Project Summary

Augustine Tang, Harold Su, Zijin Wang

1 Introduction

In this project, we select restaurants which provide pizza as our samples. The reason why we pick them is that pizza is one of the favorite foods and there are a lot of pizza restaurants so that our analysis results could help a number of business owners. On the other hand, the large number of pizza restaurants also makes our analysis more reliable.

2 Data Pre-Processing

We use R/Rstudio to program and process data. Also, for some parts requiring a lot of computations, we use CHTC of UW-Madison to run our codes. To select pizza restaurants from the whole dataset, we directly use the variable categories in business_city.json and pick those businesses with pizza in that variable. Next, we select reviews from review_city.json corresponding to those businesses. That is all data we use to perform our analysis.

3 Methodology

The aim of this project is to provide data overview and business suggestion for pizza restaurant owners. The results can be divided into two aspects. The first one is a general result for all business owners while the other is personalized and different for each business owner.

For the first general part, we mainly consider the variables related to attributes in business_city.json. Firstly, we extract the Business ID and the attributes from the business_city.json related to pizza and pick the attributes with adequate observes (non-NA). Next, to take full advantage of the stars rated by customers, we merged the "attributes" in the business data and the "stars" in review data by matching each Business ID. Then, we treat each attribute as categorical variable, observe the sample statistics (mean, standard deviation, median and interquartile range) by groups , and then conduct Kruskal–Wallis H test. It is a non-parametric method only assuming independence for different levels instead of the normal distribution of the residuals. Also, we choose it since it is more effiecient and robust on large sample size than other median test. A significant Kruskal–Wallis test indicates that at least one sample stochastically

dominates one other sample. In our scenario, significance means the variable is influential for rates.

Yet, it is easy to see that the general suggestion may not be suitable for every restaurant and we also want to make use of reviews. Thus, we want to provide more specific, personalized suggestion for each business owner. Imagine I am an owner of a restaurant and there is only one review for my restaurant which says my food is too expensive. Then I know perhaps I should low down the price. Now assuming there are 1000 reviews (and only 1000) for my restaurant and most of them say my food is too expensive, then I know I should definitely low down the price of my pizza. Unfortunately, I do not have infinite memory and time to go through every review. Our solution is to, of course, make a program which can automatically analyze reviews for each business owner.

More specifically, we first tokenize all reviews into words, drop stopwords and stem them. Then, we manually go through top 300 words with highest frequency and pick some of them as keywords and divide them into 4 categories: food, service, environment and price. All keywords can be found on our App. We pick a category and a keyword in this category, e.g. category food and keyword pizza. Secondly, we tokenize each review into sentences and select those sentences containing this keyword. Next, for each sentence i, we do sentiment analysis and obtain a sentiment score S_i and also record the stars R_i and business ID of the review to which the sentence belongs. The method we use to obtain this sentiment score is given by package 'sentimentr' in R. The link above provides detailed model and equation. This score is between -1 and 1, where positive score implies positive sentiment in the given sentence. It can also handle negation, e.g. the score for 'it is good' is 0.433, for 'it is bad' is -0.433 and for 'it is not good' is -0.375.

Then, still for each sentence, we summarize an initial score from sentiment score and stars by formula $(S_i \times 10 + 1)/2$ if $R_i = 5$, $(S_i \times 10 - 1)/2$ if $R_i = 1$ and otherwise $(S_i \times 10)/2$.

The logic of this formula is as following. If we assume the sentiment score of each sentence is totally believable, then we can directly use its sentiment score as this initial score. But in practice, we find that it is not reliable in some cases. So, we want to correct it a little bit. The sentence in the review with 5 stars tend to be compliment so we add a little bit for those scores of high-rating sentences. However we cannot add too much, since I find some reviews in 5-star review also could point out shortcomings. So by this formula, we only want to make the sentiment score more robust.

Next, we group sentences by Business ID and take average of initial scores in each group as the final score for each Business ID(restaurant) corresponding to the keyword. Finally, for this keyword, we obtain the distribution of the final scores for different Business IDs(restaurants) and for a specific restaurant we can see the rank of its final scores for this keyword among all final scores corresponding to this keyword and other restaurants. So each business owner can know his/her performances in very

detailed aspect, e.g. I can see that people do not like my pizza, since my score for pizza is at the bottom 1/4. Also, for a restaurant, we take average of its final scores for all keywords in a category, obtain its final score for this category and deal with this score with same procedure as above.

The aim of this project is surely to provide business owners with business recommendation. So our first part provides general recommendation, e.g. free WiFi could improve rate. And our second part could give a much more customized recommendation, since we know its rank among all pizza restaurant. We capture the categories in top 25% and in bottom 25% and then give targeted opinions.

4 Analysis Result

The first part is analysis on attributes. According to the results of the Kruskal–Wallis H test (see Table 1), there are 21 attributes have the p-value less than the significance level 0.05. Thus, we can conclude that these attributes are significantly influential to the customers review. Here we pick attribute "WiFi" to give a brief interpret.

Attribute	**Kruskal-Wallis rank sum statistic**	**p-value**
RestaurantsTakeOut	105.75023	1.0880×10^{-23}
GoodForKids	1.9081	0.1672
RestaurantsDelivery	812.71323	3.3233×10^{-177}
WiFi	52.3438	4.3022×10^{-12}
BusinessAcceptsBitcoin	1.5026	0.2203

Tab. 1: Results of the Kruskal–Wallis H test.

