RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Assignment

Course Code	CSC402A
Course Name	Data Mining
Programme	B.Tech
Department	CSE
Faculty	FET

Name of the Student	Satyajit Ghana
Reg. No.	17ETCS002159
Semester/Year	07/2020
Course Leader(s)	Prof. Mohan Kumar

Declaration Sheet

Student Name	Satyajit Ghana		
Reg. No	17ETCS002159		
Programme	B.Tech	Semester/Year	07/2020
Course Code	CSC402A		
Course Title	Data Mining		
Course Date		to	
Course Leader	Prof. Mohan Kumar		

Declaration

The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly.

Signature of the Student		Date	
Submission date stamp (by Examination & Assessment Section)			

Signature of the Course Leader and date	Signature of the Reviewer and date

Contents

Declaration Sheet ii

Contents iii

List of Figures iv

1 Question 1 5

1.1 Data Cleaning: Redundant and Inconsistent Data 5

1.1.1 Inconsistent Data 5

1.1.2 Univariate Analysis 6

1.2 Data Cleaning: Missing Values and Outliers 6

1.2.1 Outlier Analysis 7

1.3 Data Normalization 11

1.3.1 Min-Max Scaling 11

1.3.2 Z-Score Standardization 12

1.3.3 Decimal Scaling 13

1.4 Data Transformation 14

1.4.1 Natural Log Transform 14

1.4.2 Square Root Transform 15

NOTE: age has a skewness of 0 using Square Root Transform ! 15

1.4.3 Inverse Square Root Transform 16

1.5 EDA and Interpretation of Results 17

Bibliography 20

List of Figures

Figure 1-1 Book Ratings, before and after removing 0 ratings.. 5

Figure 1-2 Book Rating Univariate Analysis 6

Figure 1-3 Age univariate analysis 6

Figure 1-4 Isolation Forest of Original Data.. 7

Figure 1-5 Dropping book_rating using IQR.. 8

Figure 1-6 Dropping Age using IQR.. 8

Figure 1-7 Isolation forest after dropping outliers with IQR... 8

Figure 1-8 Dropping book_rating with BoxCox 9

Figure 1-9 Dropping age with BoxCox 9

Figure 1-10 Isolation Forest after BoxCox... 9

Figure 1-11 Dropping book_rating with imputation..10

Figure 1-12 Dropping age with imputation..10

Figure 1-13 Isolation Forest after imputation..10

1 Question 1

Solution to Question No. 1 Part A

This contains a brief summary of the data, and its preprocessing, refer to the Jupyter Notebook output at the end of this for a complete study of data.

1.1 Data Cleaning: Redundant and Inconsistent Data

Column	Mean	Std	Min	Max	Skewness	Kurtosis
age	36.23	10.41	5	100	0.83	1.34
book_rating	2.83	3.85	0	10	0.75	-1.21

1.1.1 Inconsistent Data

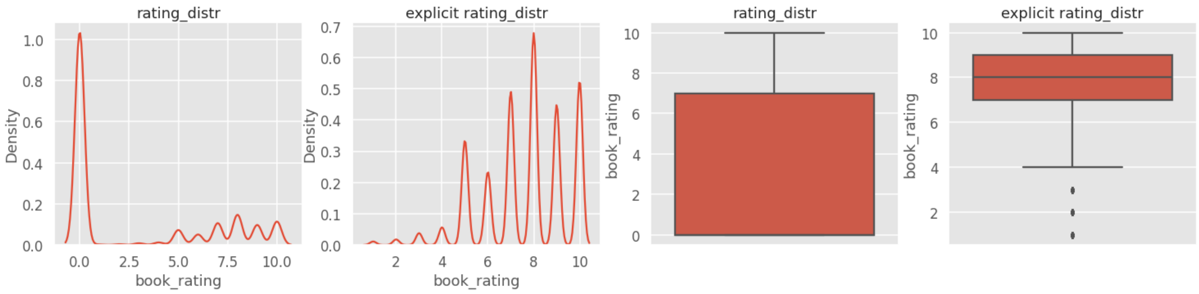


Figure 1-1 Book Ratings, before and after removing 0 ratings

Since 0 rated books done make sense they were removed, after removing our skewness and kurtosis values have changed a lot.

Column	Mean	Std	Min	Max	Skewness	Kurtosis
age	36.23	10.36	5	100	0.85	1.64
book_rating	2.83	3.85	1	10	-0.66	-0.12

