---
title: "Lab 1 v3"
author: "Emily Mingus, Aden Bhagwat, Erick Njue"
date: "2025-01-08"
output:
  pdf_document:
    keep_tex: true
    output_dir: "~/Desktop/Lab 1 v3"

```
library(edld652)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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
```

```
#set_key('aden1234')

#list_datasets()
#Sys.unsetenv("AZURE_SAS_TOKEN")  # Clear any existing token
# Set your new SAS token
#Sys.setenv(AZURE_SAS_TOKEN = "aden1234")

acgd <- get_data("EDFacts_acgr_lea_2011_2019")
```

```
## Rows: 11326 Columns: 29
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (17): ALL_RATE, CWD_RATE, DATE_CUR, ECD_RATE, FIPST, FILEURL, LEAID, LE...
## dbl  (11): ALL_COHORT, CWD_COHORT, ECD_COHORT, LEP_COHORT, MAM_COHORT, MAS_C...
## dttm  (1): DL_INGESTION_DATETIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
acgdd <- get_documentation("EDFacts_acgr_lea_2011_2019")
```

```
## https://www2.ed.gov/about/inits/ed/edfacts/data-files/acgr-sy2018-19-public-file-documentation.docx
```

```
#list_datasets()
#NOT WORKING
```

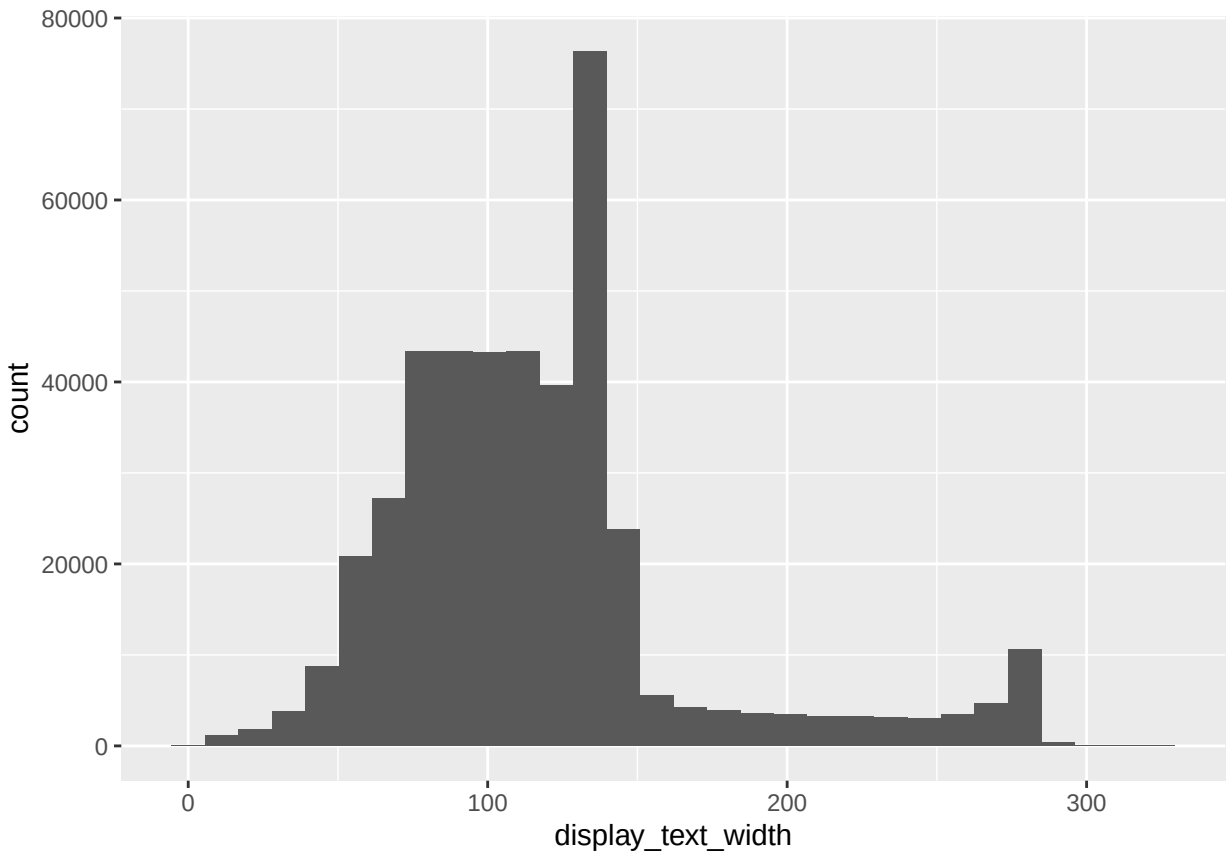
```
library(here)
```

```
## here() starts at C:/Users/adenb/OneDrive/Desktop/Git/EDLD-652-Lab-1
```

