Shruti Alavala
10/21/21

Final Project

The Dataset

The dataset I chose to analyze was exported from my Goodreads account. Goodreads is an online platform that allows readers to log books they've read, want to read, browse books, and see what their friends are reading. I'm an avid reader and have logged 800+ books (read and want to read) so I thought it would be an interesting dataset to analyze and maybe unearth some trends in my reading habits!

Preparing the Data

First, I used Goodreads' export feature to get a csv of all of the books that I have logged in the site and stored it locally. Then I simply imported the data into an R data frame using the read.csv function.

```
> data.location <- "/Users/shrutialavala/OneDrive/Documents/BU MET/T4 Fall 2021 CS 544/Final Project/goodreads_library
_export_2021_Oct.csv"
>
> my.books.orig <- read.csv(data.location, header=TRUE)
> my.books <- my.books.orig
```

A glimpse of my dataset:

```
> head(my.books)
   Book.Id                                                                    Title        Author    Author.l.f Additional.Authors        Genre
1 52128695     Ace: What Asexuality Reveals About Desire, Society, and the Meaning of Sex  Angela  Chen  Chen, Angela                      Non-Fiction
2 55789067                                                 The Authenticity Project  Clare Pooley Pooley, Clare                     Contemporary
3 43582376                                          The Body: A Guide for Occupants   Bill Bryson  Bryson, Bill                      Non-Fiction
4 55600878                                                 Instructions for Dancing   Nicola Yoon  Yoon, Nicola                          Romance
5 55643287 How the Word Is Passed: A Reckoning with the History of Slavery Across America Clint   Smith  Smith, Clint                      Non-Fiction
6 54985743                                                 People We Meet on Vacation   Emily Henry  Henry, Emily                          Romance
  Ratings.Count Reviews.Count       ISBN       ISBN13 My.Rating Average.Rating                Publisher        Binding Number.of.Pages Year.Published
1          3982           889 080701379X 9.780807e+12         5           4.40             Beacon Press      Hardcover             224            2020
2         38674          5445 1984878638 9.781985e+12         0           3.96            Penguin Books      Paperback             384            2020
3         54782          6420 0385539304 9.780386e+12         0           4.30           Doubleday Books      Hardcover             450            2019
4         16152          3717                      NA         0           4.10                  Penguin Kindle Edition             304            2021
5          6233          1169 0316492930 9.780316e+12         0           4.80 Little, Brown and Company      Hardcover             336            2021
6        198418         25801 1984806750 9.781985e+12         5           4.13            Berkley Books      Paperback             364            2021
  Original.Publication.Year Date.Read Date.Added Bookshelves Bookshelves.with.positions Exclusive.Shelf My.Review Spoiler Private.Notes Read.Count
1                      2020   10/5/21    7/13/21                                                  read        NA      NA           NA          1
2                      2020             10/2/21     to-read            to-read (#480)        to-read        NA      NA           NA          0
3                      2019             10/2/21     to-read            to-read (#479)        to-read        NA      NA           NA          0
4                      2021             10/2/21     to-read            to-read (#478)        to-read        NA      NA           NA          0
5                      2021             10/2/21     to-read            to-read (#477)        to-read        NA      NA           NA          0
6                      2021   9/28/21    5/25/21                                                  read        NA      NA           NA          1
  Recommended.For Recommended.By Owned.Copies Original.Purchase.Date Original.Purchase.Location Condition Condition.Description BCID
1              NA              NA            0                     NA                        NA                                NA   NA
2              NA              NA            0                     NA                        NA                                NA   NA
3              NA              NA            0                     NA                        NA                                NA   NA
4              NA              NA            0                     NA                        NA                                NA   NA
5              NA              NA            0                     NA                        NA                                NA   NA
6              NA              NA            0                     NA                        NA                                NA   NA
```

There are a lot of columns that are either don't have any actual data in them or I was not interested in analyzing them so I removed them so that my data frame was easier to use.

```
> # delete blank/unnecessary columns
> my.books <- subset(my.books, select = -c(Author.l.f, Additional.Authors, ISBN, ISBN13,
+                                          Bookshelves, Original.Publication.Year,
+                                          Bookshelves.with.positions,
+                                          My.Review, Spoiler, Private.Notes,
+                                          Recommended.For, Recommended.By,
+                                          Owned.Copies, Original.Purchase.Date,
+                                          Original.Purchase.Location, Condition,
+                                          Condition.Description, BCID))
```

Then I renamed the columns.

```
> # naming columns
> colnames(my.books) <- c("BookID", "Title", "Author", "Genre", "RatingsCount",
+                         "ReviewsCount", "MyRating", "AverageRating", "Publisher",
+                         "Format", "NumberOfPages", "YearPublished",
+                         "DateRead", "DateAdded", "Status", "ReadCount")
>
> colnames(my.books)
 [1] "BookID"        "Title"         "Author"        "Genre"         "RatingsCount"  "ReviewsCount"  "MyRating"
 [8] "AverageRating" "Publisher"     "Format"        "NumberOfPages" "YearPublished" "DateRead"      "DateAdded"
[15] "Status"        "ReadCount"
```

Then I checked the data types of each column and changed them where necessary. For example, the columns with dates in them were imported as “chr” and “Date” is more appropriate.

```
> str(my.books)
'data.frame':   876 obs. of  16 variables:
 $ BookID       : int  52128695 55789067 43582376 55600878 55643287 54985743 43575115 26535513 58104944 56366617 ...
 $ Title        : chr  "Ace: What Asexuality Reveals About Desire, Society, and the Meaning of Sex" "The Authenticity
 Project" "The Body: A Guide for Occupants" "Instructions for Dancing" ...
 $ Author       : chr  "Angela  Chen" "Clare Pooley" "Bill Bryson" "Nicola Yoon" ...
 $ Genre        : chr  "Non-Fiction" "Contemporary" "Non-Fiction" "Romance" ...
 $ RatingsCount : int  3982 38674 54782 16152 6233 198418 133333 4037 83316 34675 ...
 $ ReviewsCount : int  889 5445 6420 3717 1169 25801 26271 465 13597 3709 ...
 $ MyRating     : int  5 0 0 0 0 5 4 0 0 0 ...
 $ AverageRating: num  4.4 3.96 4.3 4.1 4.8 4.13 3.84 4.4 3.66 3.89 ...
 $ Publisher    : chr  "Beacon Press" "Penguin Books" "Doubleday Books" "Penguin" ...
 $ Format       : chr  "Hardcover" "Paperback" "Hardcover" "Kindle Edition" ...
 $ NumberOfPages: int  224 384 450 304 336 364 498 288 333 349 ...
 $ YearPublished: int  2020 2020 2019 2021 2021 2021 2019 2015 2021 2021 ...
 $ DateRead     : chr  "10/5/21" "" "" "" ...
 $ DateAdded    : chr  "7/13/21" "10/2/21" "10/2/21" "10/2/21" ...
 $ Status       : chr  "read" "to-read" "to-read" "to-read" ...
 $ ReadCount    : int  1 0 0 0 0 1 1 0 0 0 ...
```

DateRead Column:

```
> # change date read column to Date data type
> head(my.books$DateRead)
[1] "10/5/21" ""        ""        ""        ""        "9/28/21"
> my.books$DateRead <- as.Date(my.books$DateRead, "%m/%d/%y")
> head(my.books$DateRead)
[1] "2021-10-05" NA           NA           NA           NA           "2021-09-28"
```

DateAdded Column:

```
> # change date added column to Date data type
> head(my.books$DateAdded)
[1] "7/13/21" "10/2/21" "10/2/21" "10/2/21" "10/2/21" "5/25/21"
> my.books$DateAdded <- as.Date(my.books$DateAdded, "%m/%d/%y")
> head(my.books$DateAdded)
[1] "2021-07-13" "2021-10-02" "2021-10-02" "2021-10-02" "2021-10-02" "2021-05-25"
```

There was one book in the data frame that did not have a value for the NumberOfPages column so I chose to remove it to make my later analysis simpler.

```
> # remove rows with NA number of pages
> v <- which(is.na(my.books$NumberOfPages))
> my.books <- my.books[-v, ]
```

I also added a new calculated column – the ratio of the number of reviews (written out) and ratings (value 1 to 5) for a particular book.