1.1.2 Univariate Analysis

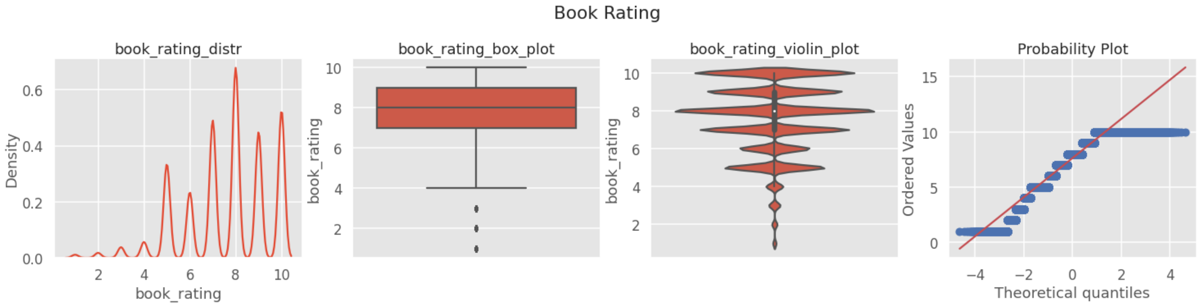


Figure 1-2 Book Rating Univariate Analysis

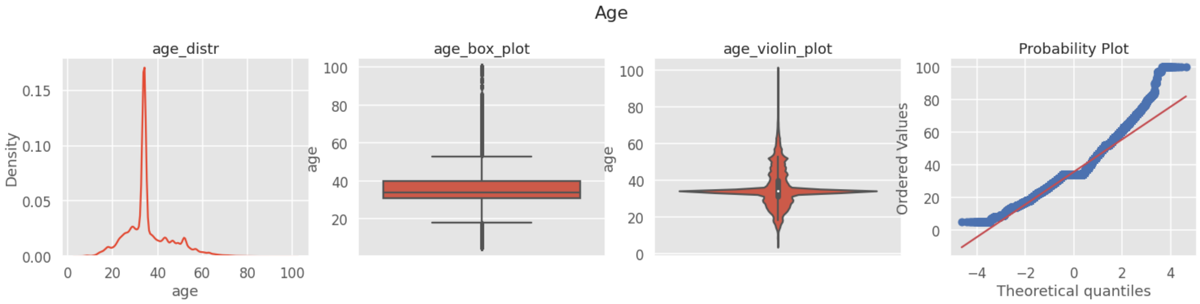


Figure 1-3 Age univariate analysis

1.2 Data Cleaning: Missing Values and Outliers

Refer Jupyter Notebook for Cleaning up Missing Values

1.2.1 Outlier Analysis

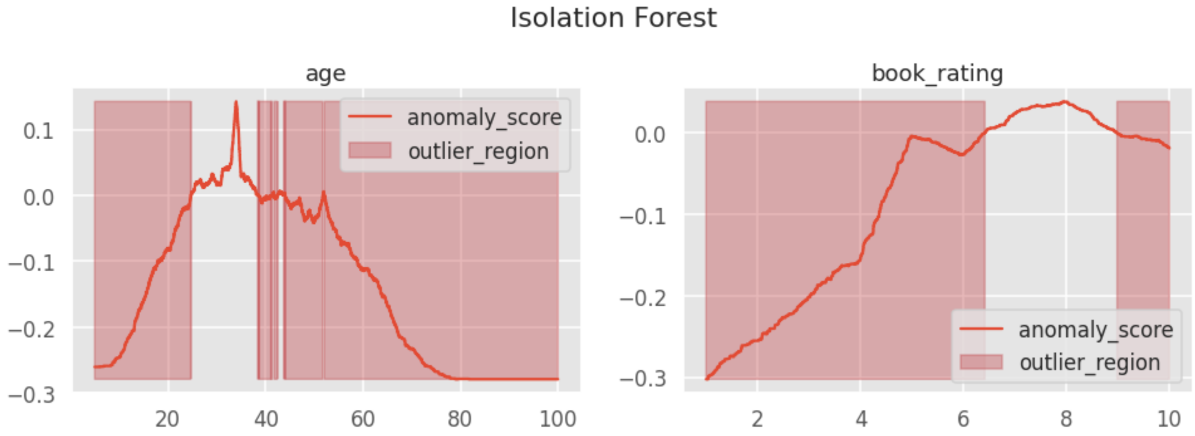


Figure 1-4 Isolation Forest of Original Data

After Dropping Outliers using IQR

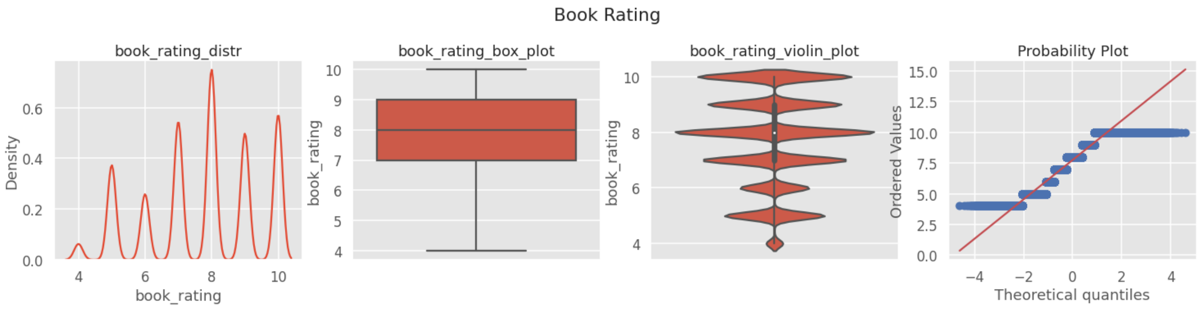


Figure 1-5 Dropping book_rating using IQR

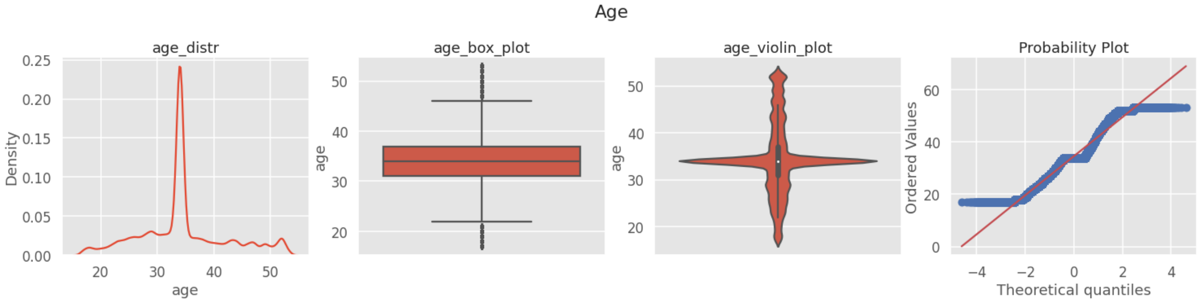


Figure 1-6 Dropping Age using IQR

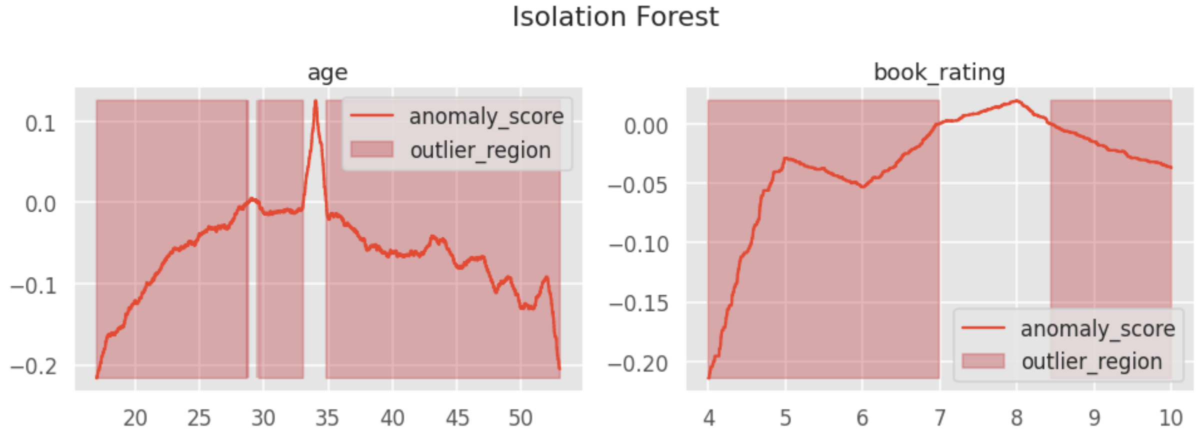


Figure 1-7 Isolation forest after dropping outliers with IQR

Column	Mean	Std	Min	Max	Skewness	Kurtosis
age	36.53	7.69	17	53	0.35	0.16
book_rating	7.74	1.66	4	10	-0.34	-0.80

