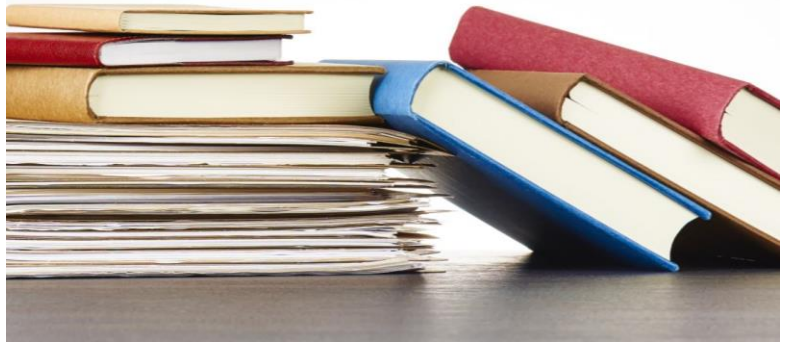
BOOKS PER MONTH ANALYSIS

NOELLE RICHARDS

BUSINESS CASE

"Americans read an average (mean) of 12 books per year, while the typical (median) American has read 4 books in the last 12 months." (Perrin, 2020)

- As the "business" in this particular case, I want to read more books because it connects me to my community (book club, mom groups, play dates). I also there is education and emotional value in reading.
- If I take the average of books read per month based on my historical data, I find that I read 4.875 books each month. However, the median is 3. I want the median and average to be closer (thus reducing the variability) and with rounding, that puts my goal at 5 books per month.
- In order to meet my goal of reading 5 books per month (60 per year), I want to understand the common themes of my reading selections and use them to optimize my future choices.



Perrin, A. (2020, May 30). Majority of Americans are still reading print books. Pew Research Center: Internet, Science & Tech. Retrieved December 8, 2021, from https://www.pewresearch.org/internet/2016/09/01/book-reading-2016/.

OPERATIONAL DEFINITION

- A book is considered read when it is marked as such in GoodReads. The date it has been marked as "Read" will be the month in which it is counted towards my goal. In order to filter out children's books, only books with over 100 pages will be considered.
- Books are automatically tracked in GoodReads if read through the Kindle app. Otherwise they need to be added manually. Through the website, I can download an CSV of my entire GoodReads library, including date, rating, author, and other information.

PROBLEM AND GOAL

Problem:

- Currently, I only read 5 books a month 25% of the time.

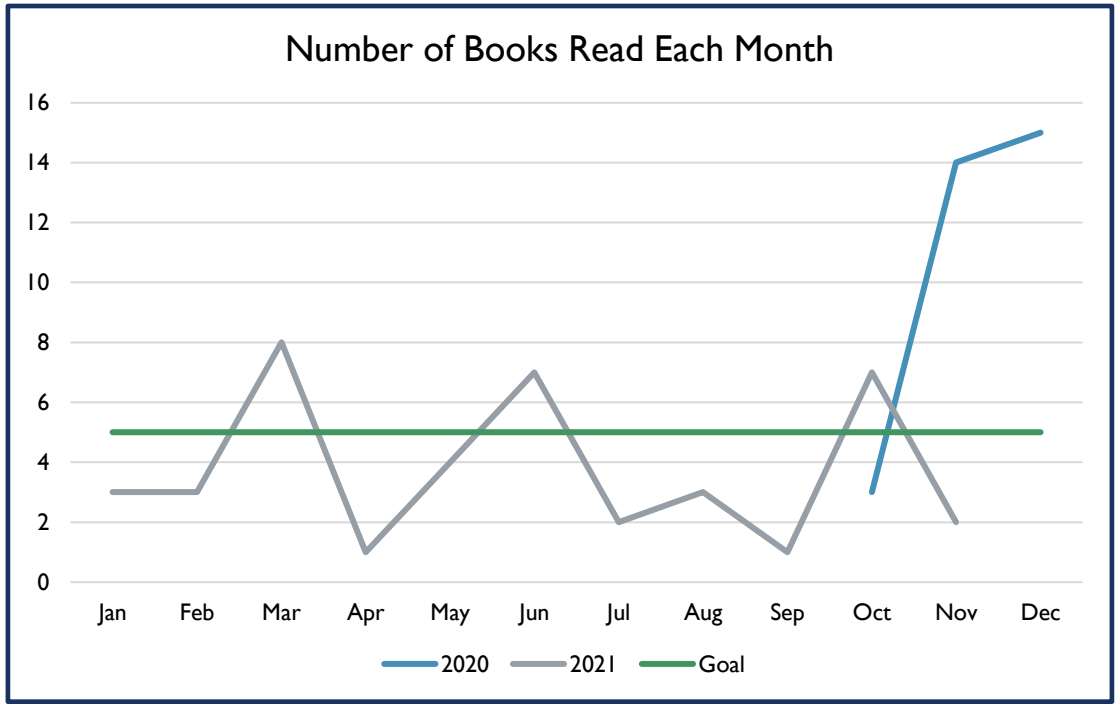
Goal:

