Yelp Reviews EDA

Jensen Harvey

2026-01-30

Load Data

```
yelp <- read_csv("Yelp Restaurant Reviews.csv")
```

```
## Rows: 19896 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Yelp URL, Date, Review Text
## dbl (1): Rating
##
## 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.
```

```
yelp <- yelp %>%
  clean_names() %>%
  rename(
    url = yelp_url,
    text = review_text
  )


head(yelp)
```

```
## # A tibble: 6 x 4
##   url                                                rating date      text
##   <chr>                                               <dbl> <chr>     <chr>
## 1 https://www.yelp.com/biz/sidney-dairy-barn-sidney       5 1/22/2022 "All I can~
## 2 https://www.yelp.com/biz/sidney-dairy-barn-sidney       4 6/26/2022 "Nice litt~
## 3 https://www.yelp.com/biz/sidney-dairy-barn-sidney       5 8/7/2021  "A delicio~
## 4 https://www.yelp.com/biz/sidney-dairy-barn-sidney       4 7/28/2016 "This was ~
## 5 https://www.yelp.com/biz/sidney-dairy-barn-sidney       5 6/23/2015 "This is o~
## 6 https://www.yelp.com/biz/sidney-dairy-barn-sidney       5 5/1/2019  "I've been~
```

```
str(yelp)
```

```
## spc_tbl_ [19,896 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ url   : chr [1:19896] "https://www.yelp.com/biz/sidney-dairy-barn-sidney" "https://www.yelp.com/bi
##  $ rating: num [1:19896] 5 4 5 4 5 5 1 5 5 2 ...
```

```
##  $ date  : chr [1:19896] "1/22/2022" "6/26/2022" "8/7/2021" "7/28/2016" ...
##  $ text  : chr [1:19896] "All I can say is they have very good ice cream I would for sure recommend
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Yelp URL` = col_character(),
##   ..   Rating = col_double(),
##   ..   Date = col_character(),
##   ..   `Review Text` = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(yelp)
```

```
##      url                rating          date               text
##  Length:19896       Min.   :1.000   Length:19896       Length:19896
##  Class :character   1st Qu.:4.000   Class :character   Class :character
##  Mode  :character   Median :5.000   Mode  :character   Mode  :character
##                     Mean   :4.131
##                     3rd Qu.:5.000
##                     Max.   :5.000
```

Check Missing Values

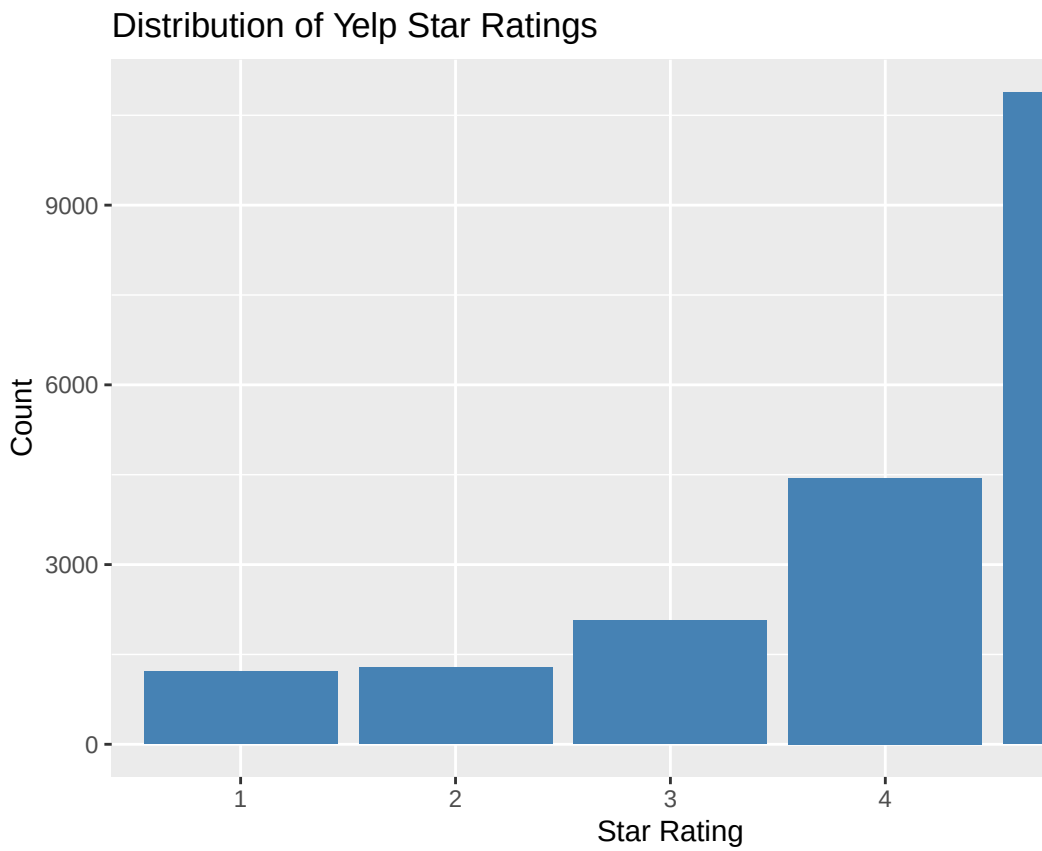
```
colSums(is.na(yelp))
```

```
##    url rating   date   text
##      0      0      0      0
```

```
class(yelp$rating)
```

```
## [1] "numeric"
```

