STAT 628 Module 3

Group 2: Fengxia Dong, Zhiyu Ji, Xinyue Wang

**Introduction/Motivation**

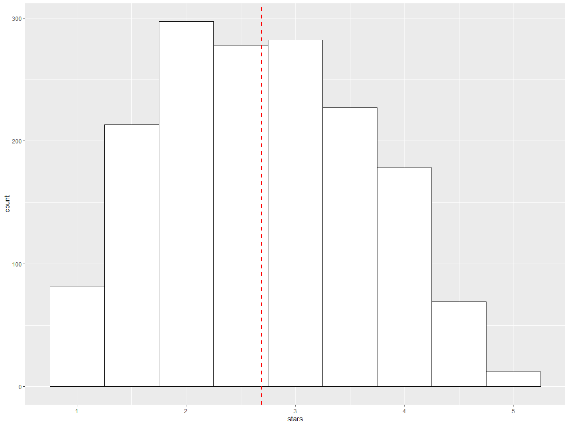
Our analysis focuses on restaurants that provide fast food to customers and had been on Yelp during the period from October 2004 to December 2019. We are motivated to analyze this category of restaurants because we find that their Yelp reviews were not as good as other categories of restaurants, the average stars only 2.69 during the period. This category of restaurants has not been serving customers well and need to have advice to improve.

We will explore what factors significantly affect customers’ review. Furthermore, we will focus on the four major areas: food, waiting time, service, and sanitary conditions of restaurants and answer the following questions: (1) will using fresh food materials be helpful to improve Yelp review and by how much? (2) how many stars will be increased if the restaurants reduce waiting time? (3) how important is staff service in Yelp review? and (4) will cleaning restaurants well be useful to improve? We will also provide advices to the restaurants accordingly.

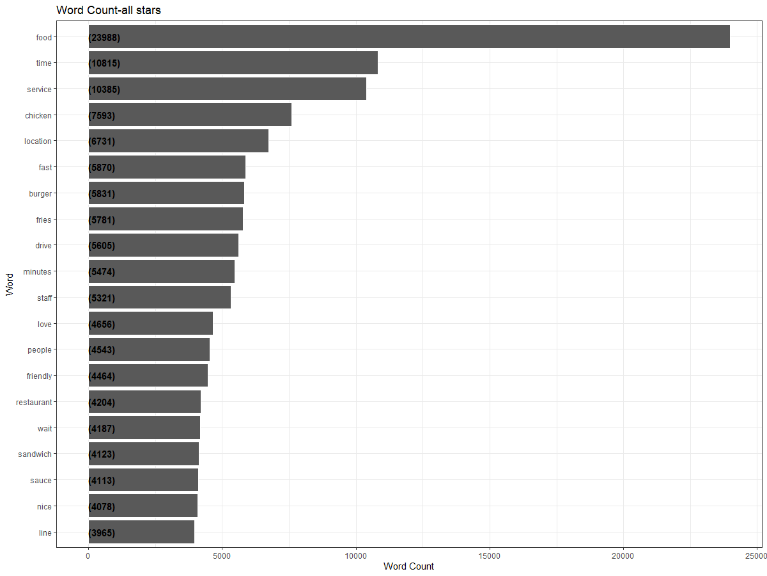
**Background Information/Data Cleaning/Data Pre-Processing**

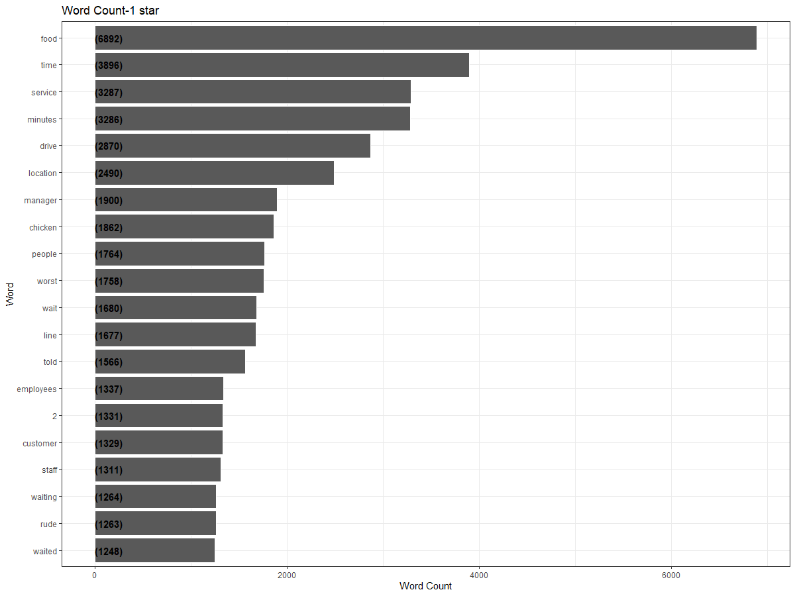
Our data set is obtained from Yelp, a company recently released some of its reviews to the public. The data are stored in JSON format, including information of reviews, businesses, users, and tips. Given our study objectives, we first select businesses in business json file with “fast food” listed in its category. Then we select reviews for the chosen businesses from review json file through the common variable “business\_id” in both business and review files. We further select users from user json file through the variable “user\_id” in the selected reviews and tips from tip json file through “business\_id” in selected businesses. We end up with 33,262 reviews by 21,741 users for 1,638 restaurants serving fast food with 5,823 tips. The four selected files are put in csv format and analyzed in R.

