Book page length and average reader ratings.

Math Analysis & Approaches SL Internal Assessment

Rawan Manjal

TABLE OF CONTENTS

BOOK PAGE LENGTHS & AVERAGE READER RATINGS

I. INTRODUCTION

An avid reader all my life, I have come to analyze my reading habits and question what makes a book enjoyable to me. After researching into the *New York Times* bestseller list and the trends it has gone through in recent years, it became clear that the average length of a standard book on the market has increased from around 350 pages in 1999 to around 400 in 2014.[1] But were the reader ratings favoring such a change?

To investigate this and see if there was any correlation between a longer page length and higher average reader ratings, I catalogued the 90 books that I have read in the past 3 years, their page count and their average *Goodreads* rating. I chose to use *Goodreads* ratings instead of any other platforms' due to its large user base of over 90 million in 2019, back catalogue of over 14 years and simple rating system where readers can review a book using a 5 star scale.[2] I also chose to use my own personal book collection as it varied enough due to the large sample, and so I could see if the trend would be prominent in an average collection.

My aim in this investigation is to firstly find if there is a linear correlation between the two variables of book length and average ratings and how strong it is if present. I would then attempt to find the most accurate way to model this relationship. However, before that, I will have find and eliminate any outliers. I will then compare how close to the data each of these models is through the calculation of percentage error and finally determine which equation should be used to model such a set of data.

II. DATA POINTS & SCATTER PLOT GRAPH

I catalogued the number of pages in the most popular edition of each book and the average rating from the site, which was given accurate to 2 decimal places on each book's page. A portion of the data is shown in *Table 1* to the side, and the full list of 90 books is seen under *Appendix 1*.

Table 1: Portion of Data

Title	*Page Length (x)*	*Average Rating (y)*
Paper	21	3.18
The Grownup	64	3.57
Caffeine	120	3.71
The Magician's Elephant	201	3.83
Big Rock	235	3.94
Everything, Everything	306	4.05
Cinder	387	4.13
A Torch Against the Night	452	4.19
Heir of Fire	562	4.26
Obsidio	672	4.37
A Game of Thrones	835	4.45

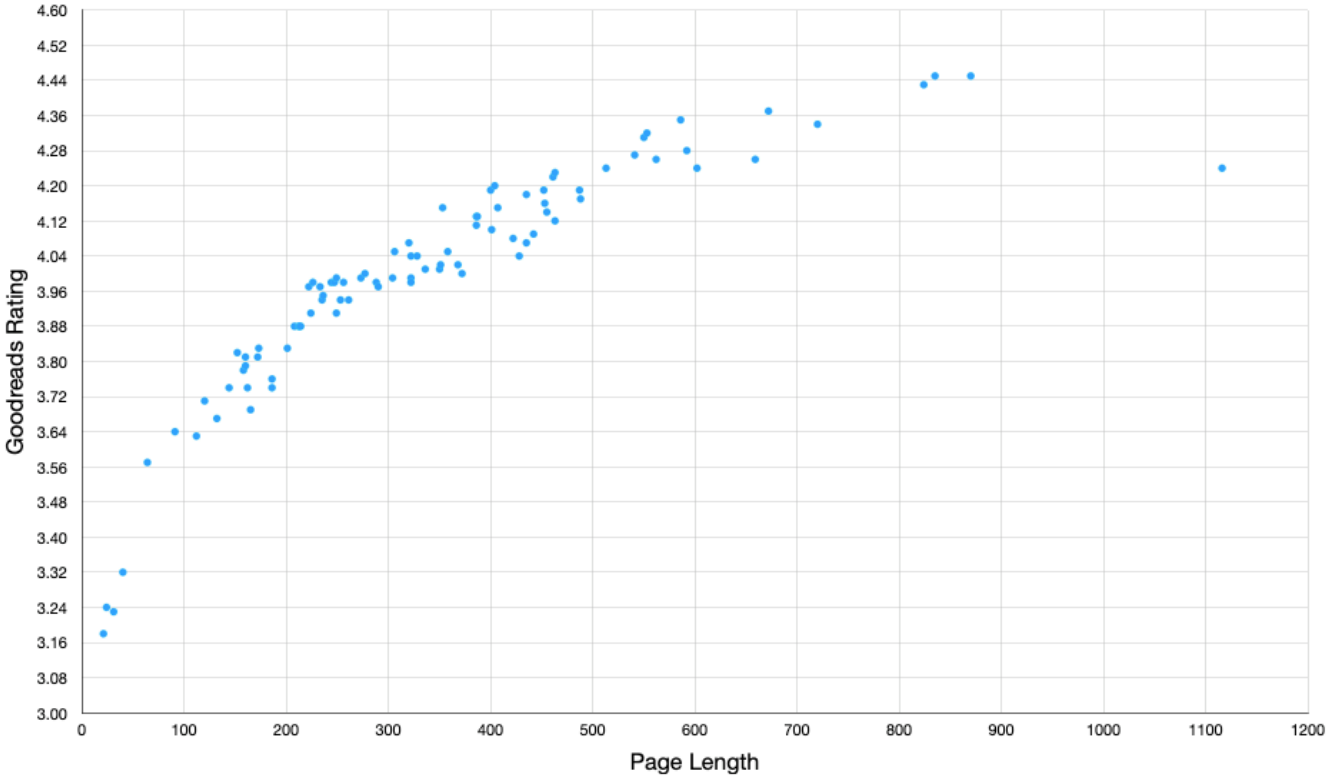
After collecting the data, I inputed my points into a spreadsheet software. It then plotted them on a scatterplot graph. These points are plotted on an x-axis representing the page length versus a y-axis

[1] Lea, Richard. "The Big Question: Are Books Getting Longer?" *The Guardian*, Guardian News and Media, 10 Dec. 2015, www.theguardian.com/books/2015/dec/10/are-books-getting-longer-survey-marlon-james-hanya-yanagihara.

[2] Clement, J. "Goodreads: Number of Registered Members 2019." *Statista*, 1 Oct. 2020, www.statista.com/statistics/252986/number-of-registered-members-on-goodreadscom/.

representing average reader reviews. The *x*-axis' scale is increasing in 75 page steps while the *y*-axis is moving in 0.08 steps. I chose to start from 3.00 on the *y*-axis and end at 4.60 to best suit my data and represent it in the best way, this is due to my data set having a minimum of 3.18 rating and 4.21 being the highest rating amongst the data. The graph is labeled *Graph 1* and is seen below.

Graph 1: Scatter Plot of Page Length vs. Goodreads Rating



Generally, when the x is increasing, the y is also increasing which suggests that there is a positive correlation . I will model the data with linear correlation and check its strength. However, firstly, I will have to find and eliminate any outliers.

III. FINDING & ELIMINATING OUTLIERS

Outliers are unusual values in a data set and can distort my later calculations for the correlation between variables and the finding of an accurate model. As such, they need to be calculated and removed. To do this, the upper and lower quartiles need to be found.

