Capstone Project -4

Book Recommendation System

Individual Project

Kalyani Nikam

Problem Description

During the last few decades, with the rise of YouTube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries).

Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors.
The main objective is to create a book recommendation system for users.

Data Description

The Book-Crossing dataset comprises 3 files.

- **Users:**

Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

Books:

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title, Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-Image-URL-M, Image-URL-L), i.e., small, medium large. These URLs point to the Amazon website.

Ratings:

Contains the book rating information. Ratings (Book-Rating) are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

Data Preparation

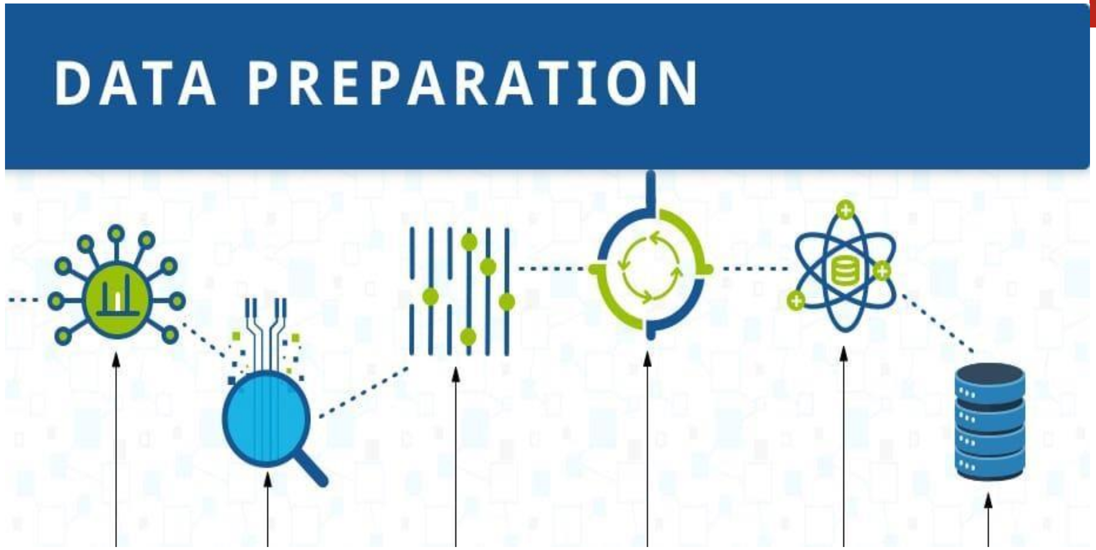
- **Handling missing values**

- Looking for percentage of null values of each column
- For columns containing large null values, replacing null with proper values.
- For columns containing small null values, dropping those nulls.

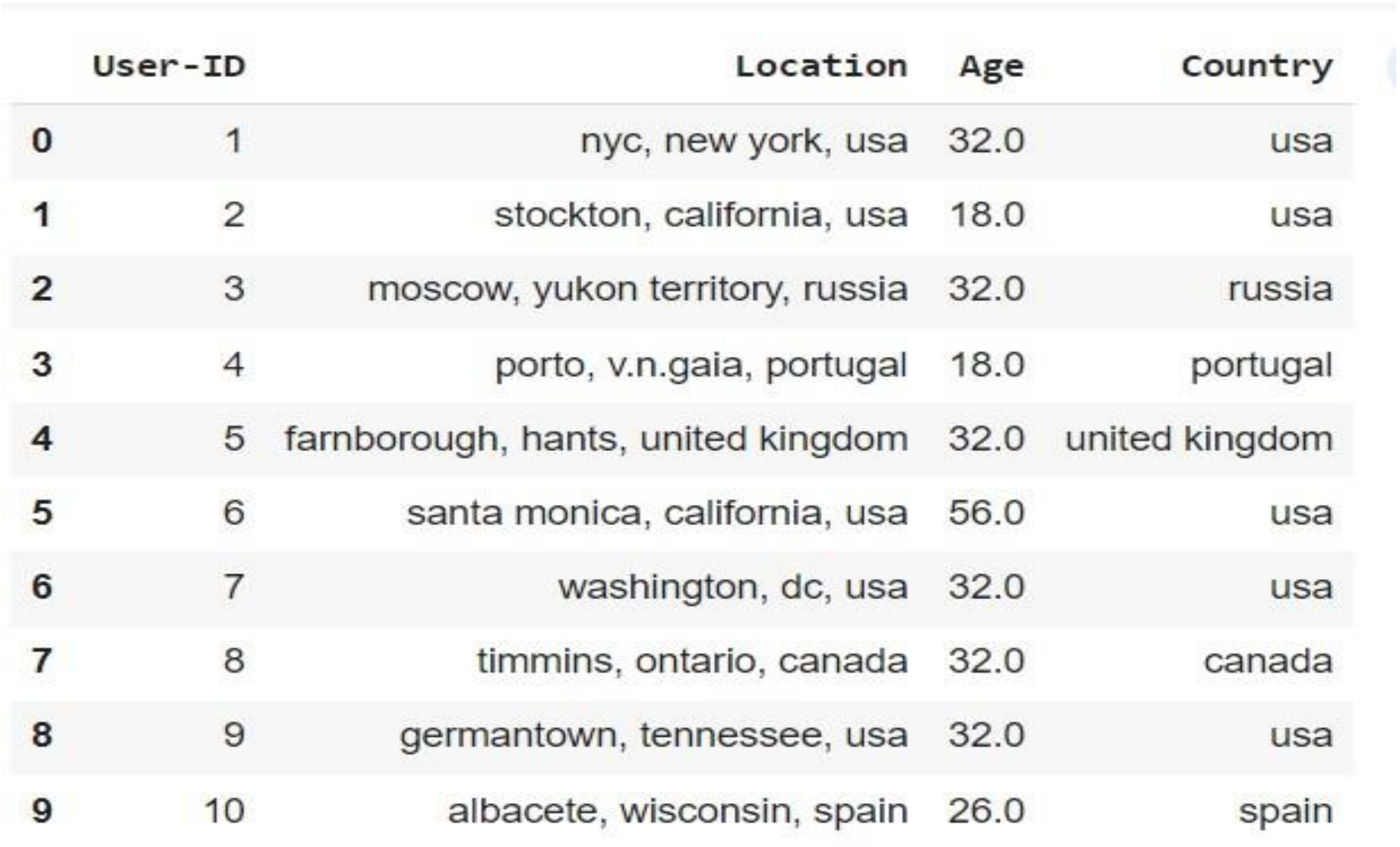
- **Handling duplicate values**

- **Making of proper features**

- Removed outliers from features **`Year-OfPublication and Age`**
- Removed unnecessary features
- Proper formatting of features
- Created new column **`Country`** from **`Location`**



