DATA NARRATIVE

Kishan Ved
Computer Science Engineering
Indian Institute of Technology Gandhinagar
Roll Number - 22110122

***Abstract*—This data narrative aims to provide a comprehensive insight into the Goodbooks dataset collection and analyze the data it contains.**

I. OVERVIEW OF THE DATASET

The Goodbooks dataset on GitHub (attached in the question) contains six million ratings for ten thousand most popular (with the most ratings) books. It has five datasets. Out of these, the datasets emphasized in this data narrative include – books.csv, to_read.csv and ratings.csv. The following is a snippet of the dataset books.csv:

```
books.head()
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer
3	4	2657	2657	3275794	487	61120081	9.780061e+12	Harper Lee
4	5	4671	4671	245494	1356	743273567	9.780743e+12	F. Scott Fitzgerald

5 rows × 23 columns

II. SCIENTIFIC QUESTIONS AND HYPOTHESIS

The dataset has been analyzed and researched upon to answer five scientific questions and validate a hypothesis. An interesting observation made has also been mentioned.

A. *Scientific Questions:*

1) In what languages are the books in the dataset books.csv written? Which language has the most books? What is the relative number of books based on language?
2) Which authors have the maximum number of books in the dataset boks.csv?
3) What is the average rating of books in the dataset? Which book in the dataset books.csv has the highest rating? Who authored it?
4) Who are the most active users? (based on the number of ratings)
5) Which books are more likely to be read in the near future? (based on the number of to-read tags)

B. *Hypothesis:*

"A book having more number of ratings has a greater number of reviews."

C. *Interesting observation:*

The average rating of books published in the year narrows down towards 4 as the year of publication increases from 1500 to 2023.

III. PYTHON LIBRARIES AND FUNCTIONS USED

The following Python libraries and their mentioned functions were used to analyze the dataset:

- NumPy:
  - np.array() – To generate a NumPy array to operate upon.
  - np.arange() –To get a NumPy array having consecutive values.
- Pandas:
  - pd.read_csv() – To read data from a csv file.
  - pd.DataFrame() – To convert data into a Pandas DataFrame.
  - describe() – To get statistical information about the data.
- Matplotlib:
  - plt.plot() – To plot data.
  - plt.pie() – To display a pie chart of the data.
  - plt.barh() – To plot a horizontal bar graph.
- SciPy:
  - linalg.lstsq() – To compute the least square solution of datapoints.

IV. ANSWERS TO THE QUESTIONS

Q1) *In what languages are the books in the dataset books.csv written? Which language has the most books? What is the relative number of books based on language?*

A) Approach:

To find which language books are present, we can count the values per entry in the language_code column of the books.csv dataset. Sorting these values in descending order gives the most popular language(s) at the beginning.

```
l = books["language_code"].dropna()
l = pd.DataFrame(l.value_counts())
```

Language code and corresponding number of books:

	Number of books
eng	6341
en-US	2070
en-GB	257
ara	64
en-CA	58
fre	25
ind	21
spa	20
ger	13
per	7
jpn	7
por	6
pol	6
en	4
nor	3
dan	3
fil	2
ita	2
vie	1
tur	1
nl	1
swe	1
rum	1
mul	1
rus	1

A pie chart has been plotted to understand the relative number of books based on language.

```
plt.pie(val,labels=names,colors=['pink','yellow','red','green','blue'],autopct='%1.1f%%',)
```

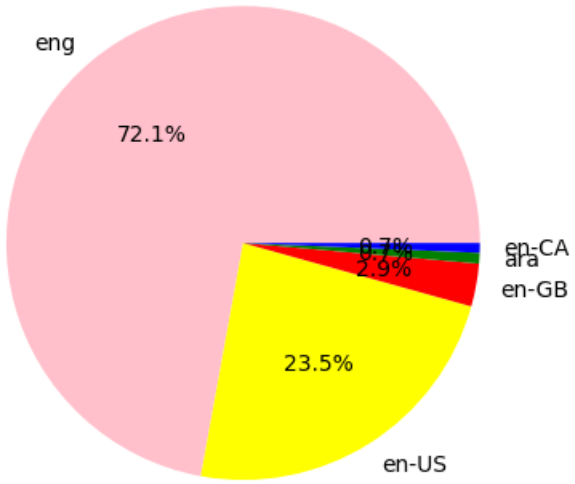

Figure 1: Pie Chart depicting the languages of books in the dataset.

Answers - The above analysis shows the languages of books and relative numbers. The language with the maximum number of books is English (having the language code eng).

Q2) *Which authors have the maximum number of books in the dataset boks.csv?*

A) Approach:

Count the values per entry in the authors column of the books.csv dataset. Sorting these values in

