Mehrdad Baradaran
DataScience Assignment03
Book Price Dataset

Report 03 (Book Price)
Prof. MR Kheradpishe
Shahid Beheshti University

Analysis of Book Price Dataset

Author: Mehrdad Baradaran

Abstract

In this task, a large dataset of books belonging to different genres and written by various authors was used. The dataset provided information on multiple book features, including the author, edition, and reviews. These features were used as predictors to forecast the price of books.

Introduction

In the dynamic landscape of data science, the ability to predict and understand complex patterns is paramount. As part of the third assignment for the CS SBU Data Science course, this competition on Kaggle invites participants to delve into the fascinating world of book price prediction. Books, as gateways to knowledge and entertainment, play a pivotal role in our lives. Accurately forecasting their prices is not only an intellectual challenge but also holds practical implications for publishers, retailers, and avid readers alike.

The dataset at the heart of this competition encompasses a rich array of features, each contributing to the intricate tapestry of a book's pricing dynamics. From fundamental details such as title and author to nuanced aspects like reviews, ratings, and genre, the dataset provides a comprehensive glimpse into the factors influencing the market value of a book. Through the lens of data science, participants are tasked with unraveling the patterns within this information to construct models that can predict book prices with precision.

Data Overview

The dataset for this competition comprises key features that encapsulate various facets of a book's identity and reception. Here's a brief overview of the primary features:

- Title: The title of the book, serving as its unique identifier.
- Author: The author(s) of the book, a crucial factor influencing its perceived value.
- Edition: Details about the specific edition of the book, which can impact its rarity and desirability.
- Reviews: User reviews provide insights into the reception and popularity of a book.
- Ratings: The overall ratings assigned by readers, contributing to the book's perceived quality.
- Synopsis: A brief summary or synopsis of the book's content, offering a glimpse into its themes and narrative.
- Genre: The genre or category to which the book belongs, influencing its target audience and market.
- BookCategory: The overarching category under which the book falls, providing additional context to its nature.
- Price: The target variable for prediction, representing the market price of the book.

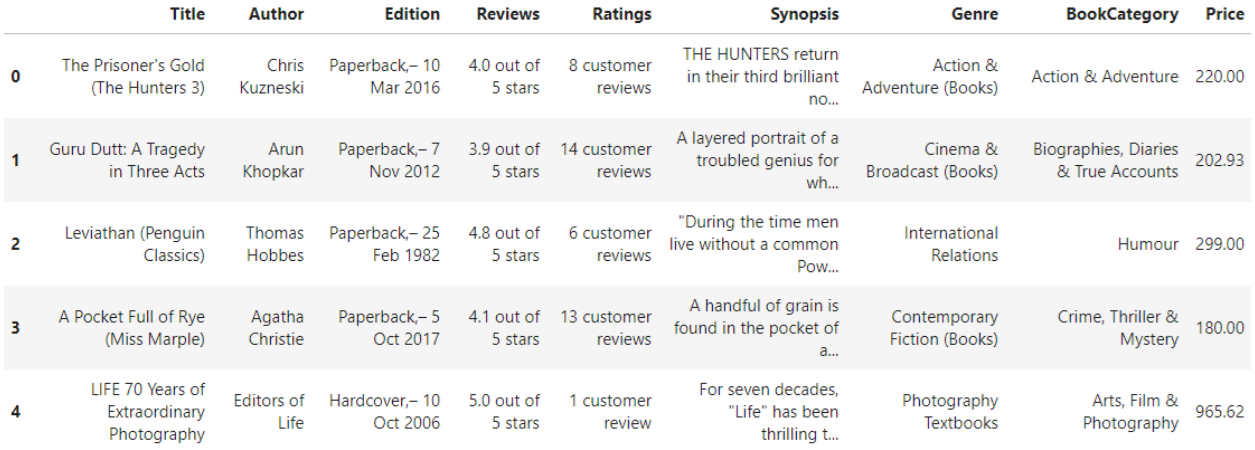
By leveraging these diverse features, participants are encouraged to apply advanced data science techniques to uncover hidden correlations, patterns, and insights that can enhance the accuracy of book price predictions.

Methodology

Some of the key methods which were used throughout the work are:

- Visualization
- TF-IDF and LDA Topic Extraction
- Text-tranlsation using Google Trasnlate Ajax API
- Cyclical feature encoding for time-based feature extraction
- Price Prediction using RandomForestRegressor from sikit-learn

The preliminary step in comprehending the dataset's structural attributes involves examining its initial rows. This exploratory analysis, often initiated by inspecting the head of the dataset, offers a concise overview of the available features and their respective values. This process aids in forming a foundational understanding of the dataset's composition, facilitating subsequent analytical procedures.



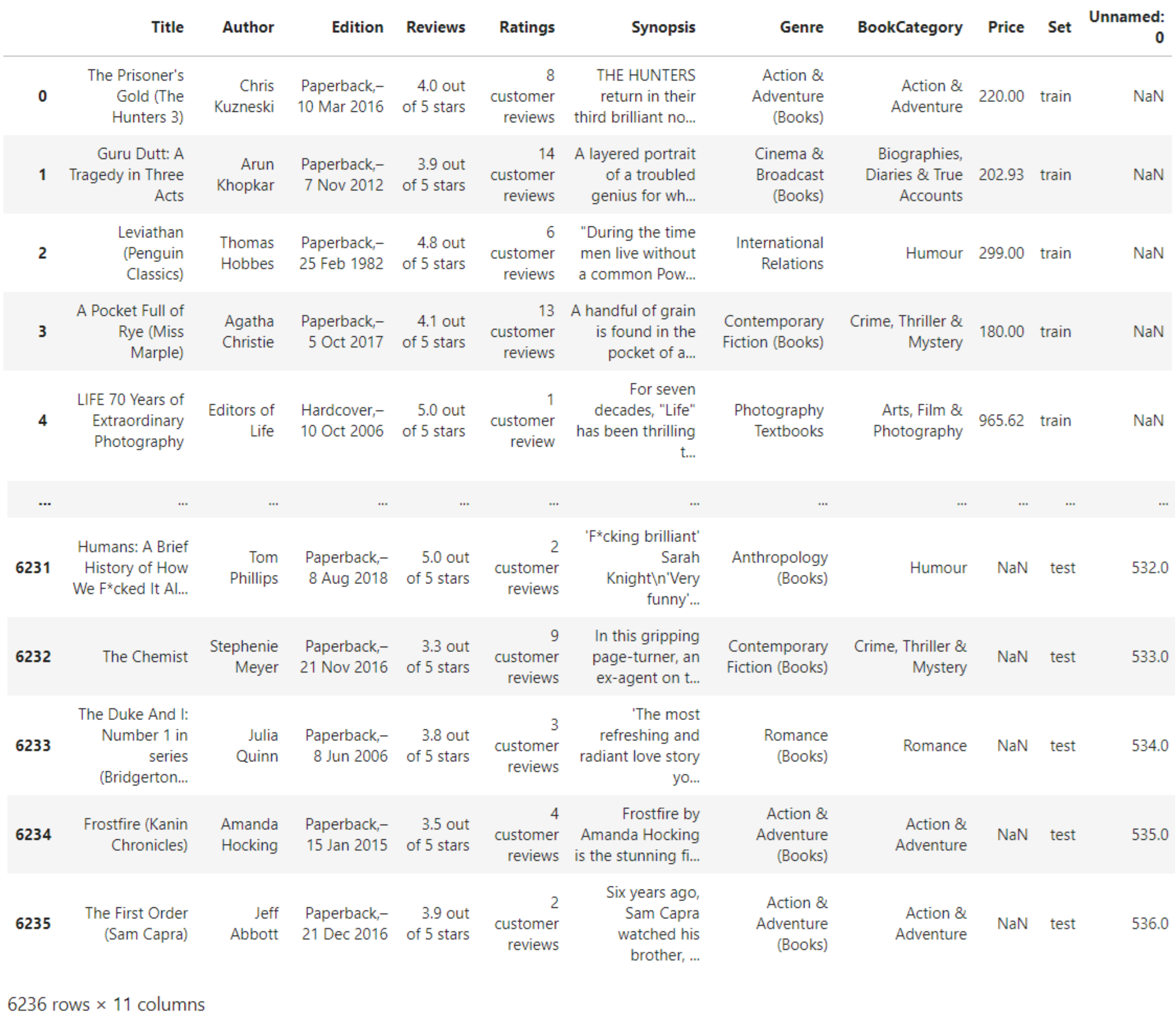
	Title	Author	Edition	Reviews	Ratings	Synopsis	Genre	BookCategory	Price
0	The Prisoner's Gold (The Hunters 3)	Chris Kuzneski	Paperback,– 10 Mar 2016	4.0 out of 5 stars	8 customer reviews	THE HUNTERS return in their third brilliant no...	Action & Adventure (Books)	Action & Adventure	220.00
1	Guru Dutt: A Tragedy in Three Acts	Arun Khopkar	Paperback,– 7 Nov 2012	3.9 out of 5 stars	14 customer reviews	A layered portrait of a troubled genius for wh...	Cinema & Broadcast (Books)	Biographies, Diaries & True Accounts	202.93
2	Leviathan (Penguin Classics)	Thomas Hobbes	Paperback,– 25 Feb 1982	4.8 out of 5 stars	6 customer reviews	"During the time men live without a common Pow...	International Relations	Humour	299.00
3	A Pocket Full of Rye (Miss Marple)	Agatha Christie	Paperback,– 5 Oct 2017	4.1 out of 5 stars	13 customer reviews	A handful of grain is found in the pocket of a...	Contemporary Fiction (Books)	Crime, Thriller & Mystery	180.00
4	LIFE 70 Years of Extraordinary Photography	Editors of Life	Hardcover,– 10 Oct 2006	5.0 out of 5 stars	1 customer review	For seven decades, "Life" has been thrilling t...	Photography Textbooks	Arts, Film & Photography	965.62

The integrity of the dataset is of paramount importance in ensuring the robustness of any analytical endeavor. In this context, it is noteworthy that the dataset under consideration has undergone rigorous quality checks, revealing the absence of any duplicate rows. The meticulous curation of this dataset, free from redundancy, underscores our commitment to providing participants with a pristine and unbiased foundation for their analyses.

In adherence to best practices in data science, the verification process confirmed that each entry is unique, eliminating concerns related to duplicated information. This assurance not only instills confidence in the dataset's reliability but also reflects our dedication to fostering transparency and precision in the exploration of book price prediction. As we embark on this academic journey, the absence of duplications serves as a testament to the meticulousness applied to every facet of this competition.

The next phase of our analytical journey involves the amalgamation of the training and testing datasets. To facilitate this unification, a new column has been introduced, meticulously indicating the origin of each observation as either belonging to the training or testing set. This strategic augmentation is paramount for preserving the identity of each data point while seamlessly merging the two distinct datasets.

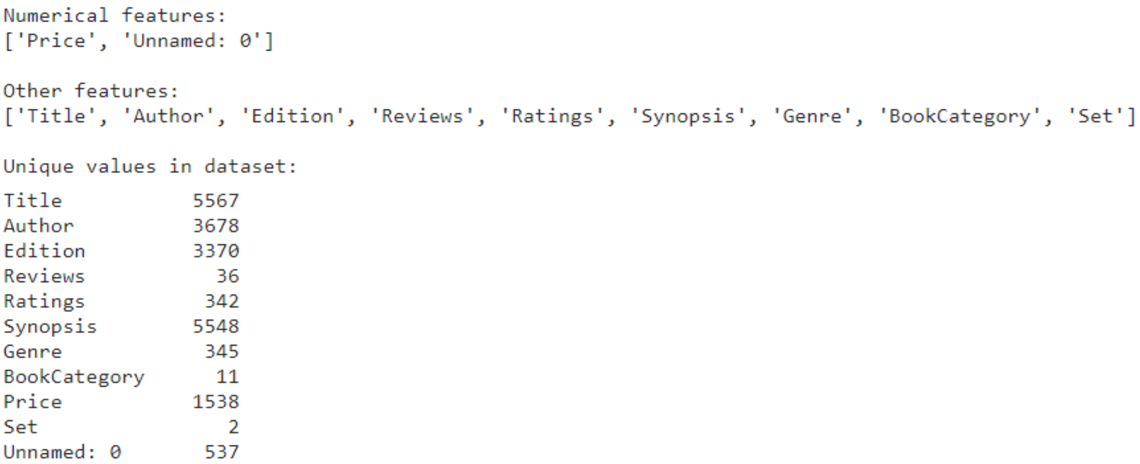
With the datasets harmoniously merged and the new column in place, the subsequent steps involve a suite of preprocessing methods. These methodologies are instrumental in refining the dataset, ensuring it aligns optimally with the requirements of diverse machine learning models. Techniques such as handling missing values, encoding categorical variables, and scaling numerical features are meticulously applied to fortify the dataset's suitability for subsequent modeling endeavors.



	Title	Author	Edition	Reviews	Ratings	Synopsis	Genre	BookCategory	Price	Set	Unnamed: 0
0	The Prisoner's Gold (The Hunters 3)	Chris Kuzneski	Paperback,– 10 Mar 2016	4.0 out of 5 stars	8 customer reviews	THE HUNTERS return in their third brilliant no...	Action & Adventure (Books)	Action & Adventure	220.00	train	NaN
1	Guru Dutt: A Tragedy in Three Acts	Arun Khopkar	Paperback,– 7 Nov 2012	3.9 out of 5 stars	14 customer reviews	A layered portrait of a troubled genius for wh...	Cinema & Broadcast (Books)	Biographies, Diaries & True Accounts	202.93	train	NaN
2	Leviathan (Penguin Classics)	Thomas Hobbes	Paperback,– 25 Feb 1982	4.8 out of 5 stars	6 customer reviews	"During the time men live without a common Pow...	International Relations	Humour	299.00	train	NaN
3	A Pocket Full of Rye (Miss Marple)	Agatha Christie	Paperback,– 5 Oct 2017	4.1 out of 5 stars	13 customer reviews	A handful of grain is found in the pocket of a...	Contemporary Fiction (Books)	Crime, Thriller & Mystery	180.00	train	NaN
4	LIFE 70 Years of Extraordinary Photography	Editors of Life	Hardcover,– 10 Oct 2006	5.0 out of 5 stars	1 customer review	For seven decades, "Life" has been thrilling t...	Photography Textbooks	Arts, Film & Photography	965.62	train	NaN
...	...	...	...	...	...	...	...	...	...	...	...
6231	Humans: A Brief History of How We F*cked It Al...	Tom Phillips	Paperback,– 8 Aug 2018	5.0 out of 5 stars	2 customer reviews	'F*cking brilliant' Sarah Knight\n'Very funny'...	Anthropology (Books)	Humour	NaN	test	532.0
6232	The Chemist	Stephenie Meyer	Paperback,– 21 Nov 2016	3.3 out of 5 stars	9 customer reviews	In this gripping page-turner, an ex-agent on t...	Contemporary Fiction (Books)	Crime, Thriller & Mystery	NaN	test	533.0
6233	The Duke And I: Number 1 in series (Bridgerton...	Julia Quinn	Paperback,– 8 Jun 2006	3.8 out of 5 stars	3 customer reviews	'The most refreshing and radiant love story yo...	Romance (Books)	Romance	NaN	test	534.0
6234	Frostfire (Kanin Chronicles)	Amanda Hocking	Paperback,– 15 Jan 2015	3.5 out of 5 stars	4 customer reviews	Frostfire by Amanda Hocking is the stunning fi...	Action & Adventure (Books)	Action & Adventure	NaN	test	535.0
6235	The First Order (Sam Capra)	Jeff Abbott	Paperback,– 21 Dec 2016	3.9 out of 5 stars	2 customer reviews	Six years ago, Sam Capra watched his brother, ...	Action & Adventure (Books)	Action & Adventure	NaN	test	536.0

6236 rows × 11 columns

In the preliminary stage, an examination is conducted to discern the various types of features or variables in the dataset.



```
Numerical features:
['Price', 'Unnamed: 0']

Other features:
['Title', 'Author', 'Edition', 'Reviews', 'Ratings', 'Synopsis', 'Genre', 'BookCategory', 'Set']

Unique values in dataset:
Title           5567
Author          3678
Edition         3370
Reviews           36
Ratings          342
Synopsis        5548
Genre            345
BookCategory      11
Price           1538
Set                2
Unnamed: 0       537
```

Observations reveal that while there are several features with numerical values, such as reviews and ratings, the distinction lies in the fact that only the 'price' feature is explicitly designated as a numerical variable.

