1. Introduction

This paper is approaching the question of whether long books (≥ 1000 pages) have a significant average rate than shorter ones (<1000 pages) or not. A pre-processed Goodreads data (n = 4173) is used to conservatively answer this question. The analysis was conducted by constructing a conservative t-distribution of the difference of means ($\mu_{(x1\ -\ x2}$). The analysis is testing the null hypothesis ($\mu_o = 0$) against the alternative hypothesis ($\mu_a \neq 0$), using a significance level $(\alpha) = 0.05$.

2. Dataset

The analysis was carried out using book reviews Goodreads website data($n = 11127$). The data consists of 12 variables: bookID, title, authors, average rating, ISBN, ISBN13, language code, pages number, rating count, reviews count, publication date, publisher.

The report examines a particular question: is there compelling evidence that long books (≥ 1000 pages) have a different rating than shorter ones (<1000 pages)?

Consequently, the research is only concerned with the average rating and the pages number variables in the dataset. Both variables are quantitative discrete variables. A quantitative discrete variable is a variable type in which there is a minimum distance between one value and the next one.

- The average rating is a quantitative discrete variable because the minimum distance between one value and the next one in the data is0.01. For instance, the next value to 4.00 cannot be less than 4.01.
- The pages number is a quantitative discrete variable because the minimum distance between one and the next one is 1. For example, the next value to 700 pages cannot be less than 701.[1]

In order to get unbiased or unreliable results, three conservative data cleaning and processing method were conducted as follows:

- Non-numerical inputs were removed from both variables using Excel in order to be able to apply statistical analyzing tools, like mean and standard deviation.
- Using Pandas library in Python, books with0number of pages were removed because this is logically flawed as there is no book with 0 pages.
- Using Pandas library in Python, books with 0 ratings were removed because they didn't have an average rating.

[1] #variables: Accurately identified the variables for the provided data and justified these identification.

3. Analysis

In order to answer the main question of whether the average rating for long books is different from short books, a difference of mean significance test was performed using Python and Pandas library (Appendix A). The books' average ratings were split into two categories: (≥ 1000 pages) and short books (<1000 pages) (Appendix B).

Two hypotheses were developed to answer the research question: null and alternative ones.

- Null hypothesis (H_o): the average rating between long books (≥ 1000 pages) and short books (<1000 pages) are the same. Consequently, the difference between their means is equal to 0: ($\mu_o = 0$).
- Alternative hypothesis (H_a): the average rating between long books (≥ 1000 pages) and short books (<1000 pages) are not the same. Consequently, the difference between their means is not equal to 0: ($\mu_a \neq 0$).

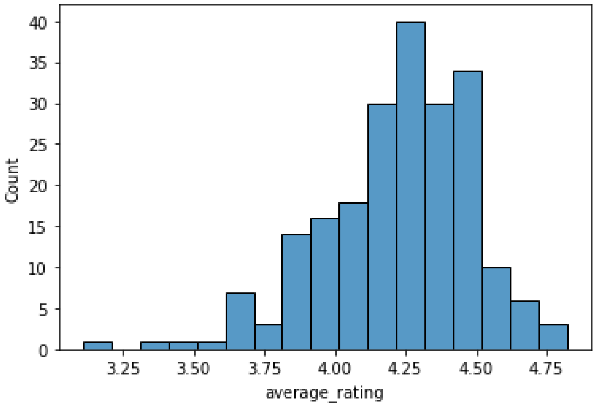
The significance level was set to the default one ($\alpha = 0.05$)because it creates the balance between the occurrence of type I and types II errors.

