KLE Society's

KLE Technological University



**Exploratory Data Analysis**

**(22ECAC210)**

**Course Project Report on**

**“Predict the prices of books”**

*Submitted in partial fulfilment of the requirement for the award of*

**Degree of Bachelor of Engineering**

**in**

**Computer Science and Engineering**

*Submitted By*

**Tarun S M – 01FE21BCI008**

**Sharad C K – 01FE21BCI011**

**Suresh H B – 01FE21BCI015**

*Submitted To*

**SCHOOL OF COMPUTER SCIENCE & ENGINEERING,**

KLE TECHNOLOGICAL UNIVERSITY

HUBLI – 580 031 (India).

Abstract

This exploratory data analysis (EDA) project focuses on predicting book prices using a comprehensive dataset comprising 6237 records for training and 1560 records for testing. The dataset contains various features such as book titles, authors, editions, customer reviews, ratings, synopses, genres, book categories, and prices, serving as inputs to build a machine learning model for price prediction.

The primary objective of this study is to develop a robust and accurate model that can predict book prices based on the provided features.

The methodology involves a systematic approach, starting with data collection from diverse genres and authors. Preprocessing steps are employed to handle missing data and outliers, ensuring data integrity. Exploratory data analysis employs visualizations and descriptive statistics to identify patterns and trends that influence book prices.

Feature selection is crucial in building the prediction model, involving dimensionality reduction techniques to choose relevant features. The data is analyzed using appropriate statistical methods and models to draw meaningful insights and correlations between book features and prices.

Through this study, we aim to gain valuable insights into the factors affecting book prices and their implications for readers and sellers. The findings and recommendations derived from the analysis will guide book pricing strategies, ultimately benefiting both book enthusiasts and industry stakeholders.

This report provides a concise and informative summary of the EDA project, allowing readers to grasp the project's essence and its significance in predicting book prices.

Table of Content:

Chapter 1 page no

Chapter 2 page no

… …

… …

**Chapter 1. Introduction**

**Overview of the EDA Project:**

The Exploratory Data Analysis (EDA) project aims to predict the prices of books using a comprehensive dataset containing various features, including book titles, authors, editions, customer reviews, ratings, synopses, genres, book categories, and prices. The dataset consists of 6237 records for training and 1560 records for testing, providing ample data for the development of a machine learning model to forecast book prices accurately.

**Importance of EDA in Data Analysis:**

Exploratory Data Analysis (EDA) plays a crucial role in the initial stages of data analysis. Its significance lies in the following aspects:

**1. Data Understanding:** EDA allows us to gain a deep understanding of the dataset by examining its structure, distribution, and characteristics. It helps us identify the types of variables present, their relationships, and potential patterns or trends.

**2. Data Cleaning:** During EDA, we detect and handle missing data, outliers, and inconsistencies in the dataset. Cleaning the data ensures that the subsequent analysis is based on reliable and accurate information.

**3. Feature Selection:** EDA aids in selecting relevant features that have a significant impact on the target variable (book prices, in this case). Identifying the most influential features improves the model's predictive power and reduces computational complexity.

**4. Relationship Identification:** EDA helps uncover correlations and dependencies between variables. Understanding these relationships can provide valuable insights into the factors influencing book prices, such as the impact of author popularity, book genre, and customer reviews.

**5. Visualization:** Visualizations, such as plots, graphs, and charts, are an essential part of EDA. They offer a clear representation of the data, making it easier to communicate findings and understand complex relationships.

**The objectives of the course project are as follows:**

**1. Price Prediction Model:** The primary objective of the project is to develop a machine learning model capable of predicting the prices of books based on various features such as book titles, authors, editions, customer reviews, ratings, synopses, genres, and book categories.

**2. Data Exploration:** Conduct thorough exploratory data analysis (EDA) to gain insights into the dataset, identify patterns, trends, and relationships between book features and prices. This exploration will help understand the data distribution and aid in feature selection.

**3. Data Cleaning and Preprocessing:** Perform data cleaning to handle missing values, outliers, and inconsistencies in the dataset. Preprocess the data to ensure its quality and suitability for model development.

**4. Feature Selection:** Select the most relevant and influential features that significantly impact book prices. Employ dimensionality reduction techniques if necessary to reduce computational complexity.

**5. Statistical Analysis:** Apply appropriate statistical methods and techniques to analyze the relationships between book features and prices. This analysis will provide a deeper understanding of the factors affecting book pricing.

**6. Model Building:** Develop a robust machine learning model using the selected features and the target variable (book prices). Experiment with various algorithms and techniques to achieve the most accurate price predictions.

**2. Data Collection**

**Description of the Dataset:**

The dataset used in the course project is a comprehensive collection of information on books, encompassing various attributes that can influence book prices. The dataset consists of 6237 records, which are utilized for training, and 1560 records for testing the machine learning model's performance in predicting book prices. Each record contains the following features:

**1. Title:** The title of the book.

**2. Author:** The author(s) of the book.

**3. Edition:** The edition of the book, including format details (e.g., Paperback, Hardcover) and publication date.

**4. Reviews:** Customer reviews and feedback about the book.

**5. Ratings:** The customer ratings given to the book.

**6. Synopsis:** A brief summary or synopsis of the book's content.

**7. Genre:** The genre or category to which the book belongs (e.g., Fiction, Mystery, Science Fiction, etc.).

**8. BookCategory:** The department or section where the book is typically available (e.g., Literature, Science, History).

**9. Price:** The price of the book (target variable), which is the value we aim to predict using the other features.

**Data Collection Process and Sources:**

The data collection process involves gathering information from various sources, such as online bookstores, libraries, and publishing platforms. Automated web scraping tools may have been used to extract data from websites that list books with their respective details.

To collect the dataset, the following steps might have been taken:

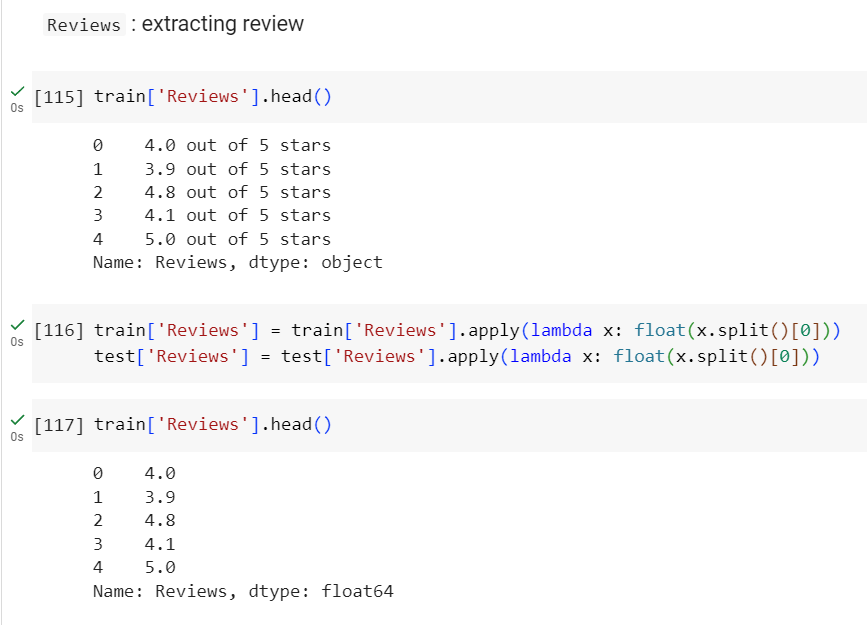
**1. Identification of Sources:** The dataset is available as a CSV file in a downloadable format from <https://machinehack.com/hackathons/predict_the_price_of_books/overview>. The CSV file contains columns representing different features like Title, Author, Edition, Reviews, Ratings, Synopsis, Genre, BookCategory, and Price.

**2. Data Extraction:** Import the pandas library and use the `read\_csv()` function to load the CSV file and create a DataFrame.

