Tiday Tuesday Week 36

Medium data science article metadata

Goal is to work with the tidytext package.

```
# LIBRARIES ------------------------------------------------------------------#
library(tidyverse)
library(tidytext)
library(lubridate)
```

```
# IMPORT DATA ----------------------------------------------------------------#
raw_metadata <- read_csv(file ="medium_datasci.csv", quote='"')
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   title = col_character(),
##   subtitle = col_character(),
##   author = col_character(),
##   publication = col_character(),
##   claps = col_double(),
##   url = col_character(),
##   author_url = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
head(raw_metadata)
```

```
## # A tibble: 6 x 21
##      x1 title subtitle image author publication  year month   day
##   <int> <chr> <chr>    <int> <chr>  <chr>       <int> <int> <int>
## 1     2 Onli~ Online ~     1 Emma ~ <NA>         2017     8     1
## 2     5 A.I.~ <NA>         0 Sanpa~ <NA>         2017     8     1
## 3    11 Futu~ From Ph~     1 Z      <NA>         2017     8     1
## 4    12 The ~ A true ~     1 Emiko~ MILLENNIAL~  2017     8     1
## 5    17 Os M~ mas per~     1 Giova~ NEW ORDER    2017     8     1
## 6    18 The ~ Origina~     1 Syed ~ Towards Da~  2017     8     1
## # ... with 12 more variables: reading_time <int>, claps <dbl>, url <chr>,
## #   author_url <chr>, tag_ai <int>, tag_artificial_intelligence <int>,
## #   tag_big_data <int>, tag_data <int>, tag_data_science <int>,
## #   tag_data_visualization <int>, tag_deep_learning <int>,
## #   tag_machine_learning <int>
```

I noticed some duplicate names, articles without names, subtitles as duplicates of titles, etc... Which means it is time to clean.

```
# CLEAN ----------------------------------------------------------------------#
metadata <- raw_metadata %>%
  distinct(title, author, .keep_all = TRUE) %>%
  drop_na(title)
head(metadata)
```

```
## # A tibble: 6 x 21
##      x1 title subtitle image author publication  year month   day
```

```
##   <int> <chr> <chr>    <int> <chr>  <chr>       <int> <int> <int>
## 1     2 Onli~ Online ~     1 Emma ~ <NA>         2017     8     1
## 2     5 A.I.~ <NA>         0 Sanpa~ <NA>         2017     8     1
## 3    11 Futu~ From Ph~     1 Z      <NA>         2017     8     1
## 4    12 The ~ A true ~     1 Emiko~ MILLENNIAL~  2017     8     1
## 5    17 Os M~ mas per~     1 Giova~ NEW ORDER    2017     8     1
## 6    18 The ~ Origina~     1 Syed ~ Towards Da~  2017     8     1
## # ... with 12 more variables: reading_time <int>, claps <dbl>, url <chr>,
## #   author_url <chr>, tag_ai <int>, tag_artificial_intelligence <int>,
## #   tag_big_data <int>, tag_data <int>, tag_data_science <int>,
## #   tag_data_visualization <int>, tag_deep_learning <int>,
## #   tag_machine_learning <int>
```

```
dim(raw_metadata)
```

```
## [1] 78388    21
```

```
dim(metadata)
```

```
## [1] 75105    21
```