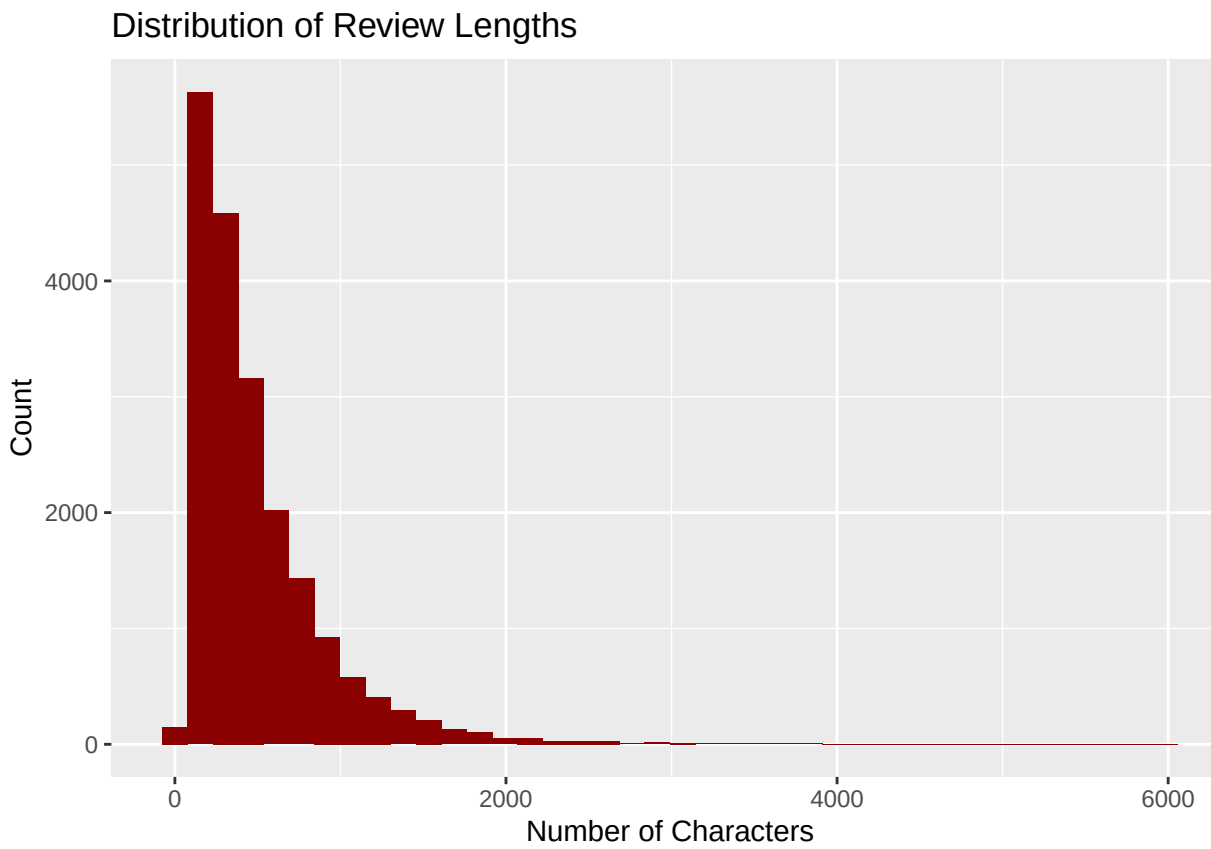
```
ggplot(yelp, aes(x = factor(rating))) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution of Yelp Star Ratings",
       x = "Star Rating",
       y = "Count")
```