Removing Outliers with BoxCox

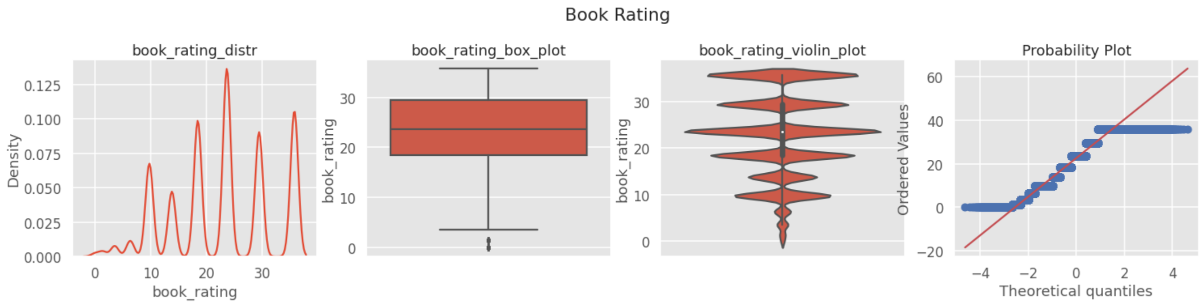


Figure 1-8 Dropping book_rating with BoxCox

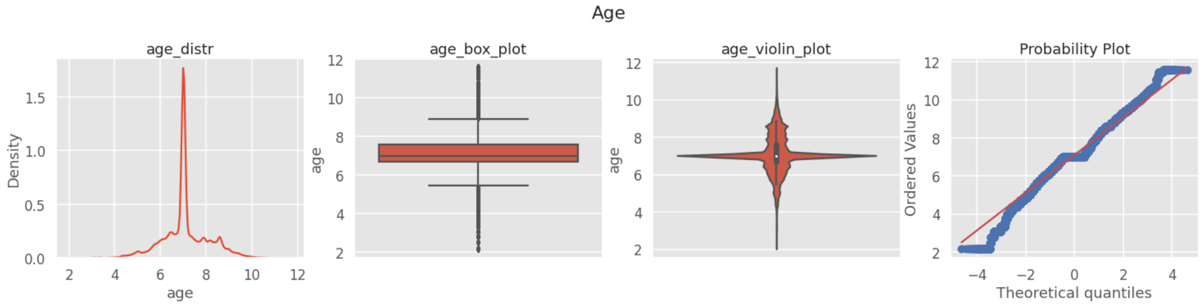


Figure 1-9 Dropping age with BoxCox

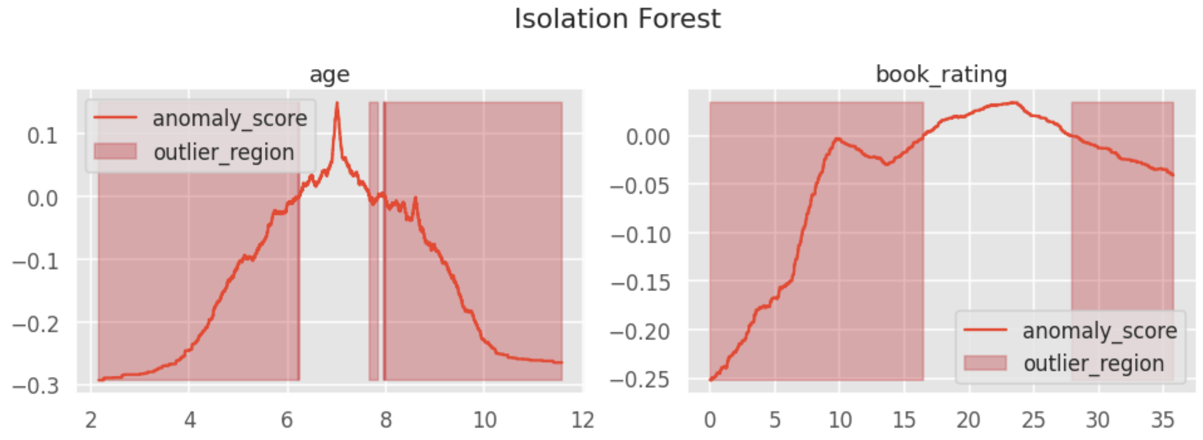


Figure 1-10 Isolation Forest after BoxCox

Column	Mean	Std	Min	Max	Skewness	Kurtosis
age	7.09	1.01	2.16	11.57	0.04	1.14
book_rating	22.62	9.14	0	35	-0.17	-0.83

