STAT 628 Module 3

Group 2: Fengxia Dong, Zhiyu Ji, Xinyue Wang

**Introduction/Motivation**

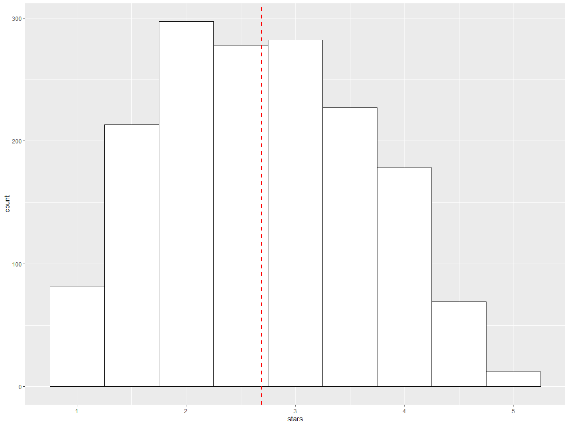
Our analysis focuses on restaurants that provide fast food to customers and had been on Yelp during the period from October 2004 to December 2019. We are motivated to analyze this category of restaurants because we find that their Yelp reviews were not as good as other categories of restaurants, the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

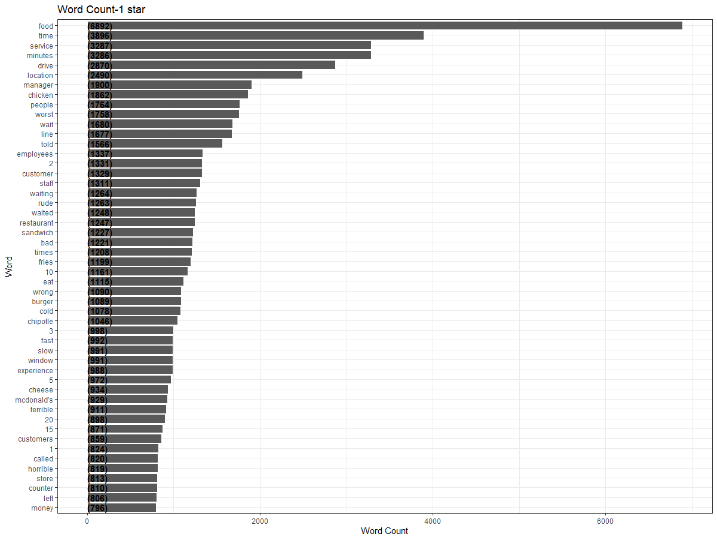
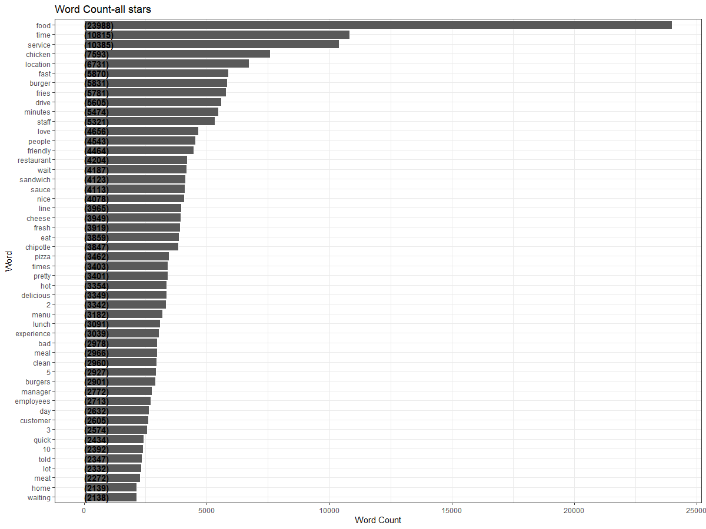
**Background Information/Data Cleaning/Data Pre-Processing**

