Report 628

# Introduction

Yelp Insight is designed to help business owners increasetheir business rating on Yelp with analytics. This project is mainly targeted forChinese/Asian foods restaurant owners. This executive report will focus on thefollowing aspects

·        Position where a single business lies among allbusiness of the same category. For example, where the rating and opening hours liein the distribution of all restaurants.

·        Features of restaurants that improve the rating.For example, whether they should have more parking space, faster WIFI or take-outservice.

·        Information from comments that may help improvethe business. For example, if any of the dishes is too spicy.

(short thesis)

# Background

Data used for the analytics in this project comes from Yelp,which is stored in JSON files for each of the businesses. The data are dividedinto 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 focus on Chineserestaurants, so we first select all the businesses with word Chinese in itsdescription.

### Attributes Analysis

In this part, we mainly use the attributes part of business data:

1. delete columns we do not use, just keep business\_id, name, stars and attributes
2. find all possible attributes
3. for each attributes, divide the data in two group, yes group(has this attribute) and no group(does not have this attribute)

### Comment Analysis

To analyze comments, there are several steps involved.

1.      Divide business in to two groups, high-rating (rating> 3) and low-rating (rating < 3).

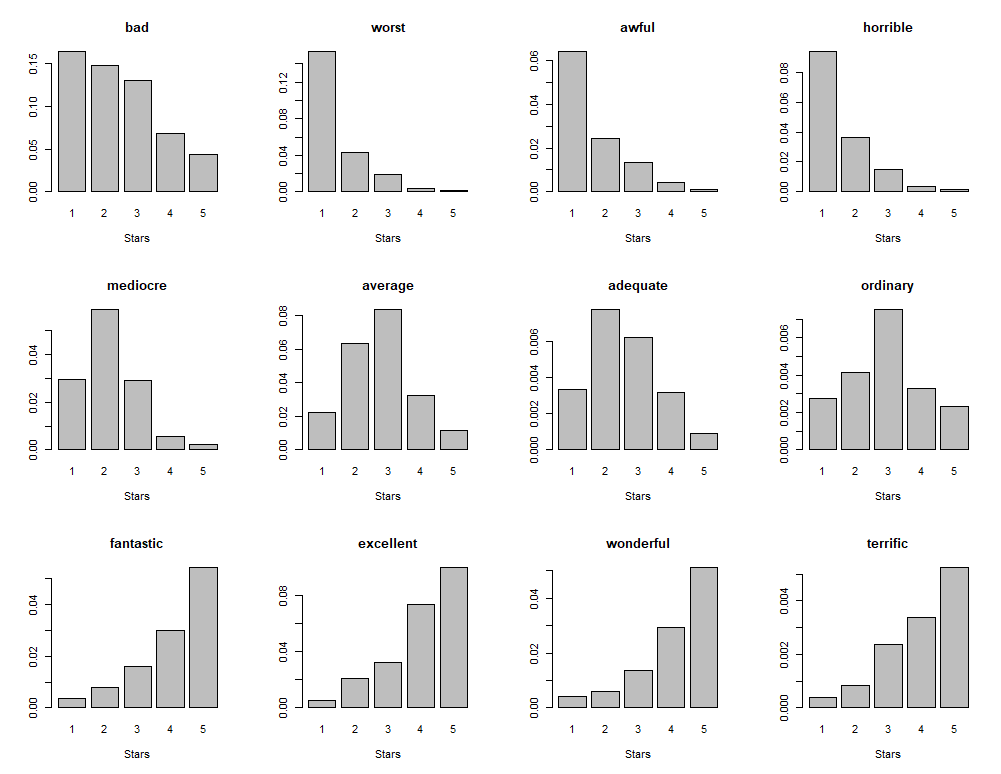
2.      Concatenate all comment texts

3.      Remove all stop words in the merged text

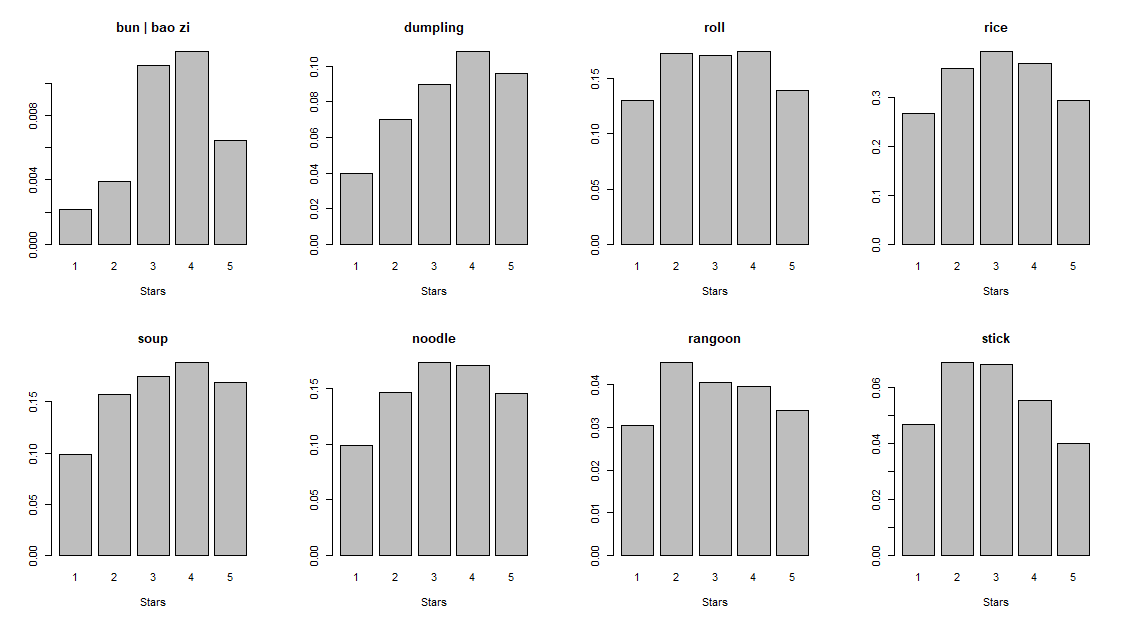
4.      Calculate the frequency of each word.

