AI Data Analysis

Darian Othman

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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
```

```r
library(readr)
library(ggplot2)
```

```r
wd<-c("C:/Users/I6240624/Documents/BISS/Master Thesis/Code/DarianOthmanMasterThesis/Notebooks")
setwd(wd)
```

```r
dfinsen <- read.csv("C:/Users/I6240624/Documents/BISS/Master Thesis/Data/dfinsenchar.csv")
dftten <- read.csv("C:/Users/I6240624/Documents/BISS/Master Thesis/Data/dfttenchar.csv")
dfyten <- read.csv("C:/Users/I6240624/Documents/BISS/Master Thesis/Data/dfytenchar.csv")

insen <- read.csv("C:/Users/I6240624/Documents/BISS/Master Thesis/Data/insenchar_cat1p1.csv")
tten <- read.csv("C:/Users/I6240624/Documents/BISS/Master Thesis/Data/ttenchar_cat1p1.csv")
yten <- read.csv("C:/Users/I6240624/Documents/BISS/Master Thesis/Data/ytenchar_cat1p1.csv")
```

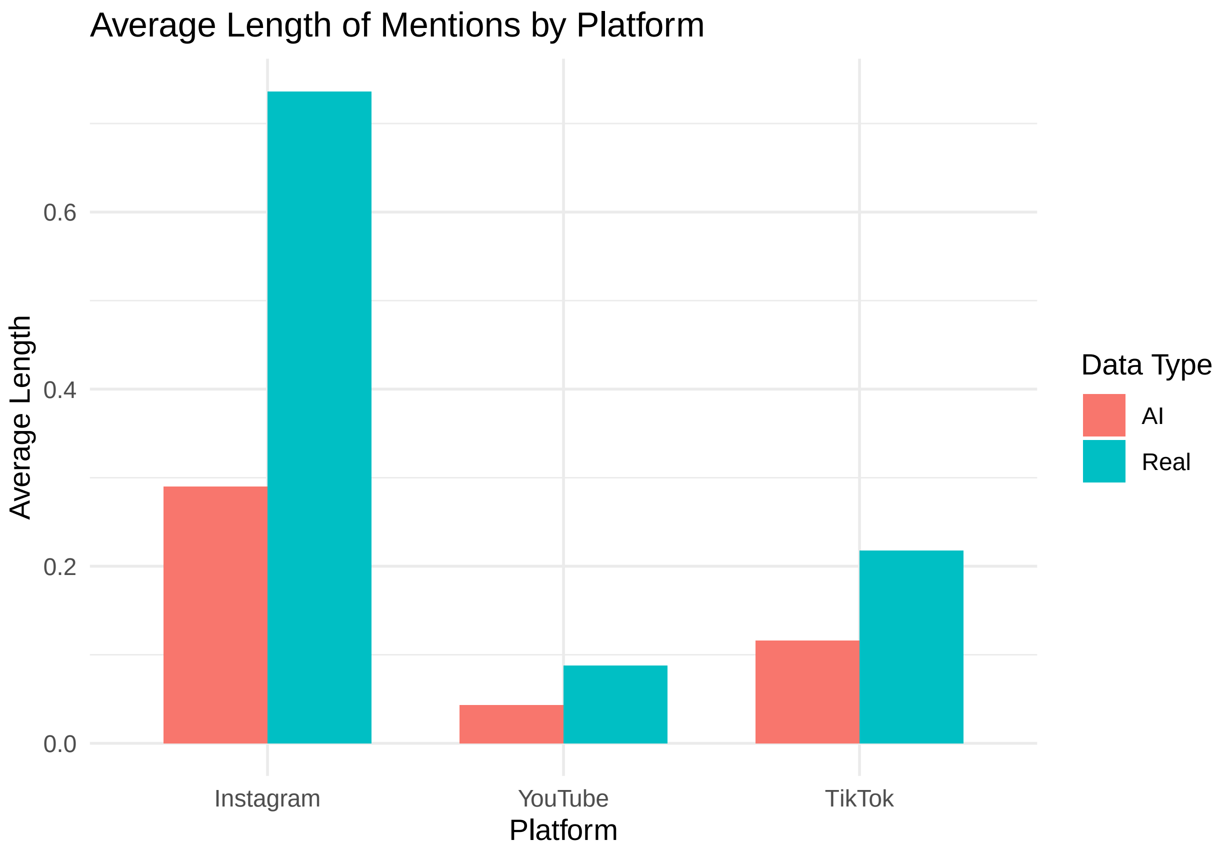
```r
combined_data_real <- bind_rows(
  # Instagram
  bind_rows(
    dfinsen %>% mutate(language = "EN")
  ) %>%
    mutate(platform = "Instagram"),