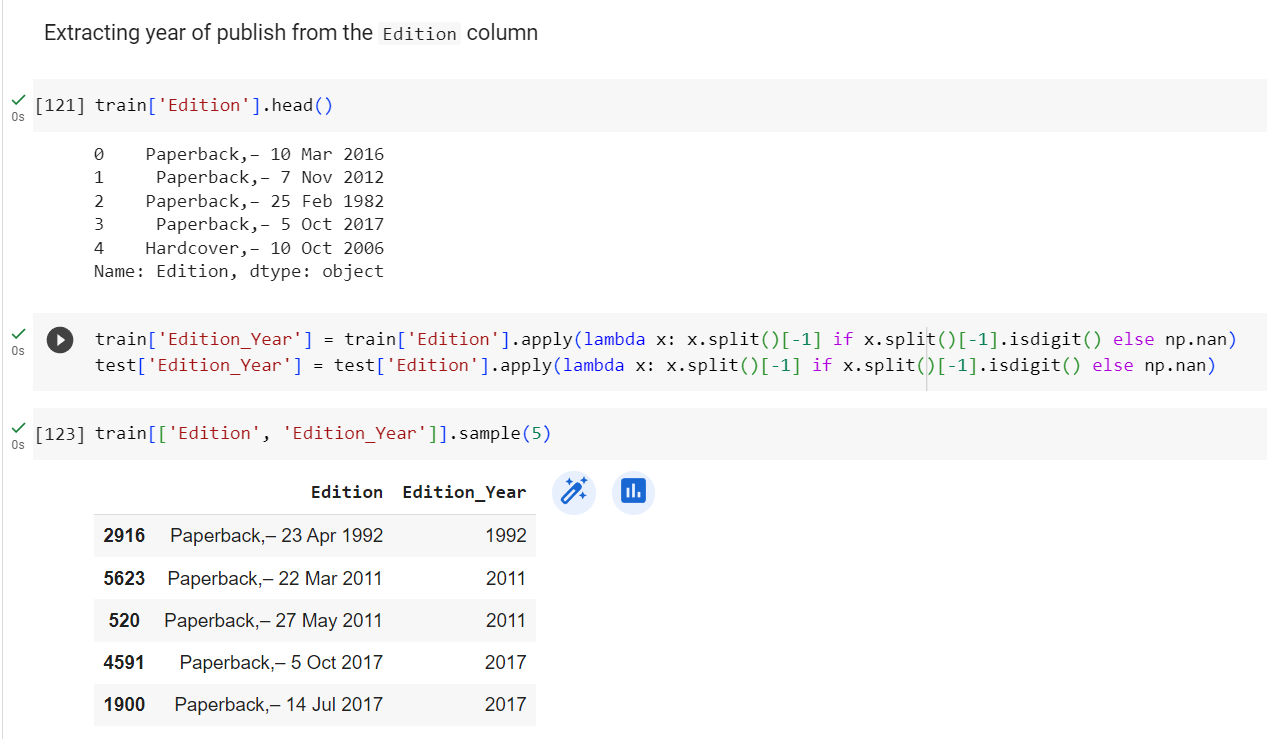
**3. Data Preprocessing:** Once the data is collected, it undergoes preprocessing to handle missing values, remove duplicates, and standardize data formats. This step ensures the dataset's quality and consistency.

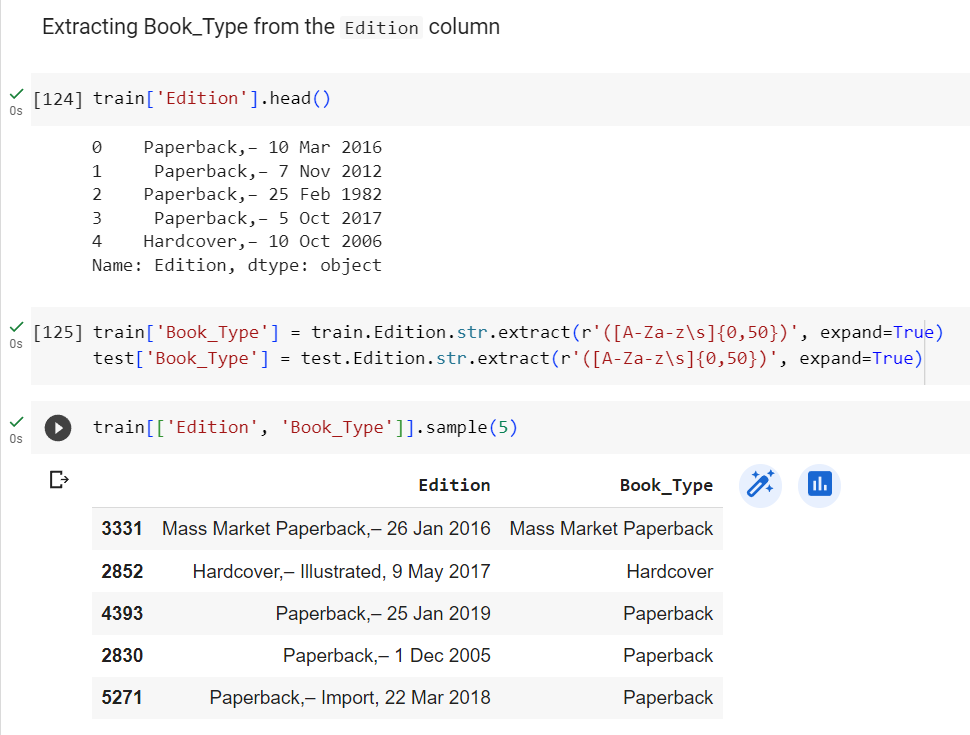
**4. Data Split:** The dataset is split into two sets: the training set, comprising 6237 records, and the test set, containing 1560 records. This division is essential to evaluate the model's performance on unseen data.

