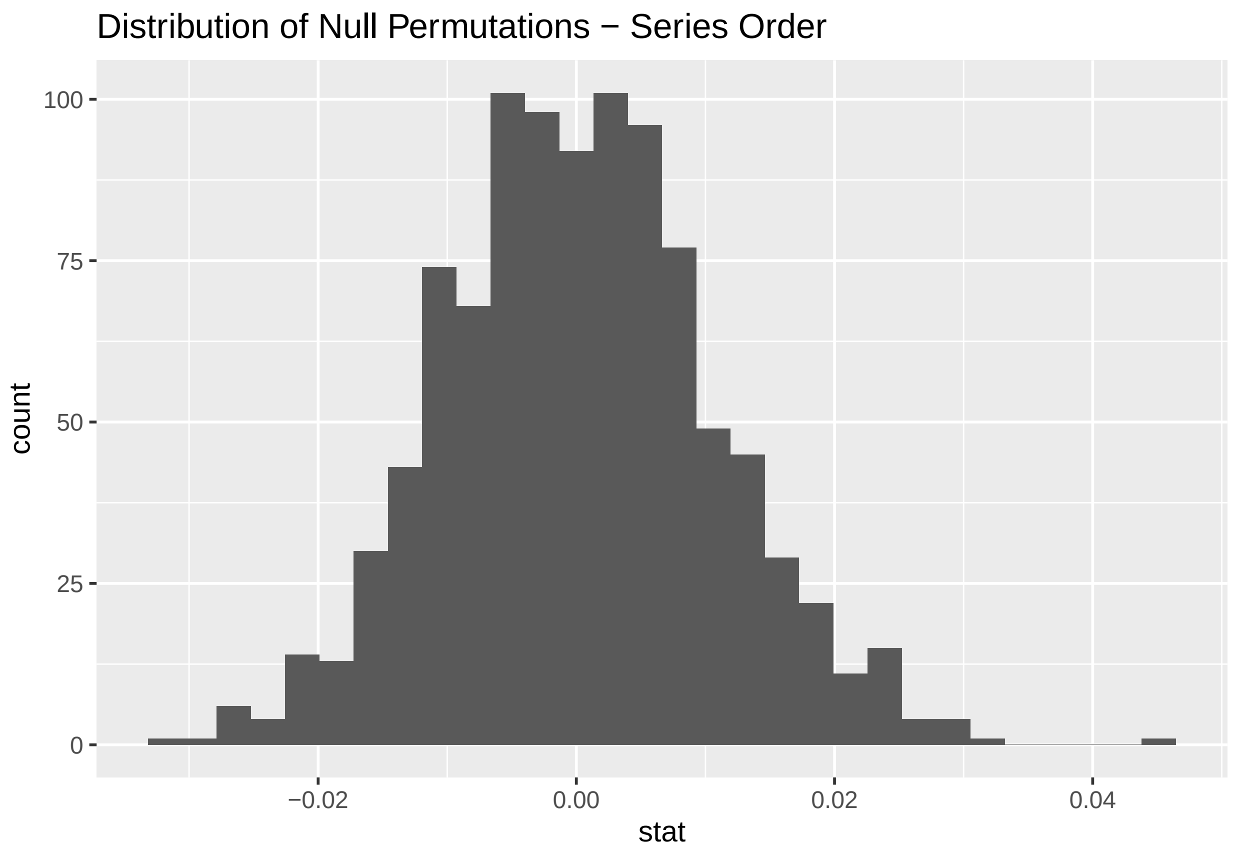
```
set.seed(99)

first_obs_diff <- series |>
  specify(rating_score ~ first_book) |>
  calculate(stat = "diff in means", order = c(TRUE, FALSE))

first_obs_diff
```

```
## Response: rating_score (numeric)
## Explanatory: first_book (factor)
## # A tibble: 1 x 1
##     stat
##    <dbl>
## 1 -0.125
```

```
first_null_dist <- series |>
  specify(rating_score ~ first_book) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in means", order = c(TRUE, FALSE))

first_null_dist |>
  visualize() +
  shade_p_value(obs_stat = first_obs_diff, direction = "two-sided")
```



```
ggplot(data = first_null_dist, aes(x = stat)) +
  geom_histogram() +
  labs(title = "Distribution of Null Permutations - Series Order")
```



Once again, the observed difference falls outside the range of the null permutations between serial and standalone ratings.

```
first_null_dist |>
  get_p_value(obs_stat = first_obs_diff, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

