Created & Presented
by Kiersten Johns

Book Rating Classifier & Recommender System

A passion project

Table of contents

01 Introduction

02 Problem Statement

03 Data & EDA

04 Recommender

05 Classification Modeling

06 Final Thoughts

01 Introduction

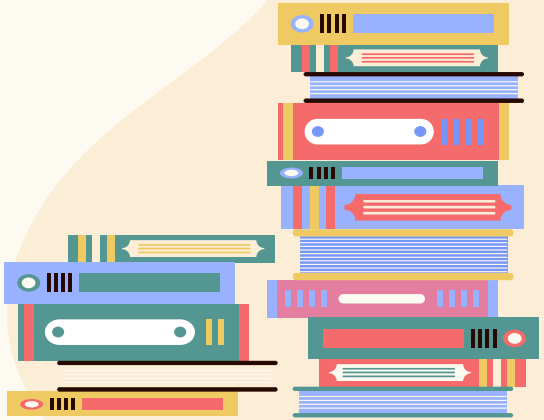
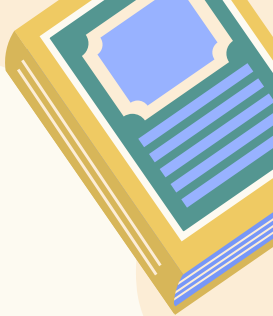


Image source: Adobe
Sound source: pixabay

BAM!

SYMPTOMS OF A BOOK HANGOVER:

TBR (To-Be-Read) List:



Book 1 by Author 1

Book 2 by Author 2

Book 3 by Author 3

Book 4 by Author 4

02 Problem Statement

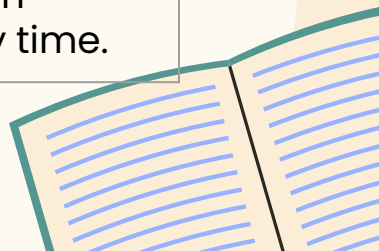
Determining the book that will aid a reader in recovering from a book hangover is the primary objective of this project. The goals of this endeavor encompass two key aspects:

1. Creation of a classification model to forecast, with at least 51% accuracy, what books from a reader's TBR list should be read next
2. Creation of a recommendation system, based off of nearly 6 million book ratings, to help readers continue to add to their TBR list

03 Data & EDA

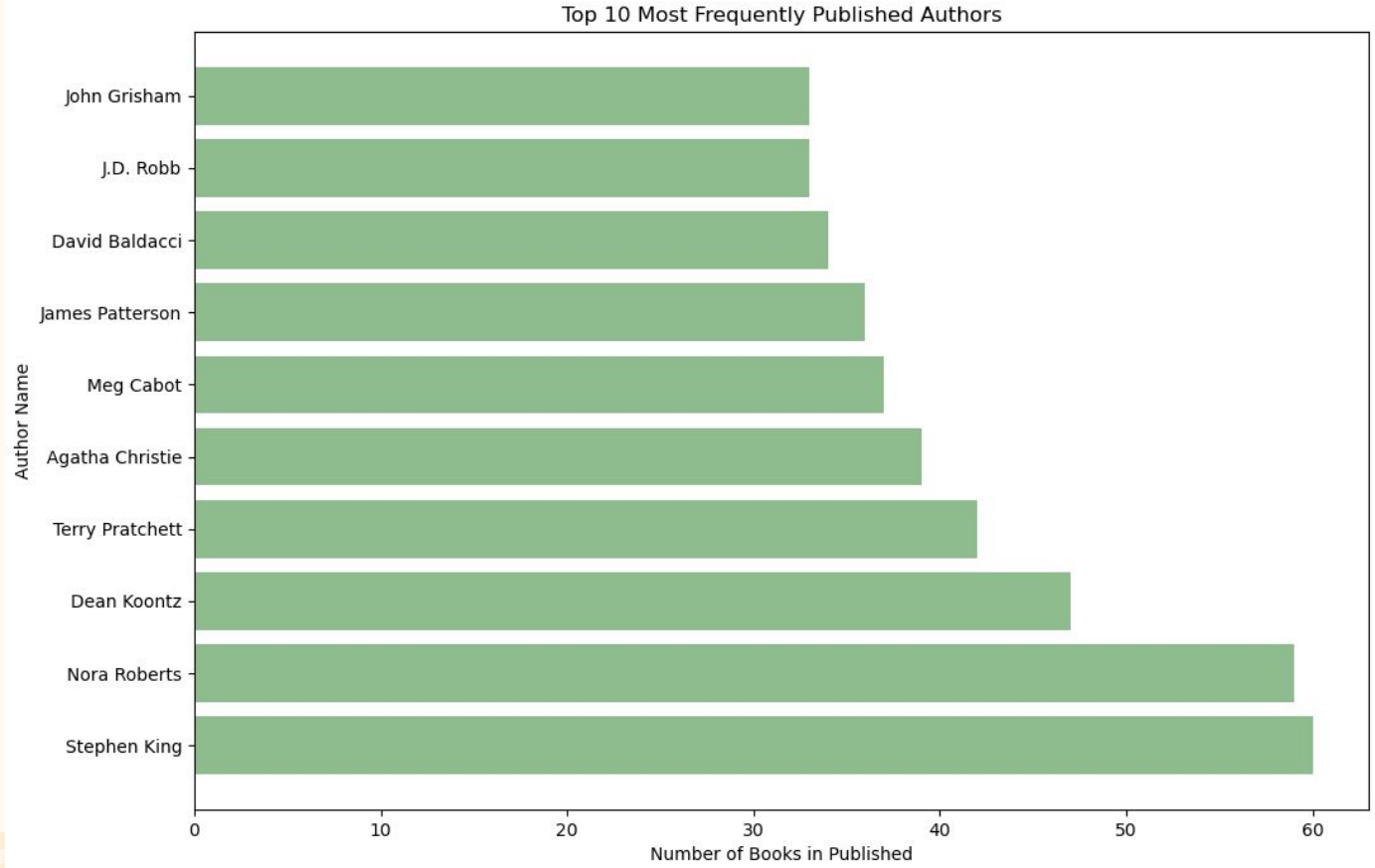
Name	Dataset	Description
books	books.csv	Top 10,000 most popular books containing metadata for each book • goodreads IDs, authors, title, average rating, etc. The metadata has been extracted from GoodReads.
ratings	ratings.csv	Contains 6 million ratings sorted by time.

