# Yelp Business

SHUBO LIN, TIANYUE LUO, NILAY VARSHNEY, HENGRUI QU

STAT 628 Module 3
Data Science Practicum FALL 21
19 November 2021

# Contents

1. General Workflow

2. Setting Goals and Data Preprocessing
   - Setting Goals
   - Data Cleaning

3. Attribute Analysis
   - Attribute Selection
   - Data Preprocessing
   - ANOVA

4. Review Analysis
   - Word Selection
   - Distribution of Specific Words

5. Preliminary Suggestions and Future Work
   - Preliminary Suggestions
   - Future Work

# Outline

1. General Workflow

2. Setting Goals and Data Preprocessing
   - Setting Goals
   - Data Cleaning

3. Attribute Analysis
   - Attribute Selection
   - Data Preprocessing
   - ANOVA

4. Review Analysis
   - Word Selection
   - Distribution of Specific Words

5. Preliminary Suggestions and Future Work
   - Preliminary Suggestions
   - Future Work

# General Workflow

- Setting Goals and Data Preprocessing
- Attribute Analysis
- Review Analysis
- Providing Preliminary Suggestions

# Outline

1. General Workflow

2. Setting Goals and Data Preprocessing
   - Setting Goals
   - Data Cleaning

3. Attribute Analysis
   - Attribute Selection
   - Data Preprocessing
   - ANOVA

4. Review Analysis
   - Word Selection
   - Distribution of Specific Words

5. Preliminary Suggestions and Future Work
   - Preliminary Suggestions
   - Future Work

# Setting Goals

Our group likes cafes, and we want to see what owners of cafes can do to improve their businesses based on Yelp reviews.

- What features are associated with successful cafes?
  - Type of coffee served?
  - Ambience?
  - Location and hours?
  - Price?
  - Presence of WiFi?
- What can cafe owners do to improve their ratings on Yelp?
  - Adding features
  - Modifying features

# General Data Cleaning

- **Original Data**

  5.36 million reviews and 155 thousand businesses

- **To get the information on cafes**

  1. Extract rows containing "cafe" in the "category" column in "business.json" and export them to a new file

  2. Extract the "business_id" column from the new file

  3. Merge "business.json" and "reviews.json" based on the "business_id" column

- **Final Data**

  303,769 reviews on 3,001 businesses.

# Outline


1. General Workflow
2. Setting Goals and Data Preprocessing
   - Setting Goals
   - Data Cleaning
3. Attribute Analysis
   - Attribute Selection
   - Data Preprocessing
   - ANOVA
4. Review Analysis
   - Word Selection
   - Distribution of Specific Words
5. Preliminary Suggestions and Future Work
   - Preliminary Suggestions
   - Future Work

# Attribute Selection

- We counted the attributes provided by the cafes using the "business.json" file. There are 35 attributes in total.
- However, not all businesses provide all 35 attributes. To avoid bias in our analysis, we chose the attributes that are provided by at least 80% of the businesses.

| Attributes | Count |
|---|---|
| RestaurantsTakeOut | 2820 |
| BusinessParking | 2744 |
| RestaurantsDelivery | 2697 |
| WiFi | 2660 |
| OutdoorSeating | 2659 |
| RestaurantsPriceRange2 | 2611 |
| RestaurantsReservations | 2488 |
| HasTV | 2462 |
| Ambience | 2405 |

# Data Preprocessing

- Remove rows that have no attributes (i.e. "attributes = NA")
- "BusinessParking" can be divided into five types, so we expanded the field to Garage/Street/Validated/Lot/Valet
- "Ambience" can be divided into nine types, so we expanded the field to Romantic/Intimate/Touristy/Hipster/Divey/Classy/Trendy/Upscale/Casual

# ANOVA

- We applied ANOVA to find the attributes that significantly affect the rating of the businesses.

| Source | DF | P-value |
|---|---|---|
| RestaurantsDelivery | 2 | 0.03271 * |
| OutdoorSeating | 2 | 8.60e-09 *** |
| RestaurantsPriceRange2 | 4 | 2e-16 *** |
| RestaurantsReservations | 2 | 0.03405 * |
| HasTV | 2 | 2e-16 *** |
| garage | 2 | 9.22e-09 *** |
| street | 2 | 2e-16 *** |
| lot | 2 | 8.43e-05 *** |
| valet | 2 | 0.00471 ** |
| intimate | 2 | 0.00928 ** |
| touristy | 2 | 0.04588 * |
| classy | 2 | 0.00864 ** |
| Residuals | 2955 | |

# Outline


1. General Workflow
2. Setting Goals and Data Preprocessing
   - Setting Goals
   - Data Cleaning
3. Attribute Analysis
   - Attribute Selection
   - Data Preprocessing
   - ANOVA
4. Review Analysis
   - Word Selection
   - Distribution of Specific Words
5. Preliminary Suggestions and Future Work
   - Preliminary Suggestions
   - Future Work