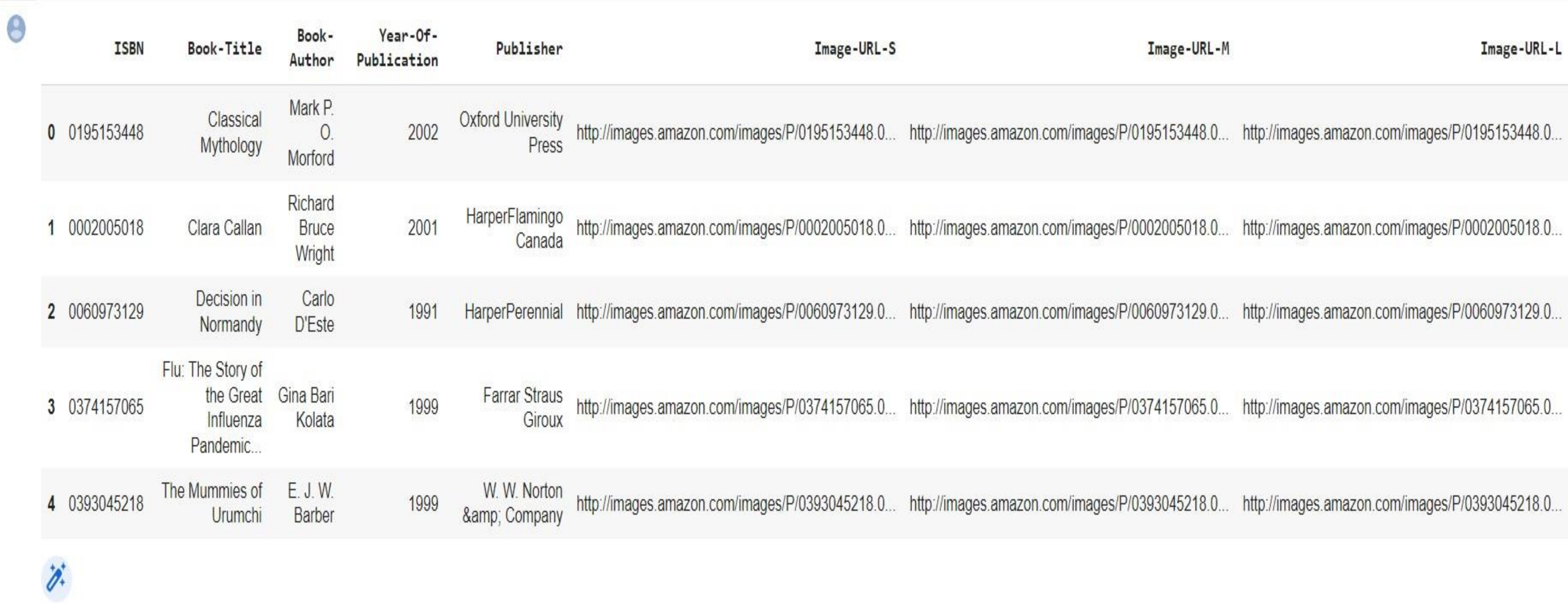
Insights Into Users Data frame



	User-ID	Location	Age	Country
0	1	nyc, new york, usa	32.0	usa
1	2	stockton, california, usa	18.0	usa
2	3	moscow, yukon territory, russia	32.0	russia
3	4	porto, v.n.gaia, portugal	18.0	portugal
4	5	farnborough, hants, united kingdom	32.0	united kingdom
5	6	santa monica, california, usa	56.0	usa
6	7	washington, dc, usa	32.0	usa
7	8	timmins, ontario, canada	32.0	canada
8	9	germantown, tennessee, usa	32.0	usa
9	10	albacete, wisconsin, spain	26.0	spain

Range Index:278858 entries, 0 to 278857 Data columns (total 3 columns):