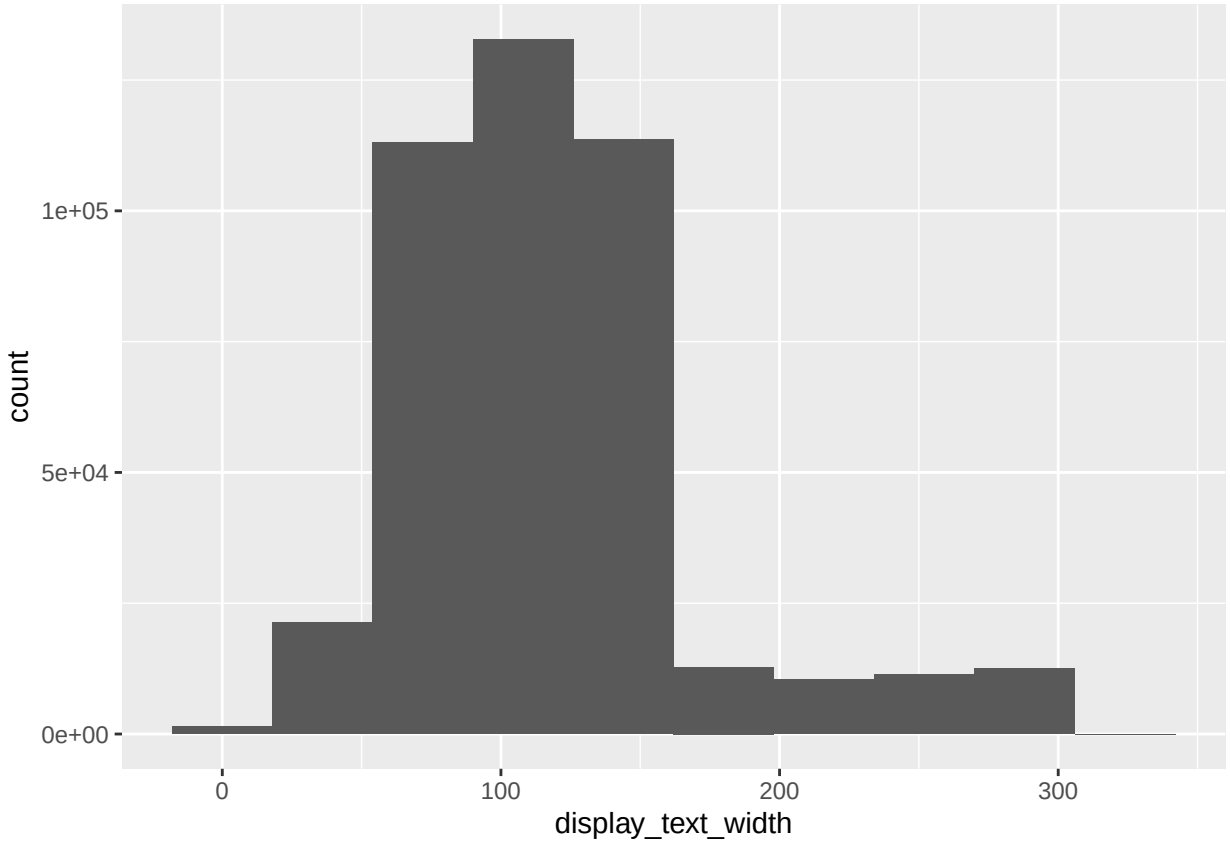
```
library(rio)
library(ggplot2)
df<-import(here("data/rstats_tweets.rds"))
```

```
## Warning: Missing `trust` will be set to FALSE by default for RDS in 2.0.0.
```

