# SENTIMENT ANALYSIS FOR WESTERN CLASSICS AND ITS CHARACTERISTICS

Alaina Rongione

# Introduction

- **Goal**: Exploratory analysis of how readers react to Western classics to gain better understanding of the dataset and how to approach the research question
  - How do certain characteristics of a classic novels influence public sentiment, and how can these factors predict if modern novels are likely to gain the same popularity and reception?
  - Author, length of novel, top 3 genres, ratings, year

# DATA RETRIEVAL

| | Titles | Authors | Years | Pages | Rating | Genre 1 | Genre 2 | Genre 3 |
|---|---|---|---|---|---|---|---|---|
| 0 | Three Men in a Boat | Jerome K. Jerome | First published January 1, 1889 | 185 pages, Paperback | 3.83 | Classics | Fiction | Humor |
| 1 | On the Road | Jack Kerouac | First published September 5, 1957 | 307 pages, Paperback | 3.61 | Classics | Fiction | Travel |
| 2 | Brideshead Revisited | Evelyn Waugh | First published January 1, 1945 | 351 pages, Paperback | 4.00 | Classics | Fiction | Historical Fiction |
| 3 | The Poisonwood Bible | Barbara Kingsolver | First published September 24, 1998 | 546 pages, Paperback | 4.11 | Fiction | Historical Fiction | Africa |
| 4 | The Woman in White | Wilkie Collins | First published November 26, 1859 | 672 pages, Paperback | 4.01 | Classics | Mystery | Fiction |

# Initial Research and Web Scraping

- Scraped Goodreads: list of the top 200 classic novels

- Received the **title** of the novel, its **author**, the **year** it was released, its **length**, its **rating**, and the **top three themes**

