1 Introduction

If you were a new owner of a pizza restaurant, how to improve the store score and the customer impression on Yelp? Briefly speaking, the **3F theory**, referring to friendly attitude, fast delivery and fresh ingredients, includes the most essential factors in increasing overall store score on the Yelp.

Our project utilize data from the Yelp that contains 76,254 reviews from 2,039 different pizza-related shops, and focuses on pizza related businesses in four state, Illinois, Ohio, Pennsylvania, and Wisconsin to solve the following questions:

–What are the distribution patterns of these restaurants? How about their feedback from customers?

–Are there any interesting information can be extracted from customers' reviewsand businesses' attributes?

–Are there any specific variables that make a great difference based on statistical model results?

2 Data Cleaning

2.1 Data Resource And Pre-processing

The dataset is from Yelp Open Dataset [1], the origin data format is JSON file, including business.json, review.json, user.json, checkin.json, tip.json, and photo.json. We selected a subset of these million reviews from businesses in Illinois(IL), Ohio(OH), Pennsylvania(PA), and Wisconsin(WI); chose business.json and review.json for our project goal.

First, we merged the business.json and review.json datasets according to unique business_id of each business. Then we selected businesses whose categories variable contains "pizza", through which 2,093 businesses were selected, corresponding with 76,254 reviews. Variables 'review_id', 'business_id', 'stars_x', 'text', 'name', 'city', 'state', 'stars_y', 'attributes', 'categories', 'hours', are selected and other 11 variables are removed. We used "nltk" in python to extract words in 'text' variable, lemmatized the words, and then filtered common English words and words which were mentioned less than 2000 times across all reviews. Finally, 390 words were selected.

Among these words, we chose 4 topics (see in table 1) that can thoroughly include informative reviews contents, and classified 79 relating words into each topic (see in figure 2(b)). In particular, all words related to attitudes are dropped (e.g. 'love' or 'disappointed') in that they are consequences instead of reasons. For the latter regression model part, each word works as a predictor vector, and the corresponding frequency of the word for each business can be seen as the input.

However, it's still time-consuming and inefficient to build model based on over 100 words with many 0s (some people might type the word 'cheese' in a review, but some might not), which forces us to perform another round of variable selection.

Table 1: Four Topics And Corresponding Words

size	service	material type	taste
little	bar	cheese	sauce
small	table	crust	fresh
large	friendly	salad	italian
...	...	...	...

2.2 Variable Selection

2.2.1 Word Selection Within Attributes

Variables in attributes are provided by the yelp, which are of great importance but also with high missing characteristic. Altogether, there are 50 variables from attributes, but some of them have too many NAs to continue our analysis. For example, "whether the restaurants open at late night" has almost half of the NAs in the 2000 restaurants, which will strongly affect our latter model fitting. After we deleted attributes considering the high missing pattern, there are 36 variables remaining, thus we employed *t-test* to check the significance level between each variable and the rating for each business. Based on this method, we managed to select four variables as *Caters, GoodForKids, RestaurantsTakeOut, RestaurantsPriceRange* to include in our model selection part.

[1] https://www.yelp.com/dataset

2.2.2 Word Selection Within Reviews

Should you improve the quality of garlic or onion as the next step to improve your review stars in Yelp? Since there are countless 0s for each word, it is not appropriate to use correlation coefficient. Instead, we define three concepts for the review data set to select vital factors:

positive review rate: count of the word in 5 star level reviews/count of the word in all star reviews

negative review rate: count of the word in 1 star level reviews/count of the word in all star reviews

average review stars: weighted average stars, where weight is the frequency of word in each star level

From these 76,254 reviews, the grand mean star count is 3.64, while the positive review rate and negative review rate are 39.18% and 14.97% respectively. Take the word "fresh" as an instance to show our selection method. "Fresh" occurs 9871 times in all reviews with average review star as 4.1817, and it's negative review rate and positive review rate are 3.98% and 50.87% respectively. Compared with the grand mean, this word increases the positive review rate by 11.69% and decreases the negative review rate by 10.99%, let alone improving the average star for 0.545. Since the absolute value of difference is over 0.5 star (a rough threshold), then we consider the effect of "fresh" important. On top of that, the importance of positive review rate and negative review rate is also included by calculate the product of two difference. Therefore, "fresh" is a vital factor in "taste" topic.

