Text Generation for Creative Writing

By Pallavi Sikha

# Data Overview

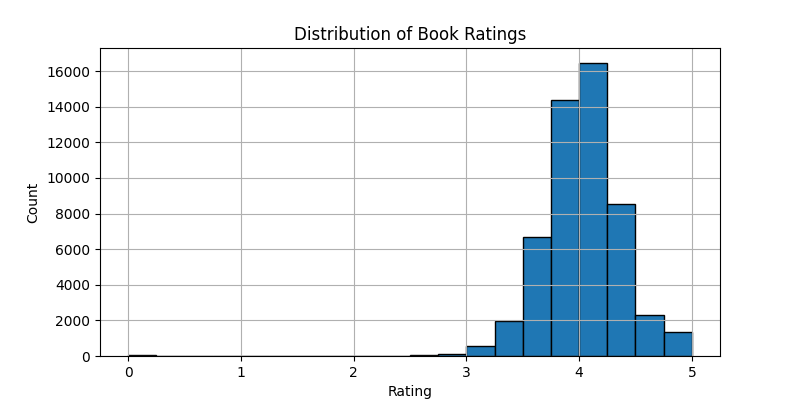
In this project, we explore a comprehensive dataset consisting of various literary works. Our dataset includes information on thousands of books, such as the title, author, rating, genres, publication date, and more. The objective of this project is to leverage the dataset for text generation in the realm of creative writing. By analyzing this dataset, we aim to understand the distribution of various features, including the diversity of genres, the distribution of ratings, and the correlation between different attributes. This analysis will serve as the foundation for developing a generative model that can assist authors in the creative process, providing suggestions for plots, characters, and literary styles.  
  
Our dataset comprises a rich collection of literary works from various genres and time periods, making it ideal for training a model that can generate diverse and coherent text. The dataset includes titles from classic literature, contemporary fiction, and niche genres, offering a wide range of content to draw upon for text generation.

Total books: 52478

Unique authors: 28227

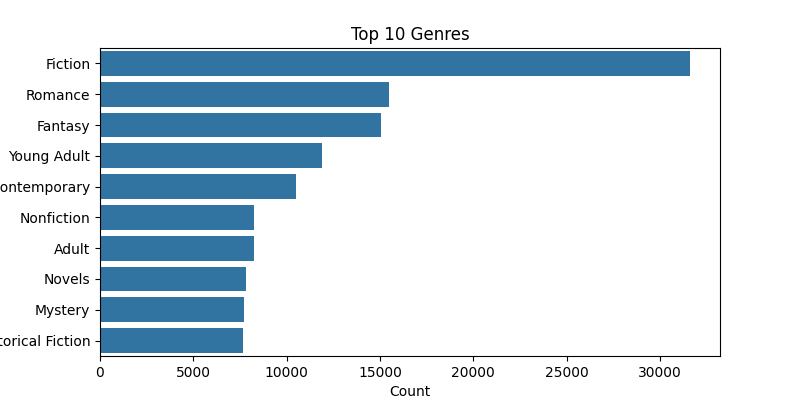
Average rating: 4.02

# Rating Distribution



The histogram above displays the distribution of book ratings within our dataset. Ratings are an essential aspect of any literary analysis, as they provide insight into how books are received by readers. In this project, understanding the distribution of ratings allows us to gauge the quality and popularity of the books in our dataset. The distribution also helps in identifying any biases or patterns in how books are rated. For instance, a large number of books receiving high ratings could indicate that the dataset contains predominantly well-received works, which might influence the text generation model to favor certain styles or genres.  
  
In this analysis, we observe that the ratings are concentrated around the 4 to 5 mark, suggesting a general trend of positive reception among the books in our dataset. This trend is beneficial for text generation, as it allows the model to learn from well-regarded examples, potentially improving the quality of the generated content.