- Read 5 books a month 80% of the time.

Rather than requiring a 75% increase in adherence to my goal, I decided on a 55% increase to make it more attainable. If this goal is met, it raises my sigma level from below 1 to above 2 and is a significant change.

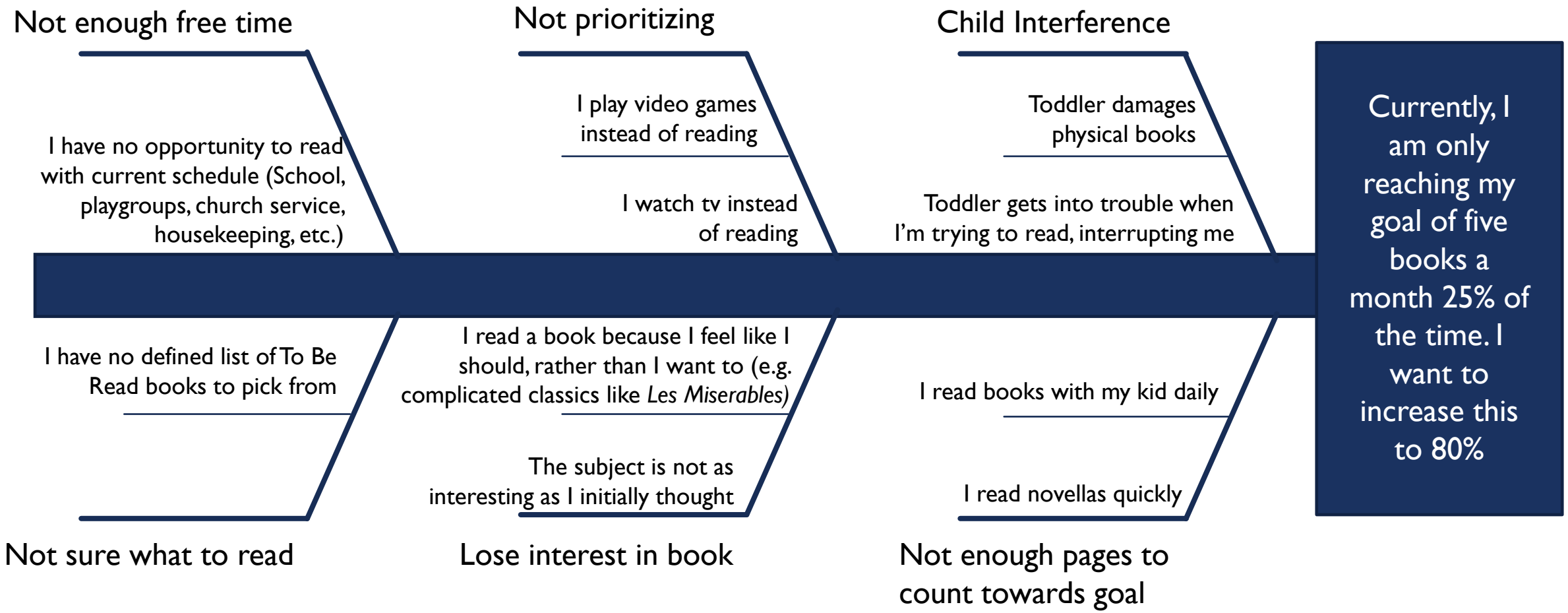
MEASURE

Process Sigma Calculation	
Number of defect opportunities per month	1
Number of months processed	16
Total number of defects made	12
Defects Per Opportunity (DPO)	0.75
Yield	25.0%
Process Sigma	0.83



A "defect" means I failed to meet my reading goal that month, meaning each month is an opportunity for a defect. I measured 16 months and only succeeded 4 times in meeting my goal, breaking down to a 25% yield and a process sigma of 0.83.