Follow this logic, we then gain these following variables: *hoagie, table, friendly, delivery, pie, drink, dessert, mozzarella, fresh, Italian, crispy, greasy, spicy.*

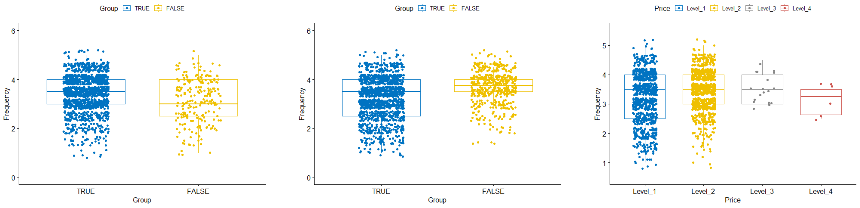
3 Part 1: Key Findings About Pizza Businesses on Yelp

3.1 Descriptive Analysis

Overall Findings:

- Offering food that is good for kids, improving delivery quality and adjusting the price appropriately is correlated with higher business rating given by Yelp.
- Fresh ingredients, clean environments, friendly and fast service are three key aspects stressed from the customer angle. Providing pizza with larger size and crispy crust, offering yummy dessert and drink will gain you a bonus. Interestingly, customers love mozzarella cheese and spicy taste.
- Higher positive review rate show an association with 'homemade' pizza and Italian style. On the other hand, negative review rate related to 'rude' reach to 68%, while 'waited' and 'delivery' bother customers as well.

Figure 1: Data Patterns From Attributes

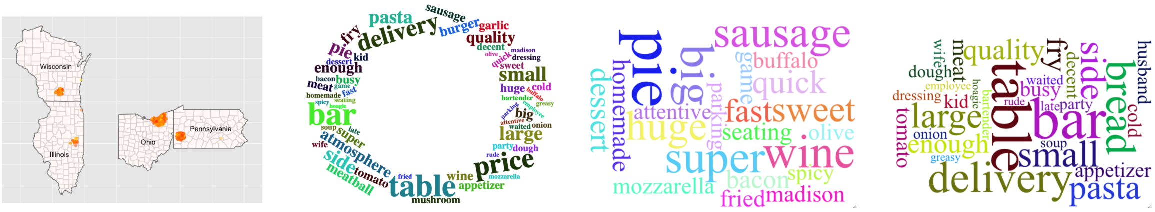


(a) GoodForKids (b) FoodDelivery (c) PriceLevel

For attributes, we drew boxplots to show the difference of rating among different level. Plots for GoodForKids, FoodDelivery, PriceLevel are displayed in figure 1. The boxplot implies that pizzeria with kids-friendly food is correlated with higher rating on Yelp. There is no doubt that delivery is a vital service in the evaluation for store rating, yet the weird fact is that pizzeria without take-out service tend to has a higher score. A reasonable explanation would be that customers are usually annoyed by the quality of delivery. Apparently, lower price caters to customer demands.

For business location and reviews, we also drew map plot and word cloud to illustrate their patterns. As shown in the map plot in figure 2(a), each point refers to a single business located in these four states, and each corresponds with color ranging from light orange to dark orange according to the review star. Clearly, pizza stores are densely distributed in more developed areas.The more density these stores are distributed, the higher score these store possess. From the word cloud of 79 words in figure 2(b), customers tend to pay great attention to food material like sauce and cheese, and services like delivery and table condition. After word selection, the results can be seen from table 3, 4 and 5 in appendix.

Figure 2: Data Patterns From Reviews



(a) Distribution of review stars (b) Most frequent words mentioned (c) "Positive" Word Cloud (d) "Negative" Word Cloud