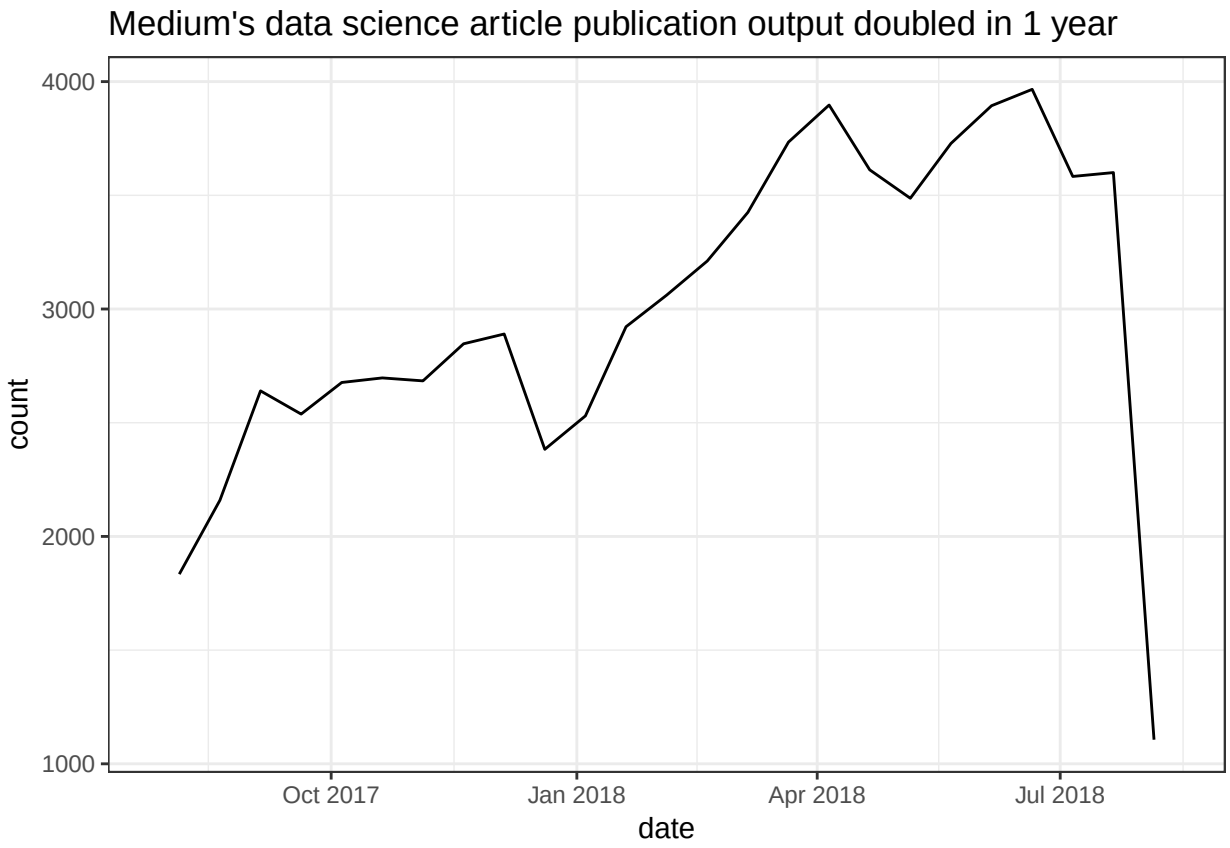
The data isn't perfect, but looking much better than before.

Let's combine the three date columns into one.

```
metadata <- metadata %>%
  mutate(date = ymd(paste(year, month, day, sep= '-')),
         weekday = wday(as.Date(date,'%Y-%m-%d'), label = TRUE, abbr = FALSE))
```

What is the trend in the number of data science articles published over the year?