  # YouTube
  bind_rows(
    dfyten %>% mutate(language = "EN")
  ) %>%
    mutate(platform = "YouTube"),

  # TikTok
  bind_rows(
    dftten %>% mutate(language = "EN")
  ) %>%
    mutate(platform = "TikTok")
)
# Set the order of factor levels for the facet
combined_data_real$platform <- factor(
  combined_data_real$platform,
  levels = c("Instagram", "YouTube", "TikTok")
)
```

```
combined_data_ai <- bind_rows(
  # Instagram
  bind_rows(
    insen %>% mutate(language = "EN")
  ) %>%
    mutate(platform = "Instagram"),

  # YouTube
  bind_rows(
    yten %>% mutate(language = "EN")
  ) %>%
    mutate(platform = "YouTube"),

  # TikTok
  bind_rows(
    tten %>% mutate(language = "EN")
  ) %>%
    mutate(platform = "TikTok")
)
# Set the order of factor levels for the facet
combined_data_ai$platform <- factor(
  combined_data_ai$platform,
  levels = c("Instagram", "YouTube", "TikTok")
)
```

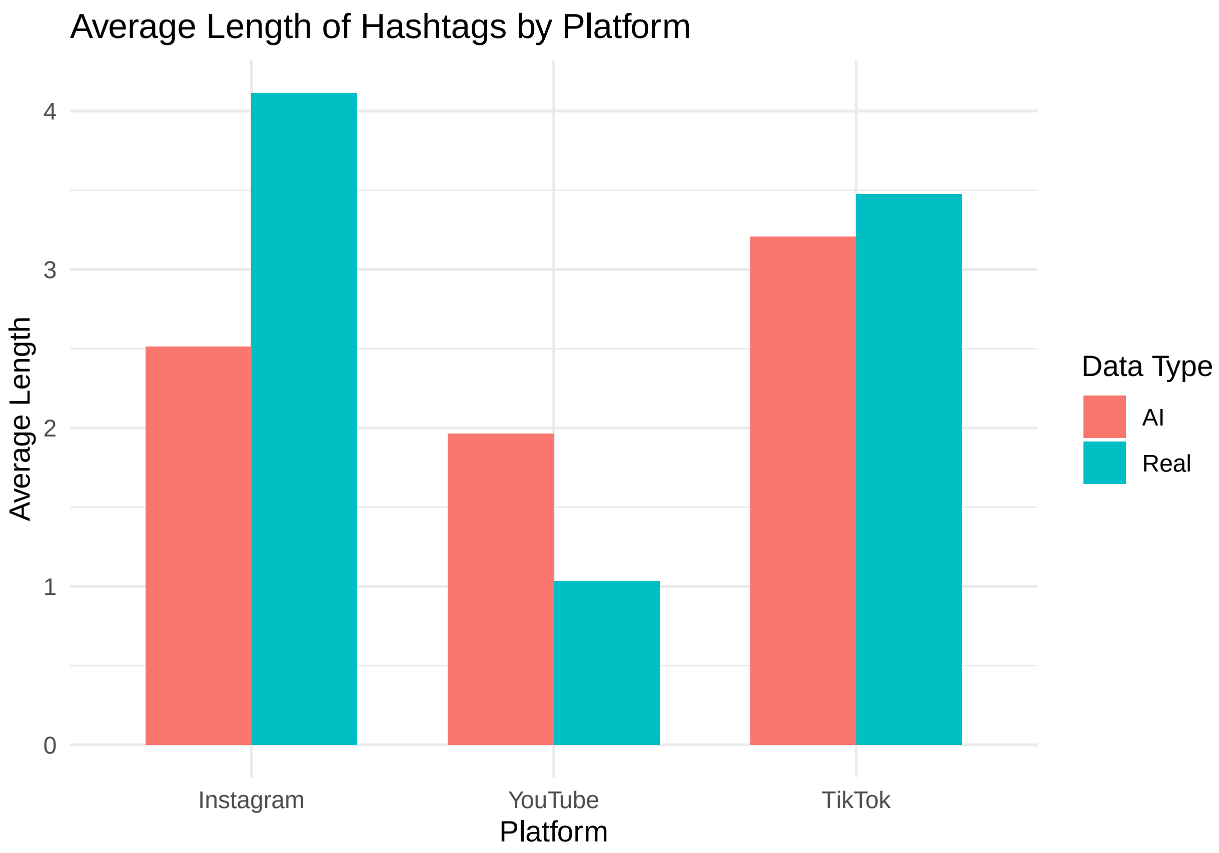
```
combined_data_real <- combined_data_real %>%
  mutate(source = "real")

