|  |  |
| --- | --- |
| **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](#_Toc58164496)

[Contents iii](#_Toc58164497)

[List of Figures iv](#_Toc58164498)

[1 Question 1 5](#_Toc58164499)

[1.1 Data Cleaning: Redundant and Inconsistent Data 5](#_Toc58164500)

[1.1.1 Inconsistent Data 5](#_Toc58164501)

[1.1.2 Univariate Analysis 6](#_Toc58164502)

[1.2 Data Cleaning: Missing Values and Outliers 6](#_Toc58164503)

[1.2.1 Outlier Analysis 7](#_Toc58164504)

[1.3 Data Normalization 11](#_Toc58164505)

[1.3.1 Min-Max Scaling 11](#_Toc58164506)

[1.3.2 Z-Score Standardization 11](#_Toc58164507)

[1.3.3 Decimal Scaling 11](#_Toc58164508)

[1.4 Data Transformation 12](#_Toc58164509)

[1.4.1 Natural Log Transform 12](#_Toc58164510)

[1.4.2 Square Root Transform 12](#_Toc58164511)

[1.4.3 Inverse Square Root Transform 12](#_Toc58164512)

[1.5 EDA and Interpretation of Results 13](#_Toc58164513)

[Bibliography 16](#_Toc58164514)

# List of Figures

[Figure 1‑1 Book Ratings, before and after removing 0 ratings 5](#_Toc58164515)

[Figure 1‑2 Book Rating Univariate Analysis 6](#_Toc58164516)

[Figure 1‑3 Age univariate analysis 6](#_Toc58164517)

[Figure 1‑4 Isolation Forest of Original Data 7](#_Toc58164518)

[Figure 1‑5 Dropping book\_rating using IQR 8](#_Toc58164519)

[Figure 1‑6 Dropping Age using IQR 8](#_Toc58164520)

[Figure 1‑7 Isolation forest after dropping outliers with IQR 8](#_Toc58164521)

[Figure 1‑8 Dropping book\_rating with BoxCox 9](#_Toc58164522)

[Figure 1‑9 Dropping age with BoxCox 9](#_Toc58164523)

[Figure 1‑10 Isolation Forest after BoxCox 9](#_Toc58164524)

[Figure 1‑11 Dropping book\_rating with imputation 10](#_Toc58164525)

[Figure 1‑12 Dropping age with imputation 10](#_Toc58164526)

[Figure 1‑13 Isolation Forest after imputation 10](#_Toc58164527)