Removing Outliers with Imputation

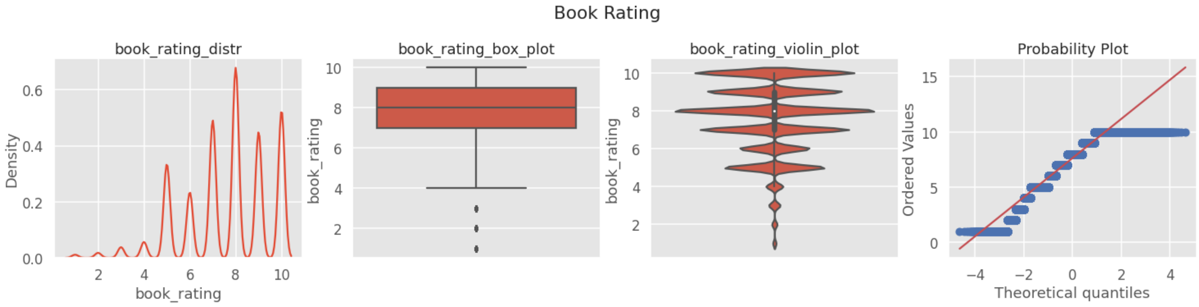


Figure 1-11 Dropping book_rating with imputation

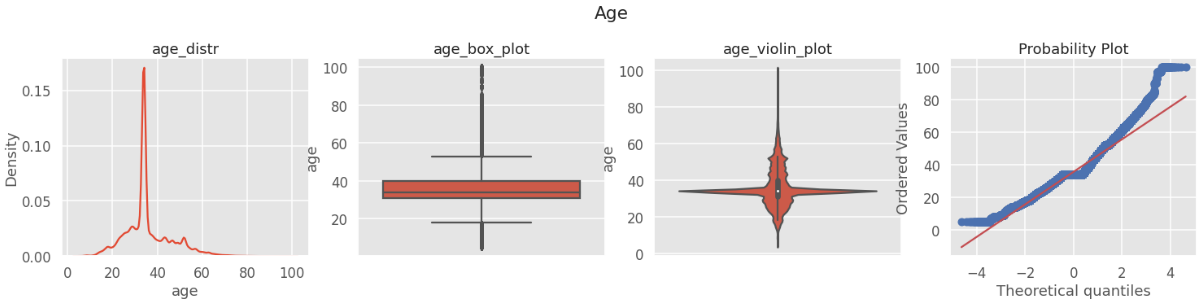


Figure 1-12 Dropping age with imputation

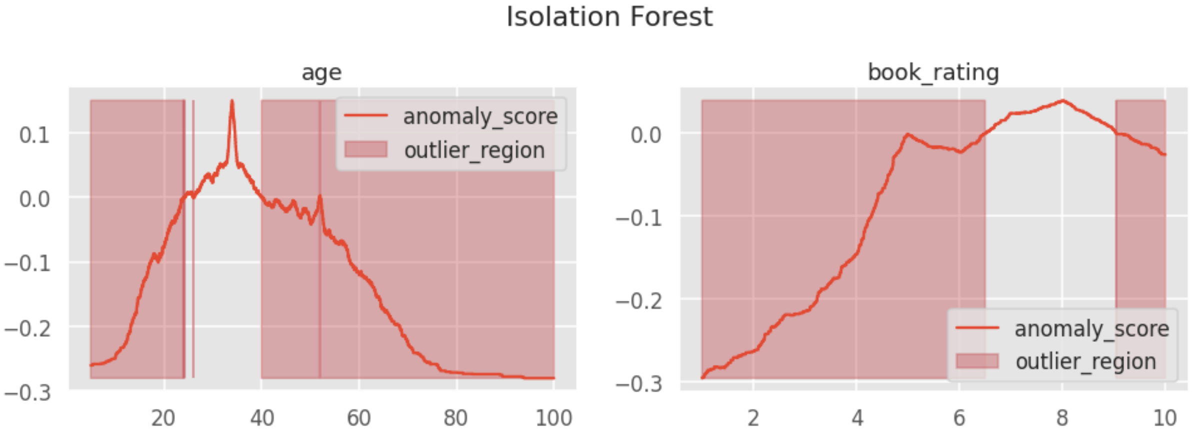
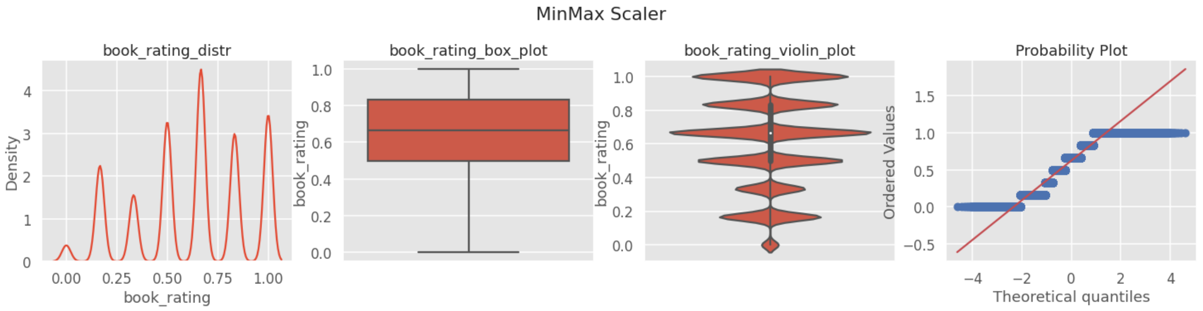


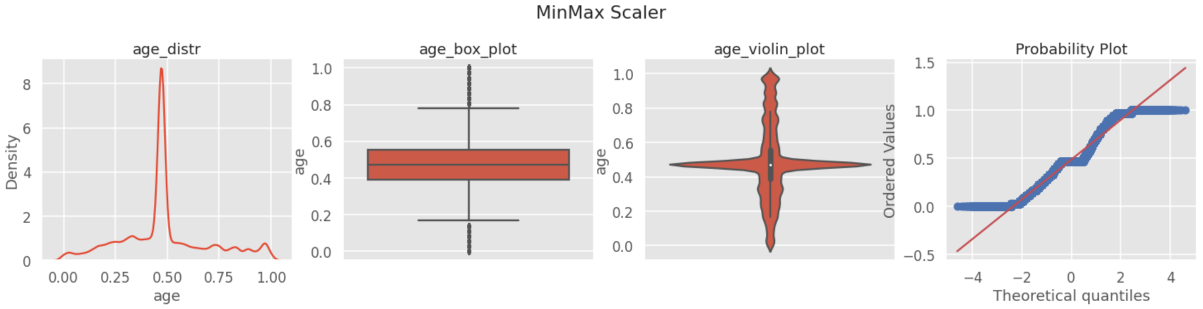
Figure 1-13 Isolation Forest after imputation

Column	Mean	Std	Min	Max	Skewness	Kurtosis
age	35.85	10.36	5	100	0.86	1.64
book_rating	7.62	1.83	1	10	-0.66	-0.12

1.3 Data Normalization

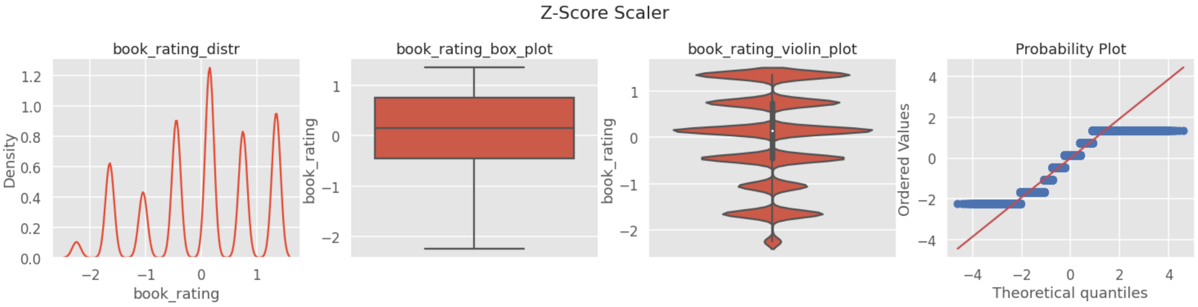
1.3.1 Min-Max Scaling

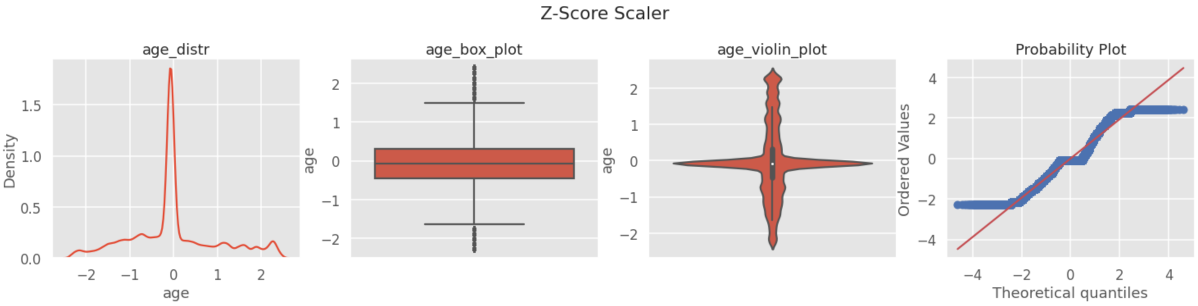




	Mean	Std	Min	Max	Skewness	Kurtosis
book_rating	0.62	0.27	0	1	-0.34	-0.80
age	0.48	0.21	0	1	0.35	0.17

