STAT 107 Final Project—Final Report

2025-11-08

Abstract

In Japanese, there are three main writing systems: Hiragana, Katakana, and Kanji. Hiragana and Katakana are both phonetic-based (i.e., every character represents a sound), while Kanji is pictorial (i.e., each character represents a concept). While each writing system represents the same set of sounds, for individuals learning how to read in Japanese, it is often confusing to know which writing system to learn first. Should an individual's goal in learning Japanese be to read novels, it is interesting to know what writing system one would encounter the most—as to most efficiently tailor their learning experience to meet their goals.

Thus, we plan to sample 20 books from the online sources Gutenberg and Aozora to analyze the overall distribution of Hiragana, Katakana, and Kanji within them.

Introduction

The purpose of this analysis is to examine the distribution of the three main Japanese writing systems: Hiragana, Katakana, and Kanji in the novels we obtained from the online libraries Gutenberg and Aozora. Specifically we aim to determine whether the frequency of each writing system differs between texts originally written in Japanese(found in Aozora), and works translates into Japanese(found in Gutenberg).

With this analysis, we are trying to answer the question that is: How do the distributions of Hiragana, Katakana, and Kanji differ between original Japanese novels and translated Japanese novels? By addressing this question, we provide insights for Japanese language learners who want to mainly focus their studies on the writing system most relevant in the type of novels they intend to read.

This analysis may benefit translators, linguists, and language learners. As it offers factual data on writing system usage in Japanese literature. In terms of methodology, we are performing a comparative analysis bt sampling and classifying characters in each test according to their writing system. This allows us to classify and compare patterns in the writing system, providing a clear image of how different types of text utilize Hiragana, Katakana, Kanji.

Part 1: Data

Part 1.1: Japanese Literature

For this project, we will observe several novels from both the (Gutenberg)[www.gutenberg.org] website and the (Aozora)[https://www.aozora.gr.jp/ (https://www.aozora.gr.jp/).] website. Gutenberg is an English-based online library consisting of several eBooks in the public domain. Likewise, Aozora is a Japanese-based online library serving the same purpose.

We chose these sites due to the types of literature within them. From what we observed, the Japanese selection on Gutenberg mostly consists of works translated TO Japanese, while Aozora mostly consists of works written ORIGINALLY in Japanese. Thus, collecting a sample from both of the sites can provide insights as to how the distribution of each of the three writings systems differ between originally-Japanese works and works translated to Japanese.

We randomly selected 10 novels from both websites. However, this was done differently for both websites. For the noveles on the Gutenberg website, a simple random selection of all the Japanese language books on the site was conducted. However, this was not necessairly possible on Aozora—as the only way to see a masterlist of all of the novels is through sorting by authors. Thus, we ran a SRS on the authors that are not under copyright, and randomly select one of their novels (if they have multiple novels listed under their authorship).

The novels we ended up selecting (and their file names) are the following:

From Gutenberg:

File Name	Novel Name	Author	Novel Source
“pg36358.txt”	お目出たき人	Saneatsu Mushanokoji	https://www.gutenberg.org/ebooks/31757 (https://www.gutenberg.org/ebooks/31757)
“pg37605.txt”	羹	Jun’ichiro Tanizaki	https://www.gutenberg.org/ebooks/36459 (https://www.gutenberg.org/ebooks/36459)
“pg34084.txt”	幽霊書店	Christopher Morley	https://www.gutenberg.org/ebooks/41325 (https://www.gutenberg.org/ebooks/41325)
“pg37626.txt”	下宿人	Marie Belloc Lowndes	https://www.gutenberg.org/ebooks/32978 (https://www.gutenberg.org/ebooks/32978)
“pg31757.txt”	友情	Saneatsu Mushanokoji	https://www.gutenberg.org/ebooks/33307 (https://www.gutenberg.org/ebooks/33307)
“pg36459.txt”	刺青	Jun’ichiro Tanizaki	https://www.gutenberg.org/ebooks/31617 (https://www.gutenberg.org/ebooks/31617)
“pg41325.txt”	火星の記憶	Raymond F. Jones	https://www.gutenberg.org/ebooks/36358 (https://www.gutenberg.org/ebooks/36358)
“pg32978.txt”	惡魔	Jun’ichiro Tanizaki	https://www.gutenberg.org/ebooks/37605 (https://www.gutenberg.org/ebooks/37605)
“pg33307.txt”	法螺男爵旅土産	Kuni Sasaki	https://www.gutenberg.org/ebooks/34084 (https://www.gutenberg.org/ebooks/34084)
“pg31617.txt”	續惡魔	Jun’ichiro Tanizaki	https://www.gutenberg.org/ebooks/37626 (https://www.gutenberg.org/ebooks/37626)

From Aozora:

File Name	Novel Name	Author	Novel Source
“onnano_kao.txt”	人間の本性	Shin, Katakami	https://www.aozora.gr.jp/cards/000492/card42252.html (https://www.aozora.gr.jp/cards/000492/card42252.html)
“ningenno_honsho.txt”	化学改革の大略	Shimizu, Usaburo	https://www.aozora.gr.jp/cards/001508/card51403.html (https://www.aozora.gr.jp/cards/001508/card51403.html)
“kagaku_kaikakuno_tairyaku.txt”	枯尾花	Sekine, Mokuan	https://www.aozora.gr.jp/cards/001358/card49253.html (https://www.aozora.gr.jp/cards/001358/card49253.html)
“kareobana.txt”	将棋の話	Shigeru, Tonomura	https://www.aozora.gr.jp/cards/001499/card52185.html (https://www.aozora.gr.jp/cards/001499/card52185.html)
“shogino_hanashi.txt”	色盲検査表の話	Shinobu, Ishihara	https://www.aozora.gr.jp/cards/001742/card55751.html (https://www.aozora.gr.jp/cards/001742/card55751.html)
“shikimo_kensahyo.txt”	父八雲を語る	Iwao, Inagaki	https://www.aozora.gr.jp/cards/001961/card58832.html (https://www.aozora.gr.jp/cards/001961/card58832.html)
“chichi_yakumoo_kataru.txt”	おくのほそ道	Matsuo, Basho	https://www.aozora.gr.jp/cards/002240/card61619.html (https://www.aozora.gr.jp/cards/002240/card61619.html)