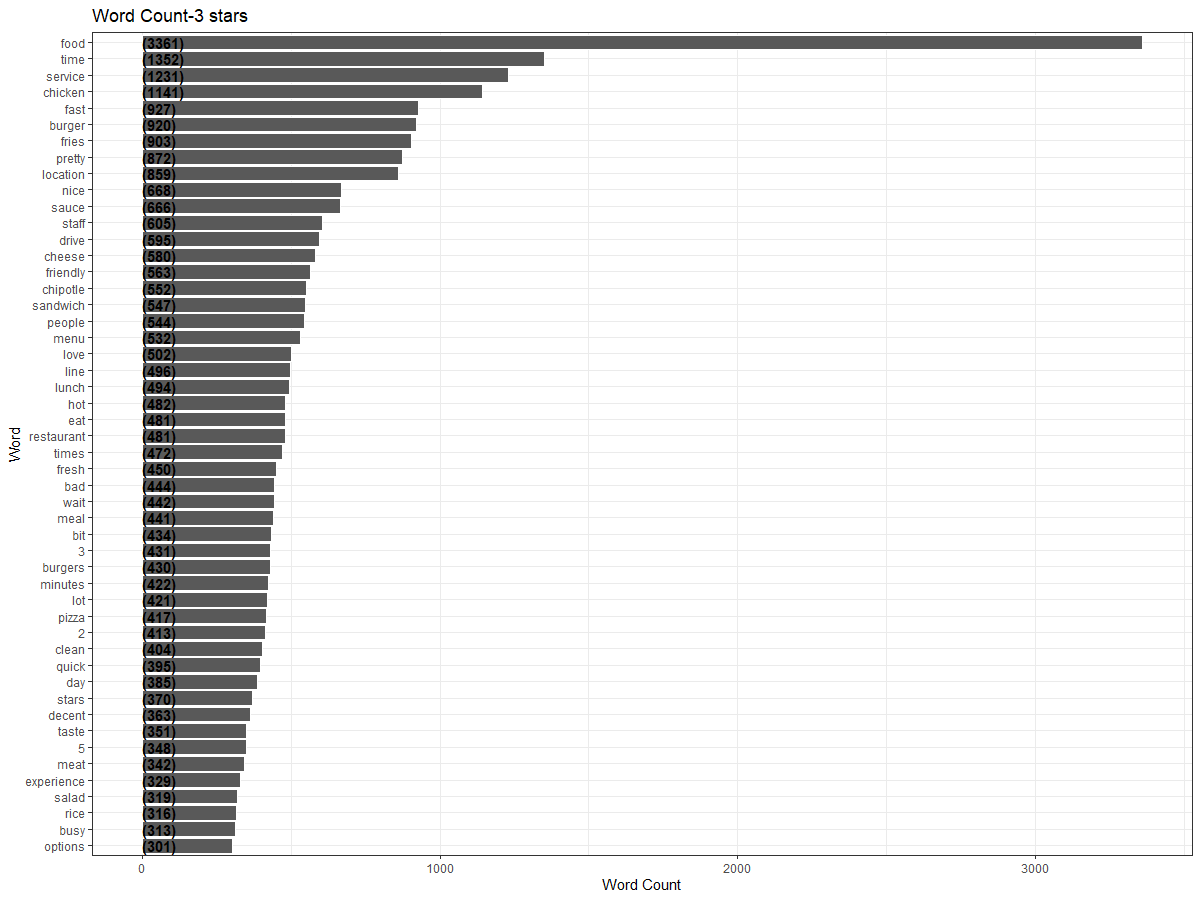
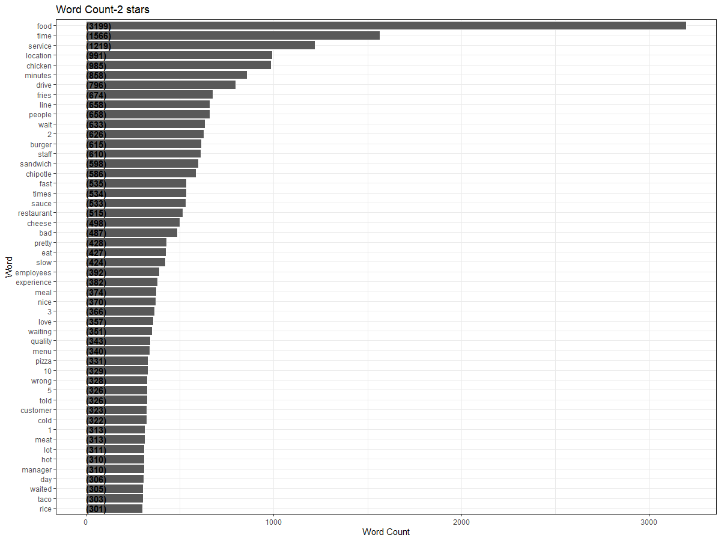
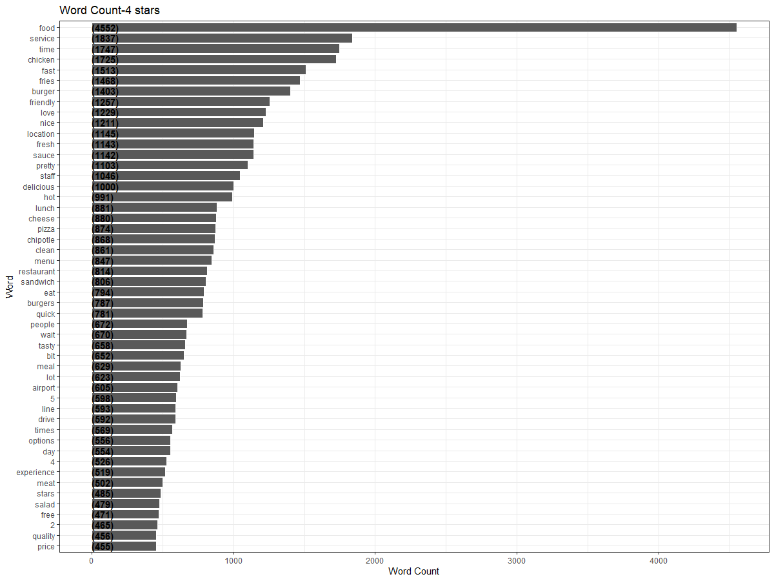
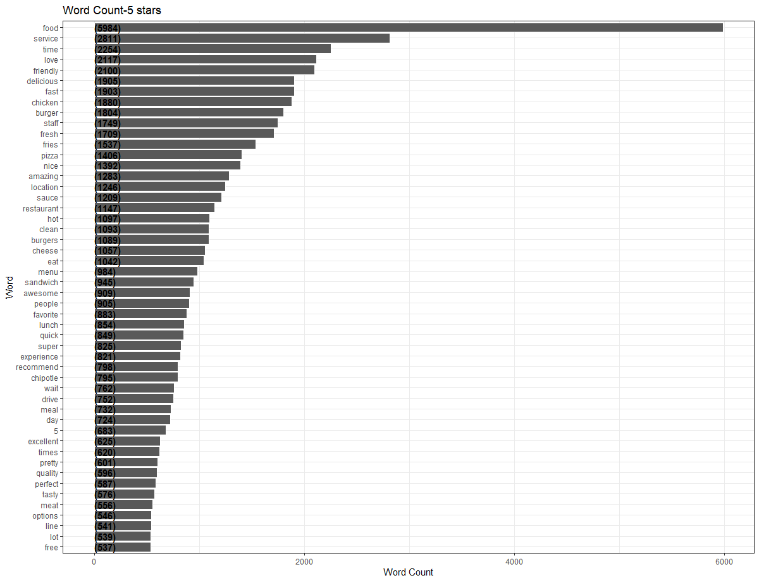
Our data set is obtained from Yelp, a company recently released some of its reviews to the public. The data are stored in JSON format, including information of reviews, businesses, users, and tips. Given our study objectives, we first select businesses in business json file with “fast food” listed in 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 serving fast foods located in the states of Illinois, Ohio, Wisconsin, and Pennsylvania and reviewed on Yelp. 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 50 words appearing in all reviews, but also for each star level, as shown below.



