

# Book Recommendation System

By: Farzad Radmehr

# Problem Statement

- Recommending top 10 books to the readers based on their rating history
- Applying 3 different recommendation filtering and compare the RMSE to determine the best method and the best features to determine the top 10 books

# Data Description

- This data was collected from Book\_Crossing community with permission from Ron Hornbaker, CTO of Humankind Systems.
- 3 different tables were used:
  - User Info
  - Book Info
  - Rating Info
- This data contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

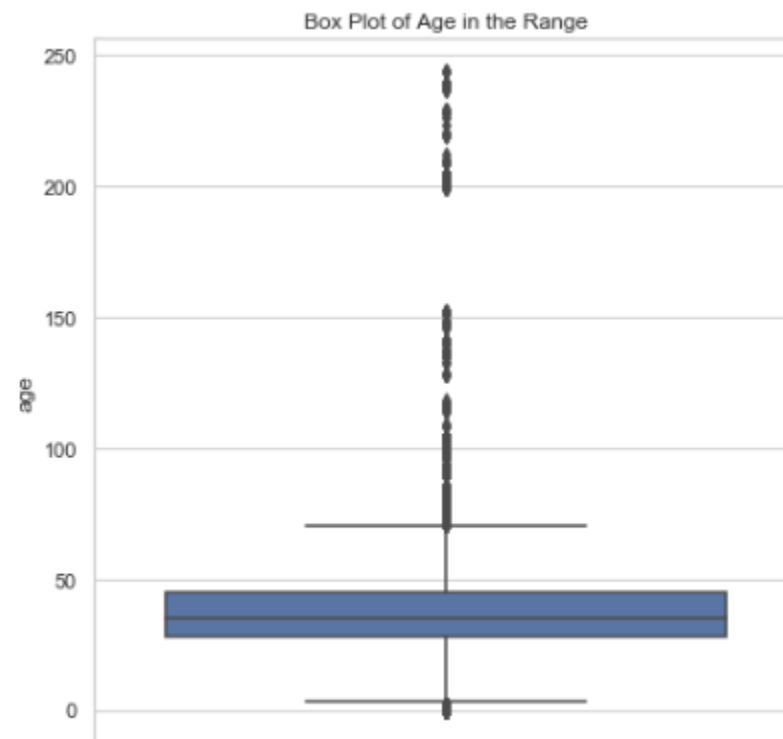
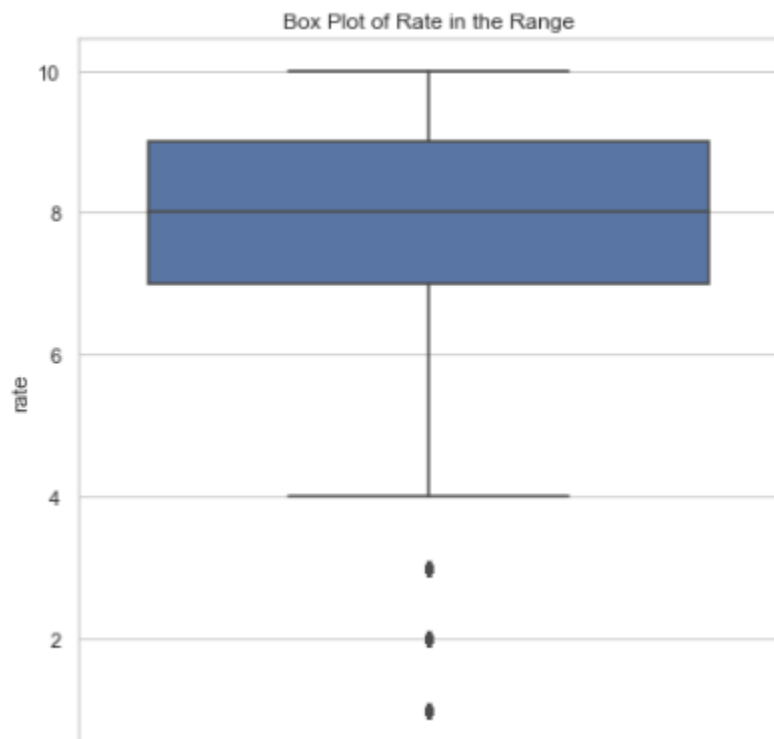
# Features