File Name	Novel Name	Author	Novel Source
“02okunohosomichi.txt”	青バスの女	Tatsuno, Kyushi	https://www.aozora.gr.jp/cards/001782/card56514.html (https://www.aozora.gr.jp/cards/001782/card56514.html)
“ao_basuno_onna.txt”	古事記物語	Otomo, Yasumaro	https://www.aozora.gr.jp/cards/000107/card1530.html (https://www.aozora.gr.jp/cards/000107/card1530.html)
“kojiki_monogatari.txt”	女の顔	Kuroda, Seiki	https://www.aozora.gr.jp/cards/000257/card1425.html (https://www.aozora.gr.jp/cards/000257/card1425.html)

Part 1.2: Dataframe

After cleaning the data and processing it (see “Part 2: Data Cleaning/Processing”), we end up with a single dataframe called `jp_novels_df` —of which contains the following variables:

1. `site` : factor. A variable describing which of the two sites (Gutenberg or Aozora) the novel came from.
2. `body_chars_total` : numerical. Total number of characters in the actual text of the literature (excluding the characters found in the header/footer).
3. `kanji` : numerical. Total number of characters from the body that are in Kanji.
4. `hiragana` : numerical. Total number of characters from the body that are in Hiragana.
5. `katakana` : numerical. Total number of characters from the body that are in Katakana.
6. `japanese_total` : numerical. Total number of characters from the body that are in Japanese (`kanji + hiragana + katakana`).
7. `non_japanese` : numerical. Total number of characters from the body that are NOT in Japanese.
8. `removed_chars_header_footer` : numerical. Number of characters removed from the novel file as part of the header/footer.
9. `removed_ascii_english` : numerical. Number of Latin-script characters (ie, characters in English) that was removed from the novel file.

Part 2: Data Cleaning/Processing:

The data cleaning/processing portion of this project was done on a separate file called `01_build_jp_novels.Rmd` . This file must be ran first in order to produce the `jp_novels_df` that will be used for the remainder of this analysis.

Should the dataframe be created correctly, it should look like the following:

```
library(ggplot2)

aozora <- read.csv("Data Processing/aozora_summary.csv")
gutenberg <- read.csv("Data Processing/gutenberg_summary.csv")
jp_novels_df <- rbind(aozora, gutenberg)

head(jp_novels_df, 10)
```

```
##      site                          file body_chars_total kanji hiragana
## 1  Aozora                onnano_kao.txt              725   251      459
## 2  Aozora            ningenno_honsho.txt             4787  1848     2811
## 3  Aozora kagaku_kaikakuno_tairyaku.txt             1508   441      588
## 4  Aozora                  kareobana.txt             5410  2647     3035
## 5  Aozora            shogino_hanashi.txt             3428  1385     2254
## 6  Aozora           shikimo_kensahyo.txt             4039  1643     2332
## 7  Aozora      chichi_yakumoo_kataru.txt             7420  3005     4538
## 8  Aozora           02okunohosomichi.txt            40045 21729    19154
## 9  Aozora             ao_basuno_onna.txt            10981  4686     6734
## 10 Aozora          kojiki_monogatari.txt            98365 25137    77682
##    katakana japanese_total non_japanese removed_chars_header_footer
## 1        84            794           21                         408
## 2       660           5319           27                         419
## 3       430           1459          226                         717
## 4       385           6067           42                         587
## 5       332           3971           97                         297
## 6       358           4333           37                         212
## 7       828           8371          130                         658
## 8      3375          44258         2047                         895
## 9      1060          12480          264                         528
## 10     6955         109774         2431                        1332
##    removed_ascii_english
## 1                     17
## 2                     17
## 3                     29
## 4                     32
## 5                     28
## 6                     21
## 7                     18
## 8                     32
## 9                     44
## 10                    35
```

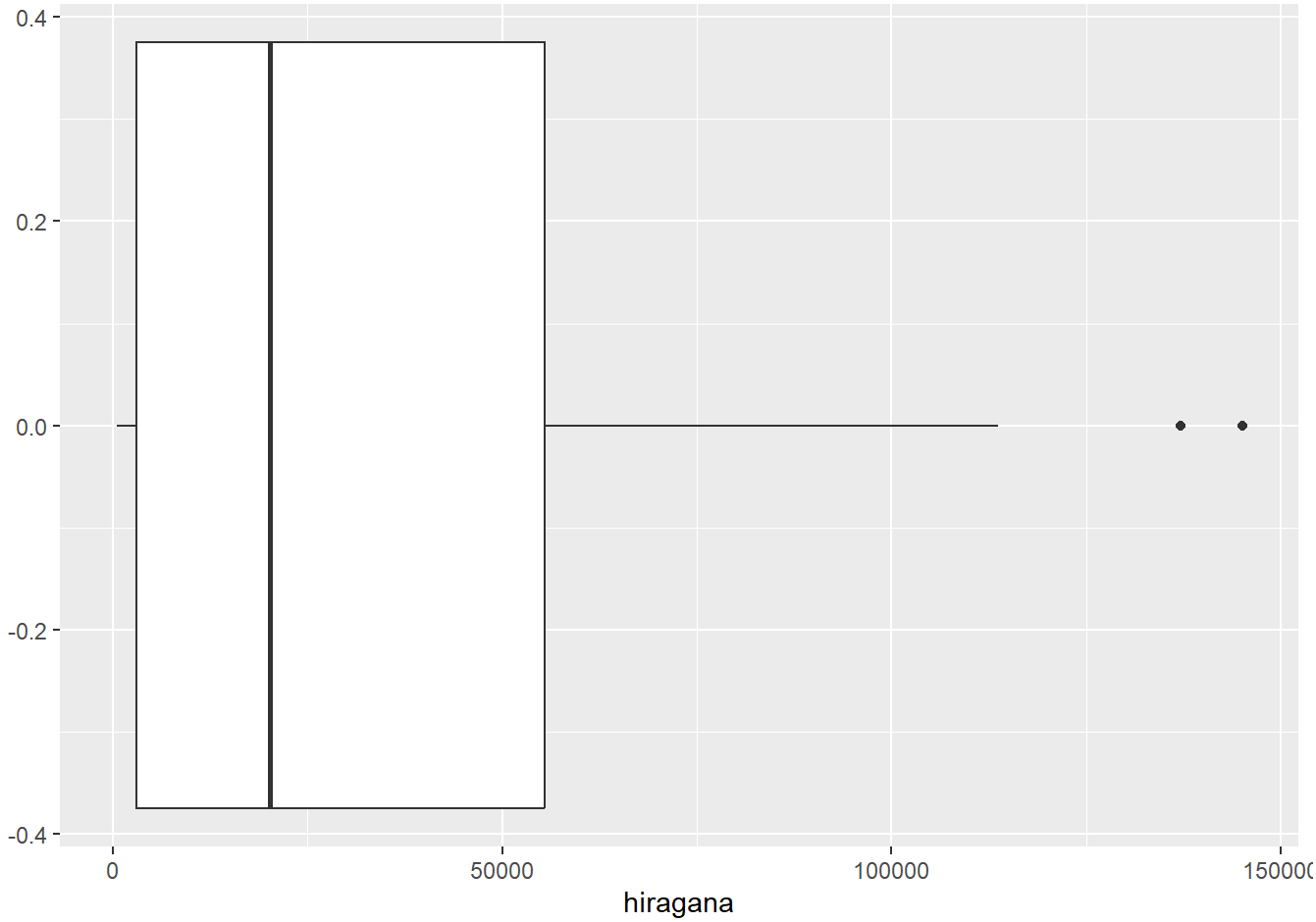
With this clean dataset, one is ready to perform analysis.