Author Preprocessing

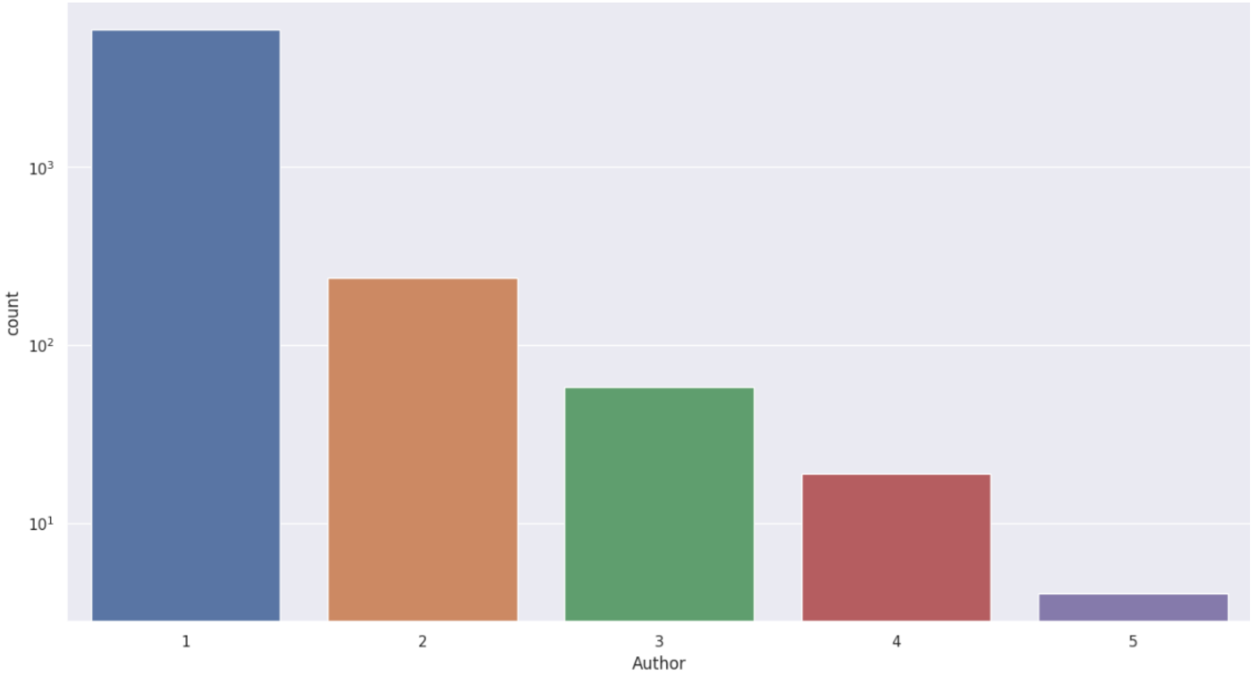
The preprocessing steps for the 'Authors' column involve addressing irregularities in the data structure. Notably, in instances where a book has multiple authors, the author names are delimited by various separators such as ',', '&', '/', or ';'. To ensure uniformity, these diverse separators have been replaced with a consistent comma.

Additionally, some author names contained unnecessary punctuation marks, and there was inconsistency in the letter case. Consequently, superfluous punctuation marks have been removed, and all author names have been standardized to lowercase. This systematic preprocessing ensures a cohesive and standardized representation of author names, facilitating a more uniform and analytically robust dataset.

There is a reduction of 23 unique authors, it means that the dataset now contains 23 fewer distinct author names compared to the original dataset. This clarification suggests that the preprocessing steps have contributed to consolidating author names and potentially resolving inconsistencies in the data.

Number of Authors (New Feature)

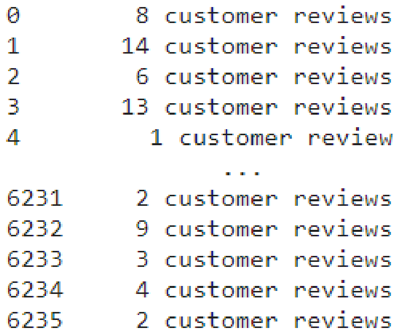
The dataset has been enriched with an insightful meta-feature: the "Number of Authors." This additional feature augments the existing information by quantifying the count of authors associated with each book. The calculation of this meta-feature provides a nuanced perspective, allowing for a more detailed exploration of books with single or multiple contributors. This enhancement sets the stage for a more comprehensive analysis, considering the collaborative nature of certain literary works and its potential impact on various aspects, including pricing dynamics and reader reception.



Rating Preprocessing

The preprocessing of the 'Rating' column was imperative due to the inclusion of text instead of the actual numerical ratings. The focus of this preprocessing step was to extract only the numeric values, aligning the data with the expected numerical format.

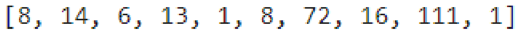
Moreover, it's noteworthy that the 'Rating' column does not directly represent the ratings of the books but rather the count of customers who reviewed each book.



```
0        8 customer reviews
1       14 customer reviews
2        6 customer reviews
3       13 customer reviews
4         1 customer review
               ...
6231     2 customer reviews
6232     9 customer reviews
6233     3 customer reviews
6234     4 customer reviews
6235     2 customer reviews
```

Before proceeding with further analysis, it is prudent to conduct an initial check for any anomalies or inconsistencies within the 'Rating' column. This entails scrutinizing the elements of the list to identify potential issues, such as non-numeric ratings or inconsistencies in the data format.

Then we Extract the actual numbers from the columns using regex



```
[8, 14, 6, 13, 1, 8, 72, 16, 111, 1]
```

Review Preprocessing

In accordance with the nuanced characteristics of the dataset, it is imperative to address the intricacies within the 'Reviews' column. This particular attribute, reflective of the average review rating of a book, inherently spans a qualitative scale from 0 to 5 stars. However, akin to the preceding preprocessing measures undertaken for the 'Ratings' column, the 'Reviews' column necessitates meticulous treatment to extract the evaluative scores from the textual representations.

Then we apply the same things that we do it for rating.

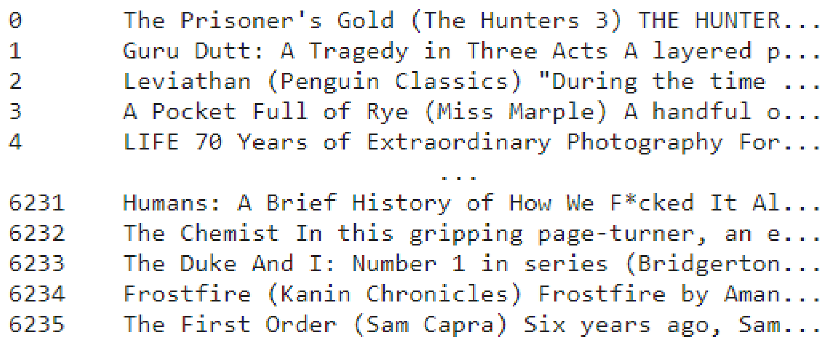
Title/Synopsis Preprocessing

The preprocessing workflow extends to the 'Title' and 'Synopsis' columns, which are anticipated to contain textual information pertaining to the titles and synopses of various books. A noteworthy challenge emerged due to the presence of non-English text in certain instances, attributable to the dataset's origin from an Indian website.

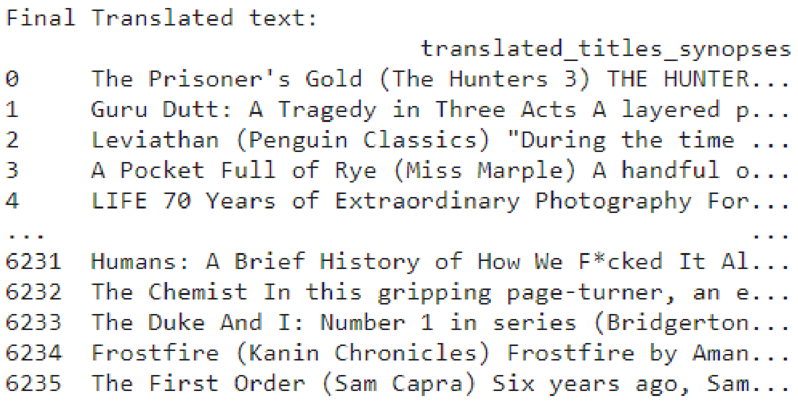
In response, a crucial preprocessing step was undertaken to ensure linguistic consistency. Books featuring non-English text underwent translation procedures to bring their content into English.

In the pursuit of linguistic homogenization within the 'Title' and 'Synopsis' columns, the Googletrans library emerged as a pivotal tool for effecting translations. Leveraging this library, the non-English textual content encountered in select books was systematically translated to English.

A strategic decision was made to consolidate the textual components within the dataset by merging both the 'Title' and 'Synopsis' columns. This amalgamation was undertaken with the overarching goal of fostering a comprehensive analysis and processing of the entirety of the textual data.



```
0       The Prisoner's Gold (The Hunters 3) THE HUNTER...
1       Guru Dutt: A Tragedy in Three Acts A layered p...
2       Leviathan (Penguin Classics) "During the time ...
3       A Pocket Full of Rye (Miss Marple) A handful o...
4       LIFE 70 Years of Extraordinary Photography For...
                               ...
6231    Humans: A Brief History of How We F*cked It Al...
6232    The Chemist In this gripping page-turner, an e...
6233    The Duke And I: Number 1 in series (Bridgerton...
6234    Frostfire (Kanin Chronicles) Frostfire by Aman...
6235    The First Order (Sam Capra) Six years ago, Sam...
```

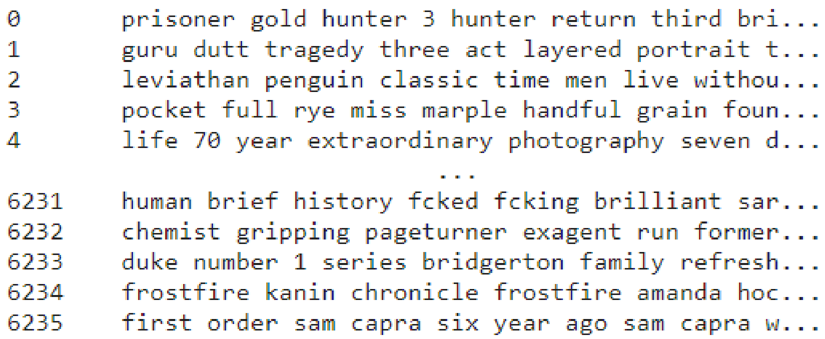


```
Final Translated text:
                              translated_titles_synopses
0     The Prisoner's Gold (The Hunters 3) THE HUNTER...
1     Guru Dutt: A Tragedy in Three Acts A layered p...
2     Leviathan (Penguin Classics) "During the time ...
3     A Pocket Full of Rye (Miss Marple) A handful o...
4     LIFE 70 Years of Extraordinary Photography For...
...                                                 ...
6231  Humans: A Brief History of How We F*cked It Al...
6232  The Chemist In this gripping page-turner, an e...
6233  The Duke And I: Number 1 in series (Bridgerton...
6234  Frostfire (Kanin Chronicles) Frostfire by Aman...
6235  The First Order (Sam Capra) Six years ago, Sam...
```

The text underwent a systematic preprocessing regimen to enhance its suitability for analytical endeavors. The sequence of operations included:

1. **HTML Code Removal:**
    - The elimination of potential HTML code aimed to purify the text by stripping away extraneous markup or formatting artifacts, ensuring the isolation of raw textual content.
2. **Contractions Expansion:**
    - Contraction expansion was executed to rectify and standardize colloquial contractions within the text. This process involved converting contracted forms into their complete expressions, fostering a uniform linguistic representation.
3. **Punctuation Removal:**
    - The exclusion of punctuation marks sought to refine the text by excising non-alphabetic characters. This deliberate omission contributed to a more focused analysis by emphasizing the core lexical components.
4. **Stop Words Elimination:**
    - Removal of stop words, common linguistic connectors with limited semantic value, was conducted to alleviate the impact of linguistic noise. This curation process aimed to enhance the discernibility of meaningful terms within the text.
5. **Lemmatization:**
    - The lemmatization procedure involved reducing words to their base or root form, harmonizing inflected variations. This standardized representation facilitates a more cohesive understanding of the underlying semantic structures within the text.

Each of these methodical preprocessing steps was tailored to refine and standardize the textual data, aligning it with the requisites of subsequent analytical or modeling tasks, particularly within the domain of natural language processing.



```
0       prisoner gold hunter 3 hunter return third bri...
1       guru dutt tragedy three act layered portrait t...
2       leviathan penguin classic time men live withou...
3       pocket full rye miss marple handful grain foun...
4       life 70 year extraordinary photography seven d...
                               ...
6231    human brief history fcked fcking brilliant sar...
6232    chemist gripping pageturner exagent run former...
6233    duke number 1 series bridgerton family refresh...
6234    frostfire kanin chronicle frostfire amanda hoc...
6235    first order sam capra six year ago sam capra w...
```

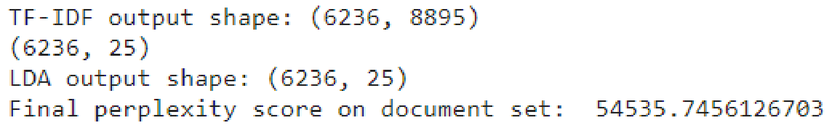
LDA (Latent Dirichlet Allocation)

To make sense of a large amount of text data, we often convert it into features. Traditional methods, like bag-of-words or TF-IDF, can result in a lot of features. To handle this, we explored an alternative called Topic Modeling, specifically using Latent Dirichlet Allocation (LDA).

Unlike simple feature methods, LDA doesn't just create clusters; it uncovers hidden topics within the text. It assumes that each document is made up of different topics, and each topic is composed of specific words. LDA works by figuring out these topics and how they contribute to each document.

The unique aspect of LDA is its ability to show, for each document, the topics it contains and the percentage of each. For example, in a manga about a Japanese ninja story, LDA might reveal a mix like 48% Manga, 31% Ninja, and 21% Japanese.

LDA usually works with TF-IDF values. To use LDA, we first encode the text into a TF-IDF matrix, making it compatible with the algorithm. This way, we can uncover meaningful topics without drowning in an excessive number of features.

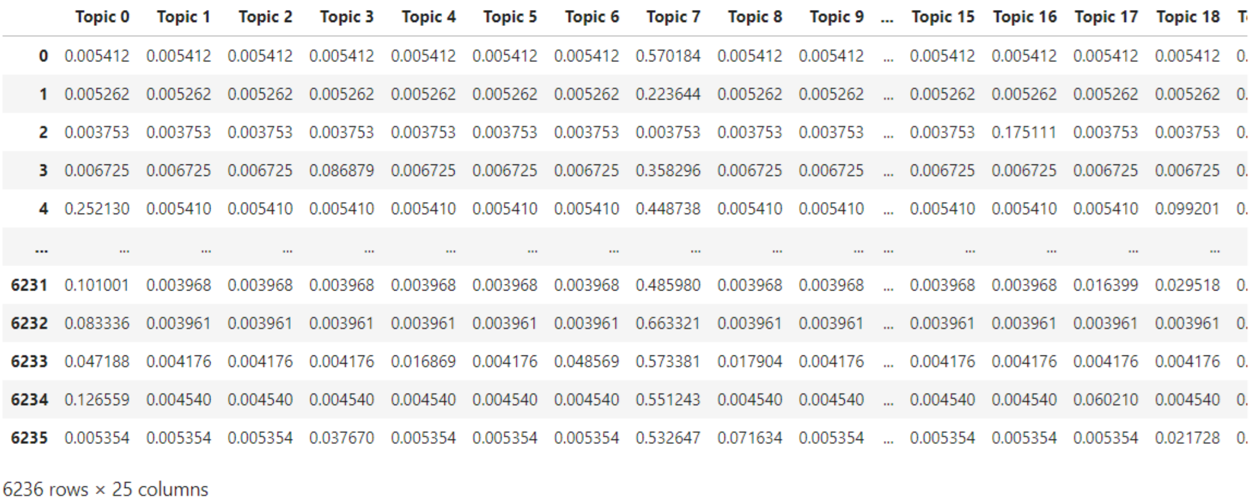


```
TF-IDF output shape: (6236, 8895)
(6236, 25)
LDA output shape: (6236, 25)
Final perplexity score on document set:  54535.7456126703
```

The decision was made to utilize 25 topics instead of the default 10 in the sklearn Latent Dirichlet Allocation (LDA) package. Typically, a higher number of topics enhances the accuracy of topic extraction. In this specific context, the choice of 25 topics was motivated by the observation that these topics exhibited a manageable degree of heterogeneity. Opting for more than 25 topics would have introduced an excessive number of features for subsequent prediction algorithms.

Printing the percentages of each topic that the synopsis and title of the first book belong to provides insights into the thematic composition of the text.

Then we save the topics of each book to a different dataframe.



	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	...	Topic 15	Topic 16	Topic 17	Topic 18	T
0	0.005412	0.005412	0.005412	0.005412	0.005412	0.005412	0.005412	0.570184	0.005412	0.005412	...	0.005412	0.005412	0.005412	0.005412	0.
1	0.005262	0.005262	0.005262	0.005262	0.005262	0.005262	0.005262	0.223644	0.005262	0.005262	...	0.005262	0.005262	0.005262	0.005262	0.
2	0.003753	0.003753	0.003753	0.003753	0.003753	0.003753	0.003753	0.003753	0.003753	0.003753	...	0.003753	0.175111	0.003753	0.003753	0.
3	0.006725	0.006725	0.006725	0.086879	0.006725	0.006725	0.006725	0.358296	0.006725	0.006725	...	0.006725	0.006725	0.006725	0.006725	0.
4	0.252130	0.005410	0.005410	0.005410	0.005410	0.005410	0.005410	0.448738	0.005410	0.005410	...	0.005410	0.005410	0.005410	0.099201	0.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
6231	0.101001	0.003968	0.003968	0.003968	0.003968	0.003968	0.003968	0.485980	0.003968	0.003968	...	0.003968	0.003968	0.016399	0.029518	0.
6232	0.083336	0.003961	0.003961	0.003961	0.003961	0.003961	0.003961	0.663321	0.003961	0.003961	...	0.003961	0.003961	0.003961	0.003961	0.
6233	0.047188	0.004176	0.004176	0.004176	0.016869	0.004176	0.048569	0.573381	0.017904	0.004176	...	0.004176	0.004176	0.004176	0.004176	0.
6234	0.126559	0.004540	0.004540	0.004540	0.004540	0.004540	0.004540	0.551243	0.004540	0.004540	...	0.004540	0.004540	0.060210	0.004540	0.
6235	0.005354	0.005354	0.005354	0.037670	0.005354	0.005354	0.005354	0.532647	0.071634	0.005354	...	0.005354	0.005354	0.005354	0.021728	0.

