Book Recommender System

DATA TRANSFORMATION SUMMARY

Contents

1. INTRODUCTION

1.1 Problem statement

The aim is to suggest random books in the first place and ask the user to select one of them. Based on the input given by the user, recommender system recommend a new collection of books based on the user's selection.

1.2 Data

In this project , we mainly use the “rating” and “user_id” data from the datasets ratings.csv and books.csv respectively.

	Id	Book_id	Best_book_id	work_id	.....	Small_image_url
0	1	2767052	2767052	2792775	.....	https://images.gr-assets.com/books/1447303603s...
1	2	3	3	4640799	.....	https://images.gr-assets.com/books/1474154022s...
2	3	41865	41865	3212258	.....	https://images.gr-assets.com/books/1361039443s...
3	4	2657	2657	3275794	.....	https://images.gr-assets.com/books/1361975680s..
. . .	. . .	. . .	. . .	. . .	.....	. . .

Table 1.1: books.csv(Columns: 0-3)

	Book_id	User_id	rating
0	1	314	5
1	1	439	3
2	1	588	5

Table 1.2: ratings.csv(Columns: 0-2)

In books.csv there are 10000 rows × 23 columns and in ratings.csv there are

981756 rows × 3 columns.

2. Methodology

2.1 Pre Processing

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

2.1.2 Missing Value Analysis

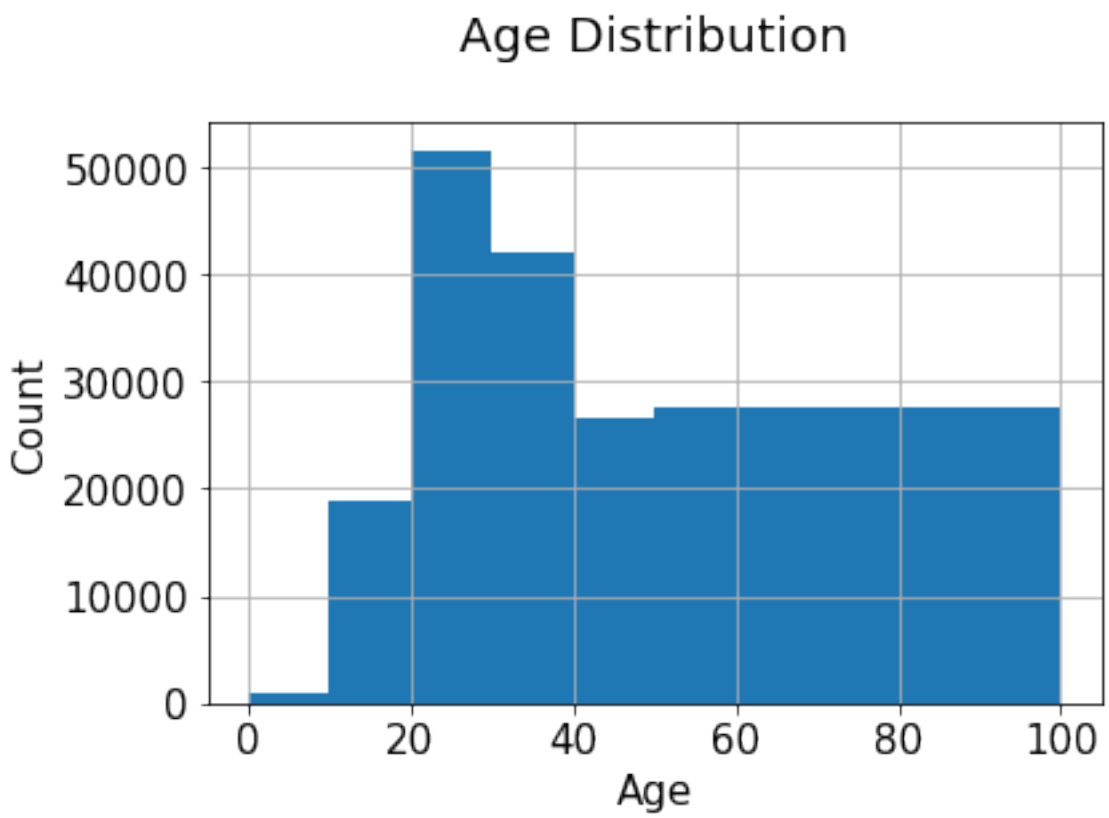
Missing value analysis is done to check is there any missing value present in given dataset. Missing values can be easily treated using various methods like mean, median method, knn method to impute missing value.