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 dependent 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. The Lasso regression selects 3,204 from 15,163 words. 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 examine the significant words that are related to the four major targeted areas. The selected estimates are listed in the table underneath.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Word | parameter estimates | p-value | Word | parameter estimates | p-value |
| Well\_done | 0.89 | 0.01 | greet | 0.34 | 0.03 |
| dissatisfied | -1.35 | 0.00 | 40\_minutes | -0.64 | 0.04 |
| Poor\_service | -1.07 | 0.01 | Quick | 2.10 | 0.04 |
| decent\_price | 0.62 | 0.02 | always\_good | 0.87 | 0.02 |
| apologies | -1.79 | 0.05 | for\_15 | -0.38 | 0.00 |
| friendliness | 2.10 | 0.05 | for\_20 | -0.79 | 0.01 |
| is\_cold | -1.64 | 0.02 | Great\_food | 0.80 | 0.05 |
| dirty\_tables | 1.29 | 0.05 | garbage | -0.75 | 0.04 |
| not\_fresh | -1.81 | 0.02 | slow\_and | -0.73 | 0.00 |
| horrible\_service | -0.73 | 0.04 | Fast | 0.68 | 0.04 |
| Huge | 0.61 | 0.02 | attitude | -0.23 | 0.02 |
| over\_15 | -0.91 | 0.01 | clean\_and | 0.21 | 0.05 |
| Delivery | 0.63 | 0.05 | minutes\_to | -0.23 | 0.02 |
| kids\_meal | 1.02 | 0.03 | in\_line | -0.31 | 0.00 |
| incompetent | -1.05 | 0.02 | rude | -1.35 | 0.02 |
| their\_job | -0.33 | 0.04 | the\_service | 0.09 | 0.02 |
| ingredients\_and | 0.74 | 0.03 | waited | -0.19 | 0.00 |
| were\_delicious | 0.37 | 0.05 | is\_always | 0.19 | 0.03 |