Part 3: Preliminary Analysis:

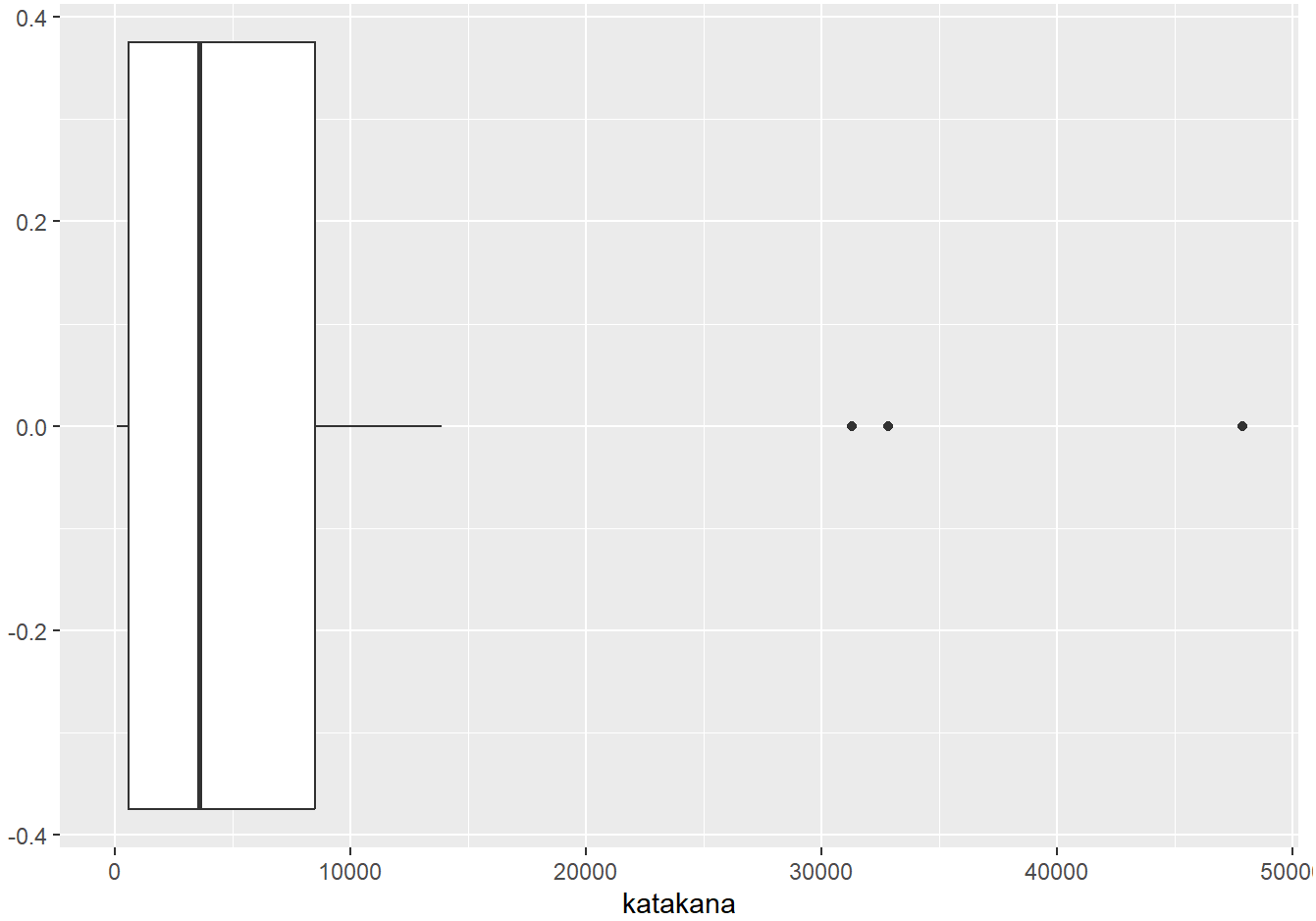
After cleaning the data and organizing it into a single dataframe, it is possible to perform a preliminary analysis of the variables.