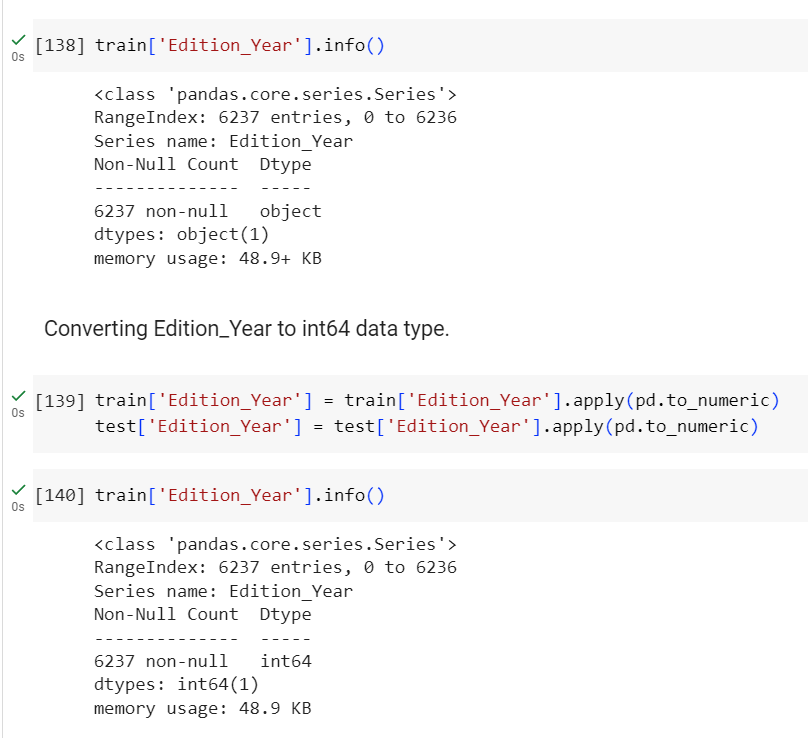
**Preprocessing Steps performed:**





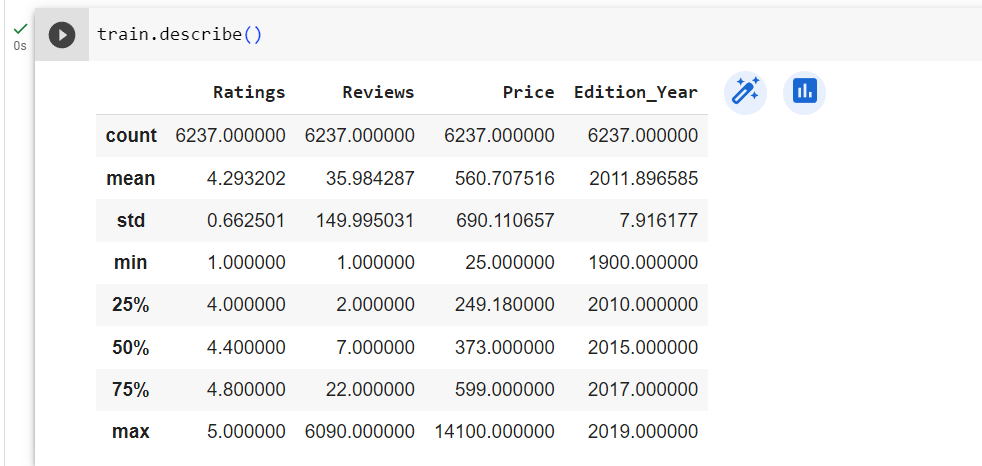




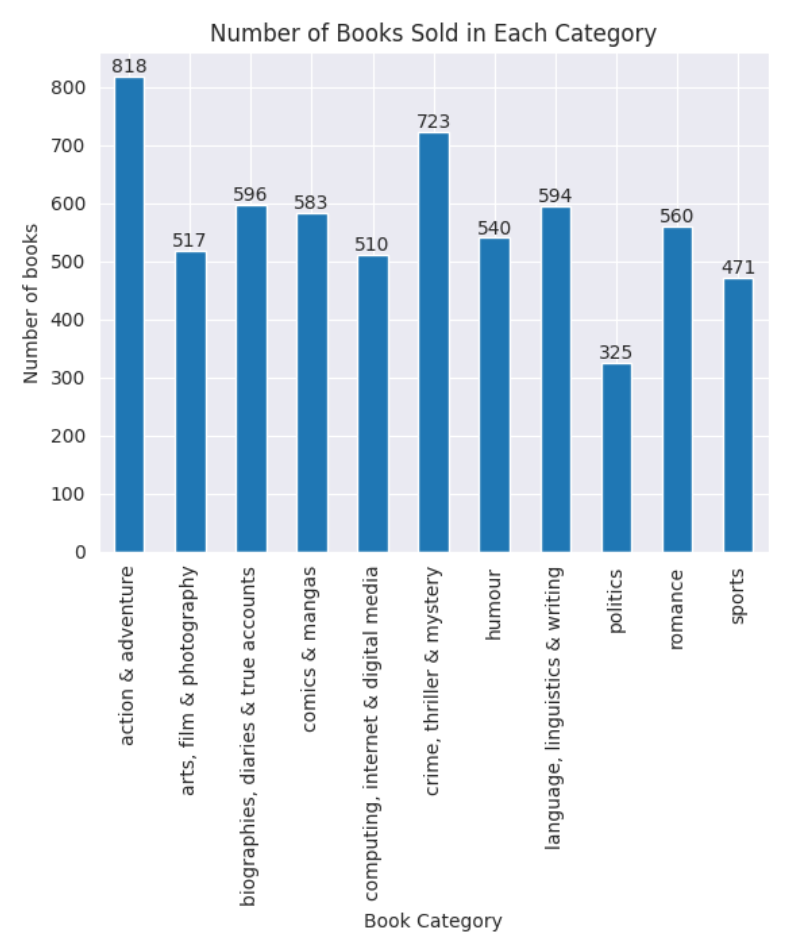


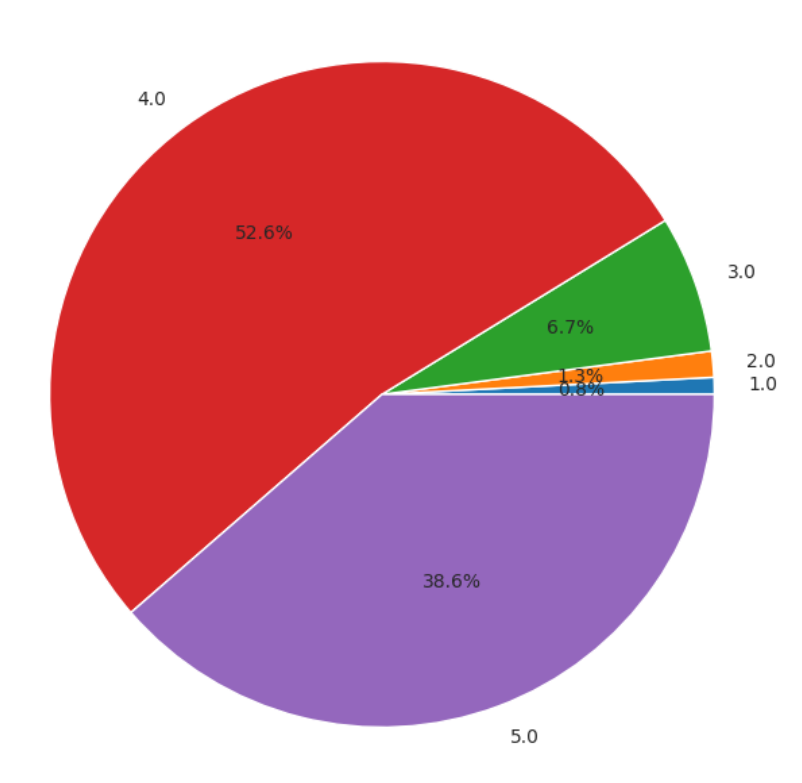
**3. Data Exploration**

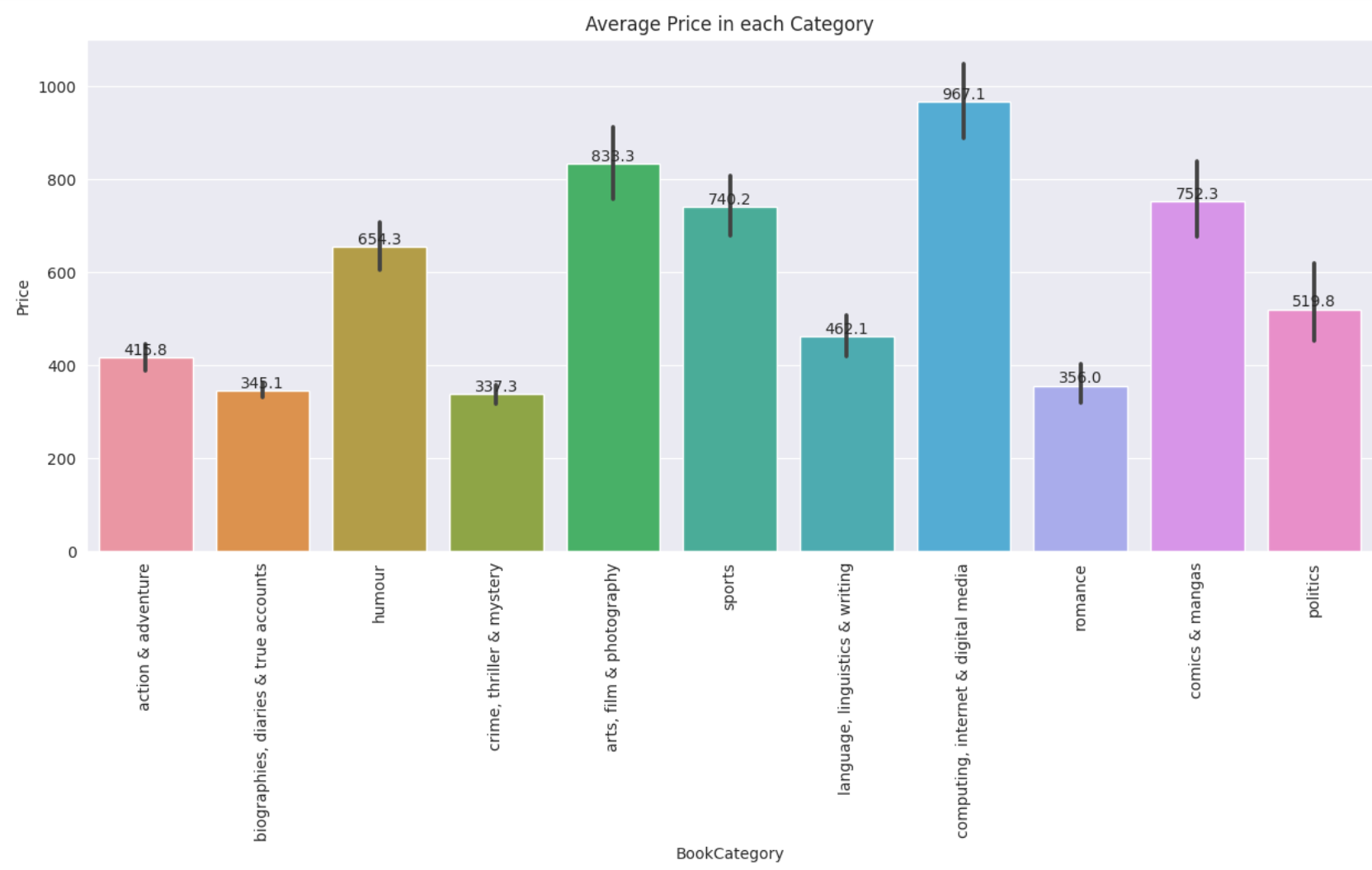
**Present descriptive statistics of the dataset:**