# 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.

## 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 |

### Inconsistent Data

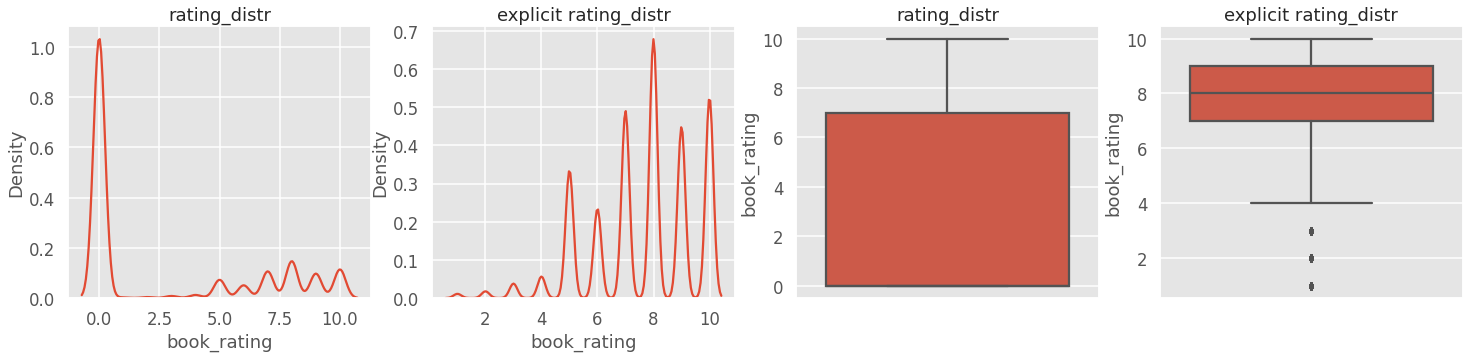


Figure ‑ Book Ratings, before and after removing 0 ratings

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 |

### Univariate Analysis

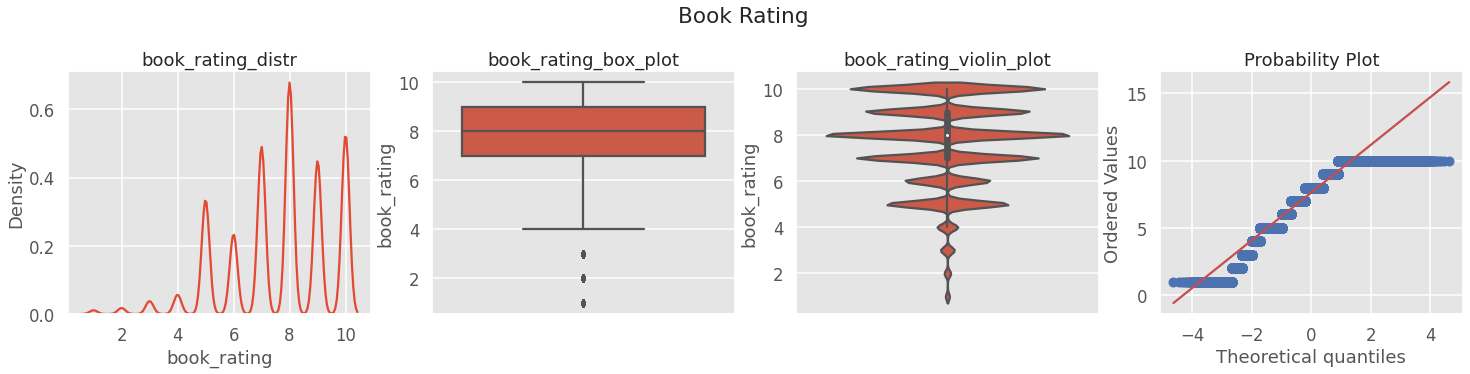


Figure ‑ Book Rating Univariate Analysis

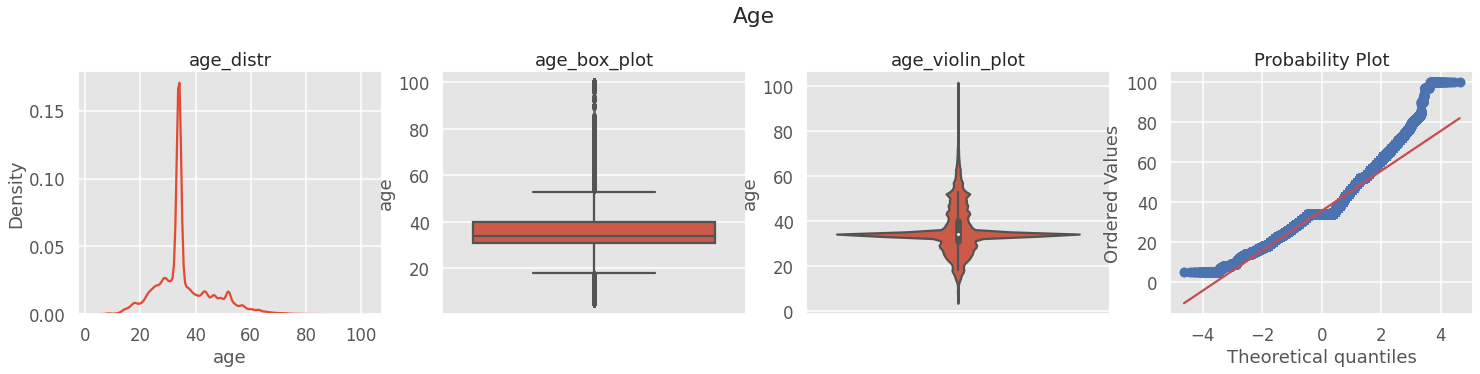


Figure ‑ Age univariate analysis

## Data Cleaning: Missing Values and Outliers

Refer Jupyter Notebook for Cleaning up Missing Values