To find the quartiles, the median is first found as follows:

$\text{median} = \left(\frac{n+1}{2}\right)^{\text{th}}$ term where n is the number of values in a data set.

$\text{median} = \left(\frac{91}{2}\right)^{\text{th}}$ term

$\text{median} = (45.5)^{\text{th}}$ term, meaning it's the average of the 45th and 46th ordered data values.

$$\text{median} = \frac{320 + 322}{2} = 321$$

Now the data can be split evenly in half and the upper and lower quartile can be found:

lower quartile = median of lower half of data

$$\text{lower quartile} = \left(\frac{46}{2}\right)^{\text{th}} \text{ term} = 23^{\text{rd}} \text{ term} = 212$$

upper quartile = median of upper half of data

$$\text{upper quartile} = \left(\frac{46}{2}\right)^{\text{th}} \text{ term} = 23^{\text{rd}} \text{ term} = 452$$

Through these values, the Interquartile Range (IQR) can be found as seen below.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

$$\text{IQR} = 452 - 212 = 240$$

And finally the upper and lower boundary can be found:

$$\text{upper boundary} = \text{upper quartile} + 1.5 \times \text{IQR} \qquad \text{lower boundary} = 212 - (1.5 \times 240) = -148$$

$$\text{upper boundary} = 452 + (1.5 \times 240) = 812$$

$$\text{lower boundary} = \text{lower quartile} - 1.5 \times \text{IQR}$$

Through this calculations, 4 outliers will be eliminated from the upper extreme as their page counts exceed 812. However, since a page count cannot be negative, the lower boundary of -148 pages did not affect the data set. The 4 outliers are found in *Table 2*.

Table 2: The Outliers Removed

Title	*Page Length (x)*	*Average Rating (y)*
Winter	824	4.43
HP & Order…	870	4.45
It	1116	4.24
Game of Thrones	835	4.45

IV. FINDING A CORRELATION

To first determine if there is a linear correlation amongst my two variables of page length (x) and *Goodreads* rating (y), I will find the Pearson correlation coefficient which is denoted by the variable r. This will allow be to determine how strong, if any, of a relationship is present. The formula for the Pearson correlation coefficient, r, is shown below and the variables are defined below it.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

r = Pearson correlation coefficient
x = values of the x-variable in the set of data (page count)
$\bar{x}$ = mean of the x values
y = values of the y-variable in the set of data (rating)
$\bar{y}$ = mean of the y values

Table 3: Portion Pearson's Coefficient Calculations

Title	*Page Length (x)*	*Actual Rating (v_A)*	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
Paper	21	3.18	-296.209	-0.795	235.590	87739.951	0.633
The Grownup	64	3.57	-253.209	-0.405	102.638	64114.951	0.164
Caffeine	120	3.71	-197.209	-0.265	52.329	38891.509	0.070
The Magician's Elephant	201	3.83	-116.209	-0.145	16.891	13504.602	0.021
Big Rock	235	3.94	-82.209	-0.035	2.906	6758.369	0.001
Everything, Everything	306	4.05	-11.209	0.075	-0.837	125.648	0.006
Cinder	387	4.13	69.791	0.155	10.793	4870.741	0.024
A Torch Against the Night	452	4.19	134.791	0.215	28.933	18168.532	0.046
Heir of Fire	562	4.26	244.791	0.285	69.680	59922.486	0.081
Obsidio	672	4.37	354.791	0.395	140.019	125876.439	0.156
A Game of Thrones	835	4.45	517.791	0.475	245.770	268107.207	0.225

In order to find this correlation coefficient, I will have to calculate the different components of it, which I have chosen to organize in a table form. The same 11 books from *Table 1* have been used to show a portion of the calculations as viewable in *Table 3*, while the calculation for all data points can be viewed under *Appendix 2*.

$$\sum \left((x - \bar{x})(y - \bar{y})\right) = 3013.254$$

$$\sum (x - \bar{x})^2 = 2142682.233$$

$$\sum (y - \bar{y})^2 = 5.073$$

Through these calculations, I am now able to find the sums I need to plug into the formula for *r*. The values are seen below. Note that all values used for calculations are rounded to 3 decimal places.

These values are now plugged into the formula to find *r*.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

$$r = \frac{3013.254}{\sqrt{(2142682.233)(5.073)}}$$

$$r = \frac{3013.254}{\sqrt{10870126.944}}$$

$$r \approx 0.913942$$

Note that the correlation coefficient was left rounded to 6 significant figures to not compromise accuracy in later calculations.

With the correlation coefficient coming out between 0.87 and 0.95, it can be said that the correlation between page length (x) and Goodreads rating (y) is a strong positive linear correlation.[3] This means I achieved my second goal of finding if there was a correlation present and its strength. Although this finding

[3] "Chapter 21: Linear Modelling." *Mathematics for the International Student: Mathematics SL*, by Robert Haese et al., Third Edition ed., Haese & Harris Publications, 2012, pp. 546–556.

does not mean that a longer page count *causes* a higher rating, it does point to a correlation which can then be used to predict a book's rating.

V. LINEAR REGRESSION

Now that I have the *r* correlation coefficient, I can use it in my first attempt to model the regression of the data. I will first try to use linear regression. The linear regression function is shown below with its variables defined.

$y = a + bx$

y = dependent values of the y-variable in the set of data (page count)
a = y-intercept
b = slope of the regression line
x = values of the x-variable in the set of data (page count)

To find the equation of the line, I will need to find the slope b and the y-intercept, a. Firstly, I will find the slope, using the formula below:

$b = r\frac{\sigma_y}{\sigma_x}$

r = Pearson correlation coefficient
σ_y= standard deviation of y data points
σ_x= standard deviation of x data points

While the *r* value is known, further calculations will be done to find the standard deviations for both the y and x values. To find the standard deviations, the following formula was used. While the formula indicates the values used for the x data points, the same formula will be used for the y points.

$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{N_x - 1}}$

σ_x= standard deviation of x data points
x = values of the x-variable in the set of data (page count)
$\bar{x}$ = mean of the x values
N = total number of points in data set

The working for both σ_x and σ_y are seen below:

$\sum (x - \bar{x})^2 = 2142682.233$
$N_x = 86$
$\sigma_x = \sqrt{\frac{2142682.233}{86}}$
$\sigma_x \approx 157.845$

$\sum (y - \bar{y})^2 = 5.073$
$N_y = 86$
$\sigma_y = \sqrt{\frac{5.073}{86}}$
$\sigma_y \approx 0.243$