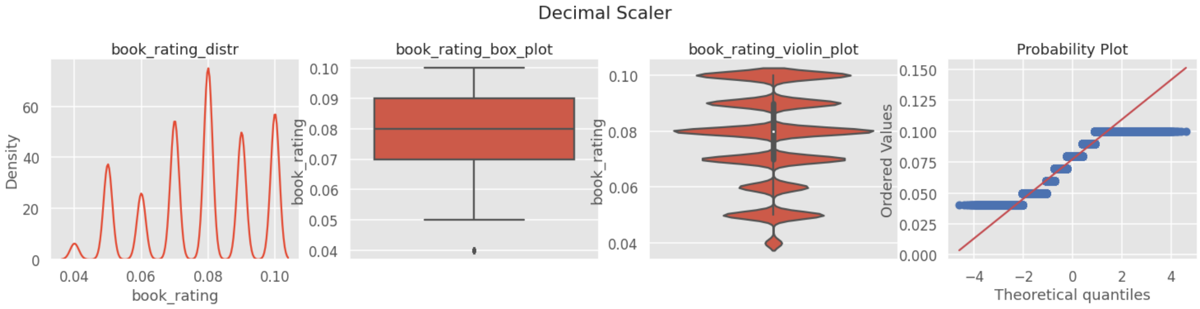
1.3.2 Z-Score Standardization

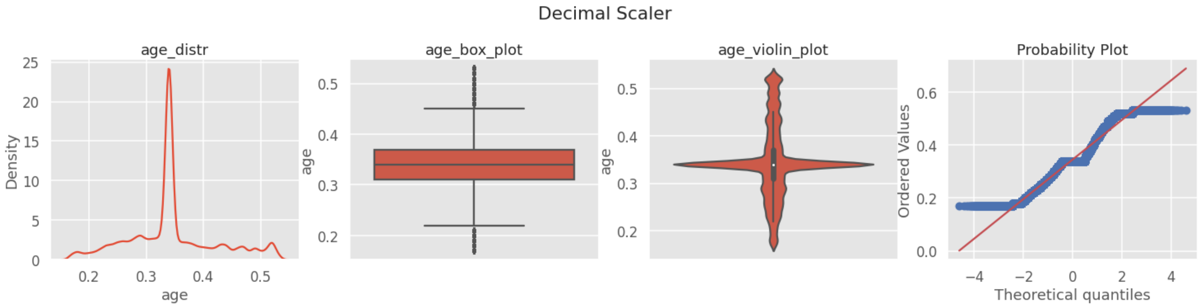




	Mean	Std	Min	Max	Skewness	Kurtosis
book_rating	~0	~1	-2.24	1.35	-0.34	-0.80
age	~0	~1	-2.27	2.39	0.35	0.17

1.3.3 Decimal Scaling

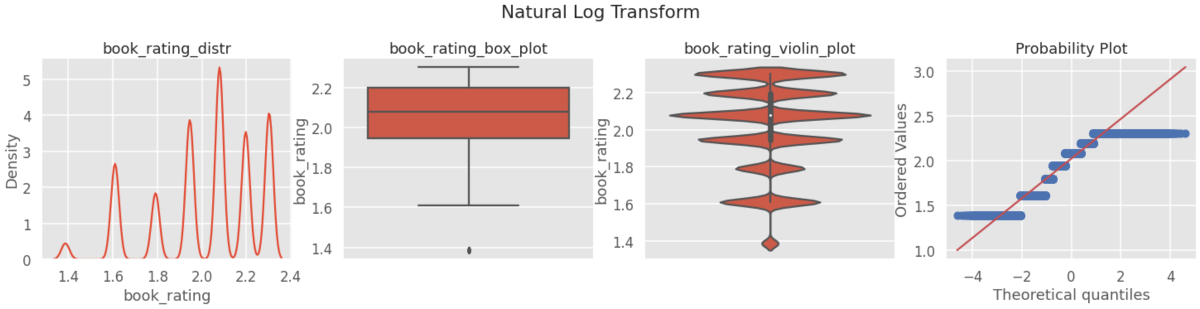


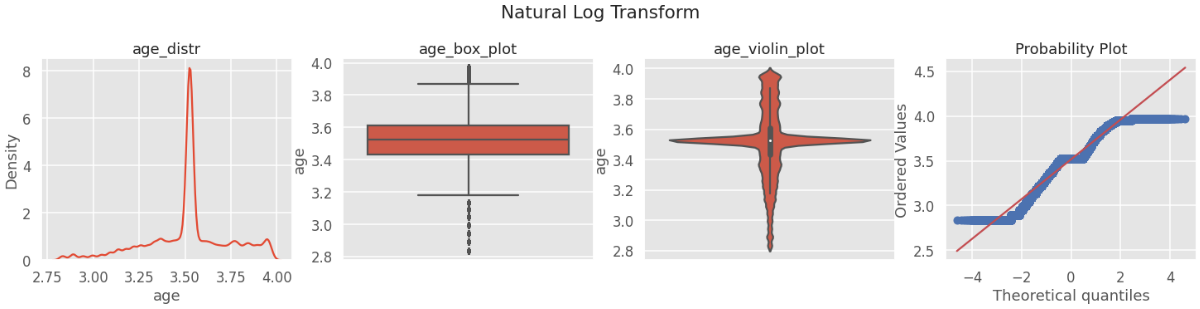


	Mean	Std	Min	Max	Skewness	Kurtosis
book_rating	0.077	0.01	0.04	0.1	-0.34	-0.80
age	0.34	0.07	0.17	0.53	0.35	0.17

1.4 Data Transformation

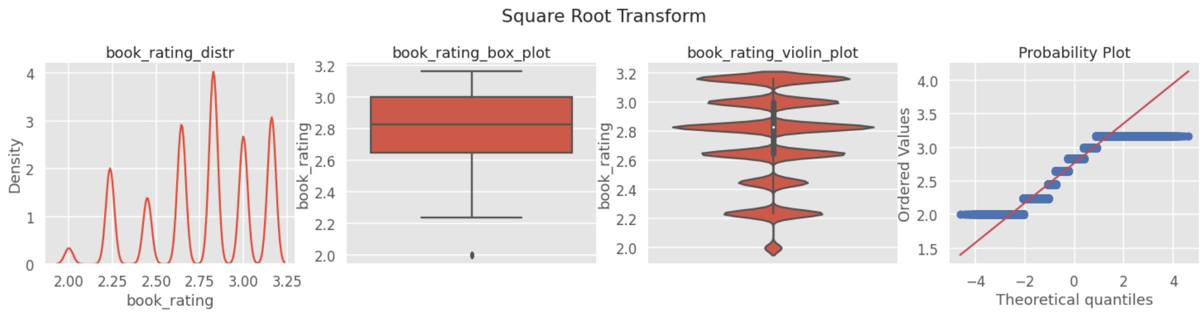
1.4.1 Natural Log Transform

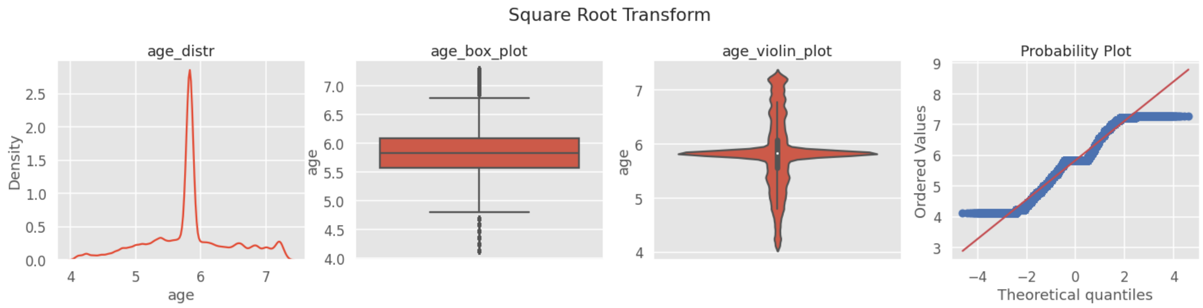




	Mean	Std	Min	Max	Skewness	Kurtosis
book_rating	2.02	0.23	1.38	2.3	-0.74	-0.23
age	3.511	0.229	2.83	3.97	-0.405	0.544