- Users: User IDs, Location, Age
- Books: ISBN, Book title, Book Author, Year of Publication, Publisher, Image\_url\_S, Image\_url\_M, Image\_url\_L
- Rating: User IDs, ISBN, Book Rating

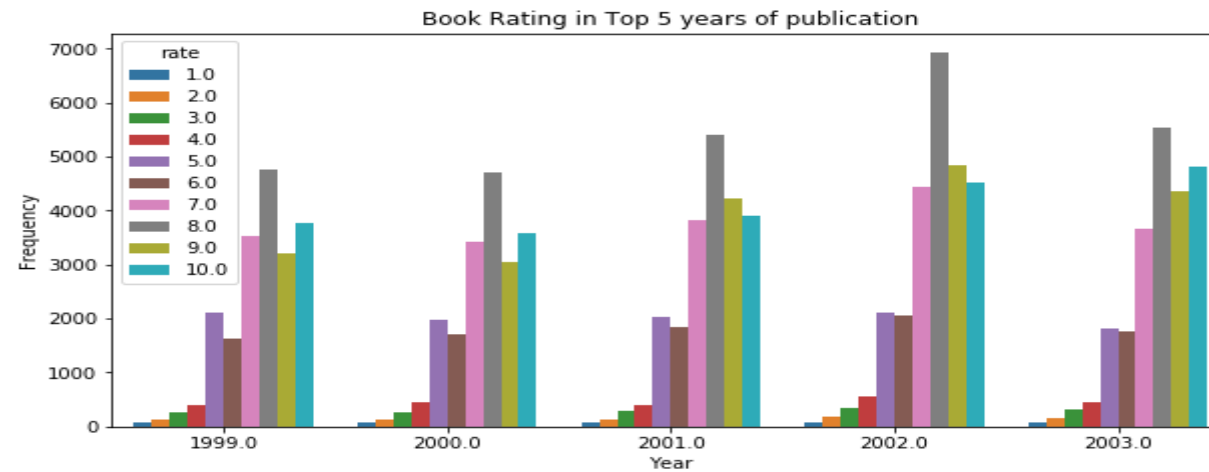
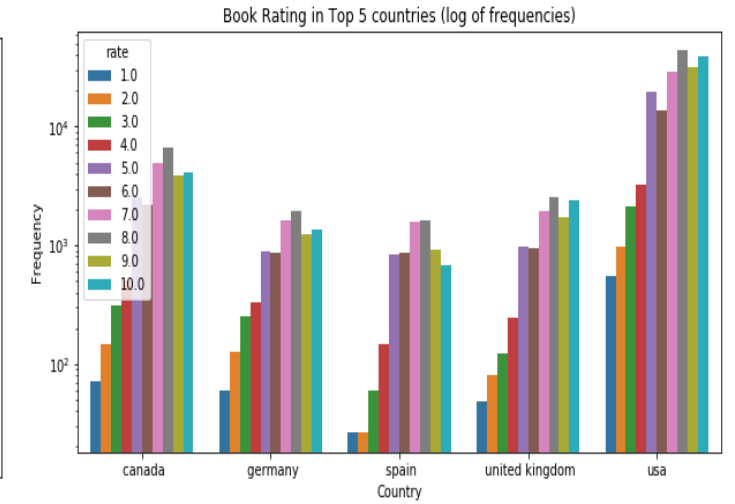
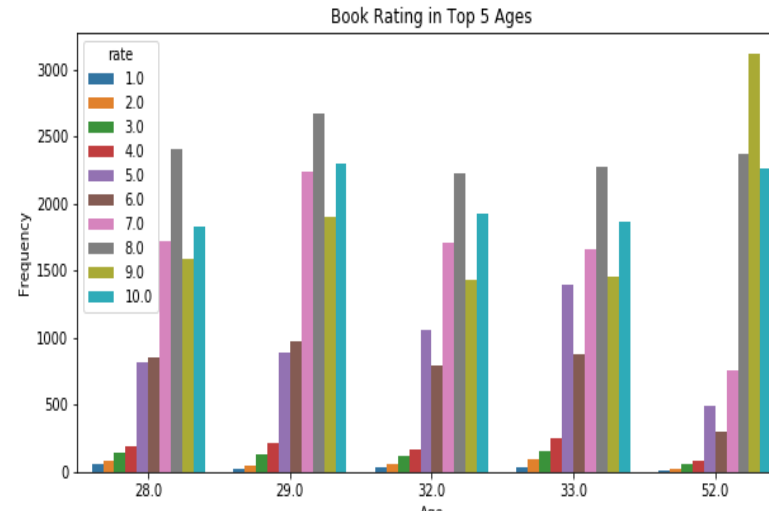
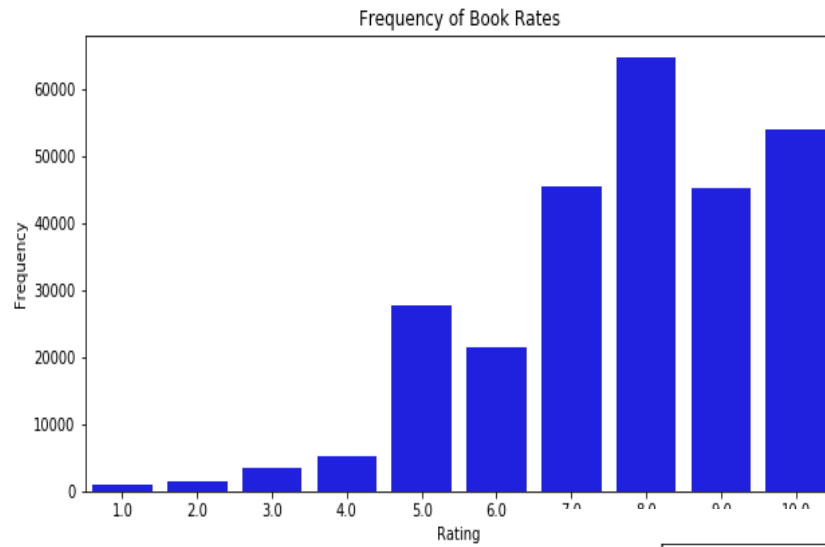
# Data Cleaning