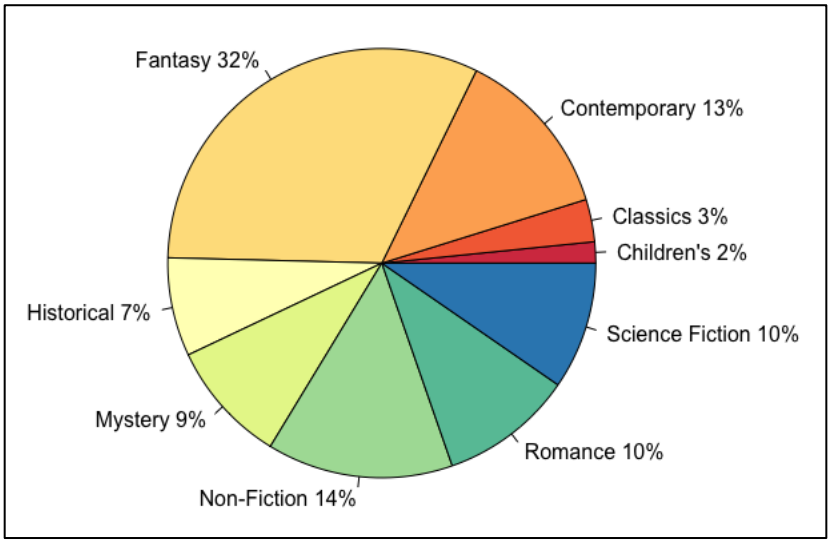
```
> my.books$ReviewRatingRatio <- round((my.books$ReviewsCount/my.books$RatingsCount), 2)
> head(my.books)
    BookID                                                                      Title        Author
1 52128695     Ace: What Asexuality Reveals About Desire, Society, and the Meaning of Sex  Angela  Chen
2 55789067                                                    The Authenticity Project  Clare Pooley
3 43582376                                                The Body: A Guide for Occupants   Bill Bryson
4 55600878                                                    Instructions for Dancing   Nicola Yoon
5 55643287 How the Word Is Passed: A Reckoning with the History of Slavery Across America Clint   Smith
6 54985743                                                  People We Meet on Vacation   Emily Henry
         Genre RatingsCount ReviewsCount MyRating AverageRating                Publisher         Format
1  Non-Fiction         3982          889        5          4.40             Beacon Press      Hardcover
2 Contemporary        38674         5445        0          3.96            Penguin Books      Paperback
3  Non-Fiction        54782         6420        0          4.30          Doubleday Books      Hardcover
4      Romance        16152         3717        0          4.10                  Penguin Kindle Edition
5  Non-Fiction         6233         1169        0          4.80 Little, Brown and Company      Hardcover
6      Romance       198418        25801        5          4.13            Berkley Books      Paperback
  NumberOfPages YearPublished   DateRead  DateAdded  Status ReadCount ReviewRatingRatio
1           224          2020 2021-10-05 2021-07-13    read         1              0.22
2           384          2020       <NA> 2021-10-02 to-read         0              0.14
3           450          2019       <NA> 2021-10-02 to-read         0              0.12
4           304          2021       <NA> 2021-10-02 to-read         0              0.23
5           336          2021       <NA> 2021-10-02 to-read         0              0.19
6           364          2021 2021-09-28 2021-05-25    read         1              0.13
```

Analyzing the Data

<u>One Categorical Variable – Genre</u>

The first categorical variable I looked into was the Genre column. I wanted to see which genres I gravitated towards so I made a pie chart to see what the biggest slices were.

```
> par(mfrow = c(1,1))
>
> genre.data <- table(my.books$Genre)
> slice.labels <- names(genre.data)
> slice.percents <- round(genre.data/sum(genre.data)*100)
> slice.labels <- paste(slice.labels, slice.percents)
> slice.labels <- paste0(slice.labels, "%")

> pie(genre.data,
+     labels = slice.labels,
+     col = brewer.pal(9, "Spectral"))
```

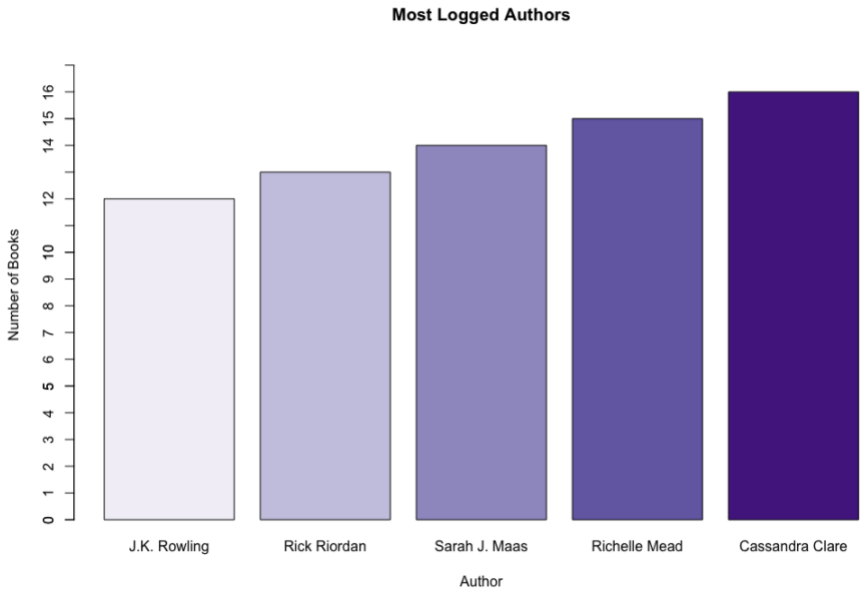


The biggest slice by far is the Fantasy genre at 32%, which wasn't surprising to me since I enjoy the genre. Additionally, series are more prevalent in Fantasy and if I start a series I tend to want to finish it, which results in logging more Fantasy books. I was surprised to see that my next highest Genre is Non-Fiction but that is partially due to the fact that Fiction was broken down into sub-categories (Fantasy, Mystery, Contemporary, etc.) but Non-Fiction was not. For a more accurate comparison, Non-Fiction should also be broken down into its sub-categories (Science, Memoirs, Politics, etc.).

One Categorical Variable – Top 5 Authors

I also analyzed another categorical variable – my top 5 logged authors.

```
> author.freq.df <- as.data.frame(table(my.books$Author))
> colnames(author.freq.df) <- c("Author", "Frequency")
> authors.ordered <- author.freq.df[order(-author.freq.df$Frequency), ]
> top.5.freq <- authors.ordered[seq(1,5),]

> top.5.authors <- subset(my.books, Author %in% top.5.freq$Author)
> t <- table(top.5.authors$Author)
> t

Cassandra Clare    J.K. Rowling   Richelle Mead    Rick Riordan   Sarah J. Maas
             16              12              15              13              14

> par(mfrow = c(1,1))
>
> barplot(sort(t),
+         main = "Most Logged Authors",
+         xlab = "Author",
+         ylab = "Number of Books",
+         ylim = c(0, max(t)*1.1),
+         col = brewer.pal(5, "Purples"))
> axis(side = 2, at = seq(0,max(t)*1.1))
```