Distribution of Star Ratings

```
yelp <- yelp %>%
  mutate(review_length = nchar(text))

ggplot(yelp, aes(review_length)) +
  geom_histogram(bins = 40, fill = "darkred") +
  labs(title = "Distribution of Review Lengths",
       x = "Number of Characters",
       y = "Count")
```



Review Length

```
tidy_reviews <- yelp %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

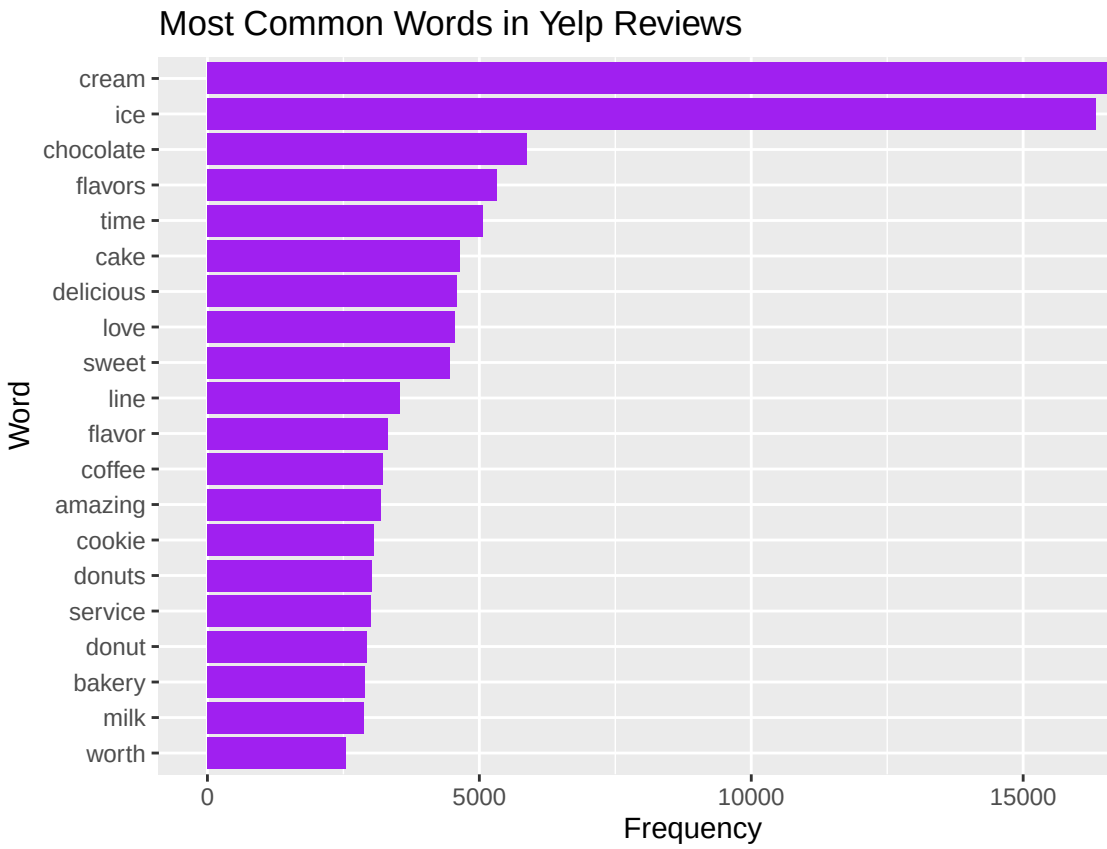
Text pre-processing Preview

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