- Part of the data frame can be seen above

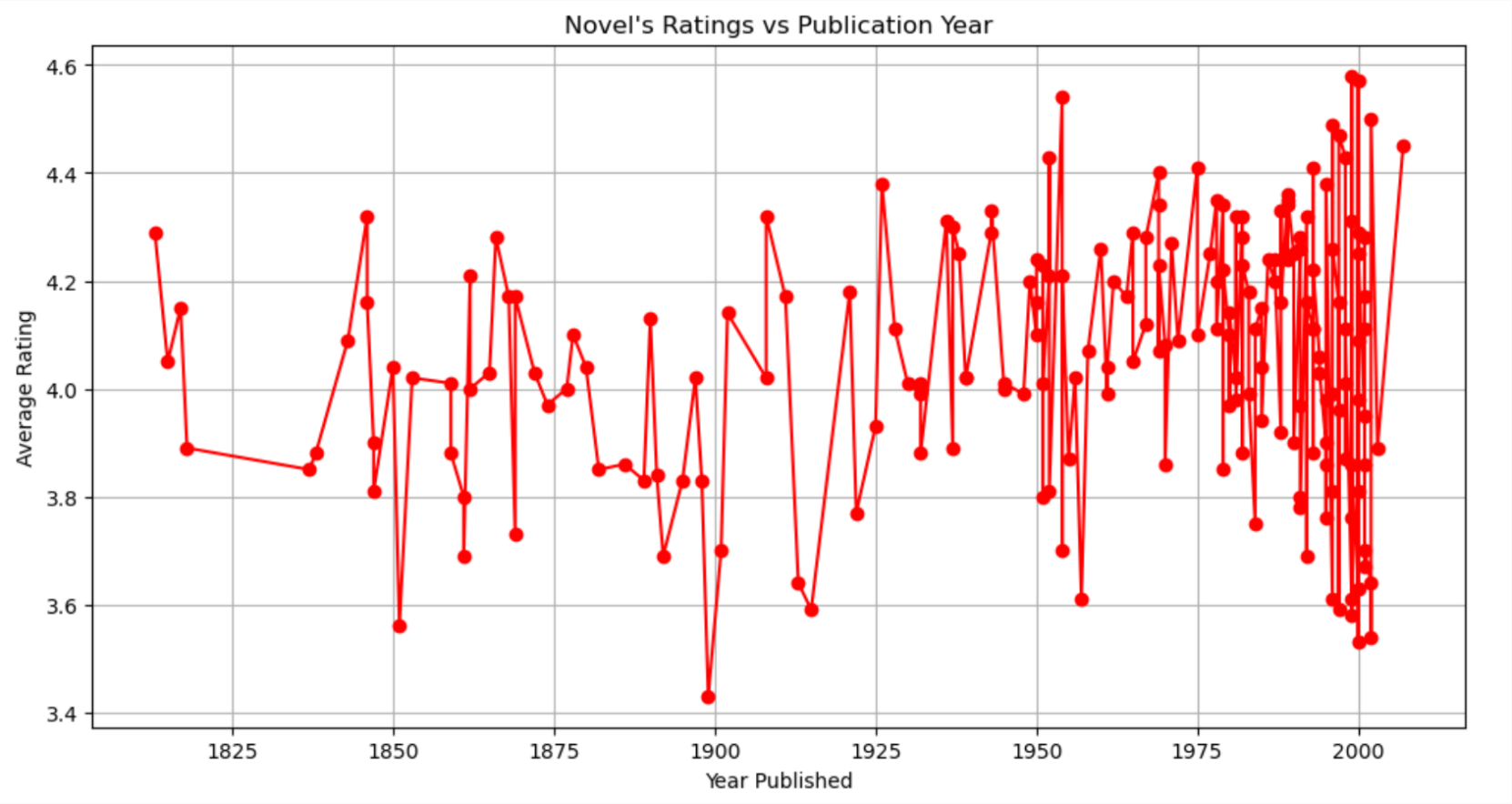
# Cleaning the Data Frame

| | Titles | Authors | Years | Pages | Rating | Genre 1 | Genre 2 | Genre 3 |
|---|---|---|---|---|---|---|---|---|
| 0 | Pride and Prejudice | Jane Austen | 1813 | 279 | 4.29 | Classics | Romance | Fiction |
| 1 | To Kill a Mockingbird | Harper Lee | 1960 | 323 | 4.26 | Classics | Fiction | Historical Fiction |
| 2 | 1984 | George Orwell | 1949 | 328 | 4.20 | Classics | Fiction | Science Fiction |
