Analysis of Amazon Best Sellers

Authors:
David Chien
Eric Luong
Siva Sai Chandra Annepu
Sri Krishna Chaitanya Kommalapati

Date:
March, 12, 2023

Course:
MKTG 2505: Marketing Analytics

Table of Contents

Abstract

This project aims to explore the relationship between book attributes and their popularity, as measured by the number of reviews they receive on Amazon. Amazon started as a book store and expanded into the ecommerce space. From the publishers point of view, this analysis could be useful in determining which categories of books are most successful to minimize the loss from picking up non-performing books/authors. Through exploratory data analysis and machine learning techniques such as natural language processing and k-means clustering, we identify several factors that are positively correlated with book popularity, including lower prices, higher ratings, and certain genres such as fiction. We also use natural language processing techniques to analyze the titles of the books, identifying the words and phrases that are most strongly associated with positive reviews. The results of this project have important implications for both authors and publishers, as they provide insights into the factors that drive book popularity and can inform decisions on book pricing, marketing, and promotion strategies.

About the Data

The dataset used in this project consists of information on the top 50 bestselling books on Amazon as of 2019. It includes 7 columns: Name, Author, User Rating, Reviews, Price, Year, and Genre. The dataset used includes information on over 351 unique books, including their titles, authors, genres, publication year, price, ratings, and reviews. The data has no null values and no duplicated rows. The Name column contains the title of each book, while the Author column provides the name of the author(s). The User Rating column contains the average rating given to the book by Amazon users on a scale of 1 to 5. The Reviews column lists the number of reviews for each book. The Price column shows the price of each book in dollars. The Year column indicates the year of publication, and the Genre column specifies the category of each book, such as fiction and non-fiction.

Exploratory Data Analysis

Chart 1 & 2: Histograms of User Ratings and Book Prices (p. A-1):

We can infer that the user rating distribution is positively skewed and the distribution of book prices is right-skewed.with the mean rating (approximately 4.6) being slightly lower than the median rating (approximately 4.7). This suggests that most books have a high rating, but there is still a significant proportion of books with lower ratings. Publishers can use this information to identify areas for improvement in their books to increase their appeal to readers. They can also use this information to benchmark their books against others in the market and to set realistic expectations for the ratings of their books.

Chart 3 & 4: User Rating and Reviews by Genre (p. A-2):

We analyzed the user ratings and reviews by genre using a boxplot and pie chart (shown in appendix A), and found that fiction books tend to have slightly higher user ratings than fiction books, with a mean rating of 4.60 compared to 4.65 for fiction. The spread of user ratings is wider for non-fiction books, with a standard deviation of 0.265 compared to 0.192 for non-fiction books. There are more non-fiction books in the dataset (301) than fiction books (240). Based on these insights, the publisher may want to consider focusing on publishing more fiction books as they tend to have higher ratings and number of reviews on average. However, the spread of ratings is wider for non-fiction books, which suggests that there may be a larger market for these types of books as well.

Chart 5 & 6: Average Reviews Over the Years by Genre (p. A-3):

The line graph of the mean number of reviews per year by genre shows that both genres experienced an increase in the mean number of reviews from 2014 to 2018. However, the mean number of reviews for fiction books is consistently higher than that for non-fiction books. Based on these insights, the publisher may want to consider focusing on publishing more fiction books as they tend to have more reviews on average. However, it is also important to note that the spread of the number of reviews is wider for fiction books.

Chart 7 & 8: Correlations of Each Feature (p. A-4):

Based on the correlation matrix, there are no strong relationships between the variables. Most correlations are weak or close to zero, indicating that each variable may be influenced by a unique set of factors. However, there are slight positive relationships between user ratings and the year, as well as between the number of reviews and the year. There is also a slight negative relationship between the price of books and the year, suggesting that prices may have slightly decreased over time. We conducted a linear regression analysis to further examine the relationships between the variables. The results showed that the correlations observed in the correlation matrix were statistically significant, but the coefficients were very small, indicating that there is very little correlation between the variables.

Chart 9: Top Authors and their Book Prices (p. A-5):

Based on the top authors with the highest mean number of reviews, it is clear that books written by well-known authors tend to receive higher numbers of reviews. Moreover, it is interesting to note that the average price of their books does not necessarily correlate with the number of reviews.

Machine Learning

NLP Analysis:

We used natural language processing (NLP) and regression analysis to identify the words that are most strongly associated with the number of reviews for books on Amazon. First, we preprocessed the book titles by removing stop words and stemming the remaining words. Next, we created a document-term matrix (DTM) that represented the frequency of each word in each book title. We applied term frequency-inverse document frequency (TF-IDF) weights to the DTM to give more weight to important words in individual documents. We then trained a linear regression model on the bag-of-words features and the number of reviews. The coefficients in the model represent the strength and direction of the relationship between each word and the number of reviews. Finally, we created a bar chart to visualize the top words and the absolute value of their coefficients (p. A-5).

Chart 10: NLP Top 30 Words in Titles (p. A-5)

Our NLP analysis revealed the top 30 words most strongly associated with the number of reviews for books on Amazon. The top words with the highest positive coefficients were "surprise," "act," "difficult," "overwhelm", "terror," "path," "fabulous," and "war," while the top 10 words with the highest negative coefficients were "drive," "brain," "god," "crave," "Twain," and "survive". These results provide insight into the words that are most likely to impact the number of reviews for books on Amazon, which can be useful for authors, publishers, and marketers looking to increase the visibility and success of their books.