```
first_diff_ci <- first_null_dist |> get_ci(level = 0.95)

first_diff_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1  -0.0200   0.0220
```

With the confidence level set to 95%, the difference in mean ratings between first books and sequels/prequels should fall between -0.02 to 0.02. Since this contains 0, once again the null hypothesis cannot be rejected. No significant difference in average rating is proven here for sequels over first books in series.

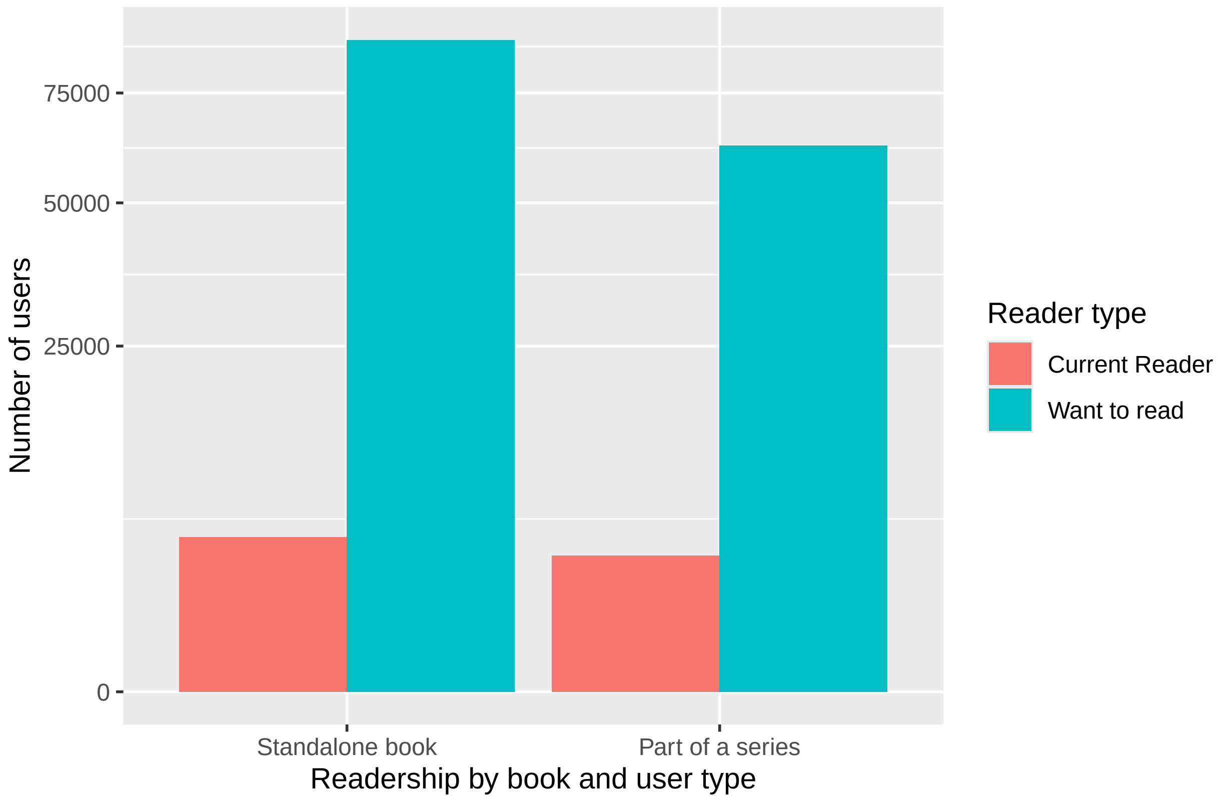
Readership

A different angle of examination than rating is readership: users who marked themselves as current readers of a title or interested/potential readers. The average number of users who are either currently reading or `want_to_read` a standalone book is much higher than for series. The idea is that it may be daunting for users to commit to reading an entire series.