| 3 | The Hobbit, or There and Back Again | J.R.R. Tolkien | 1937 | 366 | 4.30 | Fantasy | Classics | Fiction |
| 4 | Harry Potter and the Sorcerer's Stone | J.K. Rowling | 1997 | 309 | 4.47 | Fantasy | Fiction | Young Adult |

- Eliminated the extra text in the Pages column and the Years column
- Converted the numbers in the Years, Pages, and Rating column from strings to integers

# EXPLORATORY DATA ANALYSIS (EDA)

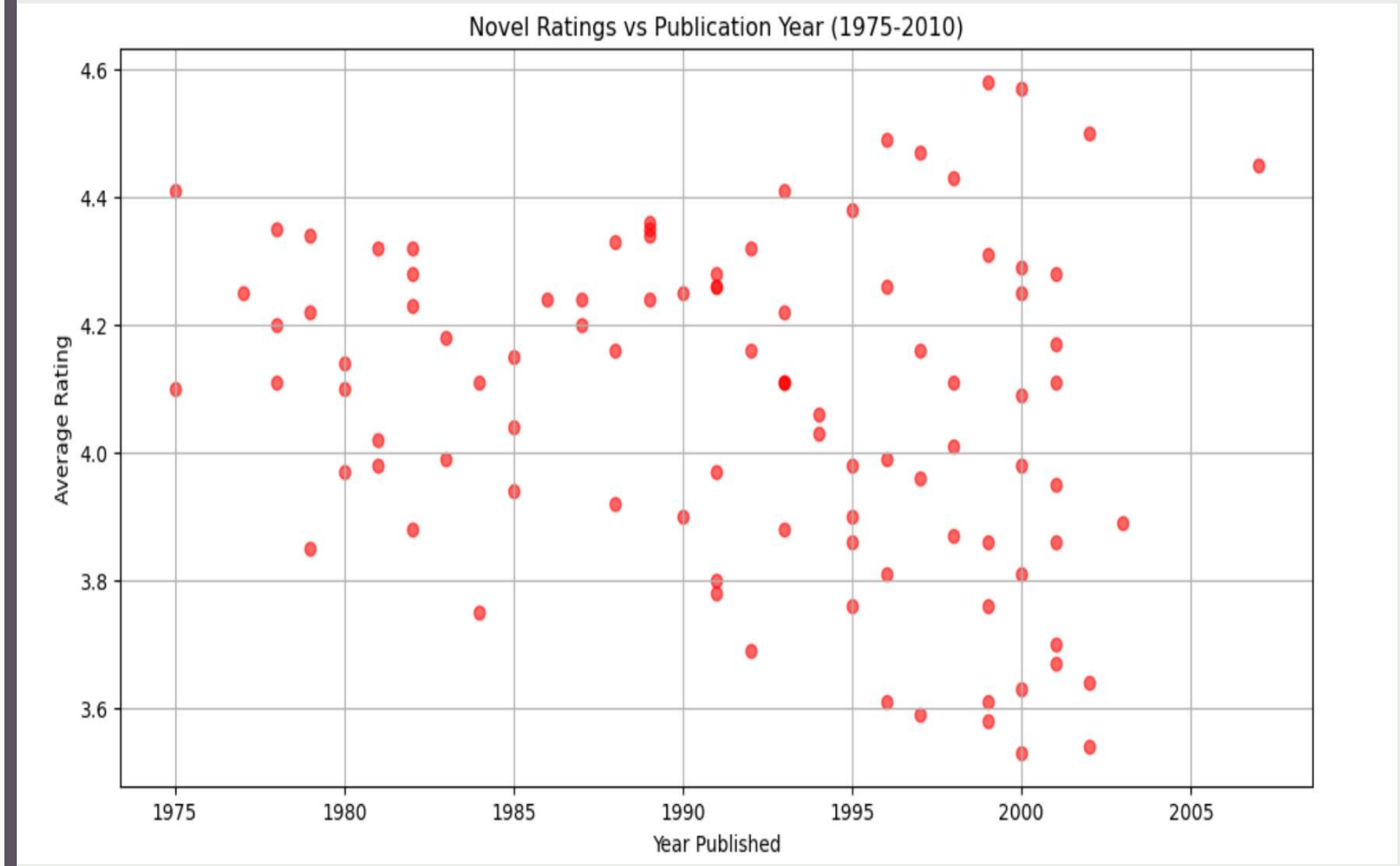
# Movie Reviews (Timeline)