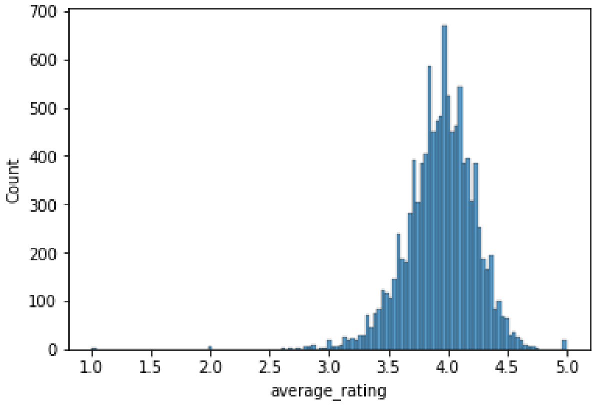
- Type I error is rejecting the true null hypothesis, which is saying that long books and short books don't have the same average rating as they truly do have.
- Type II error is failing to reject the false null hypothesis, saying that long books and short books have the same average rating while they truly don't have.

Summary Statistics

Table 1: Summary statistics for the books' average rating for the study's two groups: long books (≥ 1000 pages) and short books (<1000 pages)

Statistical measurement	Long books (≥ 1000 pages)	Short books (<1000 pages)
Count	$n = 215$	$n = 10753$
Mean	$\bar{x}_2 = 4.22$	$\bar{x}_1 = 3.94$
Median	4.25	3.96
Standard deviation	$s_2 = 0.26$	$s_1 = 0.29$
Range	1.71	4.00





Fig(1) Average rating for long books(≥ 1000 pages)

Fig(2) Average rating for short books (<1000 pages)

Data interpretation:

Both data's histograms are skewed to the left, which means that their tail is on the left side. In a left-skewed histogram, we can conclude that the mean is shifted to the left of the median. Consequently, the mean will be less than the median. This agrees with the descriptive stats results; as in the long books (≥ 1000 pages) group, the mean is equal to 4.22 while the median is equal to 4.25. Similarly, in the short books (<1000 pages) group, the mean is equal to 3.94 while the media is equal to 3.96.

Although the histograms are skewed to the left, the data is gathered between [4.00,4.50] and [3.5,4.5] bins for long books (≥ 1000 pages) and short books (<1000 pages), respectively, which indicates a low standard deviation. This is also in agreement with the descriptive stats as $s_1 = 0.29$and $s_2 = 0.26$, which is relatively low standard deviations (Appendix C, D, and E). [2] [3]

Although both groups have a large sample size($n > 60$), the analysis followed a conservative approach and used t-distribution for drawing inferences. As the study doesn't have the population's standard deviation, t-distribution is used to approximate the sampling distribution. The t-distribution criteria have been met as follows:[4]

[2] #dataviz: effectively generated a detailed data visualization appropriate for the data; effectively analyzed and interpreted the data visualization.

[3] #descriptivestats: the mean, median, and standard deviation of the data were interpreted and analyzed. Then, they were integrated with the data visualization

[4] #distributions: the using of the t-distribution was justified by analyzing the sample size and proportion from the population.

Table (2) represents the criteria for using t-score distribution and how the analysis met it.

Criteria	Long books (≥ 1000 pages) sample characteristics the met the criteria	Short books (<1000 pages) sample characteristics that met the criteria
- Slight skew sample size should be 15 - The moderate skew sample size should be 30 - The strong skew sample size should be 60	The sample size is 215	The sample size is 10753
The sample should be only 10% of the population to be independent	Books are a nearly infinite population that we don't know their actual number. Some estimates point out that it may be around 100,000,000 books existing (Madrigal, 2011). Consequently, 215 is less than 10% of the whole population.	Books are a nearly infinite population that we don't know their actual number. Some estimates point out that it may be around 100,000,000 books existing (Madrigal, 2011). Consequently, 10753 is less than 10% of the whole population.

Difference of means significance test:

Firstly, a two-tail test is used because the analysis measures any change away from $\mu \neq 0$in the positive or negative direction. In order to apply the difference of means significance test, the p-value is calculated and compared with the significance level. In order to do so, the weighted standard error for the difference between is calculated: $SE = 0.018$. Then, the T-score was calculated: $T = 15.75$.

The t-distribution has different shapes according to the used sample size, and those shapes are determined by the degrees of freedom(dof): dof = n - 1. Following the conservative approach, the calculated degree of freedom by using the minimum sample size$n = 215$is equal to 214. Finally, the p-value was calculated to be approximately 0 ($p \simeq 0$). The significance level is corrected to be 0.025 ($\alpha = 0.025$) by applying Bonferroni correction because getting two means from the same sample induces multiple comparisons problems: the actual significance level is different from the intended one.

