Yelp Project Executive Summary

Chao-Sheng Wu, Zixuan Zhao, Hongyan Xiao

I. Introduction

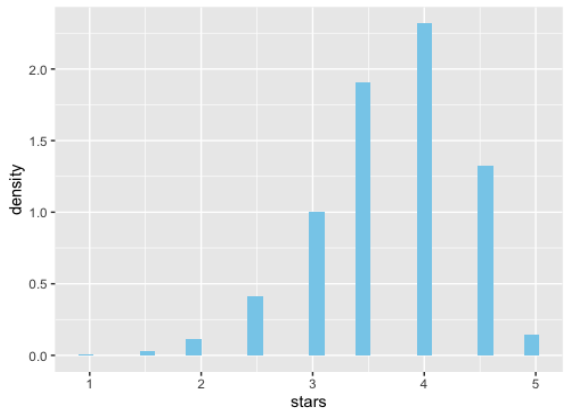
In this project, we aim to provide some suggestions for sushi restaurants based on the Yelp dataset. There are two sections of analysis, attributes analysis and review analysis. In the first section, we used Logistic Regression to see which attributes influence star ratings. In another part, we conducted some techniques to extract important words in terms of ratings and service. Finally, we summarize the above analysis to give our stakeholders a few lines of general suggestions and acknowledge the limitations of our report.

II. Data Cleaning

1. Data Selection

We filtered the restaurants from business.json by categories with SUSHI or JAPANESE and got 2459 restaurants. Based on their business ID, we extracted the reviews of these restaurants.

2. Data Distribution



This is the plot of rating distribution from cleaned "sushi.csv" data, which claims that most customers give 4 star and the next is 3.5 star, while 1 star and 5 star are the least, there are very few restaurants that make customers feel extremely satisfied or extremely unsatisfied.

III. Attributes Analysis

1. Data Processing

Firstly, we enumerated the attributes provided by cleaned "sushi.csv" data, and figured out that there are 39 attributes in total. Since not all restaurants have every attribute, we chose the following attributes that most restaurants provided for the analysis. In addition, besides 'True' and 'False', attributes also contain some 'None' values in attributes. So, we converted the 'None' as missing data (NA) to directly visualize and compare the distribution of each attribute in each rating level.

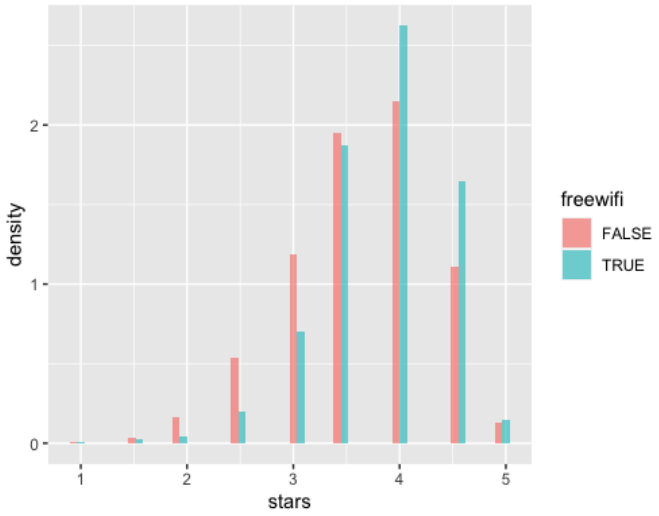
- Outdoor Seating (Service): True or False
- Restaurant Delivery (Service): True or False
- Wheelchair Accessible (Service: True or False
- Bike Parking (Service): True or False
- Noisy Level: Loud or Not loud
- Free WiFi (Service): True or False

- Alcohol (Service): True or False

2. Logistic Model

We classified restaurants into two categories based on their average star rating. A restaurant was classified as a good restaurant (numeric as 1) if its average rating was no less than 4.0, otherwise it was classified as a not good restaurant (numeric as 0). After that, we fitted a logistic regression model, tested the significant effect on the classification. From the logistic model with the significant level 0.01, we found the wheelchair accessible, noisy level, free wifi and alcohol are four significant effects for sushi restaurant ratings, which are highlighted in the table below.

