Factors Driving Short-Term Rental Review Sentiment Across Price Ranges
The Predictive Pioneers (group 2): Christine Tsai, Ben Harris, Vance Newsome (leader)
DS 4002: Data Science Project Course, 02/07/25

Data Appendix

Airbnb Dataset: each observation in this dataset represents a customer review of an Airbnb property in one of the 17 cities we analyzed. We took a random sample of reviews from each city, resulting in 12661 total observations.

Variables:

- listing_id: a unique identifier for each Airbnb property listing (integer).
    - Constructed from `id` variable in raw listings data
    - No missing observations

- review_id: a unique identifier for each customer review (integer).
    - Constructed from `id` variable in raw reviews data
    - No missing observations

- comments: the text of a review left after the customer's stay (text).
    - From review data
    - No missing observations

- price: the cost per night of the Airbnb property in United States Dollars (integer).
    - From listing data, cleaned to remove dollar signs and commas
    - No missing observations
    - Summary statistics:
        - Minimum: 10
        - 1st Quartile: 88
        - Median: 120
        - Mean: 158.2
        - 3rd Quartile: 175
        - Maximum: 10000

- city: in which city the Airbnb property is located. There are 17 possible cities (text).
    - Created for each batch of data by location
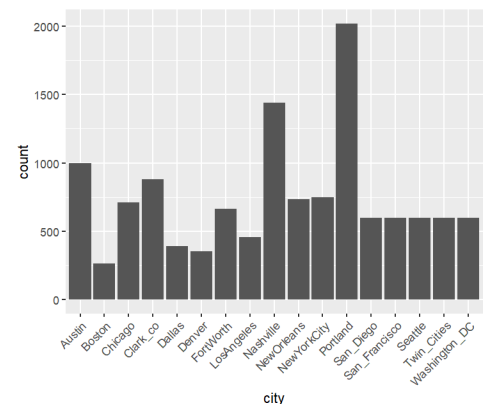   - No missing observations
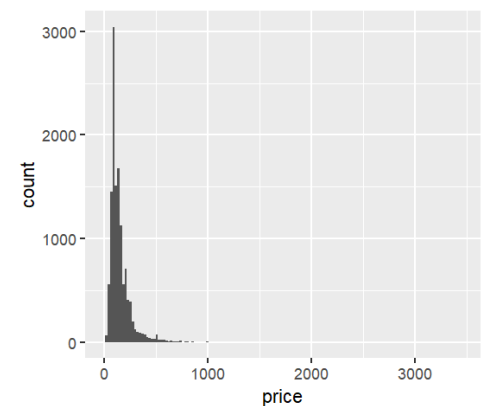
City:	Washington_DC	Twin_Cities	Seattle	San_Francisco
Frequency:	600	600	600	600

City:	San_Diego	Portland	NewYorkCity	NewOrleans
Frequency:	600	2018	751	736

City:	Nashville	LosAngeles	FortWorth	Denver
Frequency:	1440	457	664	355

| City:	Dallas	Clark_co	Chicago	Boston	Austin
Frequency:	390	881	711	263	1000

- <u>cleanliness:</u> a binary variable with value 1 for observations in which `cleanliness_sent_compound` has a value and 0 in which `cleanliness_sent_compound` is "NA".
  - ○ Created from `cleanliness_sent_compound`
  - ○ No missing observations

- <u>safety:</u> a binary variable with value 1 for observations in which `safety_sent_compound` has a value and 0 in which `safety_sent_compound` is "NA".
  - ○ Created from `safety_sent_compound`
  - ○ No missing observations

- <u>location:</u> a binary variable with value 1 for observations in which `location_sent_compound` has a value and 0 in which `location_sent_compound` is "NA".
  - ○ Created from `location_sent_compound`
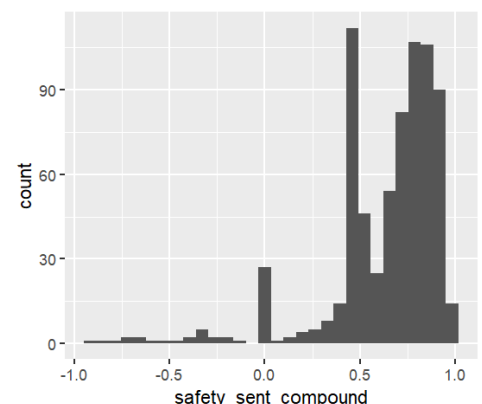 - ○ No missing observations