As $p < \alpha$ (0 < 0.025), the null hypothesis (Ho) is rejected. The reason behind this very small p-value and very large t-value is the large sample size($n = 11127$), which points out that it's very unlikely that the difference between the population means for long books (≥ 1000 pages) and short books (<1000 pages) is zero. Consequently, the null and alternative hypothesis are statistically significant.

To know the practical significance of the difference between those means, the effect size was measured using Cohen's d because it's designed to compare between two groups. Moreover, it has defined ranges to convert the effect size from numerical numbers to either low, medium, or high. Cohen's d was calculated to be equal to 0.98 using a pooled standard deviation = 0.29 (Appendix F). Consequently, according to cohen's d correction as 0.98 > 0.8, this means that the practical significance of the hypothesis is high. Finally, we can conclude that the means of long books (≥ 1000 pages) and short books (<1000 pages) are statistically significant and have high practical significance in real life.[5] [6]

Confidence interval:

The null and alternative hypotheses can be further tested by constructing confidence intervals for each of the two groups: long books (≥ 1000 pages) and short books (<1000 pages). Through these intervals, the study presents a range of values that the mean is probable to belong to it by a certain confidence level. The confidence interval that is used is 95% because of the same reasons presented when choosing the significance level (α). Moreover, the same criteria for the inference that was mentioned above validate the use of t-score in constructing the intervals.

Calculating the confidence interval is done by calculating the lower and upper limit of the interval by this formula: $\mu \pm (t - score) \times SE$while $SE = SD/\sqrt{n}$. Below is the calculated confidence intervals (Appendix F):

- Short books (<1000 pages): [3.93, 3.94]
- Long books (≥ 1000 pages): [4.19, 4.26]

In each interval, we can be 95% confident that the population means to belong to it. This means that by a probability of 95%, the population means of the short books (<1000 pages) is bigger than 3.93 and smaller than 3.94. On the other hand, the long books (≥ 1000 pages) population mean, by a probability of 95%, is bigger than 4.19 and smaller than 4.26 . Knowing that the two intervals don't interfere with each other supports rejecting the null hypothesis because it implies that the possible values set of the population mean for the two groups don't have common

[5] #probability: the p-value is explained and identified to analyze an hypothesis

[6] #significance: difference between means significance level was conducted and its usage was justified.

values. Consequently, the population means for the two groups are likely to be different $\mu \neq 0$.[7] [8]

4. Results and Conclusions

This report has investigated Goodreads books reviews data set to answer a question if long books' (≥ 1000 pages) average rating is different from the short books' (<1000 pages) average rating. Based on statistical and practical tests $p \simeq 0,\ cohen's\ d = 0.98$, there is evidence to support that the population average rating is statistically and practically different in both groups. This conclusion is supported by the non-overlapping 95% confidence intervals that were constructed for both groups [3.93, 3.94] and [4.19, 4.26], which means that the population average rating for long books probably isn't the same as the population average rating for short books.

These conclusions are generalization inductive arguments because they use statistical evidence to support a general conclusion about the population. Consequently, these conclusions go beyond the content of the sample, and their truth values cannot be guaranteed that they will be 100% true. However, their strength and reliability can be assessed.

Firstly, this argument is strong because it properly and conservatively used statistical tools (difference between means significance test, confidence interval, and t-distribution) and justified their usage. The criteria conditions to use these tools were met and verified. Moreover, the results conclusions were aligned with those tools' results and avoided hasty generalization by following the conservative approach, which is being cautious when driving conclusions about confidence intervals and hypothesis testing.