### Outlier Analysis

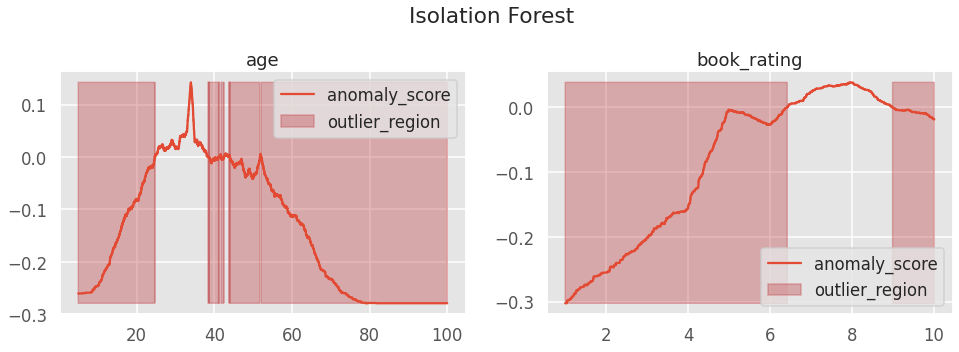


Figure ‑ Isolation Forest of Original Data

#### After Dropping Outliers using IQR

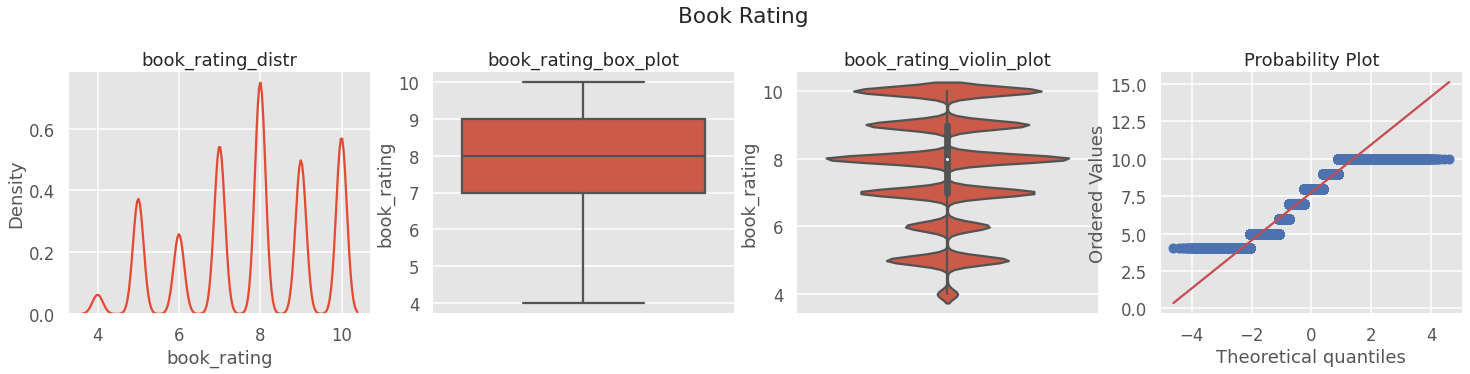


Figure ‑ Dropping book\_rating using IQR

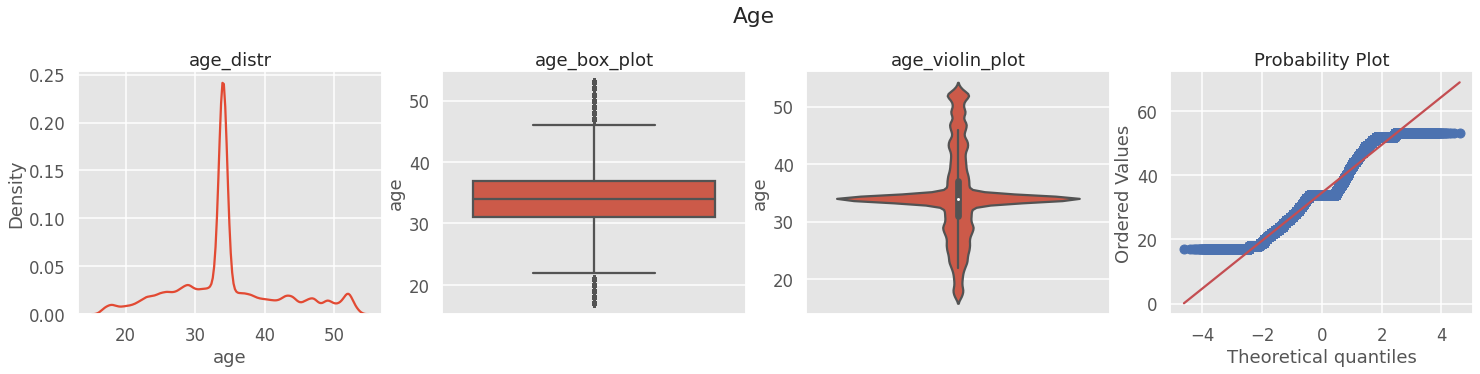


Figure ‑ Dropping Age using IQR

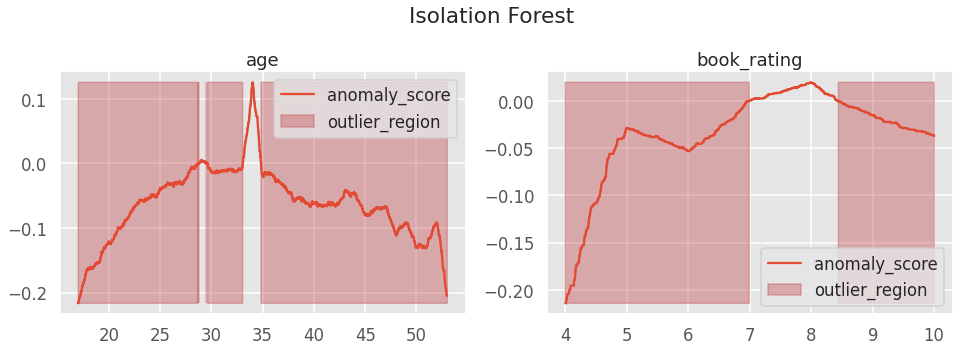


Figure ‑ Isolation forest after dropping outliers with IQR

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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

