
What is
happening?

SETTING THE EXPECTATIONS

Day 1

- ✓ Understanding the tidyverse
- ✓ Regular expressions
- ✓ Organizing unstructured text
- ✓ Frequency analysis

Day 2

- Sentiment analysis
- Word association
- Topic modeling
- Predictive modeling

organize & clean ⇢ ***describe & predict***

WARM-UPS

Enough chit-chatting, time to code!

PREREQUISITES

```
## packages we'll use
library(tidyverse)
library(tidytext)

## data we'll use
airbnb <- read_rds("data/airbnb.rds")
```

EXERCISE 1

What is the most common name in the host_name column?

SOLUTIONS

```
airbnb %>%
  select(host_name) %>%
  mutate(host_name = str_to_lower(host_name)) %>%
  count(host_name, sort = TRUE)
# A tibble: 1,334 x 2
   host_name     n
   <chr>     <int>
 1 kara        138
 2 seamless     79
 3 mike         71
 4 flatbook     58
 5 alicia       50
 6 marie        42
 7 jason        35
```

EXERCISE 2

Filter out all observations that advocate for no shoes in their house_rules

SOLUTIONS

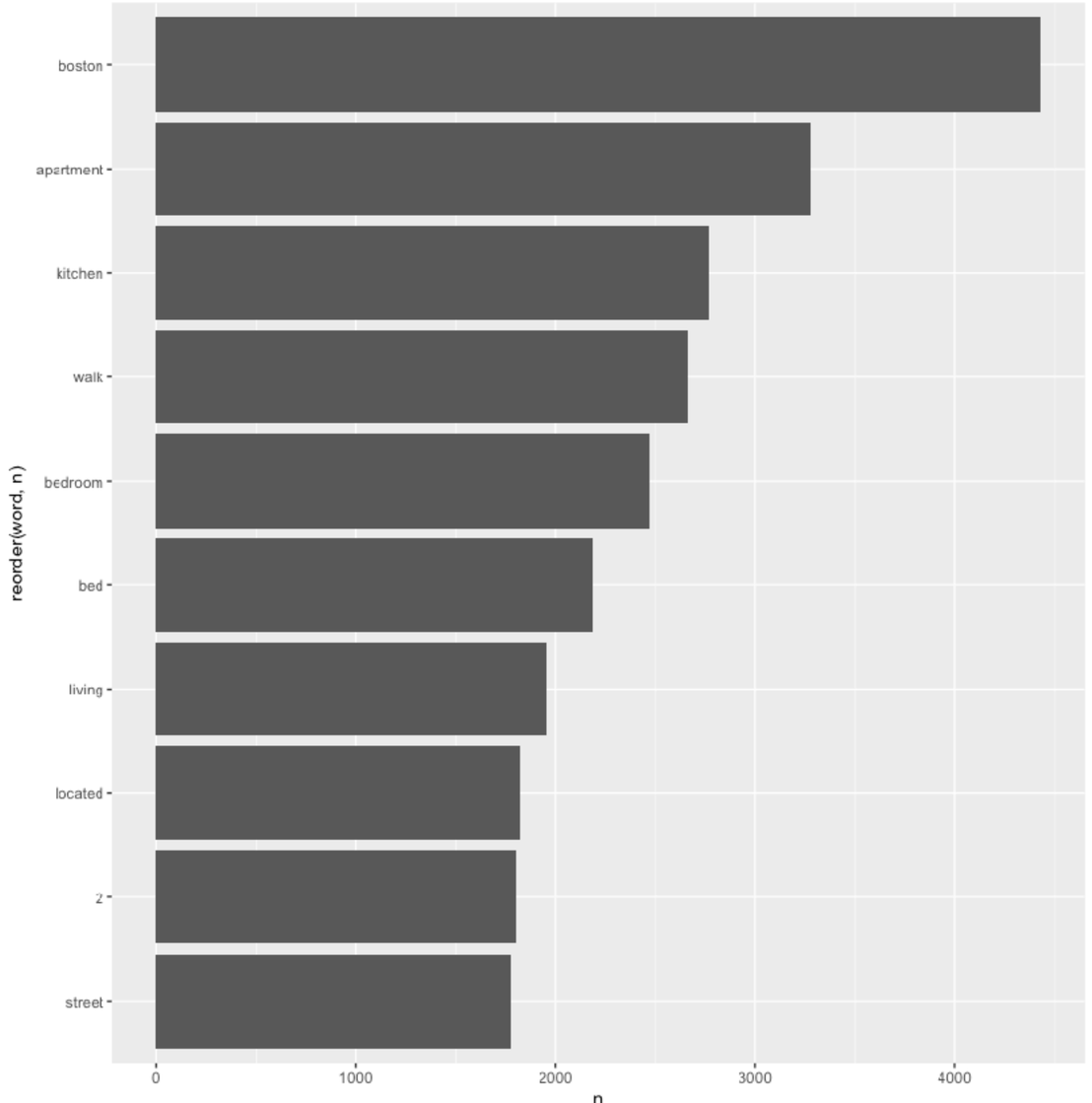
```
airbnb %>%
  filter(!str_detect(house_rules, regex("no shoes", ignore_case = TRUE)))
# A tibble: 2,326 x 95
        id listing_url  scrape_id last_scraped name  summary  space    description    experiences_off…
     <int> <chr>            <dbl> <date>       <chr> <chr>    <chr>    <chr>          <chr>
 1  1.21e⁷ https://www…   2.02e¹³ 2016-09-07   Sunn… Cozy, s… The hou… Cozy, sunny, … none
 2  3.08e⁶ https://www…   2.02e¹³ 2016-09-07   Char… Charmin… Small b… Charming and … none
 3  6.98e³ https://www…   2.02e¹³ 2016-09-07   Mexi… Come st… "Come s… "Come stay wi… none
 4  1.44e⁶ https://www…   2.02e¹³ 2016-09-07   Spac… Come ex… Most pl… Come experien… none
 5  7.65e⁶ https://www…   2.02e¹³ 2016-09-07   Come… My comf… Clean, … "My comfy, cl… none
 6  1.24e⁷ https://www…   2.02e¹³ 2016-09-07   Priv… Super c… Our sun… Super comfy b… none
 7  2.84e⁶ https://www…   2.02e¹³ 2016-09-07   "\"T… We can … "We pro… We can accomm… none
 8  7.53e⁵ https://www…   2.02e¹³ 2016-09-07   6 mi… Nice an… Nice an… Nice and cozy… none
 9  8.49e⁵ https://www…   2.02e¹³ 2016-09-07   Perf… "This i… Perfect… "This is a co… none
10  1.67e⁶ https://www…   2.02e¹³ 2016-09-07   Room… Quiet s… NA       Quiet second … none
# ... with 2,316 more rows, and 86 more variables: neighborhood_overview <chr>, notes <chr>,
#   transit <chr>, access <chr>, interaction <chr>, house_rules <chr>, thumbnail_url <chr>,
#   medium_url <chr>, picture_url <chr>, xl_picture_url <chr>, host_id <int>, host_url <chr>,
```

