Comparing Novels from the 1910s and 2010s

Group 12: Preet Patel, Dragon He, Imani Odunze

2025-11-07

```
source("00_requirements.R")
```

```
## Loading required package: stringr

## Loading required package: tidyverse

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.1.0
## v forcats   1.0.1     v readr     2.1.5
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Loading required package: readxl
##
## Loading required package: magrittr
##
##
## Attaching package: 'magrittr'
##
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
##
## The following object is masked from 'package:tidyr':
##
##     extract
##
##
## Loading required package: here

## Warning: package 'here' was built under R version 4.5.2

## here() starts at C:/Users/didih/OneDrive/Documents/stat107_final_project
## Loading required package: knitr
## Loading required package: stopwords
```

```
## Warning: package 'stopwords' was built under R version 4.5.2
```

```
source("01_getWordLengths.R")
source("02_avgWordLengths.R")
source("03_graphsAndVisuals.R")

books_1910s <- lapply(list.files("books_1910s", full.names = TRUE), read_file)
books_2010s <- lapply(list.files("books_2010s", full.names = TRUE), read_file)

names(books_1910s) <- basename(list.files("books_1910s"))
names(books_2010s) <- basename(list.files("books_2010s"))

avg_lengths_df <- compare_decades(books_1910s, books_2010s)
```

#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 two questions: whether novels from the 1910s exhibit longer sentences than novels from the 2010s, and whether the older books (1910s) tend to have longer words than newer ones (2010s).

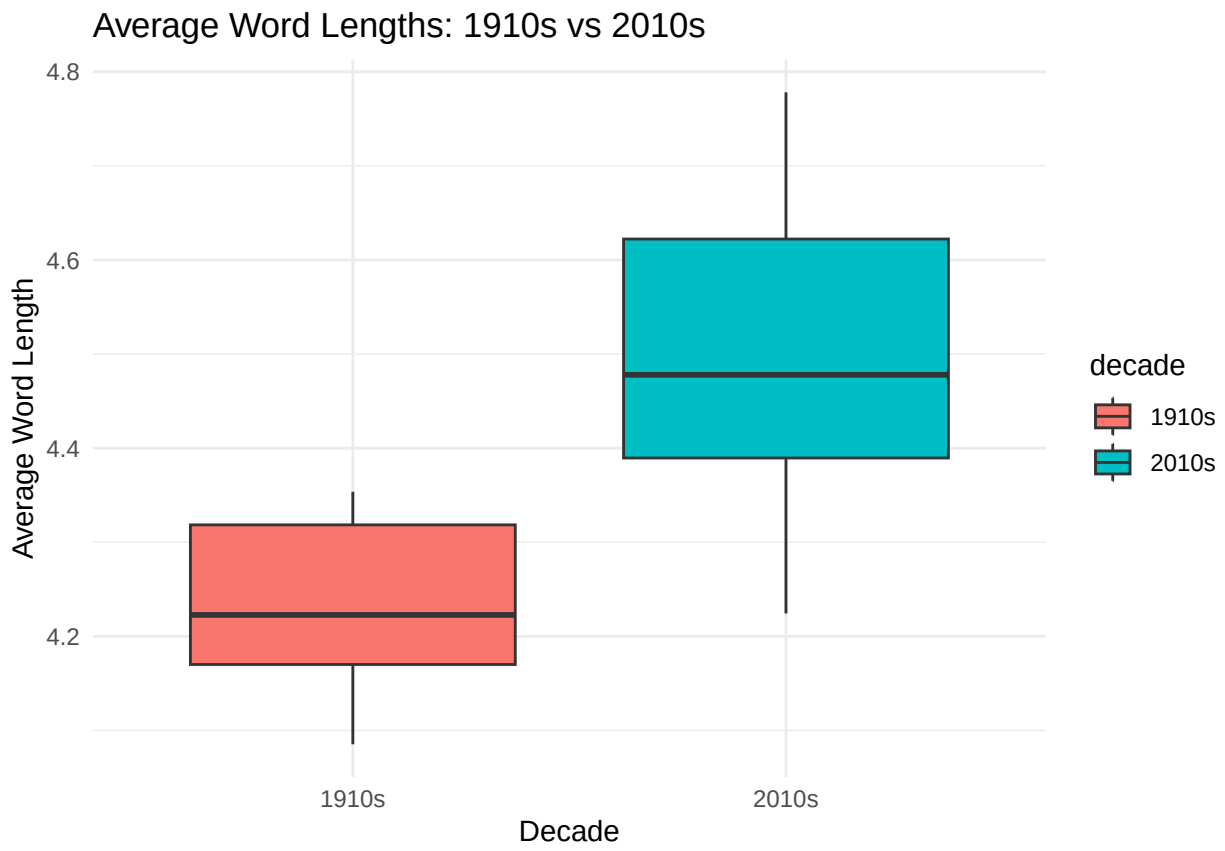
#INTRODUCTION The purpose of this analysis is to observe how writing styles have changed over time through literature. By examining the 1990s and the 2010s, through research, we aim to determine whether books from an older time period (1910s) use longer words than those from the newer period (2010s). This analysis may benefit educators and scientists who are curious about studying language trends. This research tests the hypothesis that books from the 1990s use longer words (thus having longer sentences) 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 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 20 books, 10 of which are from the 1910s and the other 10 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, etc.
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. 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 dataset 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.