Furthermore, this argument is reliable because it used a large sample-sized data (n=11127) from books reviews website that gather these data from real users entries. However, this reliability can be questioned because of the left skewness of the data, which resulted from books with a low number of rating counts. For instance, there are a lot of books that have <10 ratings, which gives a chance for biased skewness as each vote will change the average rating significantly. To further investigate this matter, an unbiased clearance process can be conducted to remove books that have an average rating below a calculated threshold. This threshold should satisfy that it won't bias the data towards certain results.[9]

5. References

1- Goodreads API (2017). Goodreads-books reviews data V2. Location: Soumik Github.

[7] #confidenceintervals: The usage of confidence intervals is justified and validated. Then, the intervals for both groups were constructed.

[8] #probability: a confidence level was set to test a hypothesis and its interpratepation was written in terms of probability.

[9] #induction: an argument was detected as an inductive one. Morever, the strength and reliability of the argument were assessed.

2- Madrigal, A. C. (2010, August 5). Google: There are exactly 129,864,880 books in the world. The Atlantic. Retrieved December 13, 2021, from https://www.theatlantic.com/technology/archive/2010/08/google-there-are-exactly-129-864-880-books-in-the-world/61024/.

6. Reflection

"Ok not bad! You are somewhat headed in the right direction here but I think you missed the key realization for this poll. When it says, "within 2 standard deviations of the mean," that means we're looking for the area between a t-score of -2 and +2. Compared to what we know about normal distributions, where the answer would be about 95%, we expect to get less than that. Review our discussion afterward where we worked out the probability to be 93.13%.

This poll question was taken directly from the recommended reading and practice in the study guide and is meant to assess your understanding of the fundamentals.

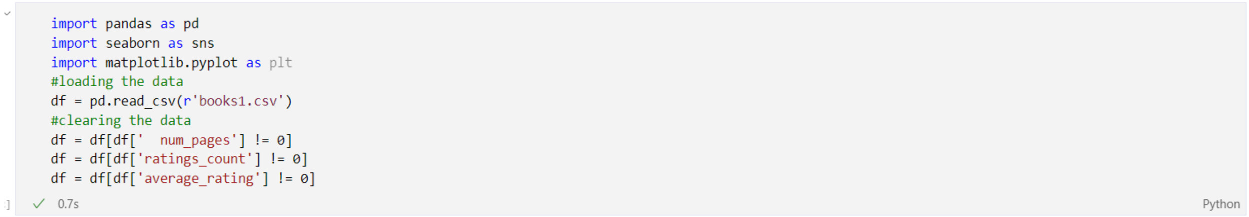
Going forward, (1) make sure that you understand how to answer this question because we'll be doing similar conversions from t-scores to probabilities for the rest of the unit, and (2) make sure that you do all of the practice questions in the study guide to prepare adequately for class. I recommend prioritizing the practice questions over the reading because it's a great way to test yourself, and they're more fun to do than reading! Remember that instructors and TAs are happy to assist you with the practice questions if you get stuck!"

I received this feedback on the HC #distribution, which encouraged me to improve on my CS level overall. Consequently, I reviewed the #scienceoflearning again to develop a strategy to help me do that. Firstly, I used the distributed practice technique to understand sample distribution and binomial distribution concepts. My practice started with recalling the definitions for both concepts and then solving some of the pre-class work problems. I distributed this practice at different time intervals that were increasing every time I did well in practice. I started with two days between practice tell reached 11 days now. After observing the success of the strategy in those topics, I used it in multiple other HCs like #deduction, #induction, and #probability.
I believe this was effective because it utilized one science of learning principle stated by Kossyln, 2017, which is spaced practice. The spaced practice principle states that it's better to use the information multiple times over a relatively long interval of time in order to learn it.[10]

[10] #Scienceoflearning: A science of learning technique was applied to improve HCs skills. Moreover, the use of this technique is analyzed and justified using science of learning principles.

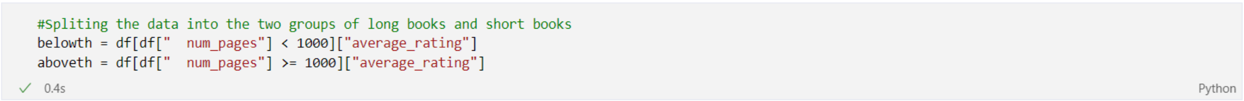
7. Appendix

A.