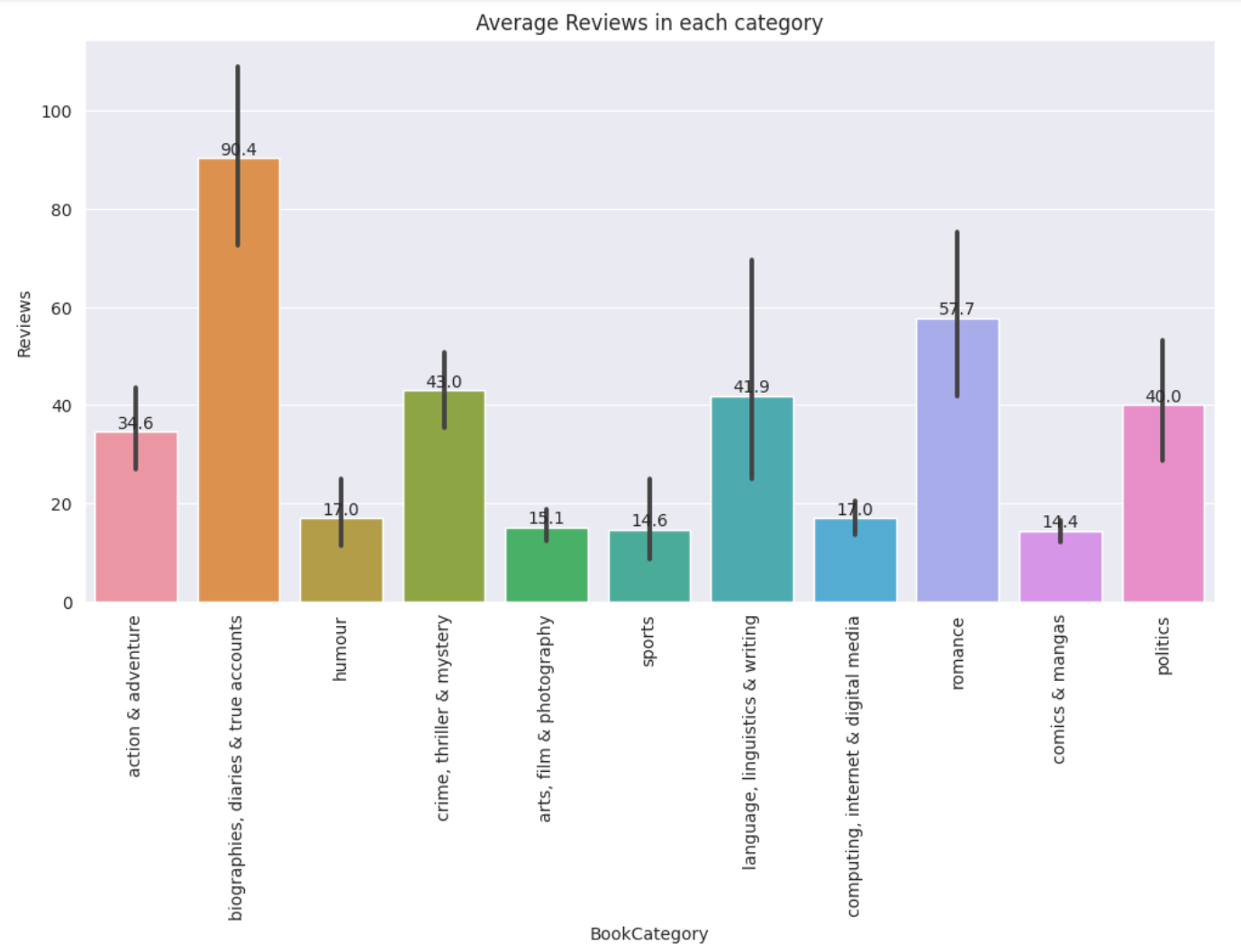


**Visualize the data through plots, graphs, and charts:**

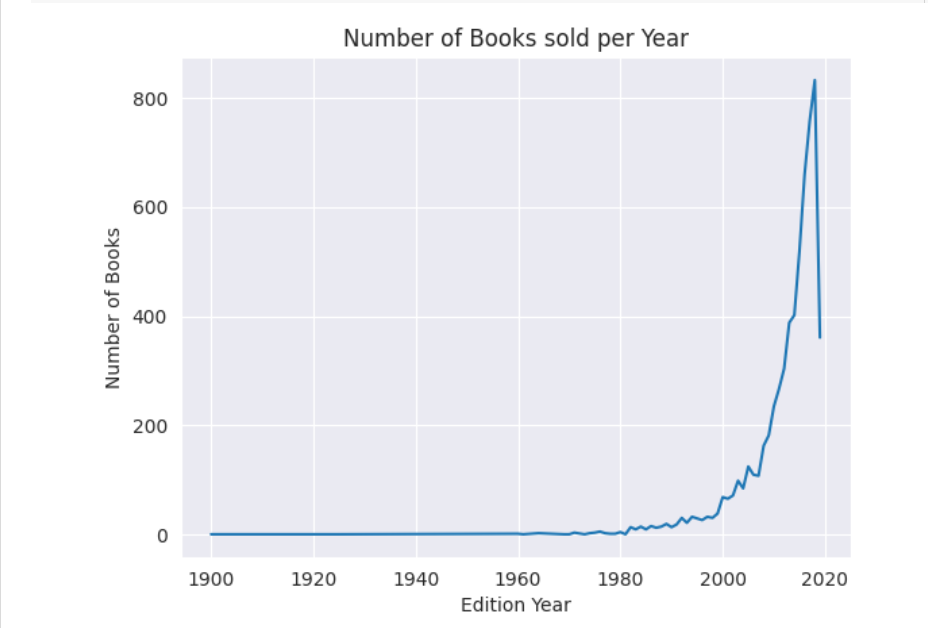


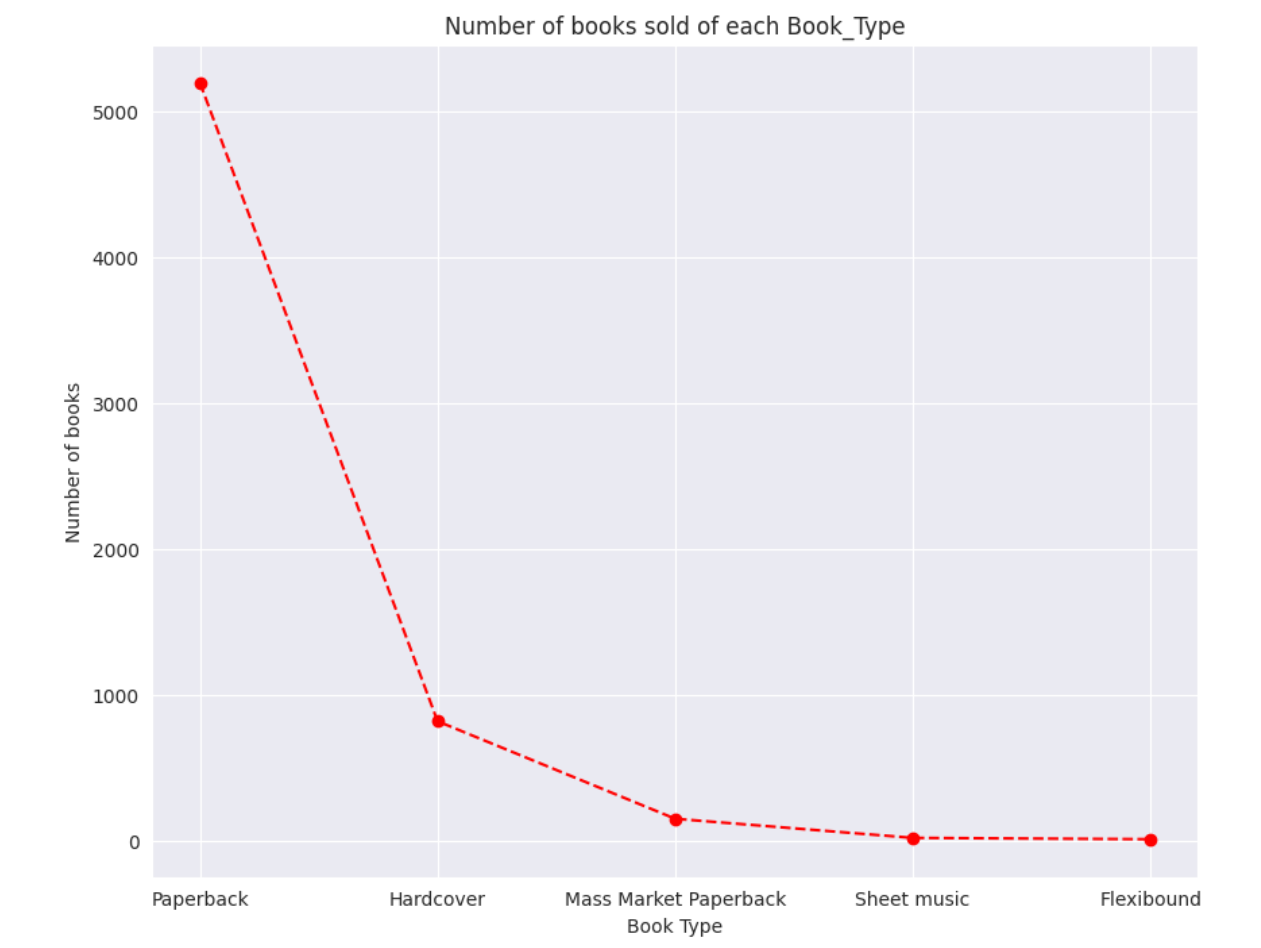


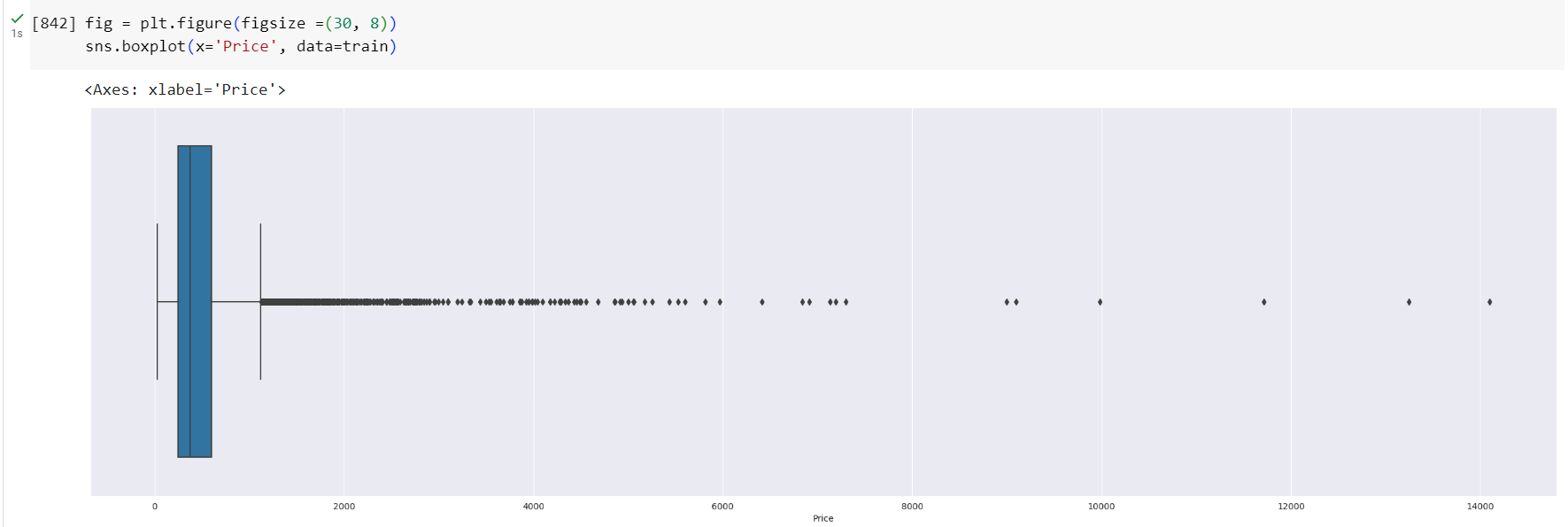