combined_data_ai <- combined_data_ai %>%
  mutate(source = "ai")

# Bind the two datasets together
final_combined <- bind_rows(combined_data_real, combined_data_ai)
```

```
average_mention_real <- combined_data_real%>%group_by(platform)%>%
  mutate(mentions_length = lengths(mentions_count)) %>%
  summarize(average_length = mean(mentions_count, na.rm = TRUE))

average_mention_ai <- combined_data_ai%>%group_by(platform)%>%
  mutate(mentions_length = lengths(mentions_count)) %>%
  summarize(average_length = mean(mentions_count, na.rm = TRUE))

average_hashtags_real <- combined_data_real%>%group_by(platform)%>%
  mutate(hashtags_length = lengths(hashtags_count)) %>%
  summarize(average_length = mean(hashtags_count, na.rm = TRUE))

average_hashtags_ai <- combined_data_ai%>%group_by(platform)%>%
  mutate(hashtags_length = lengths(hashtags_count)) %>%
  summarize(average_length = mean(hashtags_count, na.rm = TRUE))

average_urls_real <- combined_data_real%>%group_by(platform)%>%
  mutate(urls_length = lengths(urls_count)) %>%
  summarize(average_length = mean(urls_count, na.rm = TRUE))

average_urls_ai <- combined_data_ai%>%group_by(platform)%>%
  mutate(urls_length = lengths(urls_count)) %>%
  summarize(average_length = mean(urls_count, na.rm = TRUE))

average_emojis_real <- combined_data_real%>%group_by(platform)%>%
  mutate(emojis_length = lengths(emojis_count)) %>%
  summarize(average_length = mean(emojis_count, na.rm = TRUE))

average_emojis_ai <- combined_data_ai%>%group_by(platform)%>%
  mutate(emojis_length = lengths(emojis_count)) %>%
  summarize(average_length = mean(emojis_count, na.rm = TRUE))

average_caption_real <- combined_data_real%>%group_by(platform)%>%
  summarize(average_length = mean(caption_length, na.rm = TRUE))

average_caption_ai <- combined_data_ai%>%group_by(platform)%>%
  summarize(average_length = mean(caption_length, na.rm = TRUE))