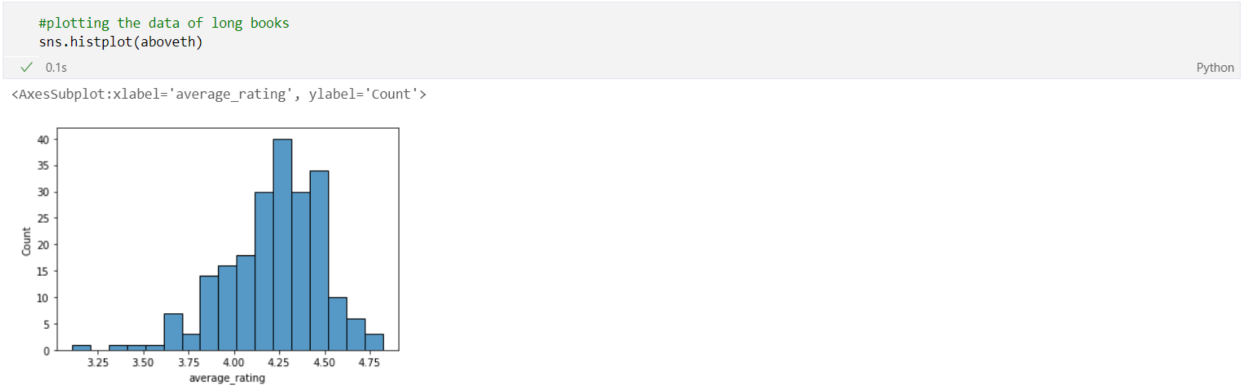
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
#loading the data
df = pd.read_csv(r'books1.csv')
#clearing the data
df = df[df['  num_pages'] != 0]
df = df[df['ratings_count'] != 0]
df = df[df['average_rating'] != 0]
```

B.



```
#Spliting the data into the two groups of long books and short books
belowth = df[df["  num_pages"] < 1000]["average_rating"]
aboveth = df[df["  num_pages"] >= 1000]["average_rating"]
```

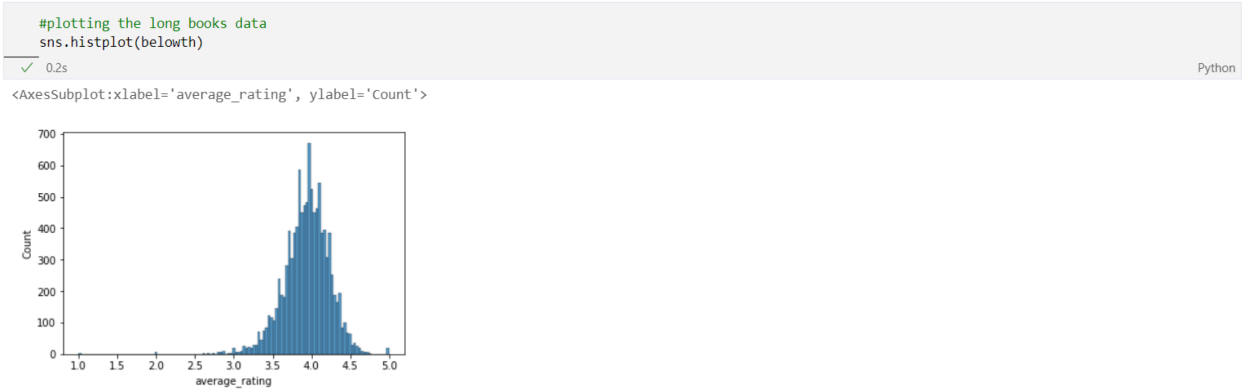
C.



