Assignment 2

List the attributes in the set S

S = {id, title, author, rating}

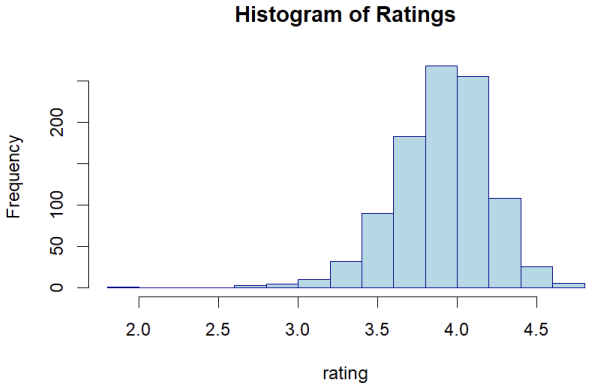
For each attribute X in S, consider only Table A and discuss the following in the report:

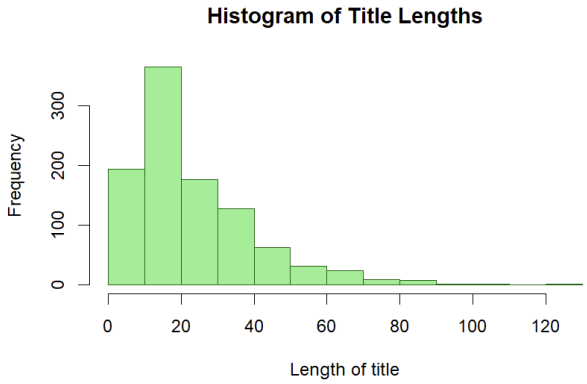
In tableA, there is 15/1000 (0.015%) missing values in rating.

One possible solution to handle missing values is by deletion. Deleting rows with missing values is simple but can lead to data loss. Another solution is to imputation replaces missing values with estimated values, such as using the mean or the median. There is more advanced method such as KNN (K-Nearest Neighbor) imputation to capture complex relationships and provide more accurate results.

The *id* attribute is numeric, as it represents a unique identifier.
The *title* attribute is textual, as it represents the title of a book.
The *author* attribute is textual, as it represents the name of an author.
The *rating* attribute is numeric, as it represents a numerical rating from 1 to 5.

The average length of the *title* attribute is 23.48 characters. The minimum length of the title is 2 characters, and the maximum length of the title is 123 characters.
The average length of the *author* attributes is 13.799 characters. The minimum length of the author attribute is 8 characters, and the maximum length of the author attribute is 47 characters.





On the first diagram, shows the histogram of ratings ranging from 1.5 to 5. We can see that there is an outlier lying at 1.5-2.0 bar. Looking at the data, there is only one book that is rated 1.99, hence, the outlier.
On the second diagram, shows the histogram of book title lengths ranging from 0 to 140. We can see that there is an outlier laying at the 120-140 bar. Looking at the data, it is the book titled

“The Magic Castle: A Mother's Harrowing True Story Of Her Adoptive Son's Multiple Personalities-- And The Triumph Of Healing” of 123 letters. The second place is “The complete â€œA Glimpse into Hellâ€ series - 6 books, 215 chapters, 1800 pages, 650K words of pure gore” of 107 letters.

None of the attributes are supposed to follow a certain format. Each of the book titles, id, author’s name, and rating do not follow a certain format.

There are no synonyms among my attributes. Each of the book title, id, author’s name, and rating is unique.

There are some book titles that might not have been captured successfully by web scraper. For example, the book titled, “The complete â€œA Glimpse into Hellâ€ series - 6 books, 215 chapters, 1800 pages, 650K words of pure gore”, there is garbled code in the data. Additionally, some book title is mixed with additional descriptions, such as “…6 books, 215 chapers…” which might affect data analyzing in the future.

RStudio is used to analyze the data above. Code will be provided in GitHub.