(Data Narrative-Assignment 1

Probability, Statistics, Data Visualisation)

-11/02/2023

- Kaushik
- IIT Gandhinagar
- Chemical Engineering
- Kaushik.kaushik@iitgn.ac.in

***Abstract*—This document comprises different results from the data which has been provided to us like the number of books published in different years, popular authors and the whole data being compressed to some graph which makes reading the data easy. Bar graph etc. has been used to present data in**

I. Overview of the Dataset

In this assignment, we have been given the following dataset of different books, which include the name of the author of the book,book_id,goodreads_book_id,best_book_id,work_id,books_count,isbn,isbn13,authors,original_publication_year,original_title,title,language_code,average_rating,ratings_count,work_ratings_count,work_text_reviews_count,ratings1,ratings2,ratings_3, ratings_4,ratings_5,image_url,small_image_url . From this data we have to make 5 scientific questions from which we can comfortably read the data.

II. Question on Dataset

Question 1) What is number of books published by each author and find which author published maximum books from the dataset and also plot its graph?

Question 2) What is the frequency distribution of books and its relation with ratings and also plot its graph.

Question 3) What is the probability that if a book is published after 1950, it will have a high rating?

Question 4)What is the graph published each year and which year has the highest publications and also visualize with graph. Derive general trend from this data.

Question 5) How are books rated on an average and what does the data on books tell u?

Question 6) What are the top ten most most rated books in dataset?

III. Details of Libraries and Functions Used in Program

The most comman libraries used in our representation of data is pandas, Numpy, matplotlib there are thousands of functions that we can use and there are hundreds of libraries which make our work easy and some of them are used in this assignment. These librariers helped us to visualize data very clearly and we also discuss in details about uses of these data. Some of the most used are mentioned below:

A. Pandas (Most used library in this assignment)

Define pandas : " Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data." Why it is used ? It is used to break down big data into a statistical manner which becomes easy to read. Functions used in ' panda' library.

- '**pd.read_csv**' is used to read the file from the given link and saves it in a variable which we declare and help us to print.
- "**df.head()**"used to define till what values we have to take the data.
- "**df.tail()**"used to define data from starting from end to maximum value of tail.
- "**values-count()**" is used to count how many time a element is present there it returns us a value in series.
- "**mean()**" is used to find the mean Equations
- "**id.max()**" is used to find the max from the present .

*B. **matplotlib.pyplot**(Second most used library in this assignment)*

Functions used in this :

- **"plot()"** is used to plot the graph between x and y .
- The " **show()"** is used to show the plotted graph.
- **"title()"** is used to print the title at the bottom of the graph.
- **"x-label()"**is used to print the x label on the graph .
- "**y-label()"** is used to print the label at y axis of the graph.

These are the main function that has been used in this assignments.

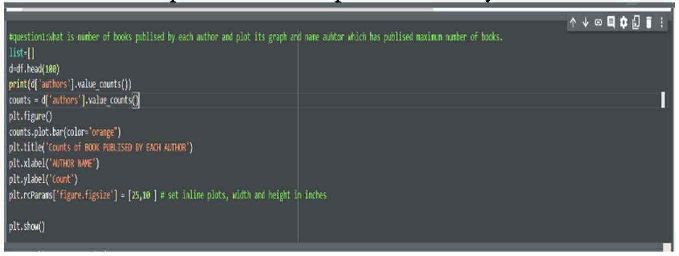
Identify applicable funding agency here. If none, delete this text box.