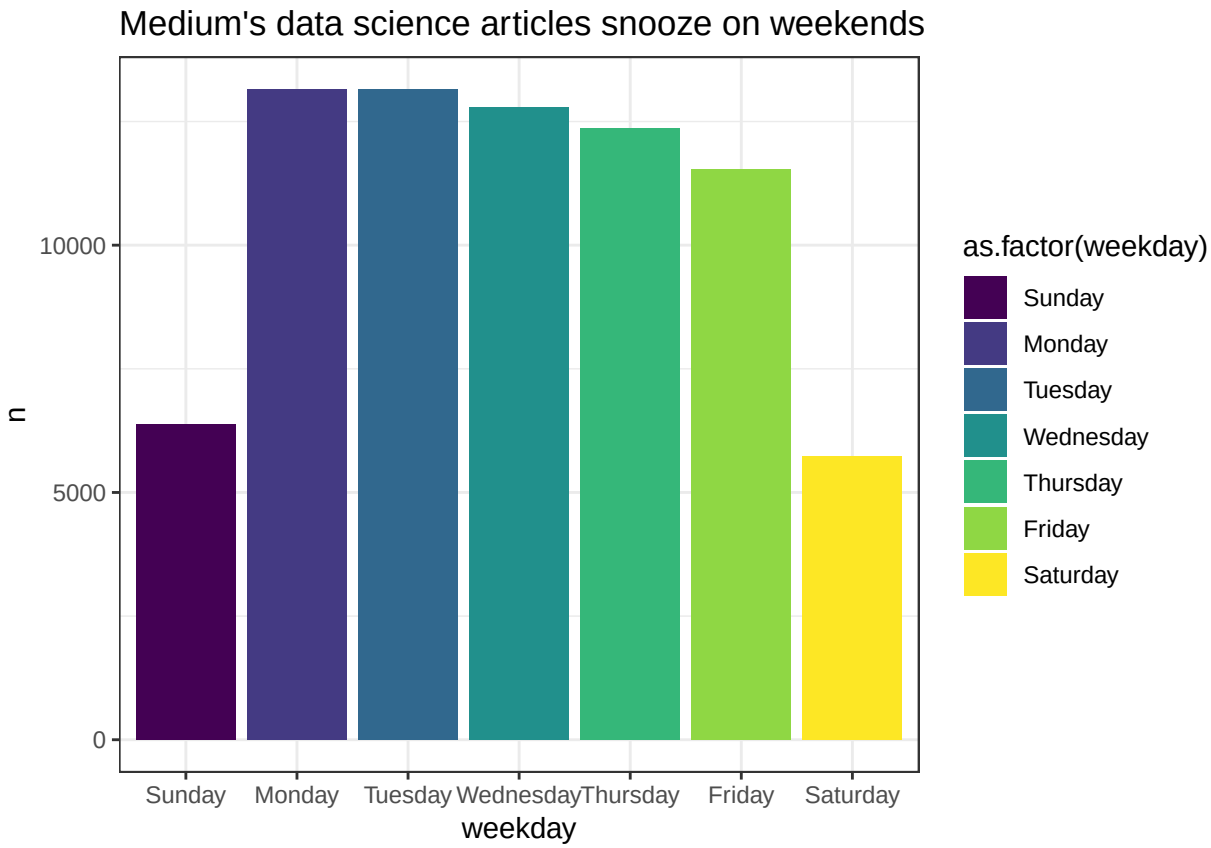
```
# ANALYSIS ----------------------------------------------------------------#
metadata %>%
  ggplot(mapping = aes(x = date)) +
  geom_line(stat = "bin", bins = 25) +
  theme_bw() +
  ggtitle(label = "Medium's data science article publication output doubled in 1 year")
```



Overall, the number of articles per day has doubled in a year.

Is there a day of the week on which people tend to publish?

```
metadata %>%
  add_count(as.factor(weekday)) %>%
  distinct(weekday, n) %>%
  ggplot(mapping = aes(x = weekday, y = n, fill = as.factor(weekday))) +
  theme_bw() +
  geom_bar(stat = "identity") +
  ggtitle(label = "Medium's data science articles snooze on weekends")
```