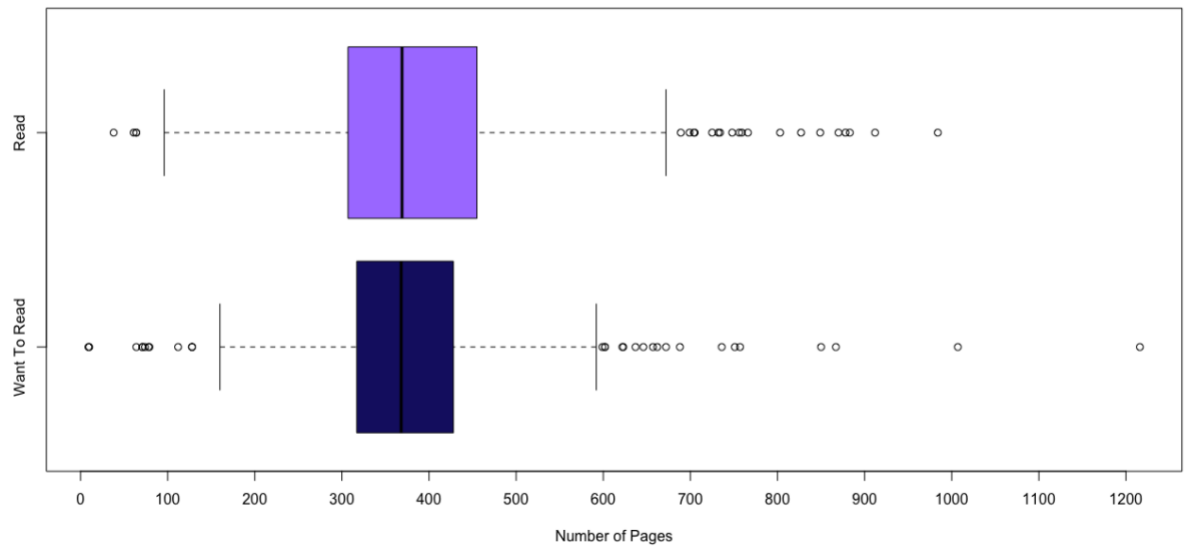


All five of these authors primarily write Fantasy books series so it makes sense that they are my most logged authors. I displayed the bars of the barplot in ascending order with a color gradient so it is easier to see the differences. It's a reasonable assumption that my most logged authors are my favorite. But, I think a better way to calculate and display that measure would be show the percentage of books I have logged from an author's entire collection of books. Then I could compare authors who have written more books with authors who have written less.

One Numerical Variable – Number of Pages

The numerical variable I analyzed was Number of Pages. First, I split up my dataset into two separate datasets so that I could compare the spread between Number of Pages that I've read and Number of Pages that I want to read.

```
> my.books.unread <- subset(my.books, Status == "to-read")
> my.books.read <- subset(my.books, Status == "read")

> boxplot(my.books.unread$NumberOfPages,
+         my.books.read$NumberOfPages,
+         names = c("Want To Read", "Read"),
+         col = c("midnightblue", "mediumpurple1"),
+         xlab = "Number of Pages",
+         horizontal = TRUE,
+         xaxt = "n")
> axis(side = 1, seq(0,1200,100))
```



Using fivenum function:

```
> fivenum(my.books.read$NumberOfPages)
[1]  38 307 369 455 984
> fivenum(my.books.unread$NumberOfPages)
[1]    9  317  368  428 1216
```

	Minimum	Q1	Q2	Q3	Maximum
Read	38	307	369	455	984
Want To Read	9	317	368	428	1216

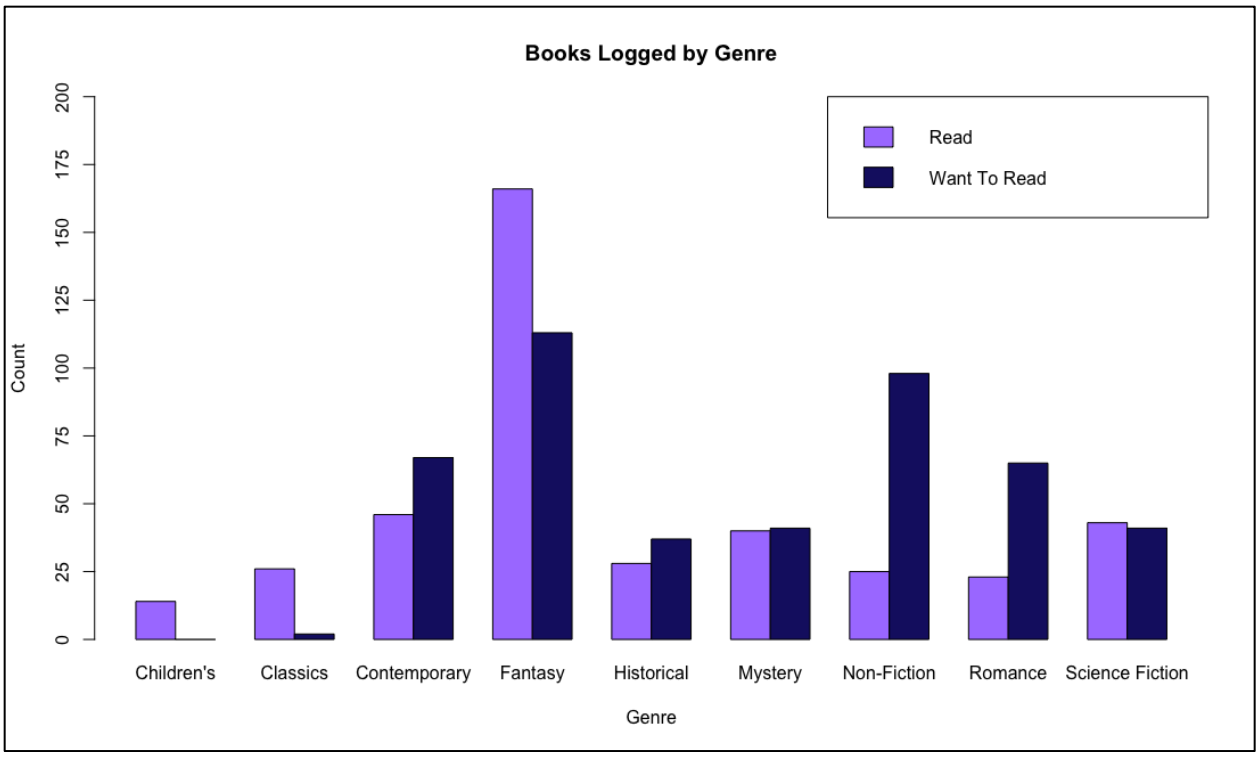
The boxplot shows that the middle 50% (Q3 – Q1) of the Read books is larger than the middle 50% of the Want to Read books, meaning there is more variation in the number of pages. However, the overall range (maximum - minimum) of Number of Pages is higher in the Want to Read subset. The medians, shown by the vertical line within the box, are only one page apart.

<u>One set of two or more variables - Categorical</u>

After creating the pie chart for genre, I wanted to see the breakdown between Genres and Status. Fantasy was my highest logged genre, but was it my most read?

```
> genre.status.freq <- table(my.books$Status, my.books$Genre)
> genre.status.freq

         Children's Classics Contemporary Fantasy Historical Mystery Non-Fiction Romance Science Fiction
  read           14       26          46     166         28      40          25      23              43
  to-read         0        2          67     113         37      41          98      65              41
```

```
> par(mfrow = c(1,1))
> barplot(genre.status.freq,
+         beside = TRUE,
+         legend.text = c("Read", "Want To Read"),
+         args.legend = c(x="topright"),
+         main = "Books Logged by Genre",
+         xlab = "Genre",
+         ylab = "Count",
+         ylim = c(0, 200),
+         yaxt = "n",
+         col = c("mediumpurple1", "midnightblue"))
> axis(side = 2, seq(0,200,25))
```