IV. Answers

The answer these questions are as following:

A. Answer 1

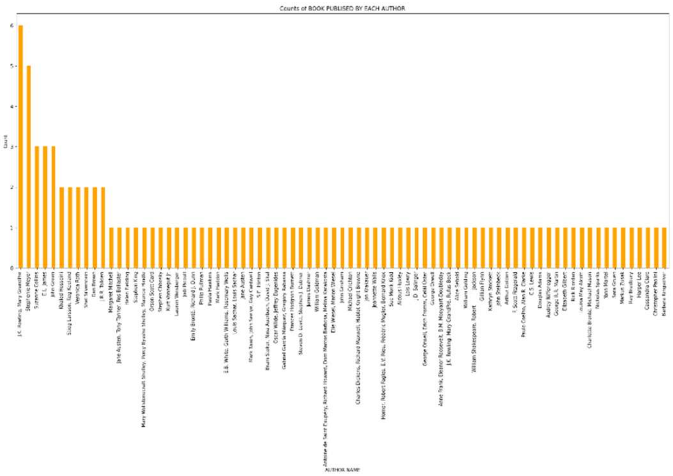
The code to plot number of publications by each author :



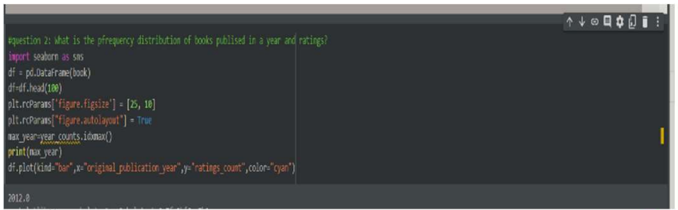
```
#question1:what is number of books publised by each author and plot its graph and name auhtor which has publised maximum number of books.
list=[]
d=df.head(100)
print(d['authors'].value_counts())
counts = d['authors'].value_counts()
plt.figure()
counts.plot.bar(color="orange")
plt.title('Counts of BOOK PUBLISED BY EACH AUTHOR')
plt.xlabel('AUTHOR NAME')
plt.ylabel('Count')
plt.rcParams['figure.figsize'] = [25,10 ] # set inline plots, width and height in inches

plt.show()
```

And the subsequent graph of the question is below, we can clearly see J.K. Rowlings and Mary GrandPre has published maximum number of books which is Six(6)

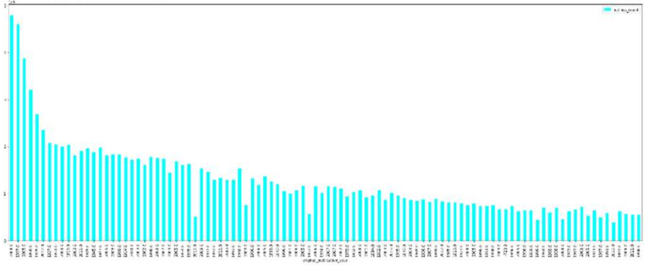


B. Answer 2



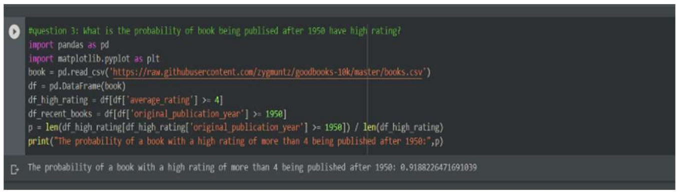
```
#question 2: What is the pfrequency distribution of books publised in a year and ratings?
import seaborn as sns
df = pd.DataFrame(book)
df=df.head(100)
plt.rcParams['figure.figsize'] = [25, 10]
plt.rcParams["figure.autolayout"] = True
max_year=year_counts.idxmax()
print(max_year)
df.plot(kind="bar",x="original_publication_year",y="ratings_count",color="cyan")
```

```
2012.0
```



The solution of second was to plot frequency distribution of books being published in each year. We can clearly see 2012 was the year in which maximum books were published.

