Using Yelp Analysis to Help Cafés

SHUBO LIN, TIANYUE LUO, NILAY VARSHNEY, HENGRUI QU

STAT 628 Module 3
Data Science Practicum FALL 21
9 December 2021

Overview

We want to see what owners of cafés can do to improve their businesses based on Yelp reviews.

- Providing suggestions based on specific business attributes:
  - Using ANOVA to find key attributes
  - Visualizing each attribute under different star ratings
- Providing suggestions based on reviews:
  - Word selection by tf-idf
  - Text analysis
- Using a Shiny app to communicate general and specific suggestions

Outline

Data Preprocessing and ANOVA

- Removed rows with no attributes (i.e. "attributes = NA")
- "BusinessParking" and "Ambience" can be divided into several types
- Applied ANOVA to find attributes that significantly affect ratings of 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	

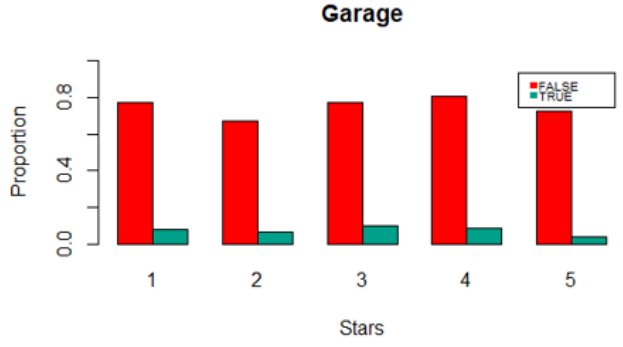
Bar Charts to Compare Scaled Ratings

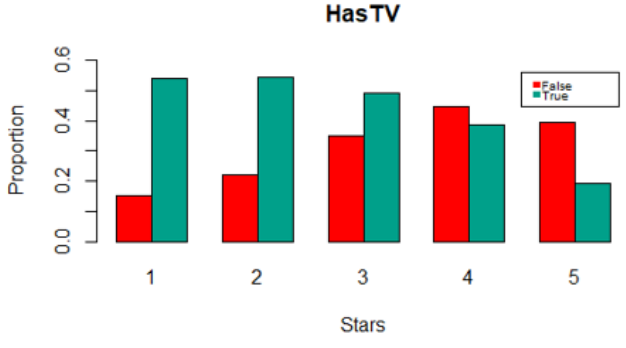
- Plotted each business attribute under different star ratings
- Visualized difference between different levels of each attribute
- Combined plots with t-tests to enhance results

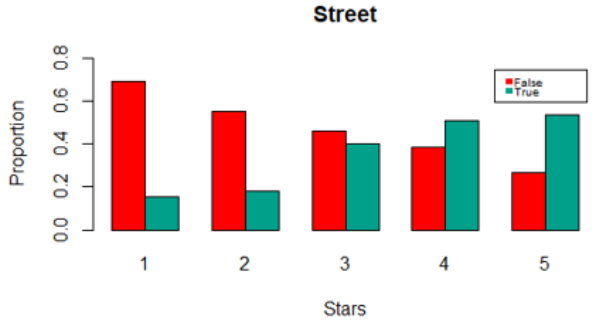
- Data Preprocessing
  - Grouped star ratings into five levels (1-5)
  - Excluded "NA" columns from charts to directly compare proportions of each level for each attribute
- Plot Information
  - X-axis is star rating
  - Y-axis is proportion of businesses under each star rating taking on a certain attribute value

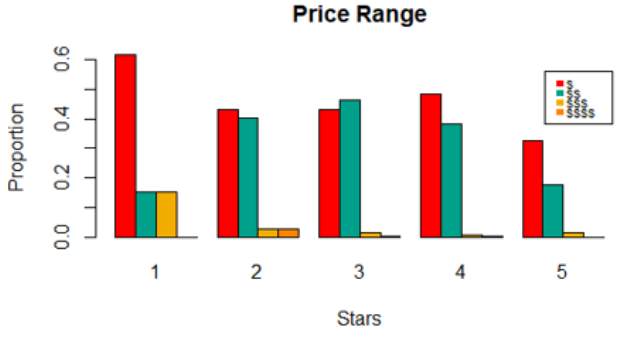
$$\text{Proportion(A, V, S)} = \frac{\#\text{ businesses where attribute A} = \text{V, star} = \text{S}}{\#\text{ businesses where star} = \text{S}}$$