#VISUALIZATION

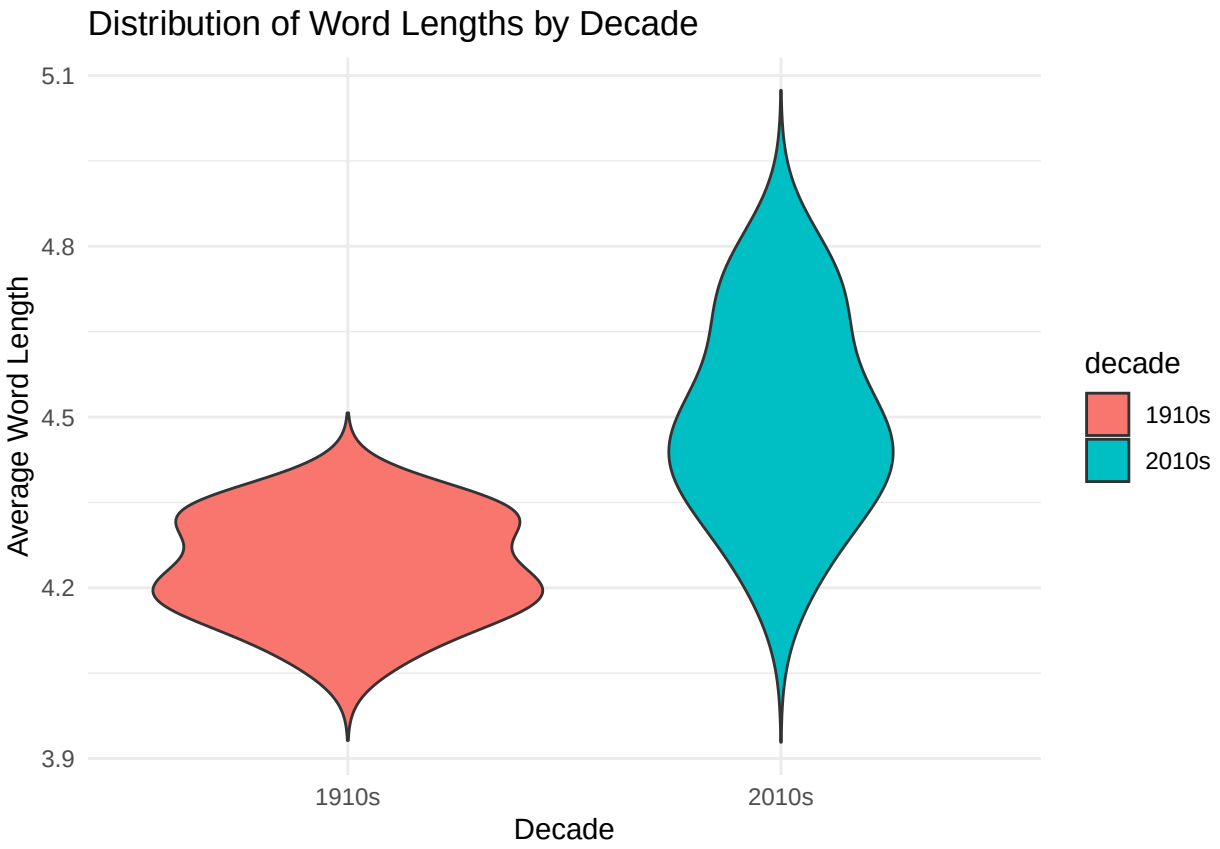
Boxplot

```
p_box <- plot_word_length_comparison(avg_lengths_df)
p_box
```



