Data Narrative -1

ES-114

Name : Aryan Sahu

Roll no. : 22110038

I. Overview of the Dataset

The given data is similar to an online catalogue for different kinds of books. It contains various books and ratings from readers with diverse interests and backgrounds. This dataset is a helpful guide for new readers who need suggestions on which books to read and for exploring the vast culture of reading.

II. Scientific Questions/Hypotheses

We can understand many inferences by analysing the dataset. Following are some valuable inferences we would want to get from the given data:

A. *What are some frequent tags associated with books written by a certain author? What do they indicate about his/her writings?*

B. *How has the production level of books changed over the course of time? What does it indicate regarding the new generation?*

C. *Many people give ratings without actually reading them. Also, some publishers use unfair means to get ratings to either gain popularity or increase sales. Provide a piece of evidence for this using the given data of the ratings.*

D. *List the authors who have written the maximum number of books. What do the numbers indicate?*

E. *Give the total number of ratings given in a year along with the year. How has this varied over the past two decades? What relation is seen between the number of ratings and the number of books published each year?*

III. Important Libraries and Functions Used

Following are some of the libraries and functions used to mine the dataset effectively and thoroughly:

A. *Pandas*

Pandas is a library offered by the python programming language, which helps analyse tabular data. It offers creating data types consisting of rows and columns and helps analyse large amounts of data.

B. *Read CSV*

In the Pandas library, we are using the read_csv function to import the CSV file to read the data from and analyse it for our benefit.

C. *Dataframe*

We are creating dataframes as a medium to represent the data inside the python program. Dataframes provide various in-built functions which are useful to derive important inferences from the given dataset.

D. *Loc function*

We have used the loc function to select the particular rows of data according to our requirements using some conditions.

E. *Matplotlib*

It is a library that offers plotting methods and data visualisation options. We have used this library to plot the diverse data and visualise it with the help of graphs.

IV. Solutions to the Questions

We are using the above-specified tools to mine the dataset and obtain various inferences. Following are the answers got from this process.

A. *What are some frequent tags associated with books written by a certain author? What do they indicate about his/her writings?*

We can create a table having various tags given to J.K. Rowling's books vs their frequencies. This will give us a brief idea of what kind of books she writes and from what perspective people see her works. We see that J.K Rowling's books have been associated with words such as 'fiction', 'fantasy' and 'favorites' etc. These indicate that her books are about fiction and fantasy and appeal to people's imaginations.

```
fiction                 8
books-i-own             8
fantasy                 8
favorites               8
to-read                 8
                       ..
harry_potter            1
have-read               1
historical              1
historical-fiction      1
1                    فانتزی
Name: tag_name, Length: 348, dtype: int64
```

Fig. 1 Different tags along with its frequency of being associated to Rowling's books.

B. How has the production level of books changed over the course of time? What does it indicate regarding the new generation?

We plot the data of the year vs the number of books published in that year and observe:

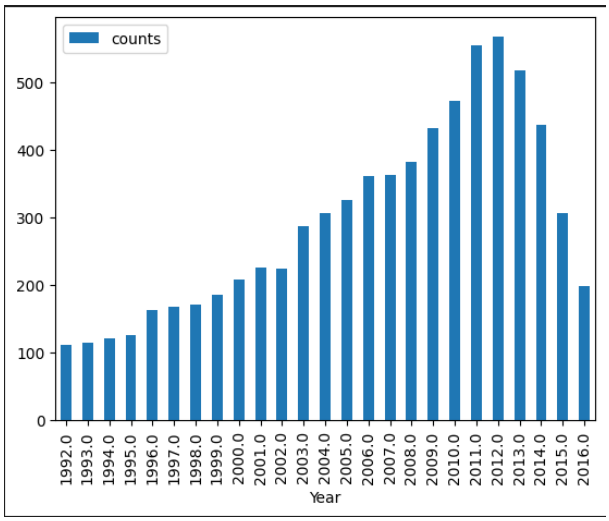

Fig. 2 Number of books published per year from 1992 to 2016.

We can see that there was an enormous increase during the late 2000s. This indicates that reading books was at its peak during this period. Afterwards, there is a sharp decline in the number of publications per year. This indicates that with the upcoming internet and other entertainment media, the publication of physical books has decreased. It means that the new generation is losing interest in the field of writing and reading novels. Writing is being seen more as a hobby than as a profession.