Now the values can be plugged in as follows to find the value of b accurate to 4 significant figures:

$$b = (0.913942)\frac{0.243}{157.845}$$

$$b \approx 0.001407$$

Finally, the y-intercept of the regression line will be found as follows:

$$a = \bar{y} - b\bar{x}$$

a = y-intercept

$\bar{y}$ = mean of the y values

b = slope of the regression line

$\bar{x}$ = mean of the x values

$$a = (3.975) - 0.001407(317.209)$$

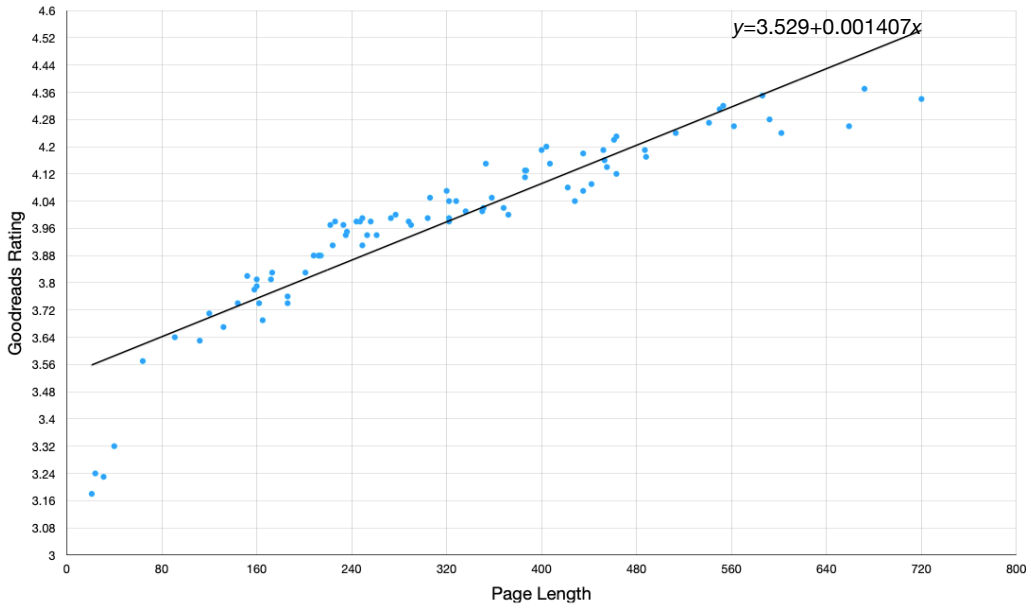
$$a \approx 3.529$$

With all the values known now, the equation representing the linear regression line can be constructed. The equation is found below labeled *Equation 1*.

Equation 1: $y = 3.529 + 0.001407x$

I returned to my original scatterplot (after the elimination of the outliers) and graphed *Equation 1* against the data points as seen with the labeled black line. This graph seen below labelled *Graph 2*.

Graph 2: Scatter Plot with Equation 1



This linear regression line seemed to represent the trend in a general sense but it misses its accuracy at both the extreme lower domain and the extreme higher domain of the x-axis due to the curved shape of the data.

While this inaccuracy can be seen through visual observation, I decided to calculate the percentage error of each point through the following formula:

$$\delta = \left| \frac{v_A - v_C}{v_C} \right| \times 100$$

δ = percentage error

v_A = Actual value (actual *Goodreads* rating)

v_C = Calculated Value (rating predicted by *Equation 1*)

I used this formula to calculate the percentage error for each of my 86 values. A portion of the percentage error calculations is seen in Table 4, with the complete calculations found under Appendix 3.

Table 4: Portion of Percentage Error Calculations

Title	***Page Length (x)***	***Actual Rating (v_A)***	***Equation 1 Rating (v_C)***	***Percentage Error***
Paper	21	3.18	3.55	11.73%
The Grownup	64	3.57	3.60	0.90%
Animal Farm	120	3.95	3.67	7.19%
Mem	184	3.65	3.74	2.44%
The Morning They Came for Us	224	4.09	3.78	7.47%
Murder on the Orient Express	274	4.18	3.84	8.10%
Sharp Objects	321	3.98	3.90	2.13%
To Kill a Mockingbird	376	4.28	3.96	7.53%
The Raven Boys	408	4.04	3.99	1.13%
Library of Souls	463	4.12	4.06	1.53%
Thunderhead	504	4.46	4.10	7.98%
Heir of Fire	562	4.46	4.17	6.50%
A Court of Mist and Fury	629	4.61	4.25	7.89%
A Little Life	816	4.13	4.46	7.99%
Harry Potter and the Order of the Phoenix	870	4.50	4.52	0.48%

After calculating all the percentage errors, I found the mean percentage error for all my data set as follows:

$$\bar{\delta} = \frac{\sum \delta}{86}$$

$$\bar{\delta} = 2.99\,\%$$

VI. POWER REGRESSION

Having calculated the simple linear regression of this data set, I wanted to attempt to model it through power regression since through visual observation, I thought this model would better represent the slight curve of the data. The power regression equation is shown below with its variables defined.

$$y = m x^k$$

y = dependent values of the y-variable in the set of data (page count)

m = y-intercept

x = values of the x-variable in the set of data (page count)

k = constant factor representing the relationship between the x value and the y value

To find the equation of the line, I will need to transform my data in order to use a linear regression formula to find the two constants k and p of the model. This will not linearize the model itself, but merely transform the data to facilitate the determination of the unknowns. This transformation is seen below:

$y = m x^k$

$\ln(y) = \ln(m x^k)$ then simplify using logarithmic laws

$\ln(y) = \ln(m) + \ln(x^k)$

$\ln(y) = \ln(m) + k \ln(x)$

To construct this equation into a linear regression form, I will have to rename components of the model. I will take $\ln(y)$ to be z, the constant $\ln(m)$ to be q, the factor k to remain the same, and $\ln(x)$ to be w moving forward. These allocations are made clearer by the labelling below.

$$\underbrace{\ln(y)}_{z} = \underbrace{\ln(m)}_{q} + \underbrace{k}_{k}\,\underbrace{\ln(x)}_{w}$$

This equation is then rewritten below to show the linear relationship that has come to be between the z and w variables, where z is the independent variable, q is the y-intercept, k is the slope of the line and w is the x value.

$$z = q + k w$$

Using this equation and the relationships, I can start equating and solving for my variables.

I will first attempt to find k. Since my equation is now transformed into a linear form and k now takes the place of the slope, I can use the formula below to find its value in the linear form as follows where the variables represent those defined previously and N is the number of data points.:

$$k = \frac{N\sum wz - \sum w \sum z}{N\sum w^2 - (\sum w)^2}$$

Before I am able to calculate k, I need to find its components. I once again organized the components I needed to calculate in a table form. A portion of the calculations are seen in *Table 5* with the rest of the calculations found under *Appendix 4*.

Table 5: Portion of Calculations for finding k

x	y	$w = \ln(x)$	$z = \ln(y)$	$w \times z$	w^2
21	3.18	3.04452	1.15688	3.52215	9.26912
64	3.57	4.15888	1.27257	5.29245	17.2963
120	3.95	4.78749	1.37372	6.57665	22.9201
184	3.65	5.21494	1.29473	6.75192	27.1956
224	4.09	5.41165	1.40854	7.62255	29.2859
274	4.18	5.61313	1.43031	8.02852	31.5072
321	3.98	5.77144	1.38128	7.97199	33.3095
376	4.28	5.92959	1.45395	8.62134	35.1600
408	4.04	6.01127	1.39624	8.39320	36.1353
463	4.12	6.13773	1.41585	8.69012	37.6717
504	4.46	6.22258	1.49515	9.30368	38.7205
562	4.46	6.33150	1.49515	9.46654	40.0879
629	4.61	6.44413	1.52823	9.84810	41.5268
816	4.13	6.70441	1.41828	9.50872	44.9492
870	4.50	6.76849	1.50408	10.18034	45.8125

$$\sum w = 480.1598$$

$$\sum z = 118.5201$$

$$\sum (w \times z) = 665.4513$$

$$\sum w^2 = 2722.0018$$

Note that the values calculated in Table 4 are left rounded to 6 significant figures and the sums are left to 4 decimal points. This was done as to not compromise the accuracy of later calculations.