IDENTIFICATION OF X'S - FISHBONE DIAGRAM



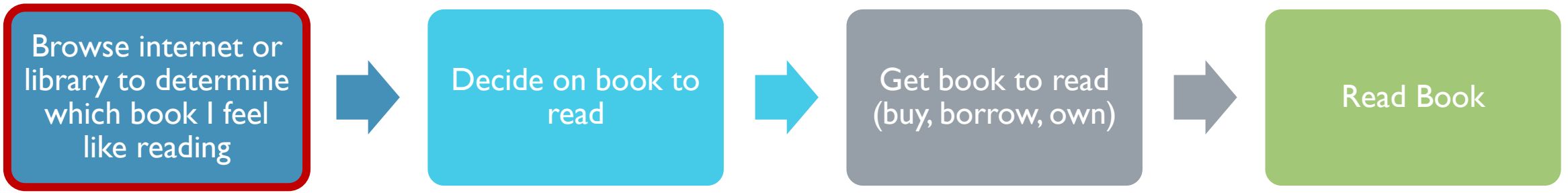
MEASURABILITY OF X'S

Not Enough Free Time	Not Prioritizing	Child Interference	Not Sure What to Read	Lose Interest in Books	Not Enough Pages to Count Towards Goal
• I am very busy with many different activities. While it is possible to measure free time, it would be time consuming, and potentially not worth the effort.	• Similar to not having enough free time, it would be time consuming to track all free time activities and determine how much of that should be reading instead.	• I could measure the number of times that my child interrupts my reading or damages my books	• I can spend up to an hour trying to decide what to read by browsing the internet and library.	• I can measure this by my "Did Not Finish" list on GoodReads. However, it is not accurate as I tend to input books into my lists after I finish them, rather than when I start them.	• This is easily measurable by the number of pages in the books I finish. I had 8 books in the time period encapsulated by my data that did not meet the page requirement.

I will propose solutions for "Child Interference", "Not Sure What to Read", "Lose Interest in Books", and "Not Enough Pages to Count Towards Goal".

PROPOSED SOLUTION - VALUE STREAM MAP

Red border: Non-Value Add Step



The first step in my current progress does not add value to the process, and instead I can end up wasting nearly an hour per book just finding what I want to read.

Proposed Solution:
Create a "To Be Read" list on GoodReads and add to it when I see or hear of interesting books. That way, when I need to decide on a book, I can simply browse a list of books I already am interested in and cut down on the time spent on this part of the process. I can use that extra time to finish more books.

PROPOSED SOLUTIONS - POKA YOKE

Problem

- Toddler damages books by ripping pages or covers
- Not enough pages to count towards goal

Proposed Solutions

- My toddler can't rip pages if there are not any to rip, so I will read books either on my phone or my Kindle Paper White. With some experimentation, I have found the Kindle to be most effective as it is not colorful and has no other functions except reading, leaving my toddler bored quickly if she does happen upon it.
- Do not start a book with less than 100 pages unless the "5 books per month" goal has already been met.

LOGISTIC REGRESSION – PROPOSED SOLUTION

One of reasons I don't meet my goal is that I start books and then lose interest. By using logistic regression with some of the variables in my data, I hope to discover which variables most impact my monthly goal completion.

The logistic regressions will be predicting the probability of goal completion

I split my variables into two pieces, one with ratings and number of pages and the other with the genres and created the regressions in R.

LOGISTIC REGRESSION #1 – RATINGS AND NUMBER OF PAGES

Code

```r
library(tidyverse)
library(fastDummies)

data <- read.csv('GoodReads_Data.csv')
data <- dummy_cols(data, select_columns = 'Genre')

ratings <- glm(Goal.Met.~My.Rating+Average.Rating+Number.of.Pages, data = data, family = "binomial")
summary(ratings)
```

Conclusions

- Only one variable is found significant in this logistic regression: the average number of pages per book. The p-value of the other variables is so high that I feel comfortable concluding that the ratings don't have a strong influence on my ability to meet my month goal.