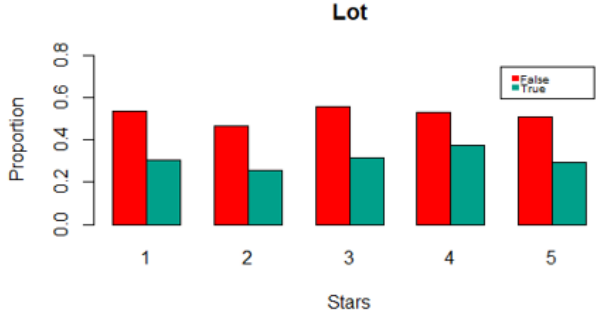
Bar Charts to Compare Scaled Ratings



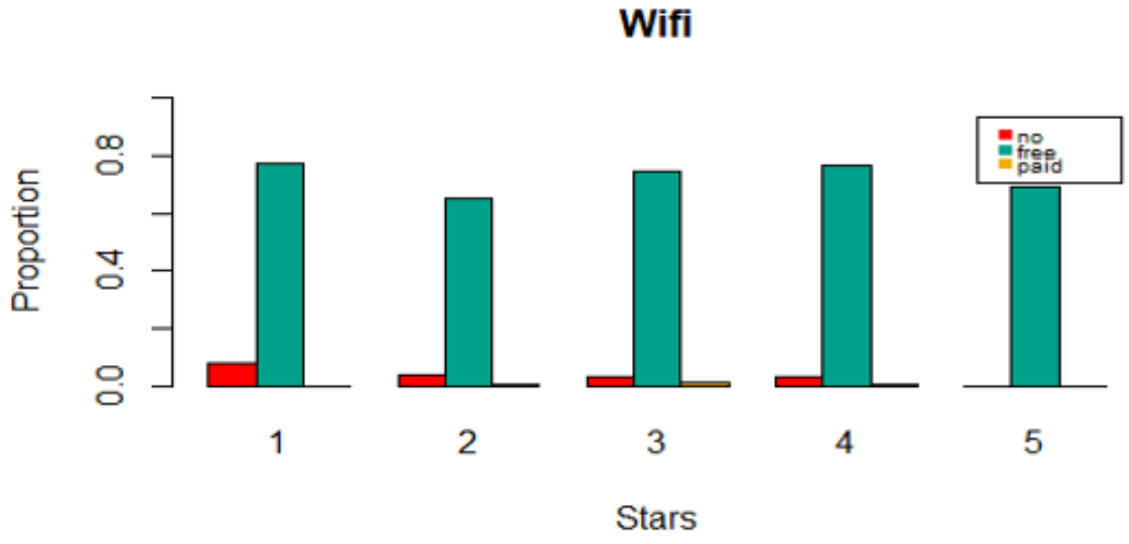


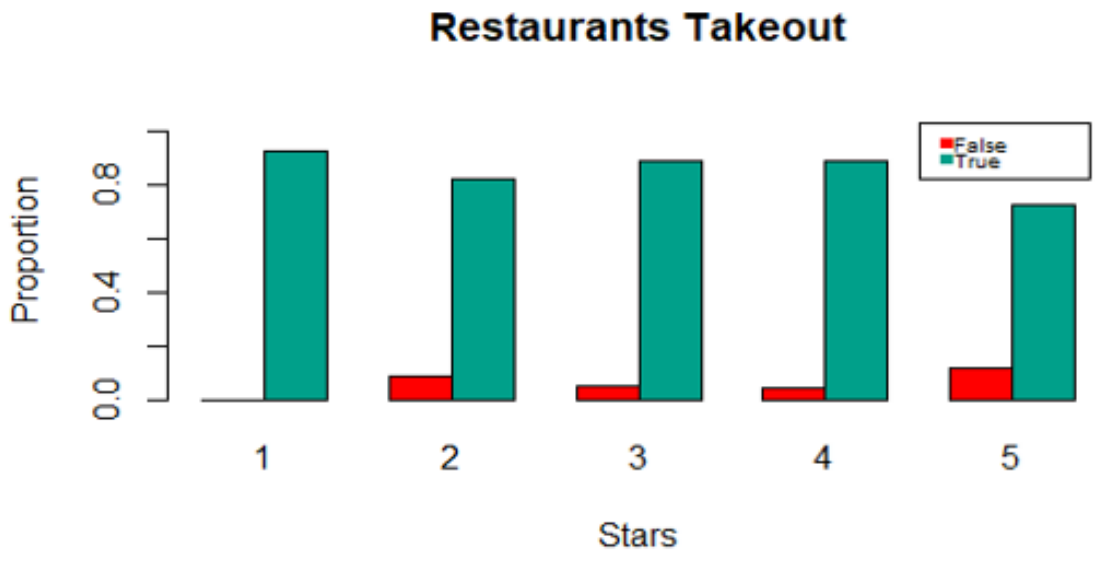






Bar Charts to Compare Scaled Ratings





Results of t-tests

Attribute	Levels	P-value
RestaurantsDelivery	True-False	0.00646 **
OutdoorSeating	True-False	1.342e-07 ***
HasTV	True-False	5.314e-16 ***
Garage	True-False	0.1597
Street	True-False	<2e-16 ***
Lot	True-False	0.06255 *
PriceRange	$-$$-$$$-$$$$	Only $-$$$$ >0.05
Valet	True-False	0.4359
Intimate	True-False	0.03744 *
Touristy	True-False	0.02825 *
Classy	True-False	1.05e-05 ***
RestaurantTakeout	True-False	0.7016
Wifi	No-Free-Paid	All p.value>0.05

Outline

Word Selection

We used tf-idf to capture the three most frequent nontrivial nouns in the reviews for each business.

- **tf**: term frequency of the word in each individual review
- **idf**: inverse document frequency — logarithm of fraction of reviews for each specific business containing the word
- **trivial words** include stop words and words not related to cafés

As a result, we had up to three nouns for each business based on the reviews and did further analysis on the selected words.

Text Analysis

- Conducted t-test between the average star rating of all reviews with the three informative words of each business and the average star rating of all reviews of the business
- Tested performance of business on informative words, to determine what the business needs to maintain or improve