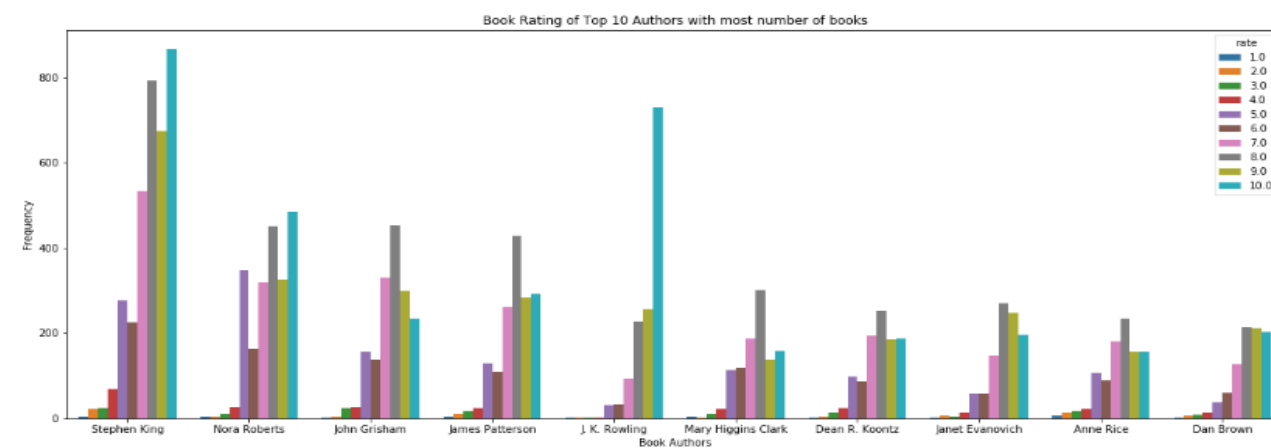
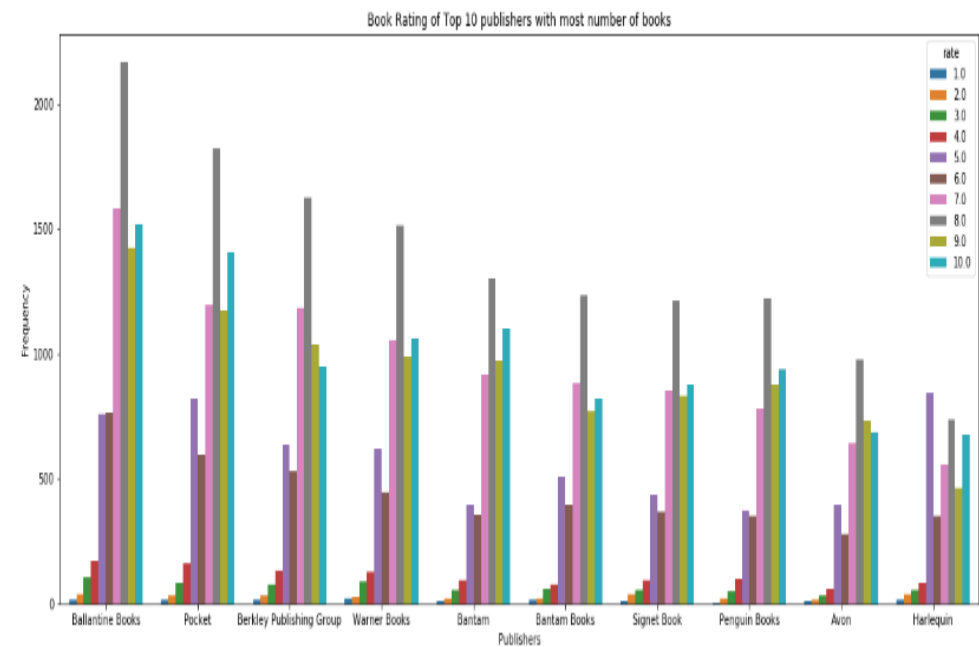
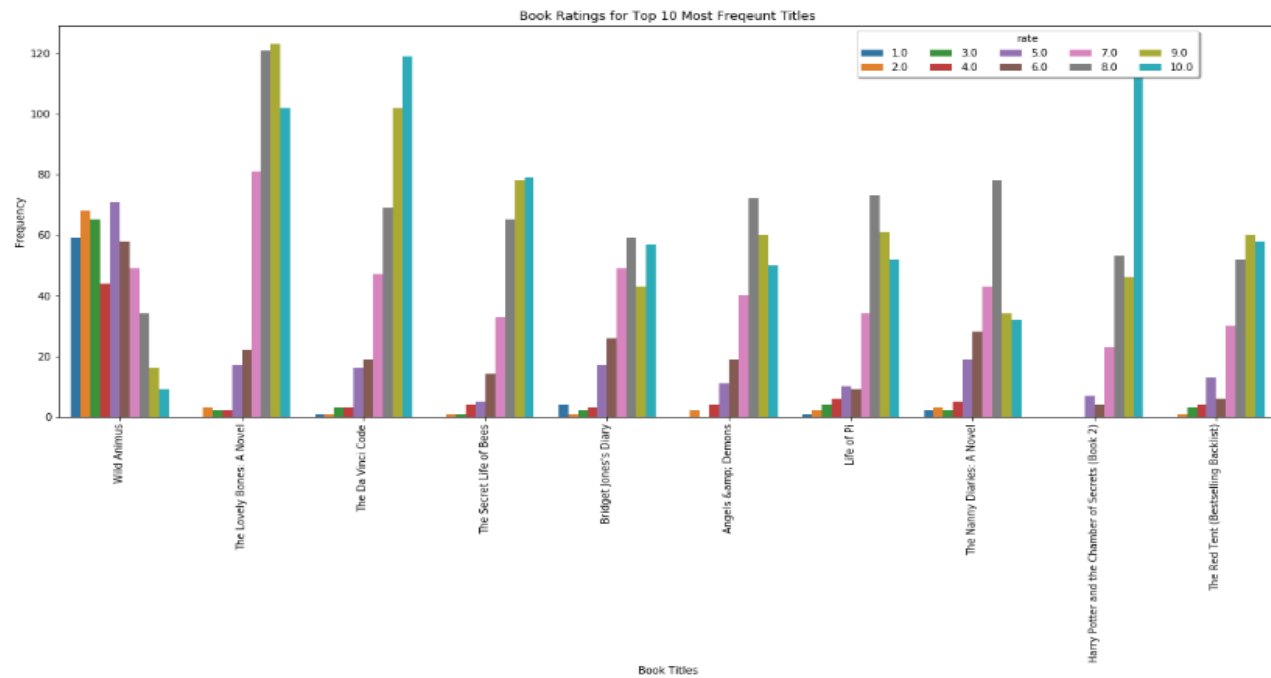
- Merged the datasets
- Removed the image columns from the Books Info dataset
- By Checking all unique values of each column, I fixed the misplacements
- Filled and fixed the missing values
- Studied the outliers