```
a = books["authors"].dropna().value_counts().head(15)
a = dict(a)

plt.barh(list(a.keys()),list(a.values()),color = 'orange')
plt.grid()
plt.show()
```

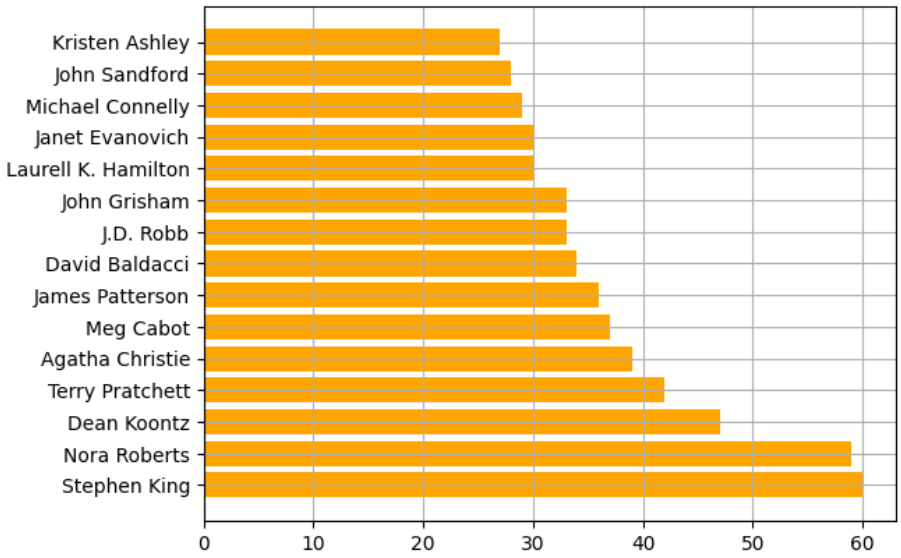

Figure 2: Horizontal Bar Graph denoting the number of books by the top 15 authors.

Answer - The top 5 authors are (based on number of books in the dataset books.csv):

authors	Number of books
Stephen King	60
Nora Roberts	59
Dean Koontz	47
Terry Pratchett	42
Agatha Christie	39

Q3) *What is the average rating of books in the dataset? Which book in the dataset books.csv has the highest rating? Who authored it?*

A) Approach:

Use describe() to analyze the mean and maximum rating. Plot it using matplotlib.