```
#plotting the data of long books
sns.histplot(aboveth)
```

```
<AxesSubplot:xlabel='average_rating', ylabel='Count'>
```

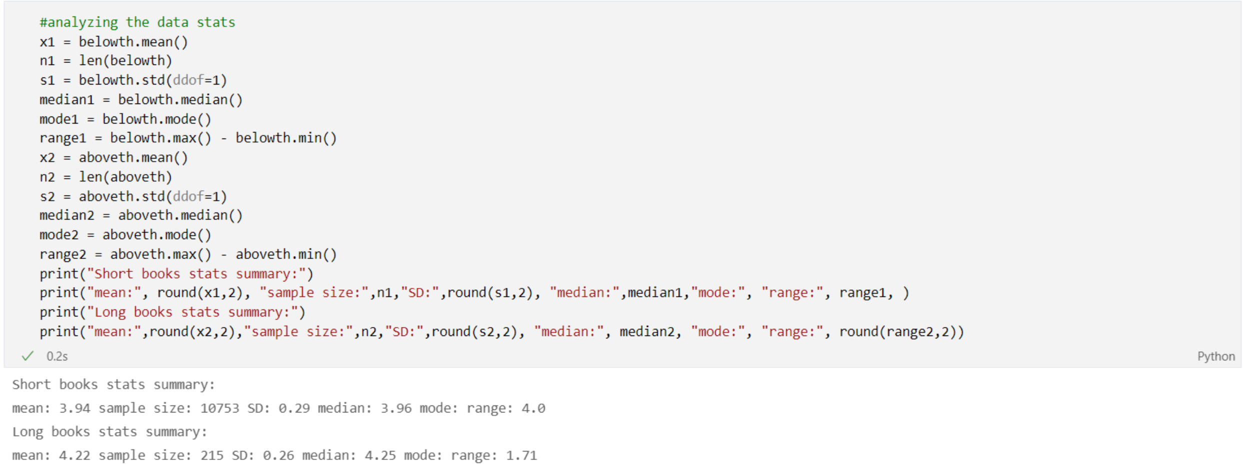
D.



```
#plotting the long books data
sns.histplot(belowth)
```

```
<AxesSubplot:xlabel='average_rating', ylabel='Count'>
```

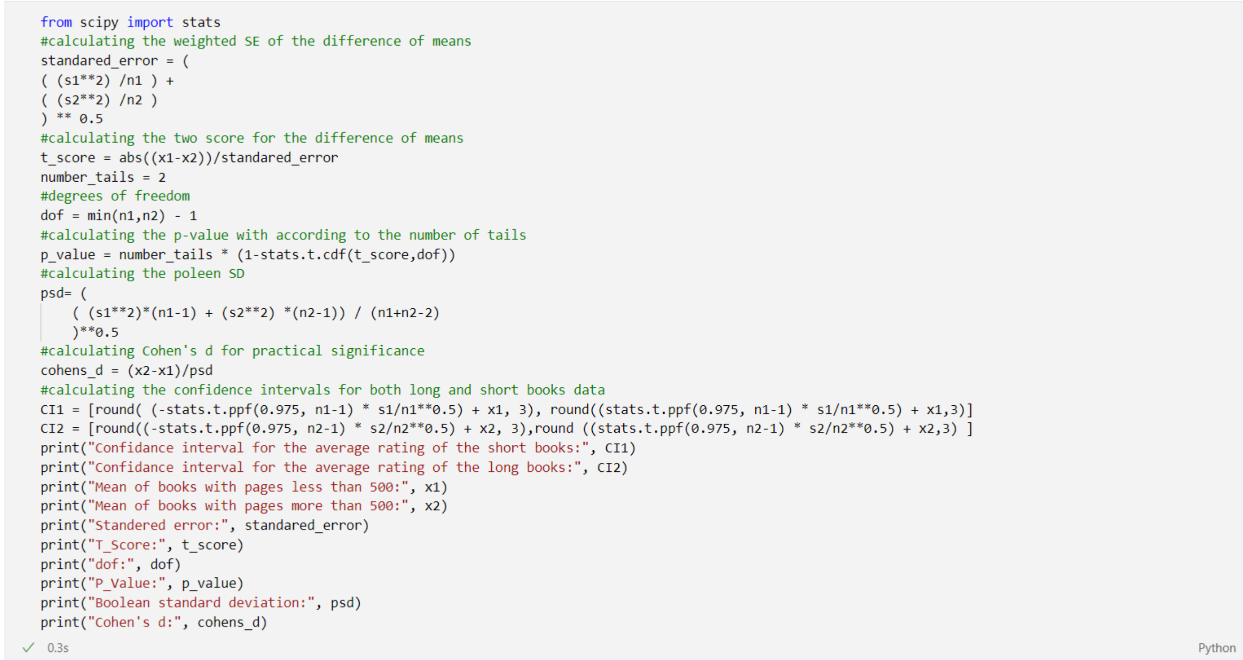
E.