a <- bind_rows(bind_rows(
  mutate(average_mention_real, data_type = "Real"),
  mutate(average_mention_ai, data_type = "AI"))%>%
    mutate(char = "mention"),
  bind_rows(
  mutate(average_hashtags_real, data_type = "Real"),
  mutate(average_hashtags_ai, data_type = "AI"))%>%
    mutate(char = "hashtags"),
  bind_rows(
  mutate(average_urls_real, data_type = "Real"),
  mutate(average_urls_ai, data_type = "AI"))%>%
    mutate(char = "urls"),
  bind_rows(
  mutate(average_emojis_real, data_type = "Real"),
  mutate(average_emojis_ai, data_type = "AI"))%>%
    mutate(char = "emojis"),
  bind_rows(
  mutate(average_caption_real, data_type = "Real"),
  mutate(average_caption_ai, data_type = "AI"))%>%
    mutate(char = "caption")
)
```

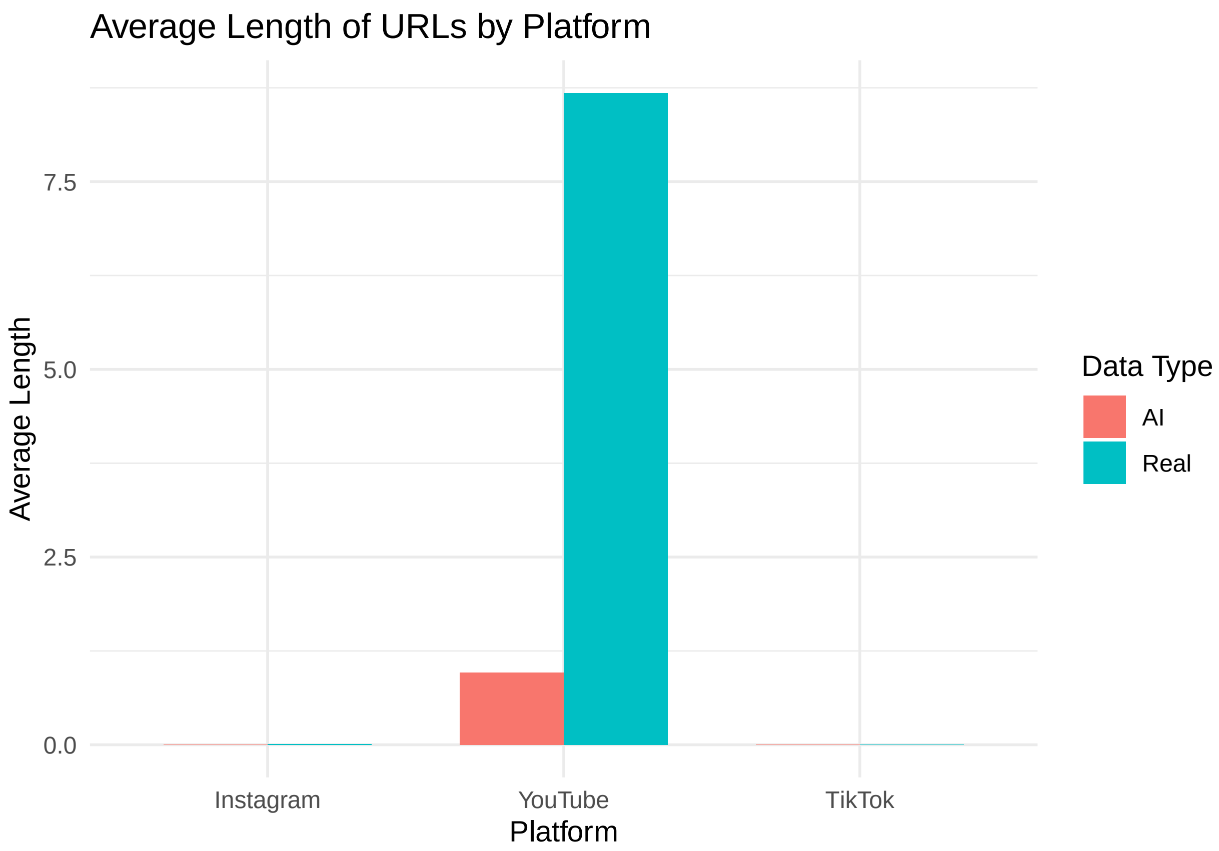
```
filtered_data <- subset(a, char == "mention")