K-Means Analysis:

In addition to our NLP analysis, we performed a K-means clustering analysis on three numerical features: user rating, reviews, and price. Before conducting the analysis, we standardized the data and removed any missing values. To determine the optimal number of clusters, we used the elbow method and chose the number of clusters at the point where the decrease in the sum of squared distances began to level off. We identified the optimal number of clusters as 3 (Chart 11, p. A-6). We performed K-means clustering with 3 clusters and assigned each book to a cluster based on its feature values. We visualized the results by creating bar charts and scatter plots. The bar charts show the mean value of each feature for each cluster and the genre distributions (Chart 12 & 13, p. A-6, A-7).

Chart 13: Analysis of K-Means Clusters (p. A-7)

Cluster 1 had the highest average user rating at 4.61 but the lowest number of reviews at 6,029. It also had the highest average price at 13.36. This cluster likely represents high-quality books that are more niche and appeal to a smaller, but dedicated, audience. Cluster 2 had an average user rating of 4.41 and the highest number of reviews at 58,490. It also had a relatively low average price of 11.69. This cluster likely represents popular books that are affordable and have a broad appeal to readers. Cluster 3 had the second-highest average user rating at 4.68 and a moderate number of reviews at 22,550. It also had the lowest average price at 10.52. This cluster likely represents books that are both popular and affordable,

but may not be as highly rated as those in Cluster 1. With this information, publishers can focus on targeting Cluster 1 to produce books with broad appeal and affordable prices, or Cluster 2 to prioritize producing high-quality books for a niche audience.

Generalized Linear Model (GLM):

We use generalized linear model (GLM), analyzing and modeling binary and count data, to predict whether a book is a bestseller based on its features. We define a book as a good bestseller if it has a user rating of 4.5 or higher and at least 10,000 reviews. Next, the code splits the data into training and testing sets and fits a logistic regression model using the GLM function in R, with "Bestseller" as the response variable and "User.Rating", "Reviews", "Price", "Year", and "Fiction" as the predictor variables.

The model makes predictions on the test set using the "predict" function, and the predicted classes are obtained by applying a threshold of 0.5 to the predicted probabilities. Finally, the code evaluates the performance of the model by computing the confusion matrix, getting an Accuracy score of 0.9018, for the predicted classes against the true labels of the test set.

To identify the factors that affect a book is to become a bestselling bestseller in the training model. We then use the variable importance plot to show the relative importance of each feature in the logistic regression model. And use a scatter plot to show the distribution of “Fiction” in "User.Rating" and "Reviews"(Chart 14 & 15, p. A-7, A-8).

Chart 14 & 15: GLM Variable Importance and Relationship (p. A-7, p. A-8)

In Chart 14, a variable importance plot shows the relative importance of each feature in the logistic regression model. In the plot, we found out that "User.Rating", "Reviews", and “Fiction” have higher ratings than the rest. "User.Rating" & "Reviews" are what really define a Bestselling Bestseller. However, the high ratings are also due to the bar we set for the dataset makes it highly correlated with the result. Thus, we need to consider “Fiction” since it has the third highest variable importance rating. From Chart 15, we found out that “Fiction” books have more top Bestsellers. Therefore, if the publishers want to lead to maximizing success, they should work on “Fiction” books more.

Conclusions and Recommendations:

Firstly, we recommend that the publisher focuses on improving the quality of their books, as the user rating distribution is positively skewed, with a significant proportion of books receiving lower ratings. Publishers can use this information to identify areas for improvement in their books to improve their ratings and increase their appeal to readers. Secondly, our analysis of user ratings and reviews by genre suggests that the publisher should consider publishing more fiction books as they tend to have higher ratings and a higher number of reviews on average. The publisher should consider their specific target audience and genre preferences when making decisions about book publishing. Thirdly, we also recommend that the publisher pay close attention to the three clusters identified in our K-means analysis. For example, the publisher may want to focus on publishing more books that fit into the most popular cluster, while still considering books from the other clusters that may have niche appeal. Finally, based on our analysis of the top authors with the highest mean number of reviews, we recommend that the publisher work on building brands. Books written by well-known/established authors tend to receive higher numbers of reviews. Additionally, we recommend that the publisher focus on creating eye-catching book titles that include keywords that appeal to readers. Our NLP analysis suggests that certain keywords in titles catch readers' attention, and this can be used to the publisher's advantage. Our NLP analysis helped us identify keywords that are commonly used in book titles that attract readers. Some of the top keywords include "love," "life," and "world," which can be useful for publishers when coming up with book titles that are more likely to grab readers' attention. Based on these findings, we recommend that the publisher consider incorporating some of the top keywords when coming up with new book titles to increase the likelihood of attracting readers. Overall, our analyses provide valuable insights that can help the publisher make more informed decisions about their book publishing strategy.

