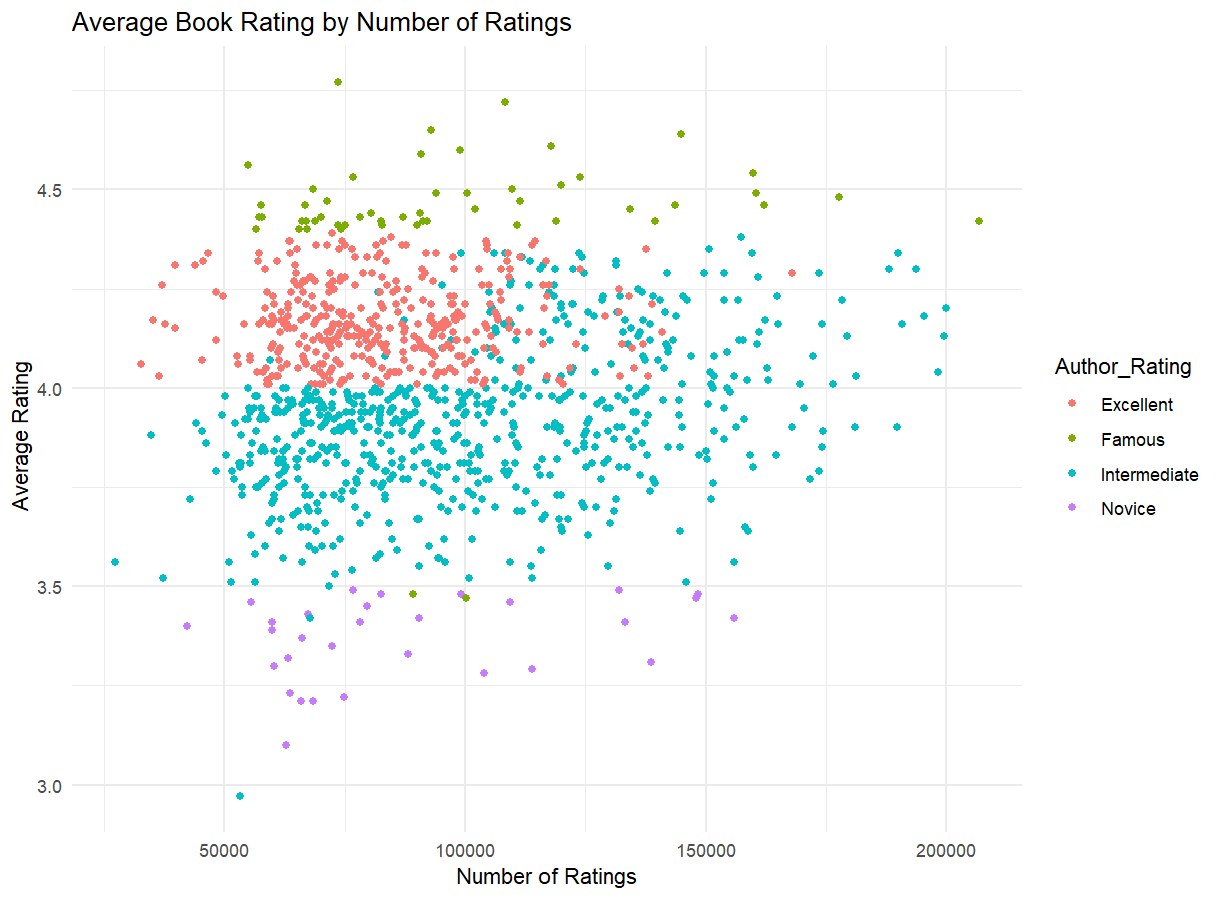
Anthony Appiani

ggplot Project

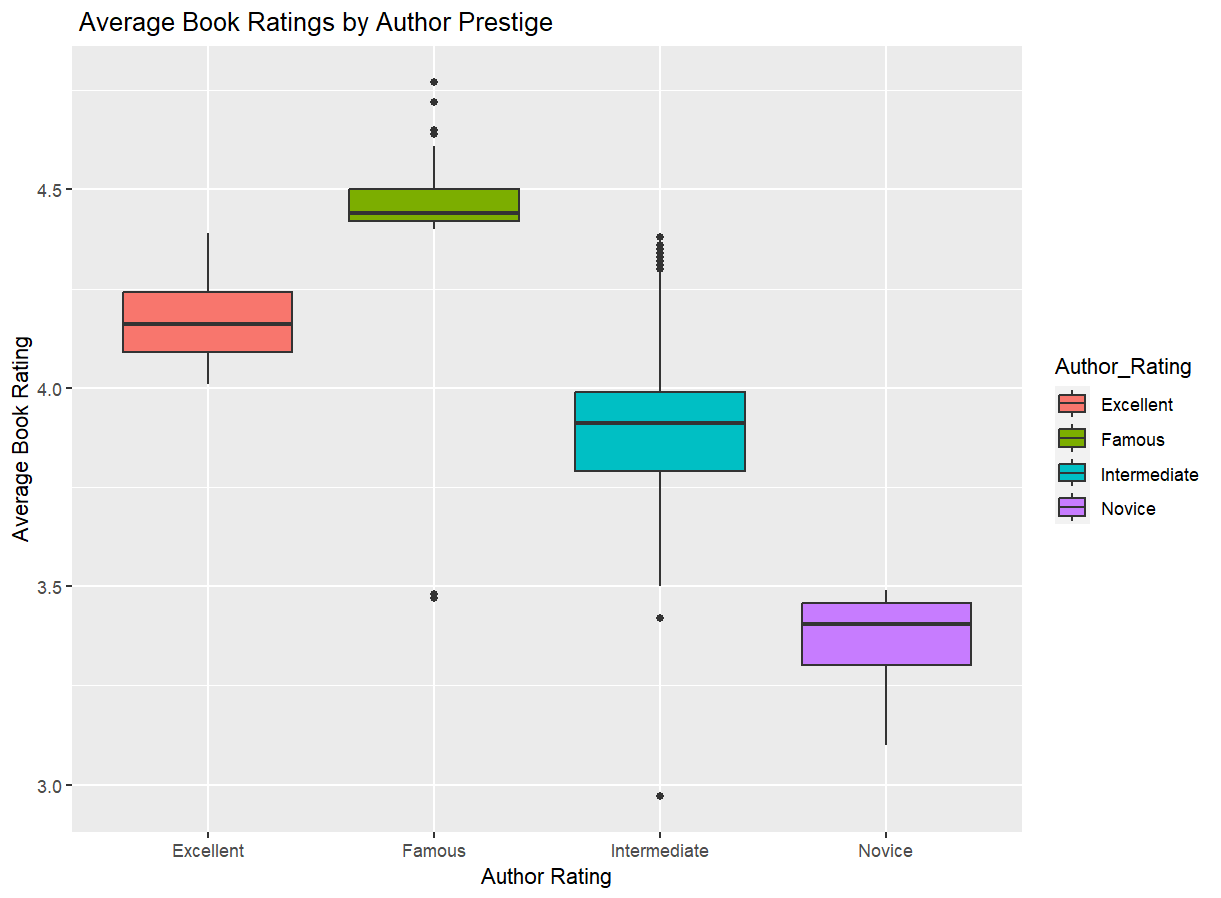
data: [Books Dataset](https://www.kaggle.com/datasets/thedevastator/books-sales-and-ratings/data)

This is a dataset about book sales. The data appears to be from a year or period of book sales the information about the data did not confirm what time frame this data was collected from. It includes the book information, the book author, sales information, and publisher information. The biggest problem with the data was in the genre column. Fiction was entered in two ways, ‘fiction’ and ‘genre : fiction’. I updated this so fiction was the only option. The data did not break down the genres in subcategories, it was only fiction, non-fiction and children. The data did have some N/As in the book title or author and some of the book names had special characters. These did not have much of an affect how the visualizations were created although one visual does have a book name with special characters.