Now having the summations needed, I solved for k accurate to 6 significant figures as follows:

$$k = \frac{n\sum wz - \sum w \sum z}{n\sum w^2 - (\sum w)^2}$$

$$k = \frac{(86)(665.4513) - (480.1598)(118.5201)}{(86)(2722.0018) - (480.1598)^2}$$

$$k \approx 0.0904921$$

I will now apply the formula for the axis intercept of a linear regression line as follows:

$$q = \bar{z} - k\bar{w}$$

q = vertical axis intercept

$\bar{z}$ = mean of the z values

k = slope of the regression line

$\bar{w}$ = mean of the x values

$\bar{z} \approx 1.37814$

$\bar{w} \approx 5.58326$

$q = (1.37814) - (0.0904921)(5.58326)$

$q \approx 0.872902$

Now both unknowns are found and the linear equation can be assembled as seen below:

$z = q + k\,w$

$z = 0.872902 + 0.0904921w$

I can now find m from my initial equation by first considering the following to be true:

$$\underbrace{\ln(y)}_{z} = \underbrace{\ln(m)}_{q} + \underbrace{k}_{k}\,\underbrace{\ln(x)}_{w}$$

$$\ln(m) = q$$

$$\therefore m = e^{q}$$

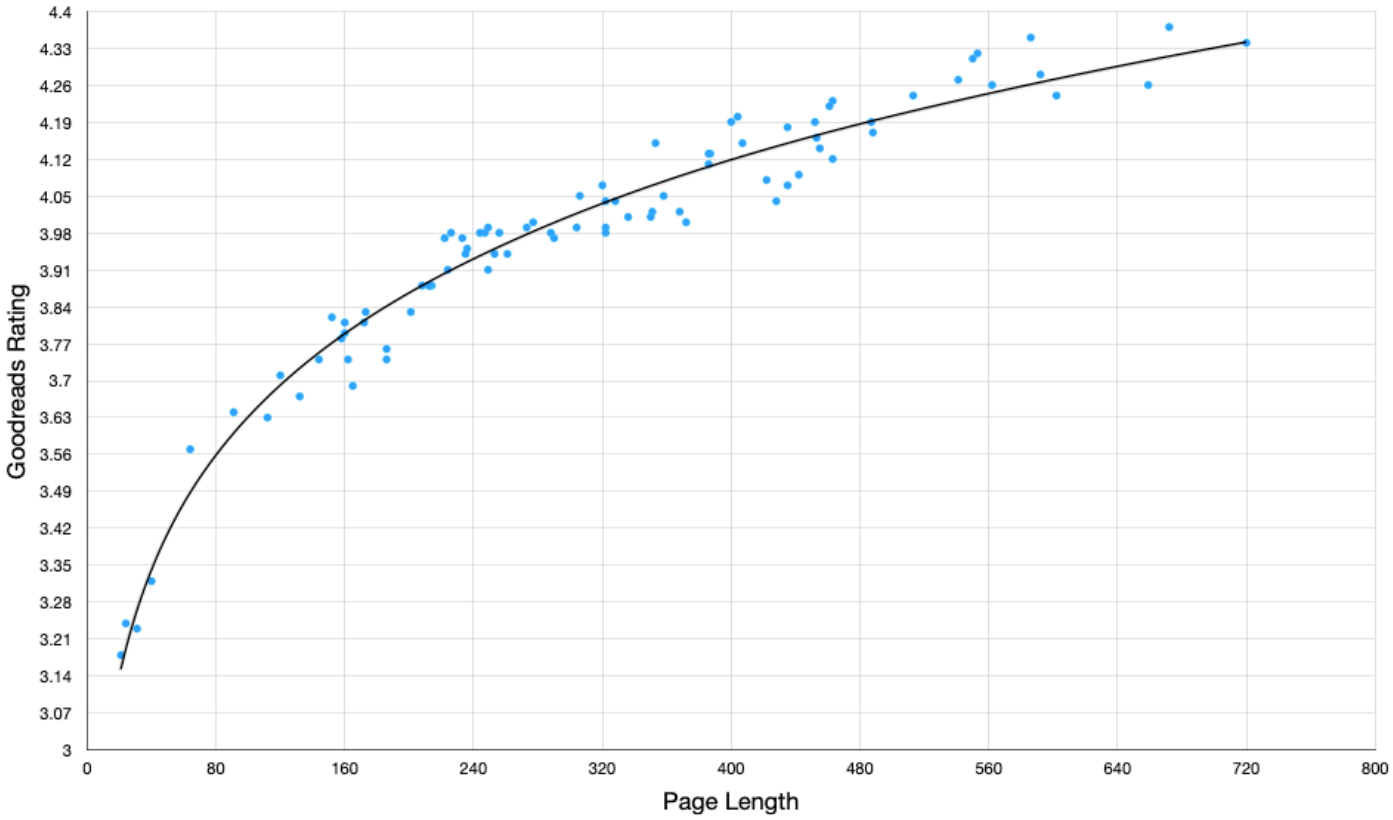
I will now work backwards towards an equation that resembles the power regression model in the form of $y = m + x^{k}$. This will lead me to *Equation 2* which models this relationship.

$m = e^{q} = e^{0.872902} \approx 2.39385$

Equation 2: $y = 2.39385x^{0.0904921}$

I now returned to the scatter plot graph and plotted the line of Equation 2 on it to see if the model represents the data set.

Graph 3: Scatter Plot with Equation 3



Visually, I could see that this model a lot more accurate at representing the shape of the data set compared to the first model since it follows the curve of the data, but to compare its accuracy to the data set, I calculated the percentage error for each data point as I did with *Equation 1* using the formula below.

$$\delta = \left| \frac{v_A - v_C}{v_C} \right| \times 100$$

A portion of the calculations for *Equations* 2's percentage errors are seen in *Table 6* and the rest are under *Appendix 5*.