Appendix A: Visualizations

Chart 1: Histogram of User Ratings

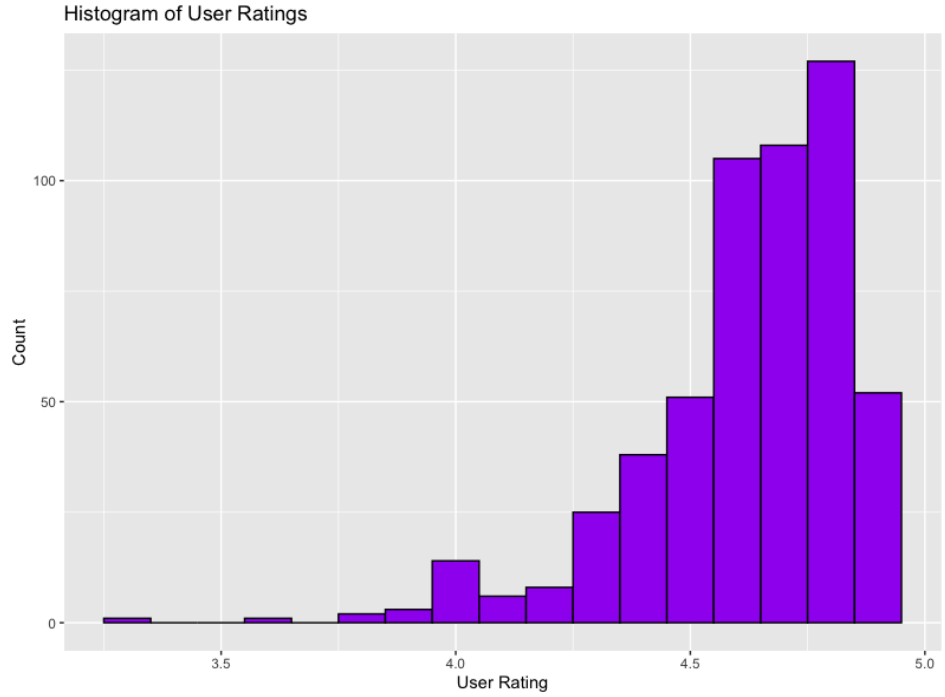

Chart 2: Histogram of Book Prices

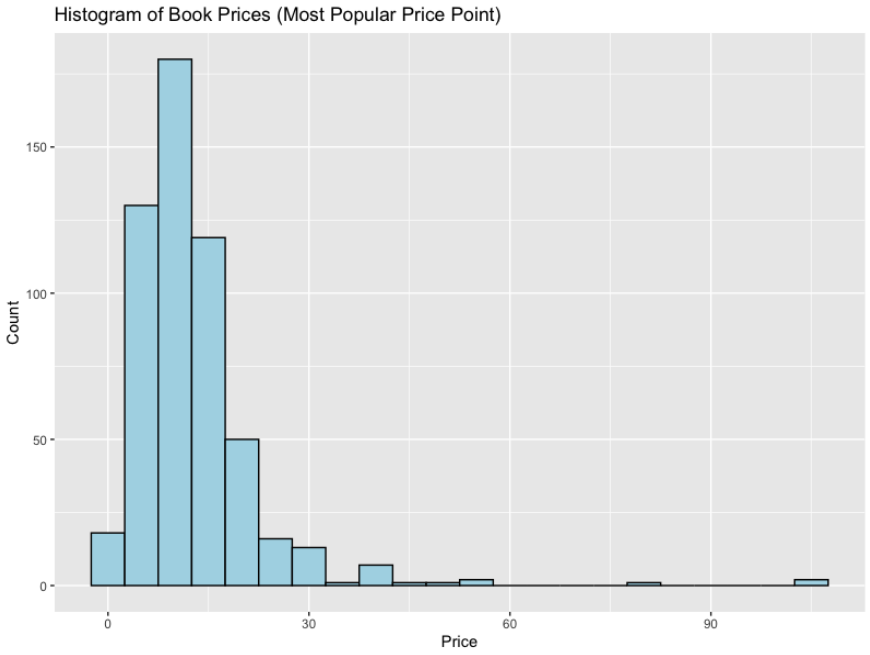

Chart 3: Boxplot of User Ratings by Genre