```
#analyzing the data stats
x1 = belowth.mean()
n1 = len(belowth)
s1 = belowth.std(ddof=1)
median1 = belowth.median()
mode1 = belowth.mode()
range1 = belowth.max() - belowth.min()
x2 = aboveth.mean()
n2 = len(aboveth)
s2 = aboveth.std(ddof=1)
median2 = aboveth.median()
mode2 = aboveth.mode()
range2 = aboveth.max() - aboveth.min()
print("Short books stats summary:")
print("mean:", round(x1,2), "sample size:",n1,"SD:",round(s1,2), "median:",median1,"mode:", "range:", range1, )
print("Long books stats summary:")
print("mean:",round(x2,2),"sample size:",n2,"SD:",round(s2,2), "median:", median2, "mode:", "range:", round(range2,2))
```

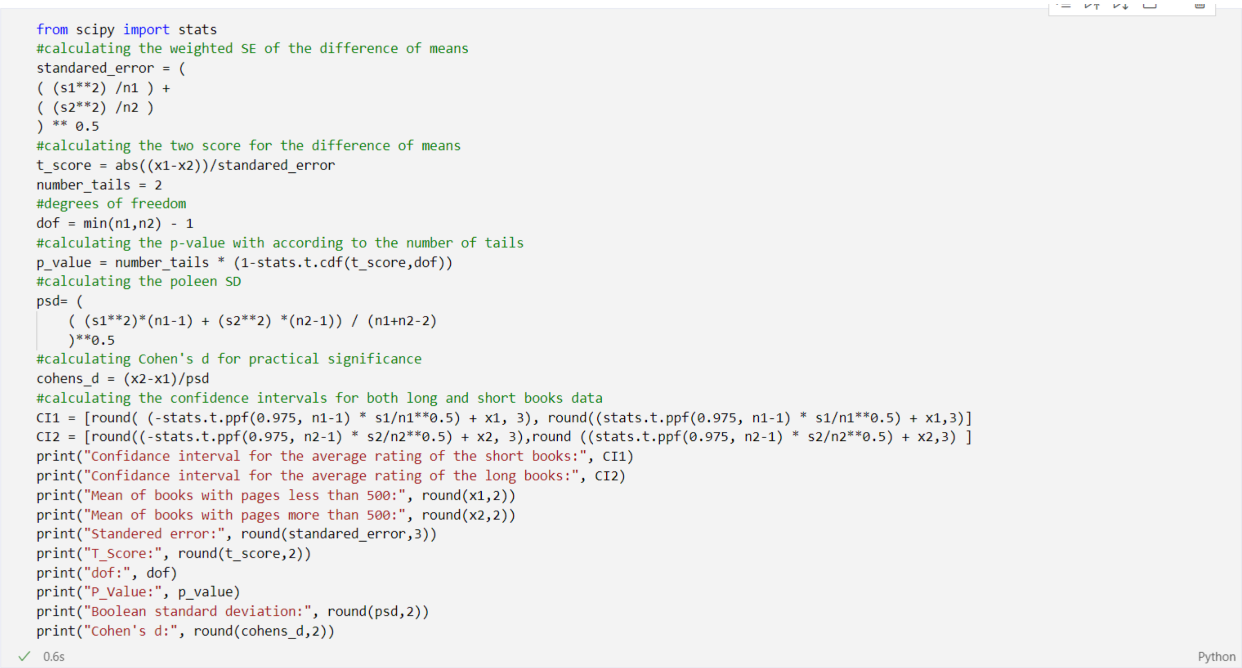
```
Short books stats summary:
mean: 3.94 sample size: 10753 SD: 0.29 median: 3.96 mode: range: 4.0
Long books stats summary:
mean: 4.22 sample size: 215 SD: 0.26 median: 4.25 mode: range: 1.71
```

F.