EXERCISE 3

Find and plot the top 10 most commonly used words in the description field

SOLUTIONS

```
airbnb %>%
  select(id, description) %>%
  unnest_tokens(word, description) %>%
  anti_join(stop_words) %>%
  count(word) %>%
  top_n(10) %>%
  ggplot(aes(reorder(word, n), n)) +
  geom_col() +
  coord_flip()
```

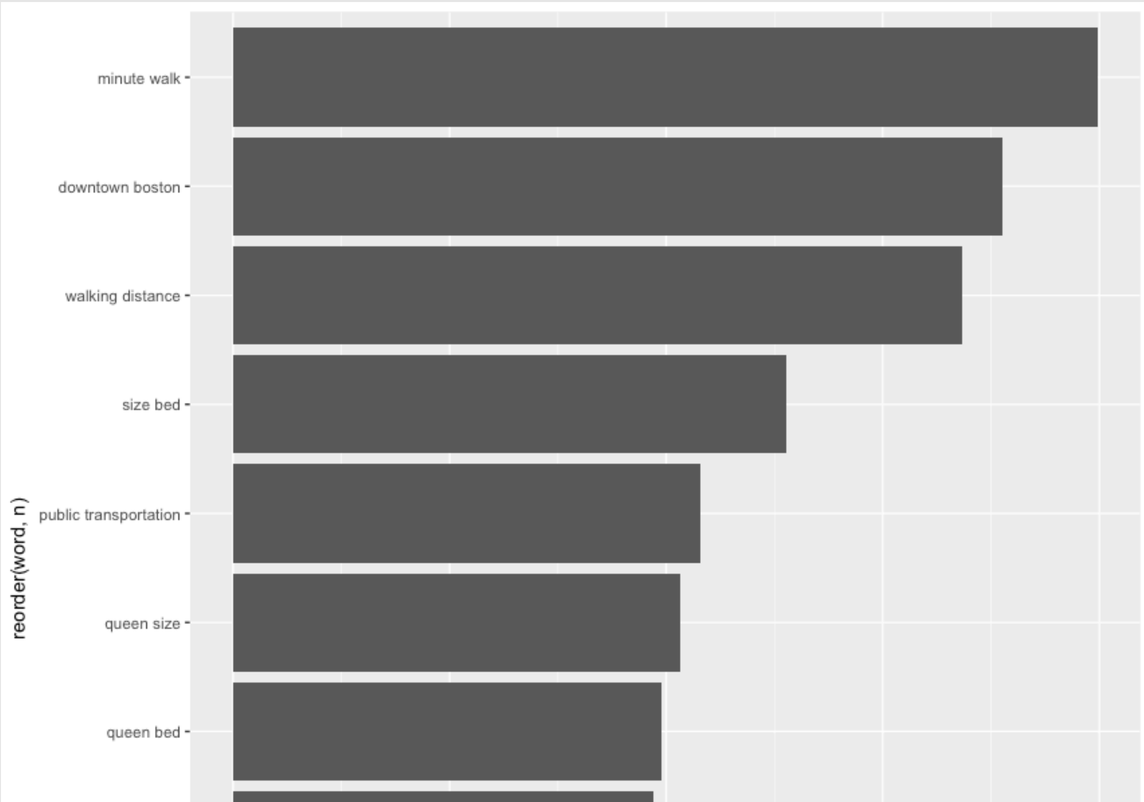


EXERCISE 4

Find and plot the top 10 most commonly used bi-grams in the description field

SOLUTIONS

```
airbnb %>%
  select(id, description) %>%
  unnest_tokens(word, description, token = "ngrams", n = 2) %>%
  separate(word, into = c("word1", "word2"), sep = " ") %>%
  filter(
    !word1 %in% stop_words$word,
    !word2 %in% stop_words$word
  ) %>%
  unite(word, word1, word2, sep = " ") %>%
  count(word) %>%
  top_n(10) %>%
  ggplot(aes(reorder(word, n), n)) +
  geom_col() +
  coord_flip()
```




LET'S GET STARTED!