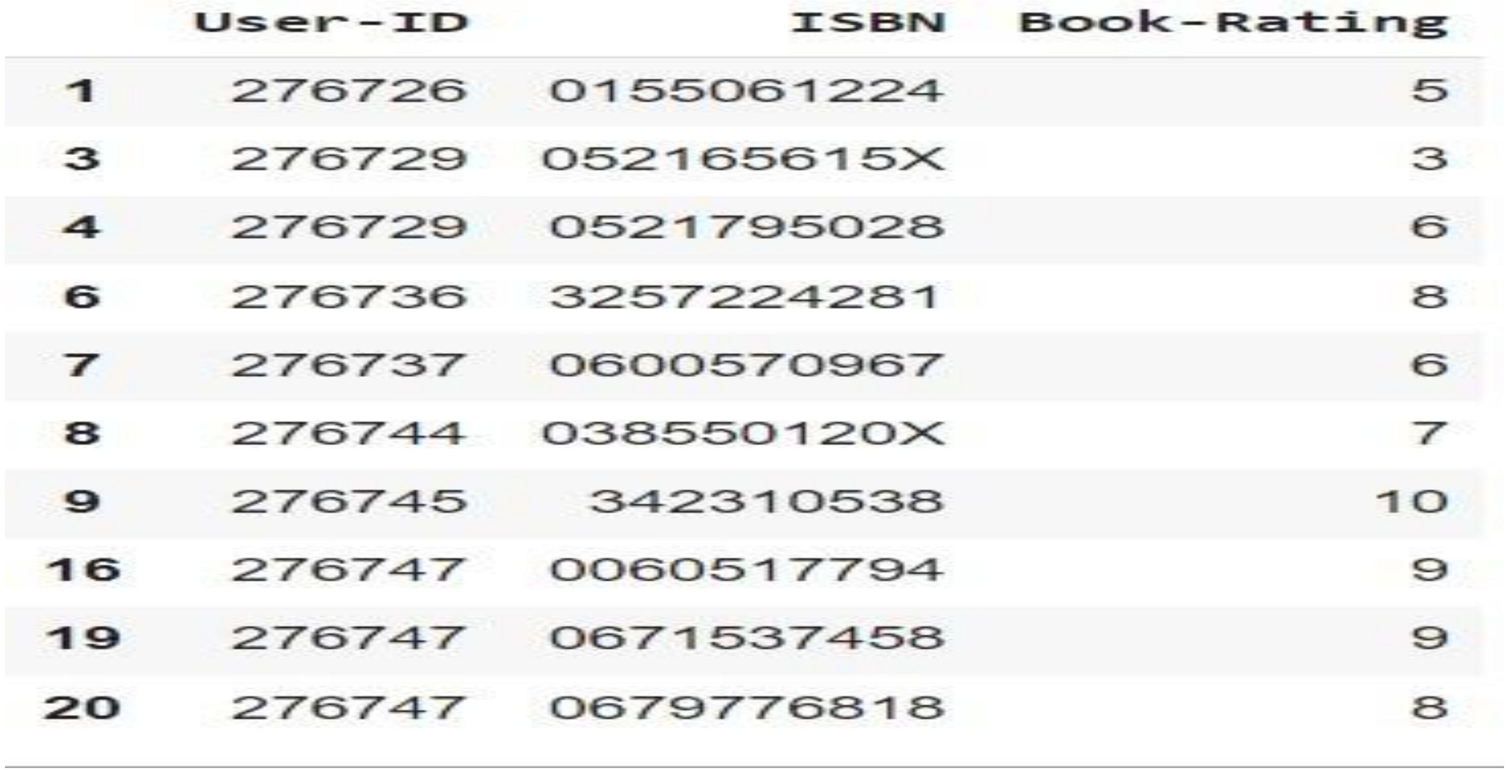
Insights Into Books Data frame



	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...
4	0393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton &amp; Company	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...	http://images.amazon.com/images/P/0393045218.0...

Range Index: 271360 entries, 0 to 271359 Data columns (total 8 columns):