Table 6: Portion of Percentage Error Calculations for Equation 2

Title	*Page Length (x)*	*Actual Rating (v_A)*	*Equation 1 Rating (v_C)*	*Percentage Error*
Paper	21	3.18	3.15	0.84%
The Grownup	64	3.57	3.49	2.30%
Animal Farm	120	3.95	3.69	6.53%
Mem	184	3.65	3.84	5.14%
The Morning They…	224	4.09	3.91	4.49%
Murder Orient Express	274	4.18	3.98	4.83%
Sharp Objects	321	3.98	4.04	1.40%
To Kill a Mockingbird	376	4.28	4.09	4.35%
The Raven Boys	408	4.04	4.12	2.08%
Library of Souls	463	4.12	4.17	1.25%
Thunderhead	504	4.46	4.20	5.74%
Heir of Fire	562	4.46	4.25	4.81%
A Court of Mist & Fury	629	4.61	4.29	6.96%
A Little Life	816	4.13	4.39	6.32%
Harry Potter & Order	870	4.50	4.42	1.85%

After calculating all the percentage errors, I found the mean percentage error for all my data set as follows:

$$\bar{\delta} = \frac{\sum \delta}{86}$$

$$\bar{\delta} = 0.97\,\%$$

Based off the average percentage error for both *Equation 1* and *Equation 2*, *Equation 2* which used power regression to model the data, is more accurate due to its average percentage error being under 1 percent with a mere 0.97% while *Equation 1*'s average percentage error was higher by 2.01%, sitting at an average error of 2.99%. Therefore, this is the equation I will use to model my data, however, since *Goodreads* allows the highest rating of a book to be 5.00 stars, there will be a limitation to my equation's application, since it will no longer be relevant after the rating reaches 5.00.

I can find when exactly *Equation 2* can no longer be used by finding when the page count stops increasing the rating, as it will have reached its maximum. I will find this page count (x) by setting my equation equal to 5.00 and solving as follows.

$$y = 2.39385x^{0.0904921}$$

$$2.39385x^{0.0904921} = 5.00$$

$$x^{0.0904921} = \frac{5.00}{2.39385}$$

$$\log x^{0.0904921} = \log\left(\frac{5.00}{2.39385}\right)$$

$$0.0904921 \log x = \log\left(\frac{5.00}{2.39385}\right)$$

$$\log x = \frac{\log\left(\frac{5.00}{2.39385}\right)}{0.0904921}$$

$$x = 10^{\frac{\log\left(\frac{5.00}{2.39385}\right)}{0.0904921}}$$

$$x \approx 3426 \text{ pages}$$

Note that the calculated value for pages is rounded to a whole number due to the inability of pages to be fractional.

This means that after a book surpasses 3,426 pages, its rating on *Goodreads* will be at its maximum of 5.00 stars out of a possible 5.00 stars, and any additional pages will no longer affect the readers' rating, so *Equation 2* can no longer be used.

VIII. CONCLUSION & EVALUATION

After the calculation of the data boundaries and removing the outliers from my set of values, I tried finding if there was a linear correlation then used 2 methods of data regression to attempt to achieve my aim of finding a model for the relationship between my books' page length and average reader rating on *Goodreads*. I additionally calculated the percentage error of each so I can determine which is the most accurate model, which turned out to be *Equation 2*.

Some of the inaccuracies in calculations could be credited to the excessive rounding. For most calculations I used values accurate to 3 decimal places, but with values that were extremely small or close together, I chose to use 6 significant figures, hoping it would decrease the inaccuracies. To make calculations more accurate in future, unrounded exact answers should be used.

Additionally, the equation cannot be said to model all of my data since I found and eliminated outliers at the beginning, meaning valuable data points of books in my collection were not taken into consideration. While they would have skewed my model and made it slightly inaccurate in regards to the majority of points, they were not mistakes in data collection, and can thus be seen as valid and correct points eliminated from calculations. However, due to the 4 points in *Table 2* not being critical, and making up a minority of the data set, I chose to remove them. Further investigations could compare the models and calculations with and without extreme points such as those removed.

All in all, considering that *Equation 2* has the lower average percentage error and visually fits the shape of the graph, I will consider it the more accurate model of my data, thus concluding that there is a strong positive power correlation between the page count of a book in my collection and the average *Goodreads* rating, and that after a book reaches 3,426 pages, no more changes will occur to its rating as it will have achieved the maximum 5.00 stars.

This approach however, is inherently flawed, since page length is not the only factor affecting the book rating as correlation as in this case does not mean there is causation, and with the millions of users on the *Goodreads* site, the chance of a book achieving a perfect rating is slim to none. So, while with this small sample of data, *Equation 2* will be able to accurately model book ratings, this may not be the case with a bigger sample.

Through this investigation, I have been able to increase my awareness regarding correlation in a more organic context such as the one of book publishing. While based in the SL Math curriculum, it gives me valuable insight into the decisions I could be facing about book length soon as an aspiring author wishing to be published. Since I am aware of this correlation, length will be a factor I'll consider.

WORKS CITED

"Chapter 21: Linear Modelling." *Mathematics for the International Student: Mathematics SL*, by Robert Haese et al., Third Edition ed., Haese & Harris Publications, 2012, pp. 546–556.

Clement, J. "Goodreads: Number of Registered Members 2019." *Statista*, 1 Oct. 2020, www.statista.com/statistics/252986/number-of-registered-members-on-goodreadscom/.

Lea, Richard. "The Big Question: Are Books Getting Longer?" *The Guardian*, Guardian News and Media, 10 Dec. 2015, www.theguardian.com/books/2015/dec/10/are-books-getting-longer-survey-marlon-james-hanya-yanagihara.

APPENDIX

Appendix 1— All data points

Title	Page Length (x)	Avg. Rating (y)	Title	Page Length (x)	Avg. Rating (y)	Title	Page Length (x)	Avg. Rating (y)	Title	Page Length (x)	Avg. Rating (y)
Paper	21	3.18	The Austere Academy	244	3.98	Throne of Glass	404	4.20	Zom-B Underground	212	3.88
Finding West	24	3.24	Stay Where You Are…	247	3.98	Gathering Darkness	407	4.15	Zom-B City	213	3.88
The Prince of Beers	31	3.23	Macbeth	249	3.91	Ruin and Rising	422	4.08	How to Train Your D…	214	3.88
What A Girl Wants	40	3.32	Fahrenheit 451	249	3.99	Hollow City	428	4.04	Fairest	222	3.97
The Grownup	64	3.57	The Gunslinger…	253	3.94	End of Days	435	4.07	The Self-Care Revol…	224	3.91
23 Anti-Procrastinat…	91	3.64	The Truth About Magic	256	3.98	Restore Me	435	4.18	Diary of a Wire	226	3.98
The Worrier's Guide…	112	3.63	If I Stay (If I Stay, #1)	261	3.94	City of Bones	442	4.09	Holes	233	3.97
Caffeine	120	3.71	The Storied Life of A.J.	273	3.99	A Torch Against the N..	452	4.19	Big Rock	235	3.94
The House on Mango…	132	3.67	The Wild Robot	277	4.00	City of Ashes	453	4.16	One of Us Is Lying	358	4.05
The Story of Tracy B…	144	3.74	The Mysterious Aff…	288	3.98	Dread Nation	455	4.14	The Silence of Six	368	4.02
Colder, Vol. 1	152	3.82	Carrie	290	3.97	Unravel Me	461	4.22	Anna and the French	372	4.00
Sinful Cinderella	158	3.78	Writer to Writer	304	3.99	Library of Souls	463	4.12	The Last Time We Say	386	4.11
My Teacher Is an Alien	160	3.79	Everything, Everything	306	4.05	A Reaper at the Gates	463	4.23	Steelheart	386	4.13
The River	160	3.81	Note to Self	320	4.07	Divergent	487	4.19	Cinder	387	4.13
The Nixie's Song	162	3.74	History Is All You Le…	320	4.04	The Darkest Minds	488	4.17	Stars Above	400	4.19
Granny's Got a Gun	165	3.69	Return to the Isle of…	320	3.99	The Son of Neptune	513	4.24	We Are Okay	236	3.94
The Miracle Morning	172	3.81	Sharp Objects	321	3.98	City of Glass	541	4.27	Illuminae	602	4.24
Every Heart a Doorway	173	3.83	Eleanor & Park	328	4.04	Cress	550	4.31	Gemina	659	4.26
Hatchet	186	3.74	The Elite	336	4.01	The Lost Hero	553	4.32	Obsidio	672	4.37
Stargirl	186	3.76	Looking for Alaska	350	4.01	Heir of Fire	562	4.26	A Court of Wings an…	720	4.34
The Magician's Eleph..	201	3.83	Queer	351	4.02	The Mark of Athena	586	4.35	Winter *	824	4.43
The Alchemist	208	3.88	Warcross	353	4.15	The Book Thief	592	4.28	A Game of Thrones *	835	4.45
Harry Potter & Order..*	870	4.45	It *	1116	4.24						