6236 rows × 25 columns

Edition Preprocessing

The examination of the Edition column revealed a wealth of information encompassing various aspects, such as the type (e.g., paperback) and the release date. Consequently, to distill this rich information into distinct features, a systematic partitioning process was employed within this notebook.

The recursive procedure adopted facilitated the creation of several new feature columns, each catering to specific facets of the original Edition column:

1. **Print (Categorical - Single value):** Denoting the printing format, this categorical feature captures singular expressions such as 'Hardcover' or 'Paperback.'
2. **Type (Categorical - Multivalue):** Representing the edition type, this categorical feature accommodates multiple values, characterizing diverse attributes within the Edition column.
3. **Year (Numerical):** Extracting the numerical component denoting the year of release, this feature provides temporal information about each edition.
4. **Month (Numerical):** Similarly, this numerical feature captures the month of release, adding a temporal dimension to the dataset.

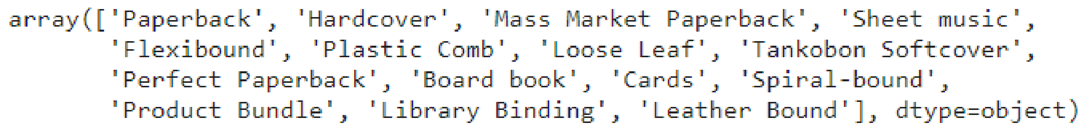
By implementing this recursive procedure, the notebook successfully disentangles the Edition column into these newly formed features, enhancing the granularity and interpretability of the dataset.

Extract language tag from edition

To enhance the dataset, an extraction of the language tag from the Edition column was performed. The language property, denoted by a language tag ('A'), was identified as the initial segment of the Edition text.

However, it was observed that this language property was present in only a limited number of rows, specifically four observations. Consequently, in the interest of maintaining dataset uniformity and addressing the scarcity of language property instances, the decision was made to remove all language tags from the Edition column.

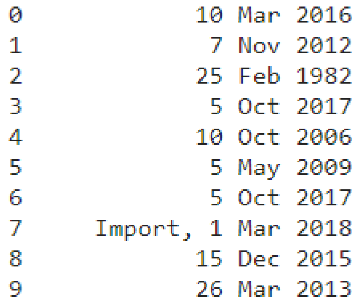
Extract print of edition



```
array(['Paperback', 'Hardcover', 'Mass Market Paperback', 'Sheet music',
       'Flexibound', 'Plastic Comb', 'Loose Leaf', 'Tankobon Softcover',
       'Perfect Paperback', 'Board book', 'Cards', 'Spiral-bound',
       'Product Bundle', 'Library Binding', 'Leather Bound'], dtype=object)
```

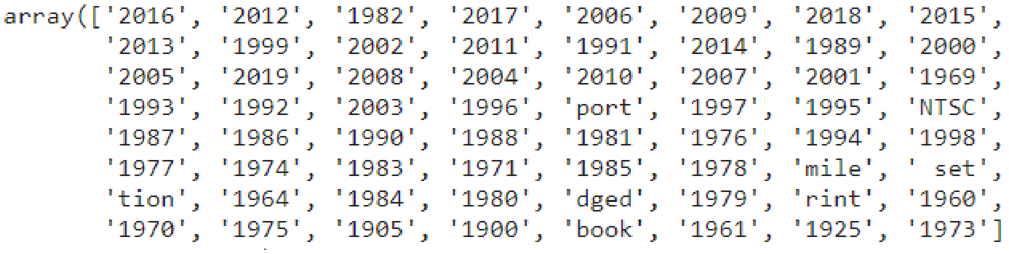
Extract the year from edition

add Codeadd Markdown
The next step involved the extraction of the publication year. However, a preliminary cleanup of the text was necessary to eliminate redundant or previously extracted information. Many rows contained punctuation marks, such as '–'. Given that the different sub-tags in the Edition column are separated by commas (','), the removal of other punctuation marks was deemed necessary for further processing. This ensures a consistent and streamlined approach to extracting the publication year from the Edition column.



```
0            10 Mar 2016
1             7 Nov 2012
2            25 Feb 1982
3             5 Oct 2017
4            10 Oct 2006
5             5 May 2009
6             5 Oct 2017
7     Import, 1 Mar 2018
8            15 Dec 2015
9            26 Mar 2013
```

Subsequently, the remaining information, encapsulated in the rest_edition_series Series object, became the focus of extracting the publication year. An essential consideration is that not all rows/books contain the year of the edition. Consequently, a marker was introduced to identify those books where the year of the edition is absent. This step ensures that the dataset retains information about the presence or absence of the publication year for each book.

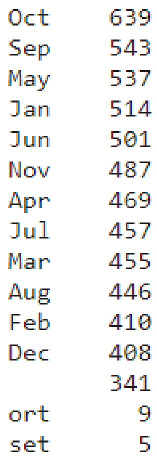


```
array(['2016', '2012', '1982', '2017', '2006', '2009', '2018', '2015',
       '2013', '1999', '2002', '2011', '1991', '2014', '1989', '2000',
       '2005', '2019', '2008', '2004', '2010', '2007', '2001', '1969',
       '1993', '1992', '2003', '1996', 'port', '1997', '1995', 'NTSC',
       '1987', '1986', '1990', '1988', '1981', '1976', '1994', '1998',
       '1977', '1974', '1983', '1971', '1985', '1978', 'mile', ' set',
       'tion', '1964', '1984', '1980', 'dged', '1979', 'rint', '1960',
       '1970', '1975', '1905', '1900', 'book', '1961', '1925', '1973']
```

Subsequent to the identification of rows/books lacking the year property, a dedicated marking process was implemented. A new series was constructed, associating each row with the corresponding observation's publication year. In instances where a row is devoid of any year information (as indicated by the marking), the series includes the 'NA' value. This construction ensures a comprehensive representation of the publication years for each observation, with due consideration for cases where the information is unavailable.

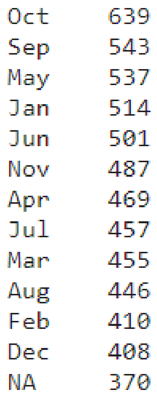
Extract the month of the edition

The subsequent task involved the extraction of the publication month. To achieve this, a preliminary cleanup of the text was undertaken to eliminate redundant or previously extracted information. Specifically, the remaining portion of the edition text was isolated, focusing on the segment that does not encompass the print or year information. This strategic isolation facilitates a more targeted approach to extracting the publication month from the Edition column.



```
Oct    639
Sep    543
May    537
Jan    514
Jun    501
Nov    487
Apr    469
Jul    457
Mar    455
Aug    446
Feb    410
Dec    408
       341
ort      9
set      5
```

After extracting the month, we marked rows or books that don't have this information. We created a series where each row corresponds to the observation's publication month. In cases where a row lacks month information, it's marked with 'NA'.



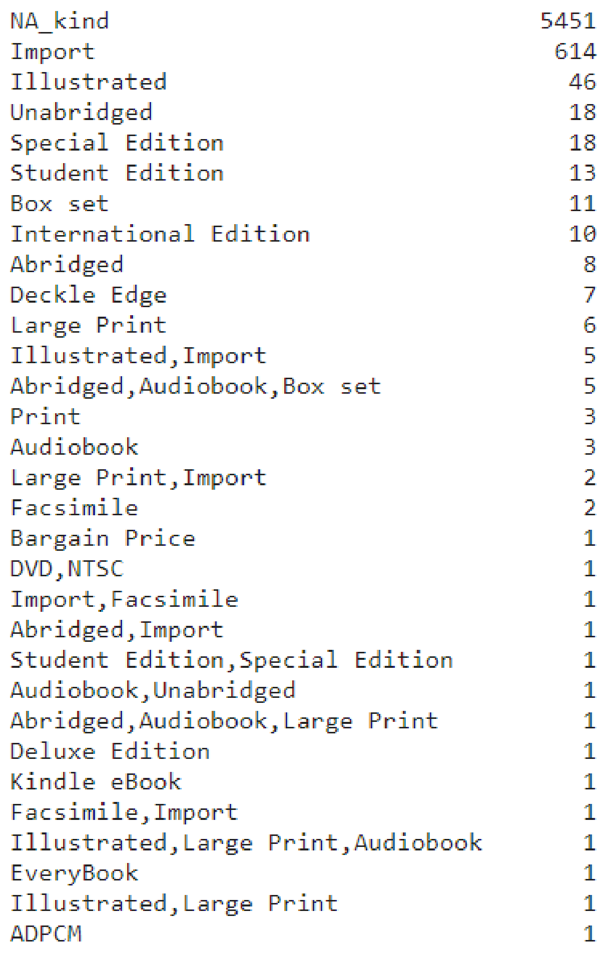
```
Oct    639
Sep    543
May    537
Jan    514
Jun    501
Nov    487
Apr    469
Jul    457
Mar    455
Aug    446
Feb    410
Dec    408
NA     370
```

Extracting Type of the edition

The subsequent task involved the extraction of the edition type. To accomplish this, a preliminary cleanup of the text was undertaken to eliminate redundant or previously extracted information. Precisely, the remaining portion of the edition text was isolated, focusing on the segment that does not encompass edition, year, or month information. This strategic isolation facilitates a more targeted approach to extracting the edition type from the Edition column.

Next, we refined the edition type information. First, we removed the day of the month property, as it didn't contribute much valuable information.

Additionally, recognizing that the type of print might have multiple tags (e.g., both imported and illustrated), the code was crafted to effectively split and capture these diverse properties within the Edition text. This ensures a more detailed representation of the edition type, accounting for various attributes associated with the print.



```
NA_kind                              5451
Import                                614
Illustrated                            46
Unabridged                             18
Special Edition                        18
Student Edition                        13
Box set                                11
International Edition                  10
Abridged                                8
Deckle Edge                             7
Large Print                             6
Illustrated,Import                      5
Abridged,Audiobook,Box set              5
Print                                   3
Audiobook                               3
Large Print,Import                      2
Facsimile                               2
Bargain Price                           1
DVD,NTSC                                1
Import,Facsimile                        1
Abridged,Import                         1
Student Edition,Special Edition         1
Audiobook,Unabridged                    1
Abridged,Audiobook,Large Print          1
Deluxe Edition                          1
Kindle eBook                            1
Facsimile,Import                        1
Illustrated,Large Print,Audiobook       1
EveryBook                               1
Illustrated,Large Print                 1
ADPCM                                   1
```

As confirmed, there are instances where books have multiple types included, such as both imported and illustrated.

It's important to note that the print_series and type_series will require further preprocessing after or during the analysis. These variables are categorical and may need special consideration or encoding techniques to effectively capture the nuances introduced by multiple types associated with the print.

Application of Feature Engineering for Dataset Enhancement

The applied feature engineering processes involve the enhancement of the dataset by modifying or introducing new features. This iterative refinement aims to improve the dataset's representational capacity and facilitate more effective modeling. The following steps were undertaken:

1. **Removal of Unnecessary Features:** Certain features such as 'Title,' 'Synopsis,' 'Author,' and 'Edition' were deemed redundant or unsuitable for modeling purposes and were consequently dropped from the dataset.

2. **Inclusion of Translated Titles and Synopses:** The translated titles and synopses, obtained through the Googletrans library, were considered as potential features. However, for the current implementation, these were replaced with clusters, as indicated by the comment.

3. **Transformation of Reviews and Ratings:** The original 'Reviews' and 'Ratings' features were processed to extract numerical values, providing more meaningful representations for subsequent analyses.

4. **Authors Column Enhancement:** The 'Authors' column underwent preprocessing to address inconsistencies, including variations in punctuation and letter case. Additionally, the number of authors for each book was calculated and introduced as a new feature.

5. **Edition Column Processing:** Information from the 'Edition' column was partitioned into distinct features, such as 'Print,' 'Type,' 'Month,' and 'Year,' providing more granular insights into the publication details.

6. **Incorporation of Topic Features:** Topic modeling using Latent Dirichlet Allocation (LDA) was applied, generating new features that represent the distribution of topics within the textual data.

The resulting enhanced dataset, denoted as **preprocessed_data_df**, reflects these modifications and additions, poised for subsequent modeling and analysis.

Fill missing values

In addressing the scarcity of missing values in the Year column, a practical approach was taken to enhance dataset completeness. Specifically, the missing values, constituting only a limited portion of the dataset, were imputed by employing the median value.

To handle missing values in the Month column, a probability-based approach was adopted. First, the probabilities of each month were calculated based on their frequencies in the dataset. Then, for each missing value, a month was assigned according to these probabilities.

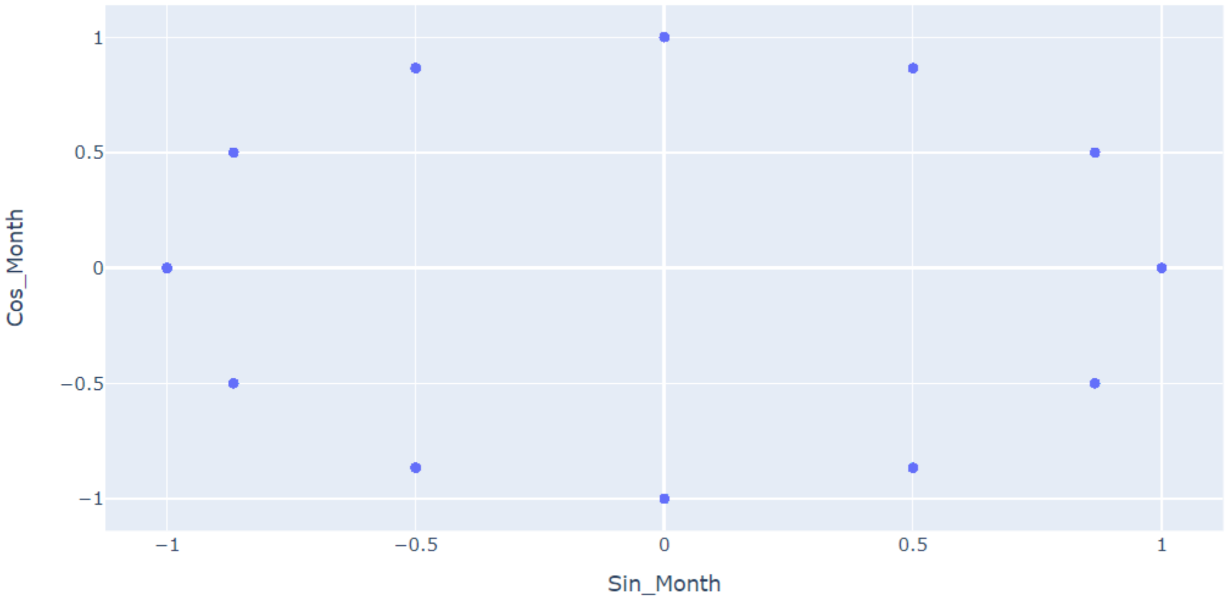
Cyclical Encoding for Month

Recognizing that the Month feature is a cyclical and temporal attribute, a specialized encoding technique was applied. Cyclical features, like months or days of the year, exhibit a cyclical pattern in their values and are treated as such through polar coordinate encoding.

