Goodreads Books Dataset 1-pager

**Description:**

The [Goodreads-books dataset](https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks) is a comprehensive collection of book-related information scraped via the Goodreads API. This dataset caters to avid readers by providing a clean and enriched dataset, offering essential features such as ratings, reviews, authors, publication details, and more. It was created with the aim of providing a well-curated dataset for book enthusiasts to explore and gain insights, which has been very useful in the course of this miniproject.

**Key features:**

* Size: the dataset contains a large number of entries, with over 11,000 books, each having detailed information.
* Authors: multiple authors are delimited by ‘/’, allowing for a clear representation of collaborative works.
* Ratings and reviews: the dataset includes solid numerical values for the number for the number of total ratings and reviews the book has, as well as its average rating between 0-5.
* Publication Details: information on publishers and publication dates is also available.

**Dataset maintenance:**

The dataset was initiated in May 2019. As of December 2020, Goodreads ceased issuing new developer keys for their public developer API, and has since been retired.

**Operations Used:**

1. Filter Non-English Books:
   1. Remove entries with books that are not in English (language\_code = “eng”, “eng-US”, “en-GB”)
   2. Save new dataset to a new csv file called EnglishBooks.csv
   3. This new dataset has 10530 observations and will be used for the rest of the project.
2. Select High-Rated Books:
   1. Choose books with a 4-star average rating or higher.
   2. Save new dataset to a new csv file called HighRatedBooks.csv
   3. This new dataset has 4664 observations.
3. Randomly Select 100 Rows:
   1. Randomly sample 100 rows from the dataset.
   2. Save new dataset to a new csv file called “RandomBooks.csv
   3. This new dataset has 100 observations.
4. Select Books by Authors Starting with J:
   1. Filter books with authors whose names start with J.
   2. Save new dataset to a new csv file called AuthorsStartingWithJ.csv
   3. This new dataset has 1432 observations.
5. Identify Highest Rated and Highest Page Book:
   1. Find a single book with the highest average rating and the highest number of pages.
   2. Output: Colossians and Philemon: A Critical and Exegetical Commentary (International Critical Commentary) by R. McL. Wilson
6. Identify Lowest Rated and Lowest Page Book:
   1. Find a single book with the lowest average rating and the lowest number of pages.
   2. Output: American Film Guide by Frank N. Magill
7. Select Books Published in the 2000s Starting with D:
   1. Choose books published between 2000 and 2009 with titles starting with D.
   2. Save new dataset to a new csv file called BooksPublished2000sStartingWithD.csv
   3. This new dataset has 266 observations.
8. Compare Average Ratings of Books A-M vs. N-Z:
   1. Create a boxplot to compare the average ratings of books with titles starting from A-M and N-Z.
   2. Output:A graph of a number of rating

      Description automatically generated with medium confidence
9. Compare Ratings of Books by Jane Austen:
   1. Create a bar graph to compare the average rating of books written by Jane Austen.
   2. Output:A graph of different colored squares

      Description automatically generated
10. Scatter Plot of Number of Pages vs. Average Ratings:
    1. Create a scatter plot to visualize the relationship between the number of pages and average ratings.
    2. Output:A graph with numbers and dots

       Description automatically generated