# Plot the filtered data
ggplot(filtered_data, aes(x = platform, y = average_length, fill = data_type)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(
    title = "Average Length of Mentions by Platform",
    x = "Platform",
    y = "Average Length",
    fill = "Data Type"
  ) +
  theme_minimal()
```



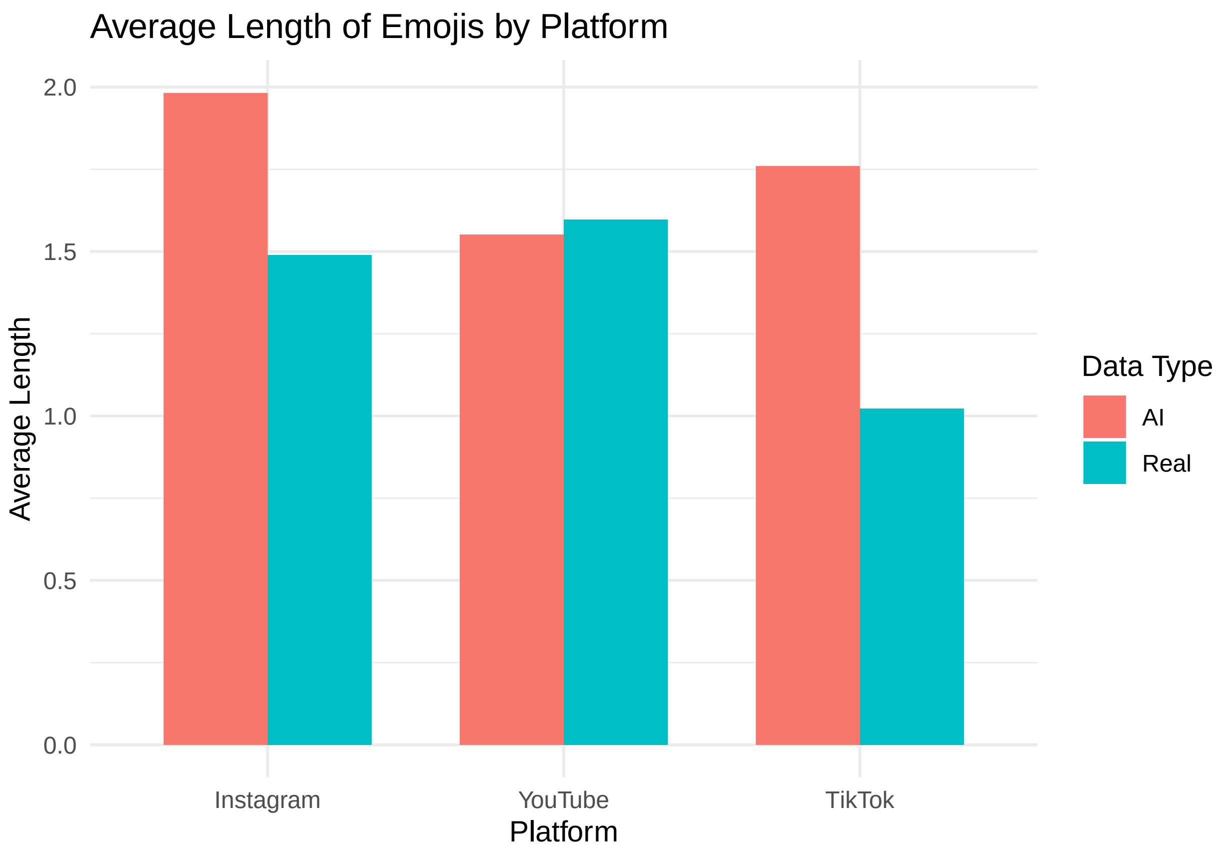
```
filtered_data <- subset(a, char == "hashtags")

