Books Dataset Analysis

Introduction:

The purpose of this report is to analyze the books dataset and present the results of the exploratory data analysis (EDA) and the machine learning (ML) models built. The dataset was pre-processed to remove outliers and improve the quality of the data. The selected features for the ML model were ['book', 'author', 'lge','num_pages', 'ratings_count', 'text_reviews_count', 'year']. The ML models used for prediction were Linear Regression and Random Forest Regressor, and their results were compared. The code and models can be explored in the interactive website https://books-eda.onrender.com/ and the source code for the project can be found in the Github repository https://github.com/zackemcee/books_analysis
The dashboard is split into two parts, ML and EDA. The ML section focuses on using the name of the book, author, year of publication, language, number of pages, number of reviews, text reviews, and type of model as filters to predict the average rating of a book. The EDA section focuses on analyzing the data in-depth based on publishers, authors, year, and language.

Exploratory Data Analysis (EDA):

The EDA was performed to explore the book's dataset and detect any outliers. Outliers were defined as rows with the number of pages greater than 1000, number of ratings greater than 500k, and number of text reviews greater than 5000. These rows were removed to improve the quality of the data and reduce the impact of extreme values. The final ML dataset contained only rows that met the criteria defined above.

Machine Learning (ML) Models:

The goal of the ML models was to predict the average rating of a book based on the selected features. Two models were used, Linear Regression and Random Forest Regressor.

Linear Regression:

The Linear Regression model uses a linear function to fit the data and make predictions. The model was trained and tested on the same dataset.

Random Forest Regressor:

The Random Forest Regressor uses an ensemble of decision trees to predict the outcome. The model was trained and tested on the same dataset.
The Random Forest Regressor is often more accurate than the Linear Regression model, and is a good choice when the relationship between the features and the target variable is complex.

Conclusion:

In conclusion, this report presents the results of the EDA and ML models built on the books dataset. The data was pre-processed to remove outliers, and the selected features for the ML model were ['book', 'author', 'lge','num_pages', 'ratings_count', 'text_reviews_count', 'year']. The ML models used for prediction were Linear Regression and Random Forest Regressor, with the latter being often more accurate. The code and models can be explored in the interactive website https://books-eda.onrender.com/ The source code for the project can be found in the Github repository https://github.com/zackemcee/books_analysis The dashboard is split into two sections, ML and EDA, allowing for a detailed analysis of the data and predictions of the average ratings of books.