In the polar representation, distinct values are assigned to each moment in time while preserving the cyclical similarities and differences inherent in temporal features. Following the encoding of months to numerical values (1 to 12), the cyclical-polar encoding introduces two new features, guided by the following formulas:

$$month_{\cos} = \cos\left(\frac{2\pi \times month}{\max(month)}\right)$$

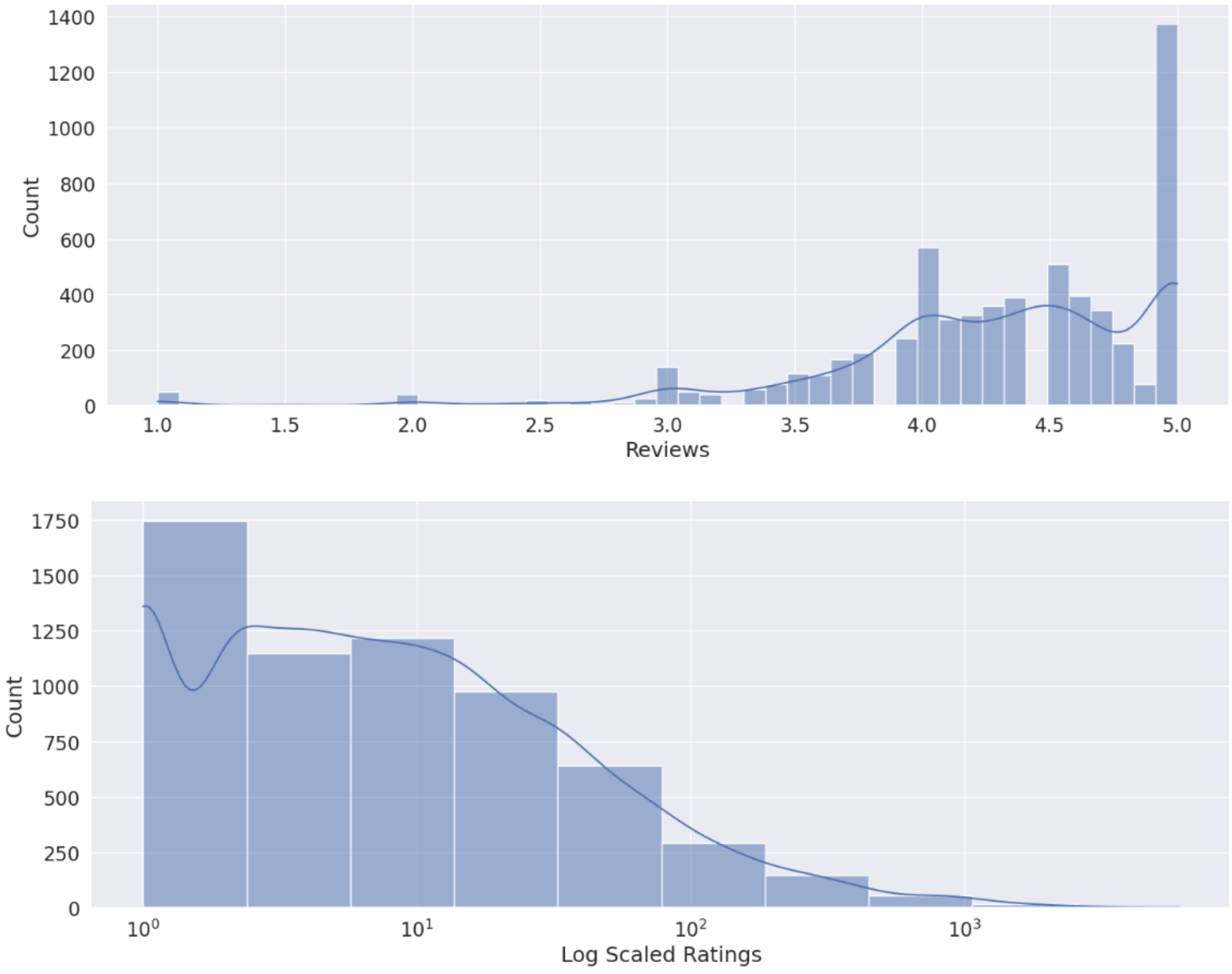
$$month_{\sin} = \sin\left(\frac{2\pi \times month}{\max(month)}\right)$$

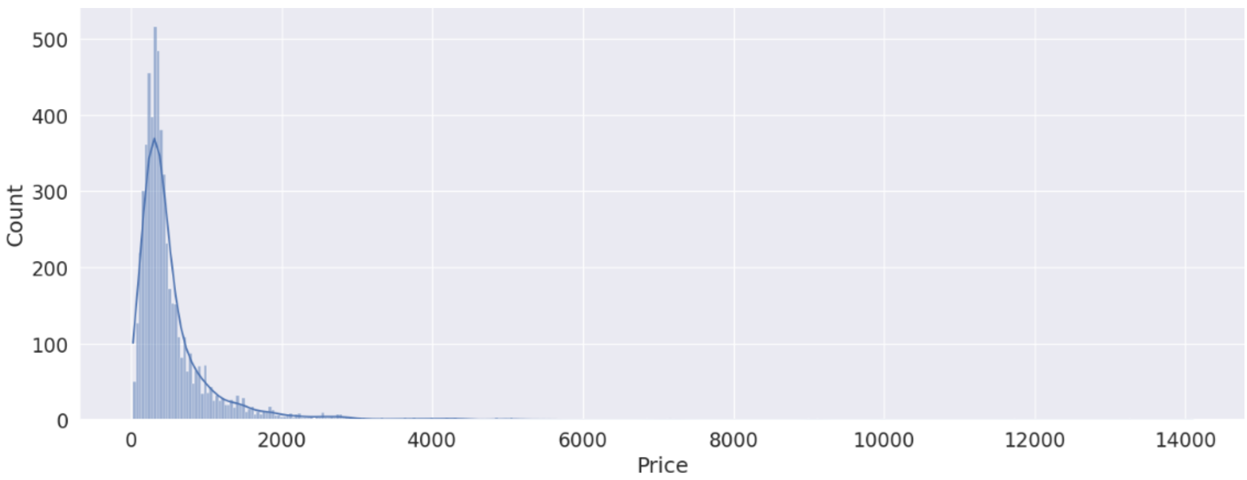


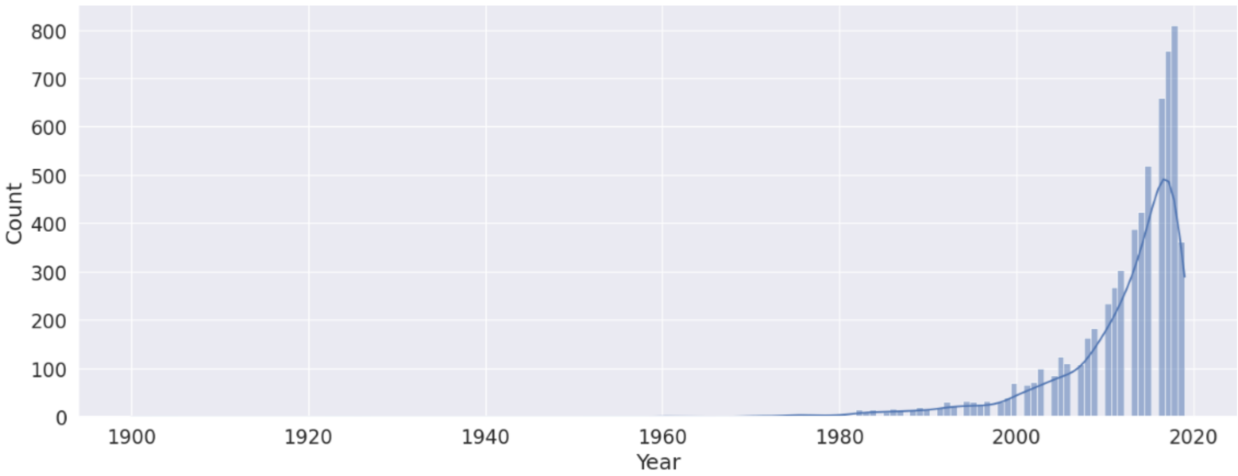
As it is shown, the two new features for the month align on the cycle.

