# ES114 Probability,Statistics and Data Visualization Report: Data Narrative 1

Keshav Sobania, *B TECH 22,Civil Engineering, ROLL NO. 22110118*

## A. Overview of The Dataset:-

The 5 Datasets and their content are as following:

- **books.csv**-Various columns representing various info such as author,average rating, year of publication,etc have been represented in this Dataset.
- **ratings.csv-** book_id,user_id and the rating have been given in the following dataset.
- **books_tags.csv-**goodreads_book_id,tag_id and count have been given in the following dataset .
- **tags.csv-** tag_id and tag_name have been given in the following dataset.
- **to_read.csv-**This Dataset contains the information regarding user_id and book_id.

## B. Scientific Questions/Hypothesis:-

- If I pick a book with language code eng what is the probability that it's average rating is 4 or greater than 4?
- If I pick a book written by Suzanne Collins what is the probability that its average rating is greater than or equal to 4?
- What is the ratio of books with an average rating of 3-3.9 to the no. of books with an average rating of 4-5?
- What is the ratio of books having language code spa to the books having language code eng?
- Which book has the highest no. of ratings count?

## C) Details of libraries and functions:-

### a)libraries:-

- **Numpy-**Python Library used for working with arrays containing various functions for working out on matrices,linear algebra and many such topics.
- **Pandas-**Very helpful in importing and analyzing data.
- **Matplotlib**-Very helpful in visualizing data by plotting it in various ways such as histograms,bar graphs,etc

### b)functions:-

- Loops, indexes and if else functions have been used by me for deducing the answers.

## D) Answers to the Questions:-

**1.**

Let ,

P(B/A) = probability that a book chosen to be of language code eng has an average rating of 4 or greater than 4

where ,
A = No. of books with language code eng
B = No. of books with average rating of 4 or above.
P(B/A) = P(B and A) / P(A)
By code , we found that
P(B and A) = 3439/10000
P(A) = 6341/10000
P(B/A) = 0.5423434789465384

**2.**
Let ,
P(B/A) = Picking a book by Suzanne CollinsThen the probability that it has a rating of 4 or above.
where ,
A = No. of books written by Suzanne Collins
B = No. of books with average rating of 4 or above.
P(B/A) = P(B and A) / P(A)
By code , we found that
P(B and A) = 8/10000
P(A) = 9/10000
P(B/A) = 0.8888888888888888

3**.**
A = no. of books with average rating ranging
From 3 to 3.9
B= no. of books with average rating ranging
From 4 to 5
A=3218
B=5334
A/B=0.6032995875515561

**4.**
A = no. of books with language_code “spa”
B= no. of books with language_code “eng”
A=20
B=6341
A/B=0.0031540766440624505

**5.**
The book with the maximum number of ratings count is “The Hunger Games”.

## E) Summary to the Observations:-

- I came to the conclusion that pandas can be used to read and interpret various datasets and important information can be used from it.

## F) References:-

- https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/books.csv
- https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/tags.csv
- https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/to_read.csv
- https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/ratings.csv
- https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/book_tags.csv