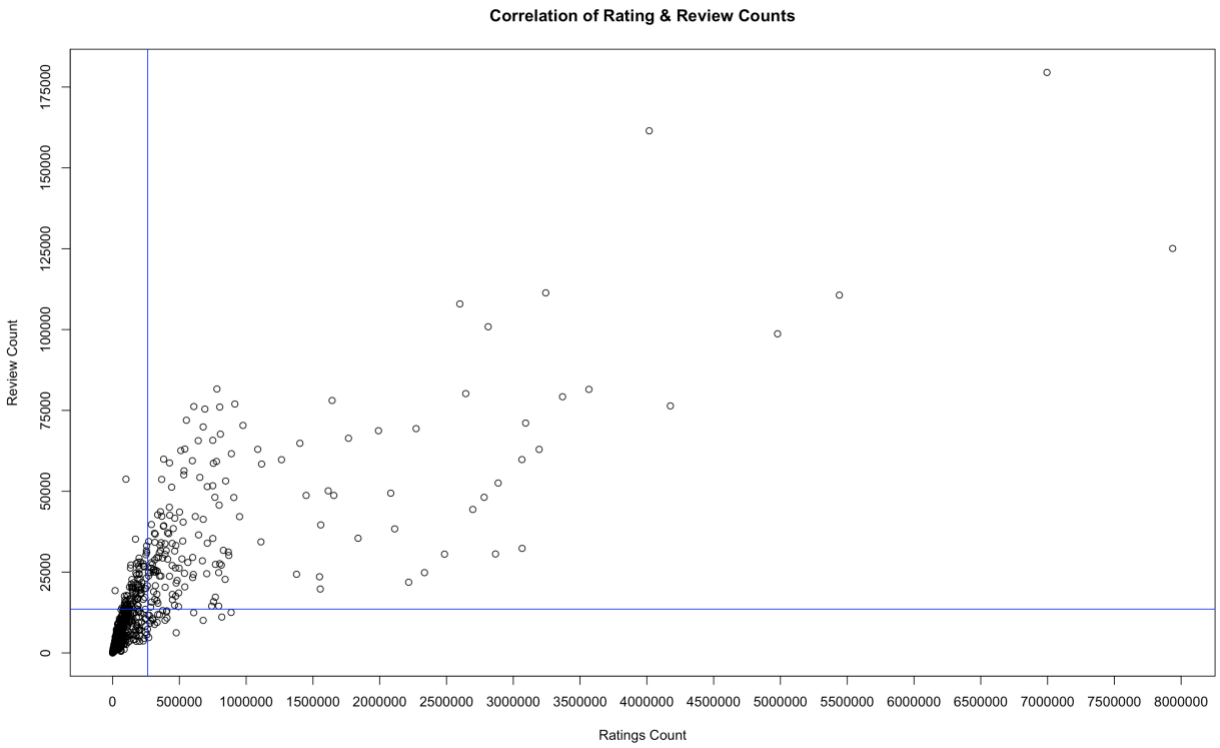


It turns out that Fantasy is my most read genre as well as most logged! It's also interesting to note that in the Non-Fiction genre I have many more books that I want to read than have actually read. It also makes sense that I have read more books in the Children's and Classics genres than I want to read. I read books in those genres when I was very young or as assigned reading in school and no longer want to read those kinds of books as an adult.

<u>One set of two or more variables - Numerical</u>

I also decided to analyze two numerical variables – the rating and review counts. In Goodreads, a rating is value from 1 to 5 that a reader can quickly give to show how much they enjoyed the book. On the other hand, a review is when a reader takes the time write out their thoughts on the book. I wanted to see the correlation between the two variables.

```
> par(mfrow = c(1,1))
> plot(my.books$RatingsCount,
+      my.books$ReviewsCount,
+      main = "Correlation of Rating & Review Counts",
+      xlab = "Ratings Count",
+      ylab = "Review Count",
+      xaxt = "n",
+      yaxt = "n")
> axis(side = 1, seq(0,8000000,500000), labels = TRUE)
> axis(side = 2, seq(0,200000,25000), labels = TRUE)
```

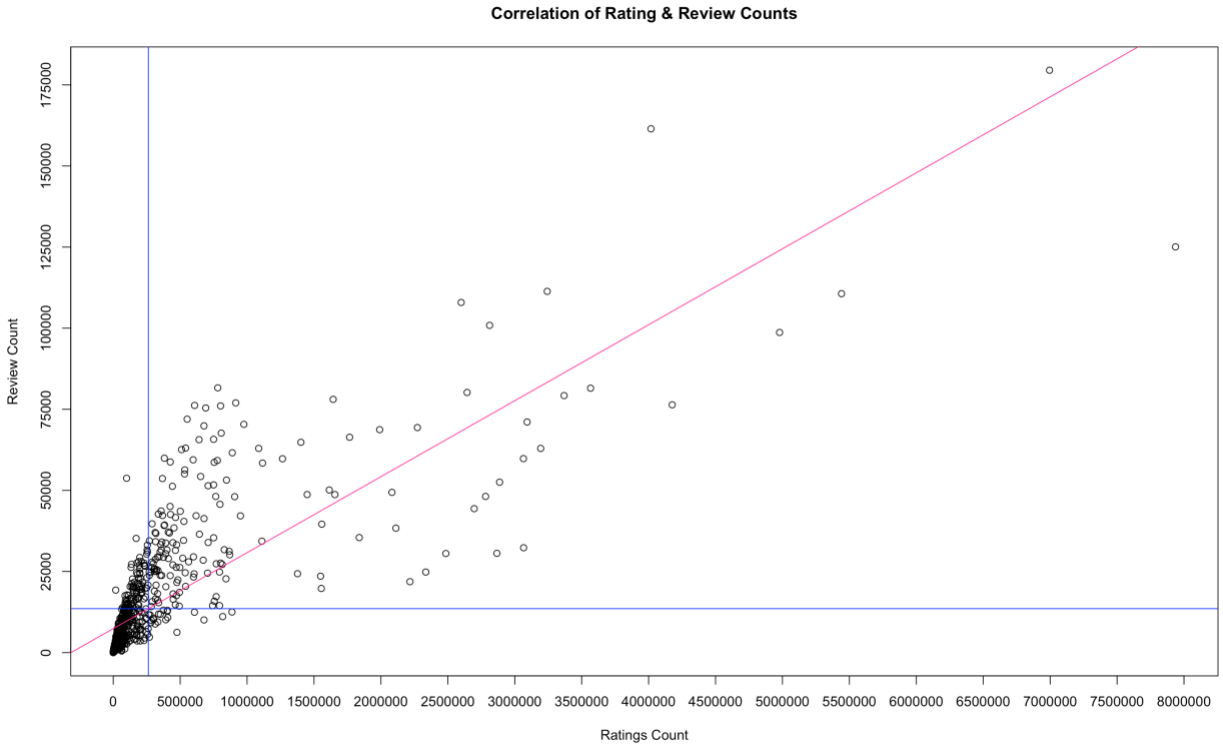


There is a general positive correlation between the two variables, a book with a lot of ratings is likely to have a lot of reviews. The blue lines represent the means of each variable. It's also interesting to note that the scales of the two variables are very different. The highest Rating Count is ~8 million while the highest Review Count is ~200k. This makes sense because someone is more likely to quickly give a book a rating than take the time to write out a review.

I also did a linear regression on these two attributes to help describe the relationship more quantitively.

```
> # linear regression
> rating.review.lm <- lm(my.books$ReviewsCount~my.books$RatingsCount, my.books)
> summary(rating.review.lm)

Call:
lm(formula = my.books$ReviewsCount ~ my.books$RatingsCount, data = my.books)

Residuals:
   Min     1Q Median     3Q    Max
-68174  -6394  -3885   2821  59987

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.392e+03  4.324e+02   17.09   <2e-16 ***
my.books$RatingsCount 2.342e-02  6.079e-04   38.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11890 on 873 degrees of freedom
Multiple R-squared:  0.6296,    Adjusted R-squared:  0.6292
F-statistic:  1484 on 1 and 873 DF,  p-value: < 2.2e-16
```

```
> abline(rating.review.lm, col = "deeppink")
```

