Yelp Data Analysis - Recommendation for Chinese Restaurants

Introduction

Yelp Insight is designed to help business owners increase their business rating on Yelp with analytics. This project is mainly targeted to help Chinese foods restaurant owners. This executive report focuses on the following aspects

1. The position where a single business lies among all business of the same category. For example, where the rating and opening hours lie in the distribution of all restaurants.
2. Features of restaurants that improve the rating. For example, whether they should have more parking space, faster WIFI or take-out service.
3. Information from comments that may help improve the business.

Background

Data used for the analytics in this project come from Yelp, which is stored in JSON files for each of the businesses. The data are divided into the following four parts:

1. User reviews of each restaurant
2. Detailed information about the business.
3. User information
4. Tips from users of each business.

Data Processing

Restaurant Selection

As is stated above, this project focuses on Chinese restaurants, so we first select all the businesses with word *Chinese* in its description.

Attributes Analysis

In this part, we mainly use the attributes of each business. These are the steps of data processing:

1. Keep only *business_id*, *name*, *stars,* and *attributes*, and delete all other variables
2. Find all attributes of a business
3. For each attribute or feature, divide all businesses into two groups, yes group, which has this attribute and no group, which does not have this attribute

Exploratory Data Analysis

First, we investigate the distribution of review ratings.

Table 1 Distribution of Ratings

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
1.00	2.00	4.00	3.53	5.00	5.00

Then we obtain a graph showing the correlation between review length and rating as below. From left to right, the i^{th} point stands for the mean rating of reviews whose length is in the range $[50 * (i-1) + 1,\ 50 * i]$. And the number corresponding to it tells the exact amount of reviews with that length. From this graph, we could see there is a negative correlation between review length and ratings.

The number of reviews whose length exceed 400 words is relatively much smaller, so they do not have much impact on the regression line as the short reviews do.

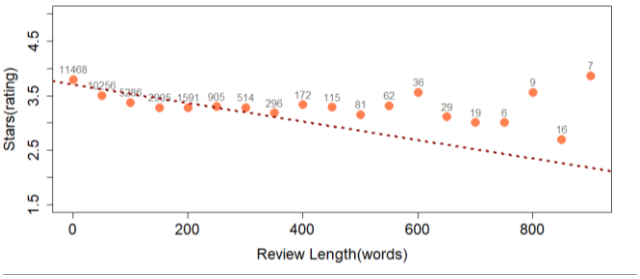

Figure 1 Correlation between Review Length and Ratings

Moreover, word frequency is important and analyzed. The following graph shows, as expected, positive words are positively correlated with ratings and negative words are negatively correlated with negative ratings.

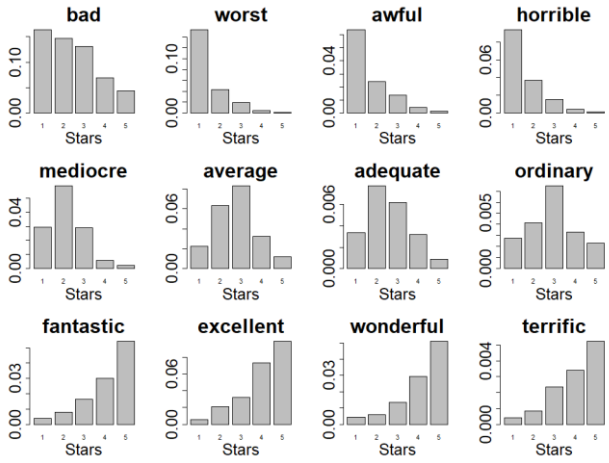

Figure 2 Distribution of Common Adjectives with Respect to Rating

Part 1. Key Findings of Chinese Restaurants

Influential Features/Attributes

We find that providing restaurants reservation, providing good kids environment, not providing food delivery and credit card acceptance increase the rating.

Table 2 Positive Attributes

	reserve.	kids	no deliv.	no card
p-val.	1.48×10^{-5}	0.025	0.023	0.014
mean diff.	0.24	0.24	0.12	0.32

Whereas, providing food take out, providing alcohol, having TV, and providing Wi-Fi do not impact the rating significantly.