Firstly, since they are the basis of this project, it is imperative to observe the distributions of Hiragana, Katakana, and Kanji within the sample data:

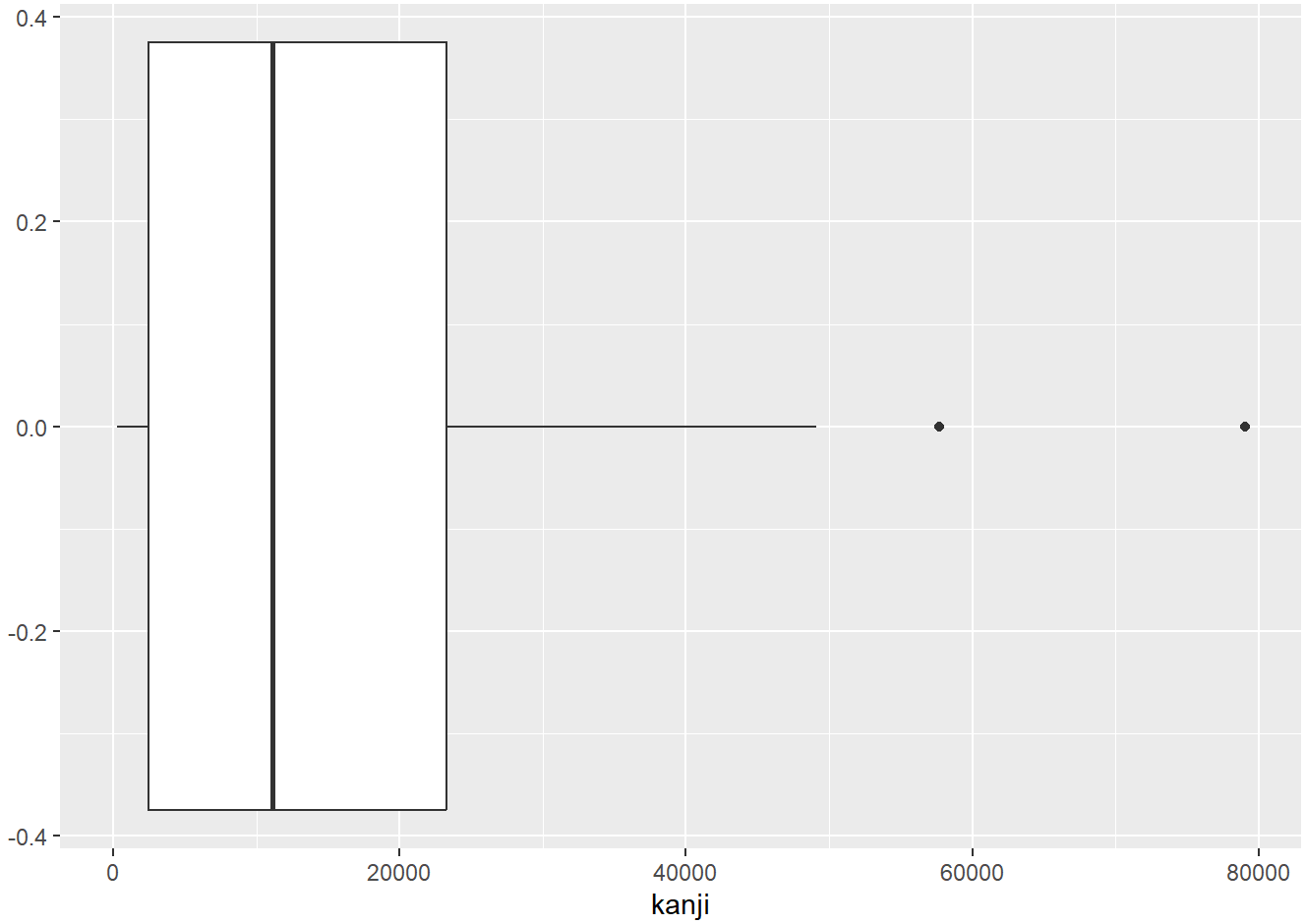
```
ggplot(jp_novels_df, aes(x = hiragana)) + geom_boxplot()
```




```
ggplot(jp_novels_df, aes(x = katakana)) + geom_boxplot()
```



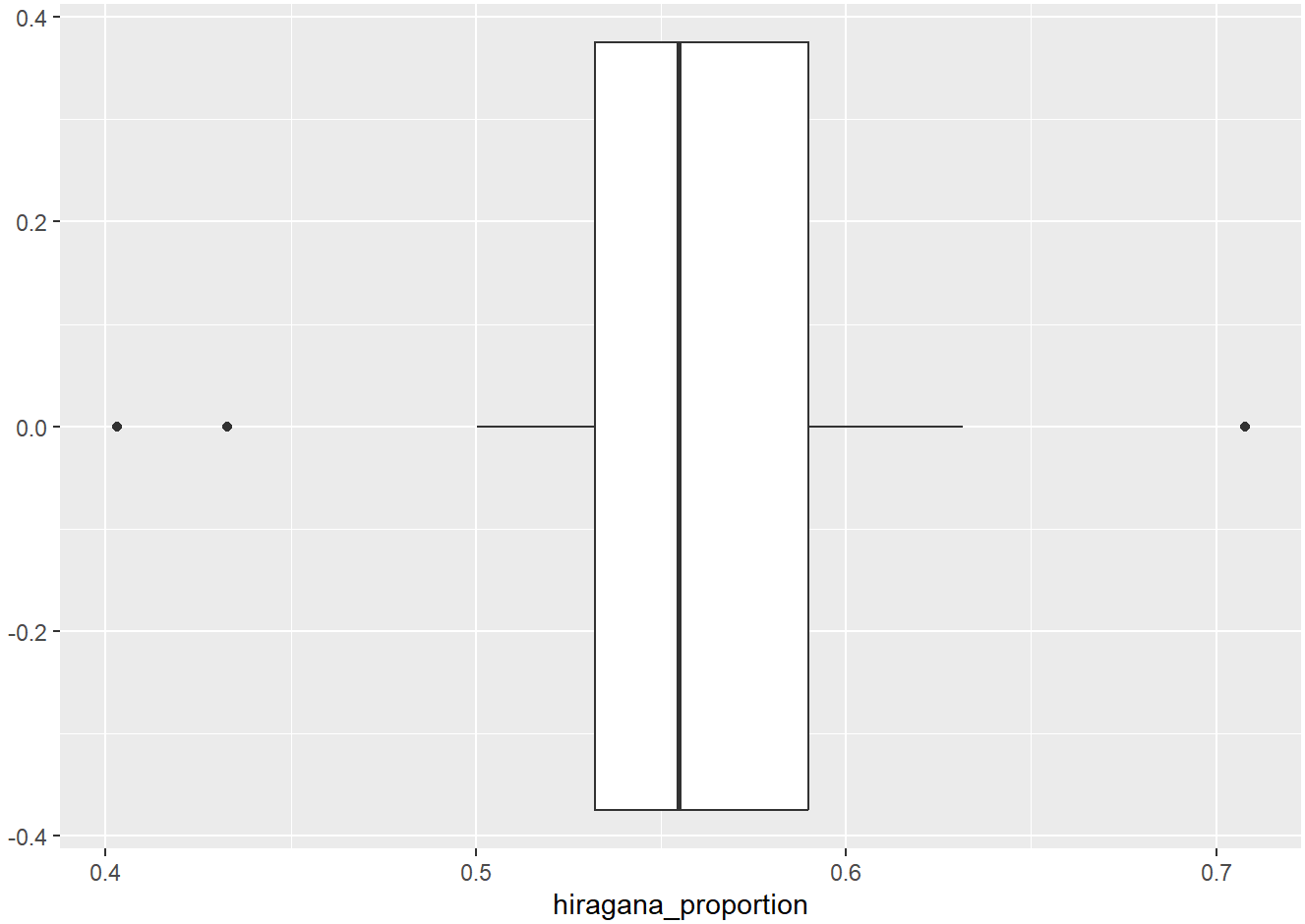

```
ggplot(jp_novels_df, aes(x = kanji)) + geom_boxplot()
```