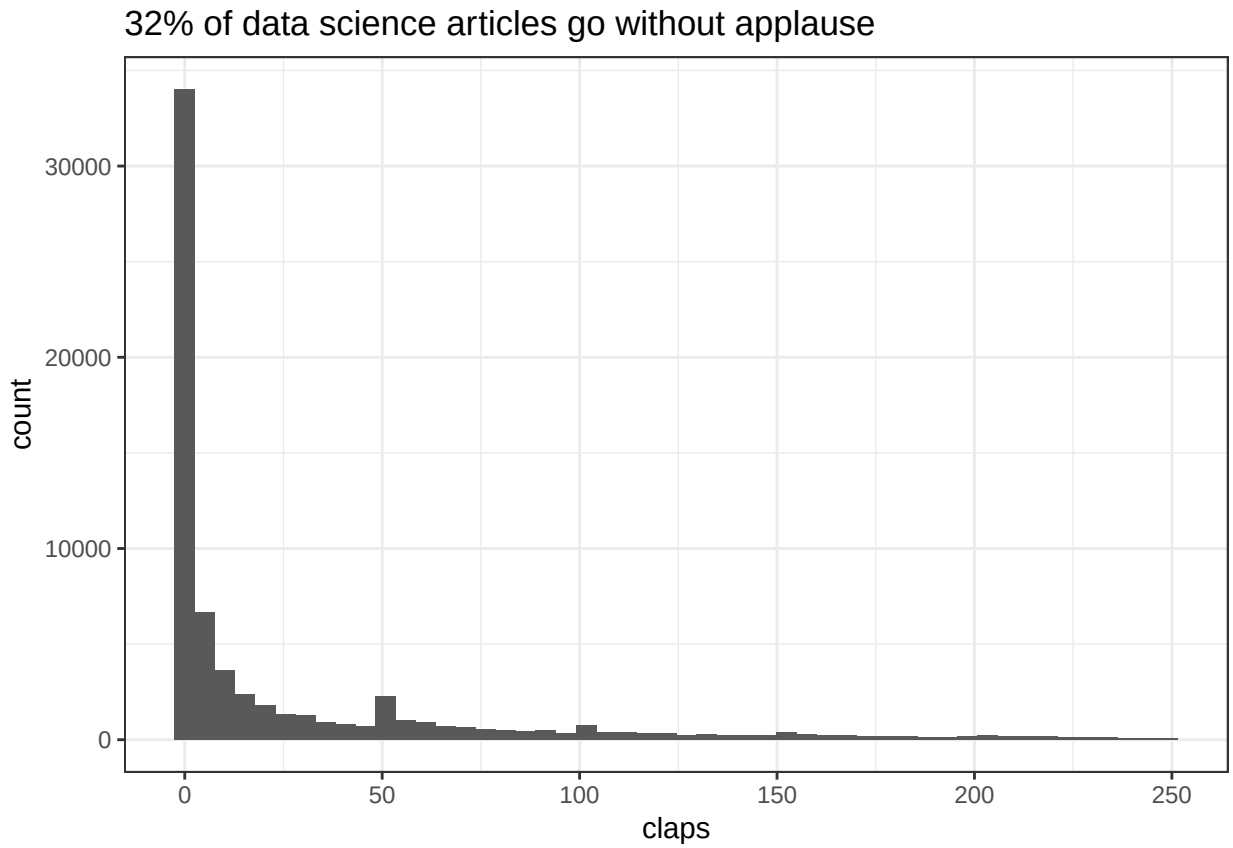


Monday's take a slight lead, with the weekends showing a huge dip in publications.