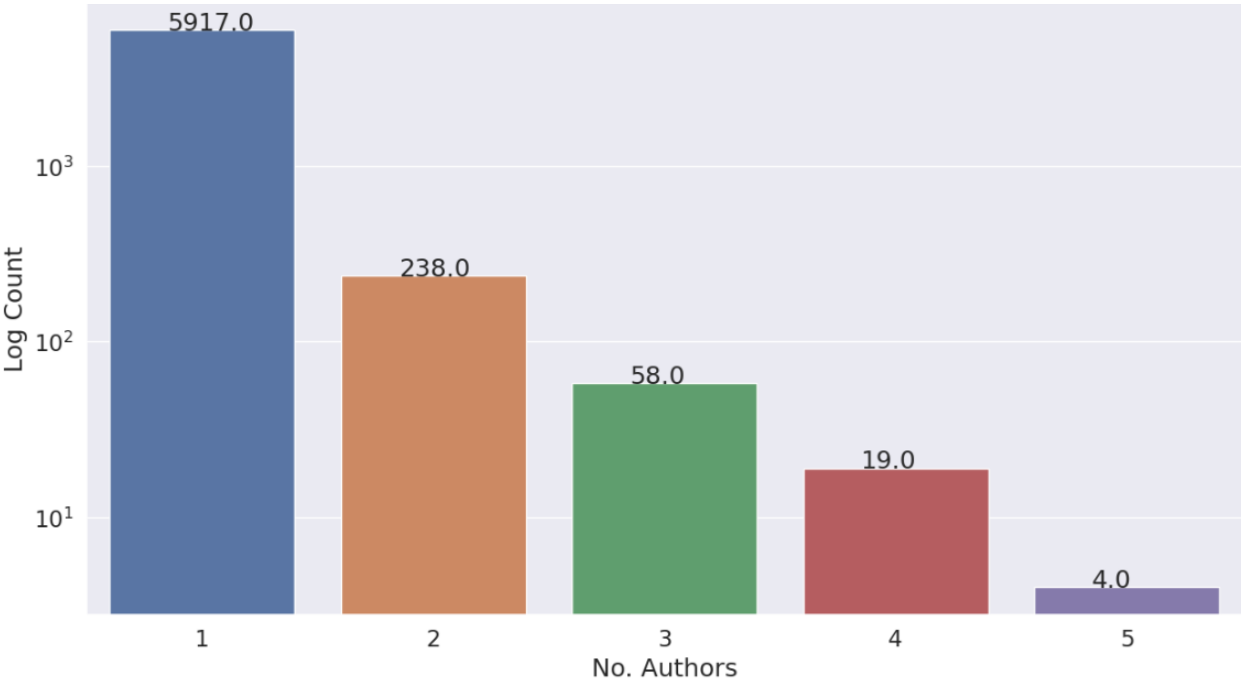
Visualization

Histograms/Barplots for numerical features

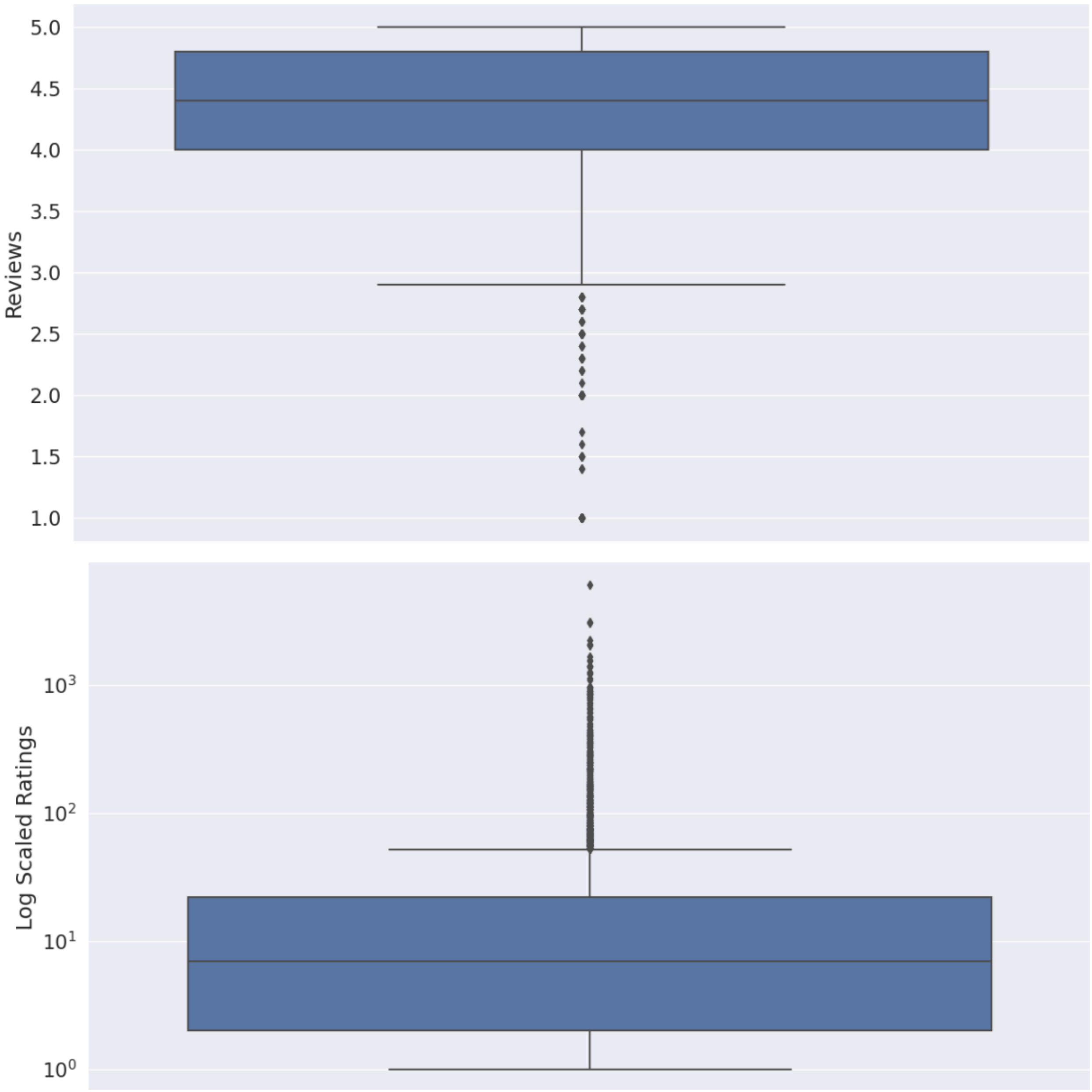


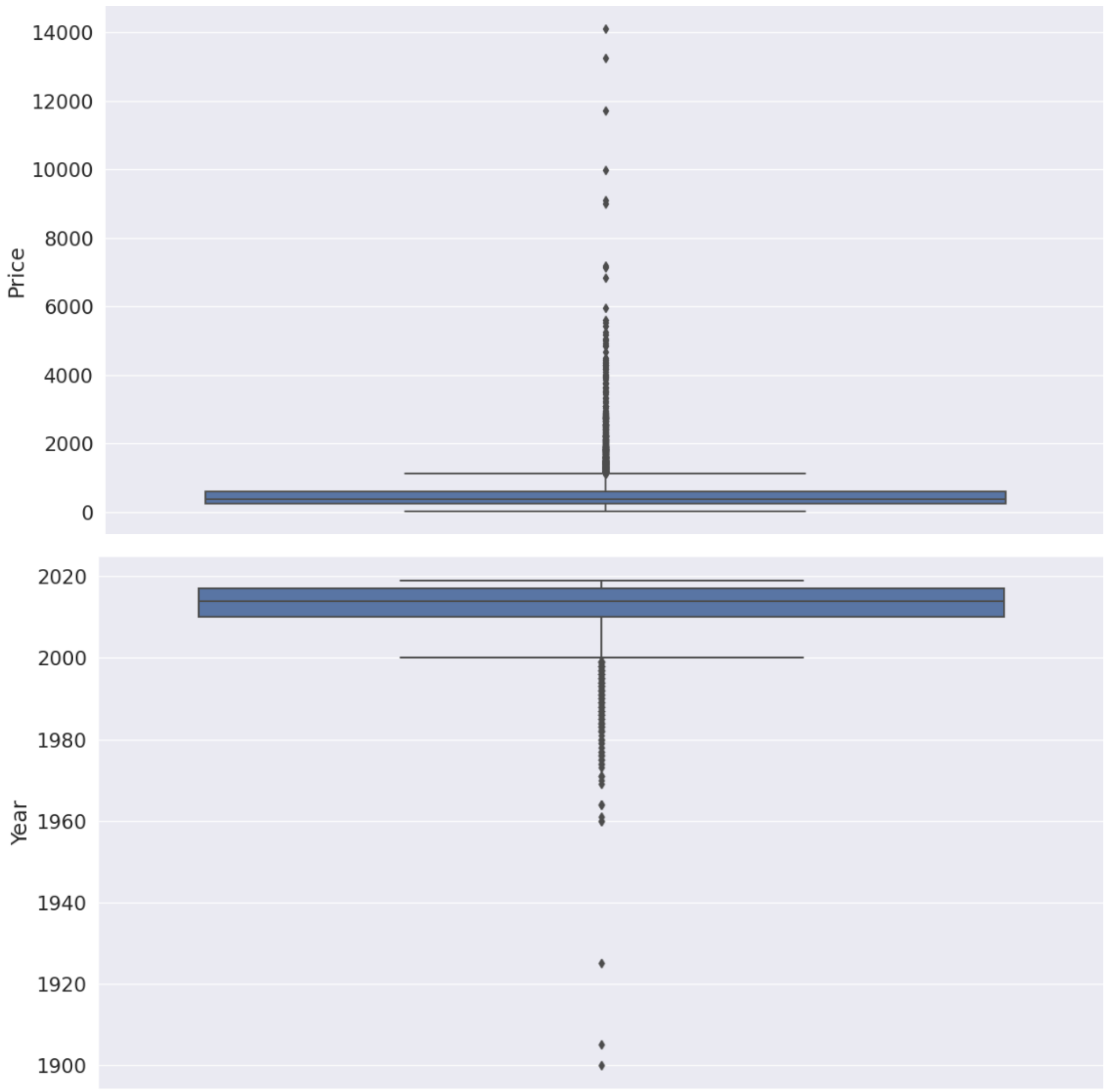

Count
500
400
300
200
100
0
0
2000
4000
6000
8000
10000
12000
14000
Price


Count
800
700
600
500
400
300
200
100
0
1900
1920
1940
1960
1980
2000
2020
Year

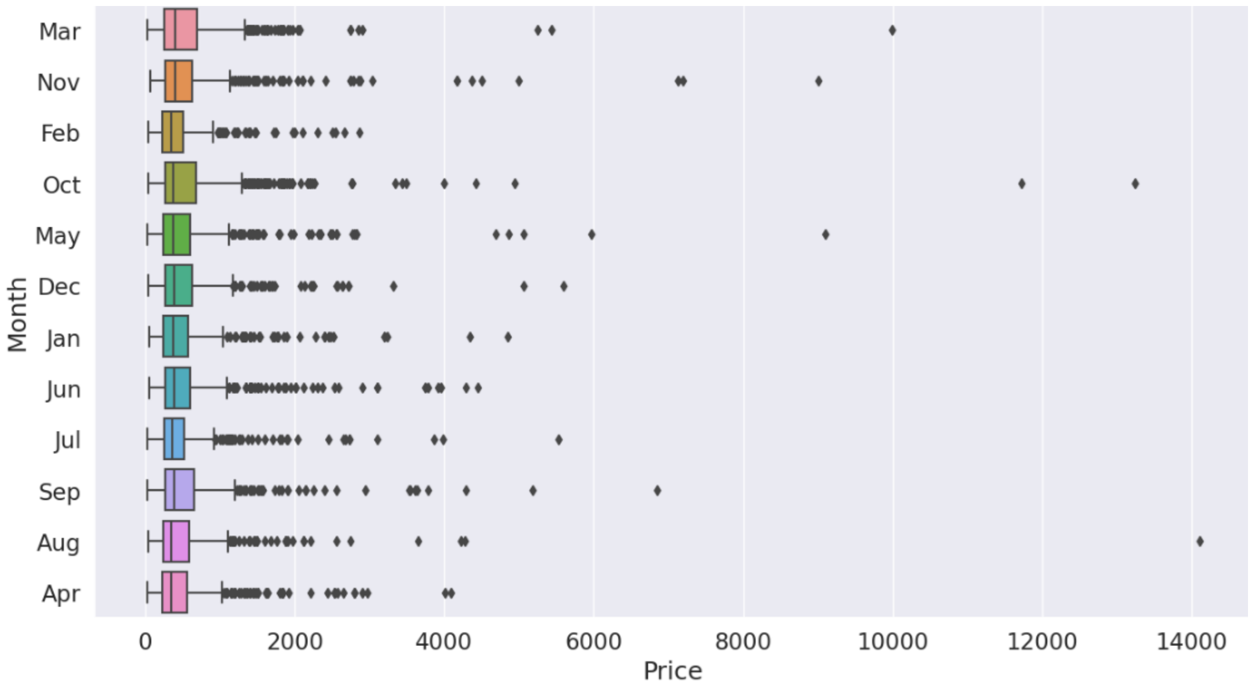

Log Count
10^3
10^2
10^1
5917.0
238.0
58.0
19.0
4.0
1
2
3
4
5
No. Authors

Boxplot for numerical features

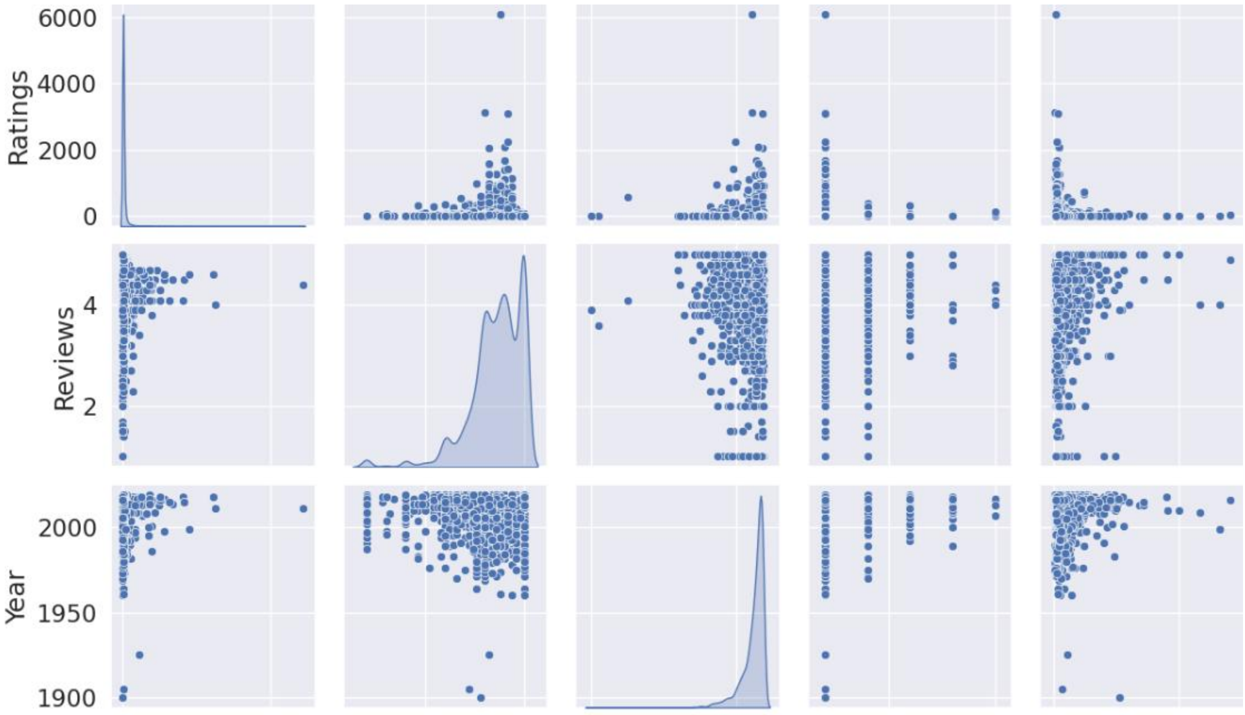


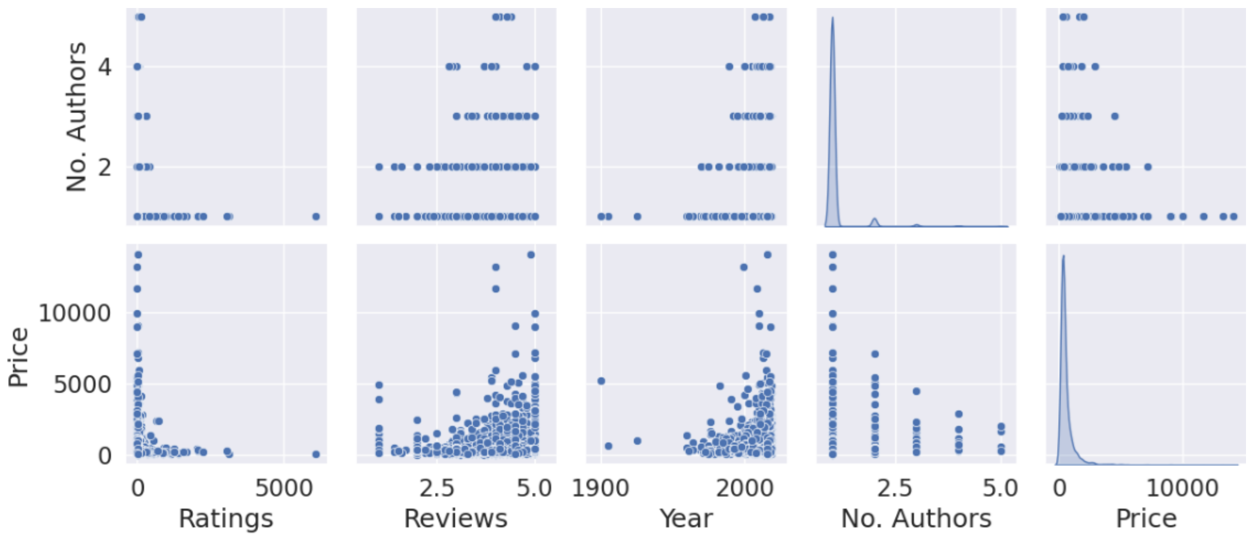

14000
12000
10000
8000
6000
4000
2000
0
Price
2020
2000
1980
1960
1940
1920
1900
Year

Boxplot between Price and Month

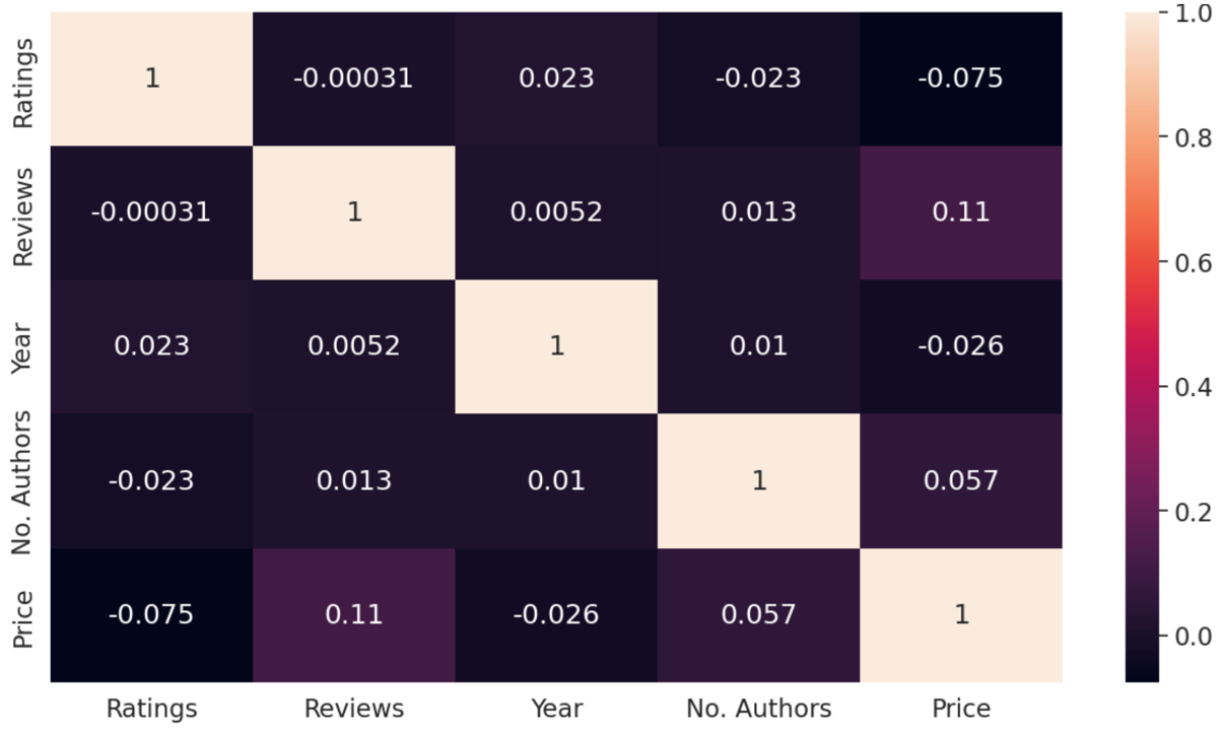


Create pairs plot for all numeric variables



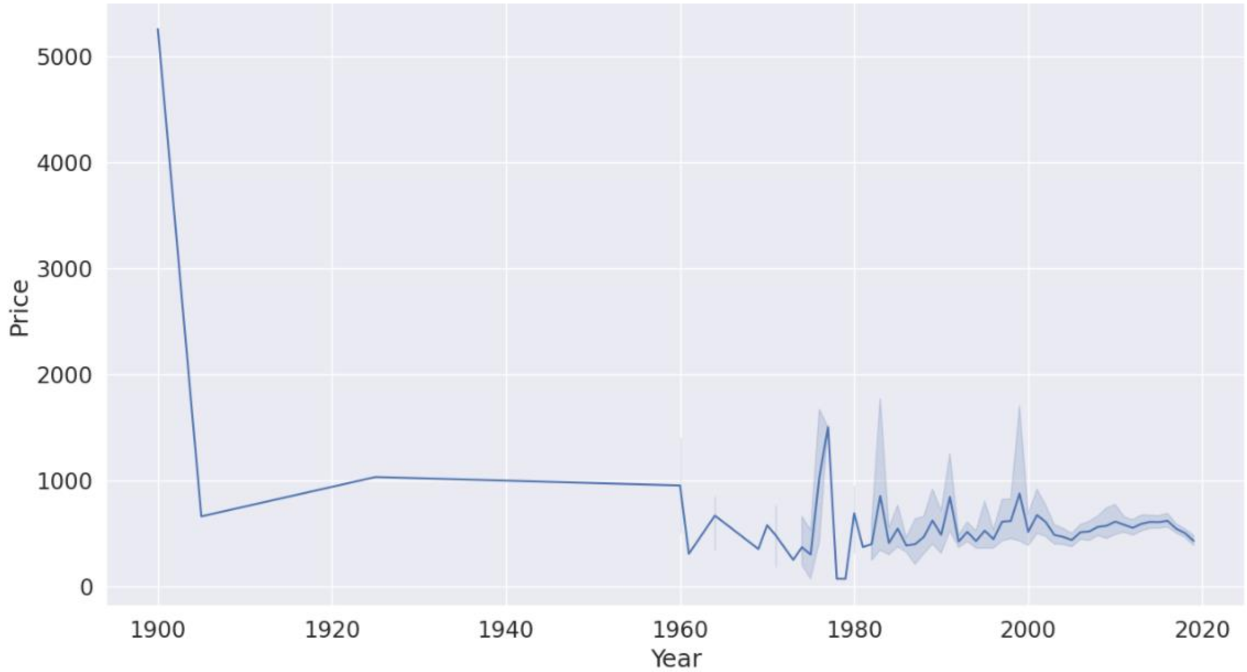


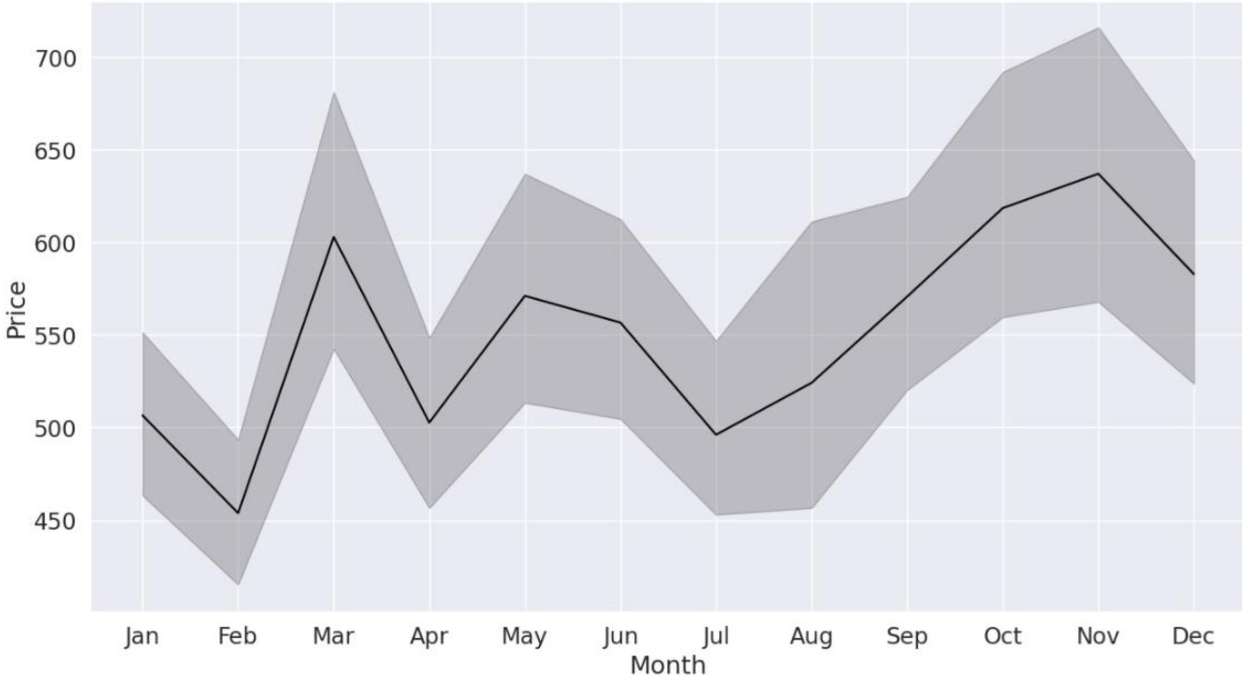
Correlation matrix



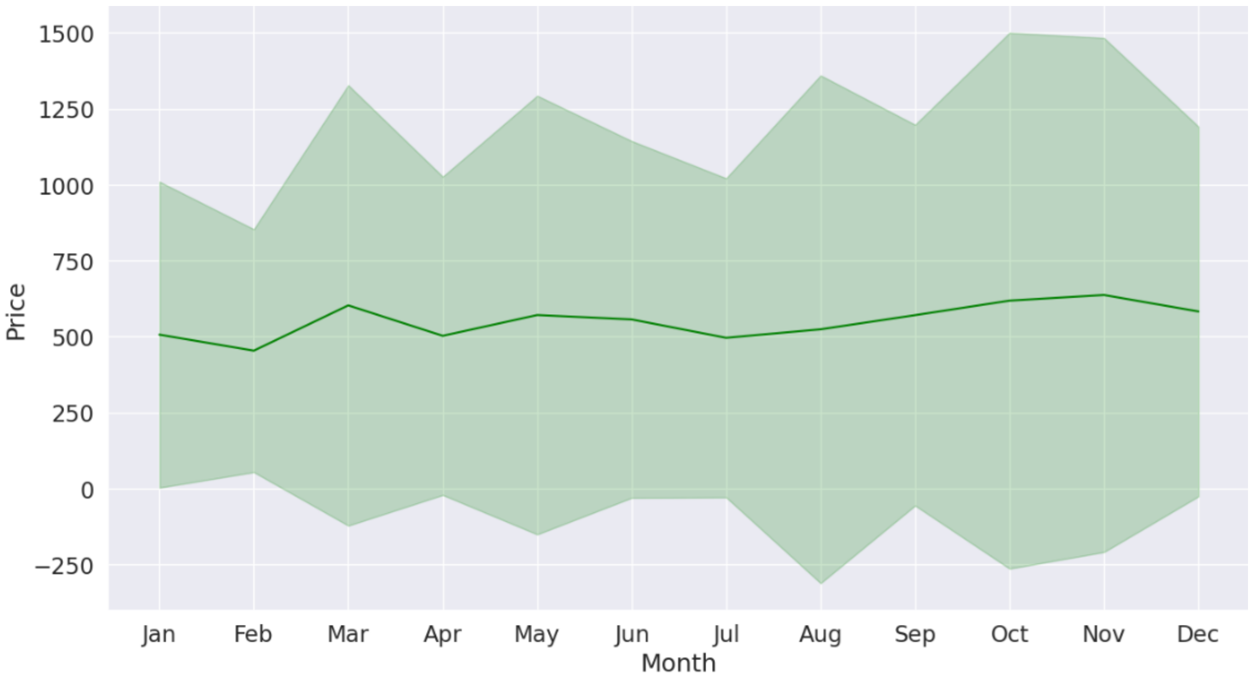
	Ratings	Reviews	Year	No. Authors	Price
Ratings	1	-0.00031	0.023	-0.023	-0.075
Reviews	-0.00031	1	0.0052	0.013	0.11
Year	0.023	0.0052	1	0.01	-0.026
No. Authors	-0.023	0.013	0.01	1	0.057
Price	-0.075	0.11	-0.026	0.057	1

The provided comment suggests that when dealing with more complex datasets in Seaborn, where there are multiple measurements for the same value of the x variable, Seaborn's default behavior is to aggregate these multiple measurements at each x value. This is achieved by plotting the mean and the 95% confidence interval around the mean. This default behavior provides a summary representation of the central tendency and the variability of the data at each x value.