Appendix 2: Pearson's Coefficient Calculations

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
720	4.34	402.791	0.365	146.878	162240.346	0.133	304	3.99	-13.209	0.015	-0.194	174.486	0.000
672	4.37	354.791	0.395	140.019	125876.439	0.156	290	3.97	-27.209	-0.005	0.146	740.346	0.000
659	4.26	341.791	0.285	97.291	116820.881	0.081	288	3.98	-29.209	0.005	-0.136	853.183	0.000
602	4.24	284.791	0.265	75.370	81105.741	0.070	277	4.00	-40.209	0.025	-0.991	1616.788	0.001
592	4.28	274.791	0.305	83.715	75509.928	0.093	273	3.99	-44.209	0.015	-0.648	1954.462	0.000
586	4.35	268.791	0.375	100.703	72248.439	0.140	261	3.94	-56.209	-0.035	1.987	3159.486	0.001
562	4.26	244.791	0.285	69.680	59922.486	0.081	256	3.98	-61.209	0.005	-0.285	3746.579	0.000
553	4.32	235.791	0.345	81.266	55597.253	0.119	253	3.94	-64.209	-0.035	2.270	4122.835	0.001
550	4.31	232.791	0.335	77.904	54191.509	0.112	249	3.91	-68.209	-0.065	4.457	4652.509	0.004

541	4.27	223.791	0.295	65.940	50082.276	0.087	249	3.99	-68.209	0.015	-0.999	4652.509	0.000
513	4.24	195.791	0.265	51.816	38333.997	0.070	247	3.98	-70.209	0.005	-0.327	4929.346	0.000
488	4.17	170.791	0.195	33.245	29169.462	0.038	244	3.98	-73.209	0.005	-0.341	5359.602	0.000
487	4.19	169.791	0.215	36.446	28828.881	0.046	236	3.95	-81.209	-0.025	2.059	6594.951	0.001
463	4.12	145.791	0.145	21.089	21254.928	0.021	235	3.94	-82.209	-0.035	2.906	6758.369	0.001
463	4.23	145.791	0.255	37.126	21254.928	0.065	233	3.97	-84.209	-0.005	0.450	7091.207	0.000
461	4.22	143.791	0.245	35.179	20675.765	0.060	226	3.98	-91.209	0.005	-0.424	8319.137	0.000
455	4.14	137.791	0.165	22.687	18986.276	0.027	224	3.91	-93.209	-0.065	6.091	8687.974	0.004
453	4.16	135.791	0.185	25.074	18439.114	0.034	222	3.97	-95.209	-0.005	0.509	9064.811	0.000
452	4.19	134.791	0.215	28.933	18168.532	0.046	214	3.88	-103.209	-0.095	9.841	10652.160	0.009
442	4.09	124.791	0.115	14.307	15572.718	0.013	213	3.88	-104.209	-0.095	9.936	10859.579	0.009
435	4.07	117.791	0.095	11.149	13874.648	0.009	212	3.88	-105.209	-0.095	10.032	11068.997	0.009
435	4.18	117.791	0.205	24.106	13874.648	0.042	208	3.88	-109.209	-0.095	10.413	11926.672	0.009
428	4.04	110.791	0.065	7.163	12274.579	0.004	201	3.83	-116.209	-0.145	16.891	13504.602	0.021
422	4.08	104.791	0.105	10.966	10981.090	0.011	186	3.74	-131.209	-0.235	30.880	17215.881	0.055
407	4.15	89.791	0.175	15.682	8062.369	0.031	186	3.76	-131.209	-0.215	28.256	17215.881	0.046
404	4.20	86.791	0.225	19.498	7532.625	0.050	173	3.83	-144.209	-0.145	20.961	20796.323	0.021
400	4.19	82.791	0.215	17.771	6854.300	0.046	172	3.81	-145.209	-0.165	24.010	21085.741	0.027
387	4.13	69.791	0.155	10.793	4870.741	0.024	165	3.69	-152.209	-0.285	43.433	23167.672	0.081
386	4.11	68.791	0.135	9.263	4732.160	0.018	162	3.74	-155.209	-0.235	36.528	24089.928	0.055
386	4.13	68.791	0.155	10.639	4732.160	0.024	160	3.79	-157.209	-0.185	29.139	24714.765	0.034
372	4.00	54.791	0.025	1.351	3002.021	0.001	160	3.81	-157.209	-0.165	25.994	24714.765	0.027
368	4.02	50.791	0.045	2.268	2579.695	0.002	158	3.78	-159.209	-0.195	31.101	25347.602	0.038
358	4.05	40.791	0.075	3.045	1663.881	0.006	152	3.82	-165.209	-0.155	25.665	27294.114	0.024
353	4.15	35.791	0.175	6.251	1280.974	0.031	144	3.74	-173.209	-0.235	40.765	30001.462	0.055
351	4.02	33.791	0.045	1.509	1141.811	0.002	132	3.67	-185.209	-0.305	56.553	34302.486	0.093
350	4.01	32.791	0.035	1.136	1075.230	0.001	120	3.71	-197.209	-0.265	52.329	38891.509	0.070
336	4.01	18.791	0.035	0.651	353.090	0.001	112	3.63	-205.209	-0.345	70.869	42110.858	0.119
328	4.04	10.791	0.065	0.698	116.439	0.004	91	3.64	-226.209	-0.335	75.859	51170.648	0.112
322	4.04	4.791	0.065	0.310	22.951	0.004	64	3.57	-253.209	-0.405	102.638	64114.951	0.164
322	3.99	4.791	0.015	0.070	22.951	0.000	40	3.32	-277.209	-0.655	181.669	76844.997	0.429
322	3.98	4.791	0.005	0.022	22.951	0.000	31	3.23	-286.209	-0.745	213.326	81915.765	0.556
320	4.07	2.791	0.095	0.264	7.788	0.009	24	3.24	-293.209	-0.735	215.611	85971.695	0.541
306	4.05	-11.209	0.075	-0.837	125.648	0.006	21	3.18	-296.209	-0.795	235.590	87739.951	0.633