**Identify any patterns, trends, or outliers in the data:**

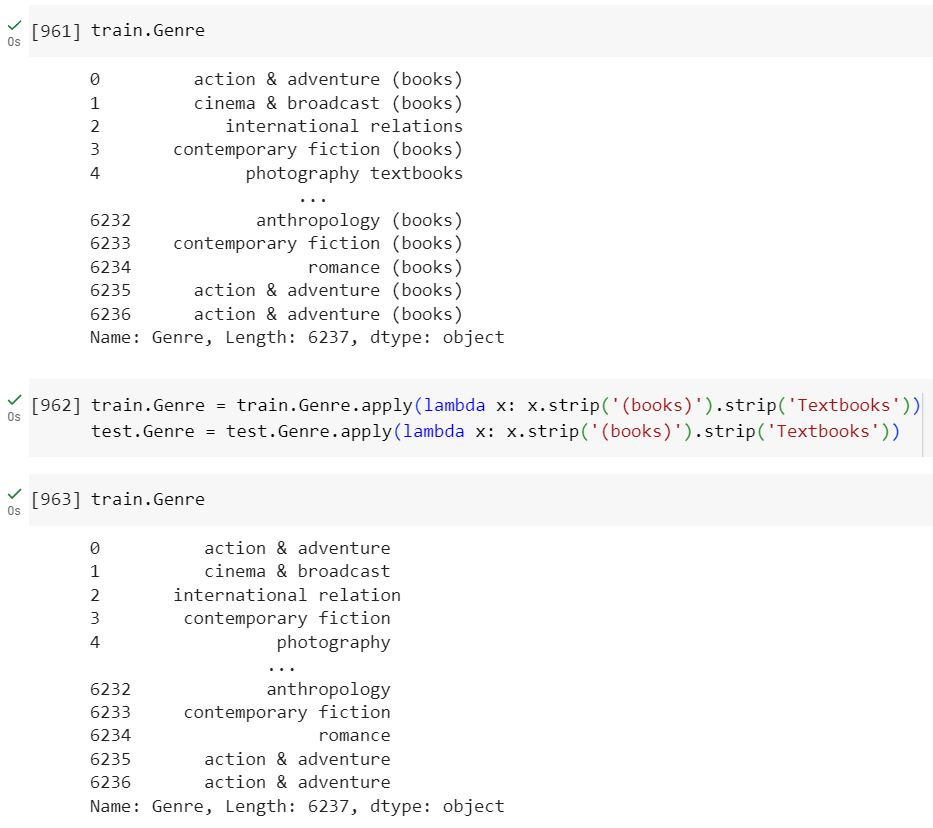






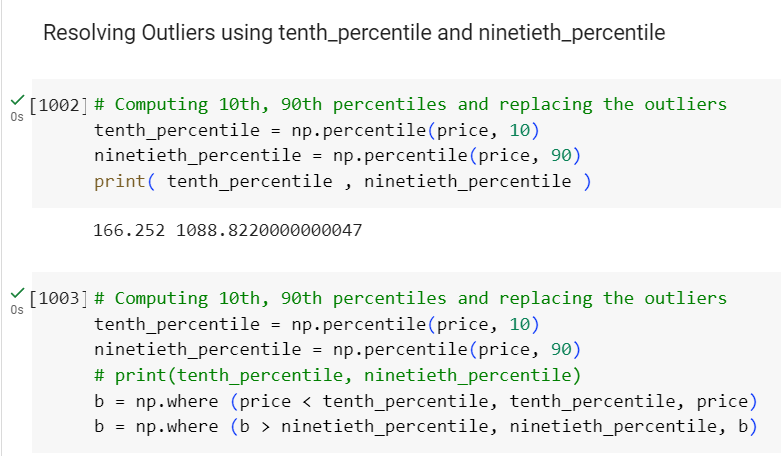
**4. Data Cleaning**

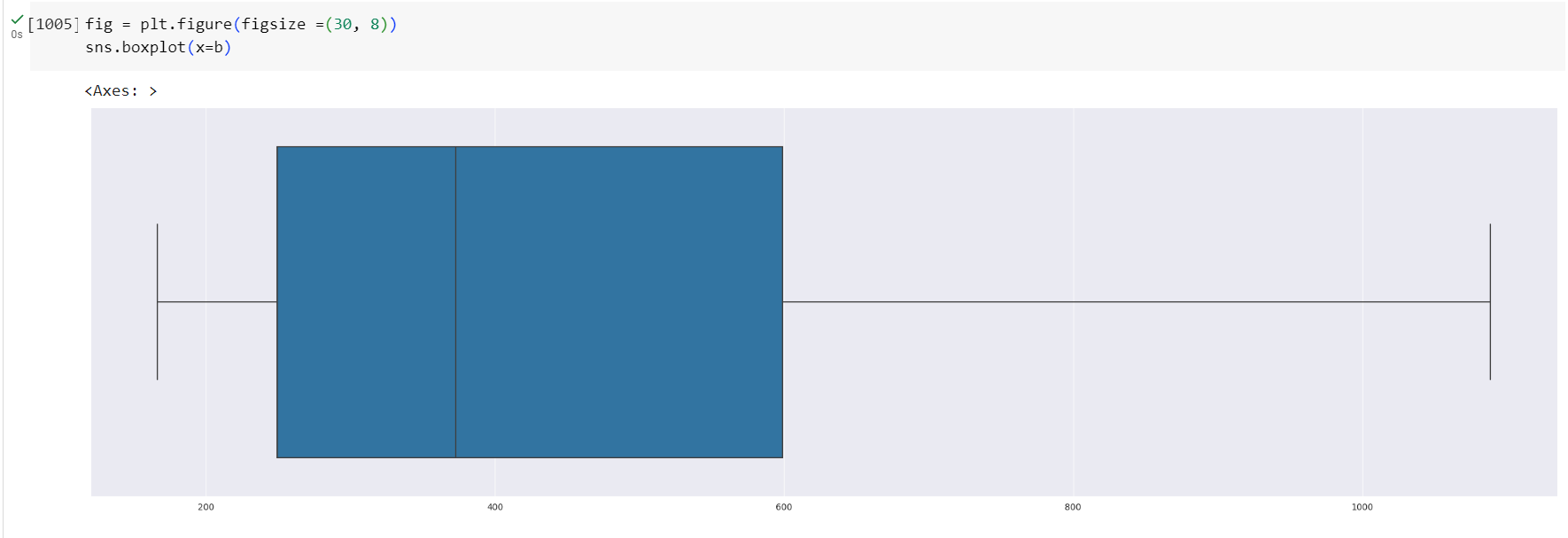
**Discuss the steps taken to clean the dataset:**