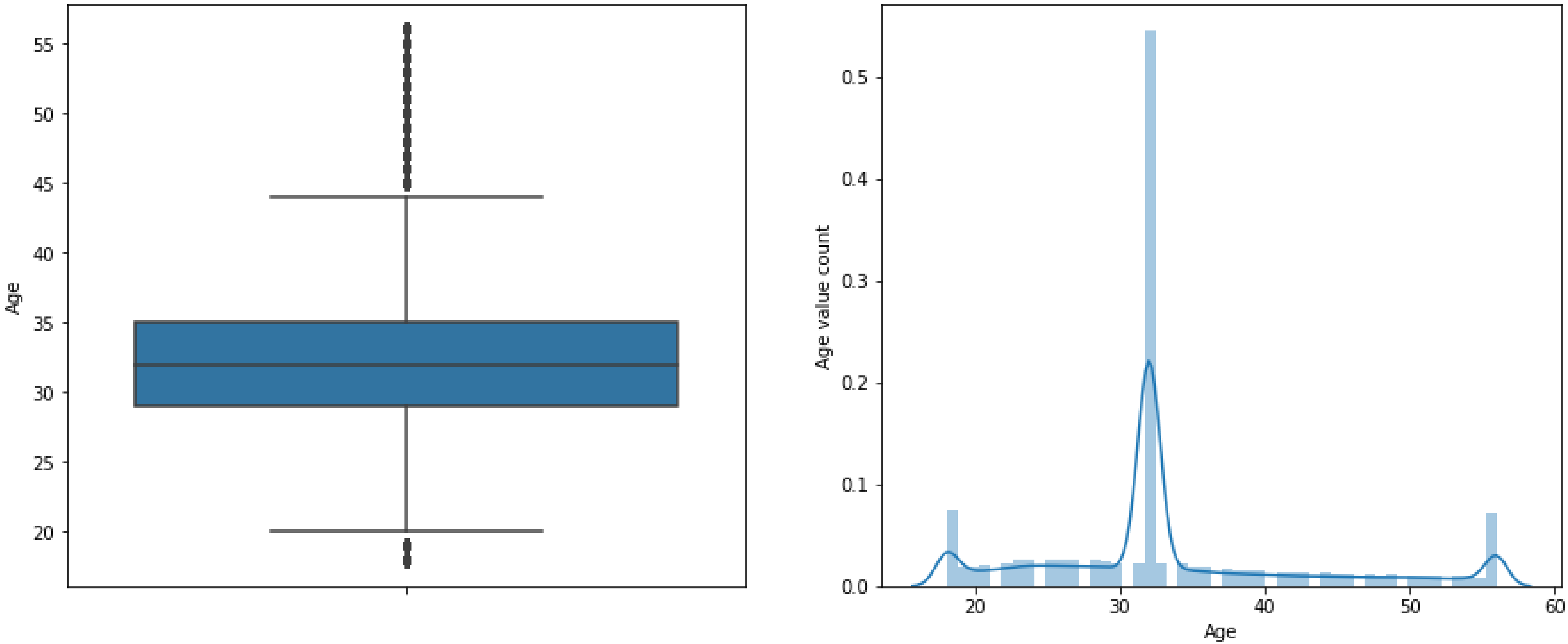
Insights Into Rating Data frame



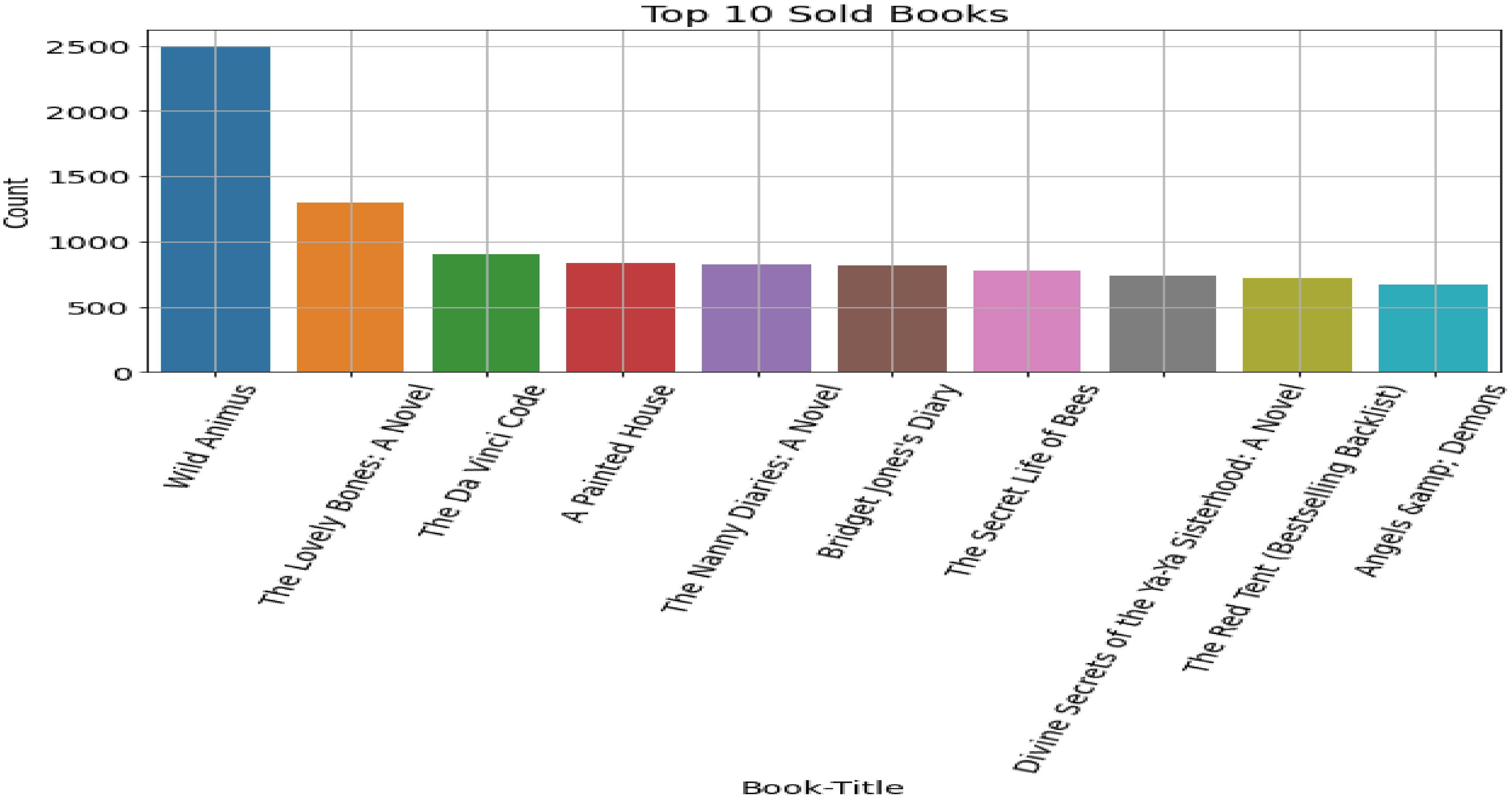
	User-ID	ISBN	Book-Rating
1	276726	0155061224	5
3	276729	052165615X	3
4	276729	0521795028	6
6	276736	3257224281	8
7	276737	0600570967	6
8	276744	038550120X	7
9	276745	342310538	10
16	276747	0060517794	9
19	276747	0671537458	9
20	276747	0679776818	8

Range Index: 1149780 entries, 0 to 1149779 Data columns (total 3 columns):