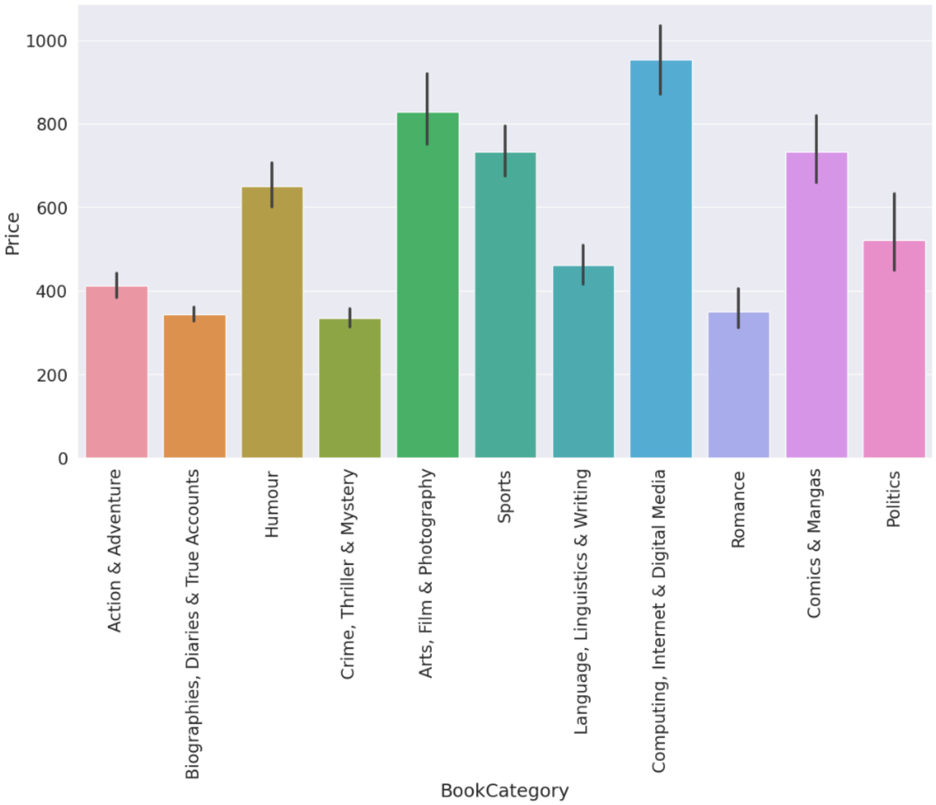


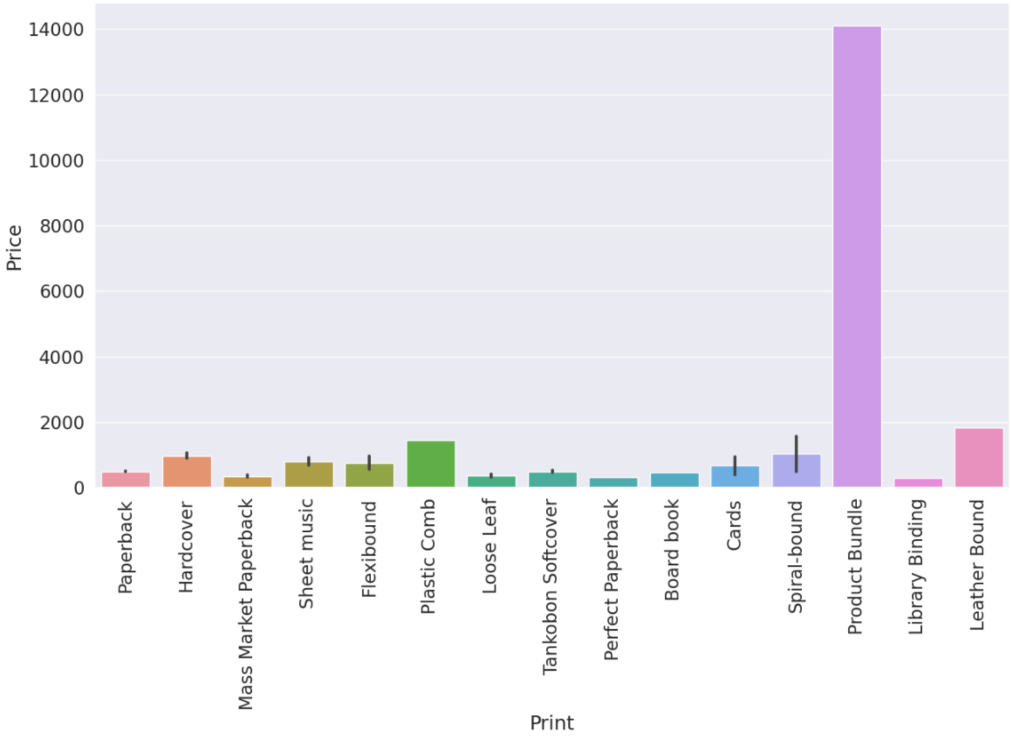


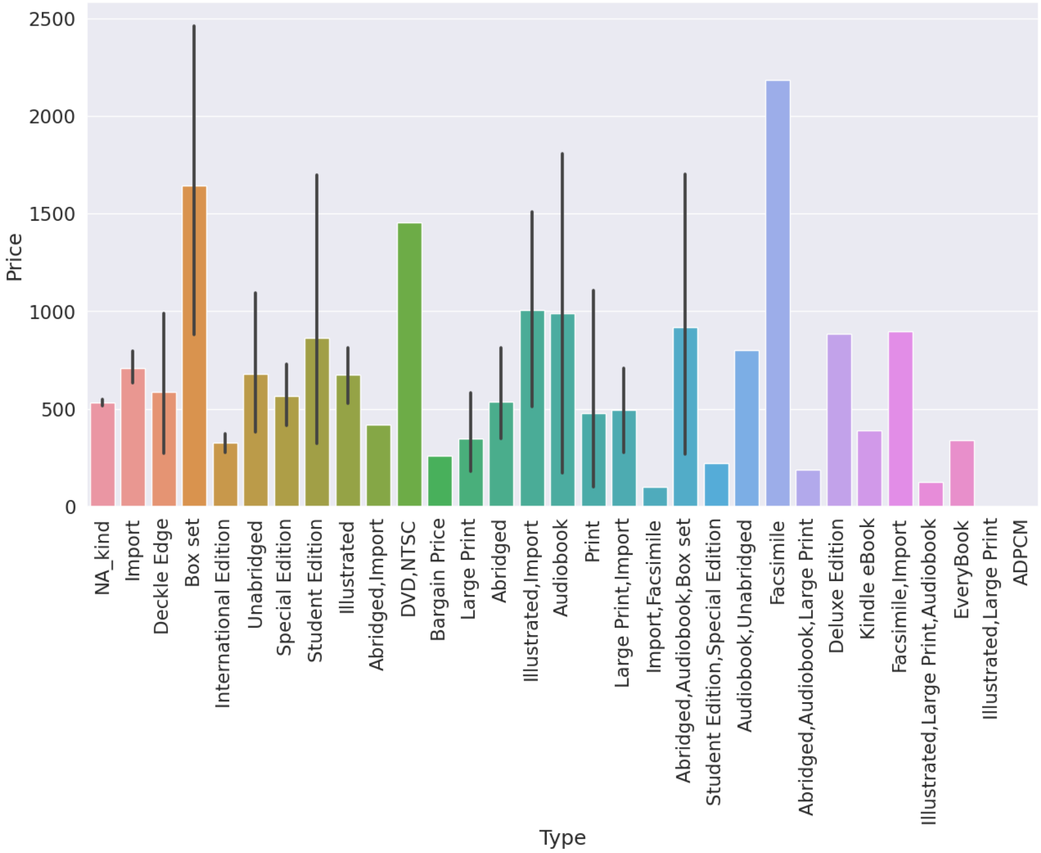
The comment suggests an alternative approach, particularly suitable for larger datasets. Instead of representing the spread of the distribution at each timepoint with a confidence interval, the recommendation is to plot the standard deviation. This alternative provides a visual depiction of the variability or dispersion of the data points at each timepoint, offering a different perspective on the dataset's distribution characteristics.

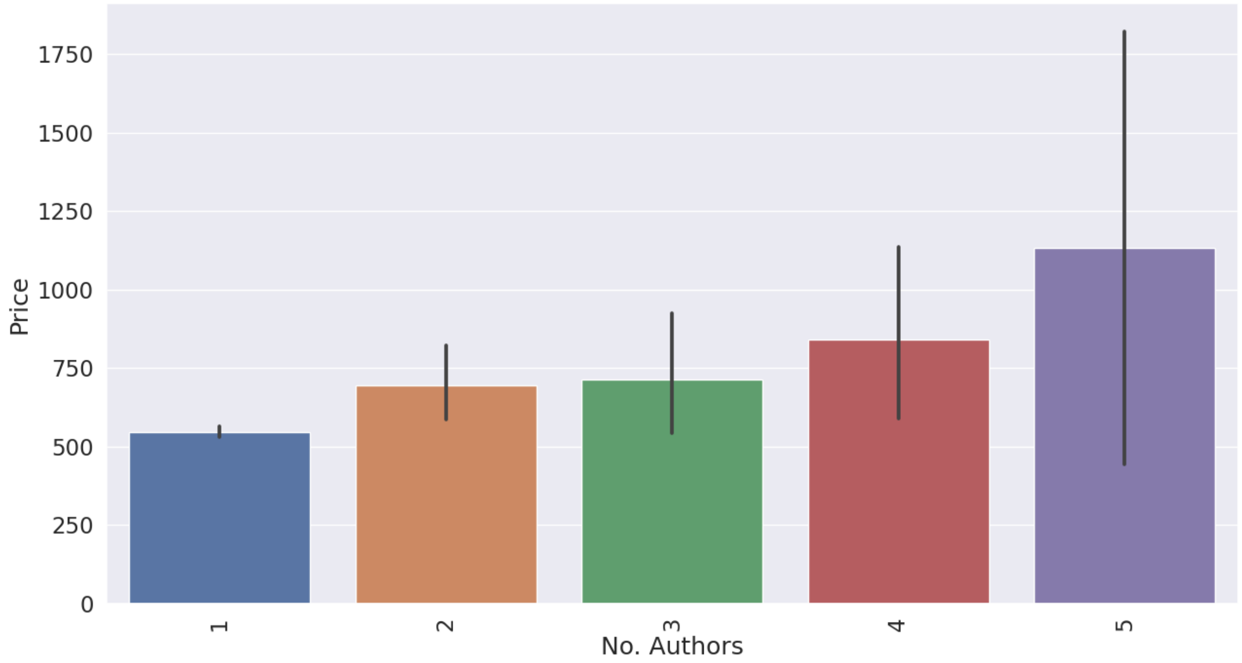


Barplots between Price and Categorical Features




Price
14000
12000
10000
8000
6000
4000
2000
0
Paperback
Hardcover
Mass Market Paperback
Sheet music
Flexibound
Plastic Comb
Loose Leaf
Tankobon Softcover
Perfect Paperback
Board book
Cards
Spiral-bound
Product Bundle
Library Binding
Leather Bound
Print


Price
2500
2000
1500
1000
500
0
NA_kind
Import
Deckle Edge
Box set
International Edition
Unabridged
Special Edition
Student Edition
Illustrated
Abridged,Import
DVD,NTSC
Bargain Price
Large Print
Abridged
Illustrated,Import
Audiobook
Print
Large Print,Import
Import,Facsimile
Abridged,Audiobook,Box set
Student Edition,Special Edition
Audiobook,Unabridged
Facsimile
Abridged,Audiobook,Large Print
Deluxe Edition
Kindle eBook
Facsimile,Import
Illustrated,Large Print,Audiobook
EveryBook
Illustrated,Large Print
ADPCM
Type



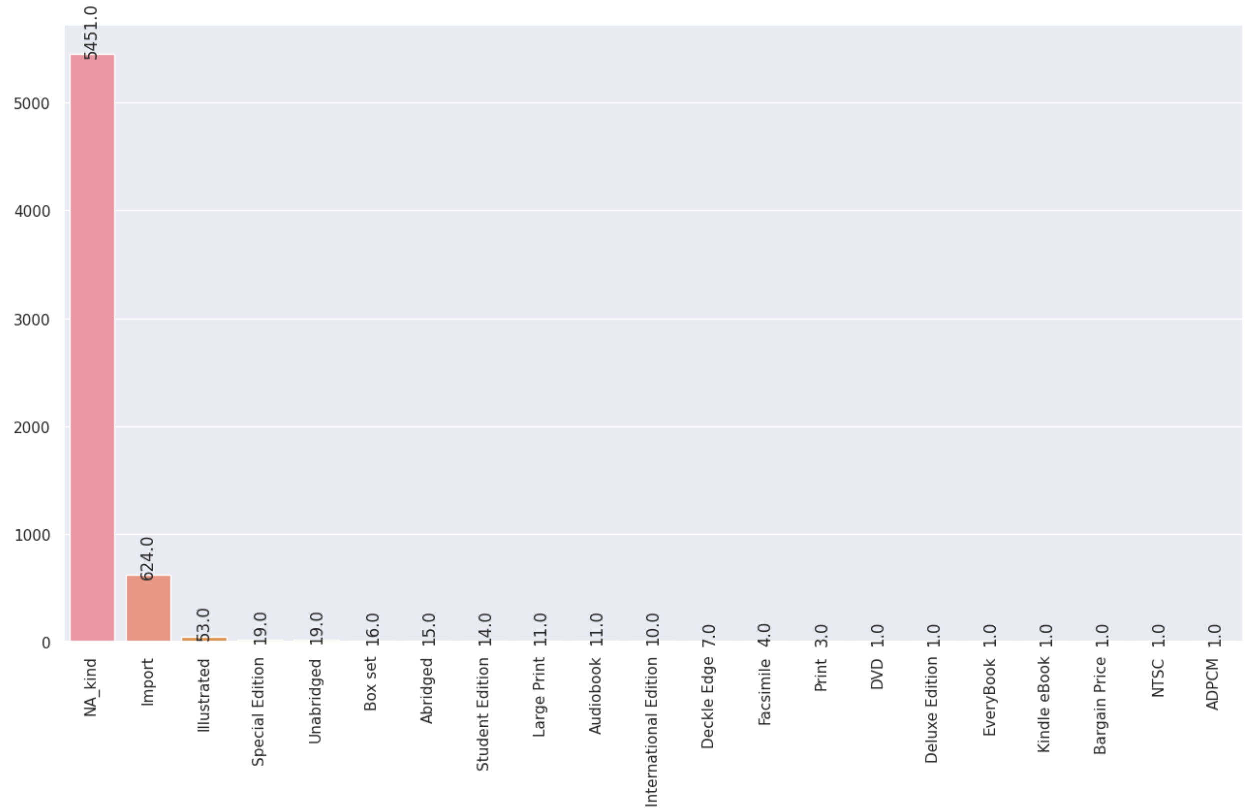
Prediction

Feature engineering for the categorical variables

In the initial preprocessing step, the decision was made to exclude categories from all categorical variables that have a low frequency of occurrence in the dataset. Specifically, a threshold of 0.2% was set, ensuring the removal of categories with minimal representation.