# Data Preprecessing

- Reformat the text in the "text" column of "review.json"
  1. Convert all letters to lowercase
  2. Delete all punctuation marks and numbers
  3. Unify word tenses (e.g. waiting = wait)
  4. Rewrite comparatives in their regular forms (e.g. worse = bad)

# Word Selection

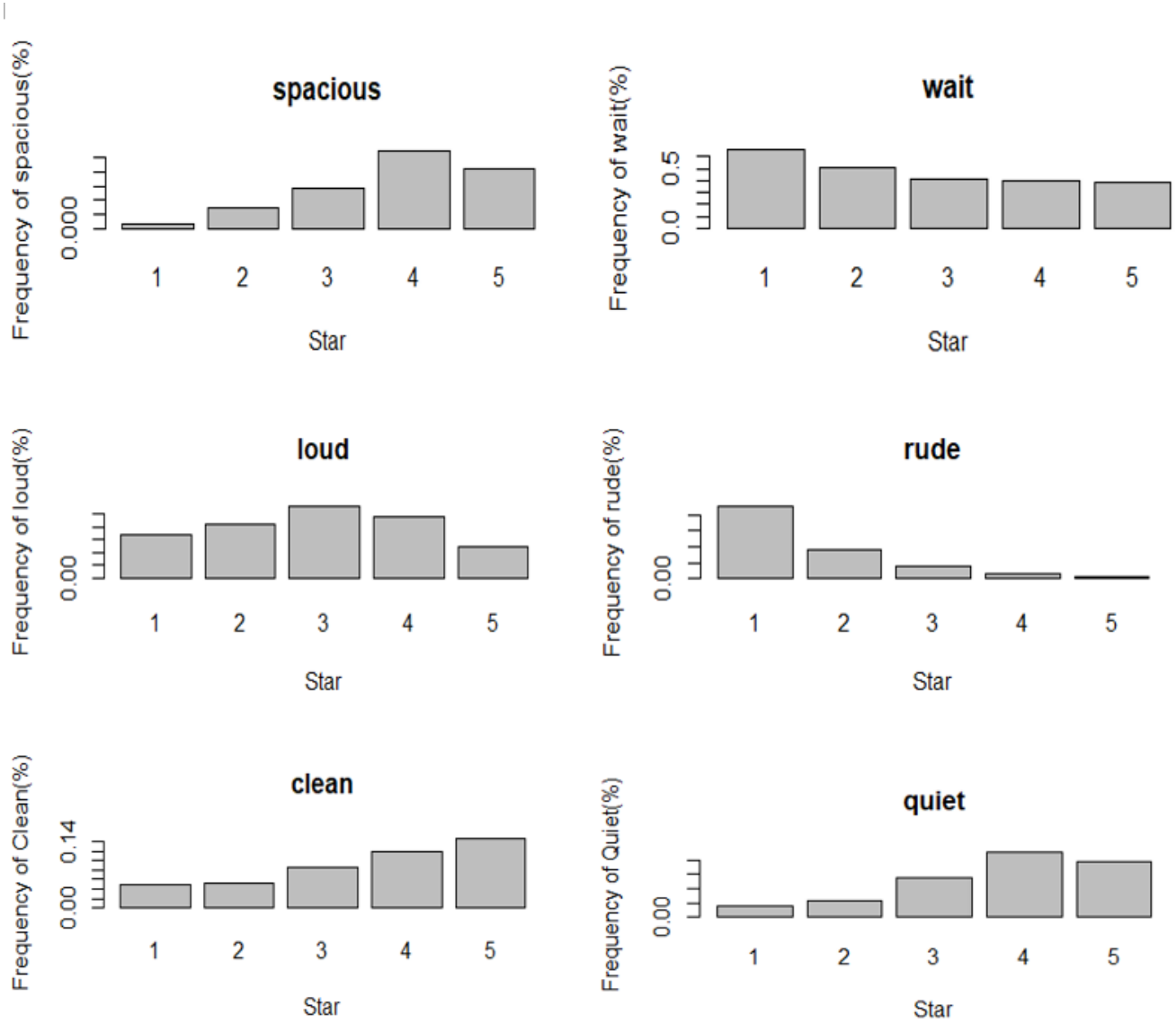
- We determined a equation that computes the frequency of a certain word across reviews with a certain number of stars.

$$\text{Frequency} = \frac{\text{Number of instances of the word}}{\text{Number of words across all reviews with same rating}}$$

- **Ambience**: clean, loud, quiet, rude, spacious, wait
- **Type of Food**: americano, brew, latte, espresso, tea, breakfast