Summary

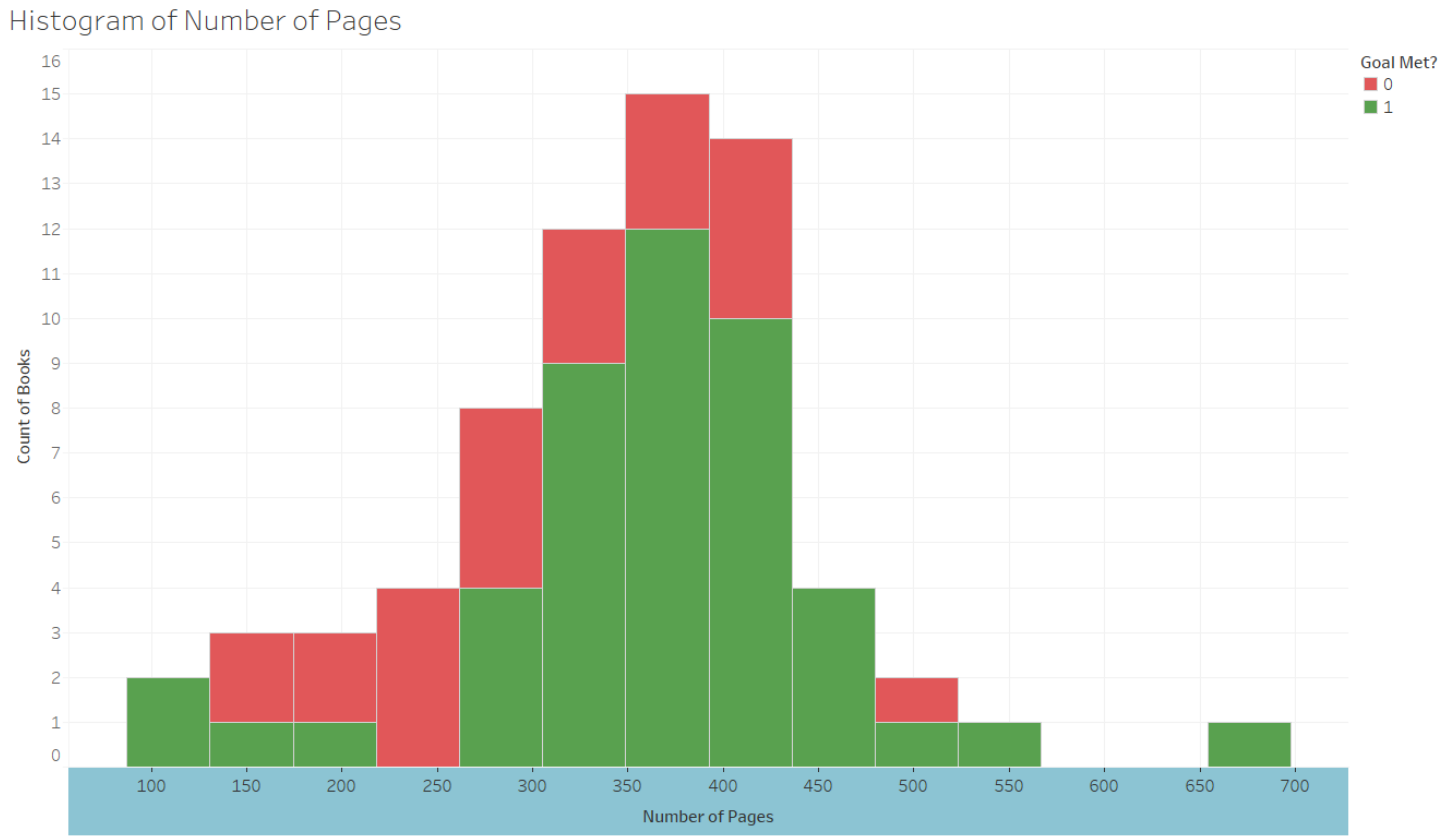
```
Call:
glm(formula = Goal.Met. ~ My.Rating + Average.Rating + Number.of.Pages,
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8752  -1.1818   0.6822   0.8438   1.6217

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.929378   4.491236  -0.207   0.8361
My.Rating       -0.241430   0.352598  -0.685   0.4935
Average.Rating   0.094818   1.115707   0.085   0.9323
Number.of.Pages  0.006799   0.003019   2.252   0.0243 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 87.839  on 68  degrees of freedom
Residual deviance: 81.144  on 65  degrees of freedom
AIC: 89.144

Number of Fisher Scoring iterations: 4
```