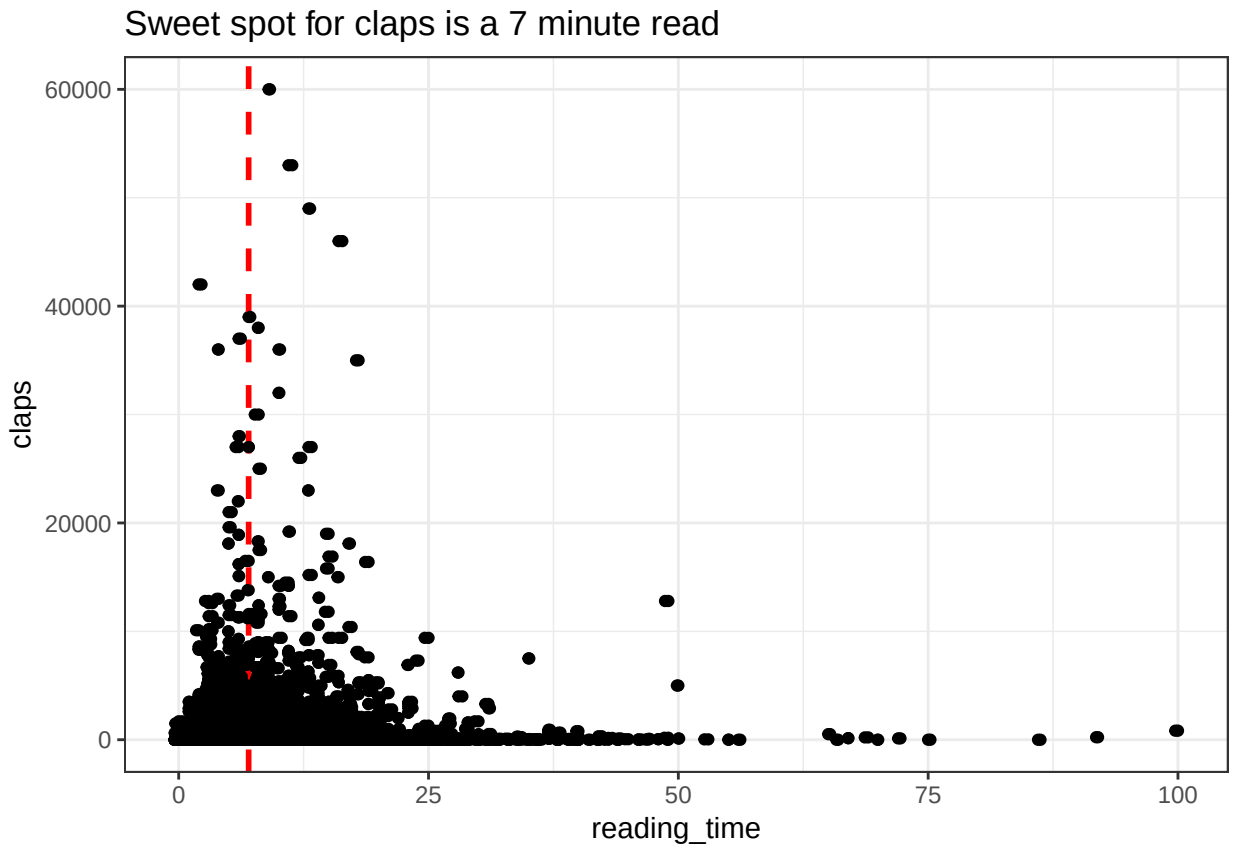
Let's get a feel for the distribution of claps per article.

```
no_clap_percentage <- round(100 * sum(metadata$claps == 0)/nrow(metadata), 0)

metadata %>%
  filter(claps < 250) %>%
    ggplot(mapping = aes(x = claps)) +
    theme_bw() +
    geom_histogram(bins = 50) +
    ggtitle(label = paste(no_clap_percentage, "% of data science articles go without applause", sep = "
```