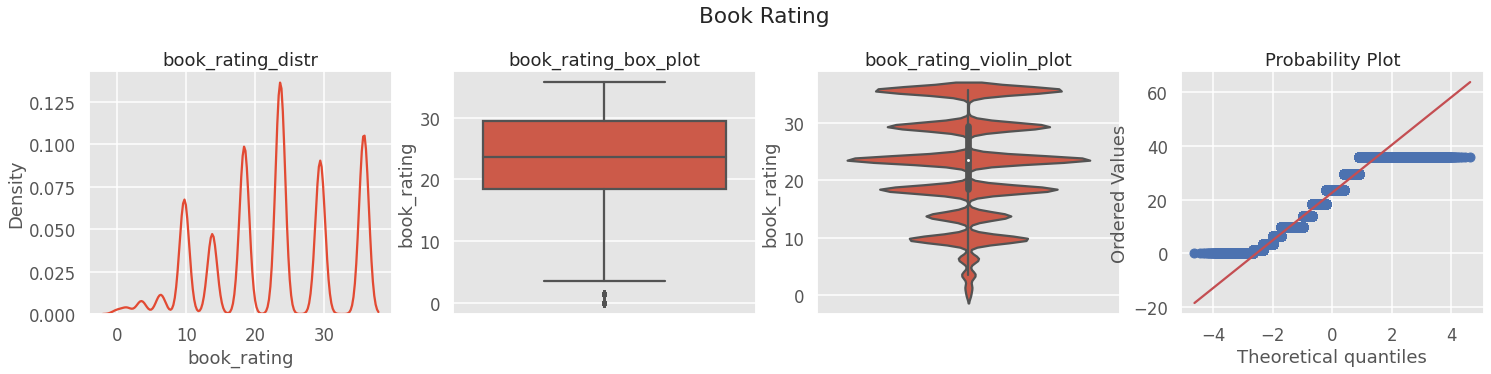


Figure ‑ Dropping book\_rating with BoxCox

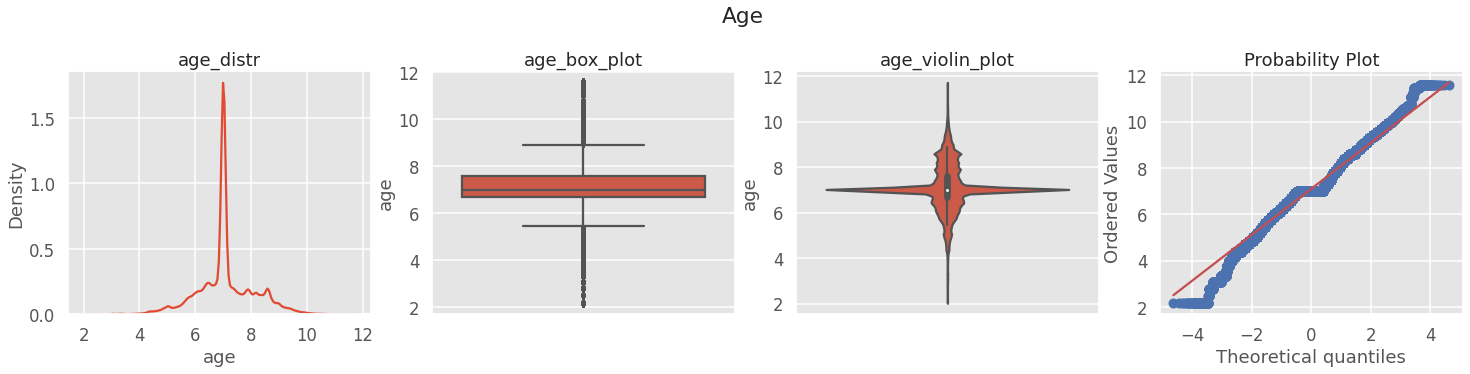


Figure ‑ Dropping age with BoxCox

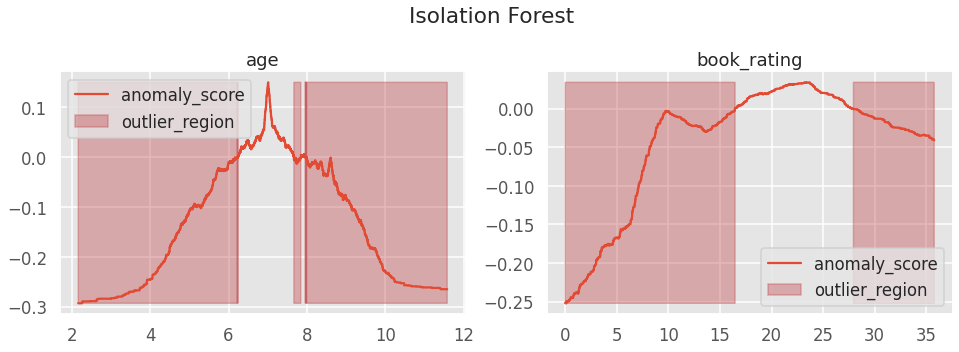


Figure ‑ Isolation Forest after BoxCox

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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