Table 3 Neutral Attributes

	take out	alcohol	TV	Wi-Fi
p-val.	0.663	0.625	0.202	0.103
mean diff.	−0.04	−0.017	0.10	0.12

Tests

We use Wilcoxon-Rank test to see whether some attributes of business are related to ratings. The null hypothesis is that the attribute is not related to the ratings. We choose a significance level of 0.05. If the p-value is less than 0.05, we think the attribute is related to the ratings. We can also get the 95% confidence interval of the average rating difference between group:

Table 4 Confidence Interval of Positive Attributes

	Reserve.	kids	no deliv.	no card
lower	2.77e-05	1.46e-05	6.32e-05	4.19e-05
upper	5.00e-01	5.00e-01	6.63e-05	5.00e-01

Table 5 Confidence Interval of Neutral Attributes

	food take out	alcohol	TV	Wi-Fi
lower	-0.50006	-4.1e-05	-4.5e-05	-4.8e-05
upper	0.49996	1.84e-06	6.79e-05	6.55e-05

Check Assumptions

We checked the normality and variance homogeneity for each attribute group of data.
Firstly, we use the Shapiro-Wilk test to check the normality of data. The p-values are all less than 0.05, so with significant level 0.05, we think the data are not normal. In this situation, we cannot use t-tests, so we use nonparametric tests Wilcoxon-Rank test to see whether there is a relationship between some attributes and ratings.

Table 6 Wilcoxon-Rank Test of Positive Attributes

p-val.	reserve.	kids	no deliv.	no card
Yes group	5.2×10^{-9}	2.9×10^{-14}	6.2×10^{-9}	0.01
No group	4.1×10^{-11}	0.012	3.5×10^{-11}	4.9×10^{-15}

Table 7 Wilcoxon-Rank Test of Neutral Attributes

p-val.	take out	alcohol	TV	Wi-Fi
Yes group	2.82×10^{-15}	1.19×10^{-6}	2.37×10^{-13}	2.17×10^{-5}

No group	0.02189	2.52×10^{-11}	1.09×10^{-5}	1.39×10^{-12}

Secondly, we use F-test to check the variance homogeneity of the two groups of data. If the p-value is less than 0.05, we think with significant level 0.05, the variance is different between two groups. According to whether the data have variance homogeneity, we will use different type of Wilcoxon-Rank test.

Table 8 F-Test of Positive Attributes

	reserve.	**kids**	**no deliv.**	**no card**
p-val.	0.00165	0.03022	0.437	0.5552

Table 9 F-Test of Neutral Attributes

	take out	**alcohol**	**TV**	**Wi-Fi**
p-val.	0.5082	0.00055	0.01737	0.2726

Findings of Keywords in Reviews

To begin with, we obtain three graphs about the correlation between some keywords from comments and the rating. The X-axis is the number of stars of each review. And in each rating level, the Y-axis tells the proportion of comments with certain keywords. For example, in the first bar plot of the first graph, the rightmost bar shows about 0.6% of 5-star rating restaurants contain the keyword "bun" or "baozi".

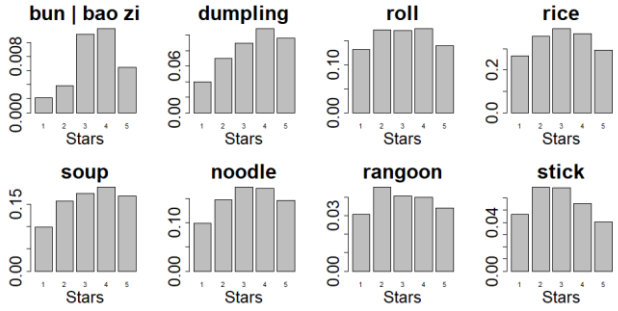

Figure 3 Distribution of Appearance of Main Dishes in Reviews with Respect to Rating

The first graph investigates the correlation between different main foods and review ratings. As we can see, bun (or baozi) and dumpling, as well as soup, have an obvious positive correlation with ratings. Meanwhile, Rangoon and sticks have a relative negative correlation with ratings.

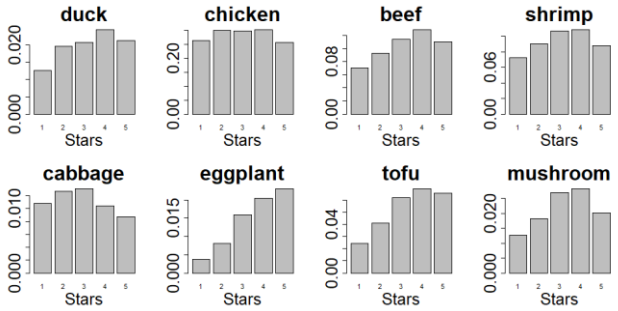

Figure 4 Distribution of Appearance of Other Dishes in Reviews with Respect to Rating

The second graph shows the correlation between different dishes and ratings. The first row consists of 4 meat dishes and the second is of 4 vegetable dishes.

