Introduction

The goal of this project is to explore book trends by analyzing multiple book data, and to experiment on predicting genres. Exploring book trends is important especially if you are an author or publisher who wants to understand the market and their customers. Looking at previous data can help an author gauge the popular genres and determine which is more profitable. The three main questions this project explored are: 1. What is the relationship between the gross sale of a book and its genre? 2. Do the top selling authors publish more than one book and are their gross sales about the same? 3. Is it possible to predict the genre of a book based on its title and summary? I'm using two datasets that I found on Kaggle. Dataset 1 has information about the general information about a book and its gross sales. This dataset was updated 2 years ago. Dataset 2 has book data from GoodReads. This dataset includes the summary of the book.

Methods

For dataset 1, I made a new copy of it and kept the columns book_id, book_title, book_details, author, genres, num_reviews, and average_rating. After that I checked for missing values such as blanks , [], and [None] and replaced them with NaN. Then I checked for the total amount of duplicates and dropped them. For dataset 2, I made a new copy of it with the columns publishing year, book name, author, language code, gross sales, publisher revenue, and sale price. Then I removed duplicates and dropped any rows that did not have a book title. I also turned the data under the column year into a whole number. After cleaning both of the datasets, I merged them through the book title using groupby and renamed the columns to make it more consistent. I also realized that there were some books with multiple genres, and I felt that it would be quite difficult to build a predictor for a book with multiple genres so I assigned those books with only the first genre in their data.

For visualization, I used multiple charts such as a pie chart and bar charts. The pie chart was made by using px.pie and it is on the relationship between gross sales and book genres. The data for this chart was made by using groupby on the columns genre and the sums of the gross sales for each of the genres. Both the pie and bar chart shows the distribution of the gross sales of book genres from the dataset. The last bar chart I have is on the top 10 authors according to the gross sales of their books. The y axis displays the top 10 authors and the y axis is the gross sales of the authors. Inside the chart, it shows the distribution of their gross sales on their published books.

I used the Naive Bayes model from the sklearn library to predict the genre of the book. My BookGenrePredictor class includes the functions tokenize_data, train, prediction_table, and predict. To prepare the text, I tokenized it by removing punctuation, spaces, and stopwords. Then the text is converted to lowercase and appended to a new list. After that I have to pass the tokenized data to where the text gets converted to numeric features (Bags of Words) by using CountVectorizer from the sklearn library. Then I encoded the genre labels into numbers. After that the data is split into training and testing data and passed to the model which is MultinomialNB() from the sklearn library.

The performance of the model is evaluated by the accuracy of its prediction of the book genre. I ran a classification report from sklearn.metrics and it gave me the details on the accuracy of its predictions. It also outputs the precision, recall, f1-score, support for each of the genres.

Results

The pie chart answers the first question, what is the relationship between the gross sales of a book and its genre. The genre that has the highest gross sales is fiction. It makes sense why fiction would rank first because within fiction there are many other genres that fall into that category. The genre with the second highest gross sales is fantasy at 17% which is very close to fiction at 23.8% of the pie chart. The bar chart shows the same thing as the pie chart but with a different visualization. The last chart answers the second question: do the top selling authors publish more than one book and are their gross sales about the same? According to the chart, most of the sales of the top ten authors from this dataset have one book that dominates their gross sales. So the answer to that question would be yes most of these authors have more than one book that are published but their gross sales are not the same. But there is also an exception, which is Stephen King. Looking at his bar, it seems that he published many books and the distribution of the sales are more even compared to the other authors.

The accuracy for the book genre predictor is very low. For prediction of the genre based on the book title, the accuracy rate is 32.9%. For the prediction of the genre based on the book summary, the accuracy rate is 45%, which is a little bit higher than the title based prediction. This may be because there is very little context that can be drawn from the book title and there are very few words in the title for the model to be able to accurately predict the genre. Some book titles only have one word and that is not enough for the model to make an accurate prediction.

Discussion

Based on the results of my genre prediction, we can try to predict the book genre based on its title and summary but it will not be very accurate. So the answer to the last question is that it's hard to predict the genre because of the length of the title and the lack of context.

If I had more time on this project, I would try other methods and models to predict the genre of the book and then compare their accuracies to each other. I would also test them on multiple datasets to see if there are limitations of models when working with different data sizes.

Conclusion

Overall this project analyzed book data and the relationship between the gross sales of a book and its genre. This project also tried to predict the genre of a book based on its title and summary but had a very low accuracy rate.