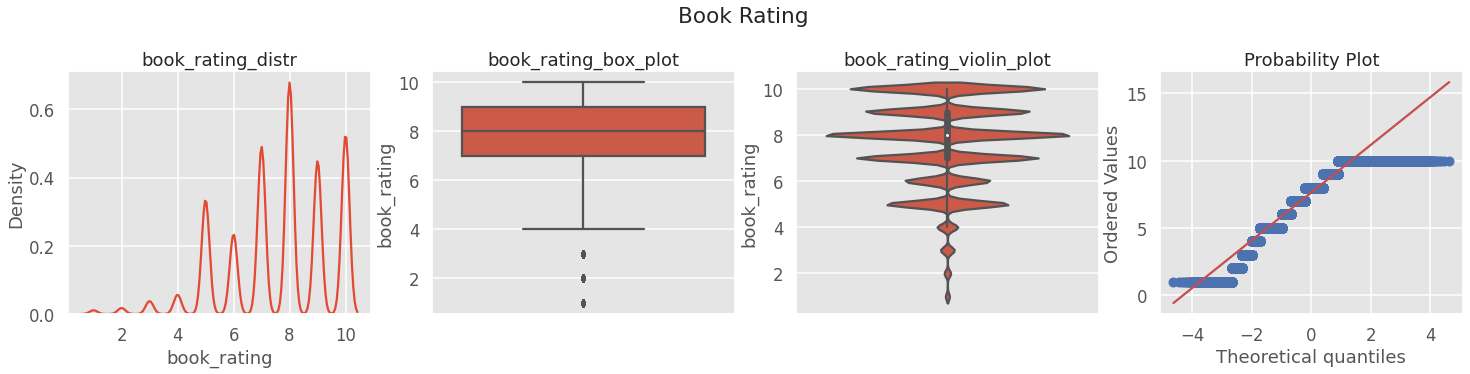


Figure ‑ Dropping book\_rating with imputation

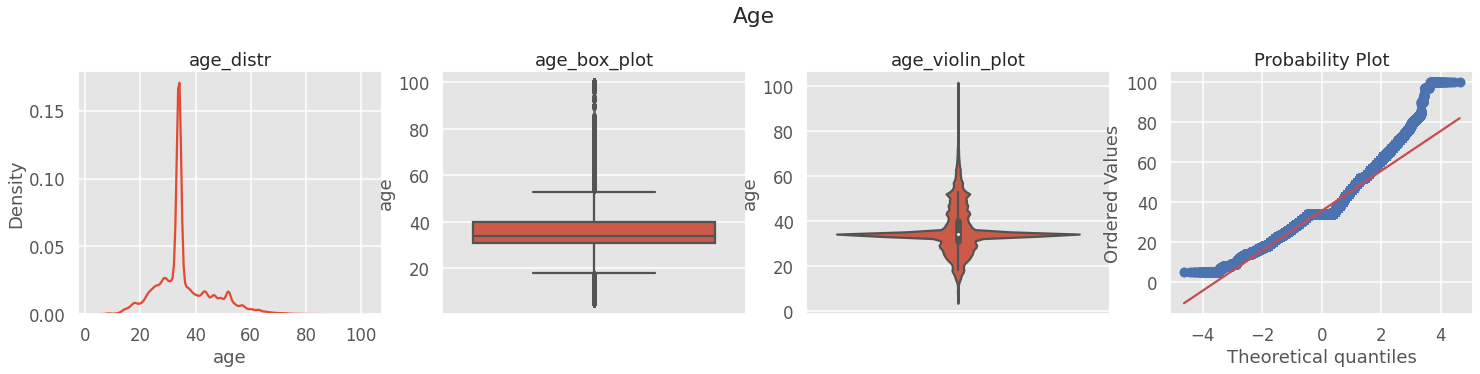


Figure ‑ Dropping age with imputation

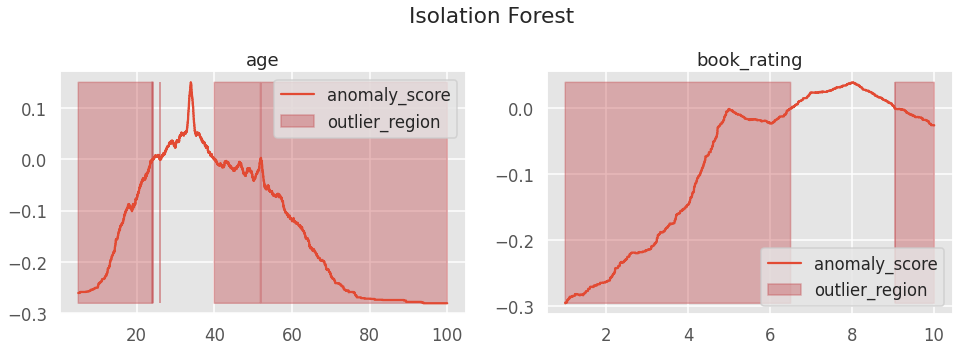
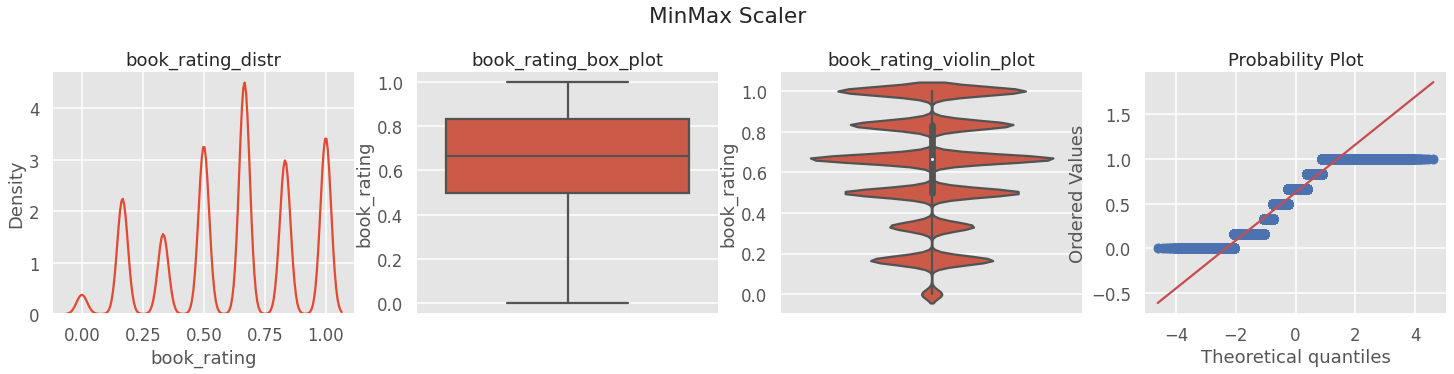


Figure ‑ Isolation Forest after imputation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

