Comparing Novels from the 1910s and 2010s

Group 12: Preet Patel, Dragon He, Imani Odunze

2025-12-04

ABSTRACT

This project conducts an analysis about literary evolution by examining how writing styles have changed from the early 20th to early 21st century. We are interested in investigating whether novels from the 1910s have higher average word lengths than novels from the 2010s. Through this analysis, we found that the average word lengths for the 2010s was longer than that of the 1910s.

INTRODUCTION

The purpose of this analysis is to observe how writing styles have changed over time through literature. By examining the 1910s and the 2010s, through research, we aim to determine whether books from an older time period (1910s) use longer words than those from the newer period (2010s). This analysis may benefit educators, scientists, and writers who are curious about studying language trends. This research tests the hypothesis that books from the 1910s use longer words than books from the 2010s. By comparing literature texts from different decades, this will expose how literature has changed over time and uncover the educational and stylistic shifts in writing.

DATA

The data available for performing this analysis is any online library that allows us to download the file of various novels as txt files. We retrieved all data from the Project Gutenberg website which is an open access library. In total we downloaded 30 books, 15 of which are from the 1910s and the other 15 from 2010s. Variables such as author name and book title are negligible in our analysis so new variables were created such as word length, decade, average word length etc. For this analysis, it is important to acknowledge whether multiple books were written by the same author, as this may influence patterns within each decade. In the data set for the 2010s, two of the selected books were written by the same author. On the other hand, all books selected from the 1910s were written by different authors. This distinction is relevant because author-specific writing styles could affect average word length and may contribute to decade variation. Each book was manually downloaded as txt files and were put into their respective folder (by decade). From there we used the stringr and stopwords packages to clean out leading/trailing white space, remove newline characters, combine with spaces, remove extra white space, make all lowercase, and filter out stop words.

