- 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;
```

6. 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;
```

7. 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;
```

8. 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.