LOGISTIC REGRESSION #1 – NUMBER OF PAGES ANALYSIS

Each book is put in the corresponding bin based on page number and colored based on completion of the monthly goal.

If I only consider this graph, I might conclude that I do best at meeting the goal when the books are between 300 and 450 pages. However, common sense tells me that if I read shorter books, I will read more in a shorter period. I believe if I had more data, especially in the 100-200 page range, my model and histogram might show a different picture.

I think the data is actually showing that most of the books I'm interested in tend to be around 300-450 pages, making it a correlation, rather than causation. There is a very steep dropoff in data after 450 pages though, implying that length of book could be a factor in book selection. To state it more simply, if it is too long, I won't even attempt it.



LOGISTIC REGRESSION #2 – GENRES

Code

```
genre <- glm(data$Goal.Met.~data$Genre_Classics+data$Genre_Fantasy+data$Genre_Fiction
             +data$`Genre_Graphic novel`+data$`Genre_Historical Fiction`
             +data$`Genre_Historical Romance`+data$Genre_Mystery+data$Genre_Nonfiction
             +data$Genre_Romance+data$`Genre_Young Adult`, family = 'binomial')
summary(genre)
```

Conclusions

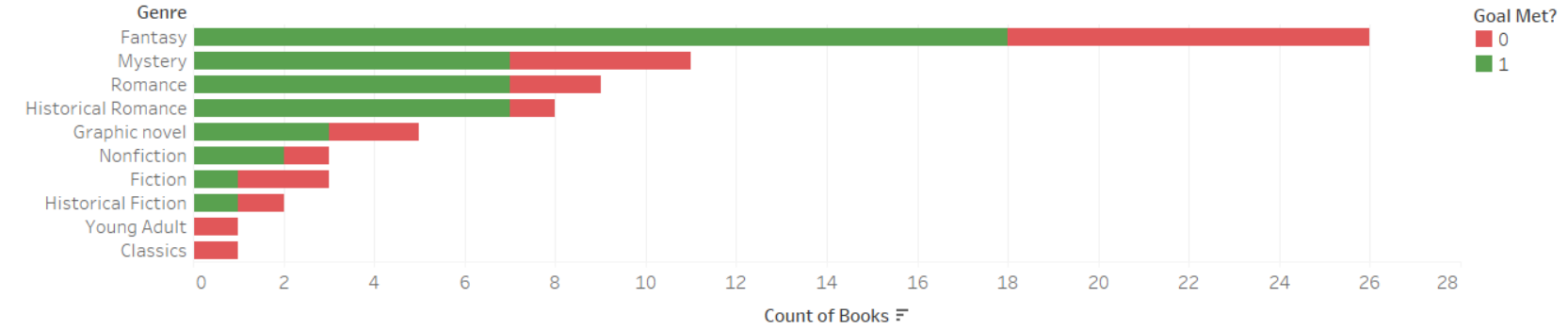
I split the "Genre" category into dummy variables for the regression. I thought that I might get bad results from genre variables because the data is so sparse and somewhat uniform, which I confirmed with the summary. No variable is significant.