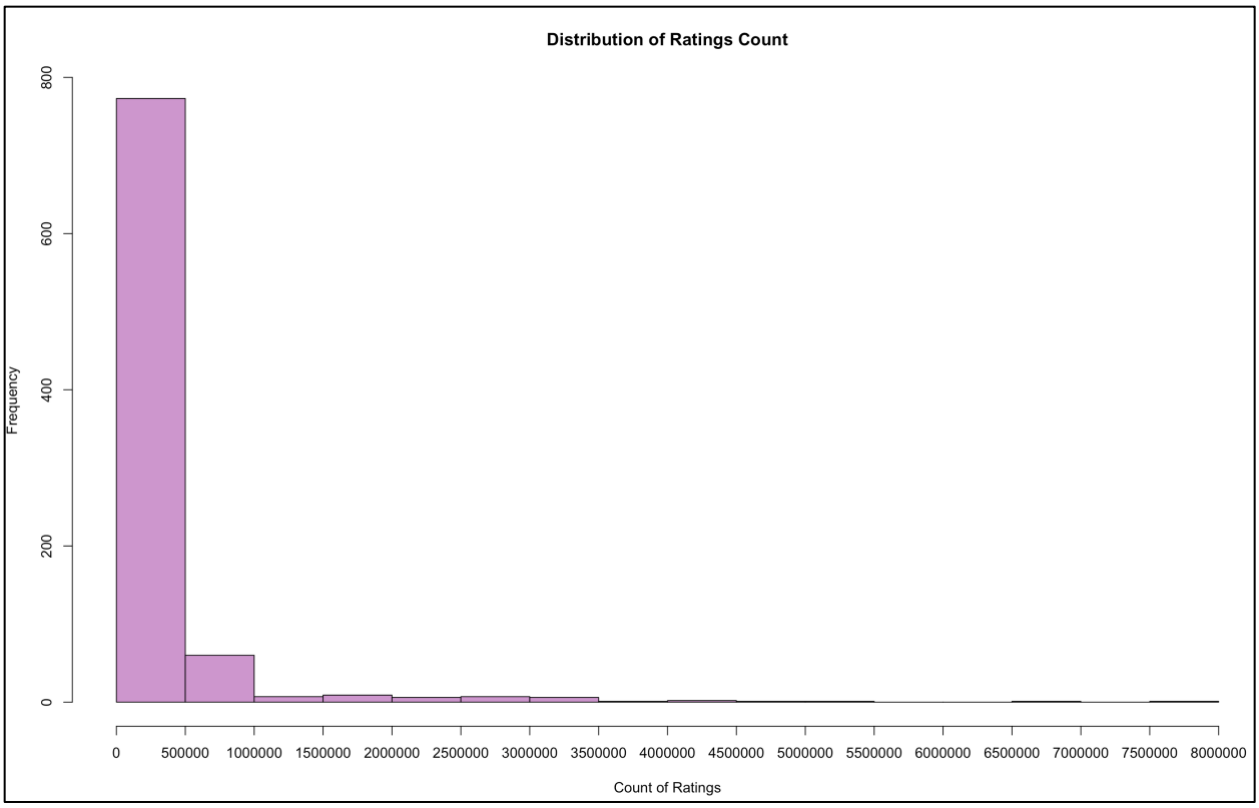


The summary function and the linear regression shows that there is indeed a positive relationship between Rating Count and Review Count. The p-value indicates how well the model fits the data where a small p-value means a strong relationship. In this case, the p-value is 2.2e-16 i.e. almost zero. Therefore, we can say that there is a very strong positive relationship between Rating Count and Review Count. An 'Estimate' of 2.342e-2 means that 0.02342 increase in Rating Count results in an increase of 1 Review Count.

Distribution of Numerical Data

I chose to examine the distribution of the Ratings Count variable.

```
> hist(my.books$RatingsCount,
+      main = "Distribution of Ratings Count",
+      xlab = "Count of Ratings",
+      breaks = seq(0,8000000,500000),
+      xaxt = "n",
+      col = "plum3")
> axis(side = 1, seq(0,8000000,500000), labels = TRUE)
```



The distribution for the Ratings Count column is extremely right skewed as shown by the vast majority of books falling in the first bucket (0 to 50k ratings) in the histogram. The mean of the distribution is 262,583 ratings, which is much higher than the median of 60,244 ratings. There are a few books with millions of ratings that are skewing the mean to be much higher than the median. The meaning of this distribution is that there are a few very popular books that I've logged, as indicated by the number of ratings, but most of the books don't have an impact of that scale on the market. The standard deviation being at 600k also show the large amount of variability in the distribution.

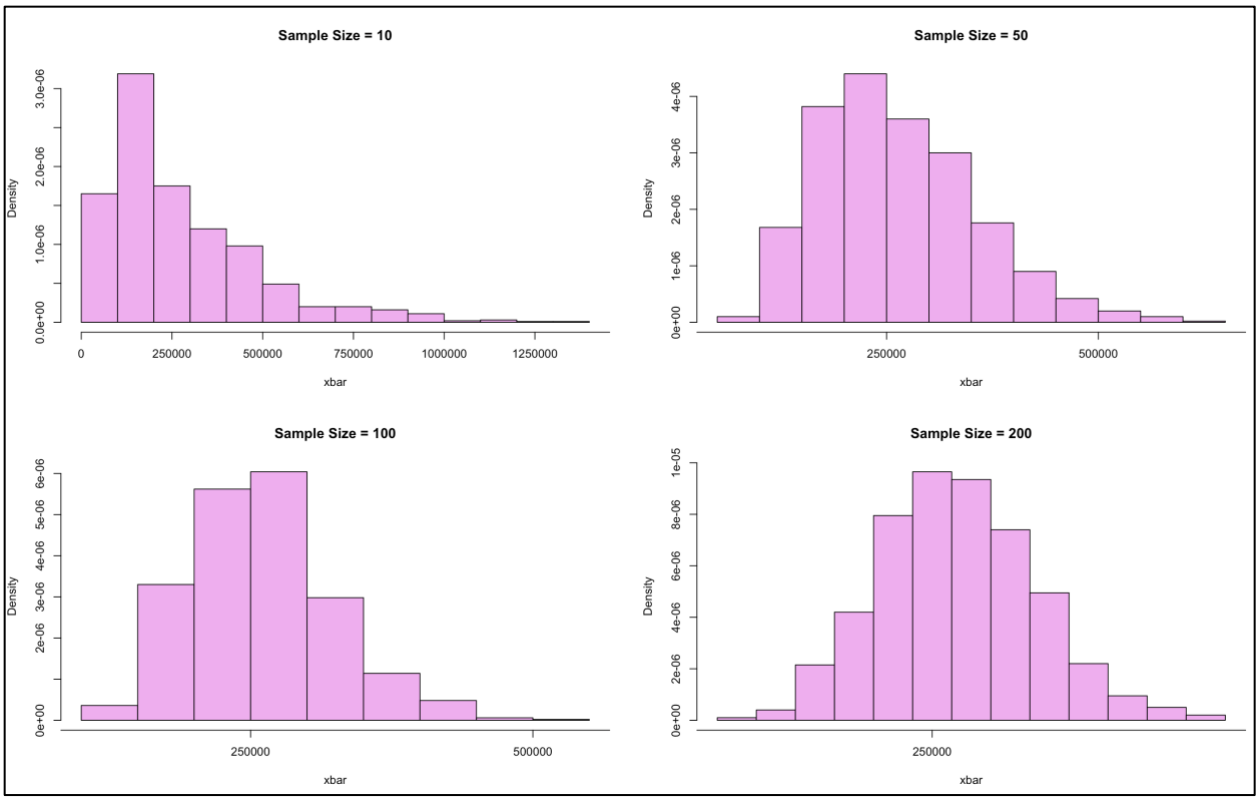
```
> median(my.books$RatingsCount)
[1] 60244
> mean(my.books$RatingsCount)
[1] 262583.2
> sd(my.books$RatingsCount)
[1] 661474.2
```

Applicability of the Central Limit Theorem

Since Ratings Count had such a skewed distribution, I thought it would be interesting to see how the Central Limit Theorem applied to this variable. I drew random samples of 1000 ratings without replacement at four different sample sizes.

```
> set.seed(0417)
> samples <- 1000
> xbar <- numeric(samples)

> par(mfrow = c(2,2))
>
> for (size in c(10, 50, 100, 200)) {
+   for (i in 1:samples) {
+     xbar[i] <- mean(sample(my.books$RatingsCount, size = size,
+                            replace = FALSE))
+   }
+
+   hist(xbar, prob = TRUE,
+        main = paste("Sample Size =", size),
+        xaxt = "n",
+        col = "plum2")
+   axis(side=1, seq(0,1500000,250000), labels=TRUE)
+
+   cat("Sample Size = ", size, " Mean = ", mean(xbar),
+       " SD = ", sd(xbar), "\n")
+ }
Sample Size =  10  Mean =  273195.8  SD =  207460.6
Sample Size =  50  Mean =  262510.4  SD =  91655.11
Sample Size =  100  Mean =  259321  SD =  62620.59
Sample Size =  200  Mean =  263053.2  SD =  39946.24
```

	Mean	Standard Deviation
Original Dataset	262583.2	661474.2
Sample Size = 10	273195.8	207460.6
Sample Size = 50	262510.4	91655.11
Sample Size = 100	259321	62620.59
Sample Size = 200	260053.2	39946.24



The Central Limit Theorem states that the distribution of the sample means for a given sample size of the population has the shape of the normal distribution. To produce the above histograms, the average of Ratings Count for a specific sample size was taken and plotted. This was repeated 1000 times (samples variable) to get the full distribution. As the sample size increased, the distributions resemble the symmetric shape of the normal distribution more and more. As shown in the table, the mean for the five distributions stayed roughly the same at 260k ratings but the standard deviation drastically reduced as the sample size increased. This means that the variability in the Ratings Count data reduced which is evident in the graphs with the data concentrated around the mean. It was very fascinating to see such skewed original distribution also be applicable to the Central Limit Theorem.