```
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.
```

```
ggplot(readership_summary, aes(fill = reader_type, x = serial, y = mean_readership)) +
  geom_bar(position = "dodge", stat = "identity") +
  scale_y_sqrt() +
  labs(y = "Number of users", x = "Readership by book and user type", title = "") +
  scale_x_discrete(labels = c("FALSE" = "Standalone book", "TRUE" = "Part of a series")) +
  scale_fill_discrete(name = "Reader type", labels = c("Current Reader", "Want to read"))
```



To isolate if the series length is a factor (is starting a 15-book series a commitment that users want to avoid?), the data can be transformed to display the readership numbers by length of series.

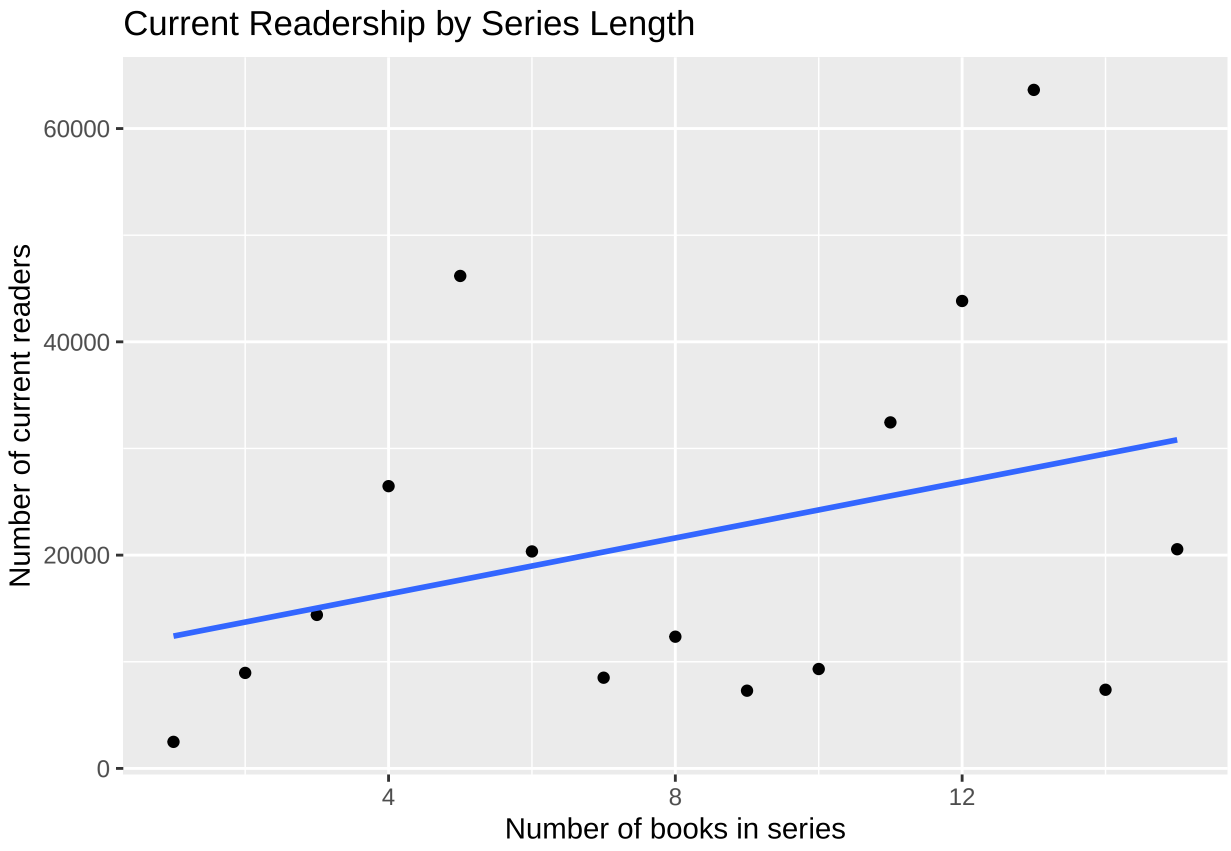
```r
# group by unique series titles, calculate the totals for series length and readership
series_length <- series |>
  group_by(series_title) |>
  summarize(
    series_len = n(),
    total_current = sum(current_readers),
    total_want_read = sum(want_to_read)) |>
  replace_na(list(total_current = 0, total_potential = 0)) |>
  select(-series_title)