1.4.2 Square Root Transform

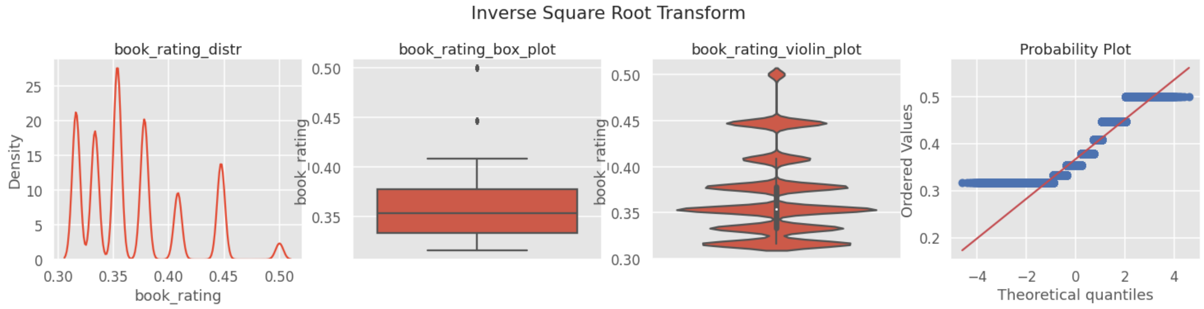


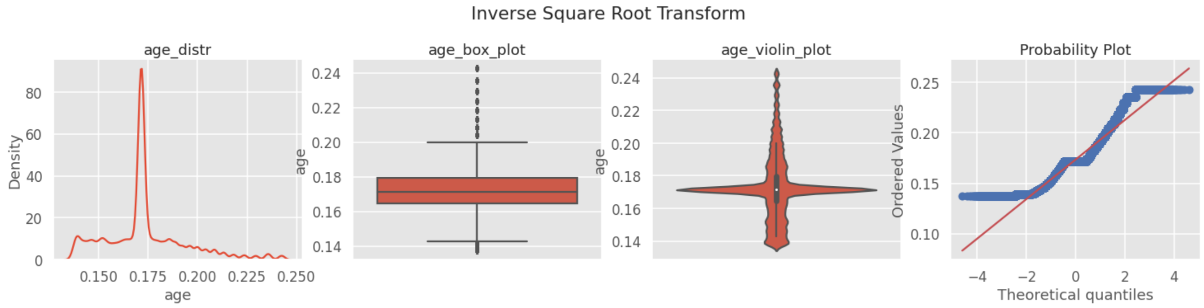


	Mean	Std	Min	Max	Skewness	Kurtosis
book_rating	2.76	0.31	2	3.16	-0.54	-0.57
age	5.84	0.65	4.123	7.28	-0.0089	-0.57

NOTE: age has a skewness of 0 using Square Root Transform !