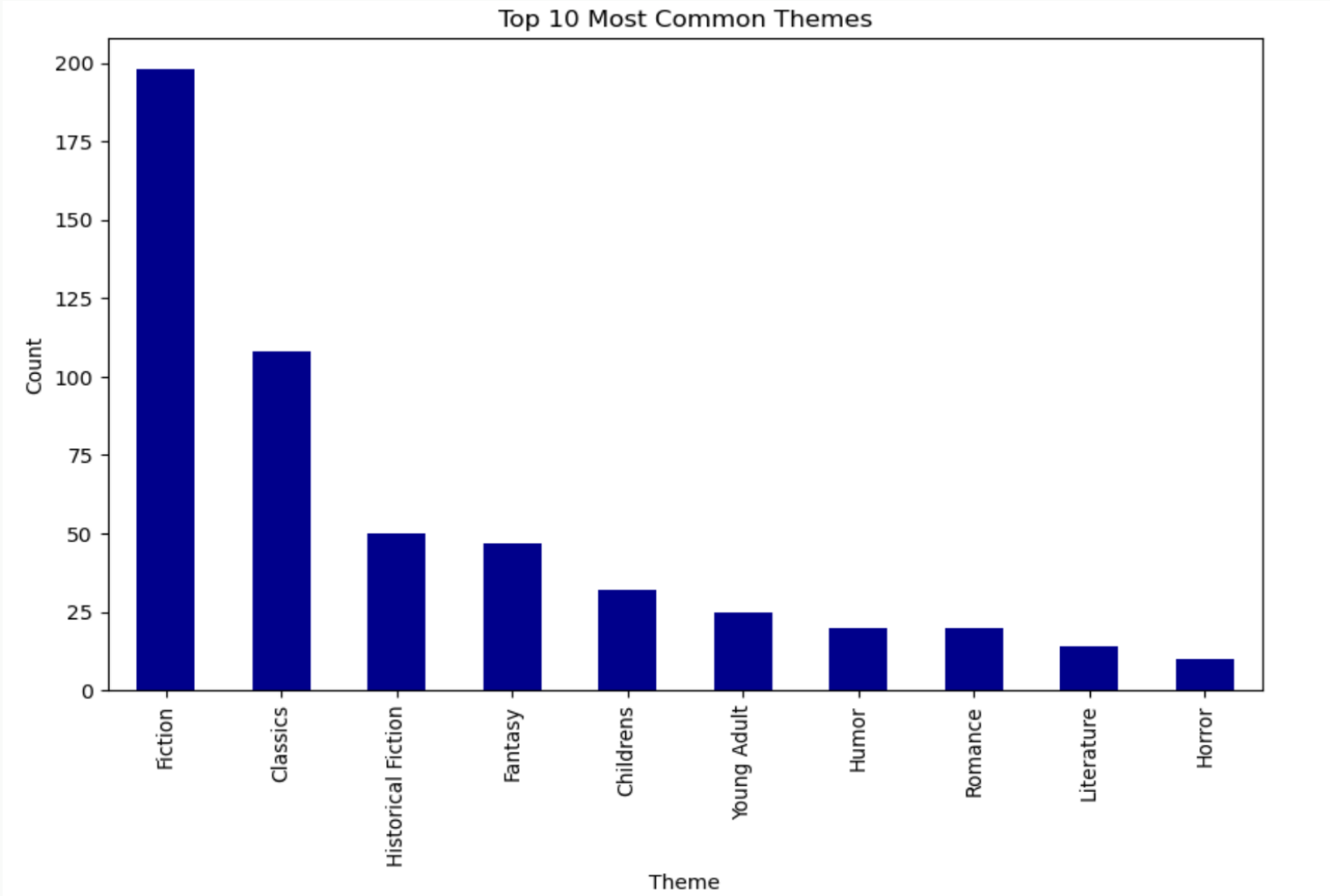

- Static
- No consistent trend with the data
- Possible that Year Published is a weak indicator of reader sentiment
- Other may have a stronger pull on the rating

# Movie Reviews (Timeline) Cont.

- Reviews expand and become more various
- Oscillating between 3.8 and 4.4 in the 1980s → Oscillating between 3.6 and 4.6 in the late 1990s