Various Sampling Methods

I'll be using various sampling methods on the dataset and showing the frequencies for selected genres. I will also analyze the Number of Pages variable and decide which sampling method produces a mean that's closest to the original dataset.

Seed and sample size that I used for all the sampling methods.

```
> set.seed(0417)
> sample.size <- 40
```

Various Sampling Methods – Simple Random Sample Without Replacement

Randomly pick 40 books out of 875. Since each book is unique and should not get picked more than once, we are sampling without replacement.

Frequency and Percentages of the Genres of the Selected Books

```
> s <- srswor(sample.size, nrow(my.books))
> table(my.books[s != 0, ]$Genre)

       Classics     Contemporary          Fantasy      Historical         Mystery     Non-Fiction
              3               10                8               2               5               5
        Romance Science Fiction
              6               1
> prop.table(table(my.books[s != 0, ]$Genre))

       Classics     Contemporary          Fantasy      Historical         Mystery     Non-Fiction
          0.075           0.250           0.200           0.050           0.125           0.125
        Romance Science Fiction
          0.150           0.025
```

Some of the selected books:

```
> simple.random.sample <- my.books[s != 0, ]
> head(simple.random.sample)
      BookID                                                 Title
6   54985743                             People We Meet on Vacation
20  34374628                                   Next Year in Havana
72  31138556                       Homo Deus: A History of Tomorrow
88  55138170 When Tara Met Farah (Bollywood Drama & Dance Society, #1)
109 42603984                                         Ace of Spades
136 43834909                                     Long Bright River
                                    Author       Genre RatingsCount ReviewsCount MyRating AverageRating
6                                Emily Henry     Romance       198418        25801        5          4.13
20                             Chanel Cleeton  Historical        96933         9517        0          3.93
72                           Yuval Noah Harari Non-Fiction       190033        12885        0          4.21
88                                 Tara Pammi     Romance          129           53        0          3.64
109 Faridah \xcbb\x92k\x8e-\xeay\x92m\x92d\x8e     Mystery        19361         5009        0          4.36
136                              Liz Moore     Mystery        76626         8502        0          4.06
          Publisher        Format NumberOfPages YearPublished   DateRead  DateAdded  Status ReadCount
6     Berkley Books     Paperback           364          2021 2021-09-28 2021-05-25    read         1
20         Berkley     Paperback           361          2018       <NA> 2021-07-12 to-read         0
72          Harper Kindle Edition           450          2017       <NA> 2021-01-29 to-read         0
88                          ebook           181          2021       <NA> 2021-01-15 to-read         0
109       Usborne     Paperback           480          2021       <NA> 2020-12-18 to-read         0
136 Riverhead Books     Hardcover           482          2020       <NA> 2020-11-24 to-read         0
    ReviewRatingRatio
6                0.13
20               0.10
72               0.07
88               0.41
109              0.26
136              0.11
```

Various Sampling Methods – Systematic Sampling, Equal Probabilities

```
> set.seed(0417)
>
> N <- nrow(my.books)
> N
[1] 875
> n <- sample.size
> n
[1] 40
> k <- ceiling(N / n)
> k
[1] 22
> r <- sample(k, 1)
> r
[1] 12
```

The overall data is divided into 40 groups of 22 books, then one book is selected from each group to get a total of 40 books (the desired sample size). Each book in the group has an equal probability of being included in the sample. The first book to be selected is the 12th so the 12th book from the remaining groups are selected. Below is the sequence of the indices of the 40 books to be selected.

```
> s.sys.eq <- seq(r, by = k, length = n)
> s.sys.eq
 [1]  12  34  56  78 100 122 144 166 188 210 232 254 276 298 320 342 364 386 408 430 452 474 496 518 540 562
[27] 584 606 628 650 672 694 716 738 760 782 804 826 848 870
```

Need to check the tail of the selected books to make sure the last index is within the dataset and actually represents a book, which it does.

```
> sys.sample.eq <- my.books[s.sys.eq, ]
> tail(sys.sample.eq)
      BookID                                                      Title           Author    Genre
761   103983                                Skeleton Key (Alex Rider, #3) Anthony Horowitz  Mystery
783    11334                                             Song of Solomon    Toni Morrison Classics
805    22205                                                This Lullaby     Sarah Dessen  Romance
827        2 Harry Potter and the Order of the Phoenix (Harry Potter, #5)     J.K. Rowling  Fantasy
849 10025305                    Clockwork Prince (The Infernal Devices, #2)  Cassandra Clare  Fantasy
871    28187    The Lightning Thief (Percy Jackson and the Olympians, #1)     Rick Riordan  Fantasy
    RatingsCount ReviewsCount MyRating AverageRating                  Publisher    Format NumberOfPages
761        50741         1528        0          4.07                Puffin Books Paperback           327
783        93963         4501        2          4.10                     Vintage Paperback           337
805       178257         5183        4          4.01                       Speak Paperback           345
827      2781798        48128        5          4.50             Scholastic Inc. Paperback           870
849       497876        26232        5          4.41 Margaret K. McElderry Books Hardcover           498
871      2272552        69363        5          4.27       Disney Hyperion Books Paperback           377
    YearPublished DateRead  DateAdded Status ReadCount ReviewRatingRatio
761          2006     <NA> 2015-12-24   read         1              0.03
783          2004     <NA> 2015-12-01   read         1              0.05
805          2004     <NA> 2015-06-16   read         1              0.03
827          2004     <NA> 2015-01-08   read         1              0.02
849          2011     <NA> 2015-01-08   read         1              0.05
871          2006     <NA> 2015-01-08   read         1              0.03
```

Frequency and Percentages of the Genres of the Selected Books

```
> table(sys.sample.eq$Genre)

      Classics    Contemporary         Fantasy      Historical         Mystery     Non-Fiction
             2               6              11               3               6               7
       Romance Science Fiction
             4               1
> prop.table(table(sys.sample.eq$Genre))

      Classics    Contemporary         Fantasy      Historical         Mystery     Non-Fiction
         0.050           0.150           0.275           0.075           0.150           0.175
       Romance Science Fiction
         0.100           0.025
```

<u>Various Sampling Methods – Systematic Sampling, Unequal Probabilities</u>

Now, we determine systematically determine the sample with unequal probabilities. I'm using the NumberOfPages variable to determine the inclusion probabilities therefore, a book with a more pages will have a higher probability of being included in the sample.

```
> set.seed(0417)
>
> pik <- inclusionprobabilities(my.books$NumberOfPages, sample.size)
> length(pik)
[1] 875
> sum(pik)
[1] 40
> s.sys.uneq <- UPsystematic(pik)
> sys.sample.uneq <- my.books[s.sys.uneq != 0, ]
```

A portion of the selected books:

```
> head(sys.sample.uneq)
      BookID                                                                    Title          Author
5   55643287 How the Word Is Passed: A Reckoning with the History of Slavery Across America   Clint   Smith
30  52879286                                                 Humankind: A Hopeful History  Rutger Bregman
53  18295861                                The First Fifteen Lives of Harry August    Claire North
73  53624358                                       Winterkeep (Graceling Realm, #4) Kristin Cashore
96  52914236                                                    Destination Wedding     Diksha Basu
118 50772888                                  No Filter: The Inside Story of Instagram     Sarah Frier
             Genre RatingsCount ReviewsCount MyRating AverageRating                  Publisher    Format
5      Non-Fiction         6233         1169        0          4.80    Little, Brown and Company Hardcover
30     Non-Fiction        27799         3418        0          4.32    Little, Brown and Company Hardcover
53  Science Fiction        77956         9052        0          4.01                      Redhook Hardcover
73          Fantasy         6820         1301        4          4.06 Dial Books For Young Readers Hardcover
96          Romance         2574          414        0          3.32             Ballantine Books Hardcover
118    Non-Fiction         6431          772        0          4.13             Simon & Schuster Hardcover
    NumberOfPages YearPublished   DateRead  DateAdded  Status ReadCount ReviewRatingRatio
5             336          2021       <NA> 2021-10-02 to-read         0              0.19
30            462          2020       <NA> 2021-06-03 to-read         0              0.12
53            416          2014       <NA> 2017-07-01 to-read         0              0.12
73            528          2021 2021-01-23 2020-07-06    read         1              0.19
96            320          2020       <NA> 2021-01-15 to-read         0              0.16
118           352          2020       <NA> 2020-11-24 to-read         0              0.12
```