VISUALIZATION

1. Box plot

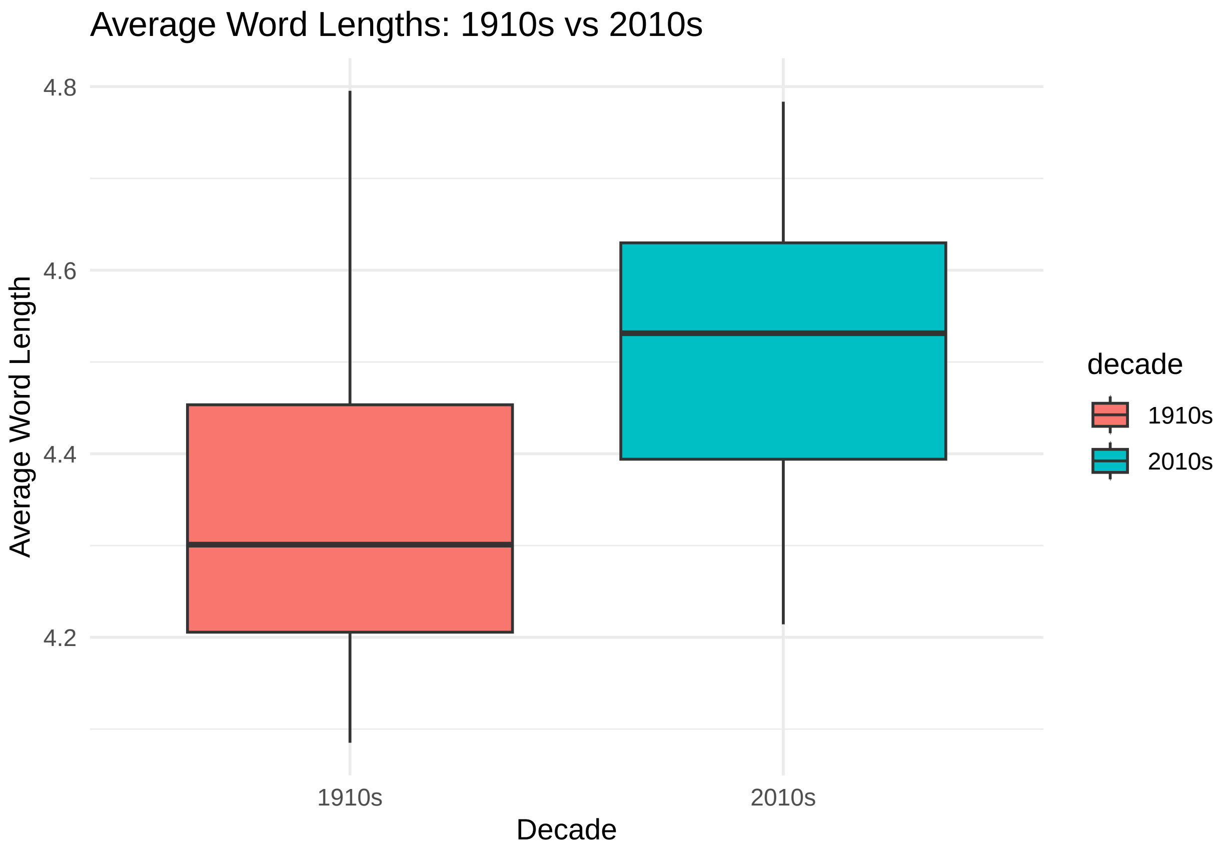

Figure 1: Box plot displaying the distribution of average word length for each decade.

2. Violin Plot

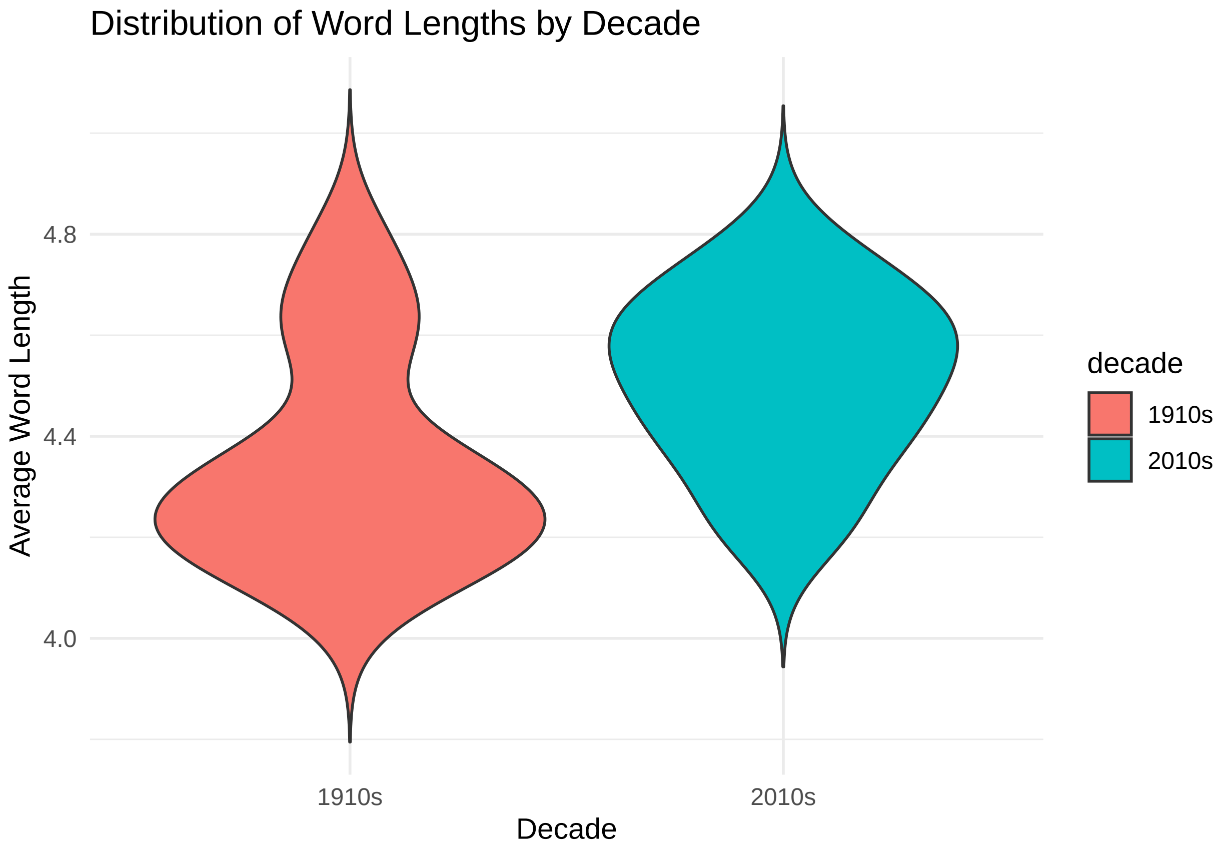

Figure 2: Violin plot showing the distribution of word lengths by decade.