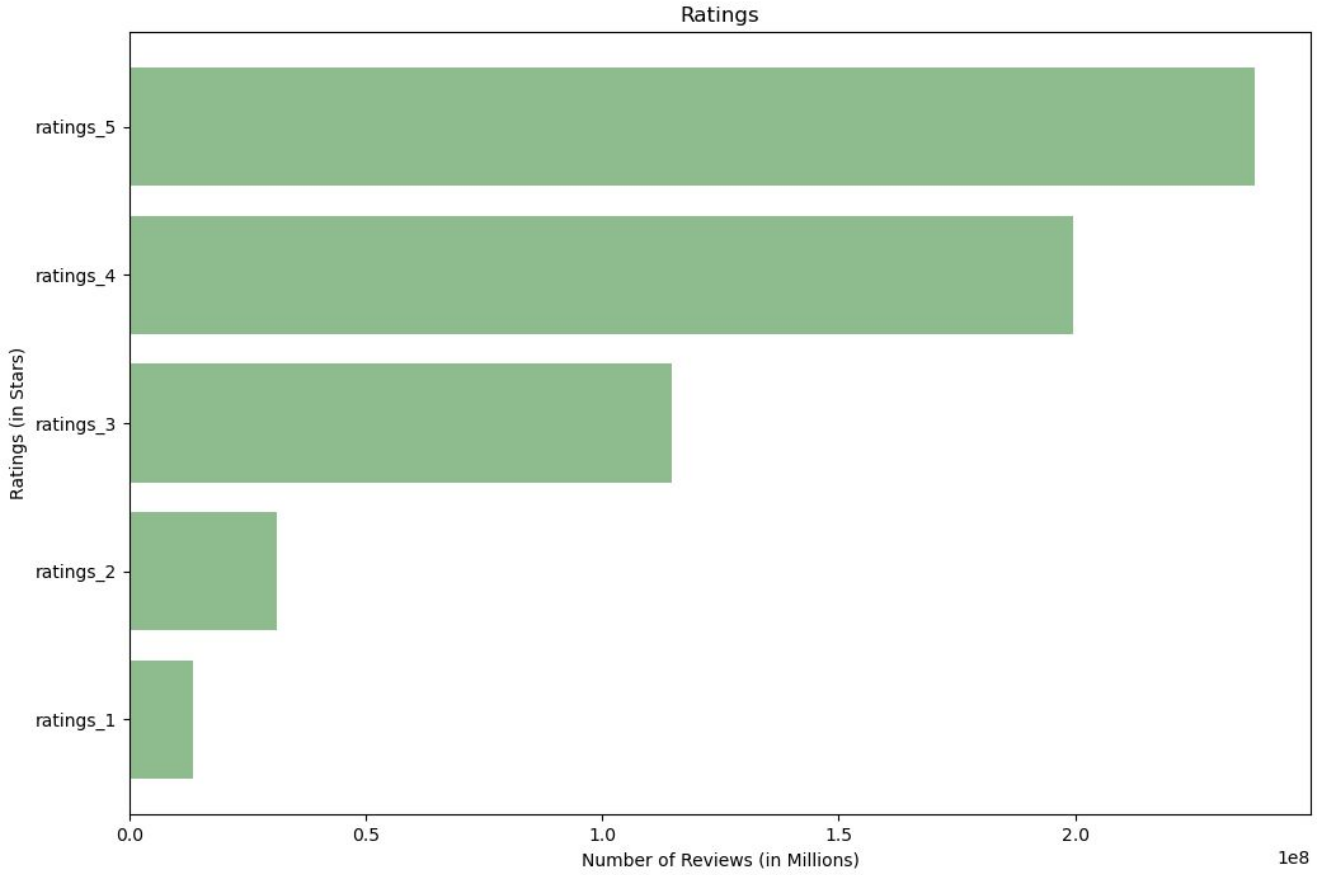
Data Source: zygmuntz's github