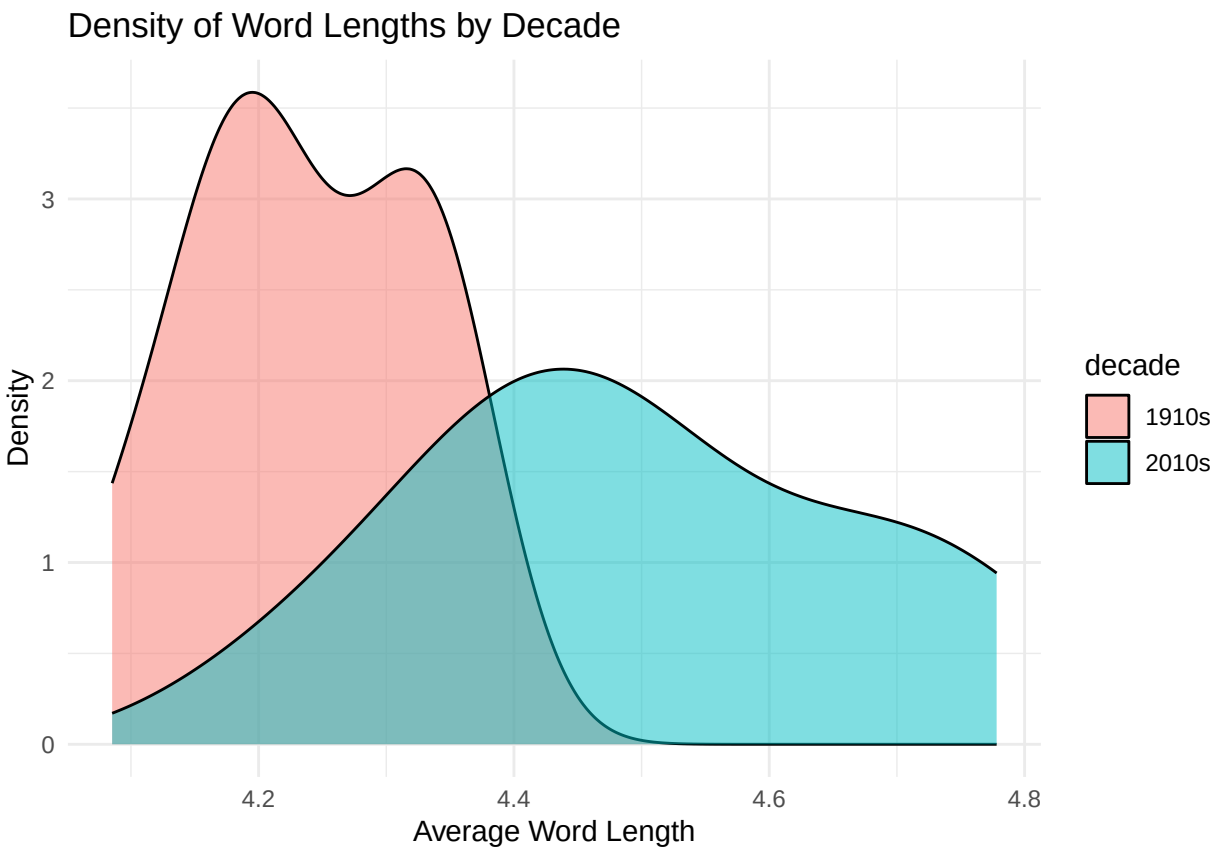
Violin Plot

```
p_violin <- plot_violin(avg_lengths_df)
p_violin
```