# Plot the filtered data
ggplot(filtered_data, aes(x = platform, y = average_length, fill = data_type)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(
    title = "Average Length of Hashtags by Platform",
    x = "Platform",
    y = "Average Length",
    fill = "Data Type"
  ) +
  theme_minimal()
```



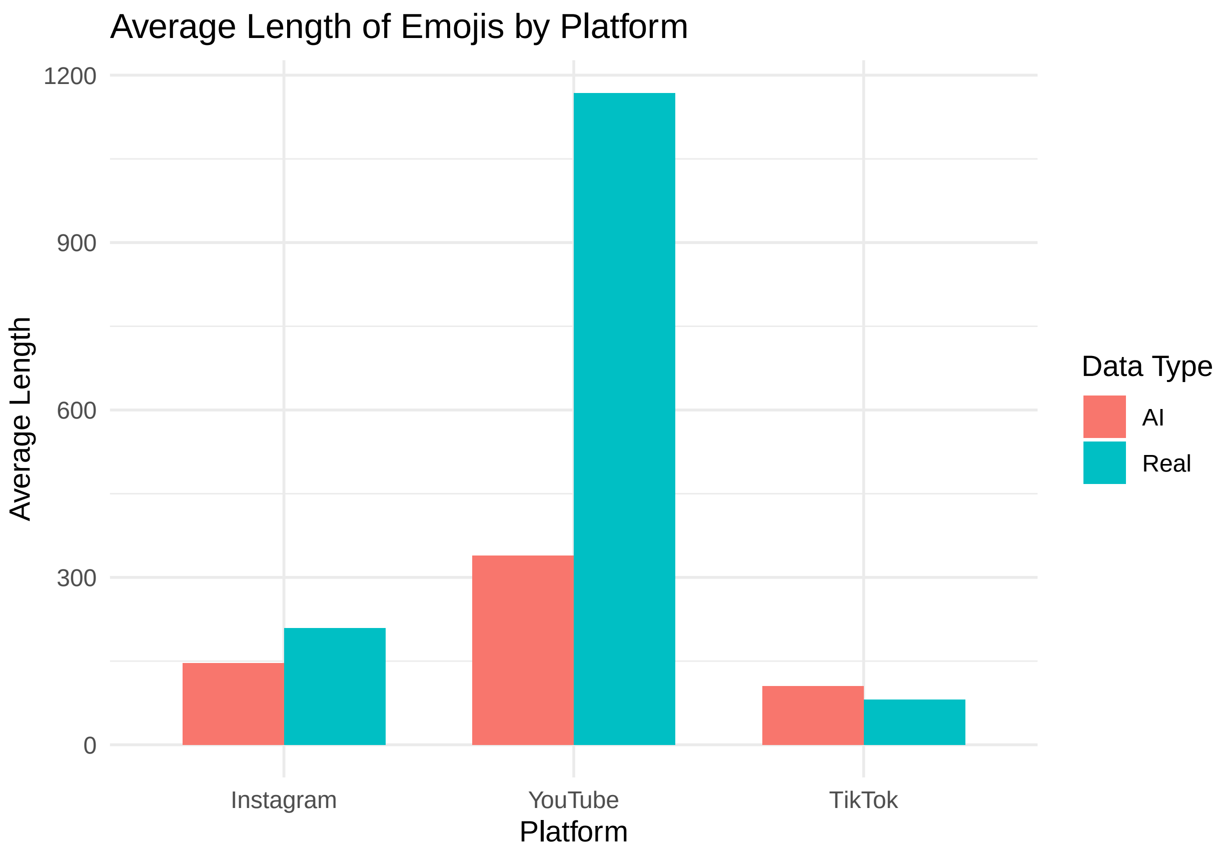
```r
filtered_data <- subset(a, char == "urls")