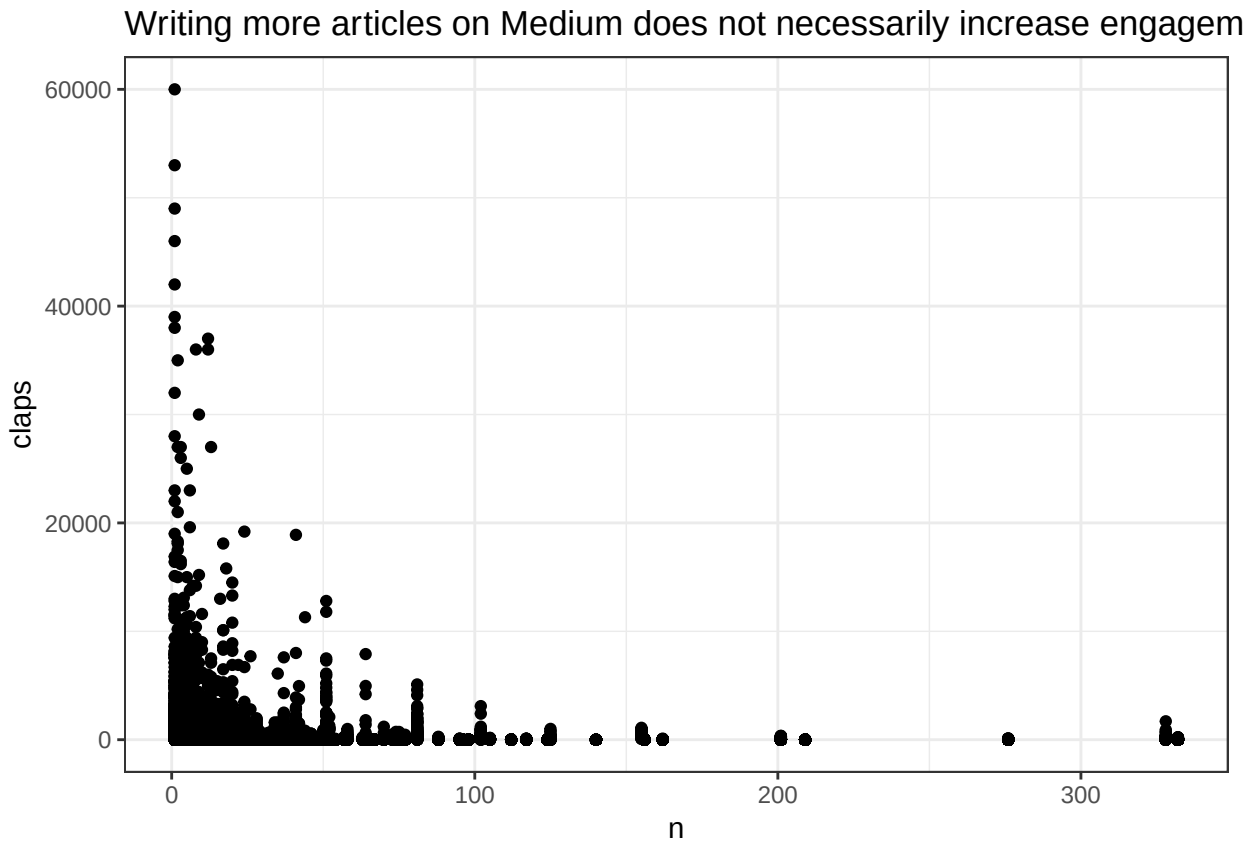

```
metadata %>%
    ggplot(mapping = aes(x = reading_time, y = claps)) +
    geom_point(na.rm = TRUE) +
    geom_vline(xintercept = 7,
               linetype = "dashed",
               color = "red",
               size = 1) +
    theme_bw() +
    geom_jitter() +
    ggtitle(label = "Sweet spot for claps is a 7 minute read")
```



I'm eye balling the read time for the above plot.

I hypothesize that more prolific authors are (1) better writers and (2) write interesting content and therefore have more applause. Is this supported by the data?

```
metadata %>%
  add_count(as.factor(author)) %>%
    ggplot(mapping = aes(y = claps, x = n)) +
    geom_point() +
    theme_bw() +
    scale_colour_gradient(low = "white", high = "red") +
    ggtitle(label = "Writing more articles on Medium does not necessarily increase engagement")
```