```
head(tidy_reviews)
```

```
## # A tibble: 6 x 5
##   url                                              rating date  review_length word
##   <chr>                                             <dbl> <chr>          <int> <chr>
## 1 https://www.yelp.com/biz/sidney-dairy-barn-s~      5 1/22~            123 ice
## 2 https://www.yelp.com/biz/sidney-dairy-barn-s~      5 1/22~            123 cream
## 3 https://www.yelp.com/biz/sidney-dairy-barn-s~      5 1/22~            123 reco~
## 4 https://www.yelp.com/biz/sidney-dairy-barn-s~      5 1/22~            123 cook~
## 5 https://www.yelp.com/biz/sidney-dairy-barn-s~      5 1/22~            123 creme
## 6 https://www.yelp.com/biz/sidney-dairy-barn-s~      5 1/22~            123 ice
```

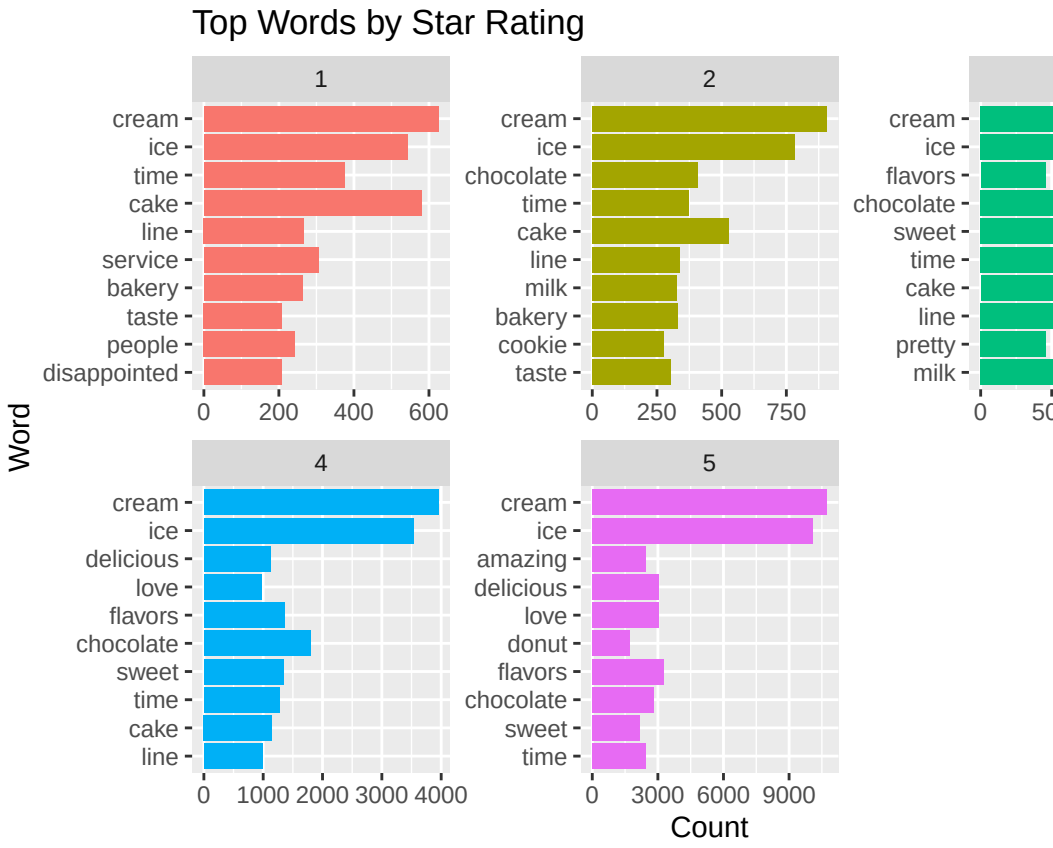
```
tidy_reviews %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20) %>%
  ggplot(aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "purple") +
  coord_flip() +
  labs(title = "Most Common Words in Yelp Reviews",
       x = "Word",
       y = "Frequency")
```



More commmon words

```
tidy_reviews %>%
  count(rating, word, sort = TRUE) %>%
  group_by(rating) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  ggplot(aes(x = reorder(word, n), y = n, fill = factor(rating))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ rating, scales = "free") +
  coord_flip() +
  labs(title = "Top Words by Star Rating",
```