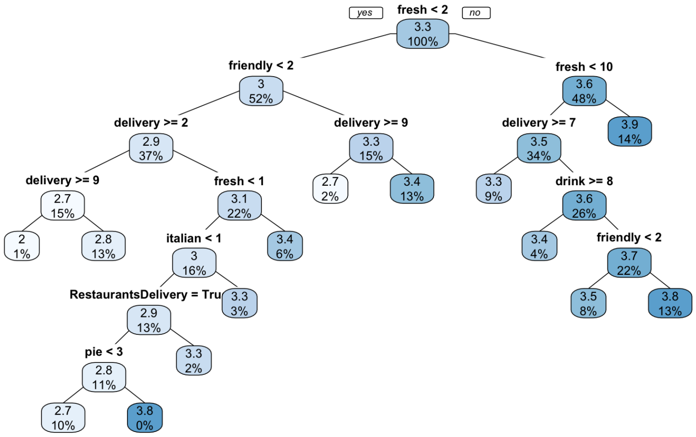
3.2 Statistical Model Analysis

3.2.1 Model Selection

To identify relative important factors effect business star evaluation, we propose a statistical model analysis. The 'stars' is an ordinal variable, so an ordinal regression is often considered. As the number of outcomes of a regression tree is limited, instead of unlimited continuous outcomes of simple linear regression. A regression tree is also applicable to this problem. Besides, since our attribute variables contain missing values, we consider regression trees as the primary model, which can deal with NAs in data. Whereas the ordinal regression need imputation for missing values.

Regression trees are methods for constructing predictions of continuous or ordered discrete dependent variables. The method creates splits based on predictor variables included in the model by recursively partitioning the dataset one predictor variable at a time. The variables at the top of the tree can be considered the most important predictors, and variables included lower down in the tree are less related to the outcome. A node is a group of similar observations, and a terminal node is the final node from a set of branches that does not split any further.

Figure 3: rpart Tree Result



3.2.2 Model Results

The outcome of the 'rpart' regression tree is shown in Figure3. At each split, an observation goes to the left branch if and only if the condition is satisfied and vice versa. For example, if the words "fresh" were mentioned more than two times from reviews for your business, your score would fall in the right half of the tree. Notice that the overall star from the right half of the tree, 3.6, are larger than the stars from the left half, 3. That indicates more "fresh" mentioned in reviews results in a higher star evaluation. From the tree

split construction, we could have a straightforward understanding about the effect of some variables. For example, if the food in your store is so fresh that impresses the customer, say, it is mentioned many times, the customer who scores tends to give you a higher star. Another example is delivery. If your store provide delivery, i.e. ”Restaurants Delivery” in attributes is true, your restaurant’s star will be higher.
’rpart’ selected variables with importance scores, referring to the extent to which the variable affects the result. A higher importance score indicates more importance. It ranked ’fresh’ as the most important variable, i.e., if you’re a boss of a pizza shop, freshly served food is always something you need to pay special attention to. Also, other important variables are given (Table 2). Details are in Section 4.

Table 2: Selected Important Variables by rpart With Important Score

variable	fresh	friendly	table	crispy	mozzarella	italian	delivery
score	229.0731882	150.6627048	92.2474379	89.2017317	78.3673421	77.9967831	73.1431726
variable	RestaurantsDelivery	drink	dessert	pie	hoagie	greasy	spicy
score	11.5655628	11.5571729	8.6959116	8.5478603	5.8732043	4.6745120	0.5960147

3.2.3 Model Limitation

There is no significance level nor coefficient to show the causal relationship between these independent variables and dependent variables.
The model may face a problem of over-fitting since we built a bigger rpart tree to include more variables for interpretation their effect on the stars.

4 Part 2: Recommendations for Pizza Businesses

As a pizza business boss who wants to improve your Yelp star, you’re strongly suggested to take following action:

- Serving customer with **friendly** attitude and **fast** delivery, as well as **clean** table.
  No matter it’s takeaway or dine-in, the attitude and speed of service are of key concern for customers, since many people eat pizza for convenience. Never being rude with your customer, nor delaying the orders especially for delivery. You may set some serving guidelines for your employees, add more tip for drivers when delivering and check the neatness of your pizzeria.
- Serve food as **fresh** as possible, try to add **homemade** taste for customers.
  Provide homemade pizza, offer taste choice like mozzarella cheese and Italian style. For size choice, people are more interested in pizza in larger size and with crispy crust. Drink, dessert, pie, and hoagie served in your shop also affect the score. If you want to serve these foods, it is recommended to investigate customers’ tastes and provide satisfactory side dishes.
- Offer pizza sets that are **good for kids** and adjust the **price** appropriately.
  A large segment of pizza customers are kids, so it would be wise to provide special children’s set, such as give free toys as gifts or packages targeting at kids’ favorite taste. Price is also important since pizza stores tend to gather together, it will be better to set the price level in 1 and 2.