Encoding of Type

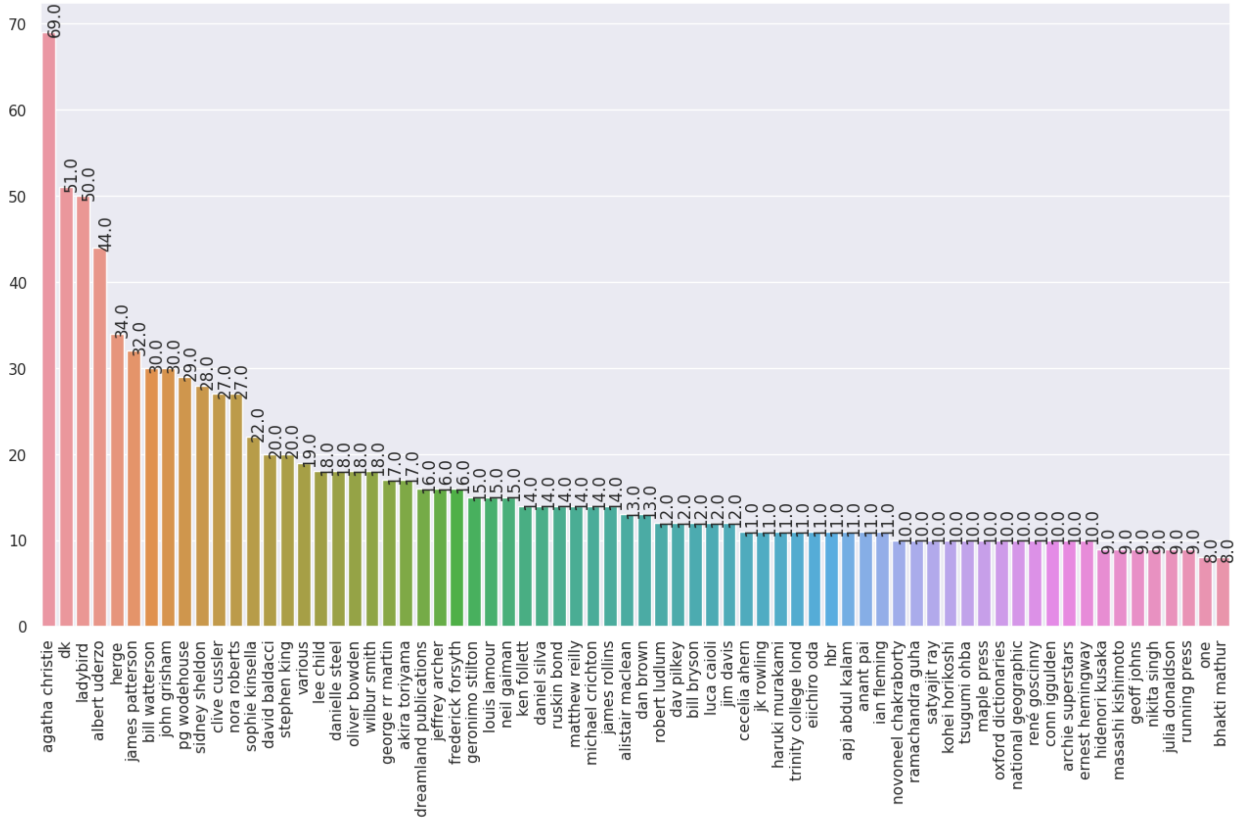
```
Category-kind frequencies:
NA_kind                 5451
Import                   624
Illustrated               53
Special Edition           19
Unabridged                19
Box set                   16
Abridged                  15
Student Edition           14
Large Print               11
Audiobook                 11
International Edition     10
Deckle Edge                7
Facsimile                  4
Print                      3
DVD                        1
Deluxe Edition             1
EveryBook                  1
Kindle eBook               1
Bargain Price              1
NTSC                       1
ADPCM                      1
```

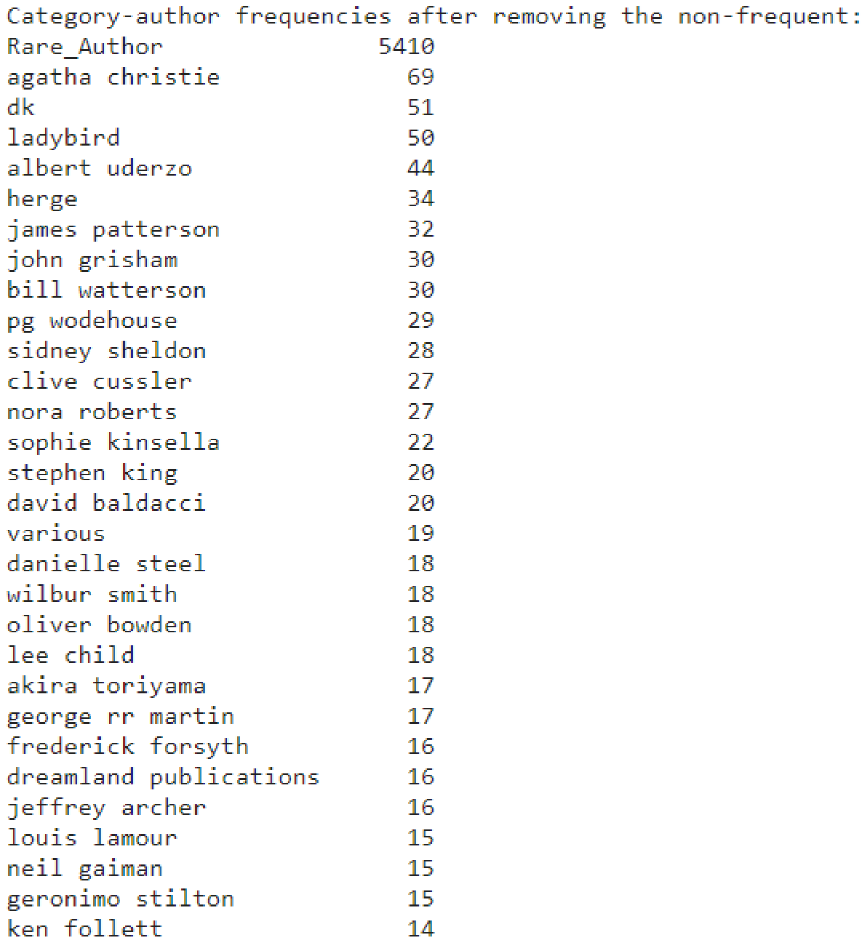


```
Category-kind frequencies after removing the non-frequent:
Import             624
Illustrated         53
Rare_Kind           50
Special Edition     19
Unabridged          19
Box set             16
Abridged            15
Student Edition     14
```

Encoding of Author

```
Category-author frequencies:
agatha christie        69
dk                     51
ladybird               50
albert uderzo          44
herge                  34
                       ..
j p dalvi               1
j p delaney             1
j s bach                1
jacek m zurada          1
zygmunt miloszewski     1
```

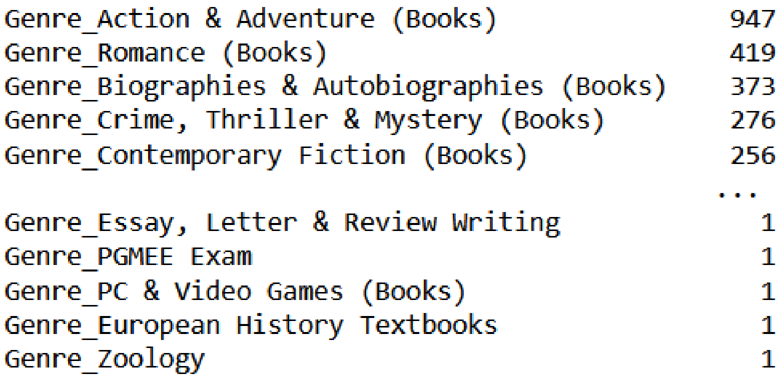




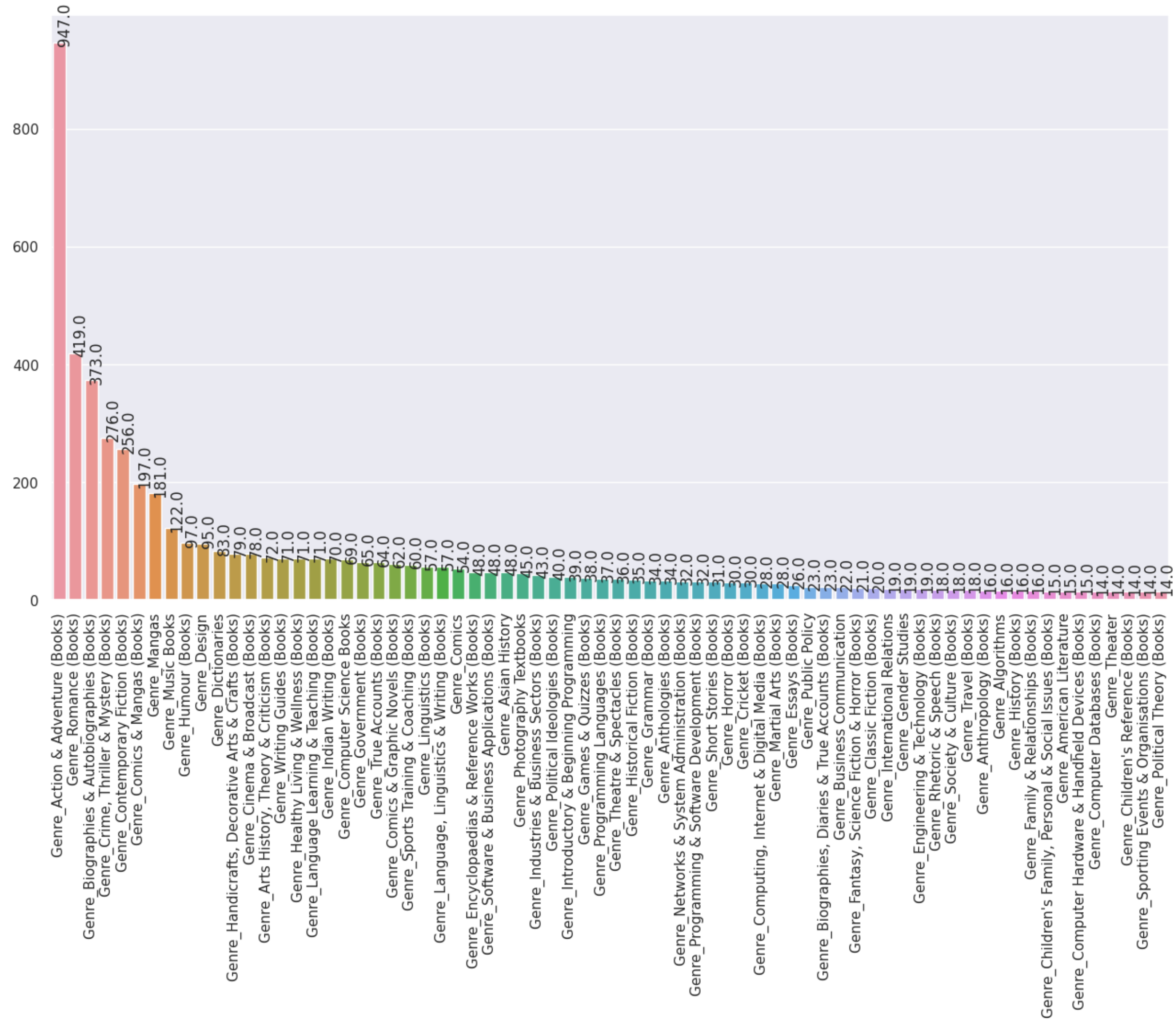
```
Category-author frequencies after removing the non-frequent:
Rare_Author              5410
agatha christie            69
dk                         51
ladybird                   50
albert uderzo              44
herge                      34
james patterson            32
john grisham               30
bill watterson             30
pg wodehouse               29
sidney sheldon             28
clive cussler              27
nora roberts               27
sophie kinsella            22
stephen king               20
david baldacci             20
various                    19
danielle steel             18
wilbur smith               18
oliver bowden              18
lee child                  18
akira toriyama             17
george rr martin           17
frederick forsyth          16
dreamland publications     16
jeffrey archer             16
louis lamour               15
neil gaiman                15
geronimo stilton           15
ken follett                14
```

Categories variables (one-hot encoding)

One-hot encoding Genre

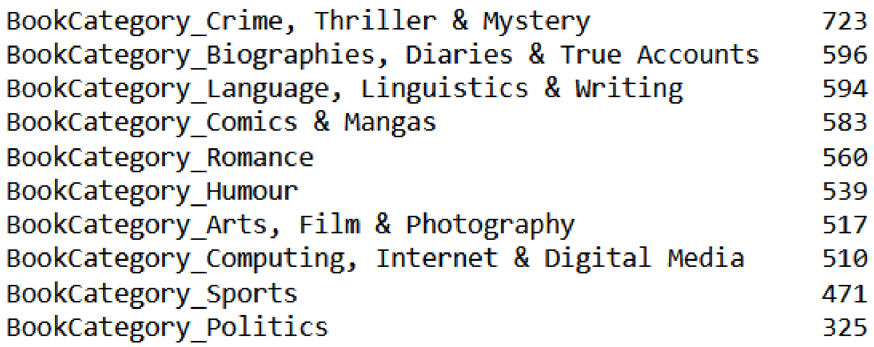


```
Genre_Action & Adventure (Books)             947
Genre_Romance (Books)                        419
Genre_Biographies & Autobiographies (Books)  373
Genre_Crime, Thriller & Mystery (Books)      276
Genre_Contemporary Fiction (Books)           256
                                            ...
Genre_Essay, Letter & Review Writing           1
Genre_PGMEE Exam                               1
Genre_PC & Video Games (Books)                 1
Genre_European History Textbooks               1
Genre_Zoology                                  1
```

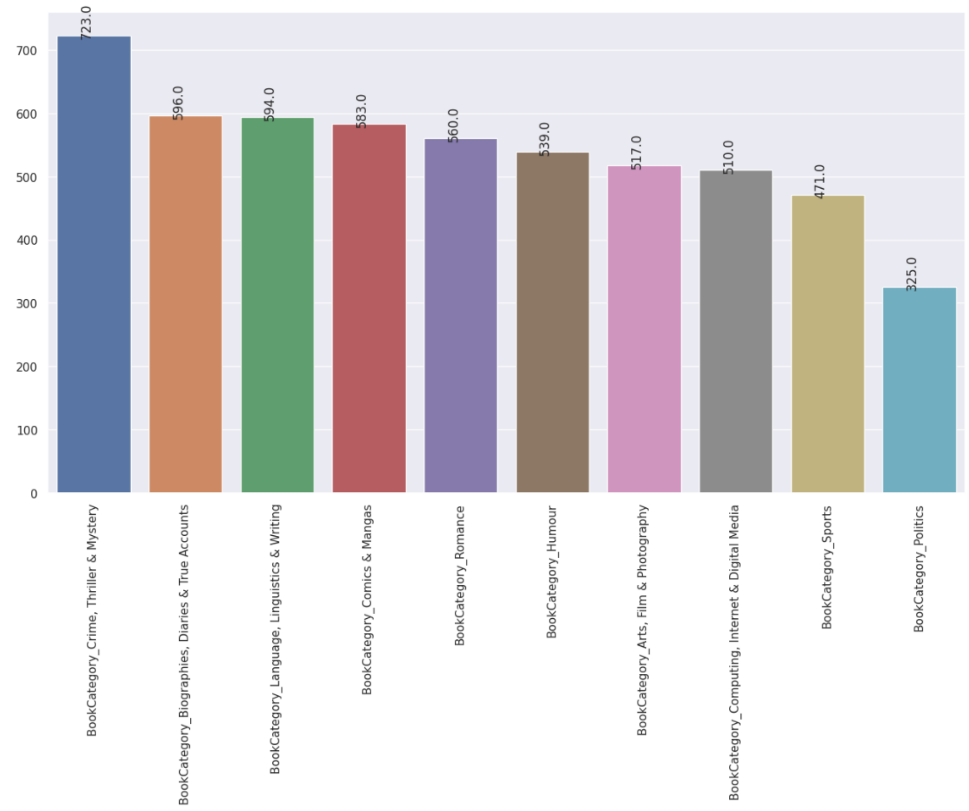


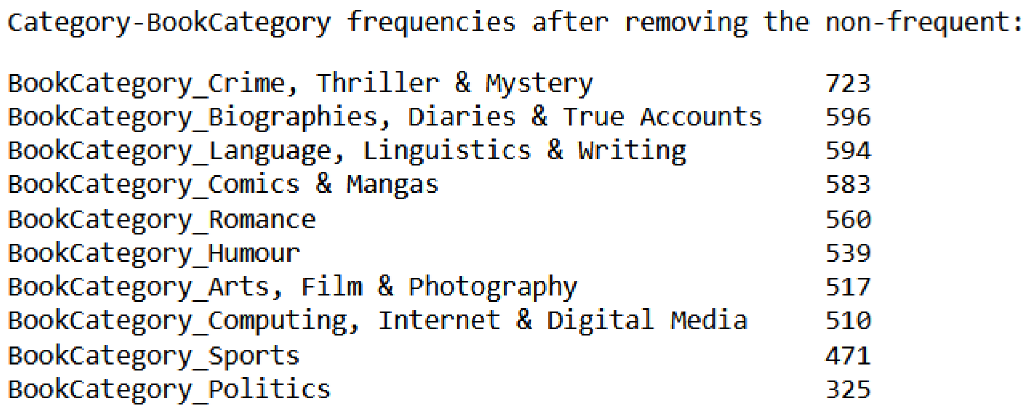
```
Category-genre frequencies after removing the non-frequent:

Genre_Action & Adventure (Books)                                      947
Genre_Romance (Books)                                                 419
Genre_Biographies & Autobiographies (Books)                           373
Genre_Crime, Thriller & Mystery (Books)                               276
Genre_Contemporary Fiction (Books)                                    256
                                                                     ...
Genre_Computer Databases (Books)                                       14
Genre_Journalism Books                                                 13
Genre_Economics Textbooks                                              13
Genre_Children's Crafts, Hobbies & Practical Interests (Books)         13
Genre_Architecture (Books)                                             13
```

One-hot encoding Book category



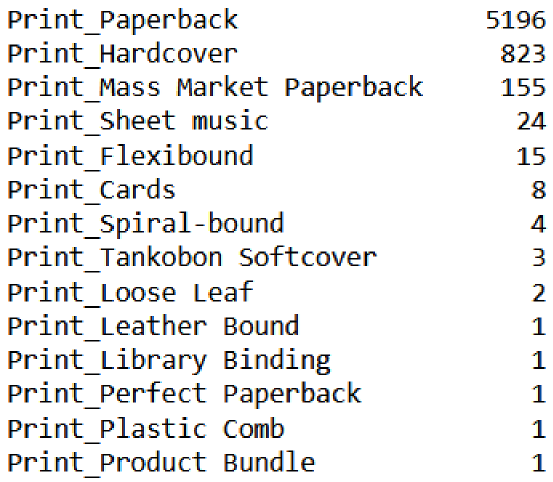
```
BookCategory_Crime, Thriller & Mystery              723
BookCategory_Biographies, Diaries & True Accounts   596
BookCategory_Language, Linguistics & Writing        594
BookCategory_Comics & Mangas                        583
BookCategory_Romance                                560
BookCategory_Humour                                 539
BookCategory_Arts, Film & Photography               517
BookCategory_Computing, Internet & Digital Media    510
BookCategory_Sports                                 471
BookCategory_Politics                               325
```