Frequency and Percentages of the Genres of the Selected Books

```
> table(sys.sample.uneq$Genre)

  Children's     Classics  Contemporary      Fantasy   Historical      Mystery
           1            1             7           12            1            2
 Non-Fiction      Romance Science Fiction
           6            5             5
> prop.table(table(sys.sample.uneq$Genre))

  Children's     Classics  Contemporary      Fantasy   Historical      Mystery
       0.025        0.025         0.175        0.300        0.025        0.050
 Non-Fiction      Romance Science Fiction
       0.150        0.125         0.125
```

<u>Various Sampling Methods – Stratified Sampling</u>

With stratified sampling, the dataset is sampled according to a subgroup, genre in this case, and books are picked from each genre. The proportions for genre in the original dataset are maintained in the sample.

```
> set.seed(0417)
>
> my.books.genres.ordered <- my.books[order(my.books$Genre),]
> freq <- table(my.books.genres.ordered$Genre)
> freq

  Children's     Classics  Contemporary      Fantasy   Historical      Mystery
          14           28           113          279           65           81
 Non-Fiction      Romance Science Fiction
         123           88            84
> sizes <- round(sample.size * freq / sum(freq))
> sizes

  Children's     Classics  Contemporary      Fantasy   Historical      Mystery
           1            1             5           13            3            4
 Non-Fiction      Romance Science Fiction
           6            4             4
```

sizes shows how many books to pick from each genre.

```
> st <- strata(my.books.genres.ordered, stratanames = c("Genre"),
+              size = sizes, method = "srswor")
> head(st)
           Genre ID_unit       Prob Stratum
12     Children's      12 0.07142857       1
19       Classics      19 0.03571429       2
108 Contemporary     108 0.04424779       3
123 Contemporary     123 0.04424779       3
129 Contemporary     129 0.04424779       3
130 Contemporary     130 0.04424779       3
```

```
> strat.sample <- getdata(my.books.genres.ordered, st)
> head(strat.sample)
      BookID                                                Title                Author RatingsCount
661    78418 The Reptile Room (A Series of Unfortunate Events, #2)         Lemony Snicket       204263
634    77013                                        As I Lay Dying       William Faulkner       144930
429 25785649                                   The Way I Used to Be          Amber   Smith        27685
538 25014114                              History Is All You Left Me           Adam Silvera        46992
594 27220730                                   We Are the Ants Shaun David Hutchinson        37080
600 24157347                    The Last Boy and Girl in the World         Siobhan Vivian         4517
    ReviewsCount MyRating AverageRating                                  Publisher         Format NumberOfPages
661         7249        0          3.98                           Scholastic, Inc.      Paperback           192
634         8180        0          3.72                                   Vintage      Paperback           288
429         3814        3          4.18                Margaret K. McElderry Books      Hardcover           367
538         7671        0          4.03                                  Soho Teen           ebook           320
594         6417        0          4.16    Simon  Schuster Children's Publishing      Paperback           464
600          796        0          3.45 Simon & Schuster Books for Young Readers Kindle Edition           320
    YearPublished   DateRead  DateAdded  Status ReadCount ReviewRatingRatio        Genre ID_unit       Prob
661          1999       <NA> 2016-08-02    read         1              0.04   Children's      12 0.07142857
634          1991       <NA> 2016-08-03    read         1              0.06     Classics      19 0.03571429
429          2016 2018-06-09 2016-11-01    read         1              0.14 Contemporary     108 0.04424779
538          2017       <NA> 2017-02-15 to-read         0              0.16 Contemporary     123 0.04424779
594          2016       <NA> 2016-11-01 to-read         0              0.17 Contemporary     129 0.04424779
600          2016       <NA> 2016-10-21 to-read         0              0.18 Contemporary     130 0.04424779
    Stratum
661       1
634       2
429       3
538       3
594       3
600       3
```

Frequency and Percentages of the Genres of the Selected Books

```
> table(strat.sample$Genre)

  Children's     Classics  Contemporary      Fantasy   Historical      Mystery
           1            1             5           13            3            4
 Non-Fiction      Romance Science Fiction
           6            4             4
> prop.table(table(strat.sample$Genre))

  Children's     Classics  Contemporary      Fantasy   Historical      Mystery
  0.02439024   0.02439024    0.12195122   0.31707317   0.07317073   0.09756098
 Non-Fiction      Romance Science Fiction
  0.14634146   0.09756098    0.09756098
```

Comparing the means of the 4 samples with the original dataset for the NumberOfPages variable.

```
> # overall data
> mean(my.books$NumberOfPages)
[1] 381.0617
>
> # simple random sample, without replacement
> mean(simple.random.sample$NumberOfPages)
[1] 346.6
>
> # systematic sampling, equal probabilities
> mean(sys.sample.eq$NumberOfPages)
[1] 437.85
>
> # systematic sampling, unequal probabilities
> mean(sys.sample.uneq$NumberOfPages)
[1] 415.175
>
> # stratified sampling
> mean(strat.sample$NumberOfPages)
[1] 383.439
```

	Difference from the Original Dataset
Simple Random Sample	|381.0617 – 346.6| = 34.4617
Systematic Sampling, Equal	|381.0617 – 437.85| = 56.7883
Systematic Sampling, Unequal	|381.0617 – 415.175| = 34.1133
Stratified Sampling	|381.0617 – 383.439| = 2.3773

The sampling method that resulted in a mean furthest away from the mean of the original dataset was systematic sampling with equal probabilities. And the sampling method that resulted in a mean closest to the mean of the original dataset is stratified sampling. This suggests that stratified sampling is the best sampling option to approximate this dataset.

Let's explore why systematic sampling with equal probabilities resulted in a much larger mean number of pages than the original dataset. Since stratified sampling using genre as the strata resulted in a much closer mean, perhaps genre and number of pages are correlated. For example, Fantasy books are generally long so it would be reasonable to think that perhaps Fantasy books were oversampled in the Systematic Sampling with Equal Probabilities sample and affected the overall mean number of pages. I wrote a function to calculate the mean number of pages per genre so it would be easier to compare each dataset.

```
> mean.by.genre <- function(df) {
+
+   genres <- c("Non-Fiction", "Contemporary", "Romance", "Fantasy", "Mystery",
+               "Historical", "Science Fiction", "Classics", "Children's")
+
+   mean.df <- data.frame(matrix(ncol = 2, nrow = 0))
+   colnames(mean.df) <- c("Genre", "Mean Number of Pages")
+
+   for(i in genres){
+     subset.df <- subset(df, Genre == i)
+     mean.pages <- round(mean(subset.df$NumberOfPages), 2)
+     mean.df[nrow(mean.df) + 1, ] = c(i, mean.pages)
+   }
+
+   mean.df[nrow(mean.df) + 1, ] = c("Overall", round(mean(df$NumberOfPages)), 2)
+   return(mean.df)
+ }
```

Overall Dataset

```
> mean.by.genre(my.books)
            Genre Mean Number of Pages
1      Non-Fiction               319.67
2     Contemporary               335.25
3          Romance               359.08
4          Fantasy               452.94
5          Mystery               358.25
6       Historical               390.09
7  Science Fiction               399.24
8         Classics               284.29
9       Children's               170.57
10         Overall                  381
Warning message:
In matrix(value, n, p) :
  data length [3] is not a sub-multiple or multiple of the number of columns [2]
```

(note – not sure what the cause of the warning message is but otherwise the function works as intended)

Systematic Sampling, Equal Probabilities

```
> mean.by.genre(sys.sample.eq)
            Genre Mean Number of Pages
1      Non-Fiction               339.14
2     Contemporary               388.83
3          Romance                340.5
4          Fantasy               540.45
5          Mystery                449.5
6       Historical               396.67
7  Science Fiction                  827
8         Classics                  393
9       Children's                  NaN
10         Overall                  438
Warning message:
In matrix(value, n, p) :
  data length [3] is not a sub-multiple or multiple of the number of columns [2]
```

Stratified Sampling

```
> mean.by.genre(strat.sample)
             Genre Mean Number of Pages
1      Non-Fiction                295.83
2     Contemporary                 347.2
3          Romance                353.25
4          Fantasy                457.08
5          Mystery                   380
6       Historical                   418
7  Science Fiction                400.25
8         Classics                   288
9       Children's                   192
10         Overall                   383
Warning message:
In matrix(value, n, p) :
  data length [3] is not a sub-multiple or multiple of the number of columns [2]
```

	Total Dataset	Systematic Sampling, Equal Probabilities Sample	Stratified Sampling Sample
Non-Fiction	319.67	339.14	295.83
Contemporary	335.25	388.83	347.2
Romance	359.08	340.5	353.25
Fantasy	452.94	540.45	457.08
Mystery	358.25	449.5	380
Historical	390.09	396.67	418
Science Fiction	399.24	827	400.25
Classics	284.29	393	288
Children's	170.57	NaN	192
Overall	381	438	383

The four genres in blue show instances where the systematic sample's mean number of pages was much higher than the total dataset's mean number of pages and the stratified sample's mean number of pages. Since the rest of the genres are comparable (except for Children's where the systematic sample's mean number of pages is effectively 0), these four genres are raising the mean number of pages for systematic sampling.