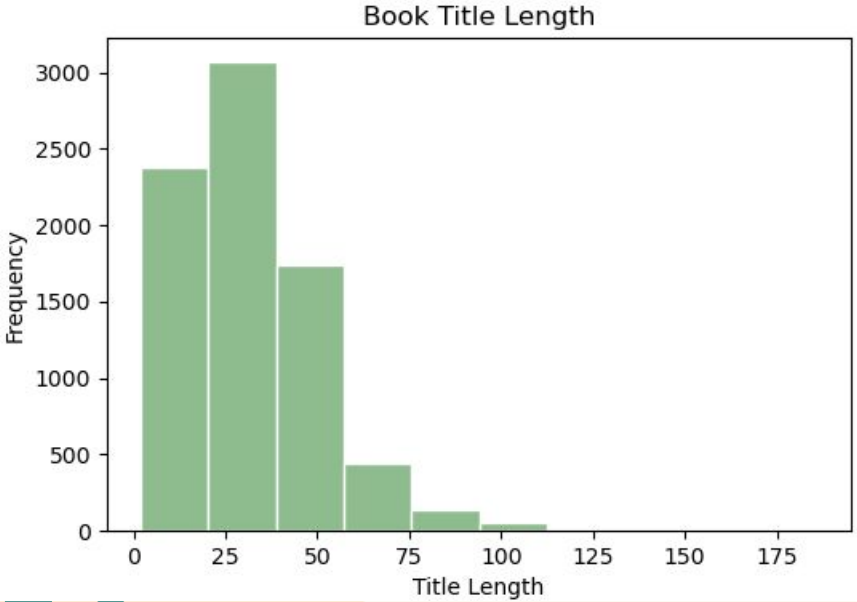


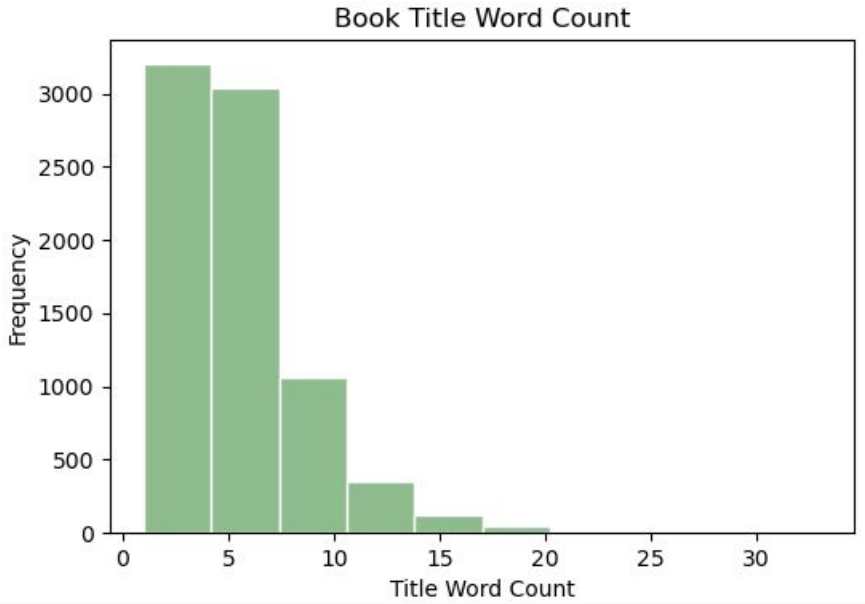
Interesting Findings:

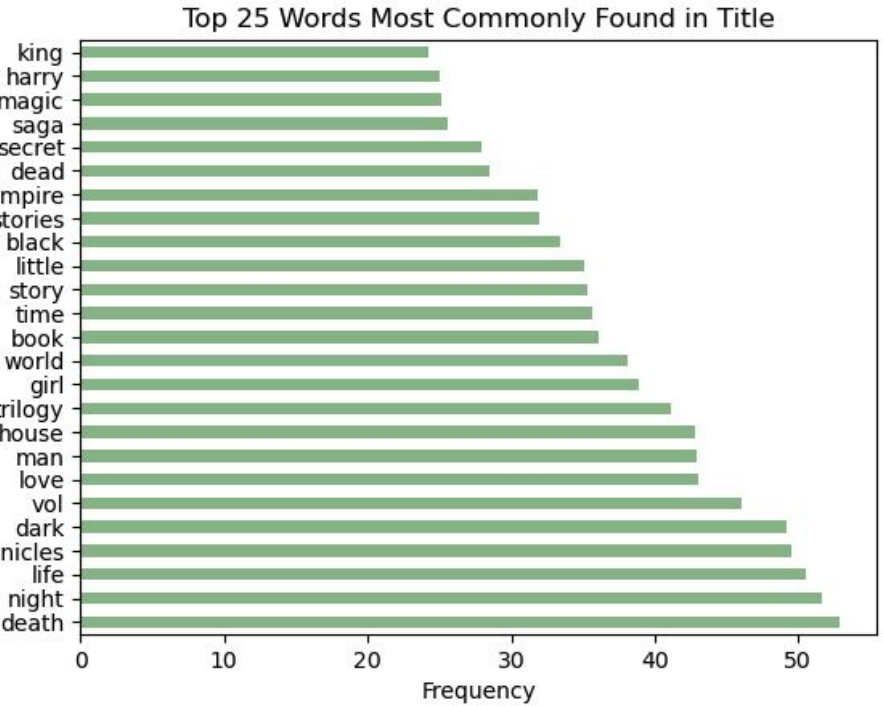
- 9,964 unique book titles
  - Some titles were duplicates, but they were still unique books due to the fact they had different authors
- The earliest publishing date is listed at -1750
- 34,252 unique tags were provided by users
- 24 language codes