```
from scipy import stats
#calculating the weighted SE of the difference of means
standared_error = (
( (s1**2) /n1 ) +
( (s2**2) /n2 )
) ** 0.5
#calculating the two score for the difference of means
t_score = abs((x1-x2))/standared_error
number_tails = 2
#degrees of freedom
dof = min(n1,n2) - 1
#calculating the p-value with according to the number of tails
p_value = number_tails * (1-stats.t.cdf(t_score,dof))
#calculating the poleen SD
psd= (
    ( (s1**2)*(n1-1) + (s2**2) *(n2-1)) / (n1+n2-2)
    )**0.5
#calculating Cohen's d for practical significance
cohens_d = (x2-x1)/psd
#calculating the confidence intervals for both long and short books data
CI1 = [round( (-stats.t.ppf(0.975, n1-1) * s1/n1**0.5) + x1, 3), round((stats.t.ppf(0.975, n1-1) * s1/n1**0.5) + x1,3)]
CI2 = [round((-stats.t.ppf(0.975, n2-1) * s2/n2**0.5) + x2, 3),round ((stats.t.ppf(0.975, n2-1) * s2/n2**0.5) + x2,3) ]
print("Confidance interval for the average rating of the short books:", CI1)
print("Confidance interval for the average rating of the long books:", CI2)
print("Mean of books with pages less than 500:", x1)
print("Mean of books with pages more than 500:", x2)
print("Standered error:", standared_error)
print("T_Score:", t_score)
print("dof:", dof)
print("P_Value:", p_value)
print("Boolean standard deviation:", psd)
print("Cohen's d:", cohens_d)
```

G.



```python
from scipy import stats
#calculating the weighted SE of the difference of means
standared_error = (
( (s1**2) /n1 ) +
( (s2**2) /n2 )
) ** 0.5
#calculating the two score for the difference of means
t_score = abs((x1-x2))/standared_error
number_tails = 2
#degrees of freedom
dof = min(n1,n2) - 1
#calculating the p-value with according to the number of tails
p_value = number_tails * (1-stats.t.cdf(t_score,dof))
#calculating the poleen SD
psd= (
    ( (s1**2)*(n1-1) + (s2**2) *(n2-1)) / (n1+n2-2)
    )**0.5
#calculating Cohen's d for practical significance
cohens_d = (x2-x1)/psd
#calculating the confidence intervals for both long and short books data
CI1 = [round( (-stats.t.ppf(0.975, n1-1) * s1/n1**0.5) + x1, 3), round((stats.t.ppf(0.975, n1-1) * s1/n1**0.5) + x1,3)]
CI2 = [round((-stats.t.ppf(0.975, n2-1) * s2/n2**0.5) + x2, 3),round ((stats.t.ppf(0.975, n2-1) * s2/n2**0.5) + x2,3) ]
print("Confidance interval for the average rating of the short books:", CI1)
print("Confidance interval for the average rating of the long books:", CI2)
print("Mean of books with pages less than 500:", round(x1,2))
print("Mean of books with pages more than 500:", round(x2,2))
print("Standered error:", round(standared_error,3))
print("T_Score:", round(t_score,2))
print("dof:", dof)
print("P_Value:", p_value)
print("Boolean standard deviation:", round(psd,2))
print("Cohen's d:", round(cohens_d,2))
```

```
Confidance interval for the average rating of the short books: [3.932, 3.943]

Confidance interval for the average rating of the long books: [4.189, 4.26]

Mean of books with pages less than 500: 3.94

Mean of books with pages more than 500: 4.22

Standered error: 0.018

T_Score: 15.75

dof: 214

P_Value: 0.0

Boolean standard deviation: 0.29

Cohen's d: 0.98
```