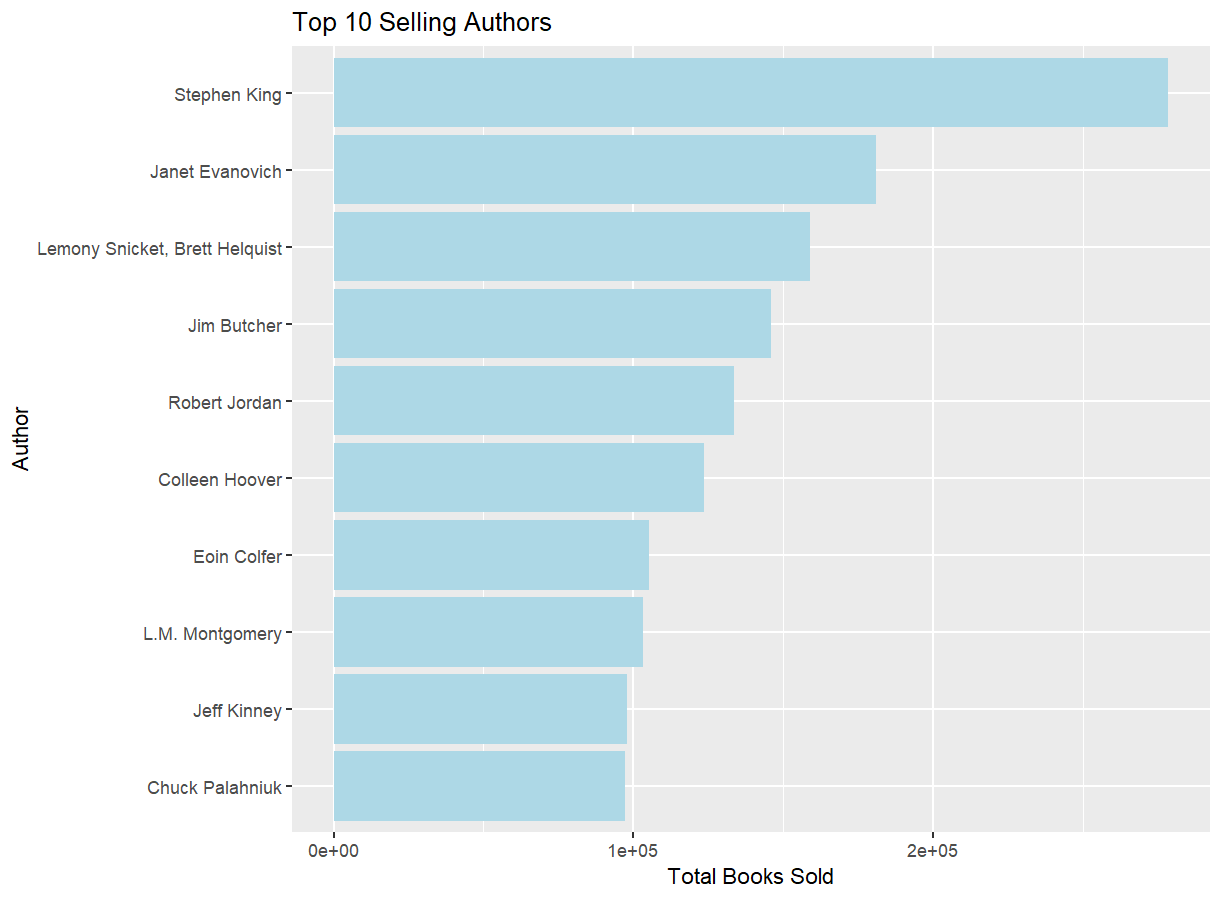
First, I chose a scatter plot to look at the relationship between the number of ratings and the average rating of a book. I also colored it by author prestige. Most of the books have 50-100k ratings and a rating between 3.75 and 4.25. Most of the points are as expected with the famous and excellent authors having higher ratings. The interesting group is the intermediate author group. They have a wide variety of points on the graph. They have the lowest rated book and the book with the lowest number of ratings. They also have some of the most rated books in the dataset. The bands are clear between author prestige and the Famous and intermediate authors get significantly higher average ratings with the same number of ratings.