# group by series length and calculate mean readership
series_length <- series_length |>
  group_by(series_len) |>
  summarize(
    mean_current = mean(total_current, na.rm = TRUE),
    mean_want_read = mean(total_want_read, na.rm = TRUE))

knitr::kable(series_length)
```

series_len	mean_current	mean_want_read
1	2489.955	52678.61
2	8955.210	154861.98

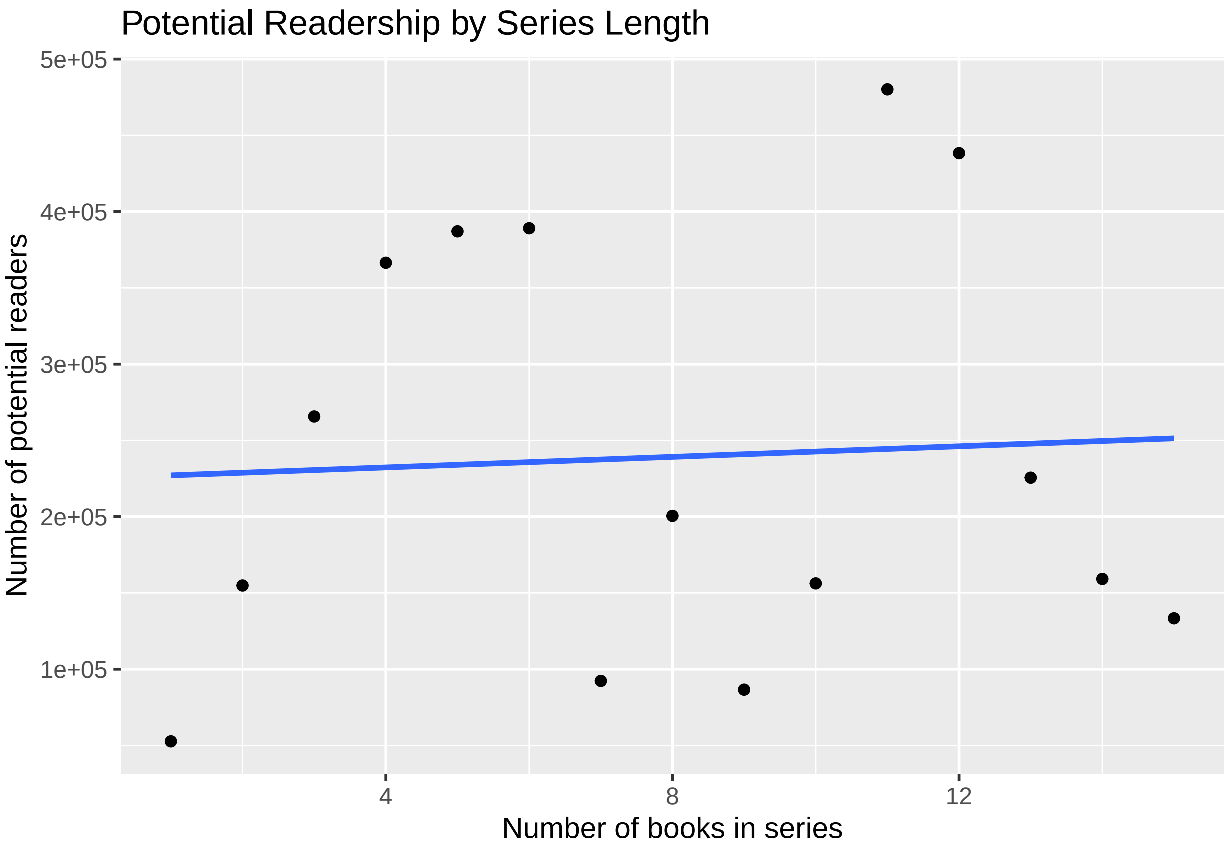
series_len	mean_current	mean_want_read
3	14395.652	265676.40
4	26467.415	366496.10
5	46178.500	387065.70
6	20345.400	389086.50
7	8506.143	92317.71
8	12351.800	200529.20
9	7289.200	86586.40
10	9321.250	156229.67
11	32446.000	480150.00
12	43831.000	438316.50
13	63623.000	225572.00
14	7376.000	159166.00
15	20554.000	133343.00

Although the user counts are higher for first books by raw totals, there does not appear to be a linear relationship between series length and mean readership.

```r
ggplot(series_length, aes(x = series_len, y = mean_current)) +
  geom_point() +
  labs(y = "Number of current readers",
       x = "Number of books in series",
       title = "Current Readership by Series Length") +
  stat_smooth(method = "lm", se = FALSE)
```



```
ggplot(series_length, aes(x = series_len, y = mean_want_read)) +
  geom_point() +
  labs(y = "Number of potential readers",
       x = "Number of books in series",
       title = "Potential Readership by Series Length") +
  stat_smooth(method = "lm", se = FALSE)
```



Below, the correlation for series length and total `want_to_read` is quite weak.