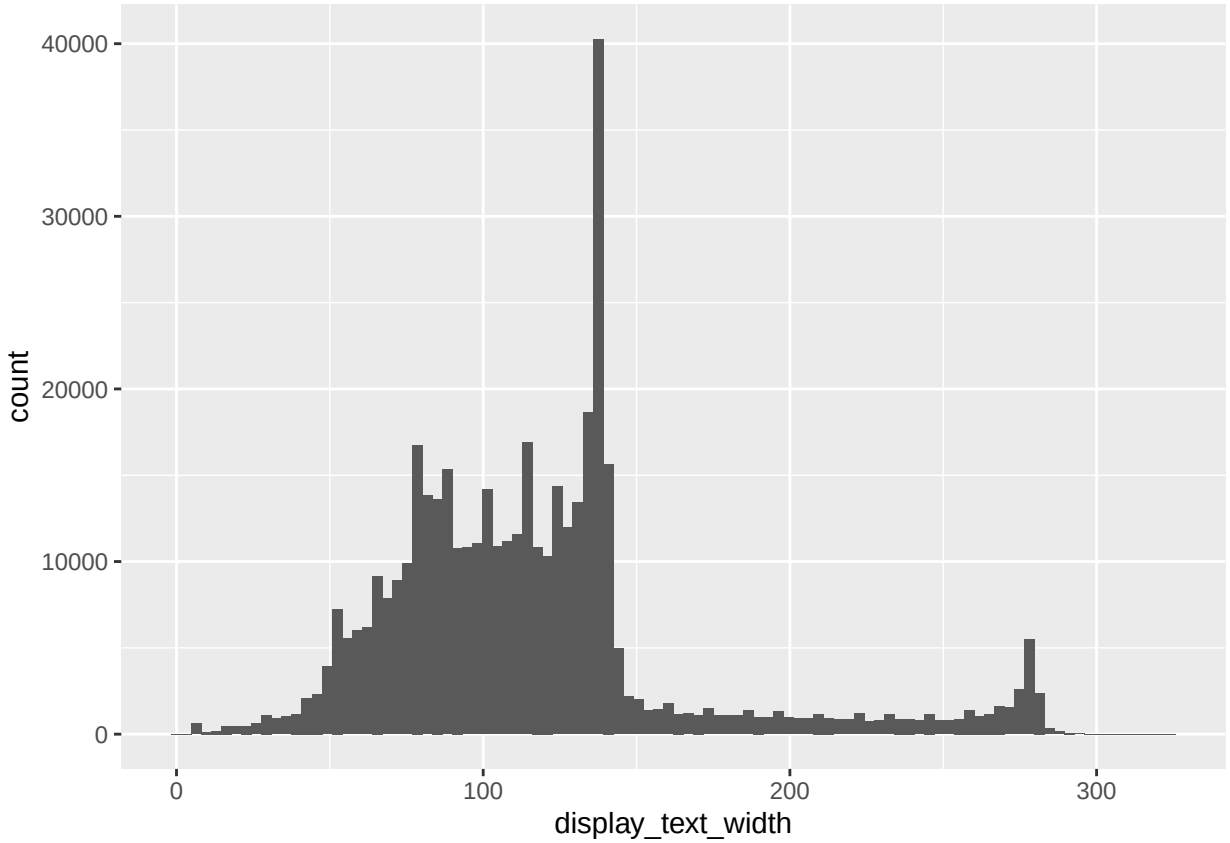
```
df %>%
  ggplot(aes(x= display_text_width))+
  geom_histogram(bins=30)
```