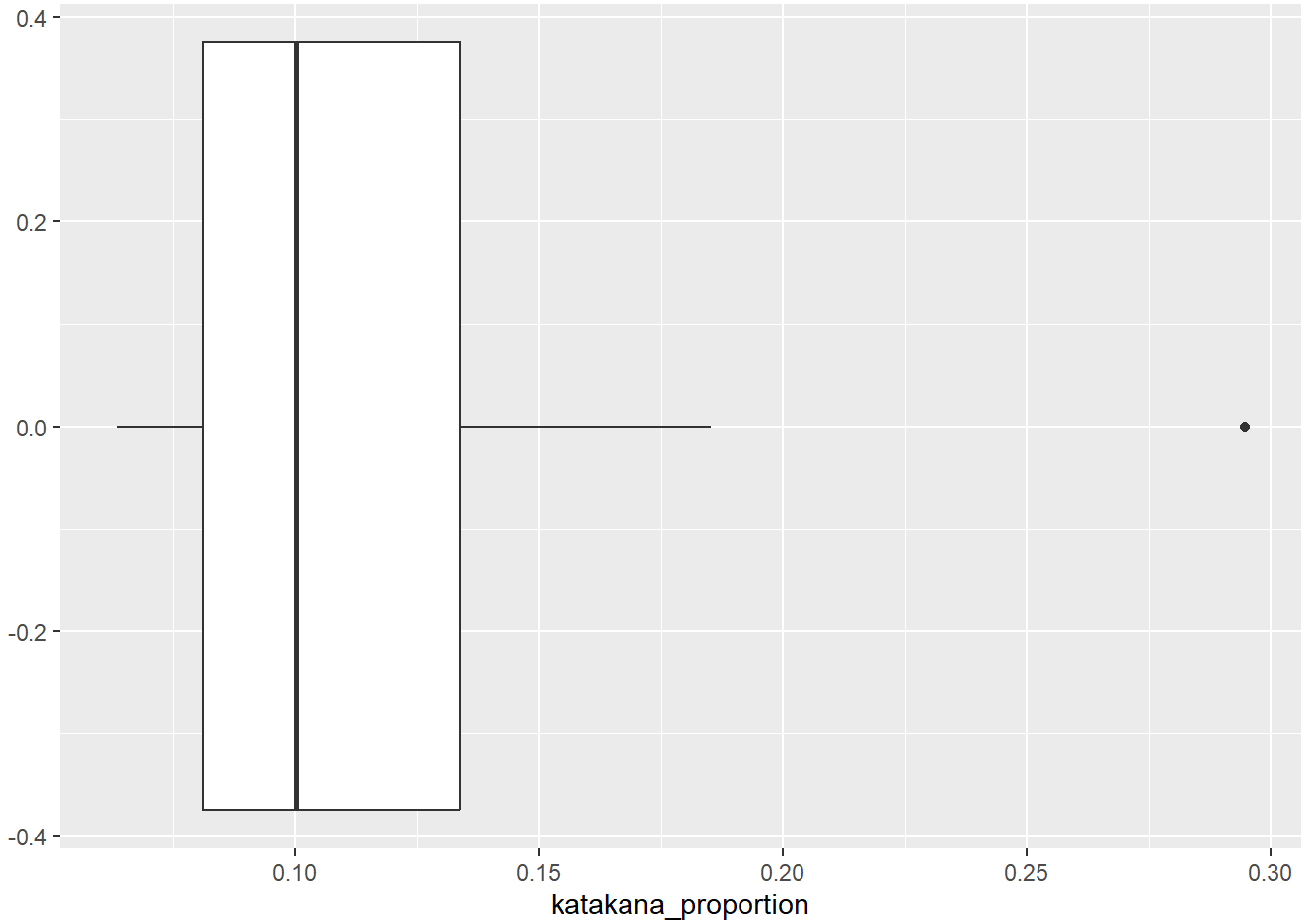
Based on these charts, it is apparent that the distributions for each of the three writing systems are quite skewed right. That is, there are many novels with a smaller counts of Hiragana, Katakana, and Kanji than there are larger counts of each writing system. Even so, it seems that there is a smaller spread of Katakana than there is of the other two writing systems.

