* Go to the Link : <https://github.com/zygmuntz/goodbooks-10k>
* Download ‘books.csv’ and ‘ratings.csv’. The downloaded data are kept in the   
  “2. Preprocessing (MySQL)/input” folder.
* Run MySQL command window and run the following commands:

1. Create a database:  
   create database book\_dataset;
2. Create a ‘rating’ table and load data from ‘ratings.csv’:  
     
   CREATE table book\_dataset.rating(user\_id INTEGER, book\_id INTEGER, rating INTEGER);  
     
   LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/book\_dataset/ratings.csv   
   INTO TABLE book\_dataset.rating   
   FIELDS TERMINATED BY ',' IGNORE 1 ROWS;
3. Create a ‘book’ table and load only 5 columns (book\_id, authors, original\_publication\_year, original\_title, & language code) from ‘books.csv’.  
     
   CREATE table book\_dataset.book(book\_id INTEGER, authors TEXT, original\_publication\_year INTEGER, original\_title TEXT, language\_code TEXT);  
     
   LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/book\_dataset/books.txt'   
   INTO TABLE book\_dataset.book   
   FIELDS TERMINATED BY ',' ENCLOSED BY '"' ESCAPED BY '"' IGNORE 1 ROWS  
   (@col1, @dummy, @dummy, @dummy, @dummy,  
   @dummy, @dummy, @col2, @col3, @col4,  
   @dummy, @col5, @dummy, @dummy, @dummy,  
   @dummy, @dummy, @dummy, @dummy, @dummy,  
   @dummy, @dummy, @dummy)  
   Set book\_id = @col1, authors = @col2, original\_publication\_year = IF(@col3 = '', NULL, @col3), original\_title = @col4, language\_code =@col5;
4. Create a ‘temp\_table’ and put 202 distinct users with highest number of ratings from the ‘rating’ table order by their number of ratings:  
     
   Create temporary table If Not Exists book\_dataset.temp\_table  
    select user\_id, count(distinct book\_id) as count\_book, '' as attr from book\_dataset.rating group by user\_id order by count(distinct book\_id) desc limit 202;
5. From the 202 users from ‘temp\_table’, create a ‘user’ table and put first two users as the coupe (attribute C), next 100 users as the other married persons (attribute M), and next 100 users as the invited friends (attribute F).  
     
   Create Table book\_dataset.user(user\_id INTEGER, book\_rated INTEGER, attr VARCHAR(1));

Insert Into book\_dataset.user  
 select user\_id, count\_book, 'C' from book\_dataset.temp\_table Limit 0,2;  
Insert Into book\_dataset.user  
 select user\_id, count\_book, 'M' from book\_dataset.temp\_table Limit 2,100;  
Insert Into book\_dataset.user  
 select user\_id, count\_book, 'F' from book\_dataset.temp\_table Limit 102,100;

drop table if exists book\_dataset.temp\_table;

1. Create a ‘rating\_filtered’ table to keep only the ratings by those 202 users in the ‘user’ table:   
     
   CREATE table book\_dataset.rating\_filtered(user\_id INTEGER,book\_id INTEGER,rating INTEGER);  
     
   Insert Into book\_dataset.rating\_filtered  
    Select r.user\_id, r.book\_id, r.rating from book\_dataset.rating r Inner Join book\_dataset.user u On r.user\_id = u.user\_id;
2. Create a ‘book\_filtered’ table to keep only the books rated by those 202 users:  
     
   CREATE table book\_dataset.book\_filtered (book\_id INTEGER, authors TEXT, original\_publication\_year INTEGER, original\_title TEXT, language\_code TEXT);  
     
   Insert Into book\_dataset.book\_filtered  
       Select b.\* From book\_dataset.book b Join book\_dataset.rating\_filtered r On b.book\_id = r.book\_id group by r.book\_id;
3. Export the Processed data:  
     
   SELECT \* FROM book\_dataset.user INTO OUTFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/book\_dataset/user.csv';  
   SELECT \* FROM book\_dataset.rating\_filtered INTO OUTFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/book\_dataset/rating.csv';  
   SELECT \* FROM book\_dataset.book\_filtered INTO OUTFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/book\_dataset/book.csv';

* The generated data are kept in the “2. Preprocessing (MySQL)/output” folder.