```
       x = "Word",
       y = "Count")
```



Word Frequency by Rating

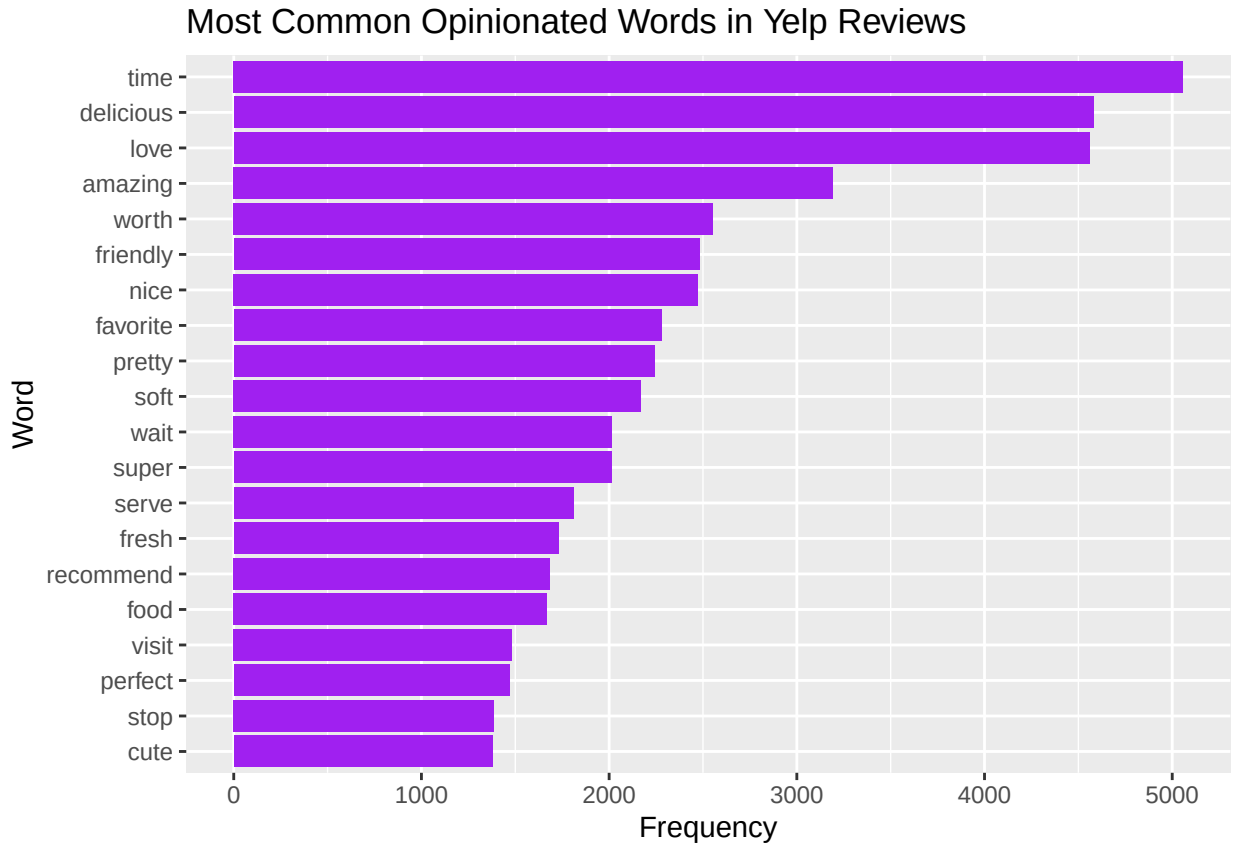
Get rid of Dessert Names:

```
food_stopwords <- tibble(word = c(
  "ice", "cream", "chocolate", "vanilla", "cookie", "cookies",
  "cake", "cakes", "flavor", "flavors", "bakery", "donut", "donuts",
  "milk", "frosting", "icing", "dessert", "desserts", "sweet", "sweets"
))


sentiment_words <- bind_rows(
  get_sentiments("bing"),
  get_sentiments("afinn"),
  get_sentiments("nrc")
) %>%
  distinct(word)


opinionated_reviews <- tidy_reviews %>%
  anti_join(food_stopwords, by = "word") %>%
  semi_join(sentiment_words, by = "word")
```

```
opinionated_reviews %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20) %>%
  ggplot(aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "purple") +
  coord_flip() +
  labs(title = "Most Common Opinionated Words in Yelp Reviews",
       x = "Word",
       y = "Frequency")
```



Segmented by Star Rating

```
opinionated_reviews %>%
  count(rating, word, sort = TRUE) %>%
  group_by(rating) %>%
  slice_max(n, n = 10) %>%
  ungroup() %>%
  ggplot(aes(x = reorder_within(word, n, rating), y = n, fill = factor(rating))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ rating, scales = "free") +
  scale_x_reordered() +
  coord_flip() +
```

```
labs(title = "Top Opinionated Words by Star Rating",
     x = "Word",
     y = "Count")
```