# Top 10 Genres



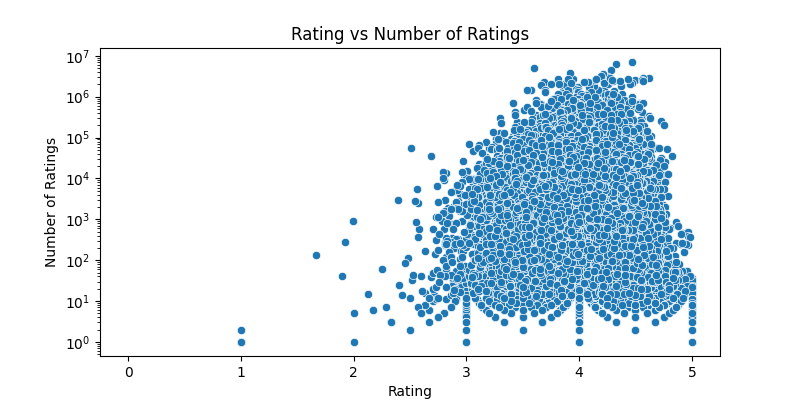
The bar chart above shows the top 10 genres in our dataset based on the frequency of their occurrence. Genres play a critical role in shaping the content and style of literary works, and understanding their distribution is vital for text generation. By identifying the most common genres in our dataset, we can tailor our generative model to focus on the predominant themes, ensuring that the generated text aligns with popular literary trends.  
  
The most common genres in our dataset include Fantasy, Young Adult, and Fiction, which are popular categories in contemporary literature. This distribution suggests that our model will likely generate content that appeals to readers who enjoy these genres. However, it is also important to consider the less frequent genres, as they offer opportunities for generating unique and niche content that can stand out in the literary market.

# Word Cloud of Book Descriptions



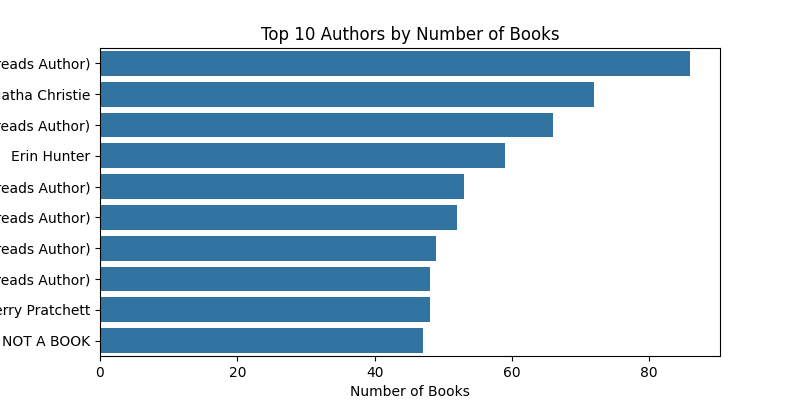
The word cloud above highlights the most frequent words found in the book descriptions of our dataset. Word clouds are a useful visualization tool for identifying prominent themes and concepts within a text corpus. In the context of text generation, understanding the frequency of certain words helps in guiding the model to focus on relevant topics and themes.  
  
The most frequent words in our dataset include common literary terms and genre-specific vocabulary, reflecting the diverse nature of the books in our dataset. By leveraging this information, the generative model can produce text that is both coherent and contextually appropriate for the genres represented in the dataset.

# Rating vs Number of Ratings



The scatter plot above shows the relationship between book ratings and the number of ratings they have received. This analysis is important for understanding how popular and well-rated books are distributed within the dataset. Books that receive a high number of ratings are often more influential, and their content may heavily influence the text generation model.  
  
Interestingly, the plot reveals that higher-rated books tend to have more ratings, suggesting a positive correlation between popularity and perceived quality. For text generation, this correlation implies that the model can benefit from learning from highly rated and popular books, which could enhance the appeal and readability of the generated content.

# Top 10 Authors by Number of Books



The bar chart above lists the top 10 authors in our dataset based on the number of books they have