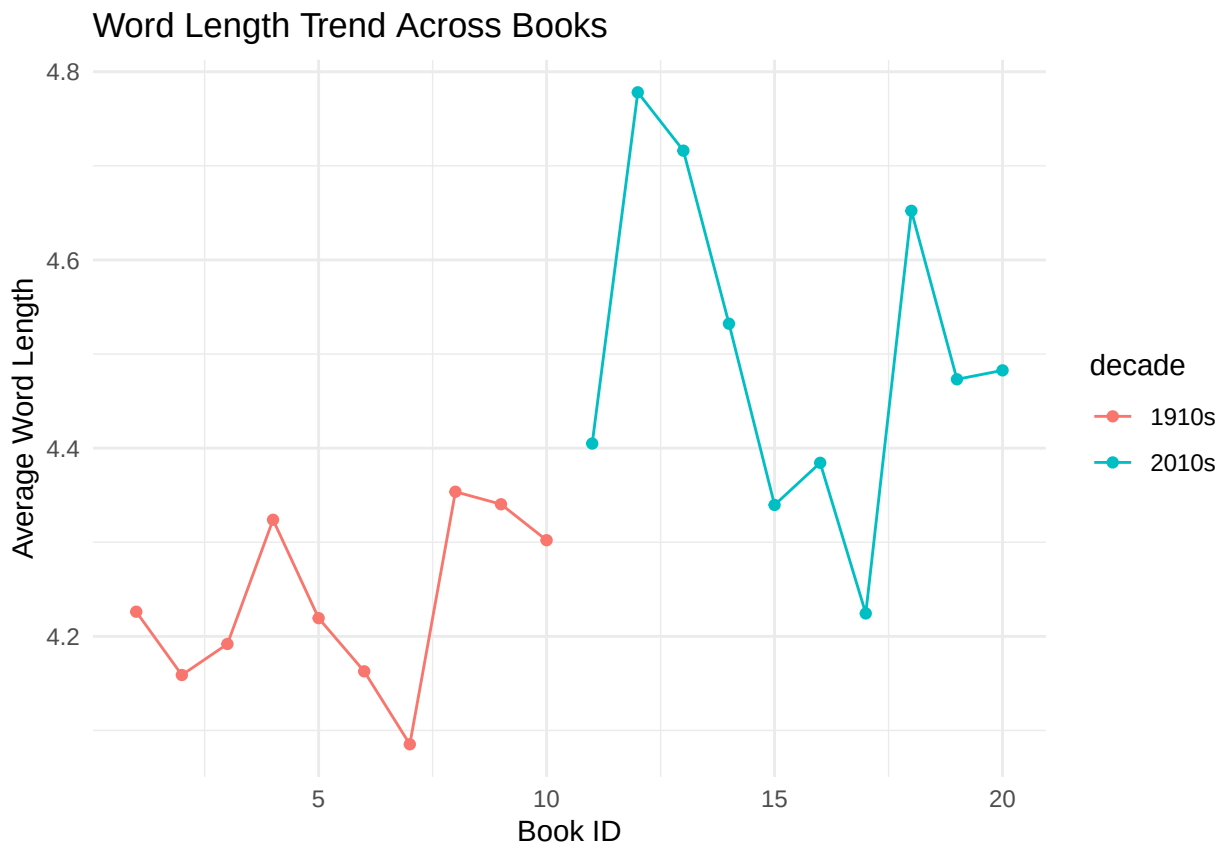
Density Plot

```
p_density <- plot_density(avg_lengths_df)
p_density
```