EDA- Boxplot And Distribution Plot Of Age

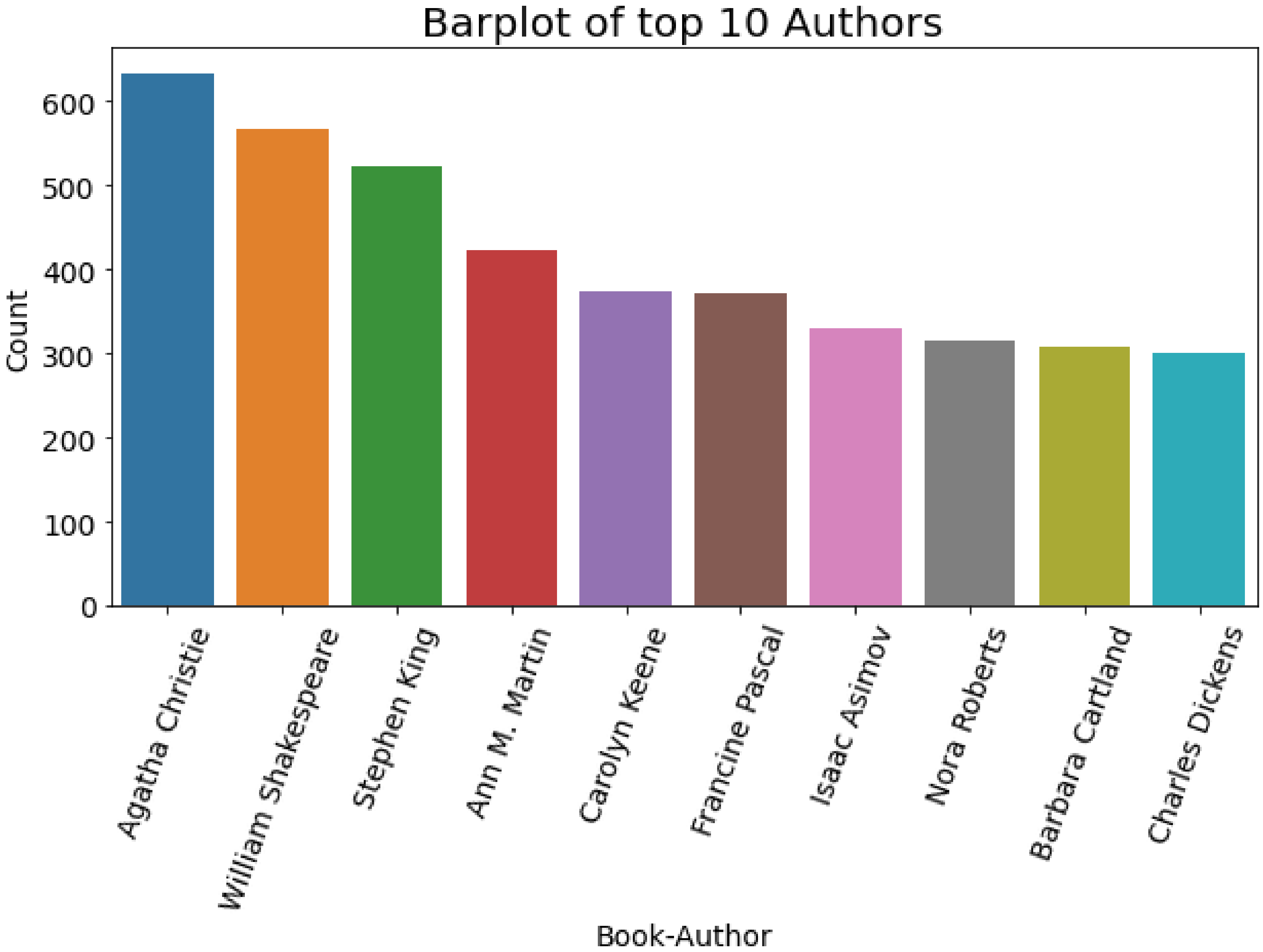


Top 10 Sold Books



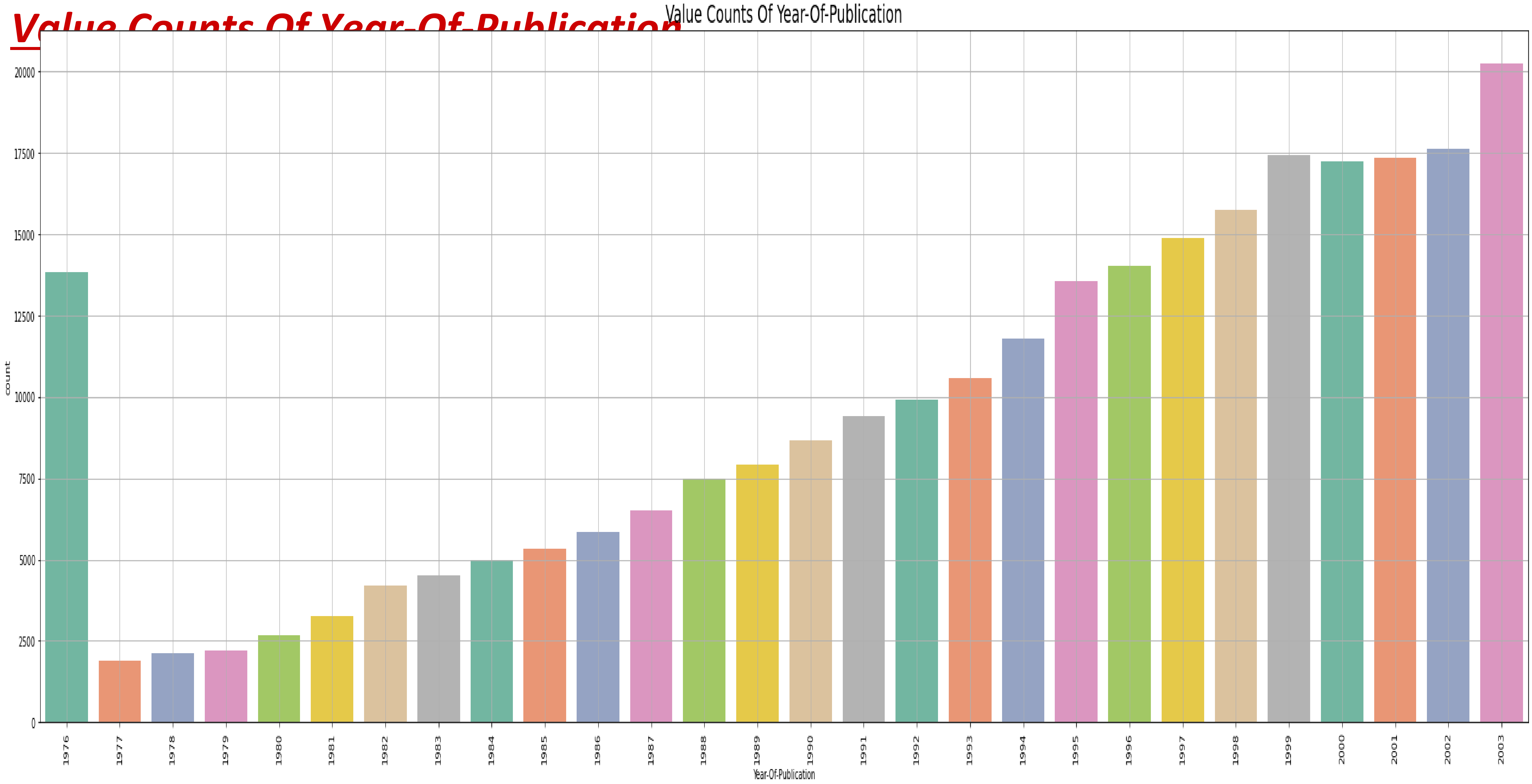
Wild Animus is the best-selling book

Top 10 Author

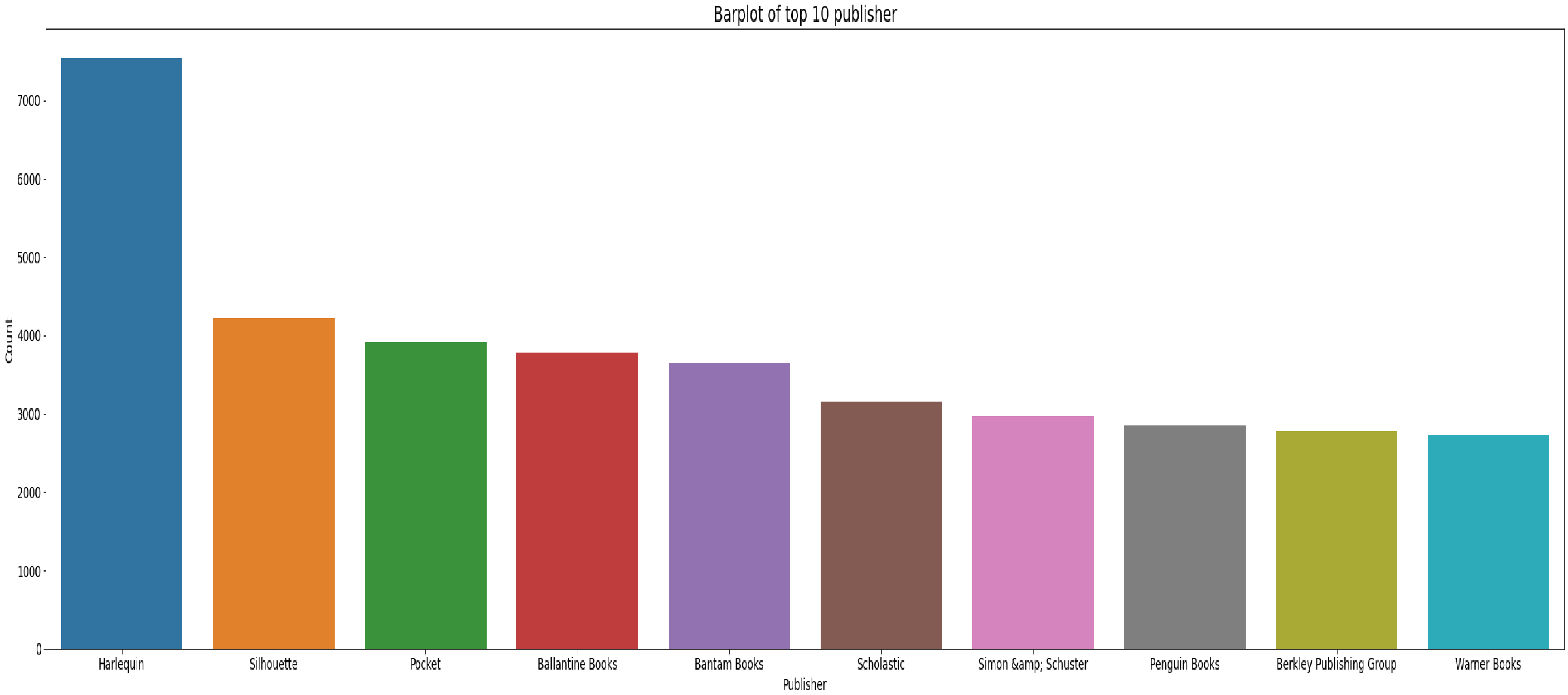


Author Agatha Christie, William Shakespeare and Stephen King wrote most of the books