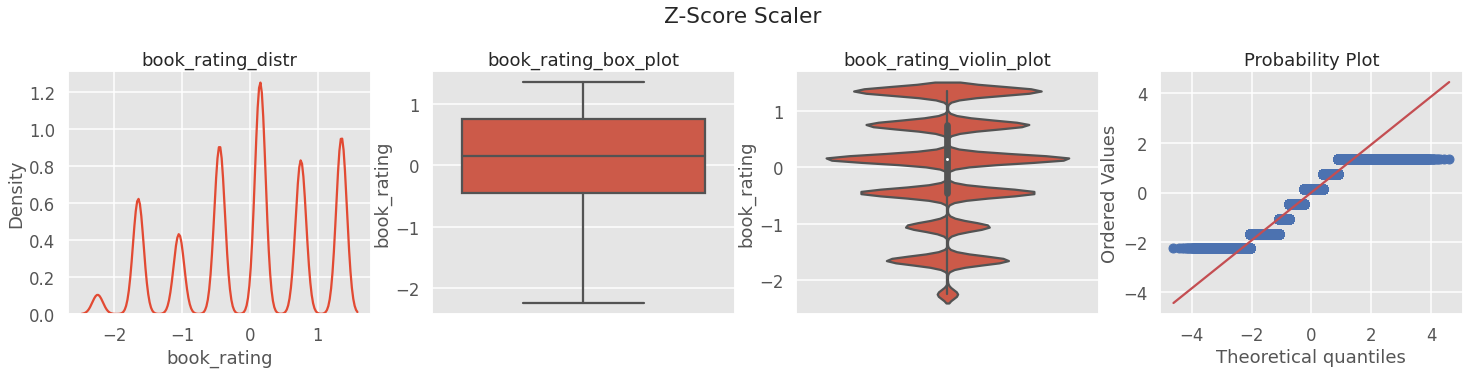
## Data Normalization

### Min-Max Scaling



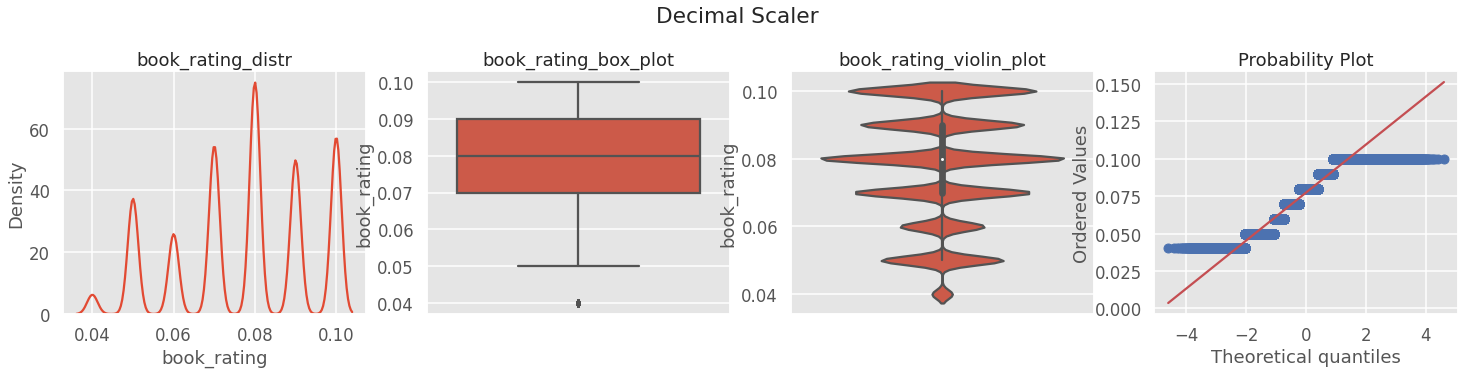
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Std | Min | Max | Skewness | Kurtosis |
| 0.62 | 0.27 | 0 | 1 | -0.34 | -0.80 |

### Z-Score Standardization



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Std | Min | Max | Skewness | Kurtosis |
| ~0 | ~1 | -2.24 | 1.35 | -0.34 | -0.80 |

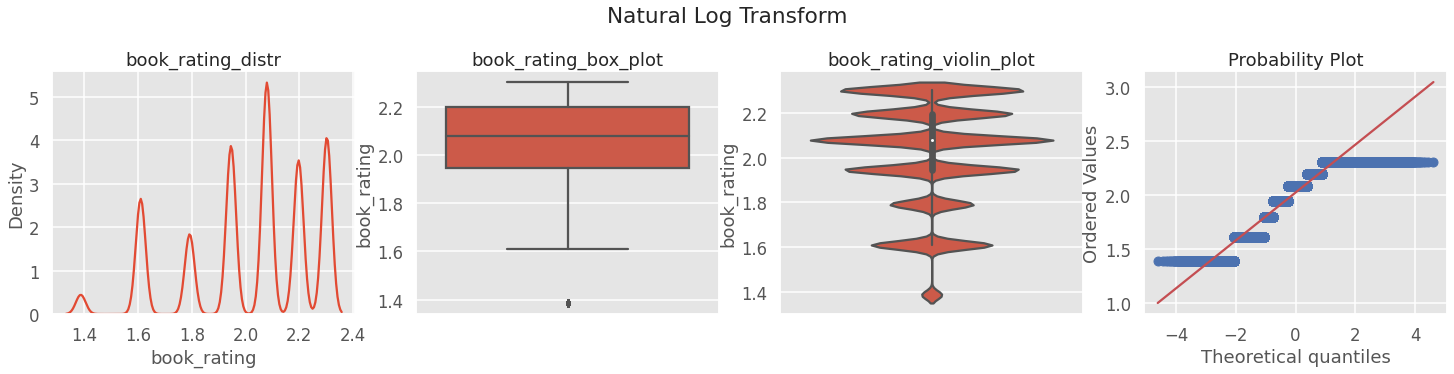
### Decimal Scaling



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Std | Min | Max | Skewness | Kurtosis |
| 0.077 | 0.01 | 0.04 | 0.1 | -0.34 | -0.80 |