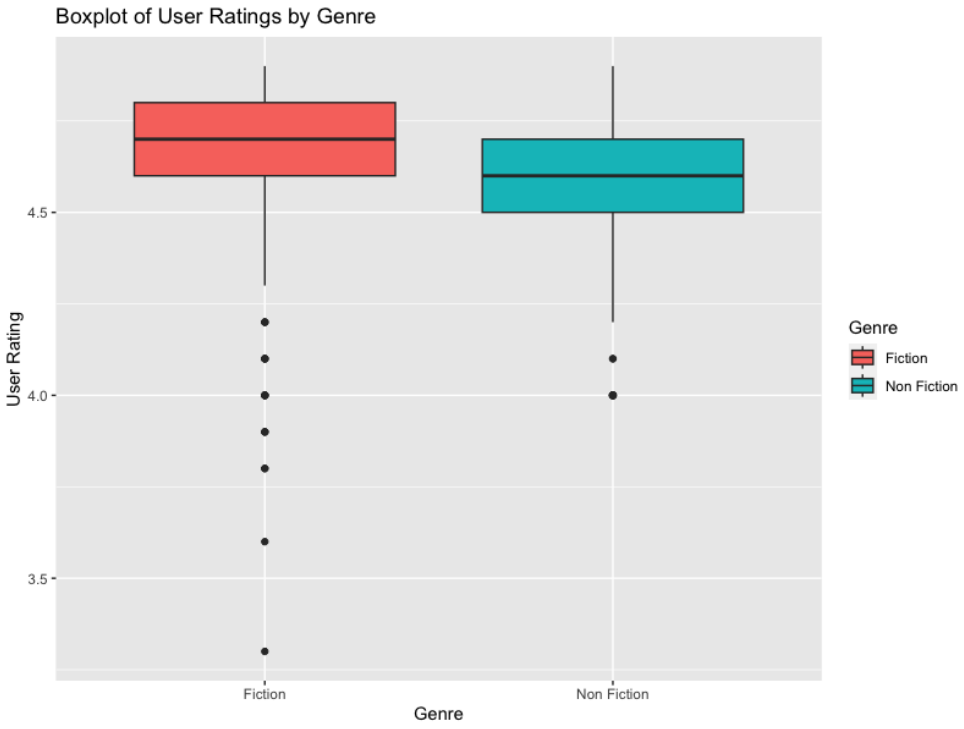

Chart 4: Sum of Reviews by Genre

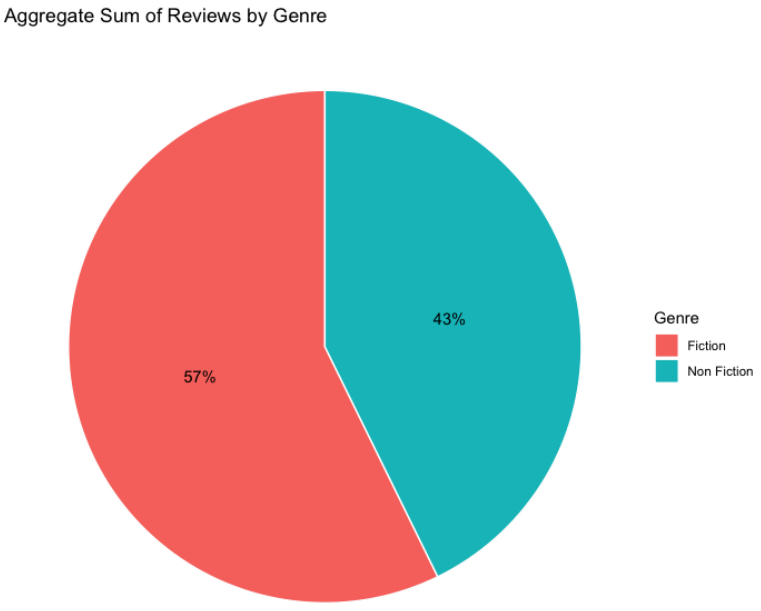

Chart 5: Mean Number of Reviews per Year by Genre

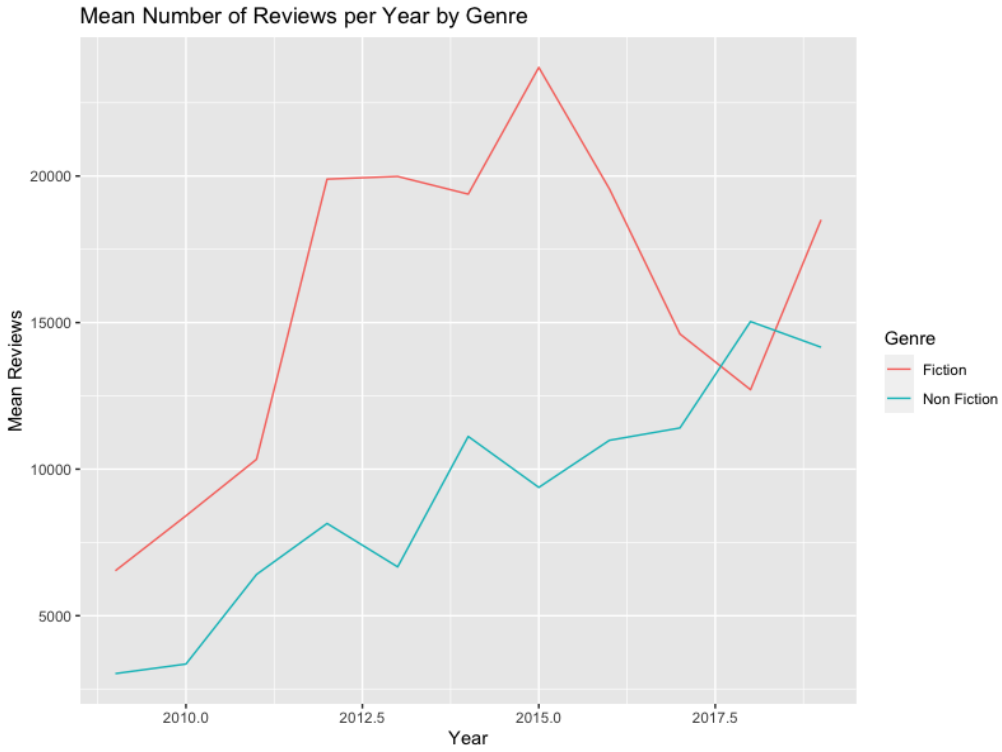