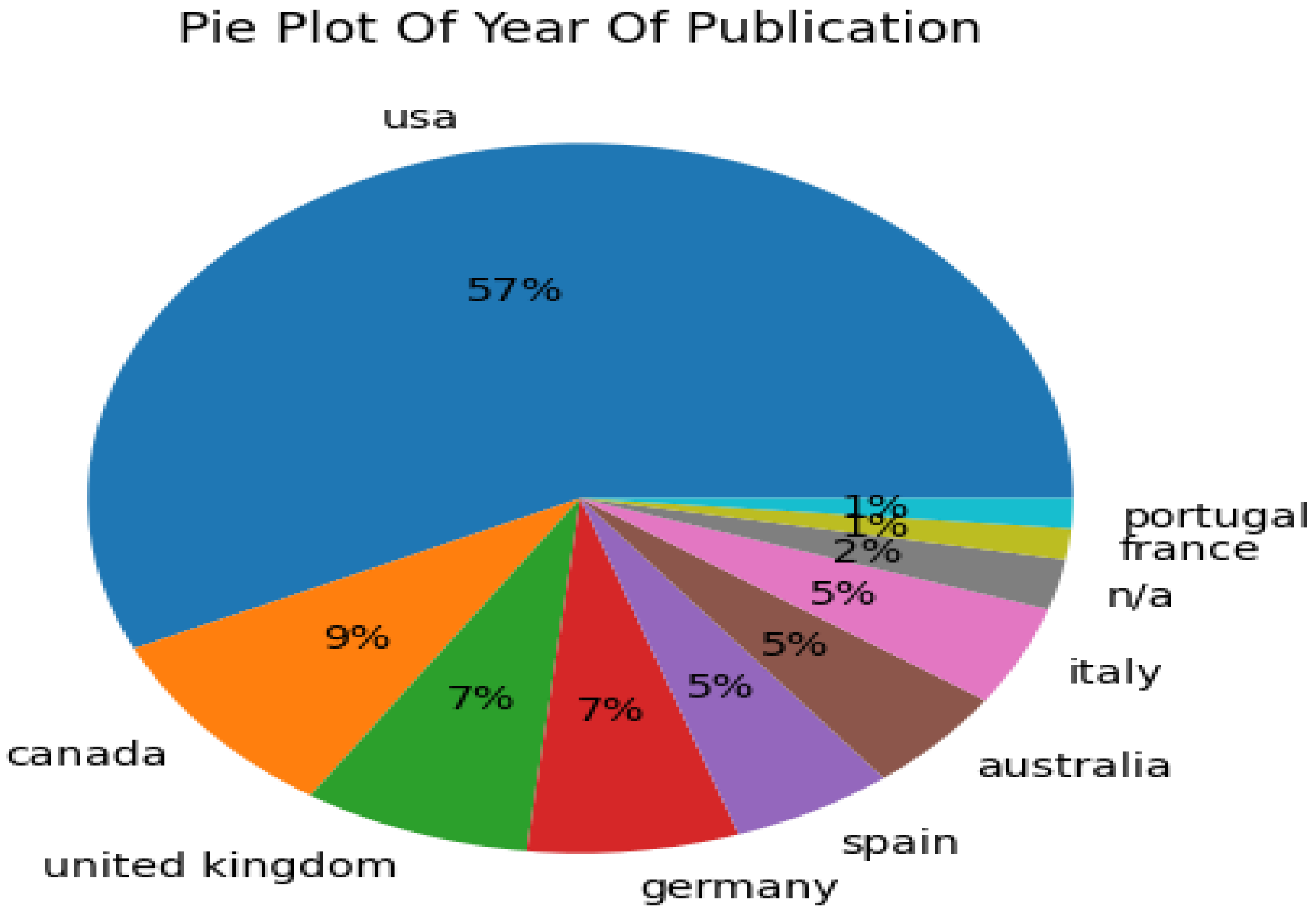
Value Counts Of Year-Of-Publication



Top 10 publisher



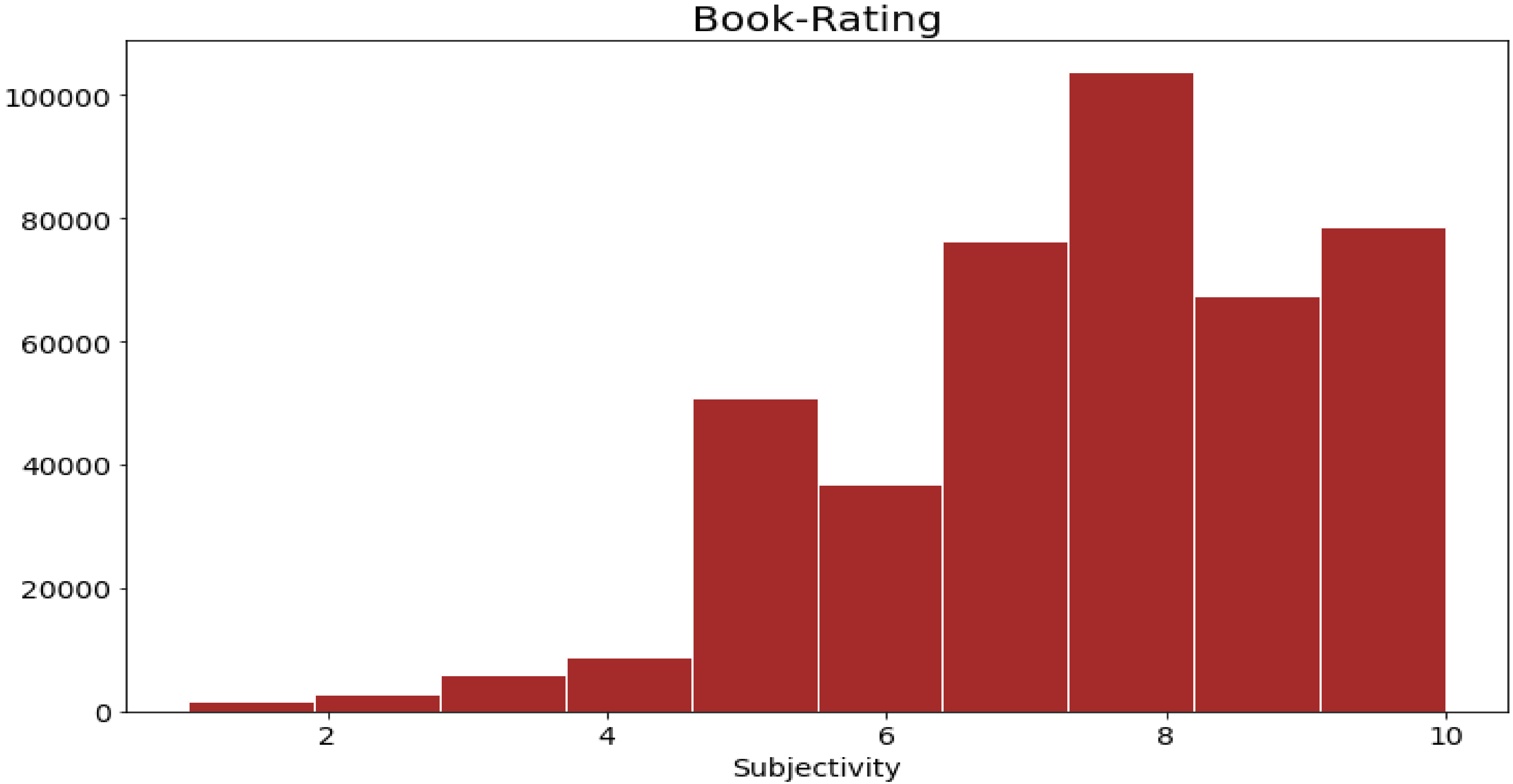
Harlequin publication published the most books



Top 10 Country

More than 50% readers are from USA

Histogram Of Book-Ratings(Excluding 0 rating count)



Book-Ratings are negatively distributed with median rating of 8

Collaborative Filtering

- To address some of the limitations of content-based filtering, collaborative filtering uses *similarities between users and items simultaneously* to provide recommendations. This allows for serendipitous recommendations; that is, collaborative filtering models can recommend an item to user A based on the interests of a similar user B. Furthermore, the embeddings can be learned automatically, without relying on hand-engineering of features.

Collaborative Filtering - Model's Used

