**Dataset Description:**

Amazon is an American multinational technology company that focuses primarily on e-commerce, online advertising, cloud computing, digital streaming, and artificial intelligence. Moreover, it has been often referred to as "one of the most influential economic and cultural forces in the world. As a result, the Amazon dataset is chosen for this research in order to establish a benchmark for the recommendation system for any company with historical data.

The model uses Collaborative filtering to evaluate the books' reviews on Amazon. The review data contain the feedback of about 3 million users on 212404 unique books and it contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. The book detail dataset contains information about 212404 unique books, the file is built by using google book API to get details information about books it rated in the first file.

The detailed attributes of the book detail dataset have been provided below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Original Attribute** | **Attribute Name** | **Format** | **Description** |
| ID | Id | object | The ID of the book. |
| Title | Title | object | Book title. |
| Price | Price | float64 | The price of the book. |
| User ID | User\_id | object | The ID of the user who rates the book. |
| Profile Name | profileName | object | Name of the user who rates the book. |
| Reviews | review/helpfulness | object | Helpfulness rating of the review, |
| Review Scores | review/score | float64 | Rating from 0 to 5 for the book. |
| Review Time | review/time | int64 | Time of giving the review. |
| Review Summary | review/summary | object | The summary of a text review. |
| Review Text | review/text | object | The full text of a review. |

Only three features ‘price’, ‘review/score’, and ‘review/time’ contain numerical values whereas rests contain categorical values.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Price** | **review/score** | **review/time** |
| Count | 481171.000000 | 3.000000e+06 | 3.000000e+06 |
| Mean | 21.762656 | 4.215289e+00 | 1.132307e+09 |
| Standard Deviation | 26.206541 | 1.203054e+00 | 1.493202e+08 |
| Min | 1.000000 | 1.000000e+00 | -1.000000e+00 |
| Max | 995.000000 | 5.000000e+00 | 1.362355e+09 |

The detailed attributes of the review dataset have been provided below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Original Attribute** | **Attribute Name** | **Format** | **Non-null Count** | **Description** |
| Title | Title | object | 212403 non-null | Book Title |
| Description | description | object | 143962 non-null | Description of book |
| Authors | authors | object | 180991 non-null | Name of book authors |
| Image | image | object | 160329 non-null | URL for the book cover |
| Preview Link | previewLink | object | 188568 non-null | Link to access this book on google books |
| Publisher | publisher | object | 136518 non-null | Name of the publisher |
| Published Date | publishedDate | object | 187099 non-null | The date of publish |
| Info Link | infoLink | object | 188568 non-null | Link to get more information about the book on google books |
| Categories | categories | object | 171205 non-null | Genres of books |
| Ratings Count | ratingsCount | float64 | 49752 non-null | The average rating for the book |

Only ‘ratingsCount’ contains numerical values whereas rests contain categorical values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Count** | **Mean** | **Standard Deviation** | **Min** | **Max** |
| ratingsCount | 49752.00 | 21.253 | 201.340431 | 1.00 | 4895.00 |