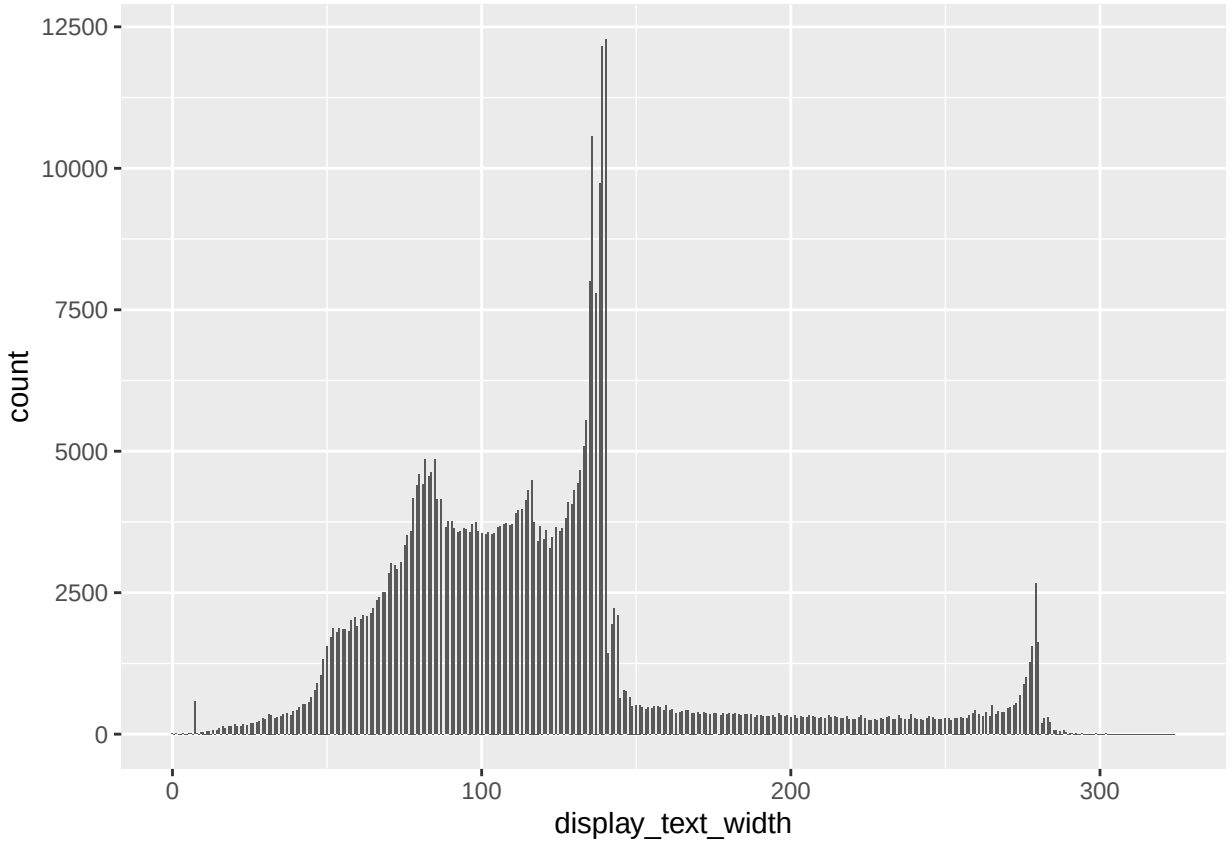
```r
df %>%
  ggplot(aes(x= display_text_width))+
  geom_histogram(bins=10)
```




```r
df %>%
  ggplot(aes(x= display_text_width))+
  geom_histogram(bins=100)
```




```
df %>%
  ggplot(aes(x= display_text_width))+
  geom_histogram(bins=500)
```



We are using 30 bins which is what R automatically selects, we think this gives the clearest pattern without being overwhelming.

```
n_plot <- sum(grepl("plot", tolower(df$text)))

n_plot/nrow(df)
```

```
## [1] 0.06834019
```

6.8% of the posts contain the word plot

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.4.2
```

```
df_text <- df %>%
  unnest_tokens(word, description)


df_text %>%
  anti_join(stop_words) %>%
  filter(!word %in% c("t.co", "https", "http", "rt", "rstats")) %>%
  count(word, sort = TRUE) %>%
  mutate(word = reorder(word, n)) %>% # make y-axis ordered by n
  slice(1:15) %>% # select only the first 15 rows
  ggplot(aes(n, word)) +
    geom_col(fill = "cornflowerblue")+
  theme_minimal()+
  labs(x="Count", y="Word", title="Word frequencies in posts", subtitle= "Top 15 words displayed", capti
```

```
## Joining with `by = join_by(word)`
```