Appendix 3: Percentage Error Calculations for Equation 1

***Pages* (*x*)**	**Actual Ratings (v_A)**	**Eq. 1 Ratings (v_C)**	**% Error (δ)**	***Pages* (*x*)**	**Actual Ratings (v_A)**	**Equation 1 Ratings (v_C)**	**% Error (δ)**	***Pages* (*x*)**	**Actual Ratings (v_A)**	**Equation 1 Ratings (v_C)**	**% Error (δ)**	***Pages* (*x*)**	**Actual Ratings (v_A)**	**Equation 1 Ratings (v_C)**	**% Error (δ)**
720	4.34	4.35	0.24%	304	3.99	3.88	2.86%	372	4.00	3.95	1.17%	407	4.15	3.99	3.78%
672	4.37	4.30	1.70%	290	3.97	3.86	2.78%	368	4.02	3.95	1.77%	404	4.20	3.99	5.00%
659	4.26	4.28	0.49%	288	3.98	3.86	3.08%	358	4.05	3.94	2.78%	400	4.19	3.99	4.89%
602	4.24	4.22	0.57%	277	4.00	3.84	3.88%	353	4.15	3.93	5.26%	387	4.13	3.97	3.86%
592	4.28	4.20	1.77%	273	3.99	3.84	3.75%	351	4.02	3.93	2.25%	386	4.11	3.97	3.42%
586	4.35	4.20	3.51%	261	3.94	3.83	2.88%	350	4.01	3.93	2.04%	186	3.76	3.74	0.50%
562	4.26	4.17	2.11%	256	3.98	3.82	3.99%	336	4.01	3.91	2.44%	173	3.83	3.73	2.71%
553	4.32	4.16	3.71%	253	3.94	3.82	3.11%	328	4.04	3.90	3.39%	172	3.81	3.73	2.23%
550	4.31	4.16	3.56%	249	3.91	3.81	2.48%	322	4.04	3.90	3.56%	165	3.69	3.72	0.74%

541	4.27	4.15	2.90%	249	3.99	3.81	4.44%	322	3.99	3.90	2.35%	162	3.74	3.71	0.70%
513	4.24	4.11	2.97%	247	3.98	3.81	4.25%	322	3.98	3.90	2.10%	24	3.24	3.56	9.76%
488	4.17	4.09	2.02%	244	3.98	3.81	4.34%	320	4.07	3.89	4.32%	21	3.18	3.55	11.73%
487	4.19	4.08	2.52%	236	3.95	3.80	3.84%	306	4.05	3.88	4.25%	386	4.13	3.97	3.89%
463	4.12	4.06	1.53%	235	3.94	3.80	3.63%	160	3.81	3.71	2.58%	160	3.79	3.71	2.07%
463	4.23	4.06	4.09%	233	3.97	3.79	4.41%	158	3.78	3.71	1.87%	435	4.18	4.03	3.70%
461	4.22	4.05	3.91%	226	3.98	3.79	4.85%	152	3.82	3.70	3.08%	428	4.04	4.02	0.56%
455	4.14	4.05	2.22%	224	3.91	3.78	3.21%	144	3.74	3.69	1.25%	422	4.08	4.01	1.71%
453	4.16	4.05	2.75%	222	3.97	3.78	4.73%	132	3.67	3.68	0.26%	208	3.88	3.77	2.93%
452	4.19	4.04	3.47%	214	3.88	3.77	2.75%	120	3.71	3.67	1.19%	201	3.83	3.76	1.87%
442	4.09	4.03	1.39%	213	3.88	3.77	2.78%	112	3.63	3.66	0.74%	186	3.74	3.74	0.03%
435	4.07	4.03	1.10%	212	3.88	3.77	2.81%	91	3.64	3.63	0.20%	31	3.23	3.56	10.35%
64	3.57	3.60	0.90%	40	3.32	3.57	7.67%								

Appendix 4: Calculations Needed for Finding k

x	y	$w = \ln(x)$	$z = \ln(y)$	$w \times z$	w^2	x	y	$w = \ln(x)$	$z = \ln(y)$	$w \times z$	w^2
720	4.34	6.57925	1.46787	9.65751	43.28655	304	3.99	5.71703	1.38379	7.91117	32.68441
672	4.37	6.51026	1.47476	9.60109	42.38346	290	3.97	5.66988	1.37877	7.81744	32.14755
659	4.26	6.49072	1.44927	9.40681	42.12949	288	3.98	5.66296	1.38128	7.82214	32.06912
602	4.24	6.40026	1.44456	9.24558	40.96330	277	4.00	5.62402	1.38629	7.79654	31.62957
592	4.28	6.38351	1.45395	9.28132	40.74916	273	3.99	5.60947	1.38379	7.76234	31.46617
586	4.35	6.37332	1.47018	9.36990	40.61921	261	3.94	5.56452	1.37118	7.62996	30.96389
562	4.26	6.33150	1.44927	9.17605	40.08792	256	3.98	5.54518	1.38128	7.65945	30.74899
553	4.32	6.31536	1.46326	9.24098	39.88375	253	3.94	5.53339	1.37118	7.58728	30.61840
550	4.31	6.30992	1.46094	9.21840	39.81507	249	3.91	5.51745	1.36354	7.52325	30.44229
541	4.27	6.29342	1.45161	9.13561	39.60713	249	3.99	5.51745	1.38379	7.63500	30.44229
513	4.24	6.24028	1.44456	9.01447	38.94104	247	3.98	5.50939	1.38128	7.61002	30.35336
488	4.17	6.19032	1.42792	8.83925	38.32000	244	3.98	5.49717	1.38128	7.59314	30.21886
487	4.19	6.18826	1.43270	8.86593	38.29461	236	3.95	5.46383	1.37372	7.50575	29.85346
463	4.12	6.13773	1.41585	8.69012	37.67169	235	3.94	5.45959	1.37118	7.48608	29.80707
463	4.23	6.13773	1.44220	8.85184	37.67169	233	3.97	5.45104	1.37877	7.51571	29.71382
461	4.22	6.13340	1.43984	8.83108	37.61857	226	3.98	5.42053	1.38128	7.48729	29.38220
455	4.14	6.12030	1.42070	8.69508	37.45804	224	3.91	5.41165	1.36354	7.37898	29.28591
453	4.16	6.11589	1.42552	8.71830	37.40414	222	3.97	5.40268	1.37877	7.44903	29.18892
452	4.19	6.11368	1.43270	8.75908	37.37711	214	3.88	5.36598	1.35584	7.27538	28.79370
442	4.09	6.09131	1.40854	8.57988	37.10406	213	3.88	5.36129	1.35584	7.26903	28.74345
435	4.07	6.07535	1.40364	8.52762	36.90983	212	3.88	5.35659	1.35584	7.26265	28.69302
435	4.18	6.07535	1.43031	8.68964	36.90983	208	3.88	5.33754	1.35584	7.23682	28.48931
428	4.04	6.05912	1.39624	8.46002	36.71297	201	3.83	5.30330	1.34286	7.12162	28.12504
422	4.08	6.04501	1.40610	8.49986	36.54209	186	3.74	5.22575	1.31909	6.89321	27.30843
407	4.15	6.00881	1.42311	8.55119	36.10584	186	3.76	5.22575	1.32442	6.92108	27.30843
404	4.20	6.00141	1.43508	8.61254	36.01698	173	3.83	5.15329	1.34286	6.92017	26.55641
400	4.19	5.99146	1.43270	8.58398	35.89765	172	3.81	5.14749	1.33763	6.88544	26.49670
387	4.13	5.95842	1.41828	8.45070	35.50282	165	3.69	5.10595	1.30563	6.66646	26.07068
386	4.11	5.95584	1.41342	8.41812	35.47200	162	3.74	5.08760	1.31909	6.71098	25.88364
386	4.13	5.95584	1.41828	8.44703	35.47200	160	3.79	5.07517	1.33237	6.76199	25.75739
372	4.00	5.91889	1.38629	8.20533	35.03330	160	3.81	5.07517	1.33763	6.78870	25.75739