# EDA

In this part, we obtained four graphs about the correlation between some key words from comments and the rating(stars). The X axis is the number of stars of each review. And in each rating level, the Y axis tells the proportion of comment with certain key word. For example, in the first bar plot of the first graph, the rightmost bar shows about 4% of 5-star ratings contains the key word "bad".

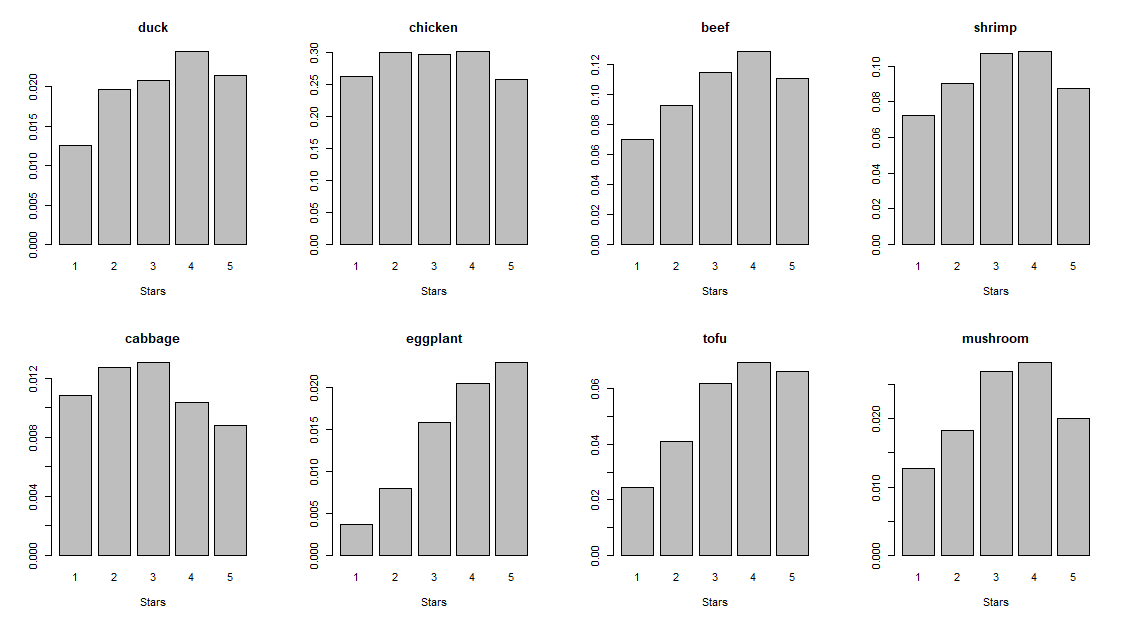


The first example graph shows, as expected, positive words are positive ratings and negative words are negatively correlated with negative ratings.



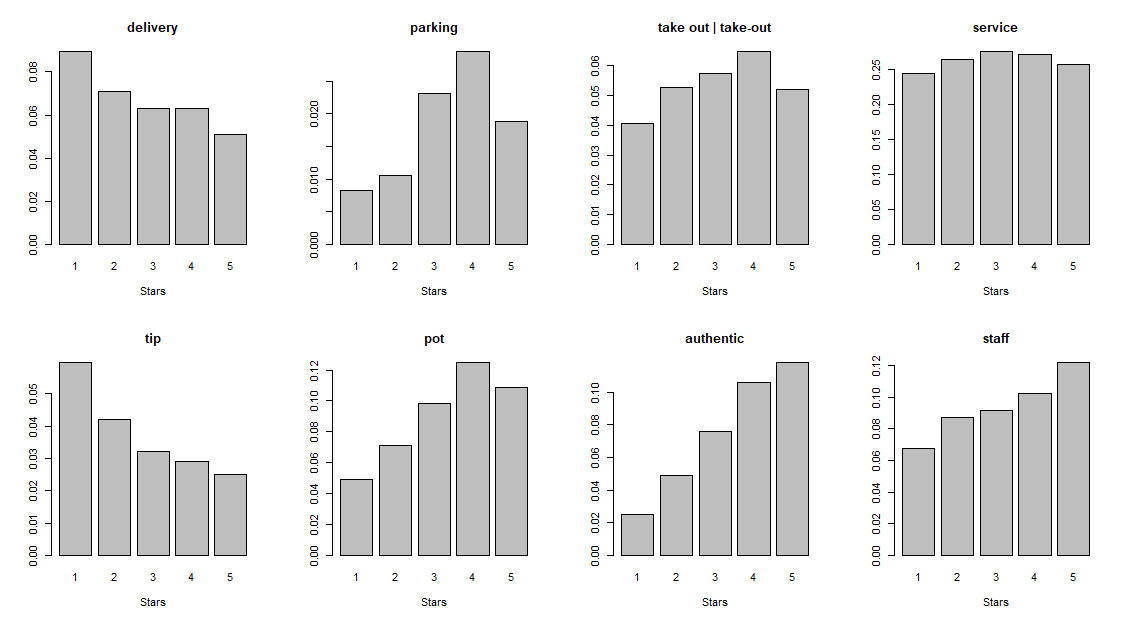
The second graph takes a look at the correlation between different main foods and review ratings.