# Outlier Study



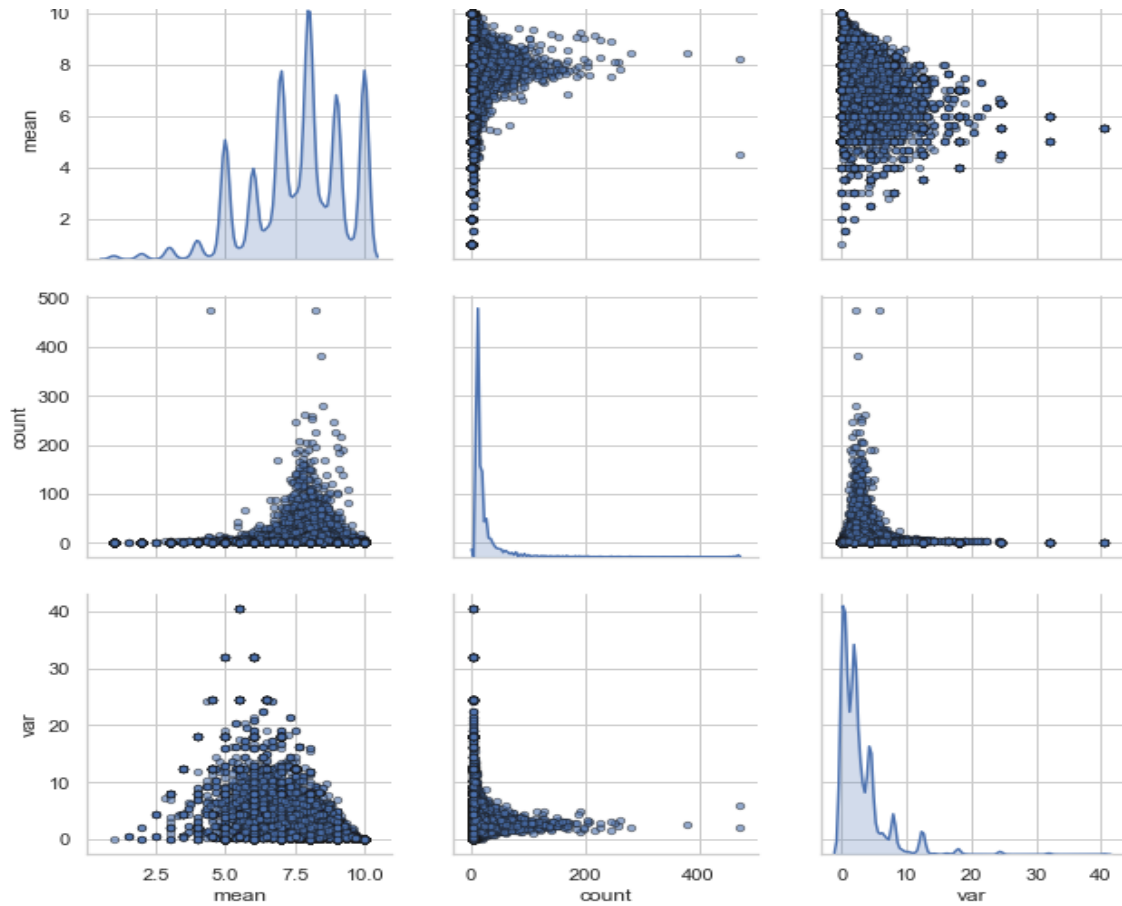
# Data Visualization



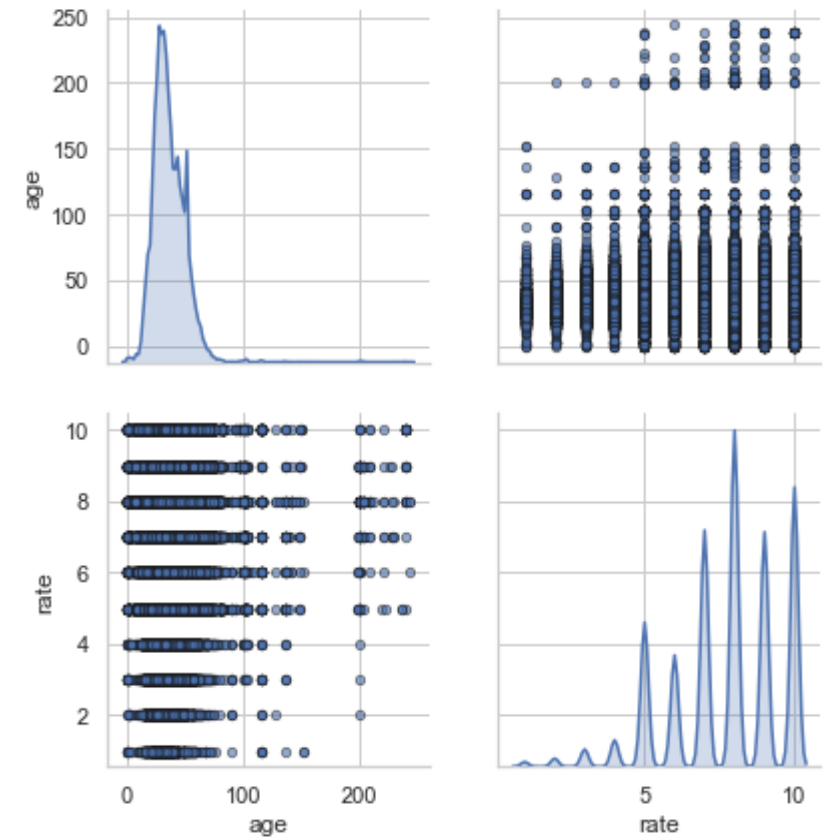