3. Density Plot

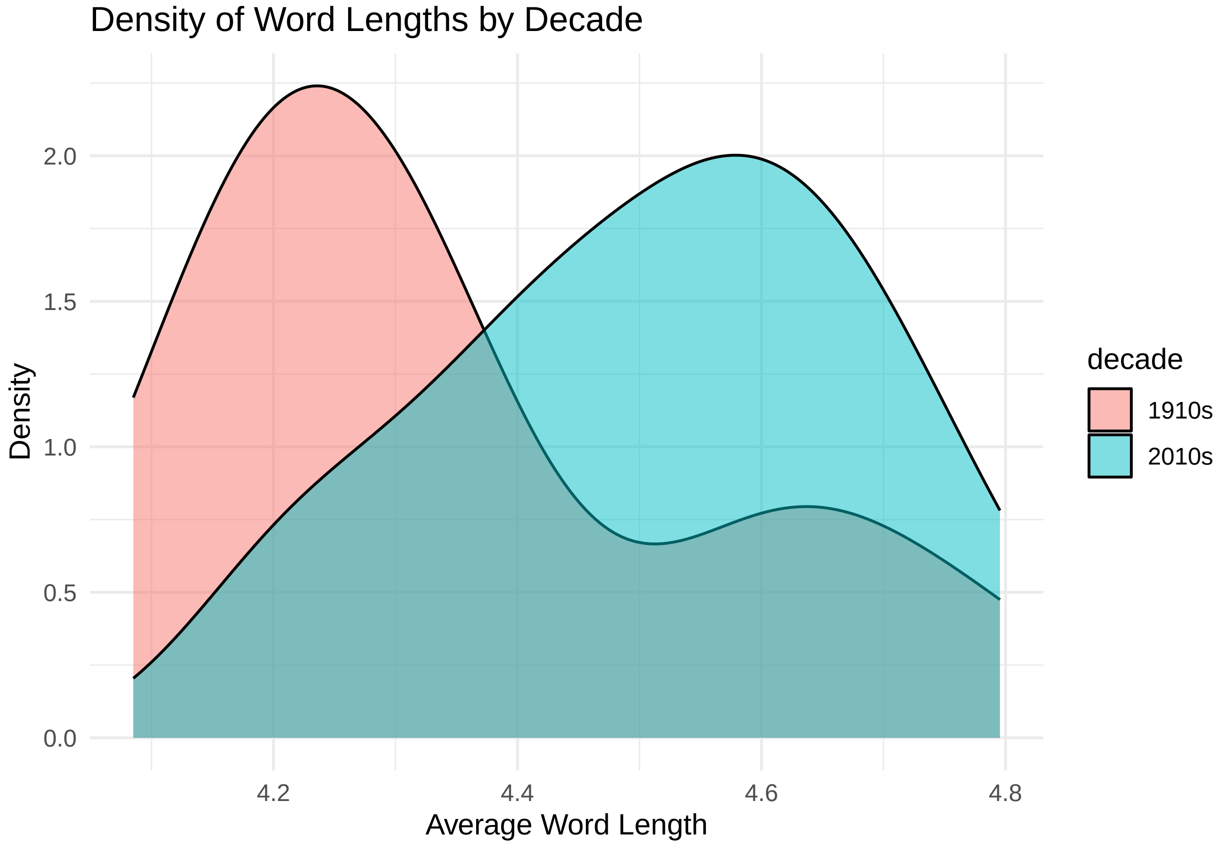

Figure 3: Density plot illustrating the density of word lengths by decade.