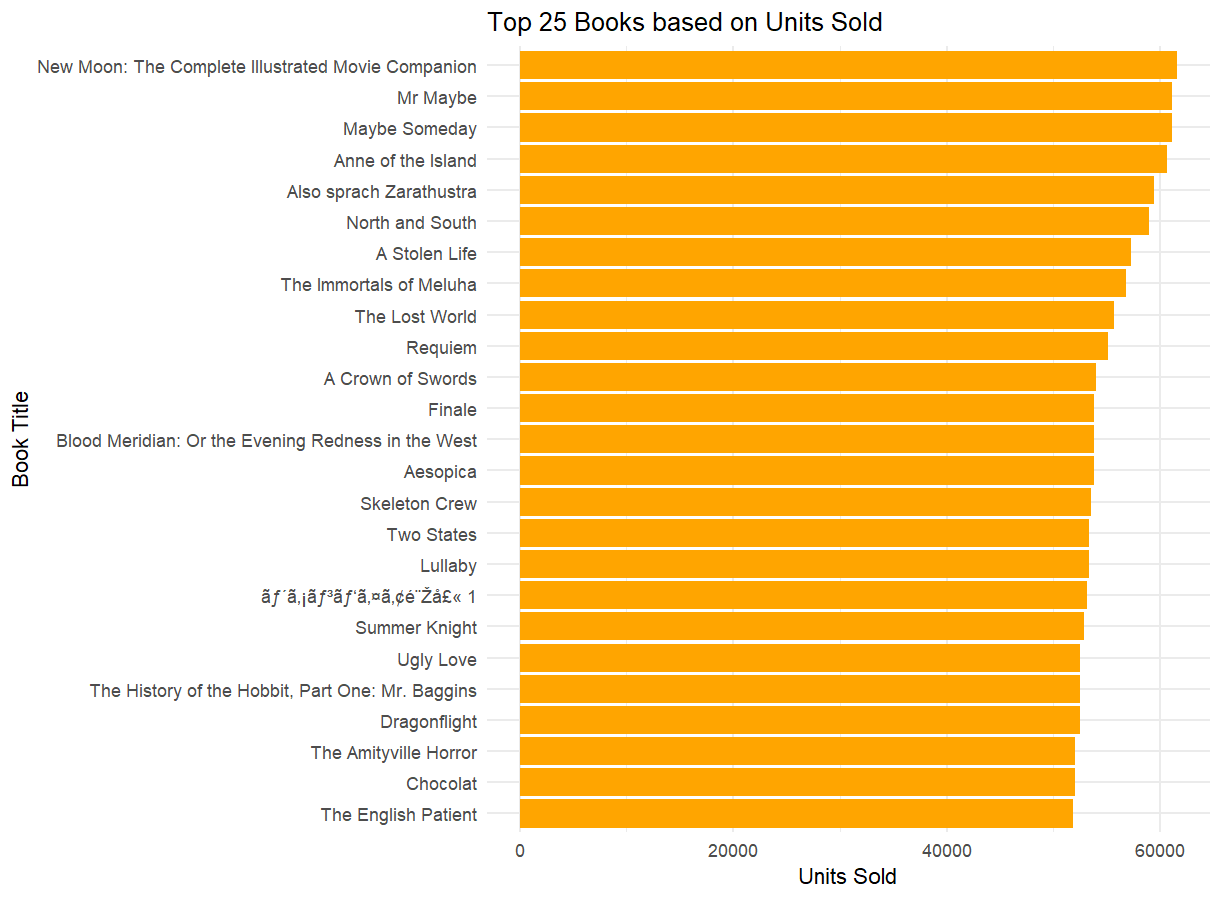
I also created a boxplot to show the average ratings for each level of author prestige. You can see similar results to the scatter plot. The Famous authors have a higher average rating with less variance. The Intermediate group has a wider box and longer whiskers, supporting the earlier idea of their wide array in ratings.