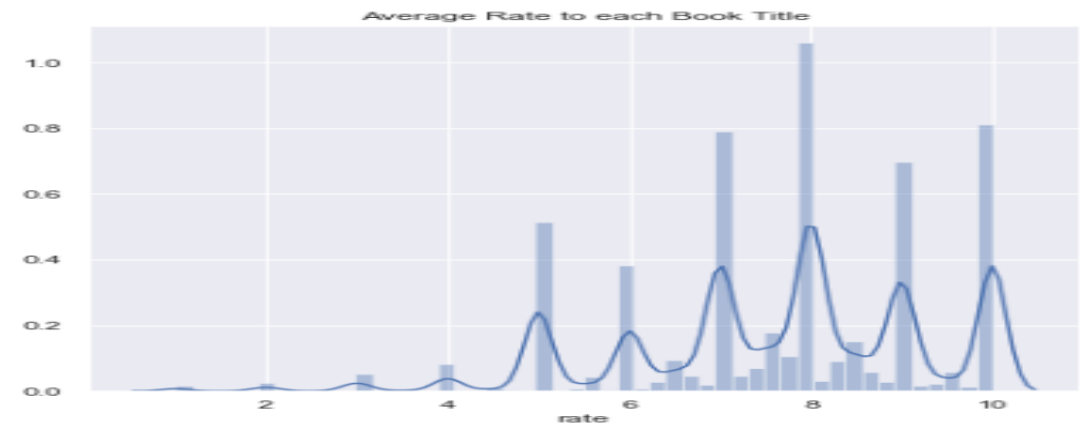
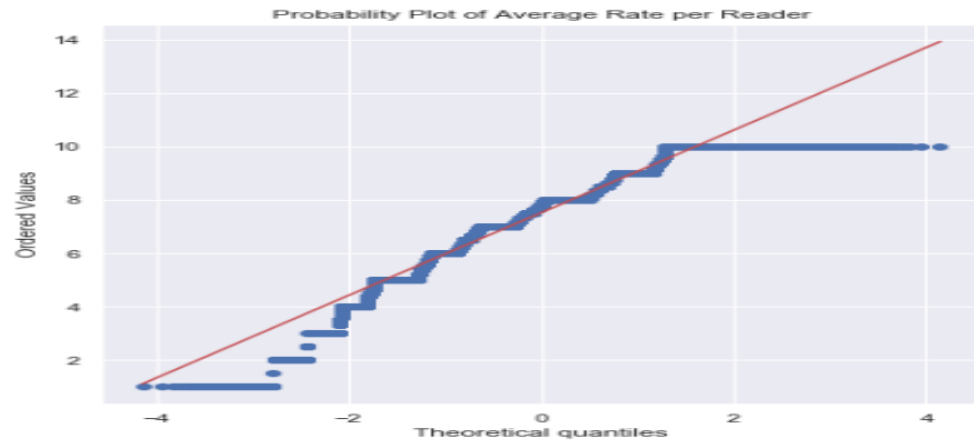