- **KNN**
- **SVD - Singular Value Decomposition**
- **SVD ++**
- **NMF - Non-negative matrix factorization**
- **Slope One**

Book Recommendations

Recommendations for Book Before and After:

1. Tishomingo Blues, recommendation score = 0.81735

2. Waiting : The True Confessions of a Waitress, recommendation score = 0.81501

3. Soul Mountain, recommendation score = 0.80676

4. Perfect Murder, Perfect Town, recommendation score = 0.80452

5. Politically Correct Holiday Stories: For an Enlightened Yuletide Season, recommendation score = 0.80446

6. A Promising Man (and About Time, Too), recommendation score = 0.8012

7. When He Was Wicked (Bridgeton Family Series), recommendation score = 0.80114

8. All-American Girl, recommendation score = 0.79553

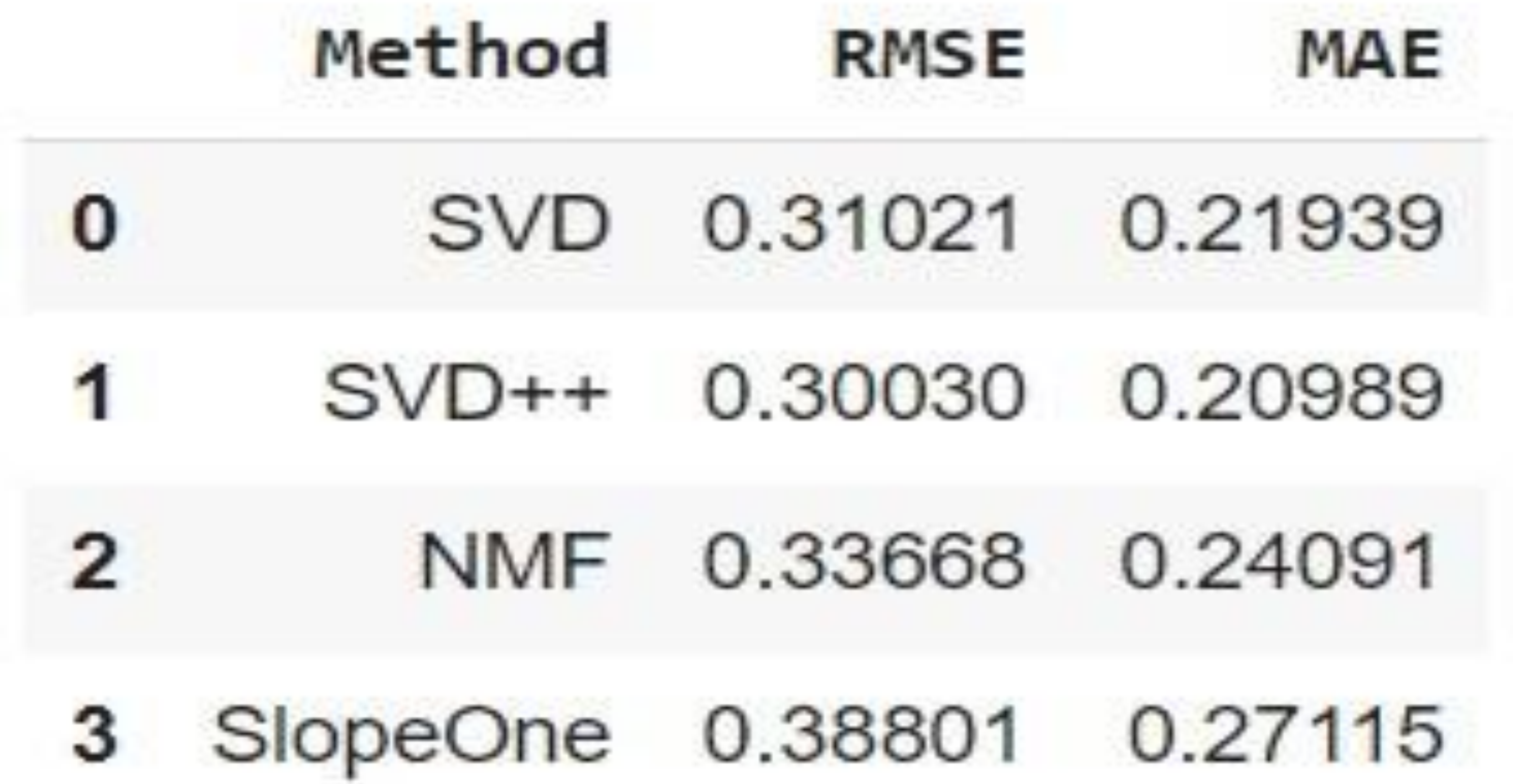
9. Night Watch, recommendation score = 0.79382