However, it is possible that these results could be skewed by higher total character counts within given novels. To alleviate part of this, one can find the percentage of each novel's Japanese characters that belong to each writing system:

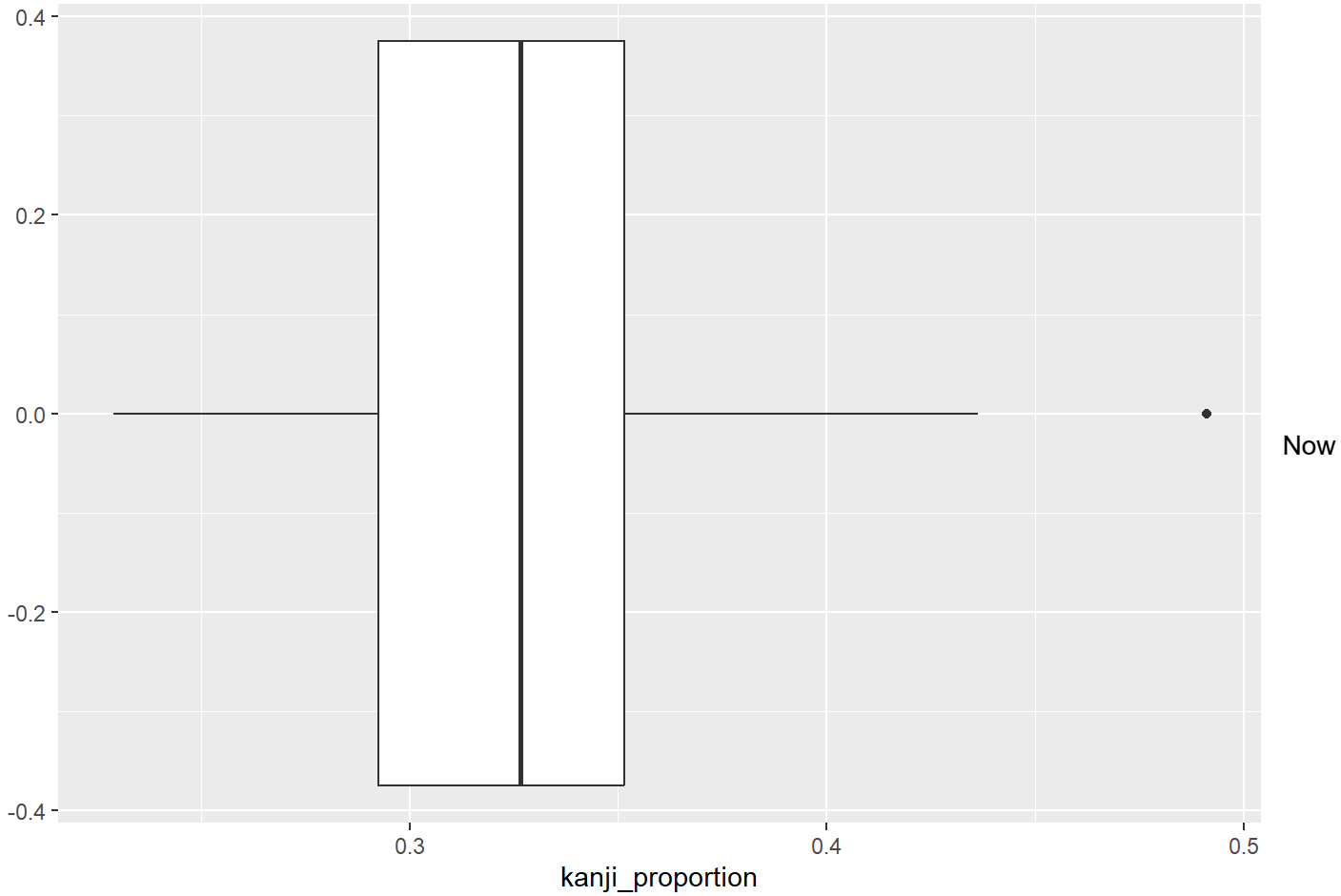
```
# [STEP 1]: Create a new column in `jp_novels_df` that represents the proportion of each novel's Japanese characters that are part of a given writing system.
jp_novels_df$hiragana_proportion <- jp_novels_df$hiragana / jp_novels_df$japanese_total
jp_novels_df$katakana_proportion <- jp_novels_df$katakana / jp_novels_df$japanese_total
jp_novels_df$kanji_proportion <- jp_novels_df$kanji / jp_novels_df$japanese_total

# [STEP 2]: Display
ggplot(jp_novels_df, aes(x = hiragana_proportion)) + geom_boxplot()
```




```
ggplot(jp_novels_df, aes(x = katakana_proportion)) + geom_boxplot()
```



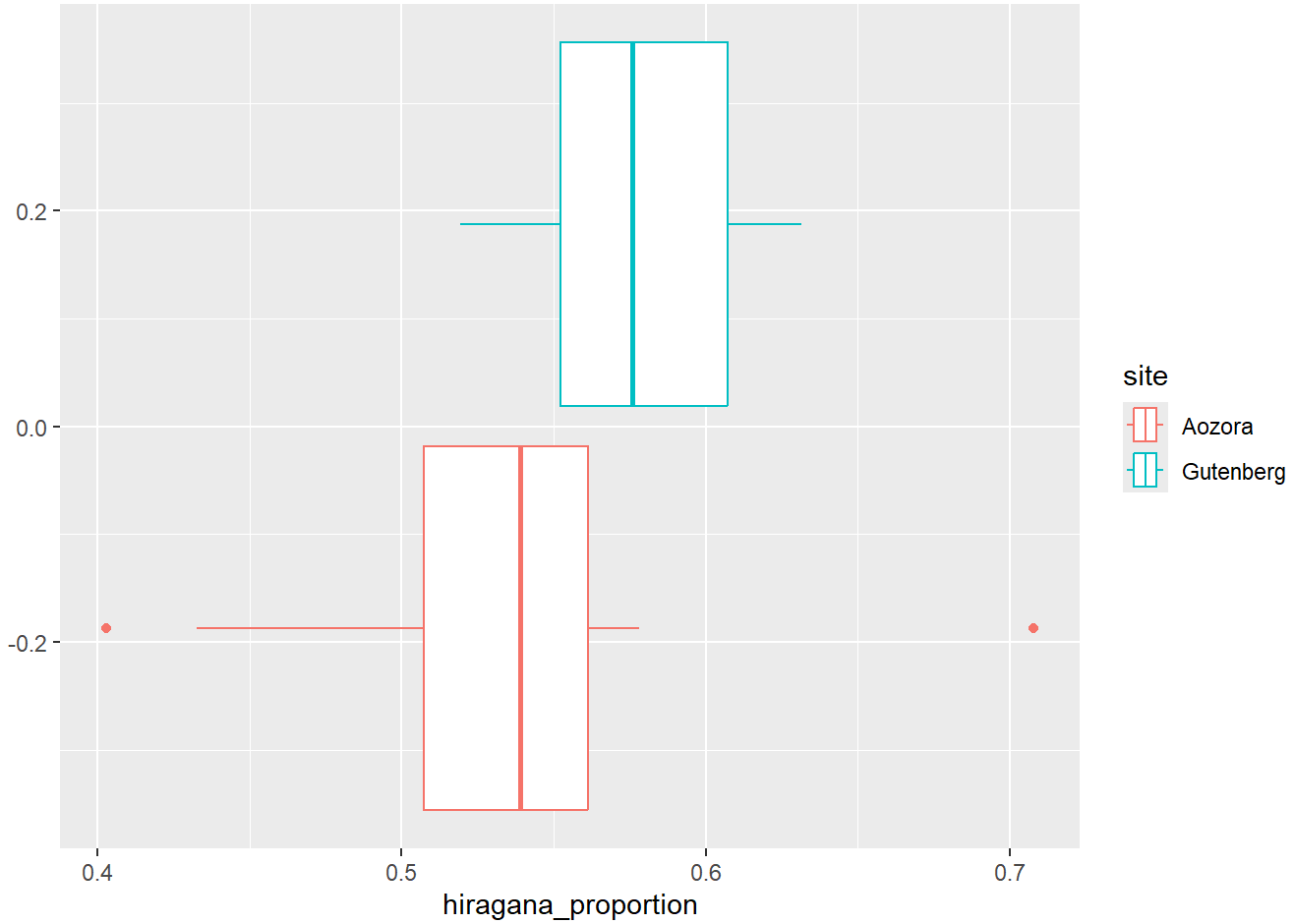

```
ggplot(jp_novels_df, aes(x = kanji_proportion)) + geom_boxplot()
```