**Address missing data and handle outliers if applicable:**

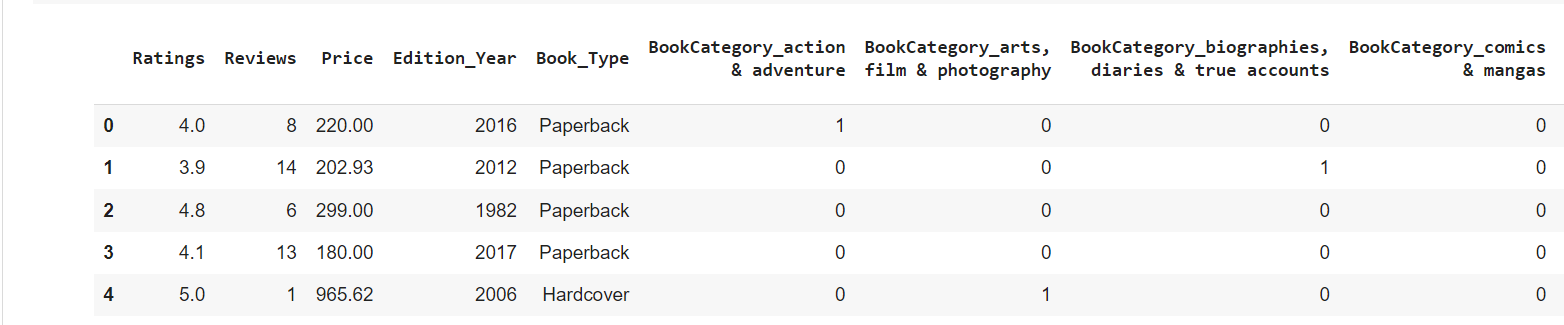






**Explain any data transformation or normalization techniques used:**

**One hot-encoding or using pd.get\_dummies()**



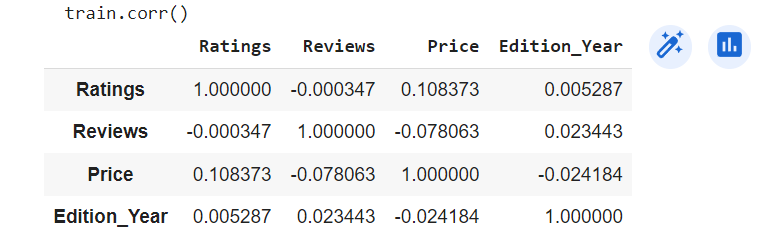
**5. Feature Selection**

**Process of Selecting Relevant Features:**

The process of selecting relevant features involves identifying the subset of features from the dataset that have the most significant impact on the target variable, which, in this case, is the book price. Here's an outline of the steps in the feature selection process:

**1. Initial Feature Exploration:** Begin by examining all the features in the dataset and their potential relevance to book prices. Consider features such as author popularity, book ratings, reviews, genre, and edition.

**2. Correlation Analysis:** Calculate correlation coefficients between each feature and the target variable (book prices). Features with high positive or negative correlations are more likely to be relevant to the price prediction.



**Discuss any dimensionality reduction techniques applied.**

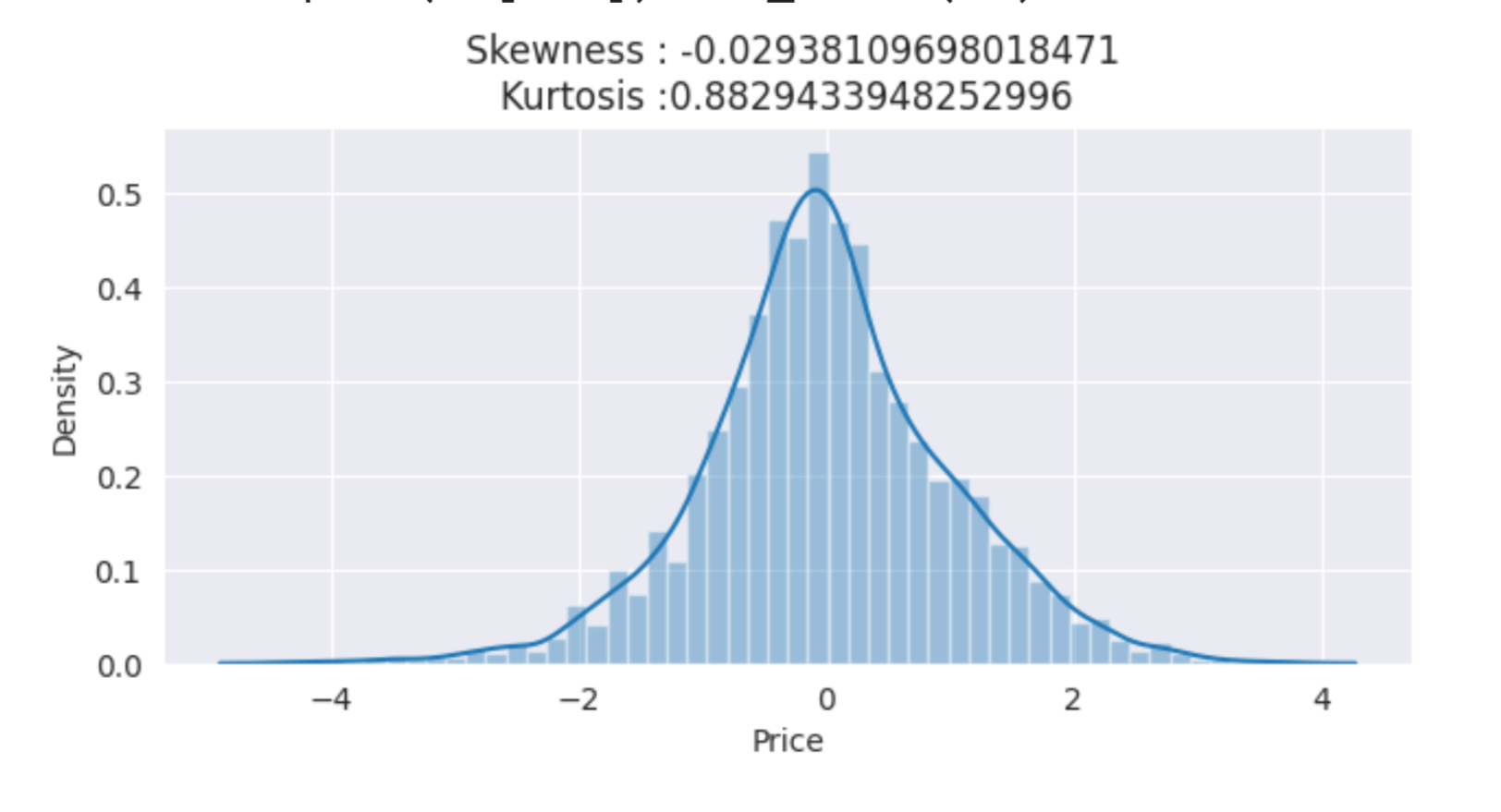
**Explain the rationale behind the feature selection process.**

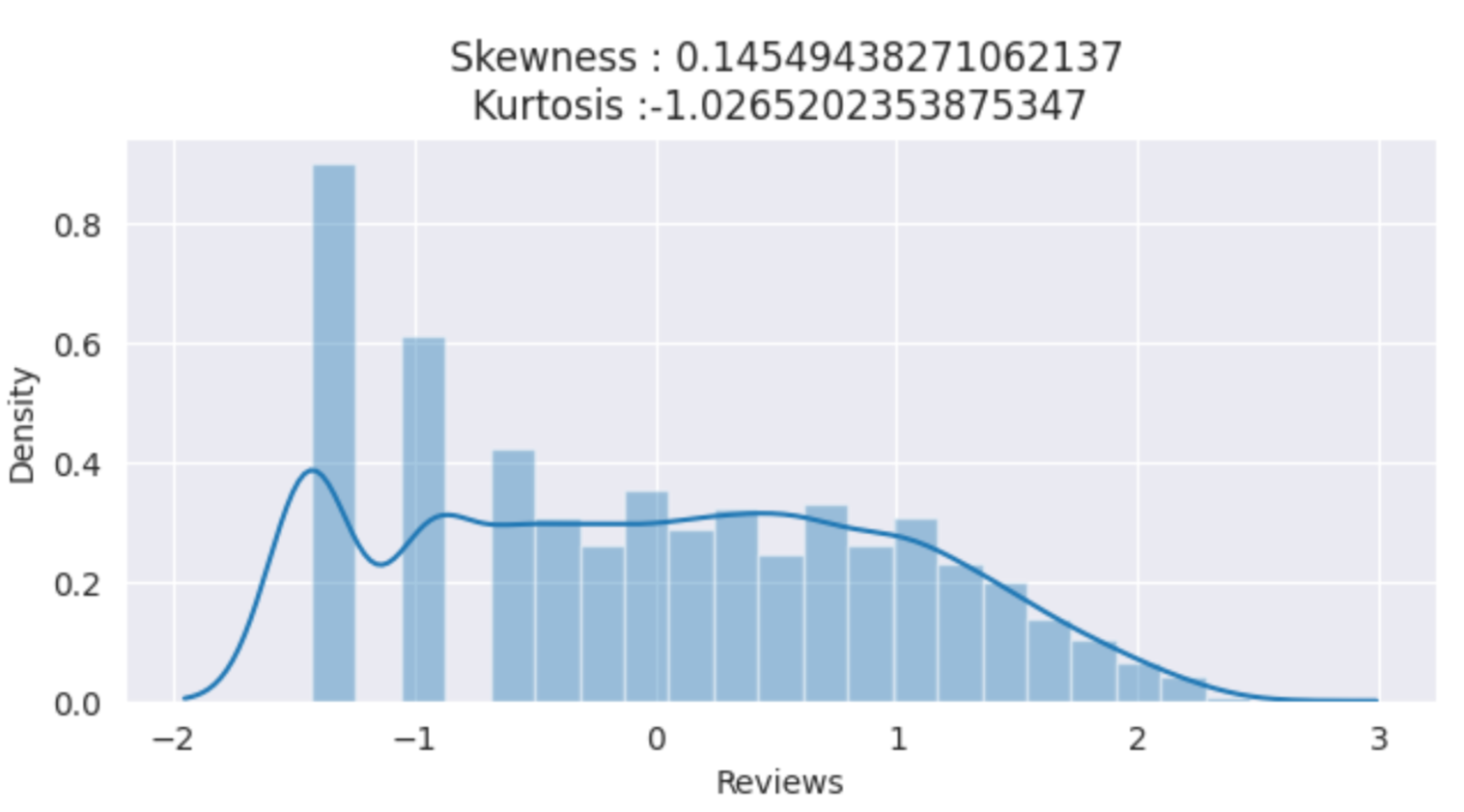
The rationale behind feature selection is to improve model performance, reduce overfitting, and enhance interpretability. By selecting relevant features, we focus the model's attention on the most influential factors affecting book prices, leading to a more accurate price prediction. Removing irrelevant or redundant features also simplifies the model, making it easier to interpret and understand.