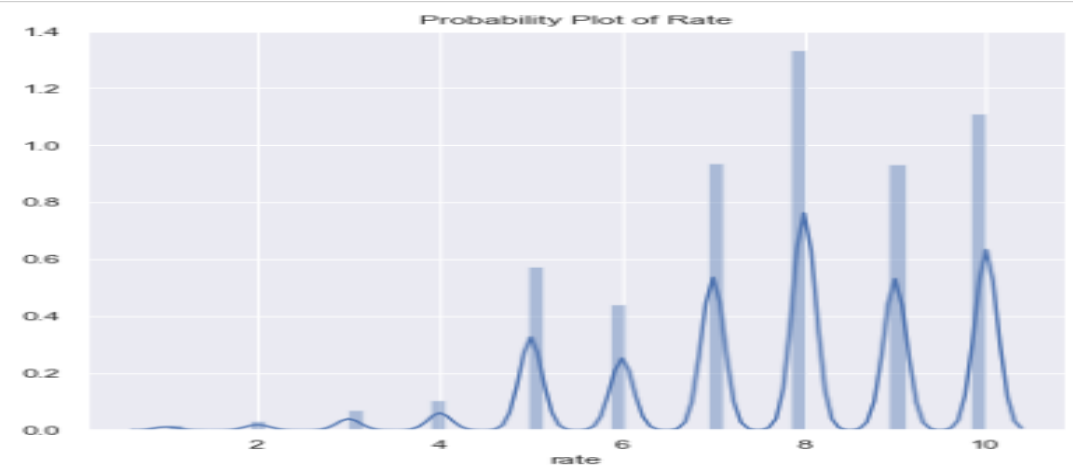
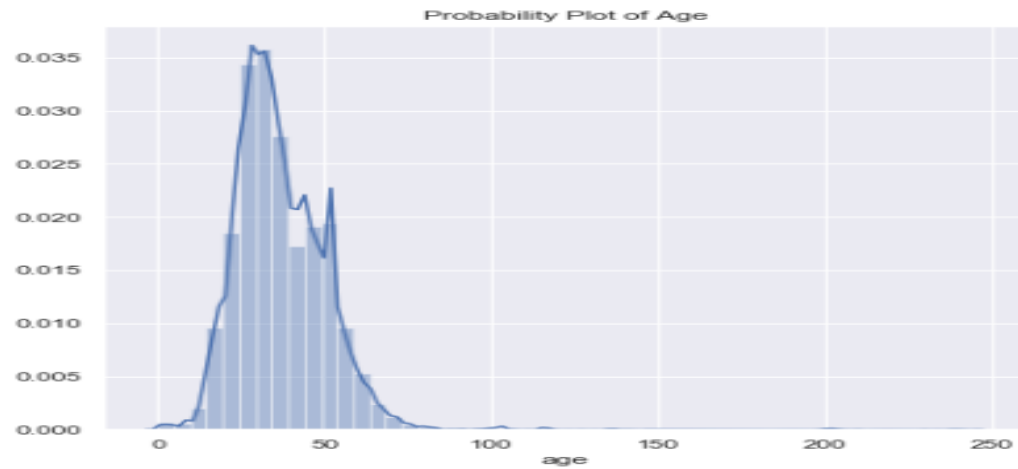
# Stats in Pair Plot