```
series_length |>
  summarise(cor(series_len, mean_want_read, use = "complete.obs"))
```

```
## # A tibble: 1 x 1
##   ‘cor(series_len, mean_want_read, use = "complete.obs")‘
##                                                     <dbl>
## 1                                                  0.0556
```

Similarly, the linear model shows a low R-squared of -0.07.:

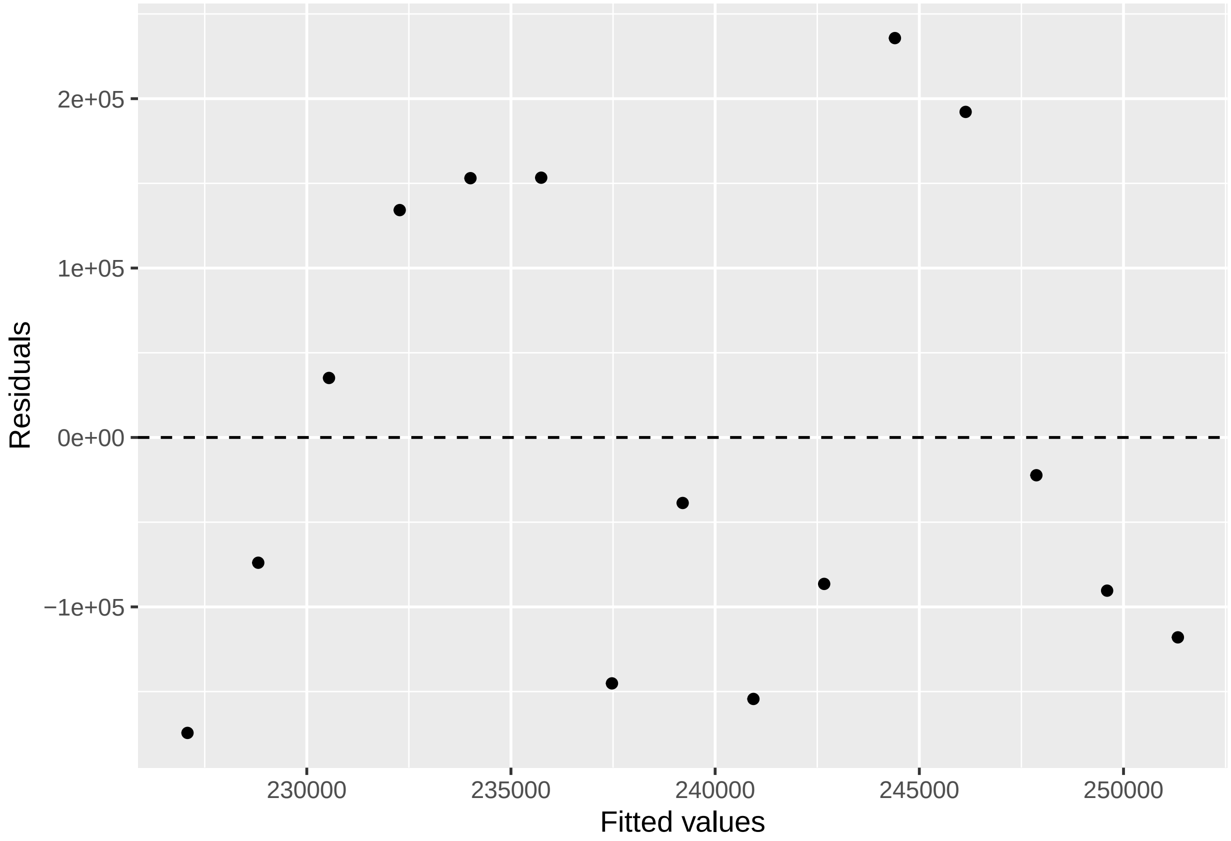
```
m_read <- lm(mean_want_read ~ series_len, data = series_length)
summary(m_read)
```