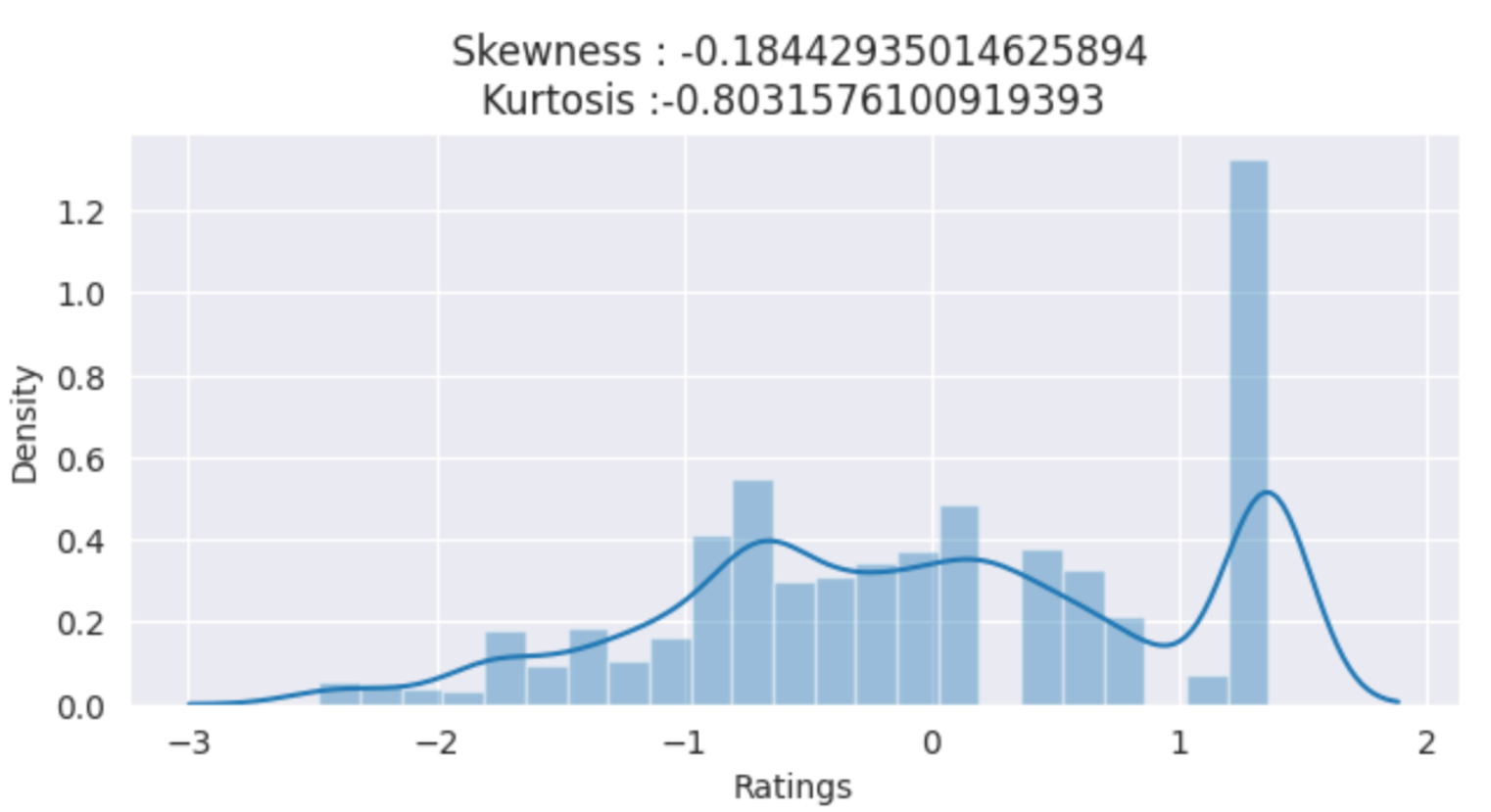
**6. Data Analysis**

**Perform statistical analysis on the dataset.**

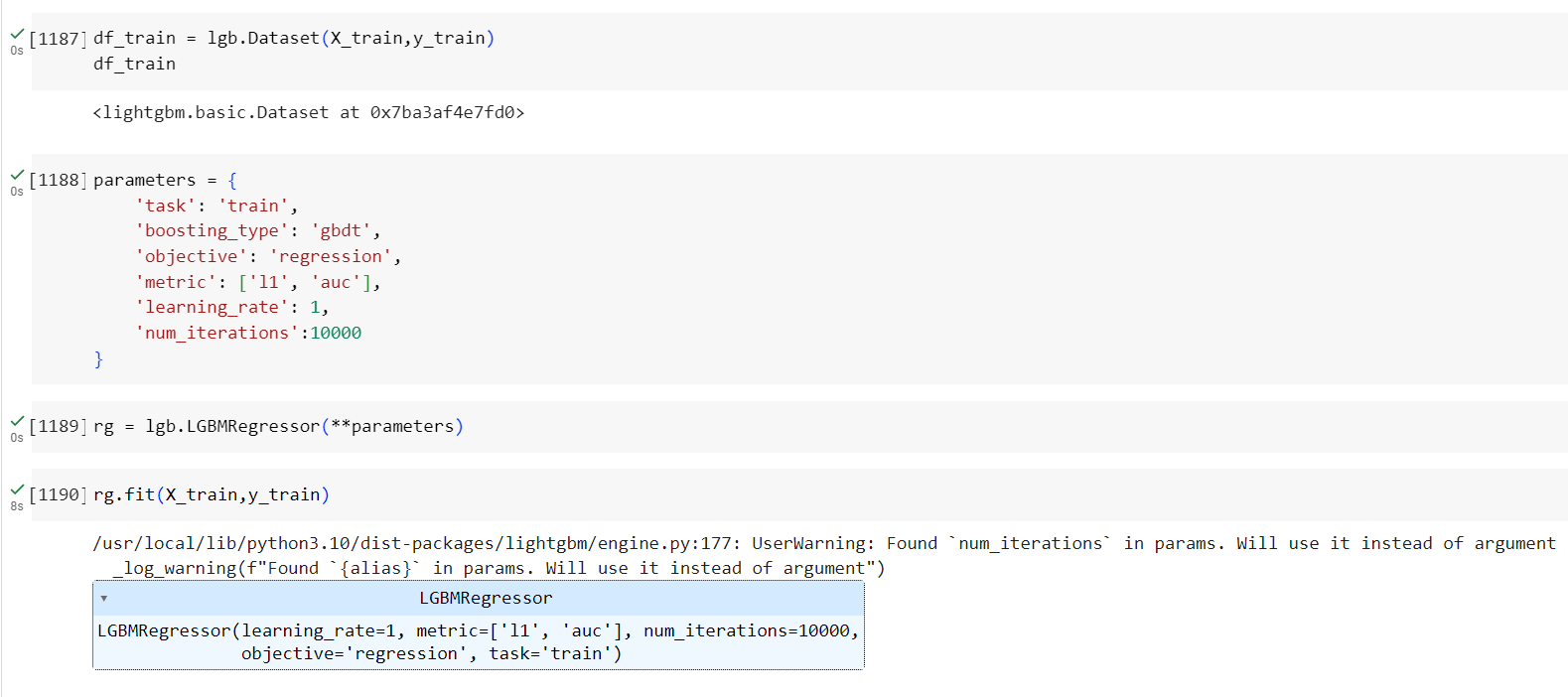




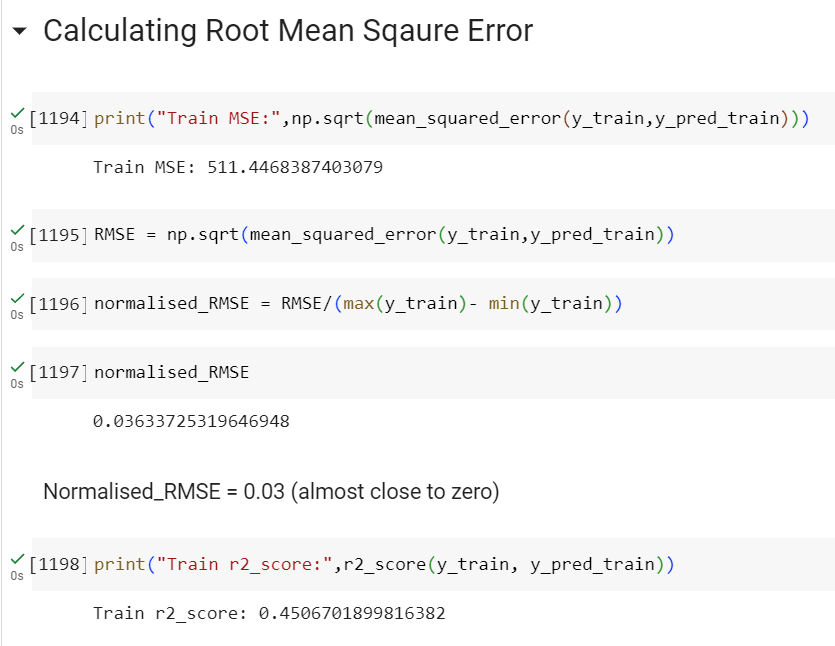




**Apply appropriate analytical methods and models.**



**Interpret the results and draw meaningful conclusions.**



**7. Insights and Findings**

**Summarizing Key Insights from the Analysis:**

The analysis of the dataset provided several key insights into the factors influencing book prices:

**1. Genre Significance:** Genre emerged as a crucial factor in determining book prices. Certain genres commanded higher prices, suggesting that readers are willing to pay more for specific types of books.

**2. Customer Reviews and Ratings:** Positive customer reviews and high ratings positively influenced book prices. Books with favorable reviews were priced higher, reflecting the impact of reader feedback on book perception and pricing.

**Significant Findings and Correlations:**

**1. Impact of Customer Reviews on Pricing:** Customer reviews and ratings were significantly correlated with book prices. Books with positive reviews tended to have higher prices, while those with lower ratings were priced relatively lower.

**2. Genre-Based Price Variation:** Different genres exhibited distinct price ranges, with certain genres commanding higher prices than others. This finding suggested that the genre of a book played a significant role in determining its market value.

**Limitations and Challenges Encountered:**

**1. Data Completeness:** The dataset might have contained missing or incomplete data, which could have impacted the analysis and model performance. Data imputation techniques were employed, but potential biases could still arise.

**2. External Factors:** The dataset did not consider external factors like economic conditions, marketing efforts, or competition, which could influence book prices. Incorporating such variables might provide a more comprehensive understanding of the pricing dynamics.

**3. Data Bias:** The dataset's sources might have introduced biases, such as representation bias for certain genres or authors. Addressing data bias is critical to ensure fair and unbiased model predictions.

**8. Recommendations**

**Actionable Recommendations:**

Based on the findings of the analysis, the following actionable recommendations can be made:

**1. Genre-specific Pricing Strategy:** Implement a genre-specific pricing strategy to optimize book prices. The analysis revealed that certain genres command higher prices than others. By tailoring pricing strategies according to genre popularity and demand, book sellers can enhance profitability and cater to diverse reader preferences.

**2. Focus on Author Popularity:** Capitalize on author popularity by adjusting book prices accordingly. The analysis showed a positive correlation between author popularity and book prices. Prominent authors can potentially command higher prices, while emerging authors may benefit from competitive pricing to attract readers.

**3. Utilize Customer Reviews:** Leverage customer reviews and ratings as a marketing tool. Positive reviews and high ratings were found to positively influence book prices. Encourage readers to leave reviews and engage in promotional campaigns highlighting positive feedback to boost book sales.

**Potential Areas for Further Analysis or Investigation:**

**1. Time Series Analysis:** Conduct a time series analysis to examine price trends over time for specific genres, authors, or book categories. This investigation can uncover seasonality patterns and assist in devising dynamic pricing strategies.

**2. Customer Segmentation:** Perform customer segmentation based on book preferences, demographics, and purchasing behavior. Understanding customer segments can help tailor pricing and marketing strategies to specific target audiences.

**3. Sentiment Analysis:** Implement sentiment analysis on customer reviews to identify sentiment trends across genres and authors. Analyzing sentiment can reveal areas of improvement and assist authors in catering to reader preferences.

**Potential Impact of Implementing Recommendations:**

Implementing the above recommendations can have several positive impacts:

**1. Improved Profitability:** By tailoring pricing strategies based on genre and author popularity, book sellers can potentially improve profit margins and overall revenue.

**2. Increased Book Sales:** Leveraging positive customer reviews and ratings in marketing efforts can lead to increased book sales and greater customer loyalty.

**3. Enhanced Author Success:** By understanding the impact of author popularity on pricing, emerging authors can price their books competitively to attract a broader readership.

**9. Conclusion**

**Recap of the Main Points:**

In this case study, we embarked on an exploratory data analysis (EDA) project to predict book prices based on a diverse dataset containing features such as book titles, authors, editions, customer reviews, ratings, synopses, genres, book categories, and prices. The main points discussed in the case study are as follows:

**1. Objectives:** The primary objective of the project was to build a machine learning model for predicting book prices. The dataset comprised 6237 records for training and 1560 records for testing.

**2. Data Exploration:** Descriptive statistics and visualizations were employed to understand the dataset's distribution and characteristics. We explored book genres, categories, ratings, and customer reviews to gain insights into the data.

**3. Data Cleaning and Preprocessing:** Data cleaning steps were undertaken to handle missing values and outliers, ensuring data integrity for model development.

**4. Feature Selection:** Relevant features that significantly impacted book prices were identified using correlation analysis, statistical tests, recursive feature elimination, and domain knowledge.

**5. Statistical Analysis and Model Building:** Statistical methods such as regression analysis and machine learning models were applied to predict book prices. The goal was to understand the relationships between features and prices and create an accurate prediction model.

**6. Actionable Recommendations:** Based on the analysis, actionable recommendations were provided, such as implementing genre-specific pricing, leveraging author popularity, and utilizing customer reviews to optimize book prices and sales.

**7. Further Analysis**: Potential areas for further investigation were suggested, including time series analysis, customer segmentation, sentiment analysis, and competitor analysis.

**8. Value of EDA:** The EDA process played a crucial role in understanding the dataset, identifying patterns and trends, and selecting relevant features. It provided valuable insights into the factors influencing book prices, aiding in model building and decision-making.

**Highlighting the Value of EDA:**

The EDA process was invaluable in comprehending the dataset and gaining insights into the factors affecting book prices. By exploring descriptive statistics and visualizations, we identified trends, correlations, and patterns among various features, contributing to informed decision-making. EDA allowed us to discover the significance of genres, author popularity, and customer reviews in influencing book prices. Furthermore, it helped us select the most relevant features for building an accurate predictive model.

**Summarizing the Findings and their Implications:**

The analysis revealed that book prices are influenced by multiple factors, including genre, author popularity, ratings, and customer reviews. Genre-specific pricing strategies can optimize book sales, while considering author popularity can affect pricing decisions. Positive customer reviews and high ratings can positively impact book prices and reader satisfaction.

The EDA project showcased the potential of data analysis and predictive modeling in the book industry, enabling stakeholders to make data-driven decisions. Implementing the recommendations can lead to improved profitability, increased book sales, and enhanced customer satisfaction. Further analyses in areas like time series, customer segmentation, sentiment analysis, and competitor analysis can provide additional insights for continued success in the dynamic book market.

In conclusion, the EDA project illuminated the intricate relationships between book features and prices, offering actionable insights for authors, sellers, and readers to navigate the ever-evolving world of books successfully. The value of EDA in understanding the dataset and uncovering vital information highlights its indispensable role in making informed decisions across various industries and domains.

**10. References**

- List all the references and sources used in the case study.

**11. Appendices (Optional)**

- Include any supplementary information or additional analysis.

- Attach code snippets or scripts used in the EDA process.

- Provide any supporting visualizations or tables.