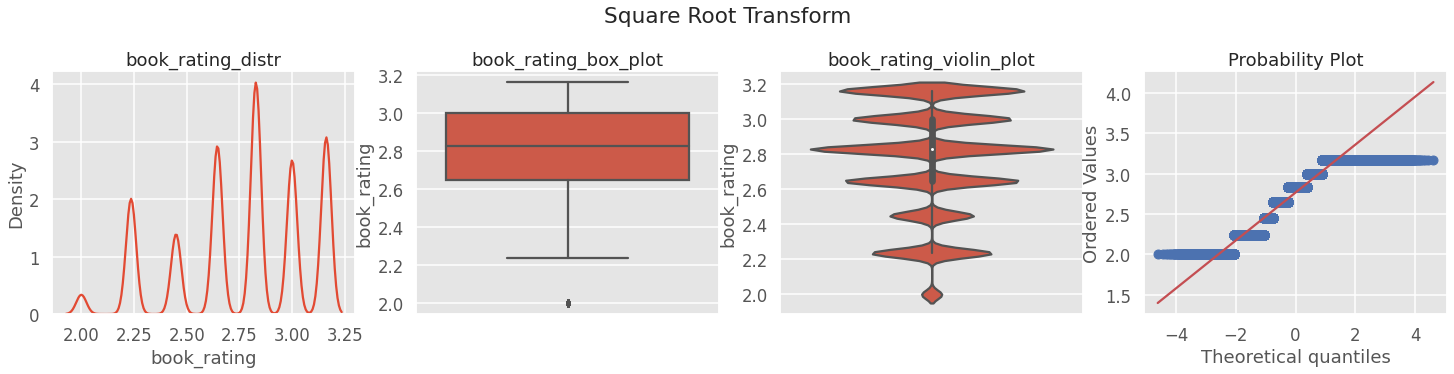
## Data Transformation

### Natural Log Transform



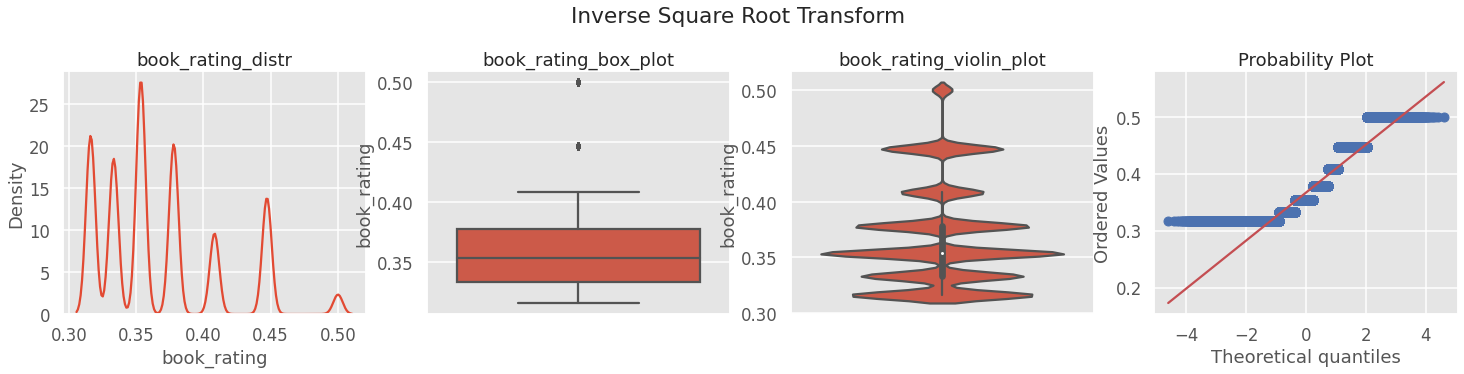
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Std | Min | Max | Skewness | Kurtosis |
| 2.02 | 0.23 | 1.38 | 2.3 | -0.74 | -0.23 |

### Square Root Transform



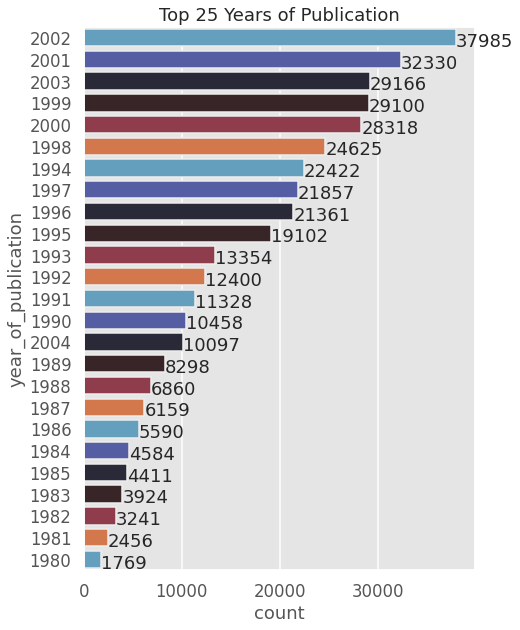
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Std | Min | Max | Skewness | Kurtosis |
| 2.76 | 0.31 | 2 | 3.16 | -0.54 | -0.57 |