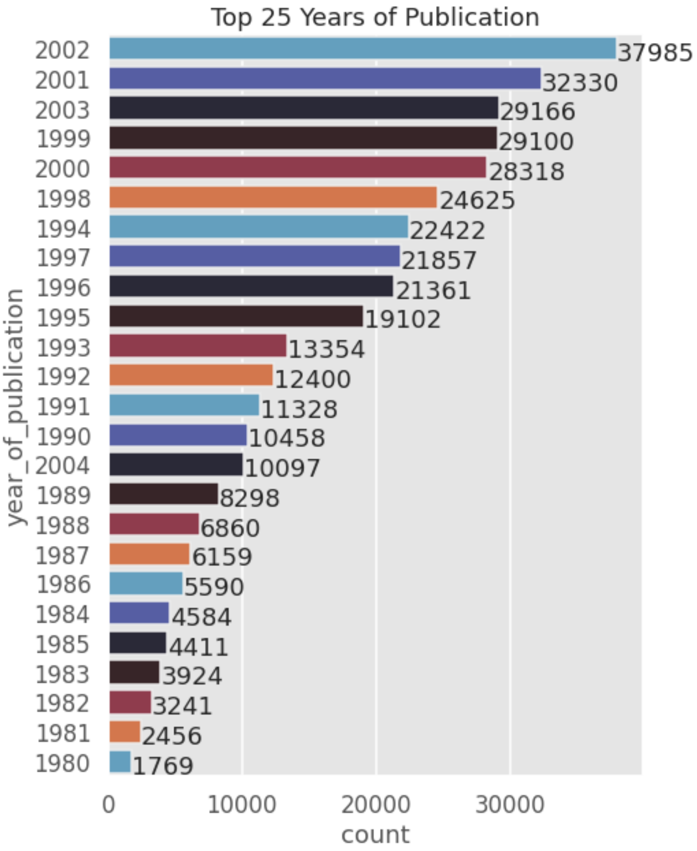
1.4.3 Inverse Square Root Transform

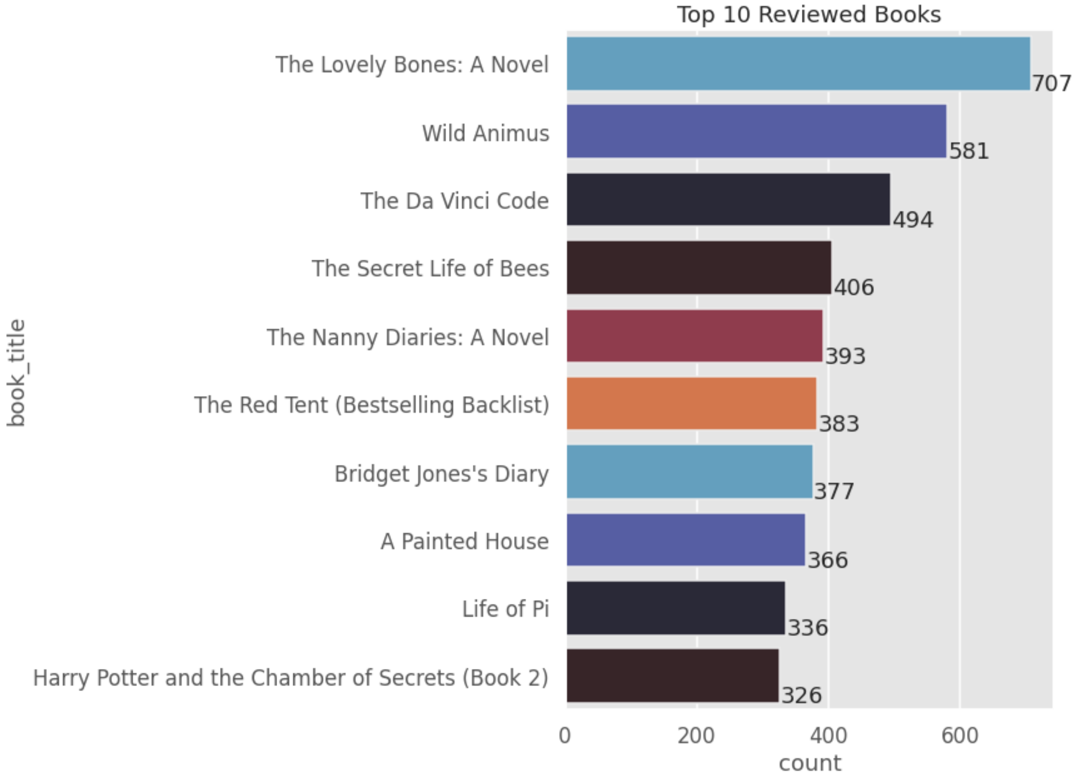


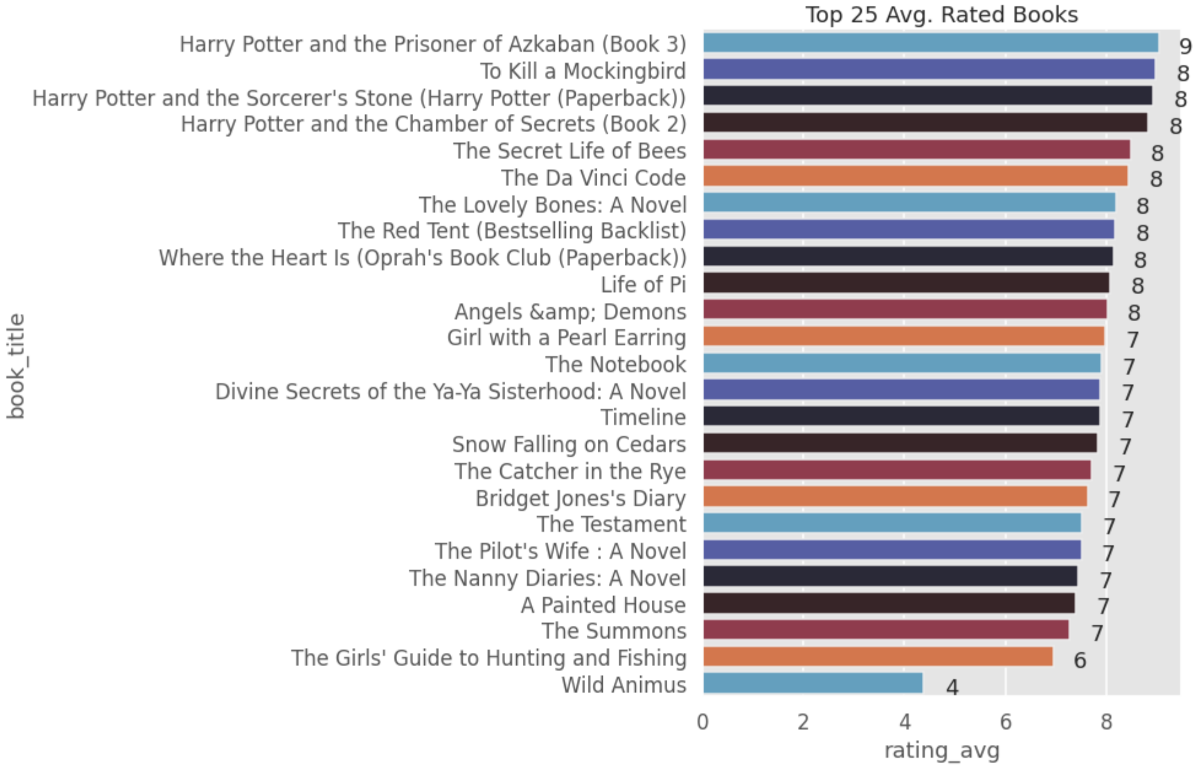


	Mean	Std	Min	Max	Skewness	Kurtosis
book_rating	0.366	0.044	0.31	0.5	0.96	0.24
age	0.17	0.02	0.13	0.24	0.82	1.21