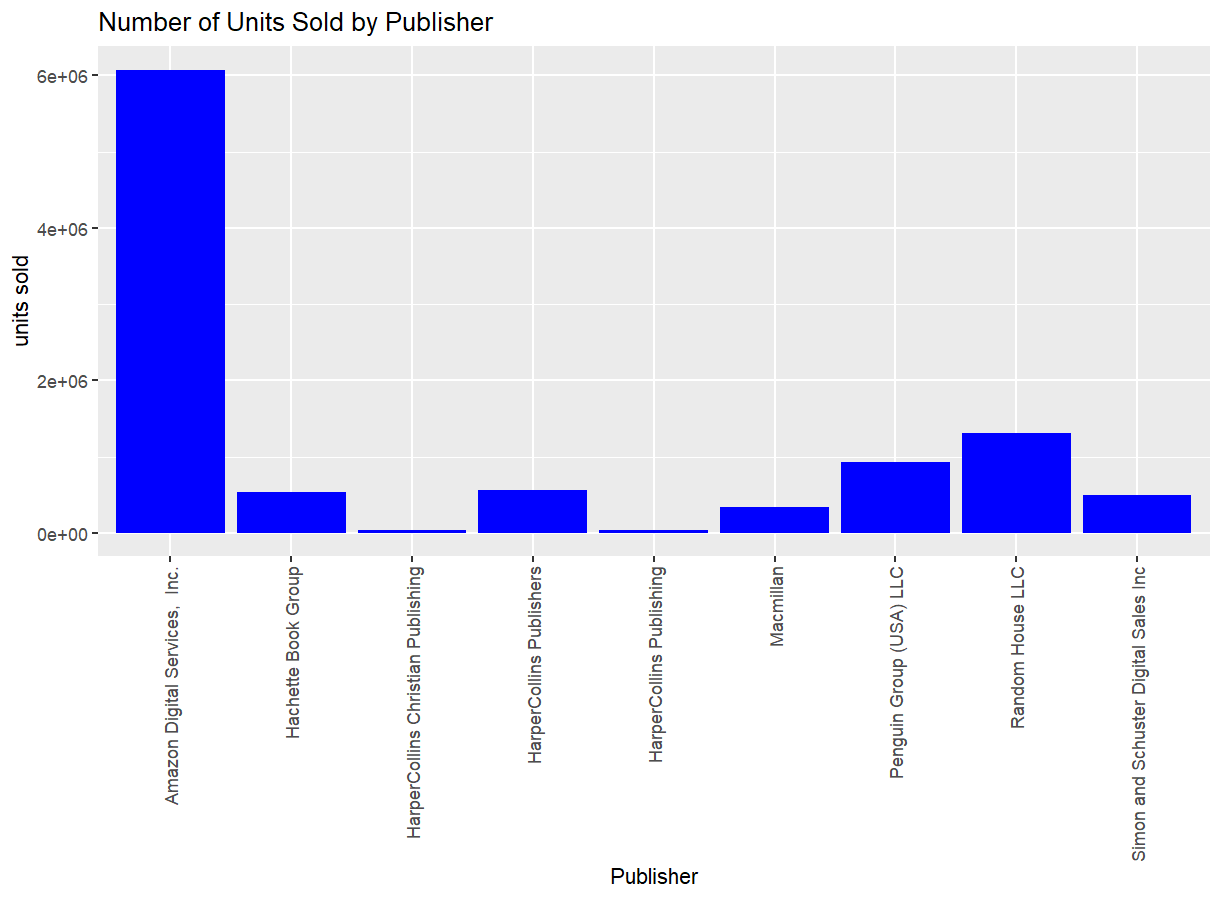
Next, I created two bar charts. The first one showing the top 10 selling authors by units sold. I am not very surprised that Stephen King tops the list and is well ahead of second place. I was a little surprised not to see Dr. Suess on here, but I did notice some data points missing and one did appear to be a Dr. Suess book. This also confirmed my suspicions that this was not an all-time list of sold books but potentially a year of books sales.