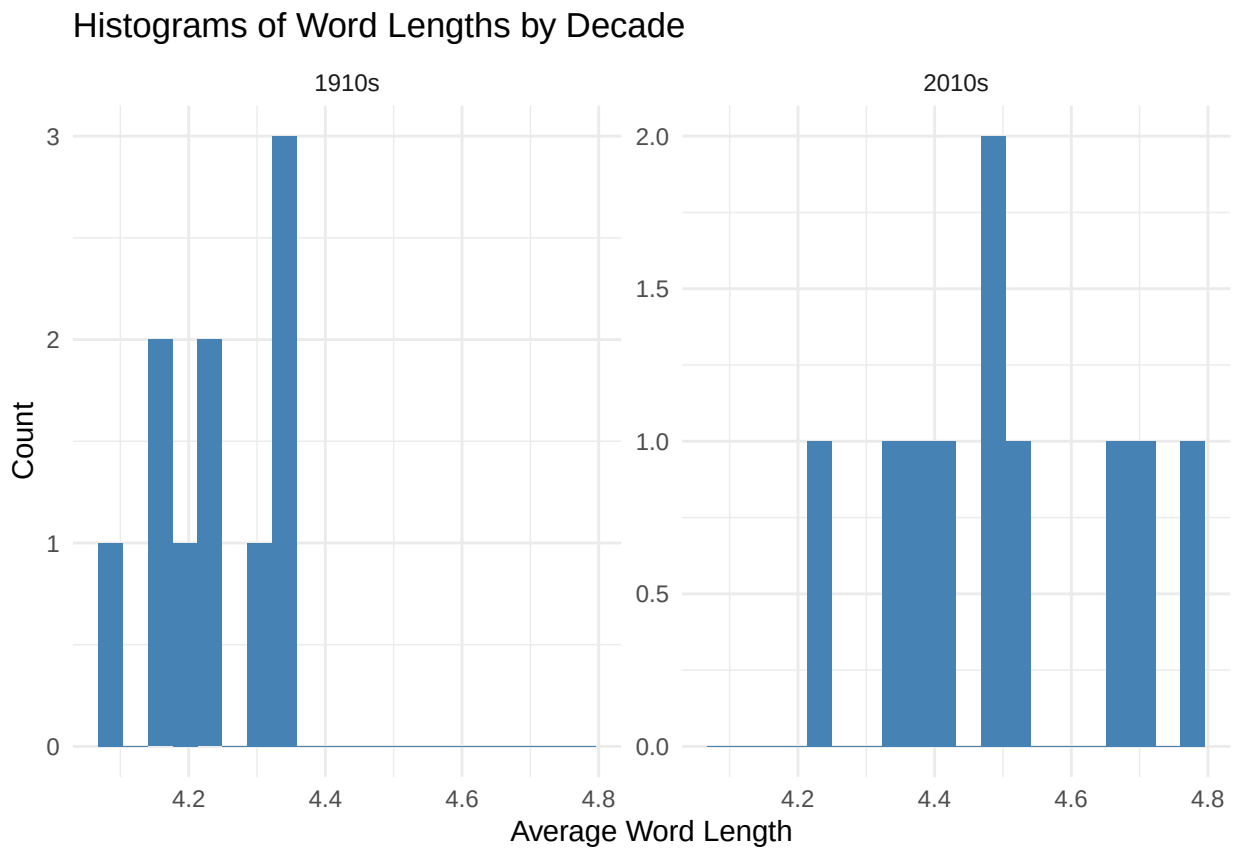
Line Plot

```
plot_line(avg_lengths_df)
```



Histogram

```
plot_histogram(avg_lengths_df)
```



#ANALYSIS For this project, we have already cleaned stop words such as “i,” “if,” and “or” and punctuation so that the focus can be on the more meaningful words. The average word length and sentence length for each ten books from either decade (the 1990s and the 2010s) will be calculated. Summary statistics such as mean and median will be used to compare texts from either decade, which will answer the question of whether books from the 1990s have longer sentences than books from the 2010s. Furthermore, to test where the differences are significant, a t-test will be used to compare the mean of the word lengths from the texts from both decades.

#TEAM CONTRIBUTIONS -Preet Patel: created Github respiratory, downloaded and organized 20 novels, loaded all novels did all data cleaning, created 00_requirement.R source, wrote abstract, data, and visualization portions on report, complied final report.

-Dragon He: created README, created R source code and functions for analysis, created plots, setup final report rmd

Imani Odunze: downloaded and organized 10 novels, wrote introduction and analysis for report.