4. Line Plot

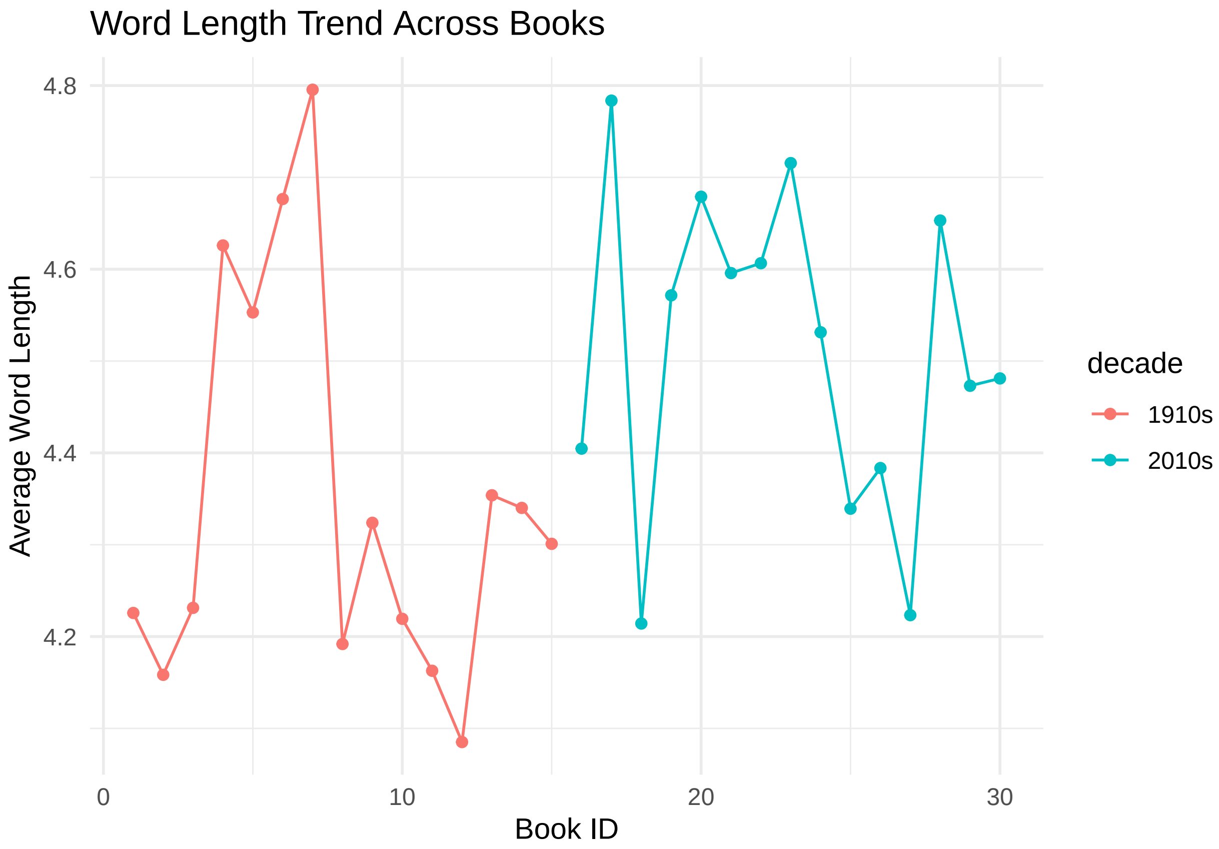

Figure 4: Line plot displaying the word length trend across books.