# Plot the filtered data
ggplot(filtered_data, aes(x = platform, y = average_length, fill = data_type)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(
    title = "Average Length of URLs by Platform",
    x = "Platform",
    y = "Average Length",
    fill = "Data Type"
  ) +
  theme_minimal()
```



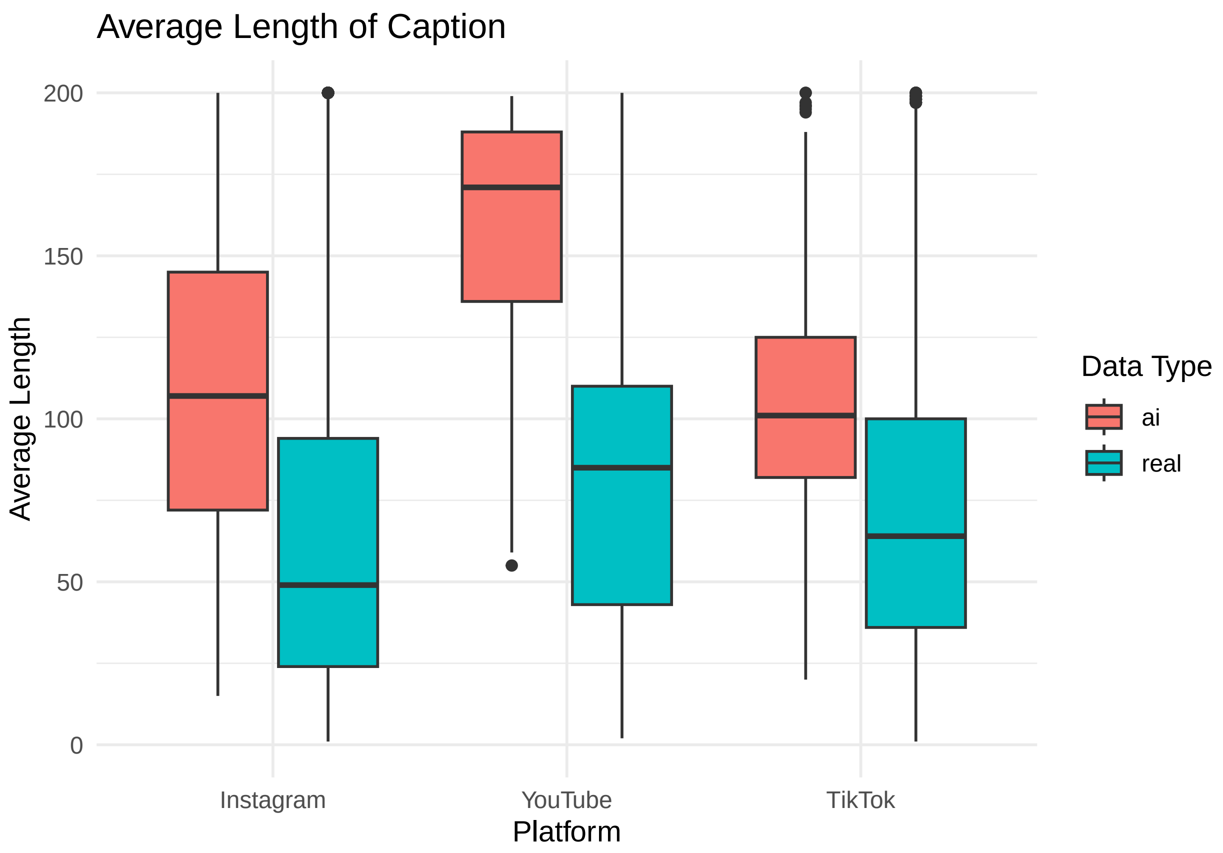
```r
filtered_data <- subset(a, char == "emojis")

# Plot the filtered data
ggplot(filtered_data, aes(x = platform, y = average_length, fill = data_type)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(
    title = "Average Length of Emojis by Platform",
    x = "Platform",
    y = "Average Length",
    fill = "Data Type"
  ) +
  theme_minimal()
```



```r
filtered_data <- subset(a, char == "caption")

# Plot the filtered data
ggplot(filtered_data, aes(x = platform, y = average_length, fill = data_type)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(
    title = "Average Length of Emojis by Platform",
    x = "Platform",
    y = "Average Length",
    fill = "Data Type"
  ) +
  theme_minimal()
```



```
ggplot(final_combined, aes(x = platform, y = caption_length, fill = source)) +
  geom_boxplot() +
  labs(
    title = "Average Length of Caption",
    x = "Platform",
    y = "Average Length",
    fill = "Data Type"
  ) +
  theme_minimal()+
  scale_y_continuous(limits = c(0, 200))
```


Average Length of Caption
Average Length
200
150
100
50
0
Instagram
YouTube
TikTok
Platform
Data Type
ai
real