Chart 6: Average Book Reviews by Year and Genre

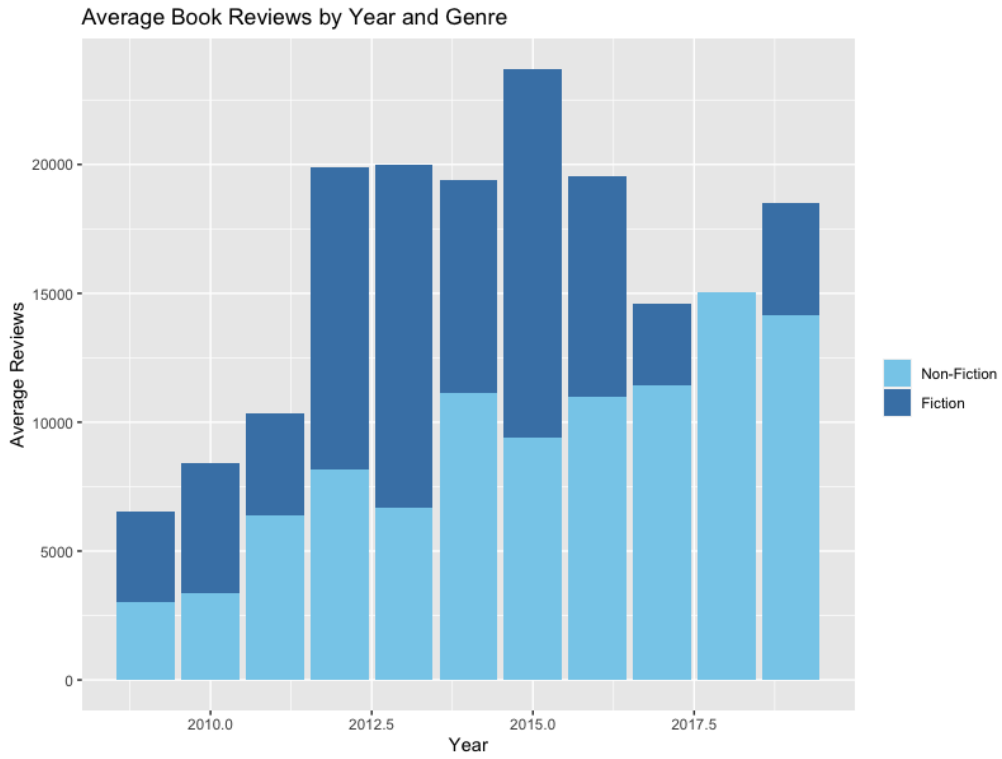

Chart 7: Correlation Matrix Between Features

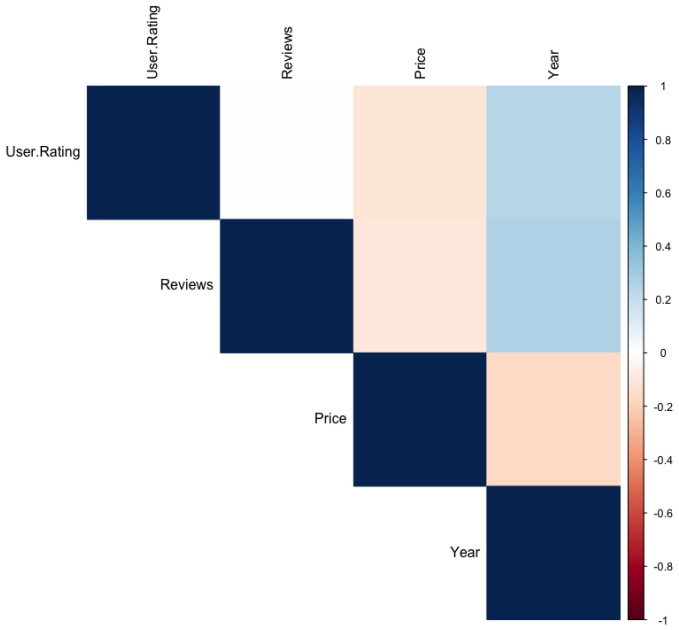

Chart 8: Correlation between User Ratings and Reviews