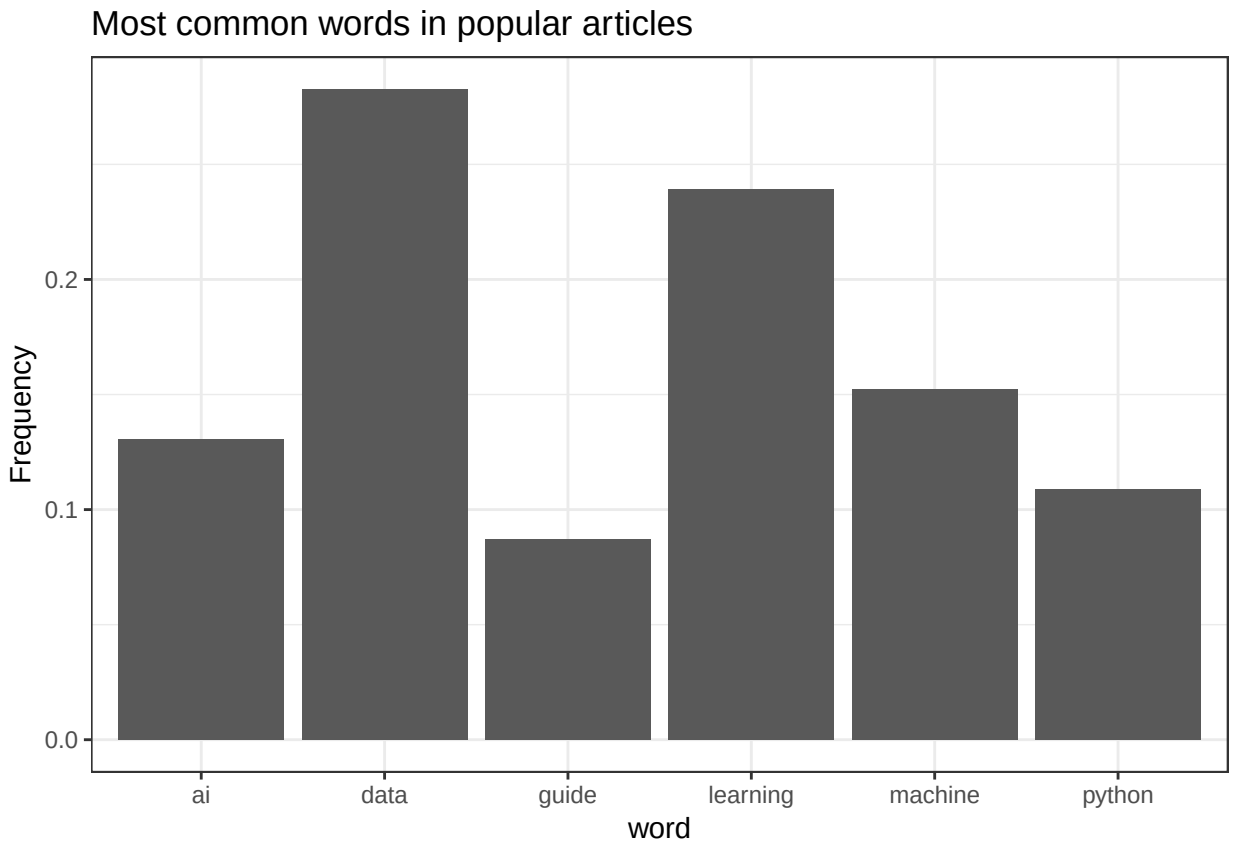


Nope! You don't have to write a lot of articles in the field to have a viral article. Perhaps highly prolific authors may be diluting their impact or prioritizing quantity over quality.

Let's investigate what's different about the topics of the most popular and least popular articles.