Attributes	Coefficient	P-Value
Outdoor Seating (True)	-0.016	0.494713
Restaurant Delivery (True)	0.042	0.040161
Wheelchair Accessible (True)	**0.196**	**< 2e-16**
Bike Parking (Ture)	0.032	0.129031
Noisy Level (Loud)	**-0.305**	**1.74e-08**
Free WiFi (True)	**0.083**	**0.000167**
Alcohol (True)	**-0.111**	**1.21e-07**



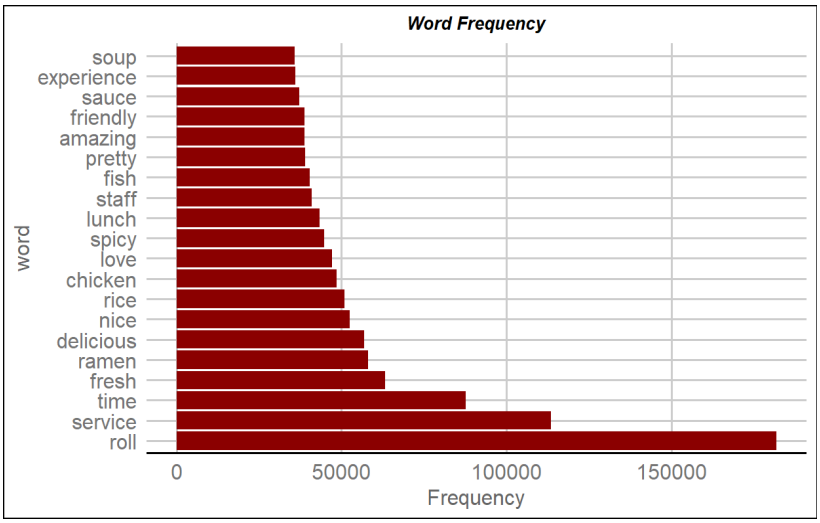
To visualize distributions of these four significant effects under different stars, we drew the four plots. They are very similar, so we show one plot as an example here. From the plot we can obviously see when the star rating greater than or equal to 4, which we identify them as 'good restaurants', the number of restaurants have free wifi are larger than not having free wifi restaurants. This shows that if a business can provide free wifi, customers would prefer to give higher a star rating.

IV. Text Mining

1. Text Cleaning

To explore the text reviews, we cleaned text by several methods. Firstly, we turned every text into lower case and removed all the punctuations and digits. Then we removed stopwords from the text. After this, we split the text into words and reduced prefixes and suffixes of words.

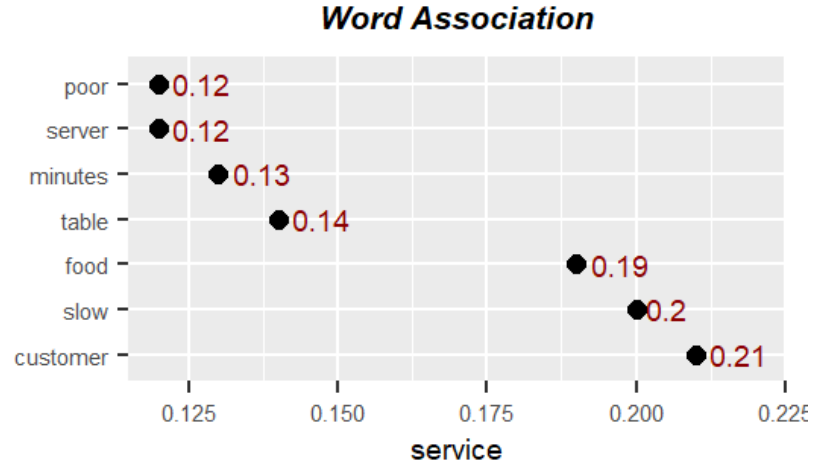
2. Count Word Frequency



After text cleaning, we counted the number of occurrences of each word and focused on the frequently occurring words. We found some words like SUSHI or RESTAURANT are frequent but useless, so we removed them from the list. The figure on the left is the final top 20 frequently occurring words.