As the boxplot in Figure 1 depicted, no matter if a business offers "free WiFi", "no WiFi" or "paid WiFi", there is barely no influence to the stars rated (except the first quantile). We might then come up with a conclusion that the attribute WiFi is not a considerable attribute to the a pizza restuarant. While the Kruskal–Wallis H test gives a p-value far more less than 0.05, meaning the WiFi offered by a pizza business owner affects the stars rated by customers significantly. Noting the null hypothesis of the Kruskal test is that the medians of all groups are equal, the reason we obtain opsite conclusion is worth thinking. From our perspective, it is due to the way the stars given. The stars varies only from 1 to 5 and our sample size is pretty large so that the boxplot under this circumstances is somewhat misleading. In other words, this extend our understanding to the power of statistical tests: we do discovered significant clues in seemingly ordinary data!

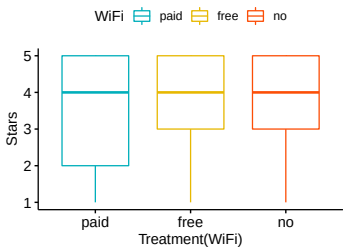
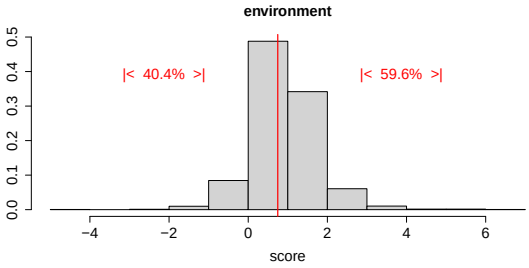


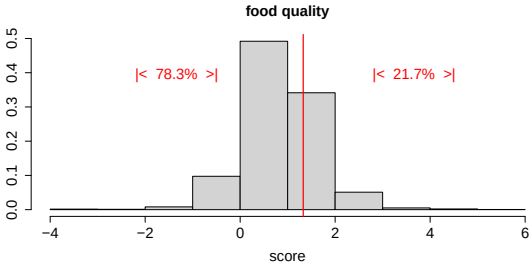
Fig. 1: Boxplot Stars by "WiFi".

The second part is analysis on reviews. As mentioned above, we have different conclusion for each restaurant. We cannot show all of them here. Take the result for restaurant "Po' Boys Restaurant" with Business ID "-5NXoZeGBdx3Bdk70tuyCw" as

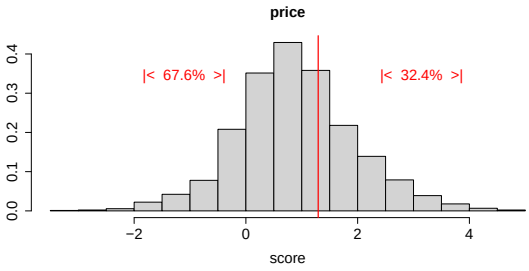
an example, the distribution of scores for each category and rank of the restaurant is shown as follows:



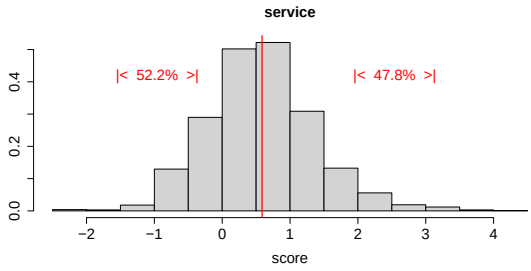
(a) environment



(b) food quality



(c) price



(d) service

Fig. 2: Distribution of Scores

We can see that, for this restaurant, its food quality is at top 20% and other ranks are ordinary. So the owner could focus on its environment, which has the lowest rank. Some of our business recommendation for this restaurant are (these recommendations are given by our program automatically)

1. According to reviews, the following food related elements in your menu are considered to be impressive:
   basil garlic honey jalapeno meatbal onion ranch rib salad spaghetti spice tea

   Reviews indicate that none of food related elements in your menu mentioned by reviews have to improve.

2. Your bartend score is 2.74 given by totally 3 reviews with this key word. It is better than 95.9 percent of all 369 businesses, while the median score is 0.59 .

3. Your pizza score is 0.74 given by totally 70 reviews with this key word. It is better than 67.5 percent of all 1991 businesses, while the median score is 0.5 .

On our App, more detailed results are shown. Our aim is to tell some facts and opinions of customers. The detailed decision is left to business owner. E.g. customers thought my price is high but I cannot low it down due to operation cost(my restaurant

is at CBD).

Contribution

Augustine Tang wrote/edited all the four sections of the summary, worked on slides pages 5-9 , created and revised code concerning review sentiment analysis, provided general feedback on the Shiny app.

Harold Su wrote/edited the methodolotgy and the analysis result of the summary, worked on slides pages 10-14 , created and revised code concerning review sentiment analysis, built up the frame and edited the first tab on the shiny App and provided general feedback on the Shiny app.

Zijin Wang wrote/edited the methodolotgy and the analysis result of the summary, worked on the slides pages 1-4 ,created and revised code concerning attributes analysis, edited the second and third tab on the shiny App and refined the shiny app so that it looks more “shiny”.