id	0
book_id	0
best_book_id	0
work_id	0
books_count	0
isbn	700
isbn13	585
authors	0
original_publication_year	21
original_title	585
title	0
language_code	1084
average_rating	0
ratings_count	0
work_ratings_count	0
work_text_reviews_count	0
ratings_1	0
ratings_2	0
ratings_3	0
ratings_4	0
ratings_5	0
image_url	0
small_image_url	0

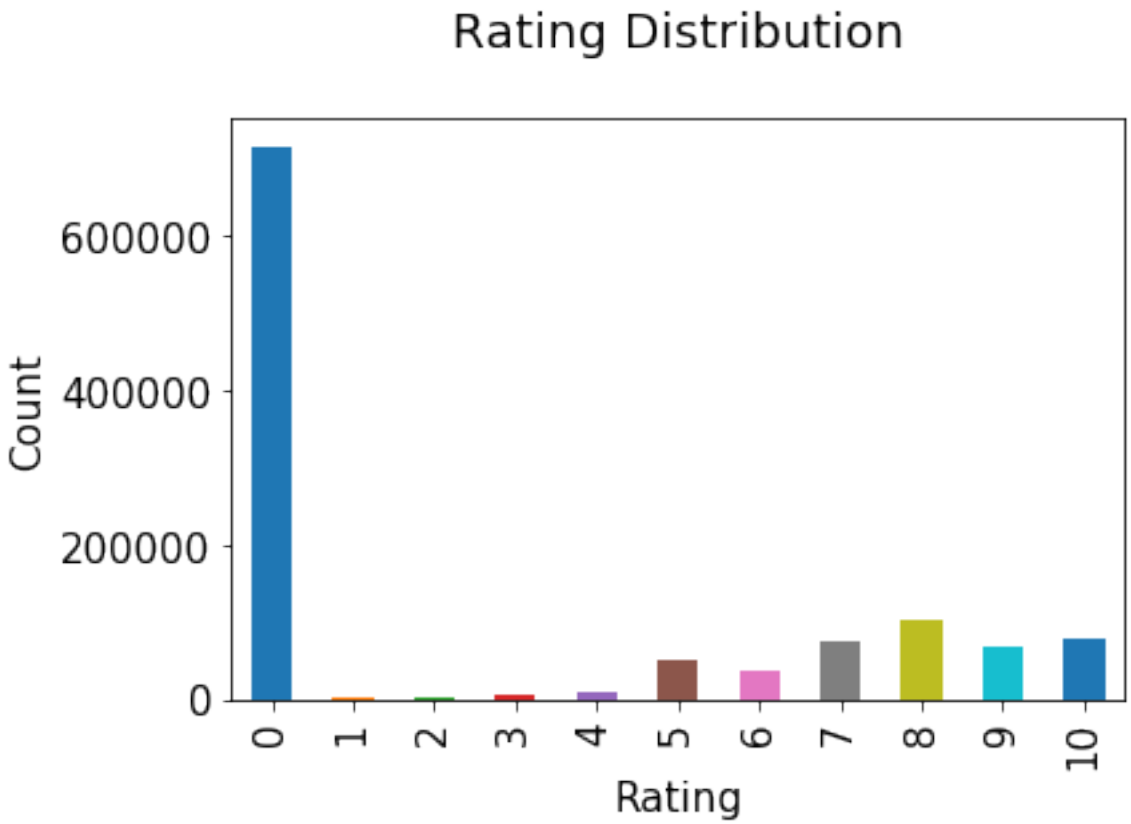
book_id	0
user_id	0
rating	0

There is no missing values in id and title column, even though remaining some columns have missing values we can neglect them as those columns are not used.

2.1.3 Age Distribution



2.1.4 Rating Distribution



Here, as in both the Age and Rating distribution the data is biased towards certain points .So, to ensure statistical significance users with less than 200 ratings and books with less than 100 ratings are excluded.