10. A Cook's Tour : Global Adventures in Extreme Cuisines, recommendation score = 0.77895

Evaluation Matric

- Explicit Feedback Recommender Systems -These are systems where the user gives explicit feedback, usually in the form of a numeric rating for each recommendation.
- Metrics used in Explicit Recommender Systems For such a system, the metrics used could be pretty similar to that used in a standard regression problem since the target is really a score that you could be predicting, and the actual score is available to measure how good the prediction is.
- **Mean Absolute Error**: Mean over all data points, absolute value of difference between actual rating and predicted rating.
- **Root Mean Square Error**: Square root of Mean over all data points, square of difference between the actual rating and predicted rating.

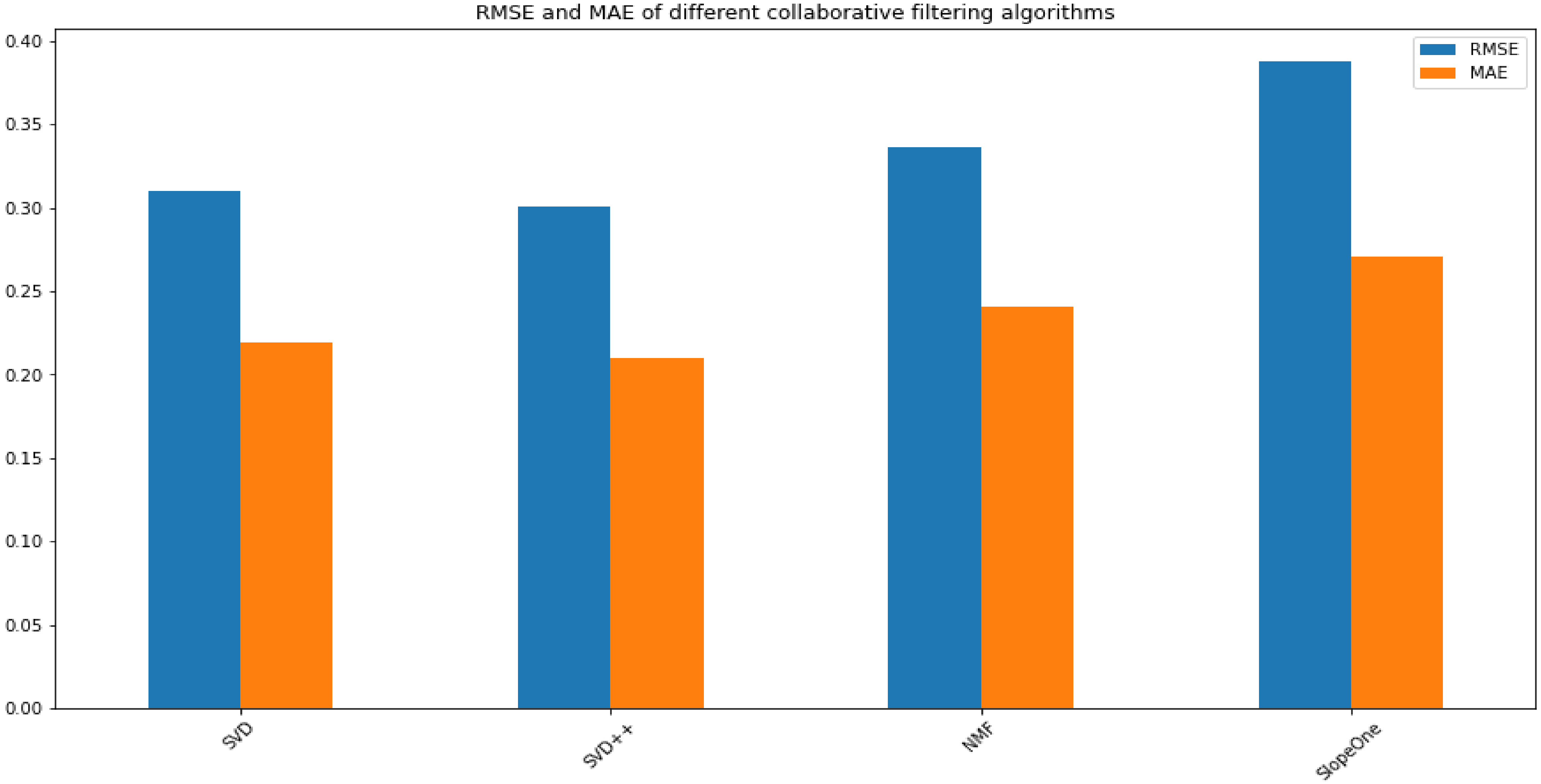
Evaluation of all models



	Method	RMSE	MAE
0	SVD	0.31021	0.21939
1	SVD++	0.30030	0.20989
2	NMF	0.33668	0.24091
3	SlopeOne	0.38801	0.27115

SVD++ is the best recommendation model with root mean squared error of 0.30 and mean absolute error of 0.20

Bar plot of evaluation of all models



Conclusion’s

- ✈ Wild Animus is the best-selling book
- ✈ Author Agatha Christie, William Shakespeare and Stephen King wrote most of the books
- ✈ Harlequin publication published the most books
- ✈ More than 50% readers are from USA
- ✈ Book-Ratings are negatively distributed with median rating of 8.
- ✈ Root mean squared error of model **SVD** is 0.31 and mean absolute error is 0.21
- ✈ Root mean squared error of model **NMF** is 0.34 and mean absolute error is 0.24
- ✈ Root mean squared error of model **Slope One** is 0.39 and mean absolute error is 0.27
- ✈ **SVD++** is the **best recommendation model** with root mean squared error of 0.30 and mean absolute error of 0.20

THANK
YOU