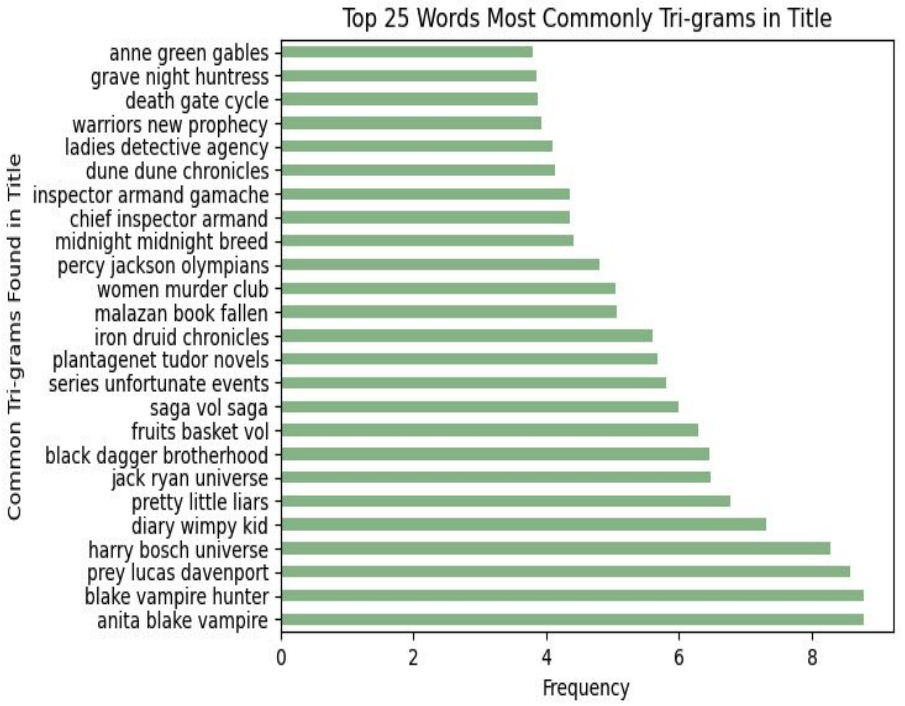

Top 10 Most Frequently Published Authors
John Grisham
J.D. Robb
David Baldacci
James Patterson
Meg Cabot
Agatha Christie
Terry Pratchett
Dean Koontz
Nora Roberts
Stephen King
Author Name
0
10
20
30
40
50
60
Number of Books in Published


Ratings
ratings_5
ratings_4
ratings_3
ratings_2
ratings_1
Ratings (in Stars)
0.0
0.5
1.0
1.5
2.0
Number of Reviews (in Millions)
1e8


Book Title Length
Frequency
3000
2500
2000
1500
1000
500
0
0
25
50
75
100
125
150
175
Title Length


Book Title Word Count
Frequency
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
Title Word Count


Top 25 Words Most Commonly Found in Title
king
harry
magic
saga
secret
dead
mpire
tories
black
little
story
time
book
world
girl
rilogy
house
man
love
vol
dark
nicles
life
night
death
0
10
20
30
40
50
Frequency


Top 25 Words Most Commonly Tri-grams in Title
Common Tri-grams Found in Title
anne green gables
grave night huntress
death gate cycle
warriors new prophecy
ladies detective agency
dune dune chronicles
inspector armand gamache
chief inspector armand
midnight midnight breed
percy jackson olympians
women murder club
malazan book fallen
iron druid chronicles
plantagenet tudor novels
series unfortunate events
saga vol saga
fruits basket vol
black dagger brotherhood
jack ryan universe
pretty little liars
diary wimpy kid
harry bosch universe
prey lucas davenport
blake vampire hunter
anita blake vampire
0
2
4
6
8
Frequency

04 Recommender

Cleaning Process

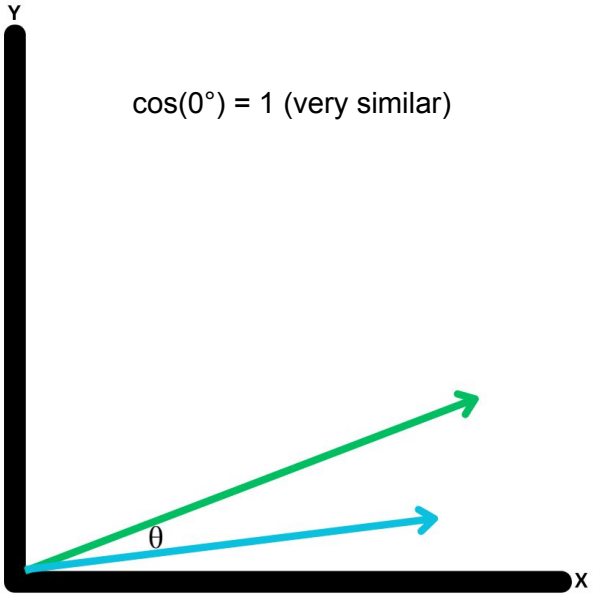
- 6 Million reviews
- Started with 50,000 reviews and increased the amount of data points
  - Checked to ensure my favorites were included
- Features:
  - user_id
  - rating
  - title
- Final system included 5,419,126 reviews of en-US books