C. Many people give ratings without actually reading them. Also, some publishers use unfair means to get ratings to either gain popularity or increase sales. Provide a piece of evidence for this using the given data of the ratings.

We get the number of books that a particular user has continuously rated the same. This data is displayed in the figure.

```
user_id  rating
53293    5         194
49288    5         193
49289    5         192
49295    5         192
49297    5         192
                  ...
20557    1           1
20555    2           1
10977    2           1
42956    1           1
26027    2           1
Length: 236277, dtype: int64
```

Fig. 3 No. of books a user has rated the same multiple times.

As we can observe, many users have rated numerous books. Some have even rated hundreds of books as five stars which seems illogical. No person could read these many books and like each one of them. This data suggests that the users have rated them wrongly. These user ids are probably being paid to increase the average ratings of their books, and they are rating them blindly without even reading them.

D. List the authors who have written the maximum number of books. What do the numbers indicate?

We can make a table suggesting the top ten authors who have written the maximum number of books as follows:

```
Stephen King        60
Nora Roberts        59
Dean Koontz         47
Terry Pratchett     42
Agatha Christie     39
Meg Cabot           37
James Patterson     36
David Baldacci      34
J.D. Robb           33
John Grisham        33
```

Fig. 4 The top ten authors with the number of books they have written.

We can see that the maximum number of books written here is 60, which is quite large, but for being the maximum, it may seem less. This indicates that authors take plenty of time to write each book with perfection, and they don't focus on quantity but on quality.

E. Give the total number of ratings given in a year along with the year. How has this varied over the past two decades? What relation is seen between the number of ratings and the number of books published each year?

We can plot the total number of ratings vs year as follows:

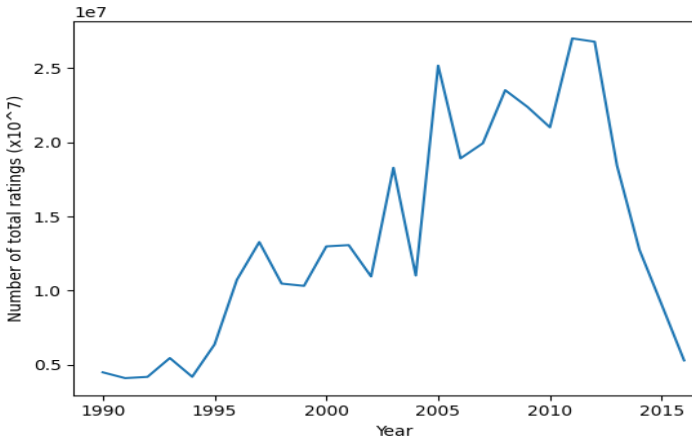

Fig. 5 Number of ratings in a year vs the year.

Observing the graph, we can interpret that the number of ratings had increased till the late 2000s and then started decreasing from the next decade. There are some sharp peaks which indicate that the rating pattern was not very regular. But, if we take a general line that traces the values approximately, it resembles the curve we had seen in Fig. 2, which was about the number of books published that year. One might think that older books would have more ratings because they have existed for longer periods, but this is not the case. Thus, the number of ratings is proportional to the number of books published.

V. Summary of the Observations

We observed a wide variety of interpretations made out of the given dataset. We saw how the user's tags could tell us about the author's writing style. We understood how the trend of writing changed over the years using the number of books published each year. We also observed that the number of ratings each year depends on the number of books published in that year and not on the period for which they have existed. We also gathered evidence indicating the presence of fake ratings and why ratings could mislead users. From this data, we also observed how most authors go for quality than quantity. All these observations help us understand the behaviour of readers and writers.

VI. Acknowledgment

I would like to express my hearty gratitude to our course instructor, Prof. Shanmuga, for allowing me to work on this project and to have this wonderful experience. I also thank all the Teaching Assistants for their guidance across this project and for helping me grasp all the learnings.

Aryan Sahu
Roll no. 22110038

VII. References

This narrative has used the following references to gain knowledge of using the libraries and functions:

1. Pandas Documentation-

   https://pandas.pydata.org/docs/

2. Stack Overflow-

   https://stackoverflow.com/

3. Geeks for Geeks-

   https://www.geeksforgeeks.org/

4. McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56)

5. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.