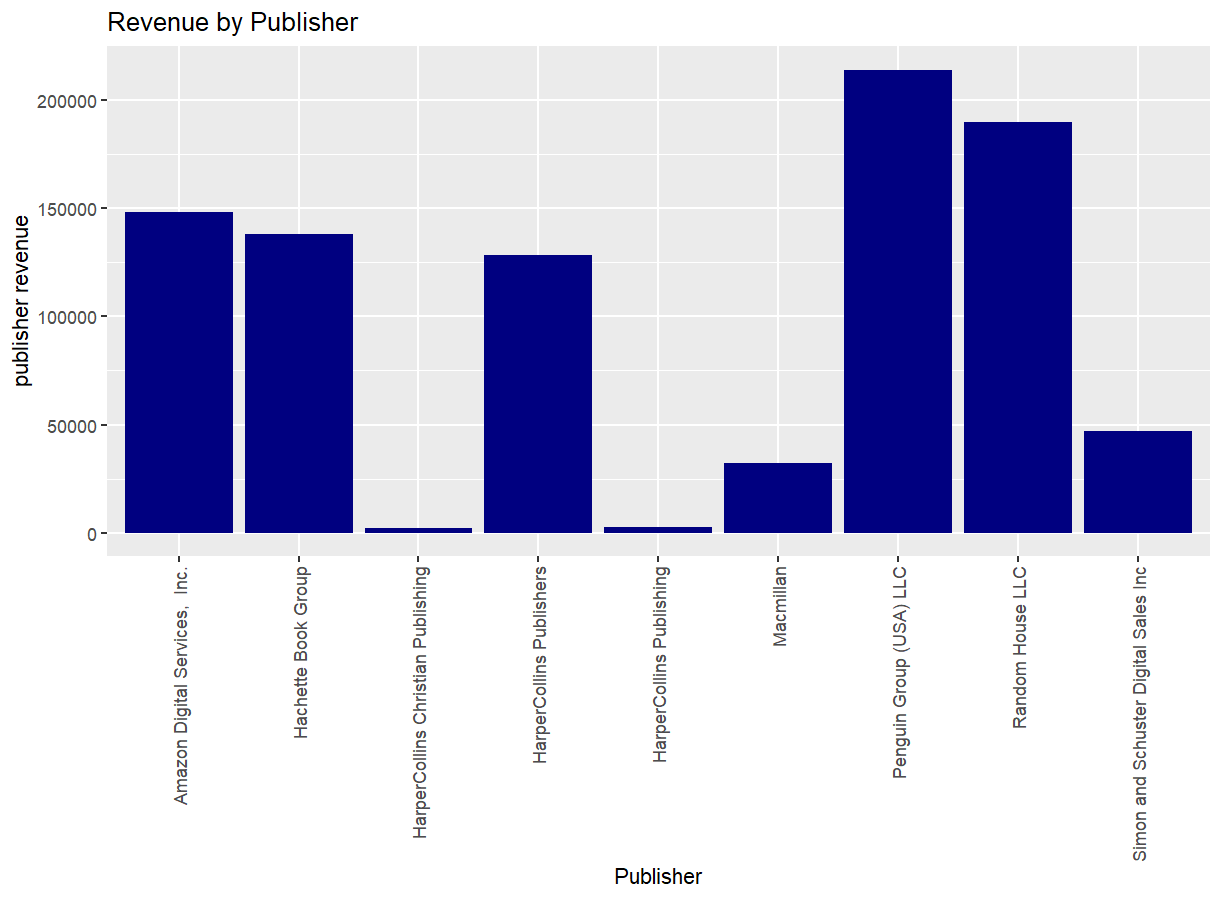


I also looked at the top 25 books sold. The top book was New Moon: The Complete Illustrated Movie Companion. This gave me a further idea that the data set was probably looking at book sales for 2009. This was in the middle of the Twilight saga fever and the movie companion was the best-selling book. This visual does highlight another potential issue with the data. There is a book with special characters on this list. After googling the authors and publishing date I believe the book is Vampire Knight. It was also interesting that Aesopica or Aesop’s Fables was in the top 25. This is the oldest book on here with a publishing data of -560. It was originally oral stories told in Greece they have since been collected and their lessons shared to teach kids valuable lessons.