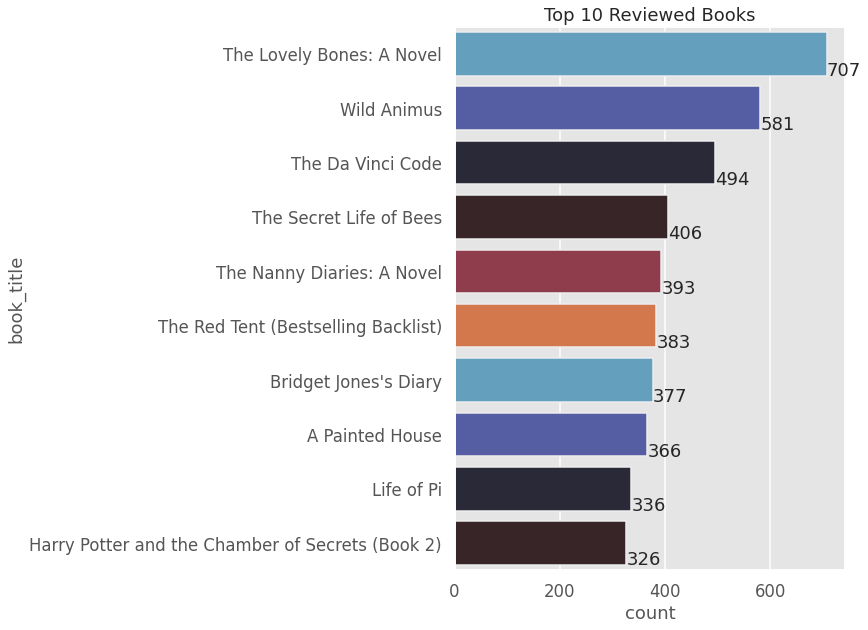
### Inverse Square Root Transform

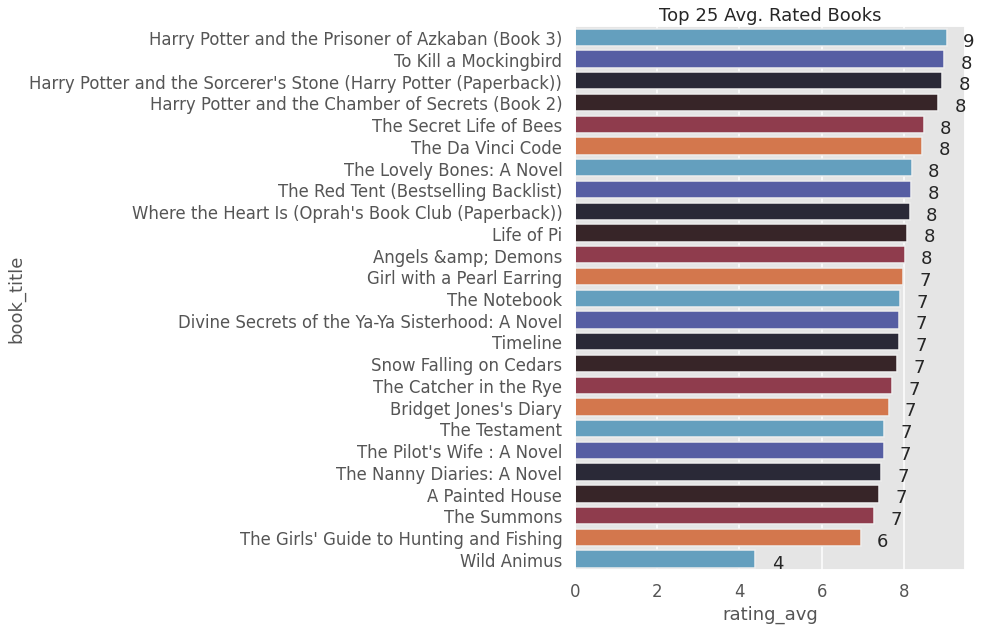


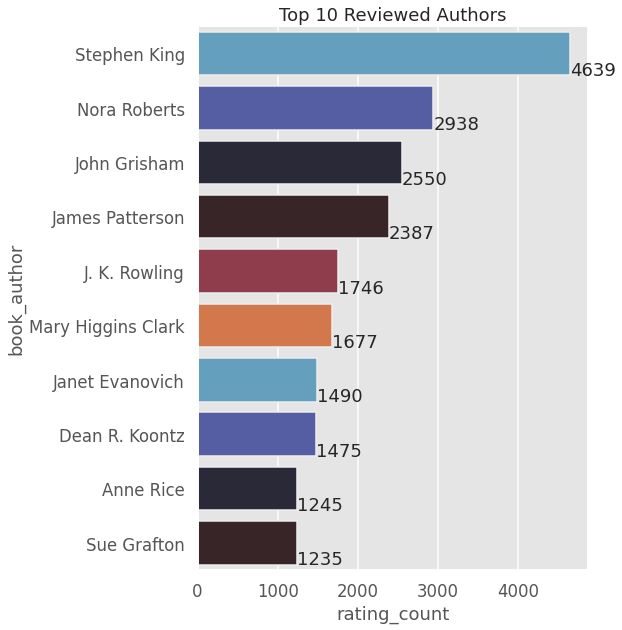
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean | Std | Min | Max | Skewness | Kurtosis |
| 0.366 | 0.044 | 0.31 | 0.5 | 0.96 | 0.24 |

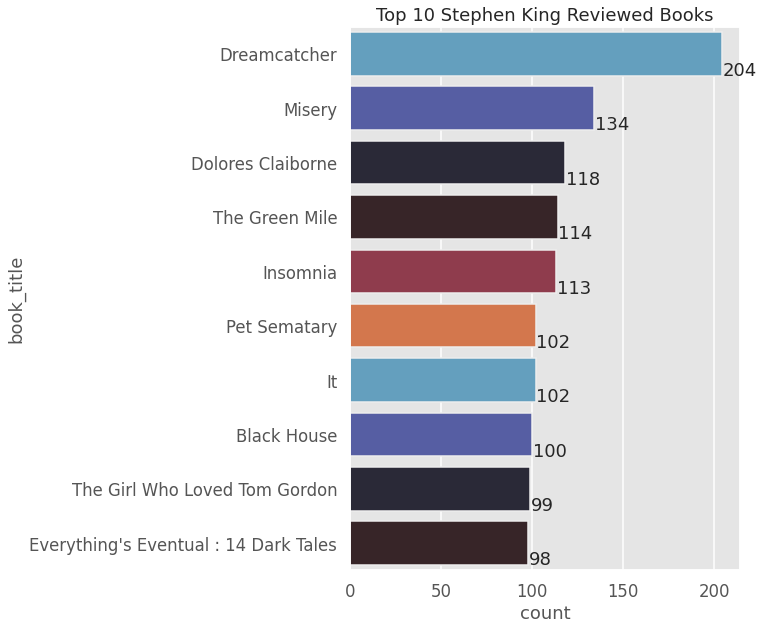
## EDA and Interpretation of Results

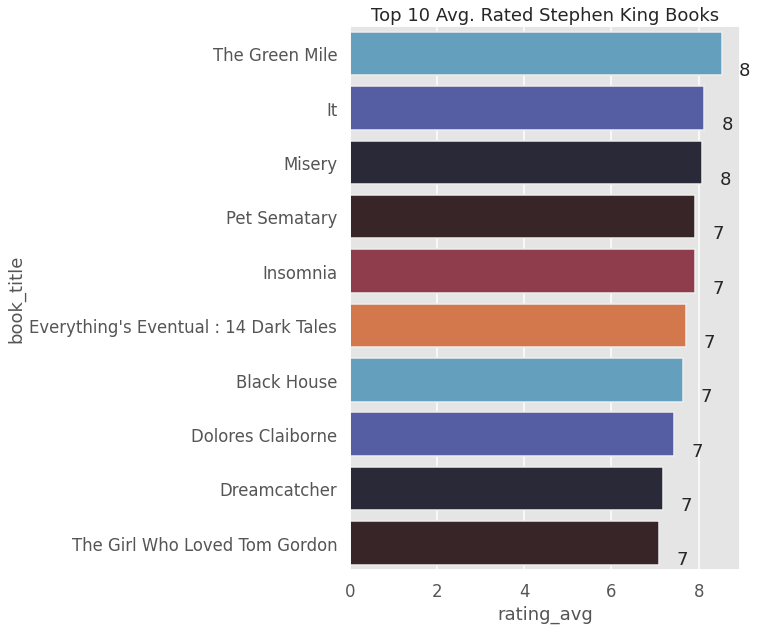












# Bibliography