Summary

```
Call:
glm(formula = data$Goal.Met. ~ data$Genre_Classics + data$Genre_Fantasy +
    data$Genre_Fiction + data$`Genre_Graphic novel` + data$`Genre_Historical Fiction` +
    data$`Genre_Historical Romance` + data$Genre_Mystery + data$Genre_Nonfiction +
    data$Genre_Romance + data$`Genre_Young Adult`, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0393  -1.3537   0.7090   0.8576   1.4823

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.657e+01  2.400e+03  -0.007    0.994
data$Genre_Classics           -7.265e-09  3.393e+03   0.000    1.000
data$Genre_Fantasy             1.738e+01  2.400e+03   0.007    0.994
data$Genre_Fiction             1.587e+01  2.400e+03   0.007    0.995
data$`Genre_Graphic novel`     1.697e+01  2.400e+03   0.007    0.994
data$`Genre_Historical Fiction` 1.657e+01  2.400e+03   0.007    0.994
data$`Genre_Historical Romance` 1.851e+01  2.400e+03   0.008    0.994
data$Genre_Mystery             1.713e+01  2.400e+03   0.007    0.994
data$Genre_Nonfiction          1.726e+01  2.400e+03   0.007    0.994
data$Genre_Romance             1.782e+01  2.400e+03   0.007    0.994
data$`Genre_Young Adult`              NA         NA      NA       NA

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 87.839  on 68  degrees of freedom
Residual deviance: 79.221  on 59  degrees of freedom
AIC: 99.221

Number of Fisher Scoring iterations: 15
```

LOGISTIC REGRESSION #2 – GENRE ANALYSIS

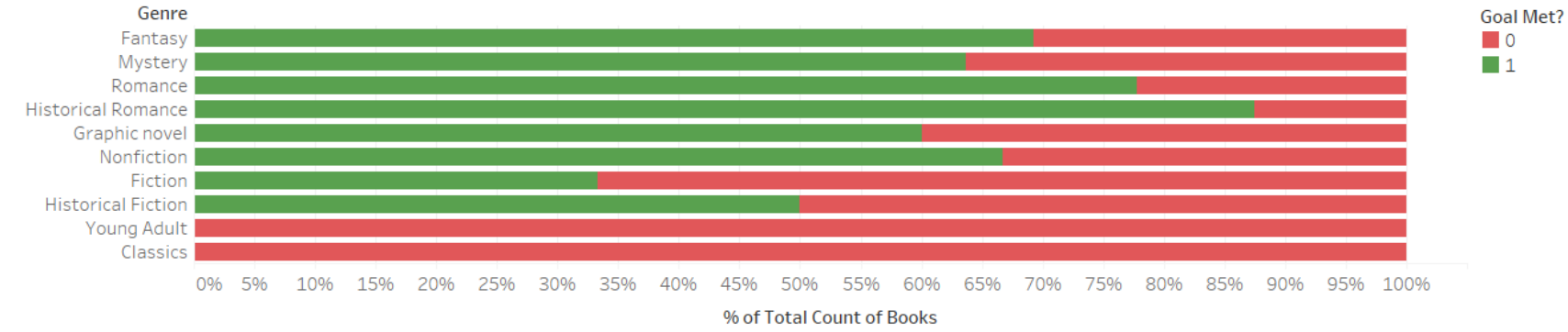
Genres



Count of GoodReads_Data.csv for each Genre. Color shows details about Goal Met?.

While my most read genre is Fantasy by far, I have the highest goal completion rate in Historical Romance.

Genres



% of Total Count of GoodReads_Data.csv for each Genre. Color shows details about Goal Met?. Percents are based on each row of the table.

SOLUTIONS SUMMARY

To improve my completion rate, I have come up with the following solutions:

- Problem 1: Not sure what to read
  - Solution 1: Create a "To Be Read" list on GoodReads and add to it when I see or hear of interesting books.
- Problem 2: Not enough pages to count towards goal
  - Solution 2: Do not start a book with less than 100 pages unless the "5 books per month" goal has already been met.
- Problem 3 : Child Interference
  - Solution 3: Read books on phone or Kindle.
- Problem 4: Lose interest in books
  - Solution 4: Pick more Historical Romance books.

Most of these fall into book selection and can be consolidated into a set of guidelines I can follow:

- From To Be Read list
- More than 100 pages and between 300-450 pages
- Focus on Historical Romance Genre

CONCLUSION

Most of these fall into book selection and can be consolidated into a set of guidelines I can follow:

- From To Be Read list
- More than 100 pages and between 300-450 pages
- Focus on the Historical Romance Genre

As the original data takes place over the course of 16 months, I will not be able to implement changes and measure them in time to prove a noticeable difference in goal completion but based on my analysis, I feel confident that this measures will work.