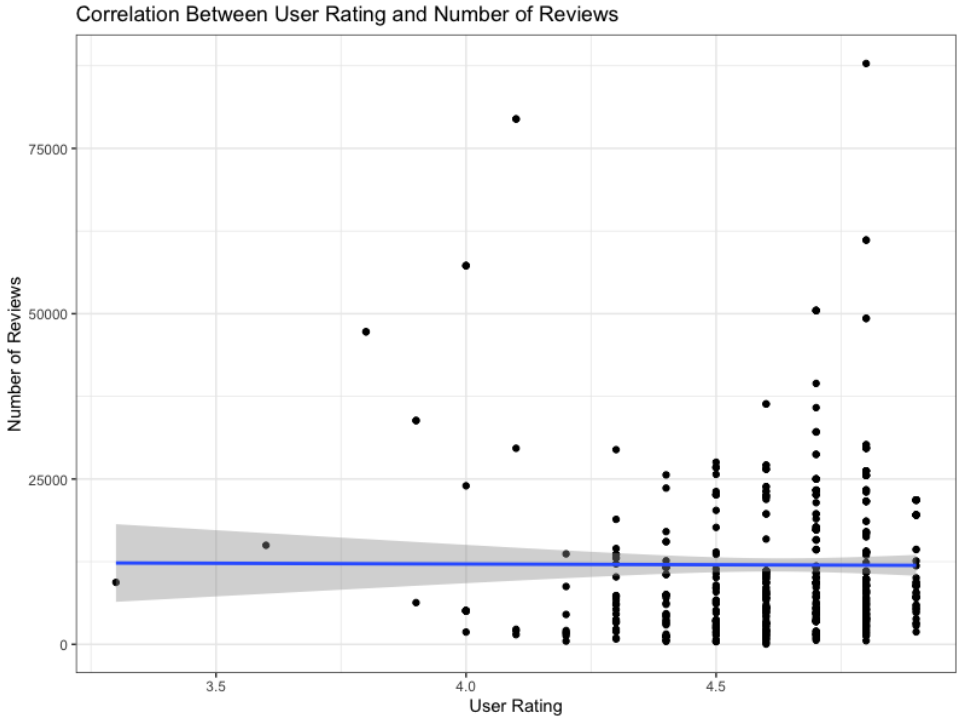

Chart 9: Top Authors with a Price Gradient

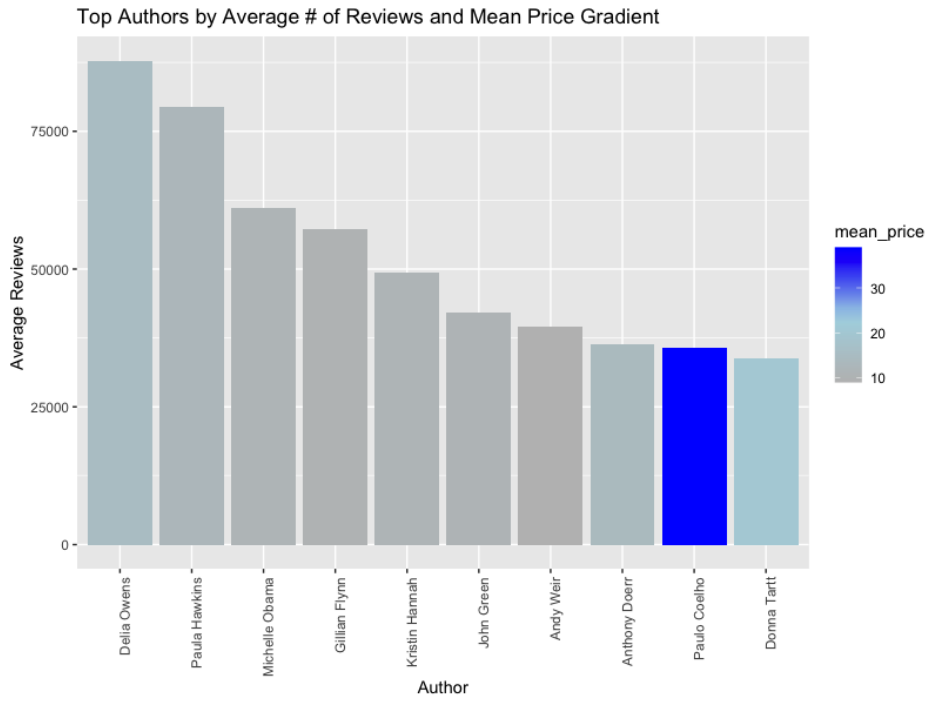

Chart 10: NLP Regression with Top Words in Titles