```
##
## Call:
```

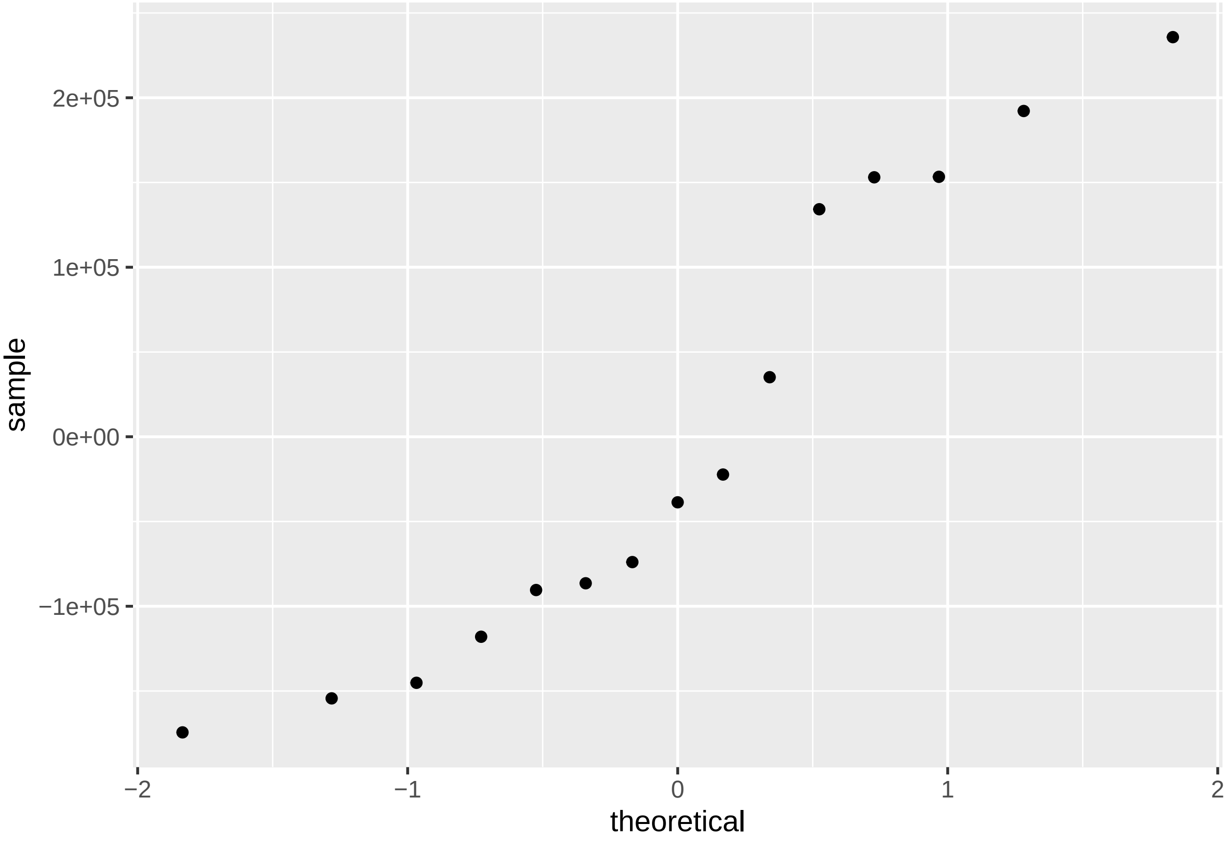
```
## lm(formula = mean_want_read ~ series_len, data = series_length)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -174400 -104210  -38676  143639  235748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   225347      78458   2.872   0.0131 *
## series_len      1732       8629   0.201   0.8440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 144400 on 13 degrees of freedom
## Multiple R-squared:  0.00309,    Adjusted R-squared:  -0.07359
## F-statistic: 0.0403 on 1 and 13 DF,  p-value: 0.844
```

To confirm the reliability of `m_read`, the scatter plot below visualizes the residuals vs predicted values. The data transformation has reduced the number of data points but they appear to be scattered fairly randomly around 0. The normal probability plot also appears distributed fairly normally.

```r
ggplot(data = m_read, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") + xlab("Fitted values") +
  ylab("Residuals")
```



```
ggplot(data = m_read, aes(sample = .resid)) + stat_qq()
```



Conclusion

Using data from Goodreads' Top 100 Books lists spanning over 40 years, this analysis reveals no statistically significant differences in mean user ratings between standalone books and book in series; neither do sequels exhibit notable differences in average ratings compared to first installments. Furthermore, the disparity in the number of current and potential Goodreads users for standalone books versus series is not significant.

These conclusions may be useful for authors and publishers, but the analysis has limitations. The dataset is based on users' subjective engagement, so ratings and `want_to_read` are fields up for individual interpretation (a reader may simply forget to update their lists). There are also potential biases not addressed here, based on authors, current trends, sales, reader demographics or genres. This information could be valuable if included. The statistical analysis was also heavily based on aggregates and means; there could be other avenues of exploration.