368	4.02	5.90808	1.39128	8.21981	34.90544	158	3.78	5.06260	1.32972	6.73185	25.62987
358	4.05	5.88053	1.39872	8.22520	34.58067	152	3.82	5.02388	1.34025	6.73326	25.23938
353	4.15	5.86647	1.42311	8.34862	34.41545	144	3.74	4.96981	1.31909	6.55561	24.69904
351	4.02	5.86079	1.39128	8.15401	34.34882	132	3.67	4.88280	1.30019	6.34858	23.84175
350	4.01	5.85793	1.38879	8.13545	34.31538	120	3.71	4.78749	1.31103	6.27655	22.92008
336	4.01	5.81711	1.38879	8.07875	33.83878	112	3.63	4.71850	1.28923	6.08324	22.26423
328	4.04	5.79301	1.39624	8.08846	33.55901	91	3.64	4.51086	1.29198	5.82796	20.34785
322	4.04	5.77455	1.39624	8.06269	33.34545	64	3.57	4.15888	1.27257	5.29245	17.29631
322	3.99	5.77455	1.38379	7.99077	33.34545	40	3.32	3.68888	1.19996	4.42653	13.60783
322	3.98	5.77455	1.38128	7.97628	33.34545	31	3.23	3.43399	1.17248	4.02629	11.79227
320	4.07	5.76832	1.40364	8.09666	33.27353	24	3.24	3.17805	1.17557	3.73604	10.10003
306	4.05	5.72359	1.39872	8.00568	32.75943	21	3.18	3.04452	1.15688	3.52215	9.26912

Appendix 5: Percentage Error Calculations for Equation 2

***Pages* (*x*)**	**Equation 2 Ratings (v_C)**	**Actual Ratings (v_A)**	**Percentage Error (δ)**	***Pages* (*x*)**	**Equation 2 Ratings (v_C)**	**Actual Ratings (v_A)**	**Percentage Error (δ)**	***Pages* (*x*)**	**Equation 2 Ratings (v_C)**	**Actual Ratings (v_A)**	**Percentage Error (δ)**	***Pages* (*x*)**	**Equation 2 Ratings (v_C)**	**Actual Ratings (v_A)**	**Percentage Error (δ)**
629	4.21	4.61	9.52%	372	4.13	4.00	3.23%	383	4.14	4.32	4.40%	236	4.07	3.95	2.97%
576	4.20	4.60	9.62%	418	4.15	4.00	3.62%	435	4.16	4.32	3.95%	429	4.15	3.95	4.91%
328	4.12	4.57	11.03%	218	4.06	3.99	1.73%	553	4.19	4.32	3.09%	432	4.15	3.95	4.93%
652	4.21	4.57	8.43%	249	4.08	3.99	2.16%	479	4.17	4.31	3.36%	235	4.07	3.94	3.20%
734	4.23	4.56	7.74%	273	4.09	3.99	2.46%	541	4.19	4.31	2.93%	253	4.08	3.94	3.44%
422	4.15	4.50	8.39%	304	4.11	3.99	2.81%	391	4.14	4.29	3.60%	261	4.08	3.94	3.54%
624	4.21	4.50	6.93%	320	4.11	3.99	2.98%	376	4.14	4.28	3.50%	286	4.10	3.94	3.83%
870	4.26	4.50	5.69%	320	4.11	3.99	2.98%	411	4.15	4.27	2.95%	338	4.12	3.94	4.37%
309	4.11	4.48	9.06%	327	4.12	3.99	3.05%	446	4.16	4.27	2.66%	131	3.99	3.93	1.58%
385	4.14	4.48	8.25%	358	4.13	3.99	3.35%	452	4.16	4.27	2.61%	323	4.11	3.93	4.47%
385	4.14	4.46	7.77%	176	4.03	3.98	1.28%	602	4.20	4.27	1.59%	324	4.11	3.93	4.48%
504	4.18	4.46	6.77%	192	4.04	3.98	1.56%	247	4.08	4.26	4.49%	400	4.14	3.93	5.16%
562	4.19	4.46	6.37%	226	4.06	3.98	2.09%	387	4.14	4.26	2.92%	320	4.11	3.92	4.68%
659	4.22	4.46	5.78%	244	4.08	3.98	2.34%	592	4.20	4.38	4.27%	381	4.14	3.97	4.04%
672	4.22	4.46	5.71%	247	4.08	3.98	2.38%	391	4.14	4.34	4.81%	144	4.01	3.96	1.13%
586	4.20	4.45	5.98%	256	4.08	3.98	2.50%	413	4.15	4.34	4.62%	320	4.11	3.96	3.71%
835	4.25	4.45	4.66%	288	4.10	3.98	2.88%	374	4.13	4.33	4.73%	120	3.98	3.95	0.79%
720	4.23	4.44	4.98%	321	4.11	3.98	3.23%	416	4.15	4.33	4.35%	176	4.03	3.95	2.02%
824	4.25	4.43	4.24%	360	4.13	3.98	3.61%	361	4.13	4.39	6.31%	233	4.07	3.97	2.43%
513	4.18	4.42	5.75%	421	4.15	3.98	4.13%	408	4.15	4.39	5.87%	290	4.10	3.97	3.15%
550	4.19	4.42	5.50%	128	3.99	3.97	0.50%	430	4.15	4.38	5.43%	307	4.11	3.97	3.33%
418	4.15	4.40	6.02%	222	4.06	3.97	2.28%								