As we can see, bun(or bao zi) and dumpling as well as soup have obvious positive correlation with ratings. Meanwhile, Rangoon and sticks have relative negative correlation with ratings.



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

We could find positive correlation with ratings when it comes to duck, beef, eggplant and tofu while there exists negative correlation for cabbage.



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 corresponds to low ratings.

From the above graphs obtained, we concluded some initial findings about the reviews and here are our corresponding suggestions in general as below.

For 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.

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.

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.

**Part 1. Key Findings About Chinese Restaurants**

1. Analytical insights
   1. We find that providing restaurants reservation and good kids environment, not providing food delivery and credit card acception increases ratings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | reservation | good kids | not food delivery | not accept credit card |
| p-value | 1.482e-05 | 0.02555 | 0.02302 | 0.0143 |
| mean difference | 0.24 | 0.24 | 0.12 | 0.32 |

* 1. We find that there is no difference in whether provide food take out, whether provide alcohol, whether have TV and whether provide WiFi with respect to ratings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | food take out | alcohol | TV | WiFi |
| p-value | 0.6636 | 0.6256 | 0.2021 | 0.1037 |
| mean difference | -0.04 | -0.017 | 0.10 | 0.12 |

1. 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 significant level 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 average ratings difference between group:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | reservation | good kids | not food delivery | not accept credit card | food take out | alcohol | TV | WiFi |
| lower | 2.77e-05 | 1.46e-05 | 6.32e-05 | 4.19e-05 | -0.50006 | -4.1e-05 | -4.5e-05 | -4.8e-05 |
| upper | 5.00e-01 | 5.00e-01 | 6.63e-05 | 5.00e-01 | 0.49996 | 1.84e-06 | 6.79e-05 | 6.55e-05 |

1. We checked the nomality and variance homogeneity for each attribute group of data.

* 1. Firstly, we use Shapiro-Wilk test to check the nomality of data. The p-value are all less than 0.05, so with significant level 0.05, we think the data are not normal. In this situation, we can not use t-test, so we use nonparametric test: Wilcoxon-Rank test to see whether there is relationship between some attributes and ratings.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| p-value | reservation | good kids | not food delivery | not accept credit card | food take out | alcohol | TV | WiFi |
| Yes group | 5.23e-09 | 2.92e-14 | 6.24e-09 | 0.01444 | 2.82e-15 | 1.19e-06 | 2.37e-13 | 2.17e-05 |
| No group | 4.10e-11 | 0.01293 | 3.48e-11 | 4.99e-15 | 0.02189 | 2.52e-11 | 1.09e-05 | 1.39e-12 |

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | reservation | good kids | not food delivery | not accept credit card | food take out | alcohol | TV | WiFi |
| p-value | 0.00165 | 0.03022 | 0.437 | 0.5552 | 0.5082 | 0.00055 | 0.01737 | 0.2726 |

**Part 2. Recommendations**

1. Recommendations for Chinese Restaurants

* Operations
  1. Providing restaurants reservation to increase ratings: on average, a chinese restaurants with reservation had 0.24 star higher rating compared to a chinese restaurants without reservation (Wilcoxon-Rank test p-value:1.482e-05).
  2. Providing special food and toy for children: on average, a chinese restaurants which is good for children had 0.24 star higher rating compared to a chinese restaurants which is not good for children (Wilcoxon-Rank test p-value:0.02555).
  3. Not accepting credit card payment: on average, a chinese restaurants which do not accept credit card had 0.32 star higher rating compared to a chinese restaurants which accept credit card (Wilcoxon-Rank test p-value:0.0143).
  4. It's not worth providing food delivery: Although food delivery is related to the ratings, the average ratings for food delivery chinese restaurant is only 0.12 higer than no food delivery chinese restaurant.
  5. It's not worth investing in food take out, alcohol, WiFi and TV: these four attributes are not related to the ratings (Wilcoxon-Rank test p-value are all greater than 0.05).