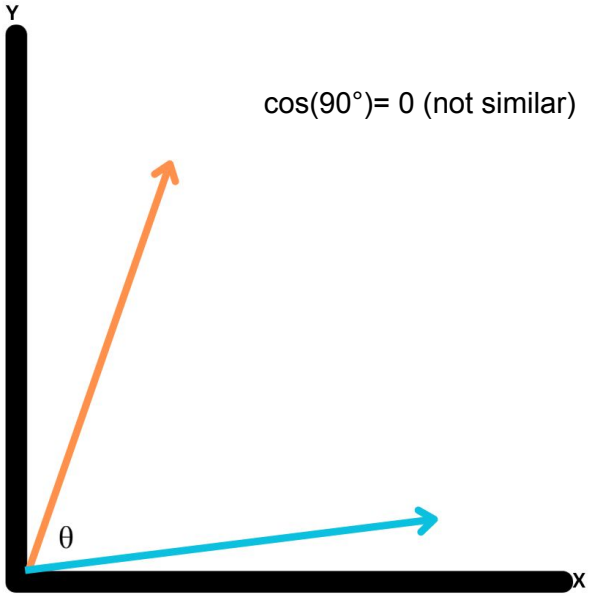
04 Recommender

- Item-based Collaborative Recommender
  - You liked Book A → Book A was rated by users similarly to Book B → You will like Book B
- Explicit Rating System
  - Interval-based star rating system
    - ⭐ - ⭐⭐⭐⭐⭐
- Pivot Table needed
  - The book title will be the index of the data frame
  - The user_id will be the columns of the data frame
  - The rating will be the values within the data frame

04 Recommender

Cosine Similarity





04 Recommendations

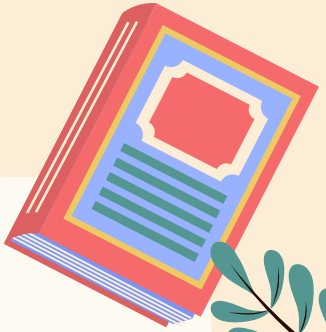
```
title
Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
Harry Potter and the Chamber of Secrets (Harry Potter, #2)
Harry Potter and the Goblet of Fire (Harry Potter, #4)
Harry Potter and the Order of the Phoenix (Harry Potter, #5)
Harry Potter and the Half-Blood Prince (Harry Potter, #6)
Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)
Name: Harry Potter and the Sorcerer's Stone (Harry Potter, #1
```

05 Classification Modeling

Cleaning Process

- Originally built multi-class prediction model
  - Very unbalanced classes
    - 95% of the data was 4 stars
- Pivot towards binary-classification model
  - Average rating >= 4 stars → 1
    - indicating that this book was worthy of breaking the book hangover
  - Average rating < 4 stars → 0
    - indicating that a reader should still read this book, but it is not going to be their next great read to pull them out of their book hangover

05 Classification Modeling



Cleaning Process

- Binary Classification Models:
  - NLP (Natural Language Processing)
    - Language code: As mentioned previously, books with language codes determined not to be US english were dropped
    - Features:
      - authors
      - title
      - should_i_read
  - Non-NLP
    - Language code: As mentioned previously, books with language codes determined not to be US english were dropped
    - Features:
      - original_publication_year
      - ratings_count
      - ratings_1_, _ratings_2_, _ratings_3_, _ratings_4_, _ratings_5
      - should_i_read_.

05 Classification Modeling

NLP Modeling

68.9% Accuracy

- TF-IDF Vectorizer
  - n_gram range: (1,2)
  - Stop words: “english”
- Logistic Regression
  - L2 penalty

Feature example:
Suzanne Collins : The Hunger Games

Non-NLP Modeling

97.4 % Accuracy

- Logistic Regression

05 Classification Modeling

Predictions: Testing Data

NLP Modeling

Actual	Predicted	Total
0	1	348
1	0	263

Non-NLP Modeling

Actual	Predicted	Total
0	1	30
1	0	22

1,963 data points in the testing dataset

05 Classification Modeling

Predictions: My TBR List

NLP Modeling

Actual	Predicted	Total
0	1	6
1	0	5

Non-NLP Modeling

Actual	Predicted	Total
0	1	2
1	0	0

28 data points in the testing dataset

05 What Do Incorrect Predictions Mean?

- **Scenario 1:** A book is predicted that I should not read it immediately (an average rating of less than 4 stars)
  - In this scenario, I could potentially opt not to read a book that could end my book hangover

- **Scenario 2:** A book is predicted that I should read it immediately (an average rating of higher than 4 stars)
  - In this scenario, I could potentially read a book that I dislike and hope that it is enough to get me through the book hangover

06 Final Thoughts

- NLP model is most practical
  - Streamlit App
- Improvements:
  - Additional Features
    - Genre
    - Book length
    - Price
    - Format (harback, paperback, e-book)
    - Written reviews
    - Book Cover Metadata

Thank You!!

A extra special thank you to:

- Tim & Rowan
- Everyone in the chatty breakout rooms