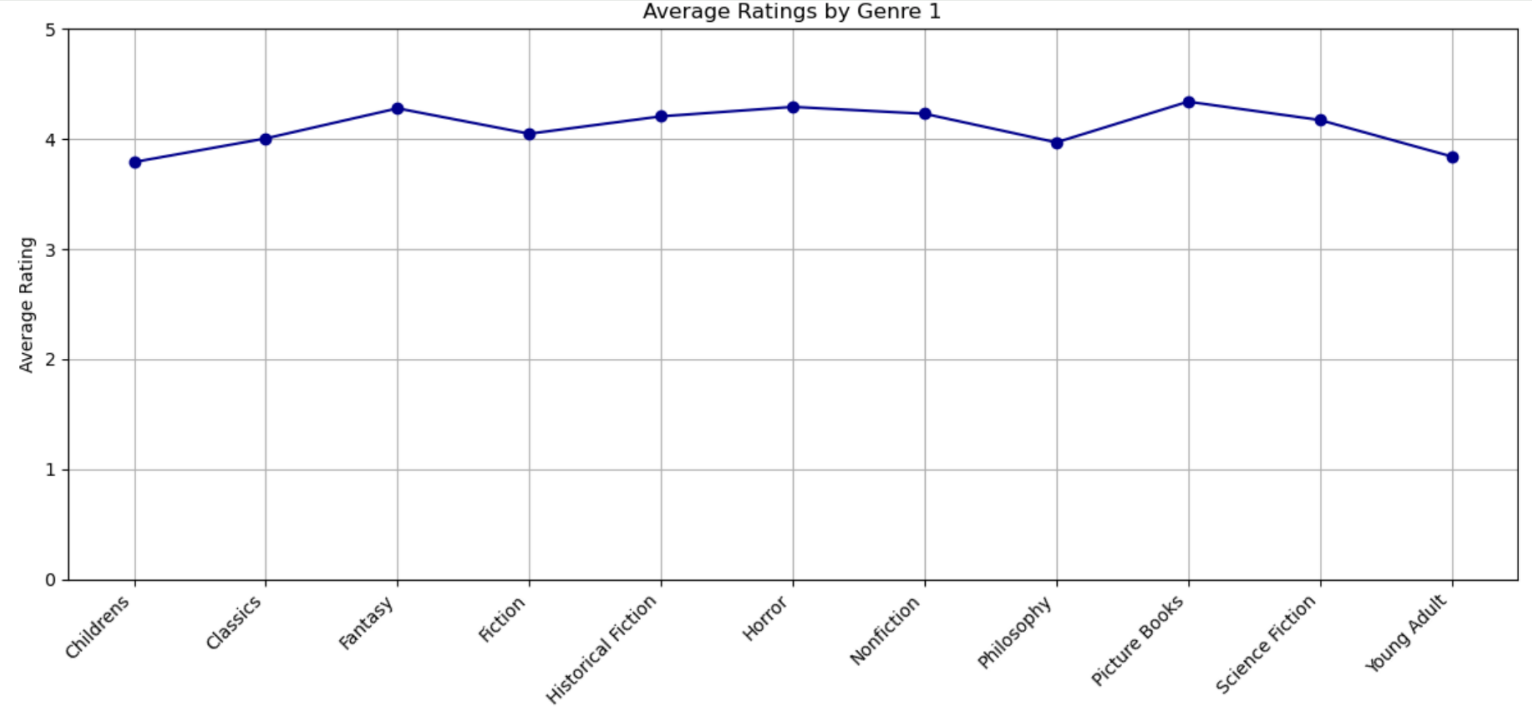
# TOP THEMES

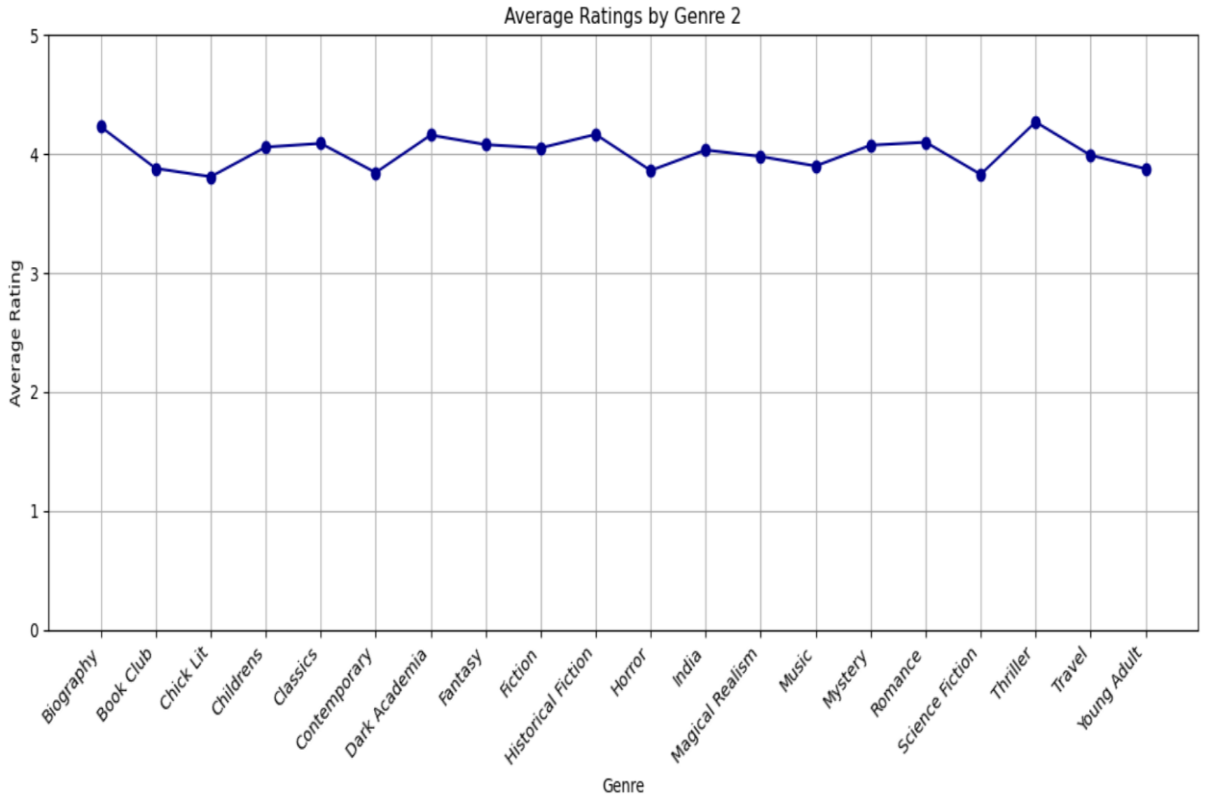


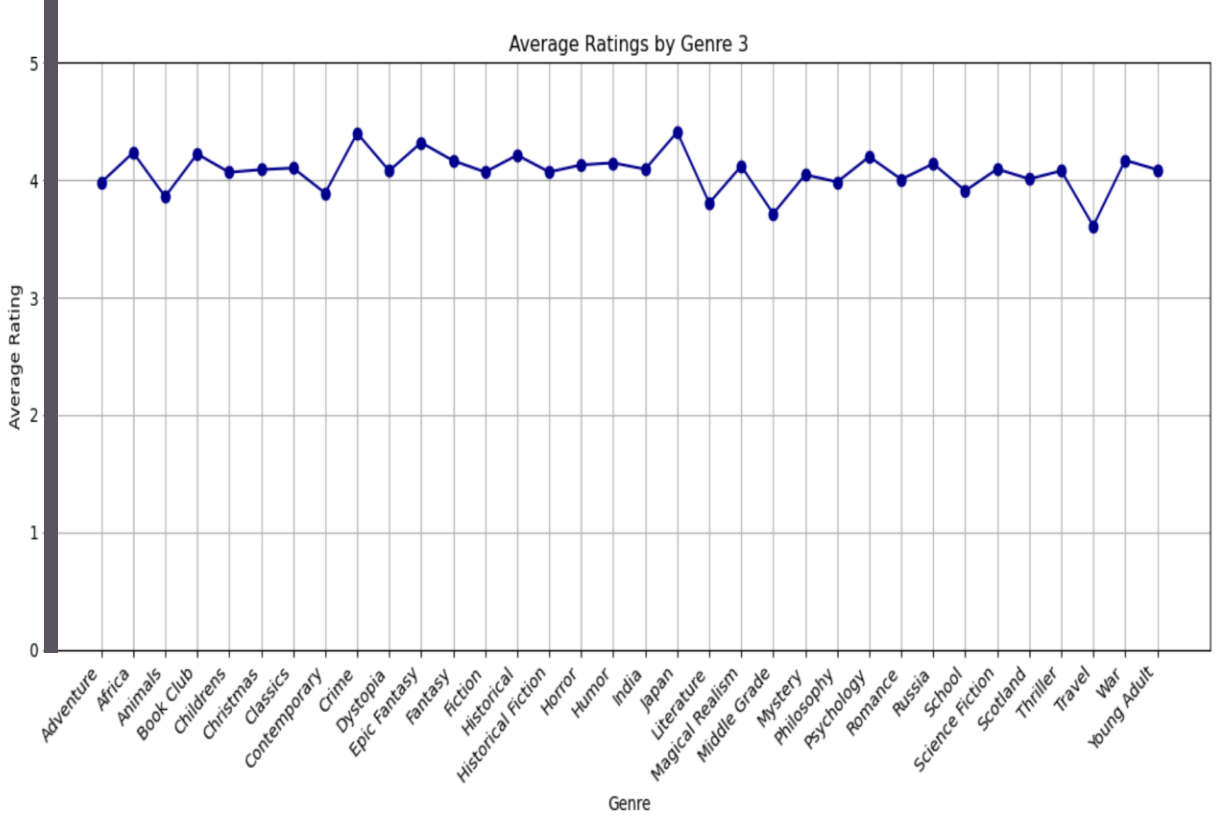

- Top 10 themes among the novels are to the left
- Will need to collect more themes for each novel rather than just three

# Themes vs Average Rating

- Overall, the initial genres the novels are described as are rated highly across the board
- There is some notable differences→ the novels initially rated as Classics, Childrens, and Philosopher receive lower reviews than those initially rated as Fantasy, Horror, and Picture Books