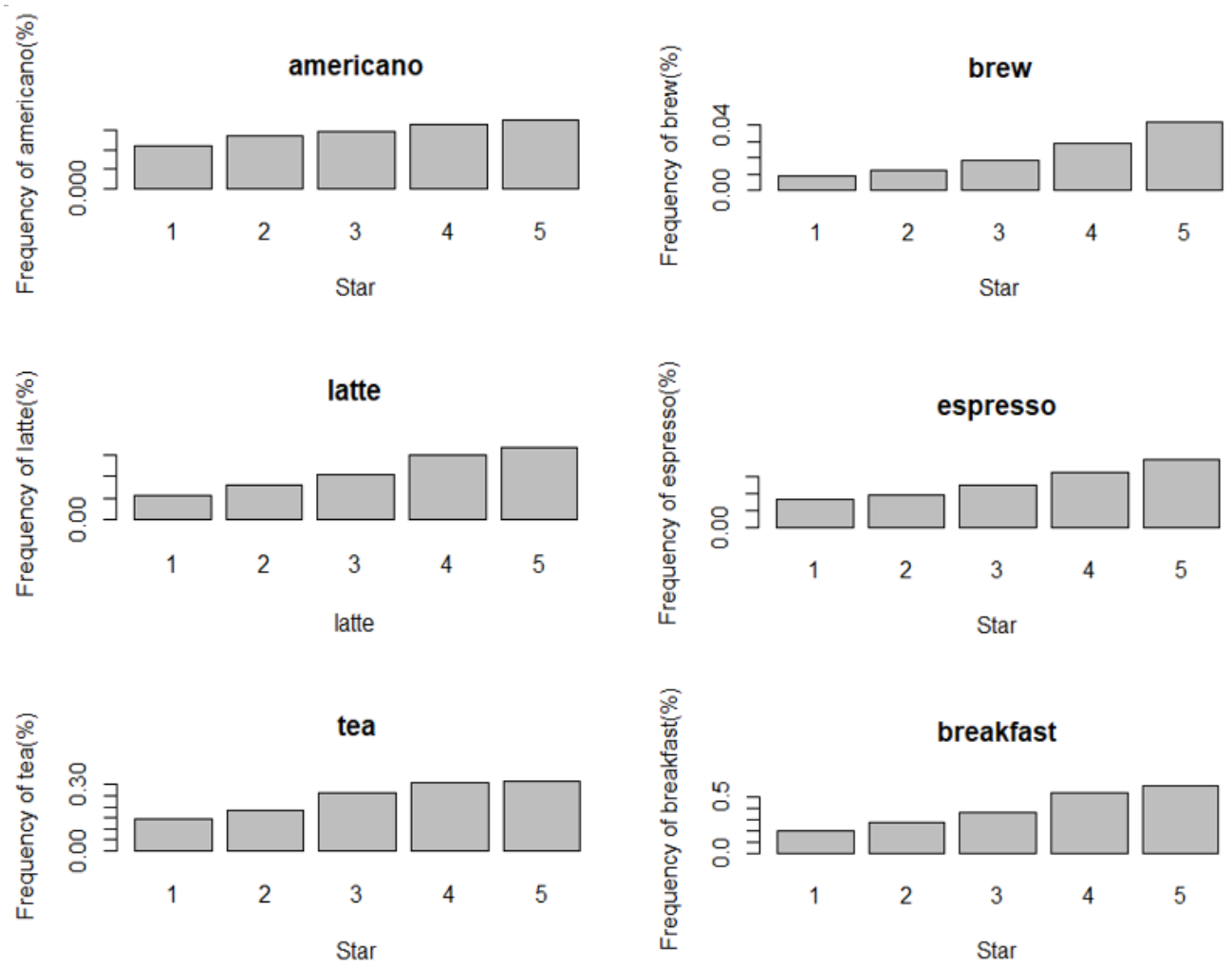
# Distribution of Specific Words

- **Ambience**



# Distribution of Specific Words

- **Type of Food**



# Outline

1. General Workflow

2. Setting Goals and Data Preprocessing
   - Setting Goals
   - Data Cleaning

3. Attribute Analysis
   - Attribute Selection
   - Data Preprocessing
   - ANOVA

4. Review Analysis
   - Word Selection
   - Distribution of Specific Words

5. Preliminary Suggestions and Future Work
   - Preliminary Suggestions
   - Future Work

# Preliminary Suggestions

**Attributes**

- Provide TV services
- Provide delivery and reservation services
- Provide outdoor seating and street/garage parking permits

**Specific Words in Reviews**

- Be spacious and clean
- Improve speed of service (e.g. use a fully automatic coffee machine)
- Add breakfast choices on the menu (e.g. sandwiches)

# Future Work

- Refine the review analysis and do sentiment analysis on key information words (e.g. counting the proportions of positive and negative nouns)
- Analyze the estimated treatment effects (treatment coefficients)
- Develop the Shiny App

# Thanks!