5 Conclusion, Strengths and Weaknesses

In summary, we managed to discover several key factors that will strongly affect the overall store rate, which are **serving attitude**, **serving speed**, and **ingredient quality**. After data cleaning, variable selection and model fitting, we also provide some data driven suggestions such as keeping your pizza store clean, providing fresh and homemade food and adding play pizza set.
There are several highlights in our project. First, we make great use of information from customer reviews, since we define three variables and use the distribution pattern of each word. Our shiny app gives suggestions from four orthogonal aspects, which is user-friendly and applicable. Second, through the descriptive analysis using *t.test*, map plot and word cloud, we visualize the current distribution and prove the statistical correctness of our findings.
However, our analysis is not perfect. First, word frequency in reviews is consequence instead of controllable variables, the relationship is not causal but association. Second, due to time limitation, we haven’t develop the prediction part for our model, which means we can not provide personalized suggestion based on model results.

Contributions

Yukun Fang: Word selection for business.json dataset and corresponding code, construction of Shiny app, slides 9-12, 20-21.
Yiran Wang: Word selection for review.json dataset and corresponding code, descriptive analysis, slides 1-6, 13-15.
Jie Sheng: Data pre-possessing and corresponding code, model selection and diagnosis, slides 7-8, 16-19.

Appendix

Table 3: Selected Words (First Round) And Their Good/negative review rate

Star Count	Total	Bad Review Rate	Diff From Grand Mean	Good Re -view Rate	Diff From Grand Mean	Opposite Impact Or Not
super	4341	5.97%	-9.00%	47.29%	8.11%	Yes
hoagie	2052	16.96%	1.99%	31.24%	-7.94%	Yes
table	10510	21.68%	6.72%	20.69%	-18.49%	Yes
friendly	9658	2.88%	-12.09%	52.22%	13.04%	Yes
delivery	8075	28.73%	13.76%	25.20%	-13.98%	Yes
pie	5928	5.31%	-9.65%	42.71%	3.53%	Yes
drink	8622	16.60%	1.63%	27.41%	-11.77%	Yes
dessert	2861	6.57%	-8.40%	42.08%	2.90%	Yes
mozzarella	2179	6.65%	-8.31%	41.99%	2.81%	Yes
fresh	9871	3.98%	-10.99%	50.87%	11.69%	Yes
italian	8730	7.42%	-7.54%	43.16%	3.98%	Yes
cold	3608	40.80%	25.83%	14.36%	-24.82%	Yes
crispy	3092	3.72%	-11.25%	40.23%	1.05%	Yes
greasy	2095	15.47%	0.50%	22.20%	-16.98%	Yes
spicy	2082	4.23%	-10.74%	39.39%	0.21%	Yes

Table 4: Selected Words (First Round) And Their Average Star

Star Count	Total	Average Star	Diff From Grand Mean	Larger Than 0.5 Star Or Not
super	4341	4.04	0.41	No
friendly	9658	4.24	0.60	Yes
delivery	8075	3.04	-0.60	Yes
pie	5928	3.97	0.33	No
drink	8622	3.35	-0.29	No
dessert	2861	3.93	0.29	No
mozzarella	2179	3.93	0.29	No
fresh	9871	4.18	0.55	Yes
italian	8730	3.90	0.27	No
cold	3608	2.43	-1.20	Yes
crispy	3092	4.03	0.39	No
greasy	2095	3.25	-0.39	No
spicy	2082	3.98	0.35	No

Table 5: Selected Words (Second Round) And Their Results

Topics	Words	Appearance Count	Average Star	Diff From Grand Mean	Larger Than 0.5 Star Or Not	Opposite Impact Or Not
Size	super	4341	4.04	0.41	No	Yes
	hoagie	2052	3.42	-0.22	No	Yes
Service	table	10510	3.09	-0.55	Yes	Yes
	friendly	9658	4.24	0.60	Yes	Yes
	delivery	8075	3.04	-0.60	Yes	Yes
Material Type	pie	5928	3.97	0.33	No	Yes
	drink	8622	3.35	-0.29	No	Yes
	dessert	2861	3.93	0.29	No	Yes
	mozzarella	2179	3.93	0.29	No	Yes
Taste	fresh	9871	4.18	0.55	Yes	Yes
	italian	8730	3.90	0.27	No	Yes
	cold	3608	2.43	-1.20	Yes	Yes
	crispy	3092	4.03	0.39	No	Yes
	gre asy	2095	3.25	-0.39	No	Yes
	spicy	2082	3.98	0.35	No	Yes