C. Answers 3



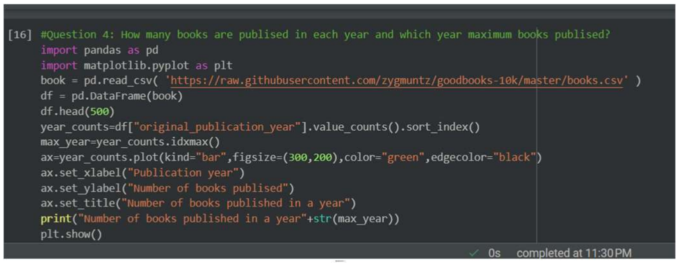
```
#question 3: What is the probability of book being publised after 1950 have high rating?
import pandas as pd
import matplotlib.pyplot as plt
book = pd.read_csv('https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/books.csv')
df = pd.DataFrame(book)
df_high_rating = df[df['average_rating'] >= 4]
df_recent_books = df[df['original_publication_year'] >= 1950]
p = len(df_high_rating[df_high_rating['original_publication_year'] >= 1950]) / len(df_high_rating)
print("The probability of a book with a high rating of more than 4 being published after 1950:",p)
```

```
The probability of a book with a high rating of more than 4 being published after 1950: 0.9188226471691039
```

In this question, it was asked to find probability of a book if it was published after 1950 and will also have high ratings. Answer came out to be `0.9188226471691039`

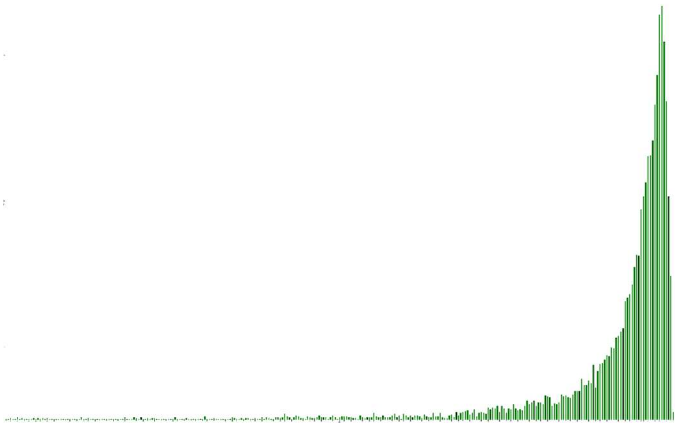
[a] footnote. (*Table footnote*)

D. Answers 4



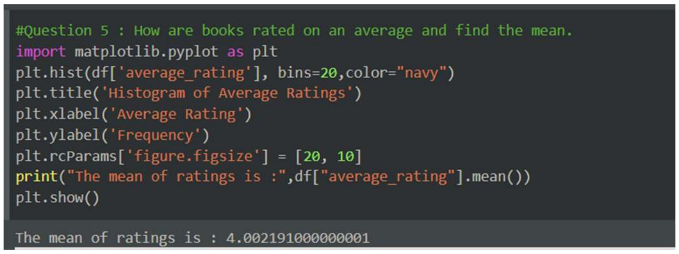
```
[16] #Question 4: How many books are publised in each year and which year maximum books publised?
     import pandas as pd
     import matplotlib.pyplot as plt
     book = pd.read_csv( 'https://raw.githubusercontent.com/zygmuntz/goodbooks-10k/master/books.csv' )
     df = pd.DataFrame(book)
     df.head(500)
     year_counts=df["original_publication_year"].value_counts().sort_index()
     max_year=year_counts.idxmax()
     ax=year_counts.plot(kind="bar",figsize=(300,200),color="green",edgecolor="black")
     ax.set_xlabel("Publication year")
     ax.set_ylabel("Number of books publised")
     ax.set_title("Number of books published in a year")
     print("Number of books published in a year"+str(max_year))
     plt.show()
```