5. Histogram

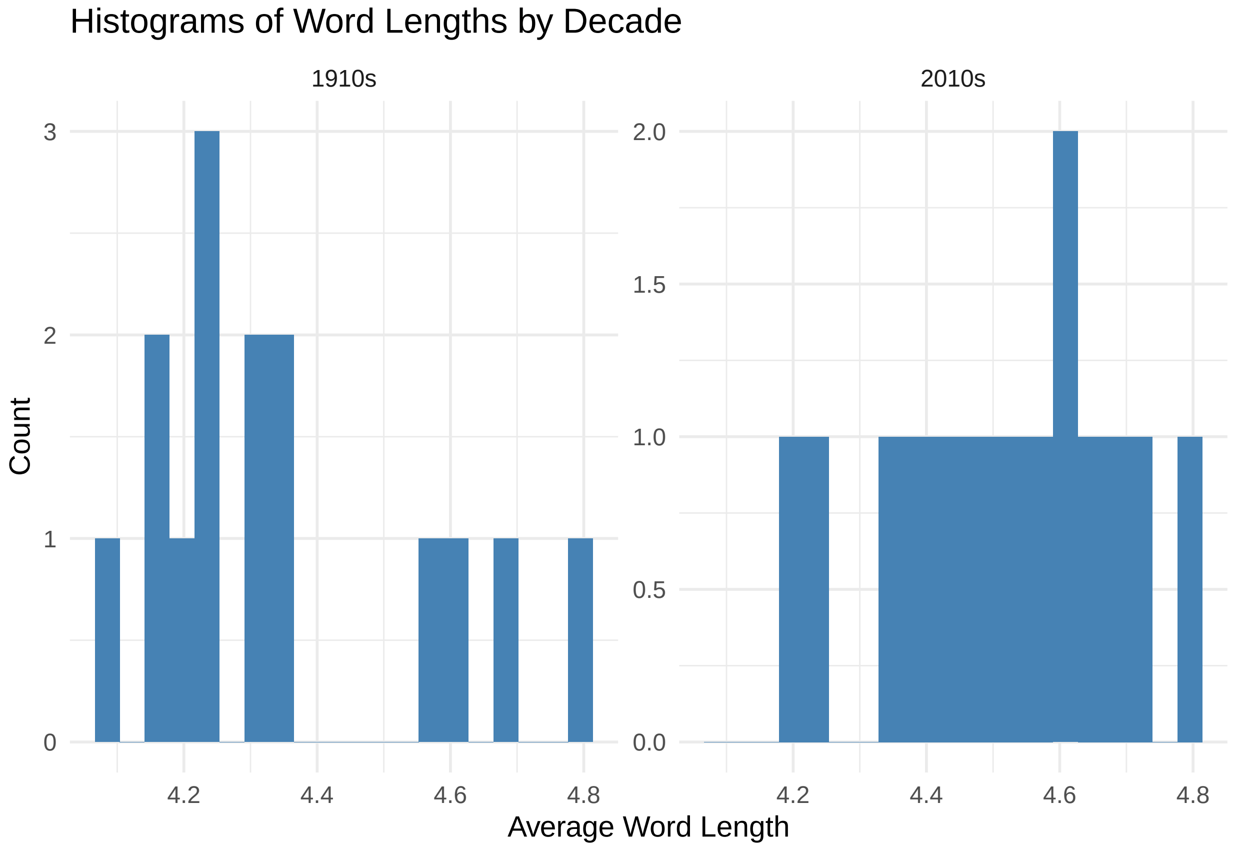

Figure 5: Histogram showing the word lengths by decade.

ANALYSIS

1. Box plot

The box plot clearly indicates that the writing styles of the 1910s and the 2010s differed significantly. In comparison to the 1910s, the entire distribution for the 2010s is shifted upward. This suggests that modern books consistently use longer words on average. The median, the interquartile range, and even the upper region all show this upward shift, indicating that longer word lengths are common throughout the dataset rather than just irregular.

In contrast, the books from the 1910s exhibit much greater variability. Older literature has a greater variety of short, medium, and long words because their box is taller and their tails of the distribution spread farther than in the 2010s. This suggests that writing styles in the early 20th century were more varied and less uniform.

On the other hand, the distribution from the 2010s seems more standardized and compact, with fewer extreme values and shorter tails of the distribution. This implies that, perhaps as a result of modern editing standards, literacy standardization, and the impact of mass-market publishing, modern writing is prone to sticking to more consistent grammar rules.

Overall, the boxplot shows that the 2010s books have a longer average word length compared to the 1910s books, where the word length varies more. Thus, it can be concluded that there is a less consistent writing style among writers and genres.

2. Violin Plot

The violin plot highlights a noticeable difference between the word length of the 1910s and the 2010s books. The thickest portion of the 2010s violin is centered around 4.6 to 4.75, whereas the 1910s violin is centered closer to 4.3 to 4.45. In general, the average word lengths of the 2010s books are longer. This demonstrates that books published in the 2010s typically use longer words.

Variability can also be inferred from the violins' shapes. The 1910s violin exhibits multiple "bulges" and is more stretched out vertically. This indicates that the word lengths from that decade vary more and include books with much longer and much shorter averages. The violin from the 2010s, on the other hand, is smaller, demonstrating that most modern books fall into a similar word-length range.

Overall, the plot indicates that while earlier writing (1910s books) is more inconsistent across books and exhibits a wider mix of shorter and longer average word lengths, modern writing (2010s books) is more consistent and uses slightly longer words on average.

3. Density Plot

The density plot does well in displaying how the distribution of average word lengths in books from the 1910s and the 2010s do, in fsct, differ. The 1910s curve is narrower and more peaked, indicating that most books are grouped around a similar word length (approximately 4.25–4.40). This suggests that the writing of the 1910s was a bit more consistent.

The density curve for the 2010s, on the other hand, is flatter and wider, ranging from roughly 4.35 to almost 4.80. This demonstrates a greater range of writing styles and more variation in contemporary word lengths. Additionally, the 2010s peak is located further to the right (between 4.65 and 4.70), indicating that modern literature (2010s) usually uses longer words.