```
r.describe()
```

```
count    10000.000000
mean         4.002191
std          0.254427
min          2.470000
25%          3.850000
50%          4.020000
75%          4.180000
max          4.820000
Name: average_rating, dtype: float64
```

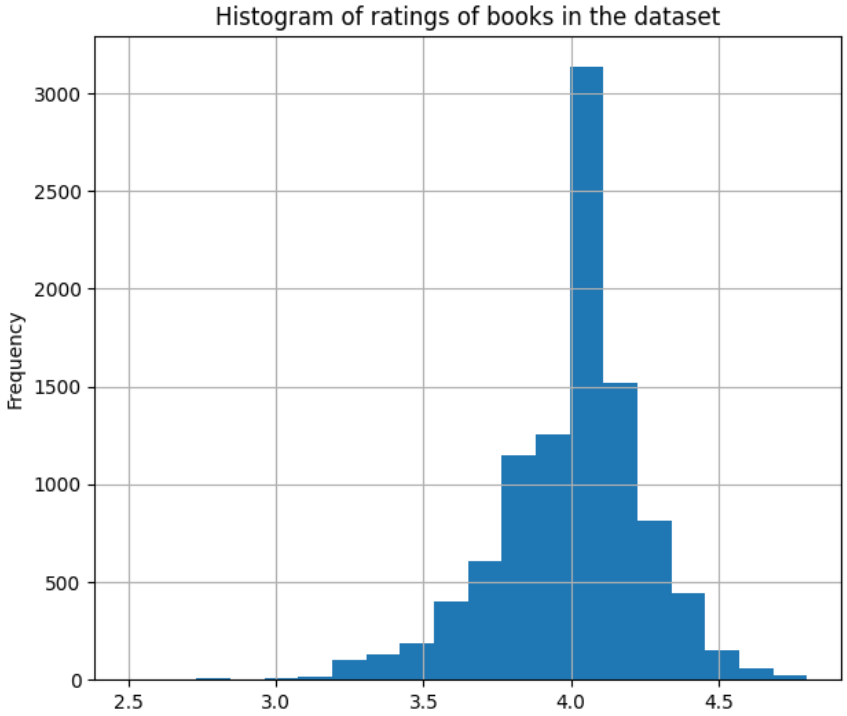

Figure 3: Histogram showing the frequency of ratings for books in the dataset.

The highest rating is 4.8, so we take the entry in columns original_title, author, and average_rating of the books.csv dataset having entry in the average_rating >= 4.8.

	Highest Rated
Author	Bill Watterson
Book Title	The Complete Calvin and Hobbes
Rating	4.82

Answer – The average rating is 4 (out of 5), the highest rating is 4.82, and the book with the highest rating is The Complete Calvin and Hobbies, written by Bill Watterson.

Q) *Who are the most active users? (based on the number of ratings)*

Approach:

Count the values per entry in the user_id column of the ratings.csv dataset. Sorting these values in descending order gives the most active users at the beginning.

```
rr = ratings["user_id"].value_counts().head(10)
rr = pd.DataFrame(rr)
```

	# of Ratings
12874	200
30944	200
52036	199
12381	199
28158	199
45554	197
6630	197
37834	196
15604	196
7563	196

Answer – The most active users are mentioned above.

Q5) *Which books are more likely to be read in the near future? (based on the number of to-read tags)*
Approach:

Count the number of to_read tags per book using the to_read dataset.