Key word	Star rating with key word	Star rating of all reviews	p-value
Breakfast	3.81	3.73	0.630
Service	3.55	3.73	0.297
Lunch	<2	3.73	<2e-16

Outline

General Suggestions

- Providing delivery service can improve star ratings by about 0.07. (p-value = 0.00646)
- Providing outdoor seating can improve star ratings by about 0.15. (p-value = 1.342e-07)
- Removing TVs can improve star ratings by about 0.21.(p-value = 5.314e-16)
- Providing street parking (or any convenient parking place) can improve star ratings by about 0.22. (p-value <2.2e-16)

Specific Suggestions

We provided three types of suggestions for each business based on the selection of high-frequency words:

- Commendable aspect
- Aspect that needs to be improved
- Aspect that needs more stable quality

Shiny app: `https://nvarshney2.shinyapps.io/recommendations/`

Strengths and Weaknesses

Strengths

- Methods used were simple and easy to interpret
- Suggestions are practical
- Used both statistical analysis and plots to provide credible and readable results
- Set reasonable word list combined with tf-idf counting results and gave personalized suggestions
- Shiny App provides great user experience

Weaknesses

- Did not study the "NA" category in business attributes
- Did not establish correlation between attitude toward certain words and star ratings of reviews
- Used star ratings in reviews to directly represent attitude of reviews

Thanks!