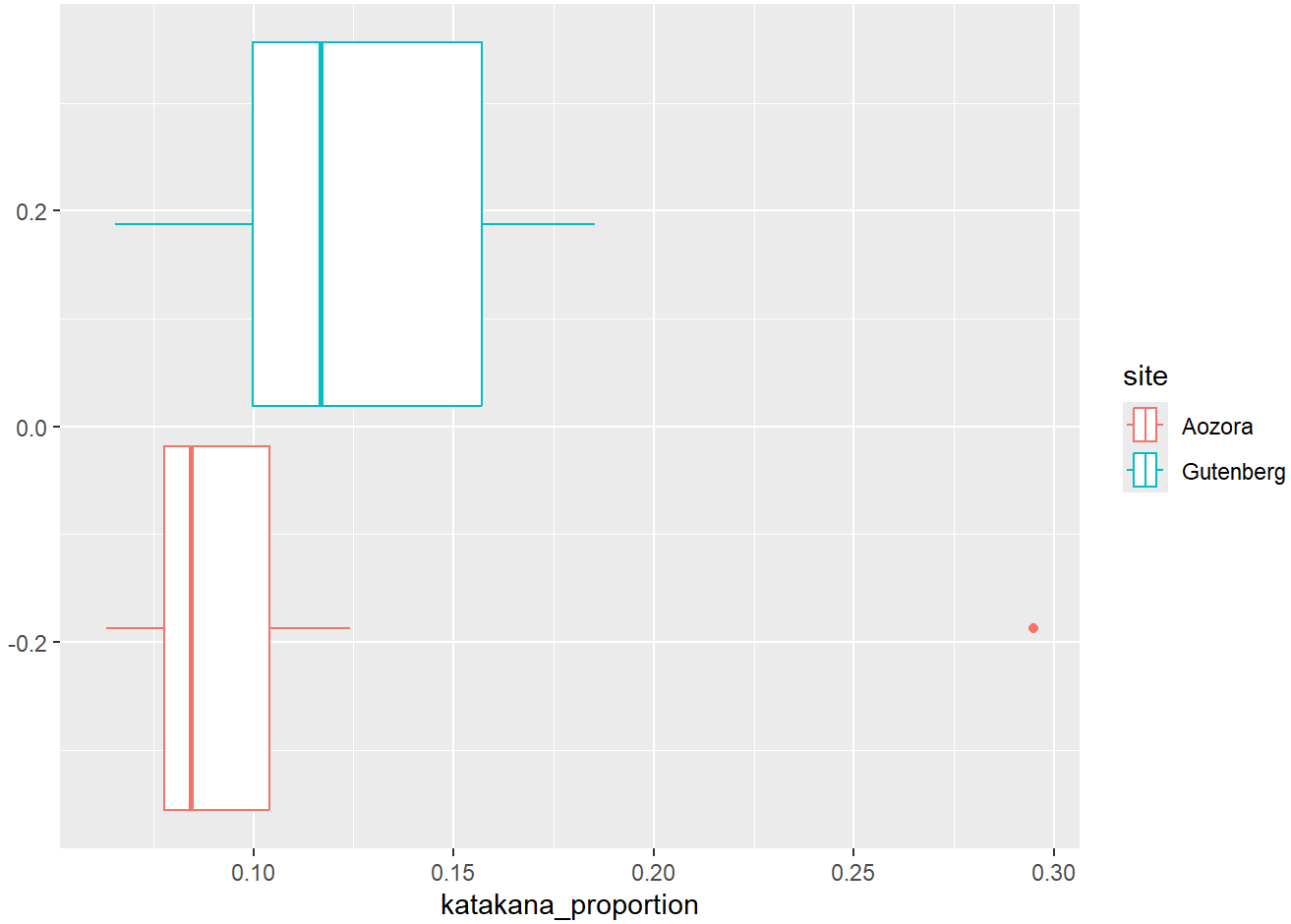
that these values are standardized, one can see that the proportion of Hiragana seems fairly normal compared to the other two systems. In addition, for the novels seen, the proportion of the novels that are written in Katakana is far smaller than the proportions of the other two systems—reaching a mean of only 10% of the novels.