We therefore create a list of words related to the four areas we are going to explore. The wordlist is as follows {"fast", "slow", "tasty", "delicious", "quick", "home", disgusting", "dirty", "fresh", "friendly", "cheap", "expensive", "stale", "quickly", "clean"}. We examine how each aspect affects customers’ review scores, for which we use Lasso method. To prepare the data for Lasso, we first transform the list of tokens into vector space and then create document-term matrix (DTM) by using the R package “text2vec”. In Lasso regression, we run the regression of stars on the DTM of the wordlist. We choose the optimal lambda by using cross validation, which is equal to 0.0012. To explore the statistical significance, we use R package “selectiveInference” to calculate p-value of parameters of each word in the Lasso regression. The regression results are present in table 1.

Table 1. Effects of selected aspects on customers’ review scores.

|  |  |  |
| --- | --- | --- |
| **word** | **estimates** | **p-value** |
| stale | -0.88764 | 0.00 |
| expensive | -0.06302 | 0.11 |
| disgusting | -1.1542 | 0.00 |
| cheap | 0.161415 | 0.00 |
| dirty | -0.8989 | 0.00 |
| quickly | 0.292146 | 0.00 |
| tasty | 0.569886 | 0.00 |
| slow | -0.84103 | 0.00 |
| clean | 0.337075 | 0.00 |
| fresh | 0.513322 | 0.00 |
| friendly | 0.830238 | 0.00 |
| fast | 0.193718 | 0.00 |