3. Word Association

The math of findAssoc() is based on the standard function cor() in the stats package of R. Given two numeric vectors, cor() computes their covariance divided by both the standard deviations. So given a DocumentTermMatrix *dtm* containing terms "word1" and "word2" such that findAssocs(*dtm*, "word1", 0) returns "word2" with a value of x, which is the correlation of the term vectors for "word1" and "word2".



Service is an important aspect according to the word frequency plot, so we tried to find out which words are related to service. The following figure shows a few most correlated words to SERVICE.

V. Shiny App

You could play around with our app: https://cwu377.shinyapps.io/sushi_analysis/

VI. Recommendation

1. Suggestions

i. General Suggestions

Based on the result of the logistic regression model, we extracted several suggestions for attributes which could produce better star ratings for businesses. Firstly, restaurants with

wheelchair accessible service have e^0.196=1.22 times the odds of the restaurants without this service to get 4 stars or higher. Also, customers prefer to have a quiet dining environment. Therefore, a quieter restaurant could be about 1.36 times the odds of a noisy restaurant to get a larger-than-4 star. Similarly, providing the free wifi service and selling alcoholic beverages in restaurants have the chance to get about 1.09 and 1.12 times the odds of the restaurants without these two services to get 4 stars or higher. So, we suggest businesses to reduce the noise decibels to maintain a quieter dining environment, and provide other three services as well.

From the Word Frequency plot, we can find what customers are most concerned about is food. Words like ROLL, FRESH, which indicates that rolls are the most ordered food, a good quality of it is the most essential. Other words like CHICKEN, RAMEN and SOUP indicate that customers also like ordering these foods in sushi restaurants. It would be better to serve nice ramen and fried chicken to attract customers. Apart from this, word LUNCH shows that lunch might be the busiest time of the day.

From the Word Association plot, some negative words showed their close relationship to service, which means when customers mention service, they tend to complain, especially the slow serving speed. It suggests that improving serving speed is necessary.

ii. Specific Suggestions

We provide a few additional customized features for each restaurant in our shiny app, including: a. Important attributes based on III.2. b. Charts of star ratings/#reviews over year c. Potential competitors in the same town.

2. Weakness

We acknowledge our users that there are some limitations of our analysis.

i. Attributes Analysis

There are few drawbacks in our attributes analysis. In the first place, the model can only explain a small part of the variance according to r-squares. Secondly, attributes may be a bit correlated. For instance, restaurants that provide alcohol tend to be more loud.

ii. Text Analysis

To extract only useful words in our word frequency chart, we manually filtered out meaningless words. A better way would be to conduct a sentiment analysis, and only reserve words with negative and positive meaning. About the word association chart, it's hard to interpret without digging into the reviews body. Nevertheless, we don't have enough time to do the above work.

3. Strength

Despite these flaws in our analysis, there are still few worth mentioning parts. First, we used a simple model (logistic regression) in the attributes analysis, which is easy to use and interpret. Also, we provide business owners an intuitive way to navigate important words in our text analysis, which helps non-technical people understand it easily. Besides, the shiny App can provide a great user experience.

VII. Conclusion

In this project, we focused on the restaurants that have keywords 'sushi' and 'Japanese'. After using the logistic regression model to analyze the attributes of business average star rating, and text mining to analyze the restaurants' reviews, we suggest to our stakeholders what facilities to have, what dishes to provide based on our attribute and text analysis, respectively. By following our instructions, they can take specific actions to improve their business.

VIII. Contributions

Chao-Sheng Wu: Shiny App, Report and Slides
Zixuan Zhao: Data Cleaning, Text Mining, Report and Slides
Hongyan Xiao: Attributes Analysis, Report and Slides