To process text data in reviews, we utilize the R package “tidy2vec” to break the text into individual tokens with ngram=1L or 2L (i.e., one word or a set of two consecutive words) and transform the list of tokens into a vector space. We prune words occurring less than 10 times and those appearing in less than 0.1% of reviews. We then create a document-term matrix (DTM) for our further analysis. In total, we have 15,163 words in the DTM.

**Exploratory Data Analysis**

In total, there are 1,638 restaurants reviewed on Yelp serving fast foods. These restaurants are located in the states of Illinois, Ohio, Wisconsin, and Pennsylvania, with average stars ranging from 2.64 to 2.82. As shown in the figure to the right, the distribution of stars for those restaurants is skewed towards the right with a mean value of 2.69 and standard deviation of 0.94. It is therefore critical for those restaurants to find out the causes of the low ratings in order to improve.

 To explore which aspects are considered important by consumers, we examine the frequency of words appearing in reviews by using “tidytext” package in R. We exclude words that are not useful for an analysis, typically extremely common words such as “a/an”, “the”, “of”, “to”, and so on stored in “tidytext” package. We plot not only the frequency of top 20 words appearing in all reviews, but also for each star level. But due to space limitation, we only show the plots of words for all reviews and reviews with 1, 2 and 5 stars.



The above word frequency justifies the areas on which we target in this module, including food quality, waiting time, service, and sanitary conditions. For example, the most frequently occurring words “food”, “chicken”, “burger/burgers”, “fries”, “sandwich”, “sauce”, “cheese”, “fresh”, “pizza”, “hot”, “delicious”, “menu”, and “meat” are all related to food; “time”, “fast”, “minutes”, “wait/waiting”, “line”, and “quick” are related to waiting time; “service”, “staff”, “friendly”, “experience”, “manager”, “employees”, “home” may be related to services; and “clean” is related to restaurant sanitary conditions.

**Key Findings About Restaurants Serving Fast Foods**

Given the large number of words occurring in reviews, we utilize Lasso regression to do the word selection and find aspects significantly affecting customers’ review scores. A dummy variable is created which is equal to 0 if the star is 1 or 2 and equal to 1 otherwise. Subsequently, a binomial model with the dummy variable as the response variable and the created DTM as the covariates is run for the Lasso regression. The optimal “lamda” that minimizes MSE is selected through cross validation which is equal to 0.0015. The Lasso regression selects 3,204 from 15,163 words and R2 is as high as 0.81, indicating a satisfactory prediction. To estimate the statistical significance of those non-zero parameters, we use R package “selectiveInference” to calculate their p-values. Among these 3,204 non-zero parameters, 244 are statistically significant at 95% level. As our main purpose is to provide useful advices to the restaurants, we exclude those general words which cannot provide insights into any related areas, such as “finally”, “well done”, “dissatisfied”, “order them”, “items in”, “were like”, “supposedly”, “began to”, and “ordered some”, etc. We thus focus on the significant words that can be directly linked to one of the four targeted areas. The selected significant words grouped in targeted areas with parameter estimate and p-values are listed in the table below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | word | parameter estimates | p-value |  | word | parameter estimates | | | p-  value |
| Service | Poor\_service | -1.07 | 0.01 | waiting- | 40\_minutes | -0.64 | | | 0.04 |
|  | decent\_price | 0.62 | 0.02 | time | Quick | 2.10 | | | 0.04 |
|  | apologies | -1.79 | 0.05 |  | for\_15 | -0.38 | | | 0.00 |
|  | horrible\_service | -0.73 | 0.04 |  | for\_20 | -0.79 | | | 0.01 |
|  | friendliness | 2.10 | 0.05 |  | slow\_and | -0.73 | | | 0.00 |
|  | attitude | -0.23 | 0.02 |  | Fast | 0.68 | | | 0.04 |
|  | rude | -1.35 | 0.02 |  | minutes\_to | -0.23 | | | 0.02 |
|  | the\_service | 0.09 | 0.02 |  | in\_line | -0.31 | | | 0.00 |
|  | incompetent | -1.05 | 0.02 |  | over\_15 | -0.91 | | | 0.01 |
|  | Delivery | 0.63 | 0.05 |  | waited | -0.19 | | | 0.00 |
|  | greet | 0.34 | 0.03 |  |  |  | | |  |
|  | their\_job | -0.33 | 0.04 |  |  |  | | |  |
| food | is\_cold | -1.64 | 0.02 | Sanitary | clean\_and | | 0.21 | 0.05 | |
|  | not\_fresh | -1.81 | 0.02 condition | condition | dirty\_tables | | 1.29 | 0.05 | |
|  | Huge | 0.61 | 0.02 |  |  |  | | |  |
|  | kids\_meal | 1.02 | 0.03 |  |  |  | | |  |
|  | ingredients\_and | 0.74 | 0.03 |  |  |  | | |  |
|  | were\_delicious | 0.37 | 0.05 |  |  |  | | |  |
|  | Great\_food | 0.80 | 0.05 |  |  |  | | |  |