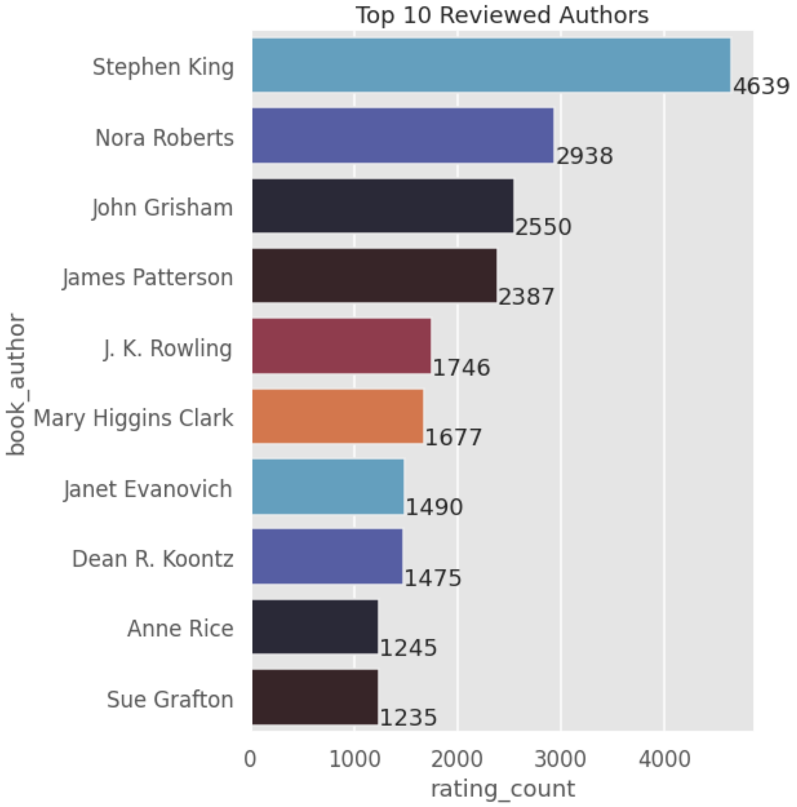
1.5 EDA and Interpretation of Results



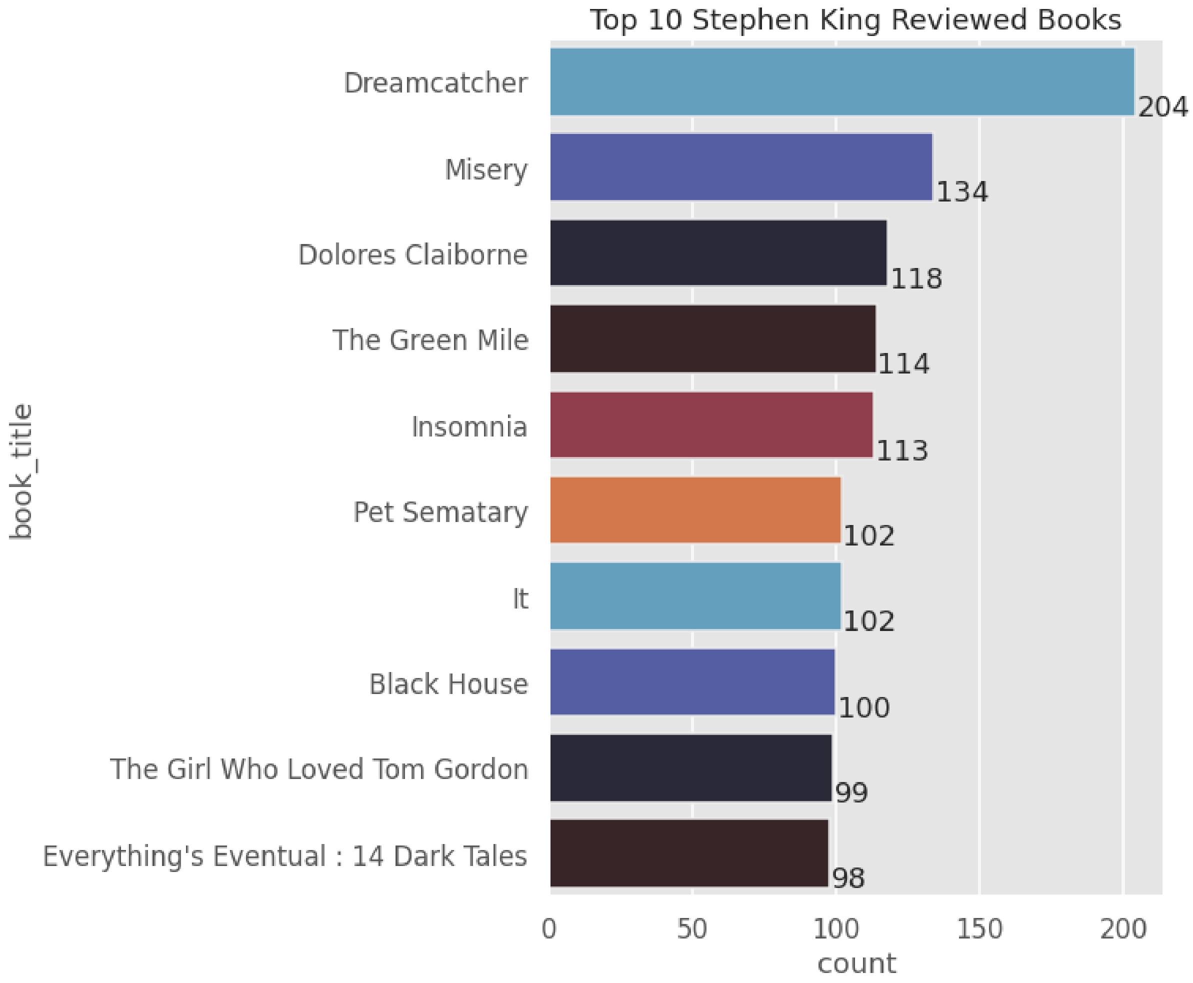


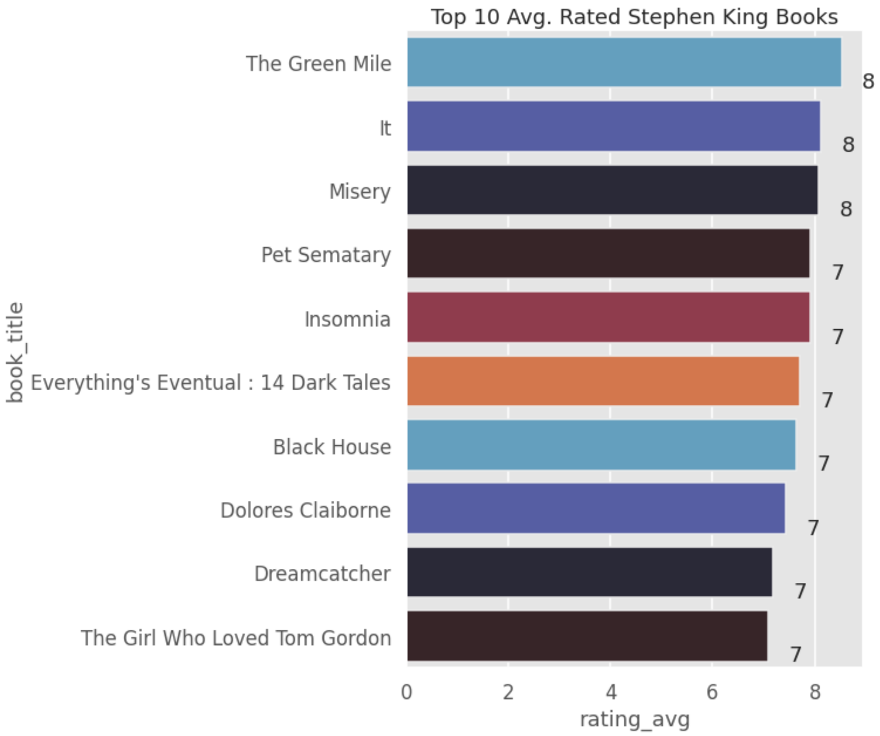


Top 25 Avg. Rated Books
book_title
Harry Potter and the Prisoner of Azkaban (Book 3)
To Kill a Mockingbird
Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
Harry Potter and the Chamber of Secrets (Book 2)
The Secret Life of Bees
The Da Vinci Code
The Lovely Bones: A Novel
The Red Tent (Bestselling Backlist)
Where the Heart Is (Oprah's Book Club (Paperback))
Life of Pi
Angels &amp; Demons
Girl with a Pearl Earring
The Notebook
Divine Secrets of the Ya-Ya Sisterhood: A Novel
Timeline
Snow Falling on Cedars
The Catcher in the Rye
Bridget Jones's Diary
The Testament
The Pilot's Wife : A Novel
The Nanny Diaries: A Novel
A Painted House
The Summons
The Girls' Guide to Hunting and Fishing
Wild Animus
9
8
8
8
8
8
8
8
8
8
8
7
7
7
7
7
7
7
7
7
7
7
7
6
4
0
2
4
6
8
rating_avg



Top 10 Reviewed Authors
book_author
Stephen King
Nora Roberts
John Grisham
James Patterson
J. K. Rowling
Mary Higgins Clark
Janet Evanovich
Dean R. Koontz
Anne Rice
Sue Grafton
4639
2938
2550
2387
1746
1677
1490
1475
1245
1235
0
1000
2000
3000
4000
rating_count


Top 10 Stephen King Reviewed Books
book_title
Dreamcatcher
Misery
Dolores Claiborne
The Green Mile
Insomnia
Pet Sematary
It
Black House
The Girl Who Loved Tom Gordon
Everything's Eventual : 14 Dark Tales
204
134
118
114
113
102
102
100
99
98
0
50
100
150
200
count


Top 10 Avg. Rated Stephen King Books
book_title
The Green Mile
It
Misery
Pet Sematary
Insomnia
Everything's Eventual : 14 Dark Tales
Black House
Dolores Claiborne
Dreamcatcher
The Girl Who Loved Tom Gordon
8
8
8
7
7
7
7
7
7
7
0
2
4
6
8
rating_avg

Bibliography