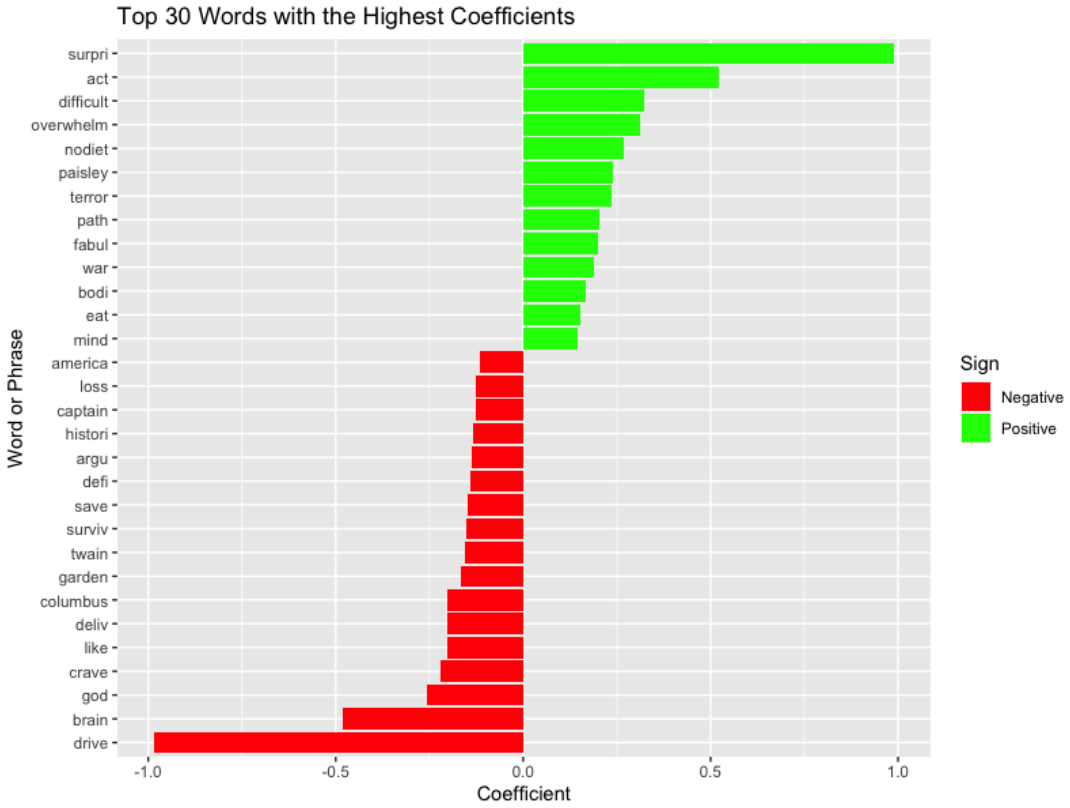

Chart 11: Elbow Plot for K-Means Clustering

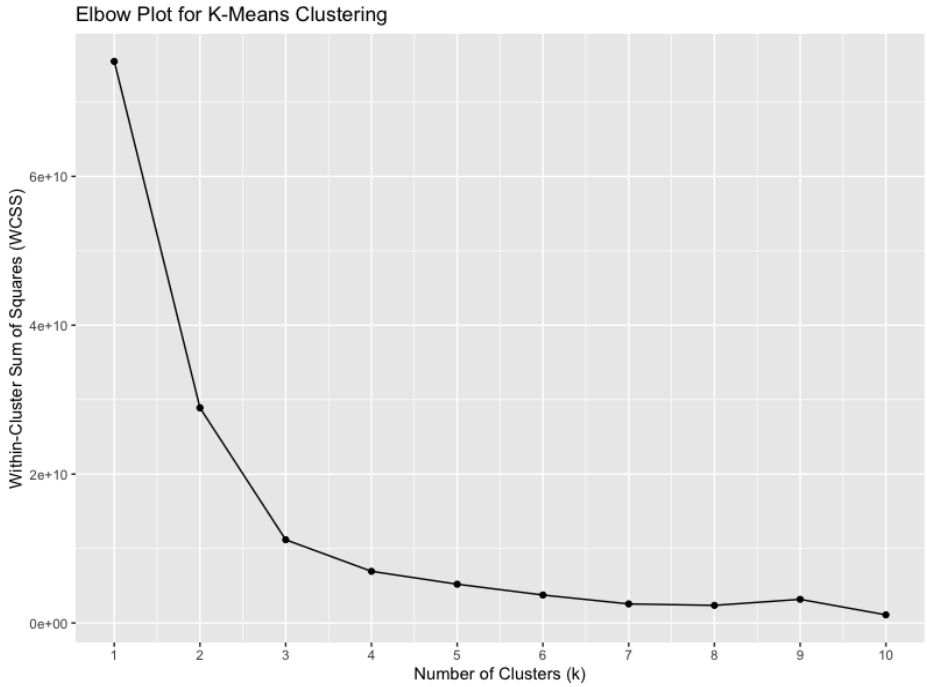

Chart 12: Genre Distribution of Clusters

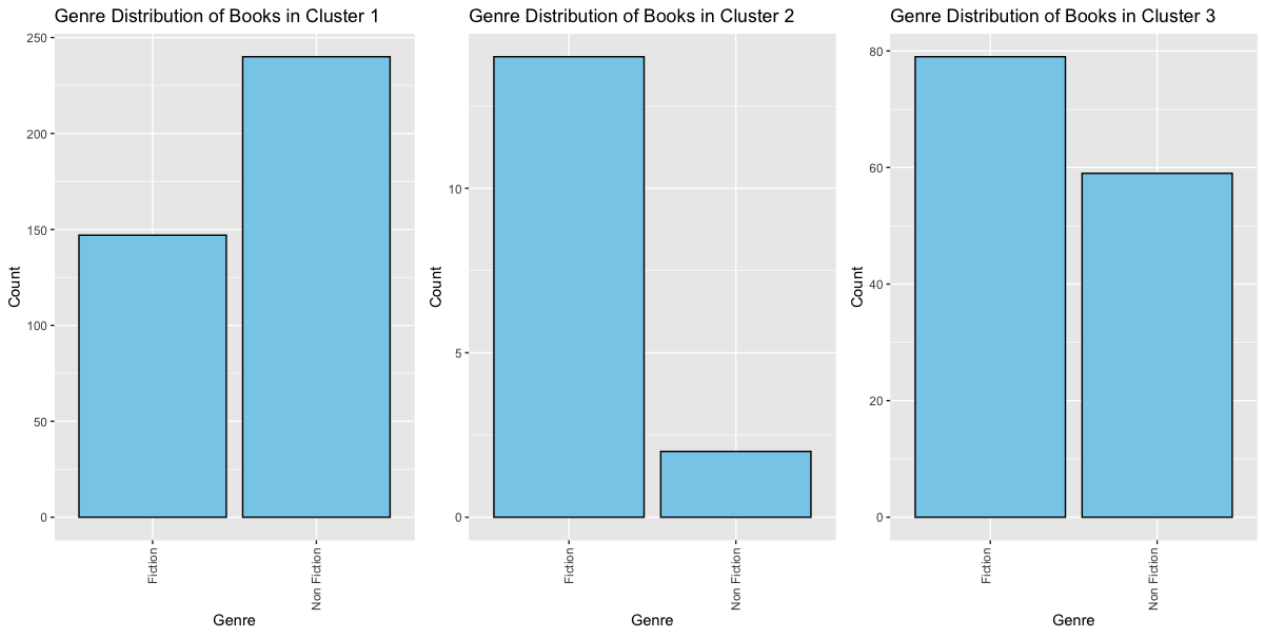

Chart 13: K-Means Cluster Centers