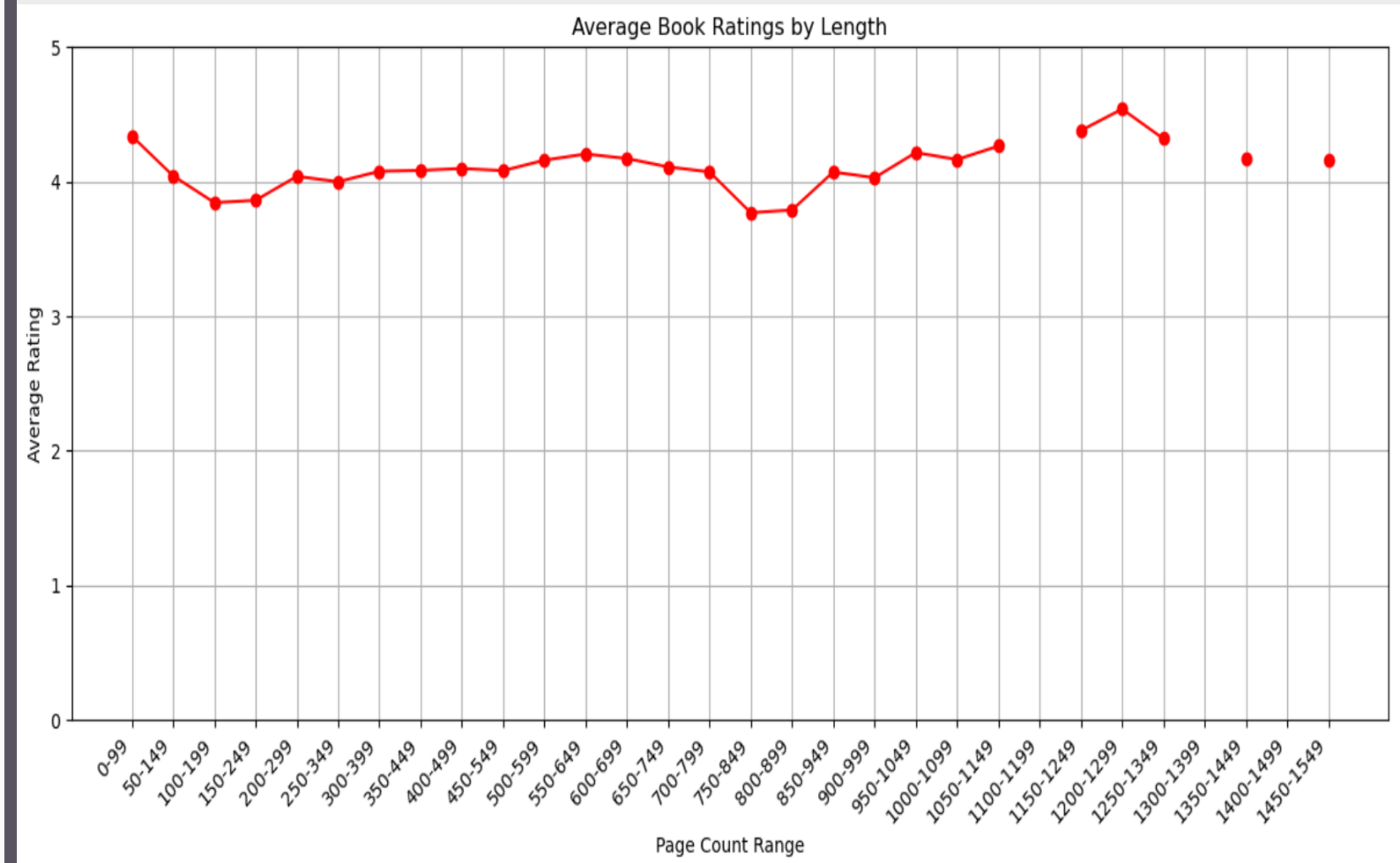

# Theme vs Average Ratings Cont.

# Authors and Themes


- Authors stay consistent with genres they incorporate into the novels
- Certain genres are dominated by specific authors calling into question the independence of themes regarding authors

# Novel Length vs Ratings

- Like the themes and ratings, the ratings are consistent across the graph
- We see a quick drop for the novels between 50 and 300 pages, and then a spike around the early 1200s



# IMPLICATIONS

# Stakeholders

- As a result of the EDA, authors can target themes that are less popular, but are rated highly amongst readers
- Readers receive more original content in thematic areas that have historically been written about less
    - A result from authors targeting less popular, but appreciated genres
- Cultural shift in modern literature and possible evolution in literary appreciation
- Literary scholars and studies will possibly be affected

# Societal, Legal, and Ethical

- With only focusing on a limited novels for “classic” literature, calls into question what novels that are being excluded
- Careful with the reviews of readers and avoid skewing their beliefs due to unintentional, model bias → generalization error
- Shed light on how historically appreciated genres are viewed, currently → ongoing cultural appreciation

# NEXT STEPS

# Further Analysis (Steps Towards Prediction)

- Model → **Logistic Regression** to discover what factors predict a novels success with the public
- Refine the model through Odds Ratio and cross-validation
- Expand my dataset (novels and genres)
- Scrape the written reviews of the novels and conduct a word cloud to visualize what is most talked about in each novel's review