Overall, the density plot confirms that 1910s books are more visually consistent and focused on shorter words, whereas 2010s books use longer words and show more vocabulary diversity.

4. Line Plot

The average word length for each book, broken down by decade, is displayed in the line plot. There is a lot more variation in the 1910s books (highlighted in red), with notable increases and decreases from book to book. This indicates that there was no single, consistent writing style used in the early-century books; some authors used much longer vocabulary, while others used noticeably shorter words. The pattern appears irregular and scattered, indicating greater variation in writing complexity during that decade.

The books from the 2010s, on the other hand, make up a closer band (in blue). The majority of the books cluster around similar values (roughly between 4.45 and 4.75), despite minor ups and downs in the overall range. This demonstrates the tendency for contemporary literature to be more language consistent. Today's authors appear to employ a more uniform vocabulary, which makes book differences less pronounced.

The books from the 2010s consistently have longer average word lengths than those from the 1910s in both decades. More significantly, though, the smoother trend in the 2010s indicates that, in contrast to the irregular patterns observed in the early 1900s, modern writing has become more consistent across various authors and genres.

5. Histogram

Looking at the histogram, between the two decades, the average word length differs noticeably. The books from the 1910s form a compact group with very little spread around shorter word lengths. Mainly between 4.2 and 4.4. This implies that writers in the 1910s tended to use a smaller vocabulary and maintained relatively more consistent word lengths throughout their works.

The 2010s histogram, on the other hand, spreads out more and moves noticeably to the right. The distribution covers a larger range overall, with the majority of the bars falling between 4.4 and 4.75. This suggests that writers from the 2010s often use longer words and exhibit greater vocabulary diversity. Longer average word lengths are highly prevalent in modern books, as indicated by the taller bars in that right-shifted area.

The histogram highlights that the 2010s books have a longer average word length than the 1910s books. The visual graph reveals that the 2010s books have a more diverse average word length across books, indicating that there are more flexible vocabulary patterns compared to the uniform writing style of the 1910s.

CONCLUSION

The goal of this project was to observe how writing styles have changed over the decades through literature to satisfy the curious notions of educators and scientists studying language trends. The hypothesis was stated that the word lengths in the 1910s books are longer than the 2010s books.

After evaluating the data with five different graphs (boxplot, density plot, histogram, violin plot, and a line plot), it is determined that the 2010s books have a longer average word length than the 1910s books, therefore rejecting our hypothesis. The distribution for the entire 2010s is shifted upward in the boxplot, indicating longer words overall. This pattern is supported by the density plot, where the 2010s curve peaks at a longer word length. According to the histogram, books from the 2010s are grouped around longer word-length bins, while those from the 1910s are grouped around shorter bins. The 2010s have a taller, more centered shape, which consistently represents longer word lengths, as the violin plot further demonstrates. Lastly, the line plot highlights the variations at the decade level, displaying more consistent and smooth word-length trends for contemporary books as opposed to the unpredictable patterns in the 1910s.

A second pattern can be seen in all five graphs: books from the 1910s exhibit much greater variability, whereas books from the 2010s have more uniform vocabulary. This implies that earlier writers exhibited more literary variety, alternating between writing that was simpler and more intricate. In contrast, modern books appear to have a more consistent writing style with less noticeable word length variation.

This analysis of sets of books from different decades is crucial as word length can reflect general changes in vocabulary, comprehension, education, and the development of modern writing expectations.

TEAM CONTRIBUTIONS

-Preet Patel: Created the Github repository. Sourced, downloaded, and organized 20 novels for data. Loaded all novels and wrote the code for all data cleaning. Created 00_requirement.R source, Wrote introduction, abstract, data, and visualization portions on final report. Complied final report and submitted.

-Dragon He: Created, updated, and finalized README file. Created, coded, tested, and debugged 00_requirements.R for future code, 01_getWordLengths.R to store word lengths, 02_avgWordLengths.R to average the word lengths, and 03_graphsAndVisuals.R to make graphs. Tested and debugged 00_dataLoadingAndCleaning.RMD. Tested and debugged finalProjectReport.RMD. Coded and debugged every graph. Displayed graphs in finalProjectReport for analysis. Debugged knitting issues, allowing the final project report to be knitted.

-Imani Odunze: Sourced, downloaded, and organized 10 novels. Wrote introduction, analysis, and conclusion for report.