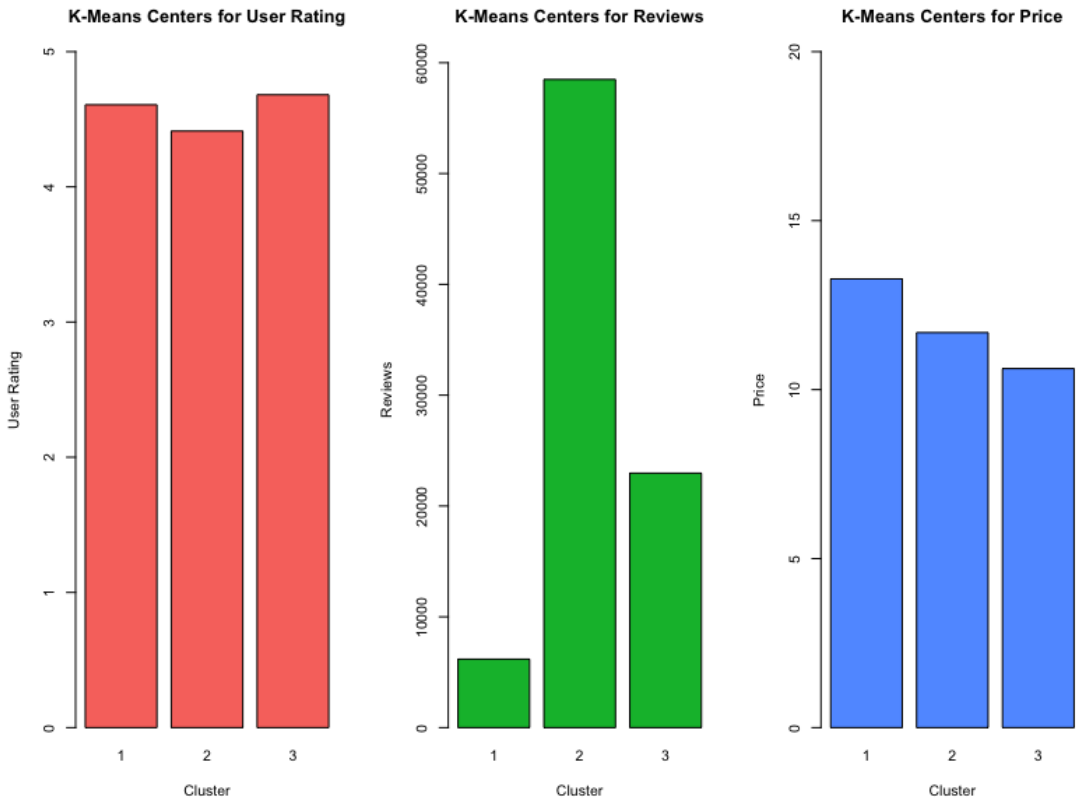

Chart 14: GLM Variable Importance for Bestselling Bestsellers

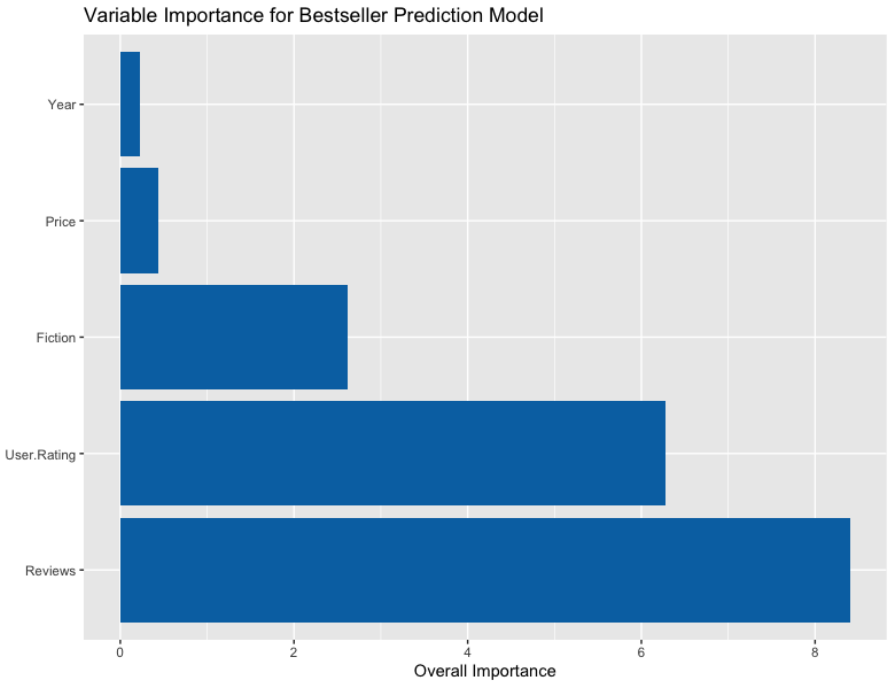

Chart 15: Relationship Between Features from GLM Results