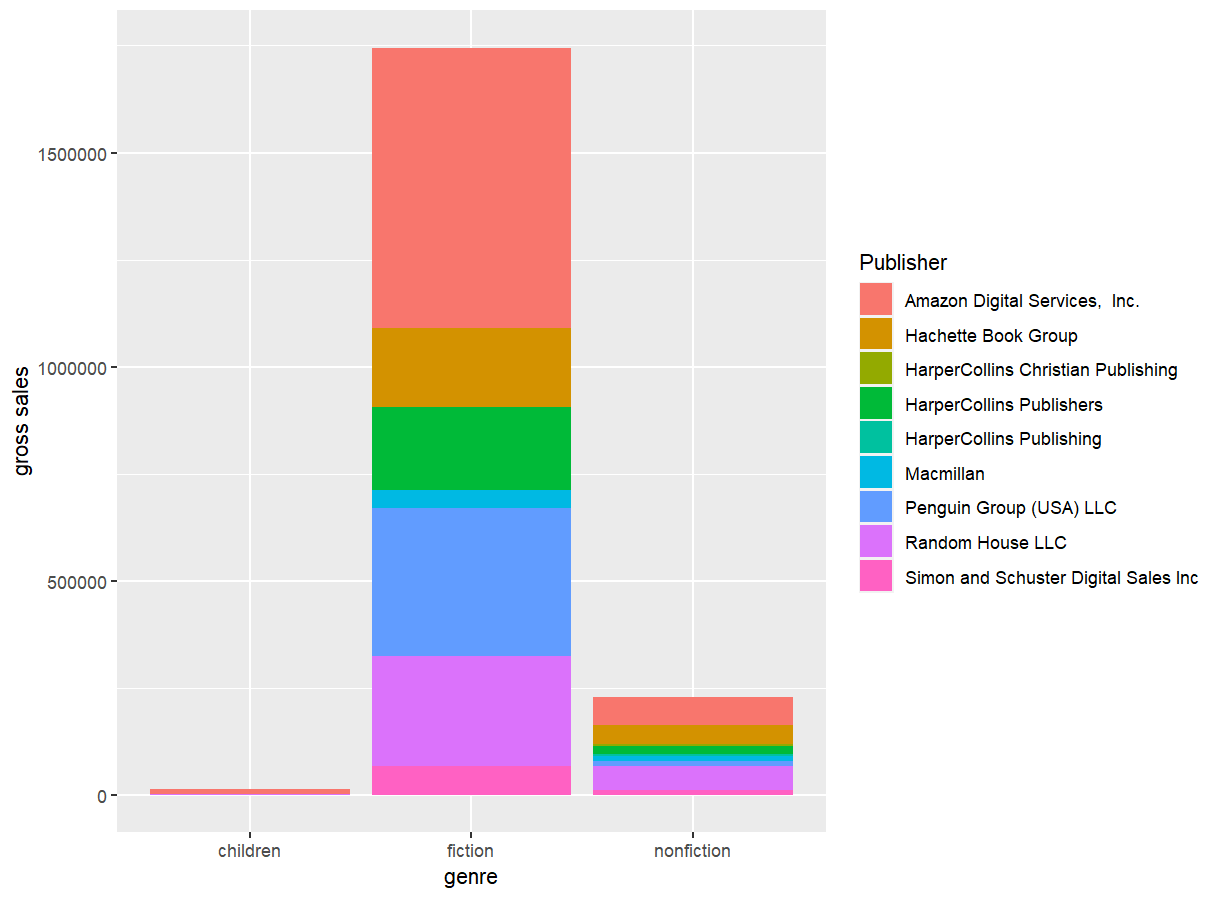


There was info regarding publishers that I wanted to explore. I created bar charts to show total units sold by each publisher and the revenue for each publisher. Although Amazon sold the most units by a wide margin, they did not have the highest revenue. Upon closer inspection I noticed a lot of books on Amazon were $0.99. While these sell quickly there is not a lot of profit to be had on a book that sells for under $1.





Lastly, I created a stacked bar chart showing the gross sales by genre while showing each publisher’s slice of the sales by genre. Fiction is the dominant genre sold by a substantial amount. It is hard to determine the proportions for each publisher in the other genres due to how small the sales in non-fiction and children’s books are compared to fiction.



Code

A screenshot of a computer code

Description automatically generated