- <u>cleanliness_sent_compound:</u> the normalized, weighted sum of the sentiment scores of words in reviews related to cleanliness of the Airbnb property. A score above 0 indicates positive sentiment and below 0 indicates negative sentiment.
  - ○ Created with VADER package, based on the `comments` variable
  - ○ Observations: 12661(8823)
  - ○ Summary statistics:
    - ■ Minimum: -0.920
    - ■ 1st Quartile: 0.458
    - ■ Median: 0.743
    - ■ Mean: 0.661
    - ■ 3rd Quartile: 0.848
    - ■ Maximum: 0.992
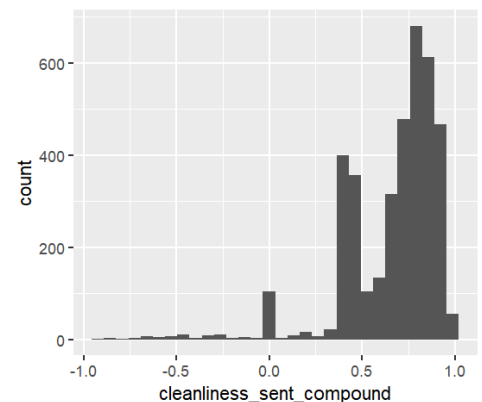
- <u>location_sent_compound:</u> the normalized, weighted sum of the sentiment scores of words in reviews related to location of the Airbnb property. A score above 0 indicates positive sentiment and below 0 indicates negative sentiment.
  - ○ Created with VADER package, based on the `comments` variable
  - ○ Observations: 12661(9670)
  - ○ Summary statistics:
    - ■ Minimum: -0.928
    - ■ 1st Quartile: 0.000
    - ■ Median: 0.572
    - ■ Mean: 0.460
    - ■ 3rd Quartile: 0.758
    - ■ Maximum: 0.991
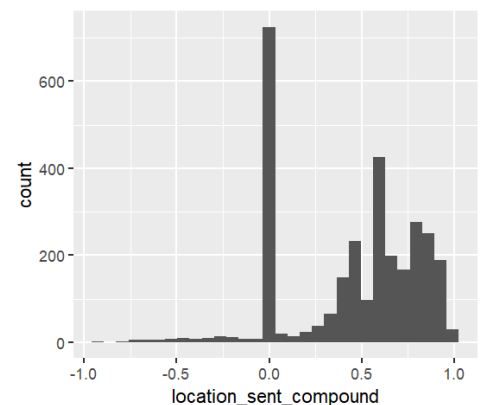
- <u>safety_sent_compound:</u> the normalized, weighted sum of the sentiment scores of words in reviews related to safety of the Airbnb property. A score above 0 indicates positive sentiment and below 0 indicates negative sentiment.
  - ○ Created with VADER package, based on the `comments` variable
  - ○ Observations: 12661(11947)
  - ○ Summary statistics:

- Minimum: -0.915
- 1st Quartile: 0.482
- Median: 0.710
- Mean: 0.632
- 3rd Quartile: 0.840
- Maximum: 0.990

- **overall_sent_compound:** the normalized, weighted sum of the sentiment scores of words in reviews of the Airbnb property. A score above 0 indicates positive sentiment and below 0 indicates negative sentiment.
  - Created with VADER package, based on the `comments` variable
  - Observations: 12661(0)
  - Summary statistics:
    - Minimum: -0.9912
    - 1st Quartile: 0.7767
    - Median: 0.9059
    - Mean: 0.8010
    - 3rd Quartile: 0.9587
    - Maximum: 0.9995