I think this has to do more with the particular books that were sampled in the systematic sample rather than a particular genre being oversampled. For example, one of the Fantasy books that was sampled was the second longest book in the whole dataset. Additionally, the one Science Fiction book that was included was the third longest Science Fiction book in the whole dataset. (Code for these two examples is below.) If I used a different seed value, perhaps the sample that resulted from Systematic Sampling with Equal Probabilities would approximate the number of pages of the overall dataset better.

```
> # second longest book in the whole dataset is in the systematic sampling with
> # equal probabilities sample
> sorted.by.pages <- my.books[order(-my.books$NumberOfPages),]
> sorted.by.pages[seq(1:5), c(2,11)]
                                                 Title NumberOfPages
654                              The Lord of the Rings          1216
585      The Way of Kings (The Stormlight Archive, #1)          1007
392               Kingdom of Ash (Throne of Glass, #7)           984
373 Queen of Air and Darkness (The Dark Artifices, #3)           912
440                              Bhagavad-Gita As It Is           883

> sys.sample.eq[sys.sample.eq$NumberOfPages == max(sys.sample.eq$NumberOfPages), c(2,11)]
                                        Title NumberOfPages
585 The Way of Kings (The Stormlight Archive, #1)          1007

> # third longest science fiction book in the whole dataset is in the systematic sampling
> # with equal probabilities sample
> sci.fi.overall <- subset(my.books, Genre == "Science Fiction")
> sorted.sci.fi.overall <- sci.fi.overall[order(-sci.fi.overall$NumberOfPages),]
> sorted.sci.fi.overall[seq(1:5), c(2,11)]
                              Title NumberOfPages
112        To Sleep in a Sea of Stars           878
492                          Seveneves           867
717 Winter (The Lunar Chronicles, #4)           827
416              Ringer (Replica, #2)           672
861           The Host (The Host, #1)           624

> sci.fi.sys.sample.eq <- subset(sys.sample.eq, Genre == "Science Fiction")
> sci.fi.sys.sample.eq[,c(2,11)]
                                    Title NumberOfPages
717 Winter (The Lunar Chronicles, #4)           827
```

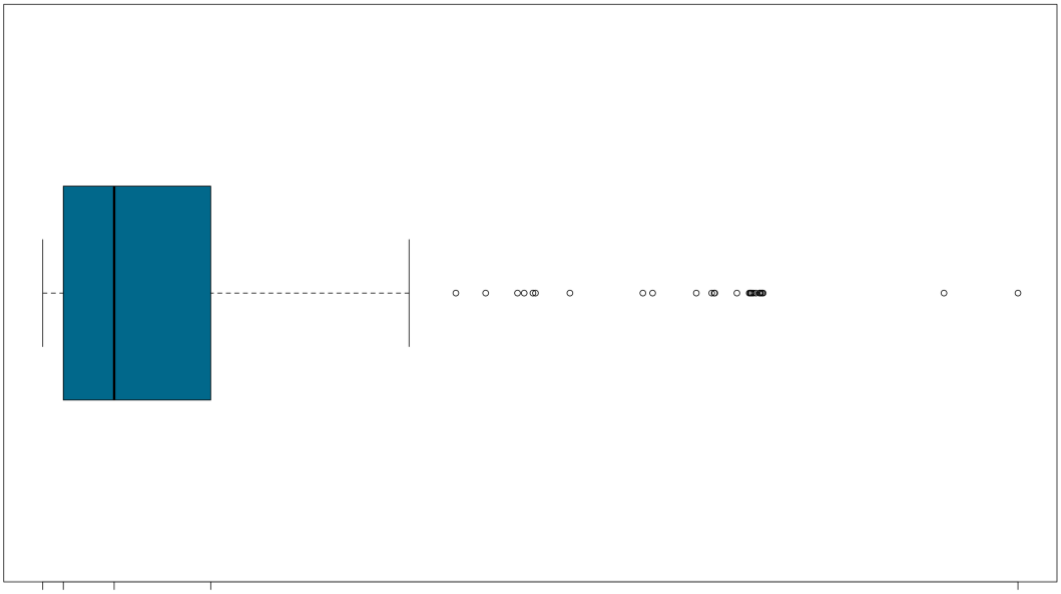
An Extra Graph

Using the two Date columns, DateRead and DateAdded, I wanted to see what the spread for how long it took me to read a book I had added as “Want to Read” was.

```
> # new column
> my.books.read$DaysToRead <- as.numeric(my.books.read$DateRead -
+                                          my.books.read$DateAdded, unit = "days")
```

```
> head(my.books.read)
     BookID                                                                  Title              Author
1  52128695 Ace: What Asexuality Reveals About Desire, Society, and the Meaning of Sex        Angela  Chen
6  54985743                                                People We Meet on Vacation         Emily Henry
7  43575115                                                          The Starless Sea    Erin Morgenstern
11 55404546                                                             Malibu Rising Taylor Jenkins Reid
12 42201995                                                       The Stationery Shop       Marjan Kamali
15 52973514                                                       Days of Distraction     Alexandra Chang
          Genre RatingsCount ReviewsCount MyRating AverageRating        Publisher    Format NumberOfPages
1   Non-Fiction         3982          889        5          4.40     Beacon Press Hardcover           224
6       Romance       198418        25801        5          4.13    Berkley Books Paperback           364
7       Fantasy       133333        26271        4          3.84  Doubleday Books Hardcover           498
11   Historical       199061        24297        5          4.13 Ballantine Books Hardcover           369
12   Historical        35923         5107        4          4.21    Gallery Books Hardcover           322
15 Contemporary         4012          622        3          3.61    HarperCollins Hardcover           312
   YearPublished   DateRead  DateAdded Status ReadCount ReviewRatingRatio DaysToRead
1           2020 2021-10-05 2021-07-13   read         1              0.22         84
6           2021 2021-09-28 2021-05-25   read         1              0.13        126
7           2019 2021-08-29 2020-01-18   read         1              0.20        589
11          2021 2021-08-05 2021-01-31   read         1              0.12        186
12          2019 2021-07-29 2021-05-05   read         1              0.14         85
15          2020 2021-07-18 2021-04-21   read         1              0.16         88
```

```
> boxplot(my.books.read$DaysToRead,
+         horizontal = TRUE,
+         col = "deepskyblue4",
+         xlab = "Days To Read",
+         xaxt = "n")
> axis(side = 1, fivenum(my.books.read$DaysToRead), labels = TRUE)
```




```
> fivenum(my.books.read$DaysToRead)
[1]    0   40  139  327 1898
```

Since the maximum value was so high (1898 days = 5 years), I was curious to see which book that was.

```
> # Book that was shelfed for the longest period
> subset(my.books.read, DaysToRead == max(my.books.read$DaysToRead, na.rm = TRUE))
      BookID                            Title          Author   Genre RatingsCount ReviewsCount MyRating
280 17699853 Chain of Gold (The Last Hours, #1) Cassandra Clare Fantasy        63383        10270        3
    AverageRating                  Publisher    Format NumberOfPages YearPublished   DateRead  DateAdded
280          4.46 Margaret K. McElderry Books Hardcover           592          2020 2020-03-20 2015-01-08
    Status ReadCount ReviewRatingRatio DaysToRead
280   read         1              0.16       1898
```

This particular book was announced long before it was released so I must have shelfed it as “Want To Read” when it was first announced. This would explain the long time period between adding and reading.