Age and Rate Plots



# Probability Plots



- Majority of readers gave the rate of 8 by and 10,7 and 9 are the next popular rates.
- Age range of 28-33 and 52 is the age range of readers. they give higher rates to the books as well.
- story and novel books have the highest popularity. Da Vinci Code is one of the popular books. the rates for each one of the top 10 books have been visualized as well.
- Stephen King is the most popular author with rates of 10 from many of the readers. top 10 Authors and their rates have been presented as well.
- Ballentine book is the most frequent publisher. Top 10 publishers and their rates have been presented as well.

- USA has the highest number of books. Top 5 countries have been presented.
- Majority of books have been written between 1999-2003.
- few outliers have been detected by using box plots, but they are in normal range, so they are not removed from dataset.
- Mean, Count and Variance plots show they are skewed to one side.
- Probability plots don't show any distribution.
- Majority of books have just few rates in dataset. some of the book have many rates.

# Hypothesis Testing Conclusions

- Overall, younger Adults have significantly more interest to read the books. for example, 19-25 range have significantly more interest (and highest rates) compared to 25-35 and 35-45. range of 9-19 has less interest to read the books.
- People don't have interest to older books. the books that published more recently have the highest rates. we can see a pattern by studying the year of publication. Full results have been presented in the table.

# Modeling

- I used Content Based Filtering, Collaborative Based Filtering and Hybrid Based Filtering Methods
- Following Features are studied in each method:
  - Content Based Filtering: Book Title, Book Author, Year of Publication, Publisher
  - Collaborative Based Filtering: User, Age, State, Country
  - Hybrid Based Filtering:
    - ✓ Age, State, Country
    - ✓ All Features

# Modeling Results

Method	Features	RMSE
Content Based Filtering	Book Title	1.434
Content Based Filtering	Year of Publication	1.529
Content Based Filtering	Publisher	1.504
Content Based Filtering	Book Author	1.389
Collabrative Based Filtering	User	4.223
Collabrative Based Filtering	Age	4.231
Collabrative Based Filtering	State	4.229
Collabrative Based Filtering	Country	4.229
Hybrid Based Filtering	Age, State, Country	1.749
Hybrid Based Filtering	All Features	1.783

# Modeling Conclusion

- In each model, I started with a basic model (with no extra features) and then added more inputs.
- Content based model creates the best results. that's because the rate is more likely determined by variables in user table.
- Users give the highest rate considering Book Author, Year of publication and Publisher. location and age of users are not as important in the rating. this conclusion seems logical.
- Out of all variables, Book Author has the highest importance in book rating. Next in the list is Book Titles.



# Future Studies

- The fact that in this dataset, Book Author/Titles (and content based Filtering) play important roles, shows that we can add more features to do the research.
- For the future research, we can add reader reviews and use NLP approaches to see how we can recommend books to the readers based on all these features and their review history.
- It's been great to create an app to produce the recommendations to each user by adding the info (Books/Rates) to this dataset. this way, we can expand the dataset as well.

**Thank you**

<https://github.com/fradmehr/Springboard-Capstone-Project-1>

<https://www.linkedin.com/in/farzad-radmehr-97132143/>