```
x = pd.DataFrame(tr["book_id"].value_counts().head(5))
top_tr_id = tr["book_id"].value_counts().head(5).index
tr_df = pd.DataFrame([[None,None,None]]*5)
tr_df.columns = ["Book Title","Book Id","To Read Tags"
tr_df.index = np.arange(1,6)
tr_df.iloc[:,1] = top_tr_id
for i in range(10000):
    b_id = books.iloc[i,0]
    for j in range(5):
        if top_tr_id[j] == b_id:
            tr_df.iloc[j,0] = books.loc[i,"title"]
for i in range(5):
    tr_df.iloc[i,2] = x.iloc[i,0]
```

Answer – The 5 books most likely to be read and their book ids are:

	Book Title	Book Id	To Read Tags
1	The Book Thief	47	2772
2	All the Light We Cannot See	143	1967
3	Catch-22	113	1840
4	1984	13	1812
5	The Kite Runner	11	1767

V. Hypothesis

"A book having more number of ratings has a greater number of reviews."

Approach – Plot using matplotlib, the ratings count versus the reviews count for all 10,000 books in the dataset, and then find the least square solution using SciPy's linalg.lstsq() to analyze the trend.

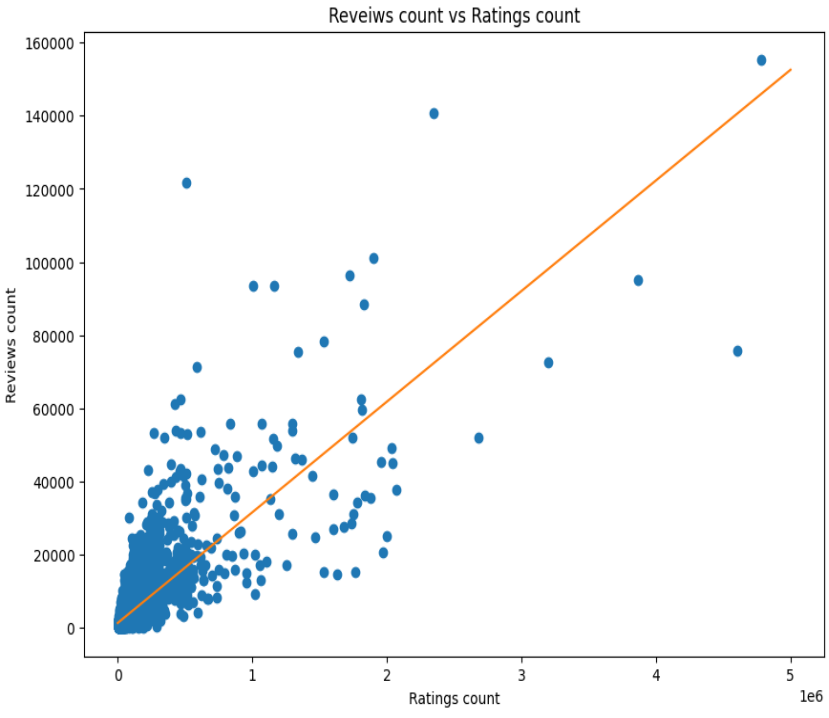

Figure 4: Plot of reviews count v/s ratings count and its least square solution.

Conclusion – The graph shows a positive correlation (the trend is increasing). Hence, a book having more number of ratings is likely to have a greater number of reviews.

VI. Interesting Observation

The average rating of books published in the year narrow down towards 4 as the year of publication increases from 1500 to 2023.

This is clearly evident from the graph of the average rating of books in the year versus the publishing year:

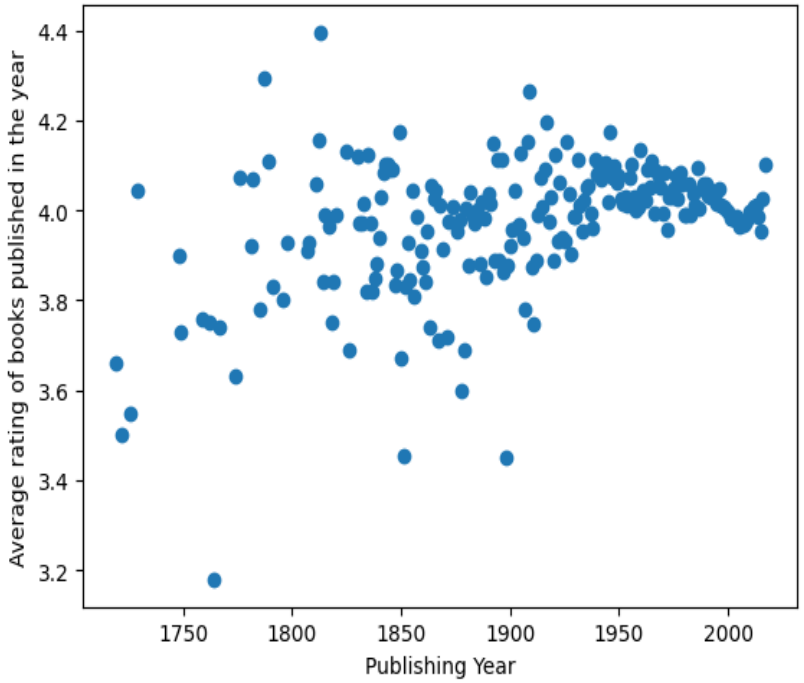

Figure 5: Plot of average rating of books in the year v/s publishing year for books in the dataset.

This unusual behavior makes the observation interesting.

VII. Summary of the Observations

The observations show that most of the books in the dataset are written in English language. The average rating of books is 4 with the highest being 4.82 out of 5. Some books have been tagged 'to read' and are likely to be read in near future. The author with the maximum number of books in the dataset is Stephen King.

VIII. Acknowledgments

I referred to examples on https://www.kaggle.com/datasets to know how data is analyzed.