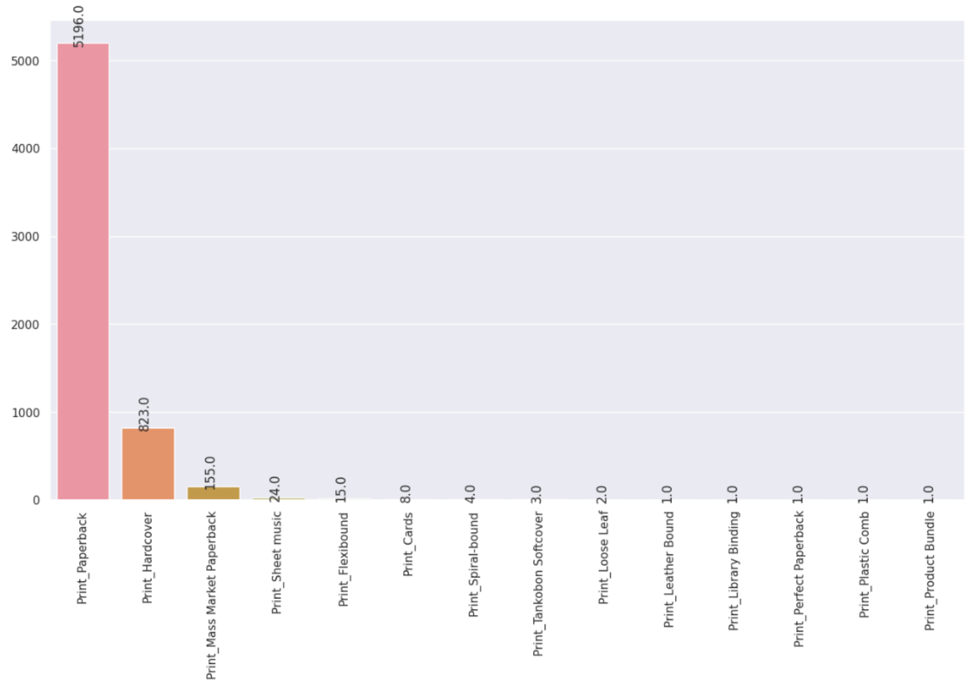


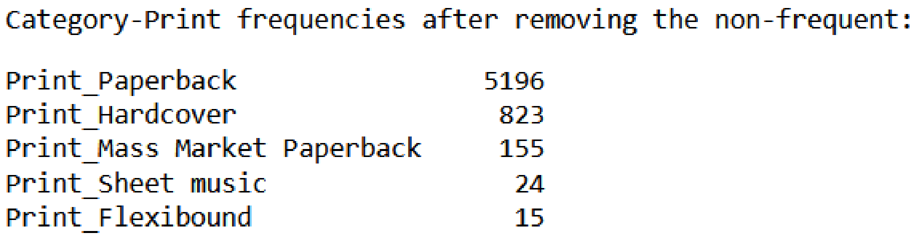
```
Category-BookCategory frequencies after removing the non-frequent:

BookCategory_Crime, Thriller & Mystery              723
BookCategory_Biographies, Diaries & True Accounts   596
BookCategory_Language, Linguistics & Writing        594
BookCategory_Comics & Mangas                        583
BookCategory_Romance                                560
BookCategory_Humour                                 539
BookCategory_Arts, Film & Photography               517
BookCategory_Computing, Internet & Digital Media    510
BookCategory_Sports                                 471
BookCategory_Politics                               325
```

One-hot encoding Print



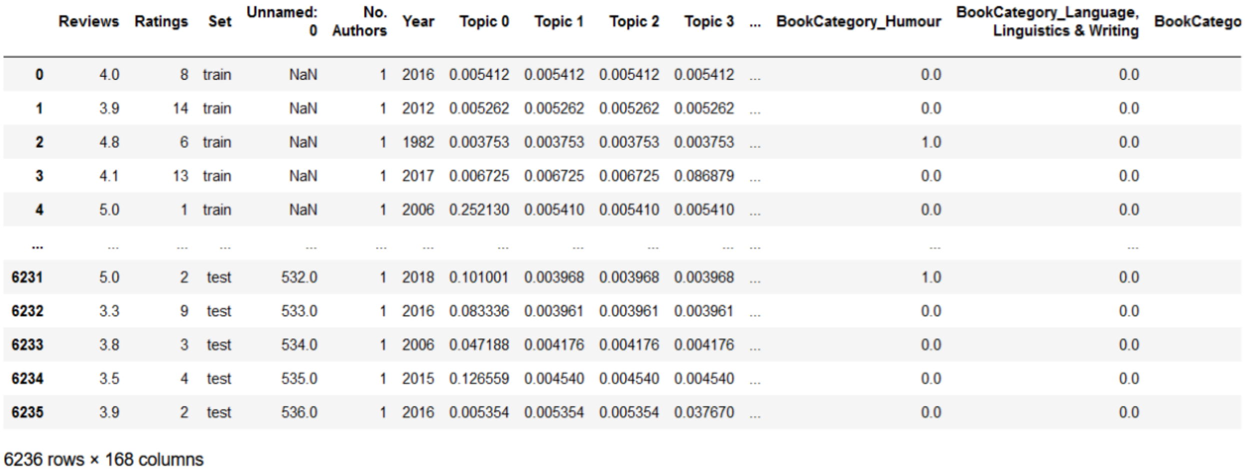
```
Print_Paperback                5196
Print_Hardcover                 823
Print_Mass Market Paperback     155
Print_Sheet music                24
Print_Flexibound                 15
Print_Cards                       8
Print_Spiral-bound                4
Print_Tankobon Softcover          3
Print_Loose Leaf                  2
Print_Leather Bound               1
Print_Library Binding             1
Print_Perfect Paperback           1
Print_Plastic Comb                1
Print_Product Bundle              1
```





```
Category-Print frequencies after removing the non-frequent:

Print_Paperback                5196
Print_Hardcover                 823
Print_Mass Market Paperback     155
Print_Sheet music                24
Print_Flexibound                 15
```

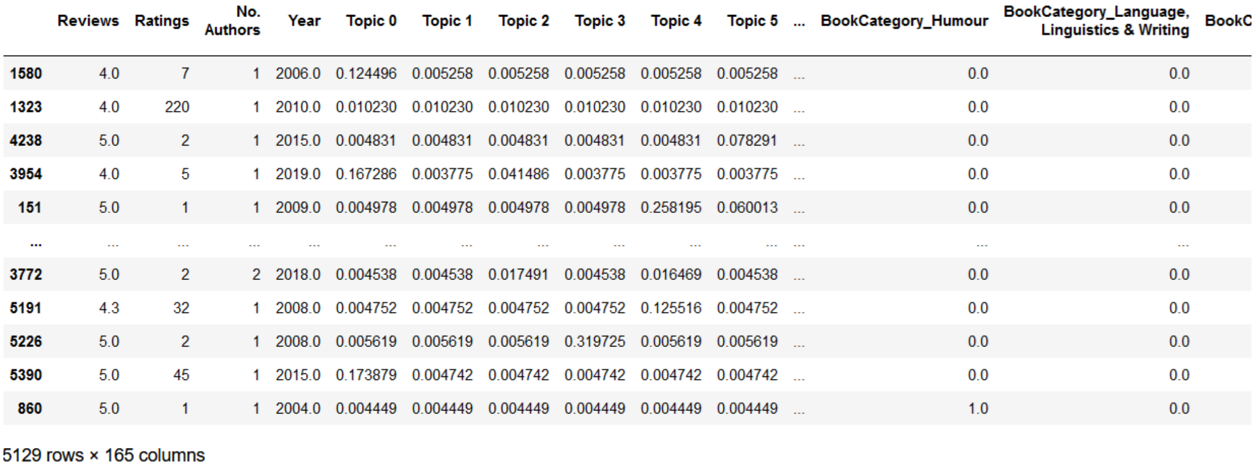
During this phase, the original categorical variables are eliminated, and new "one-hotted" columns are introduced into the dataframe.



	Reviews	Ratings	Set	Unnamed: 0	No. Authors	Year	Topic 0	Topic 1	Topic 2	Topic 3	...	BookCategory_Humour	BookCategory_Language, Linguistics & Writing	BookCatego
0	4.0	8	train	NaN	1	2016	0.005412	0.005412	0.005412	0.005412	...	0.0	0.0	
1	3.9	14	train	NaN	1	2012	0.005262	0.005262	0.005262	0.005262	...	0.0	0.0	
2	4.8	6	train	NaN	1	1982	0.003753	0.003753	0.003753	0.003753	...	1.0	0.0	
3	4.1	13	train	NaN	1	2017	0.006725	0.006725	0.006725	0.086879	...	0.0	0.0	
4	5.0	1	train	NaN	1	2006	0.252130	0.005410	0.005410	0.005410	...	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
6231	5.0	2	test	532.0	1	2018	0.101001	0.003968	0.003968	0.003968	...	1.0	0.0	
6232	3.3	9	test	533.0	1	2016	0.083336	0.003961	0.003961	0.003961	...	0.0	0.0	
6233	3.8	3	test	534.0	1	2006	0.047188	0.004176	0.004176	0.004176	...	0.0	0.0	
6234	3.5	4	test	535.0	1	2015	0.126559	0.004540	0.004540	0.004540	...	0.0	0.0	
6235	3.9	2	test	536.0	1	2016	0.005354	0.005354	0.005354	0.037670	...	0.0	0.0	

6236 rows × 168 columns

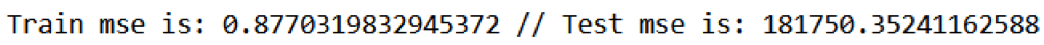
In this phase of the analysis, the dataset was partitioned into training and testing sets. The training set, denoted by 'Set=train', was used to train a predictive model employing the RandomForestRegressor algorithm. The model's performance was assessed using the Mean Squared Error (MSE), a metric that gauges the accuracy of the model's predictions.



	Reviews	Ratings	No. Authors	Year	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	...	BookCategory_Humour	BookCategory_Language, Linguistics & Writing	BookC
1580	4.0	7	1	2006.0	0.124496	0.005258	0.005258	0.005258	0.005258	0.005258	...	0.0	0.0	
1323	4.0	220	1	2010.0	0.010230	0.010230	0.010230	0.010230	0.010230	0.010230	...	0.0	0.0	
4238	5.0	2	1	2015.0	0.004831	0.004831	0.004831	0.004831	0.004831	0.078291	...	0.0	0.0	
3954	4.0	5	1	2019.0	0.167286	0.003775	0.041486	0.003775	0.003775	0.003775	...	0.0	0.0	
151	5.0	1	1	2009.0	0.004978	0.004978	0.004978	0.004978	0.258195	0.060013	...	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3772	5.0	2	2	2018.0	0.004538	0.004538	0.017491	0.004538	0.016469	0.004538	...	0.0	0.0	
5191	4.3	32	1	2008.0	0.004752	0.004752	0.004752	0.004752	0.125516	0.004752	...	0.0	0.0	
5226	5.0	2	1	2008.0	0.005619	0.005619	0.005619	0.319725	0.005619	0.005619	...	0.0	0.0	
5390	5.0	45	1	2015.0	0.173879	0.004742	0.004742	0.004742	0.004742	0.004742	...	0.0	0.0	
860	5.0	1	1	2004.0	0.004449	0.004449	0.004449	0.004449	0.004449	0.004449	...	1.0	0.0	

5129 rows × 165 columns

The training set was utilized to fit the RandomForestRegressor, and subsequently, predictions were made on the designated test set. The MSE values for both the training and test sets were computed to quantify the model's predictive accuracy.



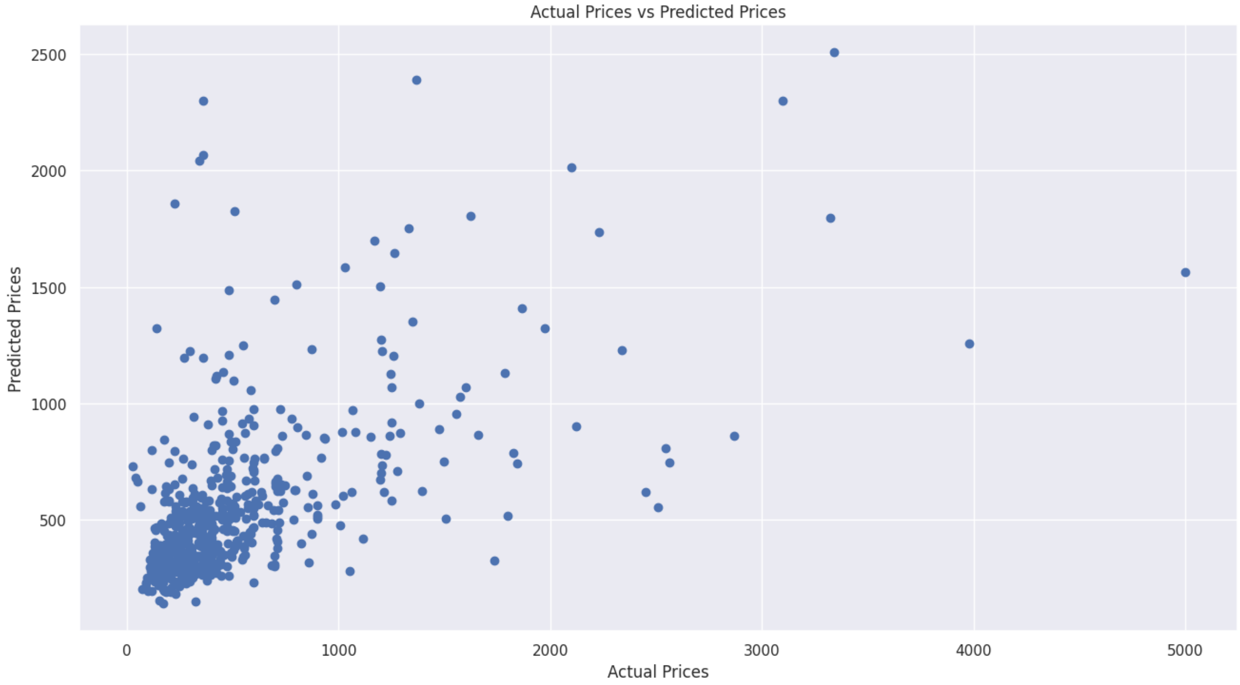
```
Train mse is: 0.8770319832945372 // Test mse is: 181750.35241162588
```

Post processing

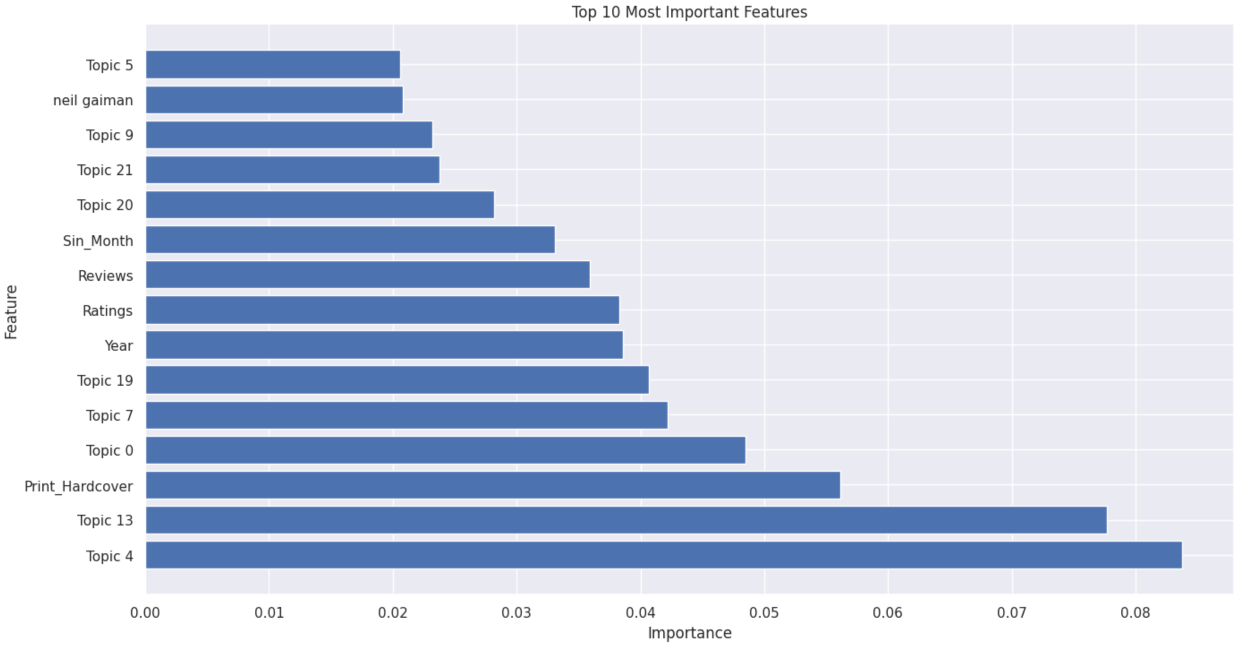
Error Analysis: Analyzing the model's errors on the test data can provide valuable insights into its strengths and weaknesses. This involves identifying common error patterns, understanding the causes of errors, and exploring strategies to mitigate them.

Feature Importance Analysis: Identifying the most important features contributing to the model's predictions can help in feature selection and dimensionality reduction. This can lead to a more efficient and interpretable preprocessing pipeline.

Error Analysis



Feature Importance Analysis


Top 10 Most Important Features
Feature
Topic 5
neil gaiman
Topic 9
Topic 21
Topic 20
Sin_Month
Reviews
Ratings
Year
Topic 19
Topic 7
Topic 0
Print_Hardcover
Topic 13
Topic 4
Importance