As we discussed above, many positive or negative words are selected by Lasso, such as “dissatisfied”, “well done”, “terrible”, and “fun”. While those words are significantly correlated with stars, they are not useful for generating advices. The speed of our statistical analysis could be much improved if we added those words into stop-words list. In our statistical analysis, we use a binomial model with the dummy variable indicating more than 2 stars or not. The limitation of using binomial model is that specific effects cannot be captured for each star level.

**Recommendations for Businesses**

A positive parameter suggests that the restaurant becomes more likely to get more stars, while a negative one suggests that the restaurant becomes less likely to get more stars. From the category of “service”, we can reasonably deduce from words “apologies” and “incompetent” that mistakes made by staff members can negatively affect consumers’ review, indicating the importance of employee training. Consumers are also concerned with ingredients in foods, exampled by words “not fresh” and “ingredients”. Food portion may be appealing to consumers as indicated by word “huge”. While long waiting in line can result in negative reviews, “quick” and “fast” service can improve reviews.

By calculating the odds ratio, we can find how much the related aspects can affect the likelihood of consumers’ review stars falling into {3, 4, 5}. For example, the likelihood of the restaurant with “friendly” staff getting 3, 4, or 5 stars is 8 times of that without “friendly” staff; the likelihood of the restaurant with “rude” staff getting more stars is only ¼ of that without rude staff. In addition, the likelihood of restaurants with “quick” service getting more stars is 8 times of that without “quick” service; and the likelihood of restaurants with “delicious” food or “great food” is 1.4 or 2.2 times of that without “delicious” or “great food”, respectively. We note that restaurants with “dirty tables” still have higher likelihood to get more stars. One reason may be that better food or service in those restaurants outweighs “dirty tables”.

Based on above analysis, we have the following advices for restaurants serving fast food.

1. Train staff to provide good services, be friendly, reduce mistakes as much as possible. Keep price at a reasonable level. If possible, provide delivery service.

From the analysis, we can see that friendliness or rudeness can significantly affect customers review; and incompetent staff can make customers frustrated.

1. Use fresh ingredients in food. Be kids friendly by providing kids meal. While providing delicious food is important, consider increasing the portion of some food (maybe one or two costing less) to appeal customers.
2. Reduce waiting time. Customers going to restaurants serving fast food may highly likely prefer a quick meal. Prepare well ahead of rush hours and train employees to be more efficient.
3. Clean the restaurants as complete as possible. While consumers may be attracted by good service or food with a trade-off with sanitary condition, being clean can still help improve customers’ reviews and avoid health issues.

Our suggestions are based on a statistical analysis using a binomial model. It cannot provide an estimate of an increase in stars if adopting any of the advices.

**Visualize Our Analysis for Business Owners**

To make it easier for business owners to understand our data analysis, we build an app based on R and Shiny. Our analysis focus on two parts, the first part is to show the top 20 most frequent word in different star-level restaurants, the second part is to show the key words in different aspects which will influence the star-level, like service, waiting time, etc. Business owners can click the selection button on the top-left and select one, then on the right image part will show a statistics graph for the selected stars. Business owners also can select an aspect they interest and on the right plot part will give them a bar plot, which shows the coefficients of words obtained by regression. It will tell the owners which words and how they do effect the final stars for their restaurants.

**图表

描述已自动生成图表, 瀑布图

描述已自动生成**

**Conclusions**

In this module, we utilized natural language processing along with Lasso regression to analyze aspects that are critical to customers’ reviews on Yelp. Based on our analysis results, we provide useful advices for restaurants to improve their reviews.

**Contributions**

XW and FD cooperated on NLP and R coding, while ZJ worked on Shiny App. FD drafted the Executive Summary and XW and ZJ made edits. Every member contributes to the narrated presentation and Github page maintenance.