✓ 0s completed at 11:30 PM



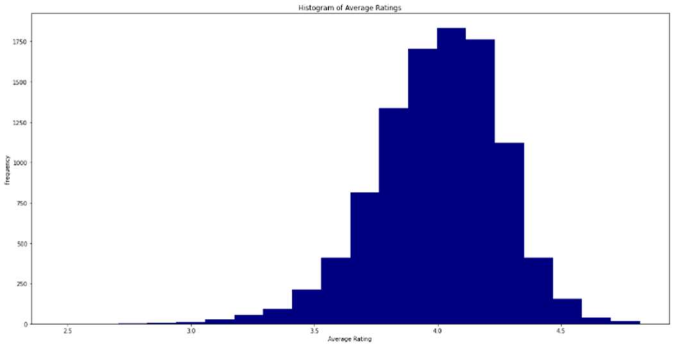
The solution of second was to plot No. of books being published in each year. We can clearly see 2012 was the year in which maximum books were published

E. Answers 5



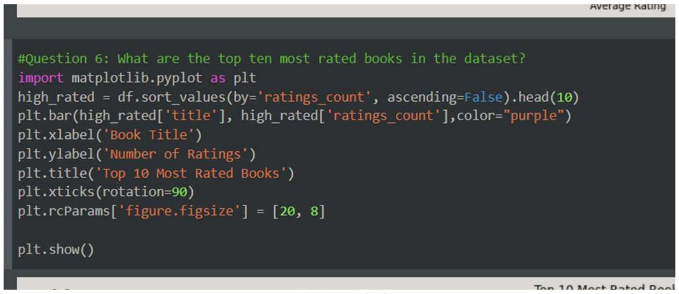
```
#Question 5 : How are books rated on an average and find the mean.
import matplotlib.pyplot as plt
plt.hist(df['average_rating'], bins=20,color="navy")
plt.title('Histogram of Average Ratings')
plt.xlabel('Average Rating')
plt.ylabel('Frequency')
plt.rcParams['figure.figsize'] = [20, 10]
print("The mean of ratings is :",df["average_rating"].mean())
plt.show()
```

```
The mean of ratings is : 4.002191000000001
```



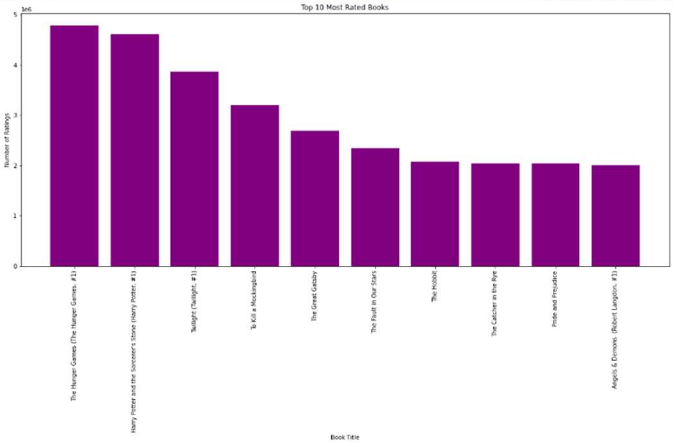
This graph shows are books were critiqued by the readers and we can clearly see on an average books in the dataset is good. This is pretty good data if you want to start with books as priority.

F. *Answer 6*



```
#Question 6: What are the top ten most rated books in the dataset?
import matplotlib.pyplot as plt
high_rated = df.sort_values(by='ratings_count', ascending=False).head(10)
plt.bar(high_rated['title'], high_rated['ratings_count'],color="purple")
plt.xlabel('Book Title')
plt.ylabel('Number of Ratings')
plt.title('Top 10 Most Rated Books')
plt.xticks(rotation=90)
plt.rcParams['figure.figsize'] = [20, 8]

plt.show()
```


Top 10 Most Rated Books
1e6
Number of Ratings
The Hunger Games (The Hunger Games, #1)
Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
Twilight (Twilight, #1)
To Kill a Mockingbird
The Great Gatsby
The Fault in Our Stars
The Hobbit
The Catcher in the Rye
Pride and Prejudice
Angels & Demons (Robert Langdon, #1)
Book Title