Additionally, it is interesting to note the difference in the distributions of Hiragana, Katakana, and Kanji between both the Gutenberg and Aozora sites:

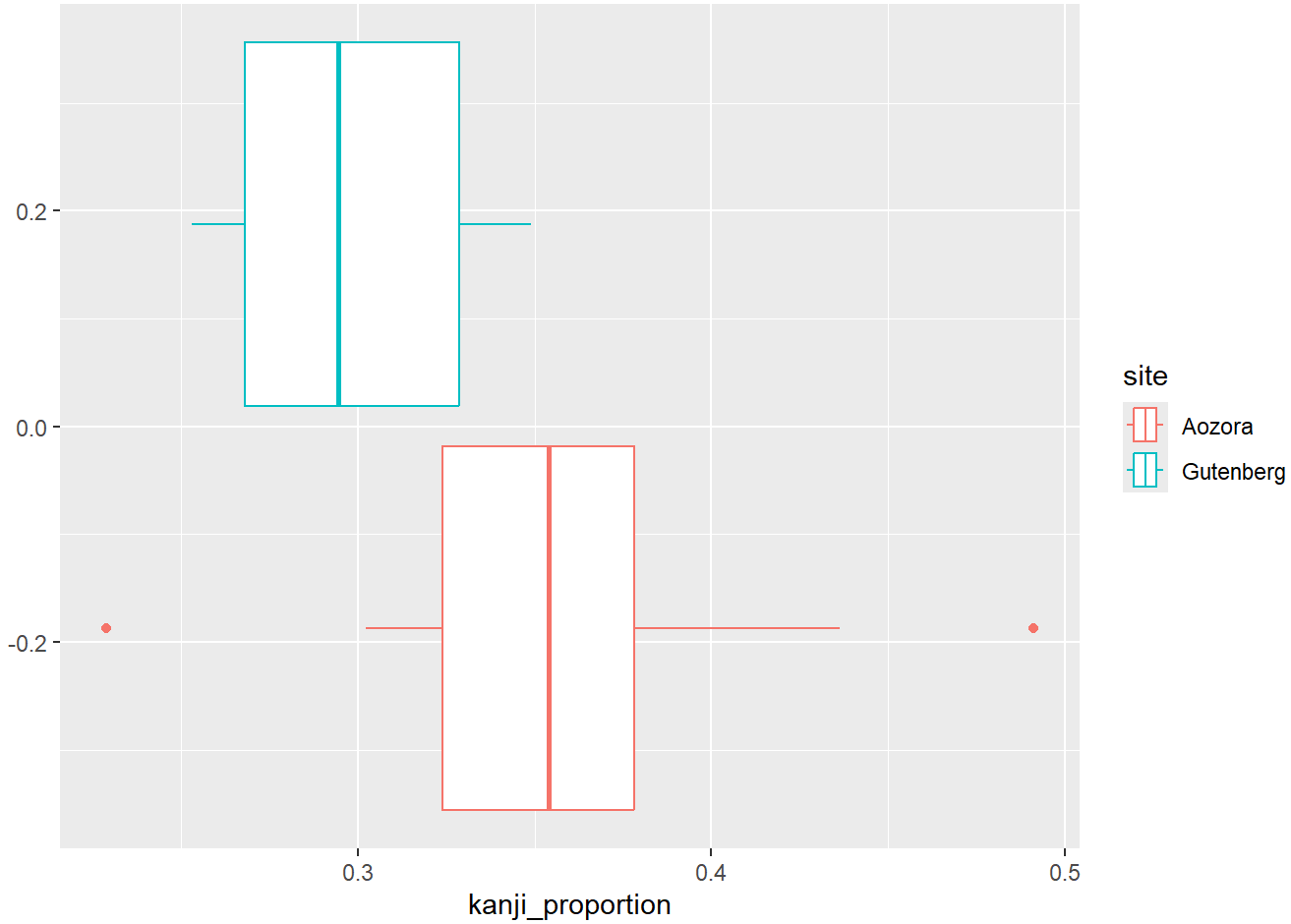
```
ggplot(jp_novels_df, aes(x = hiragana_proportion)) +
 geom_boxplot(aes(col = site))
```




```
ggplot(jp_novels_df, aes(x = katakana_proportion)) +
  geom_boxplot(aes(col = site))
```




```
ggplot(jp_novels_df, aes(x = kanji_proportion)) +
  geom_boxplot(aes(col = site))
```



These graphs seem to suggest a very clear difference in the distributions of the three systems between Aozora and Gutenberg—indicating merit in analyzing the differences between these two websites' novels further.

With this preliminary analysis complete, it is now possible to move to more complicated analysis.

Part 4: Future Analysis:

The goal of this project is to determine if there are any significant differences in the distributions of Hiragana, Katakana, and Kanji within Japanese works. In order to definitively determine this, one can run a collection of two-sample t-tests between each of the three writing systems (i.e., comparing Hiragana and Katakana, Katakana and Kanji, and Hiragana and Kanji).

In addition, to determine if there is a significant difference between the distribution of the writing systems in originally-Japanese works and works translated to Japanese, one can run a two-sample t-test for each writing system on the novels that are from Aozora and the novels that are from Gutenberg.

Via these tests, it is possible to answer the questions that were proposed at the beginning of this report.

Part 5: Team Member Contributions:

1. Pankshi Parekar: Created/Organized the GitHub repository. Created/Organized the Final Report. Ran the Simple Random Selections for choosing the novels for analysis, and prepared the `data` folder in the GitHub repository accordingly. Wrote the Abstract, as well as parts 1, 3, and 4 in the report.
2. Isabella Perez: Wrote the Introduction section of the report.

3. Ulices Ramirez Lopez: Built an R Markdown workflow to automatically clean the Japanese text files from Aozora and Gutenberg. The main goal was to remove unnecessary parts like headers, footers, and any non-Japanese or English characters. Did this by reading each .txt file, converting the text to UTF-8, cutting out lines before and after certain markers (like —— and 底本：), and deleting inline Aozora markup such as ［＃…］, 《…》, and ｜. After cleaning, the script counted only Japanese characters (kanji, hiragana, and katakana) to measure how much of the text remained. Finally, it saved two CSV summaries, one for Aozora and one for Gutenberg — showing the cleaned results for each file.