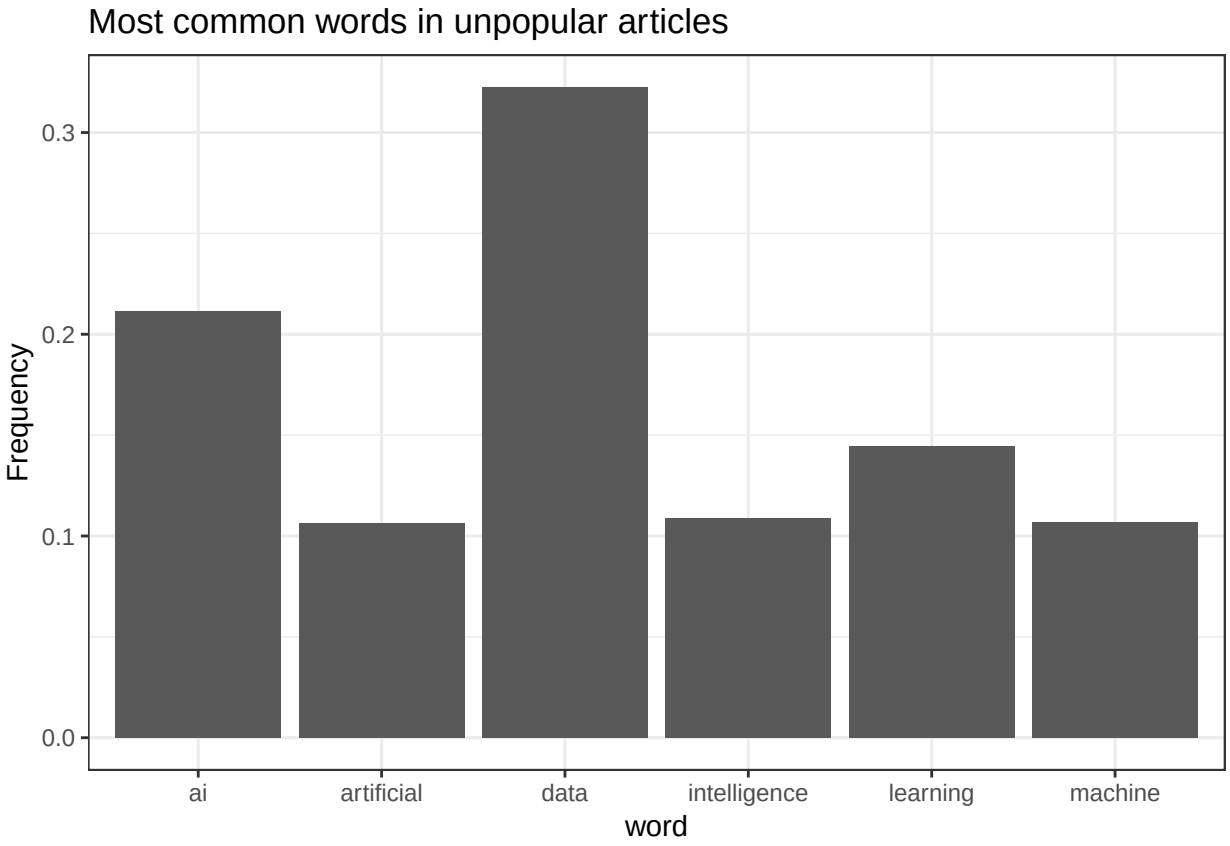
```
metadata %>%
  filter(claps > 10000) %>%
  unnest_tokens(word, title) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE) %>%
  filter(n > 3) %>%
  ggplot(mapping = aes(x = word, y = n/sum(n))) +
  theme_bw() +
  geom_bar(stat = "identity") +
  ylab("Frequency") +
  ggtitle(label = "Most common words in popular articles")
```

```
## Joining, by = "word"
```




```
metadata %>%
  filter(claps < 10) %>%
  unnest_tokens(word, title) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE) %>%
  filter(n > 1500) %>%
  ggplot(mapping = aes(x = word, y = n/sum(n))) +
  theme_bw() +
  geom_bar(stat = "identity") +
  ylab("Frequency") +
  ggtitle(label = paste("Most common words in unpopular articles"))
```

```
## Joining, by = "word"
```



Wow, almost no difference. Maybe how-to articles ("guide") have broad audiences because they are for amateurs?
Perhaps some of the difference in popularity is not because unpopular articles exclude buzz words, but that they include words that drive away views.

Learning moments

- TidyTuesday is great fun! I want to continue using this challenge to improve my speed & confidence using the tidyverse grammar and tools.
- I used lubridate, distinct(), and add_cout() for the first time today.