A positive correlation with ratings is revealed when it comes to duck, beef, eggplant and tofu, while there exists a negative correlation for cabbage.

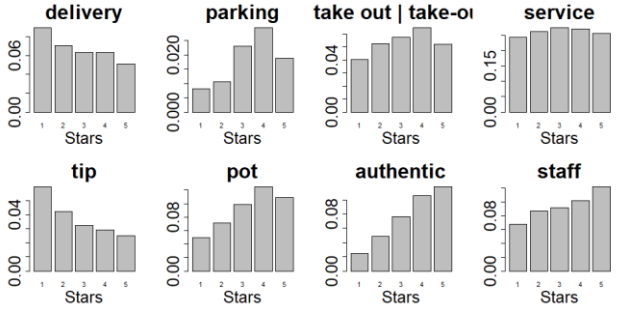

Figure 5 Distribution of Appearance of Other Keywords in Reviews with Respect to Rating

And the last graph focuses on some other interesting words. It seems "parking", "pot", "authentic" and "staff" are more relevant to high ratings and "delivery", "tip" are likely to lead to low ratings.

Part 2. Recommendations

Recommendations from Attributes

1. **Provide reservation service.**
   On average, a Chinese restaurant with reservation service has a 0.24-star higher rating, comparing to a Chinese restaurant without reservation service.
   (Wilcoxon-Rank test p-value:1.48×10^{-5}).
2. **Provide special food and toy for children.**
   On average, a Chinese restaurant which is good for children has a 0.24-star higher rating, comparing to a Chinese restaurant which is not good for children
   (Wilcoxon-Rank test p-value:0.026).
3. **Do not credit card payment.**
   On average, a Chinese restaurant which does not accept credit card has a 0.32-star higher rating, comparing to a Chinese restaurant which accepts credit card
   (Wilcoxon-Rank test p-value:0.014).
4. **It is not worth providing food delivery.**
   Although food delivery is related to the ratings, the average rating for food delivery Chinese restaurant is only 0.12 higher than no food delivery Chinese restaurant.
5. **It is not worth investing in food take out, alcohol, Wi-Fi, and TV.**
   These four attributes are not related to the ratings
   (Wilcoxon-Rank test p-value are all greater than 0.05).

Recommendations from Reviews

From word frequencies (in reviews) analysis in part 1 (4), we could also say:

1. For the main food, we encourage Chinese restaurant owners to expand the production of bun, dumplings, soup and consider carefully when it comes to Rangoon and sticks.
2. For dishes other than staple food, it seems duck, beef, eggplant and tofu are more appetizing for customers, especially the last two, while cabbage might not be a preference.
3. Also, parking space, food-in-pot, authentic meals, and behavior of staff are considerable characters for higher ratings. And delivery, as well as tips, might be a common issue for Chinese restaurants.

Conclusion

In brief, we have analyzed commercial data from Yelp to provide owners of Chinese restaurants with inspiring insights of their businesses.

First, all Chinese restaurants are selected for further analysis.

Second, each of the features of a restaurant are tested to uncover the relationship between rating and the feature. The result is that providing reservation service, being good for kids, having food delivery and not accepting credit card have positive influence on the rating, with other features having insignificant impact on rating.

Third, keywords from the customer reviews are investigated to find out best dishes. For meals, bun, dumplings, and soup will be most likely to boost the business, while Rangoon and sticks are not. For other dishes, duck, beef, eggplant and tofu are the best choices.

Contributions

Wenyi Wang:

1. Coding: Attribute Analysis
2. Report/Presentation: Attribute Analysis (Part 1), Recommendations from Attributes
3. Application: Suggestions

Hao Jiang:

1. Coding: Processing Data, Part of Comment Analysis
2. Report/Presentation: Introduction, Conclusion, Revision
3. Application: Development of App

Zixiang Xu:

1. Coding: Comment Analysis
2. Report/Presentation: Comment Analysis, Recommendations from Comments
3. Application: Suggestions