Course: DSCI 510

Assignment: Final Project

Name: Liaoliao Wei

Data Analysis of Restaurants around USC

Introduction and motivation

Los Angeles is famous for its culturally diverse and delicious food. As a student at the University of Southern California (USC), the advantageous geographical position of the school provides me with tons of food options. Therefore, a food recommendation App, such as Yelp, plays a key role when deciding which restaurant I should go to, and the rating of every restaurant is a useful indication for customers. In this project, I collect information about 500 restaurants within five kilometers of USC from Yelp. Through analyzing the reviews of every restaurant, the first question I want to explore is (1) what makes a restaurant score high/low? (2) Are people more willing to write detailed reviews for restaurants with good reputations?

Also, with the development of the influencers culture, customers are willing to choose a restaurant that has been trending online/recommended by some Internet celebrities. This kind of restaurant always has more reviews on the food recommendation App. So, what are the characteristics of a popular restaurant? Specifically, (3) do restaurants with low prices are the most popular type?

At the same time, food safety is an issue that more and more customers pay attention to. However, it is hard to tell a restaurant's food safety or not through shortly having dinners at it or ordering a pick-up. Thus, the health inspection for restaurants and food markets conducted by the

public health department becomes an important reference. In this project, I focus on (4) whether restaurants that are more popular (have more reviews on Yelp) or have a better reputation (have a higher rating) put more effort to make their place and food healthier (have a higher health inspection score).

I hope the results from this project can provide some meaningful insight for customers, especially students of USC, to choose restaurants around our school.

Description of data sources

Three essential datasets are used in this project.

Firstly, the Yelp_restaurant_USC. csv dataset shows information about 500 restaurants around USC (within five kilometers[1]), including the name, location, review count, price, rating, and URL. All the information was collected by the Yelp Fusion API (https://www.yelp.com/developers/documentation/v3/business_search).

Secondly, by using a web scraper, reviews of the 500 restaurants in the first dataset were collected and written into the Yelp_Review.csv document. However, since Yelp restricts the max entries of the web scraper, this project only scraped ten newest reviews of every restaurant. The word count and the number of pictures of every review was calculated.

Thirdly, the Restaurant_health _score.csv dataset was downloaded from the County of Los Angeles Open Data (https://data.lacounty.gov/Health/LOS-ANGELES-COUNTY-RESTAURANT-AND-MARKET-INSPECTIO/6ni6-h5kp). This document records environmental health inspection results (specifically the health score column in the dataset) for

[1] According to Yelp, "the actual search radius may be lower than the suggested radius (specifically 5,000 meters) in dense urban areas, and higher in regions of less business density."

restaurants. Since this is a very large dataset with some redundancy or irrelevant data, it was cleaned by pandas before being used: (1) Duplicated rows were dropped. (2) The restaurants located beyond Los Angeles city (for example Long beach) were dropped. (3) Many restaurants have more than one inspection result and this project only kept the up-to-date one. The cleaned dataset is saved as HealthScore_Clean_Data.csv.

Besides these four datasets, dataset 1 and dataset 3 are merged with the key 'name' (name of the restaurant), 'address', and 'zip' so that the restaurant can be matched with its health score. However, for some unknown reason, not all the restaurants have an inspection result and there are some missing values of the health score. The fifth dataset is named as Score_add.csv.

Data analysis

(1) What makes a restaurant score high/low?

In this project, restaurants with a rating higher than 4.5 are defined as high-rating restaurants. Restaurants with a rating lower than 3.5 are defined as low-rating restaurants. To answer this question, I did a text analysis for their reviews[2].

Through the word clouds of these two types of restaurants, high rating restaurants and low rating restaurants both have positive word clouds (both have comments like good, great, nice, and amazing). However, a big "don't" is shown in the word cloud for low rating restaurants. For customers, flavor, service, location, and price are most important evaluation dimensions. High rating restaurants may have easier parking than low rating restaurants.

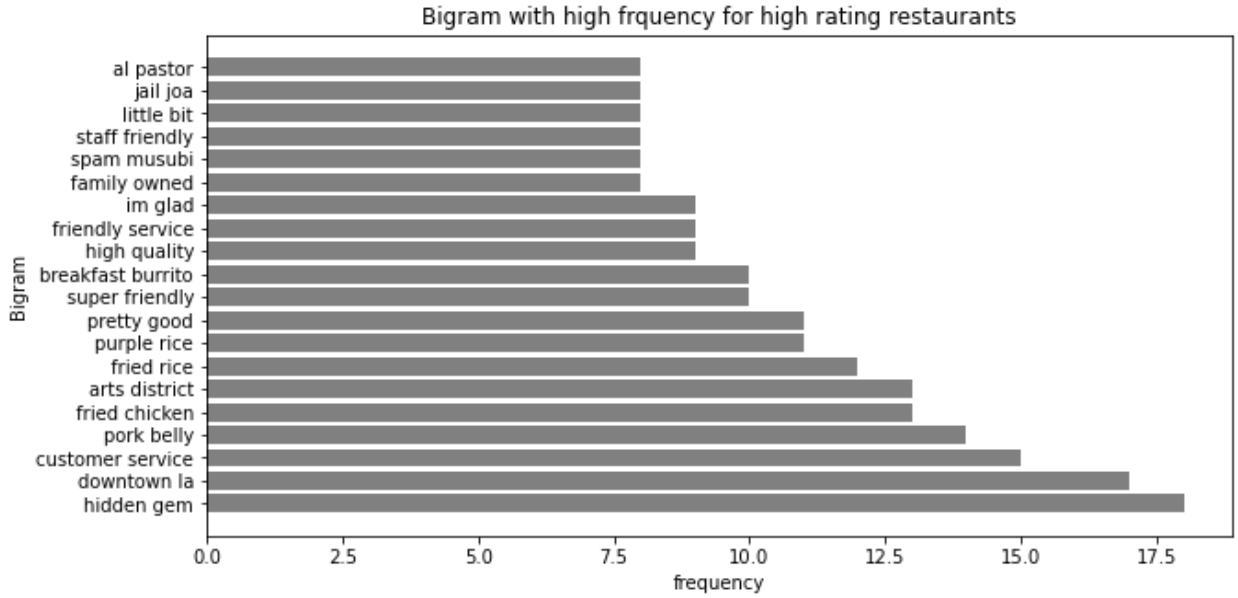
[2] The method of text analysis that I refer to: https://towardsdatascience.com/text-mining-and-sentiment-analysis-for-yelp-reviews-of-a-burger-chain-6d3bcfcab17b

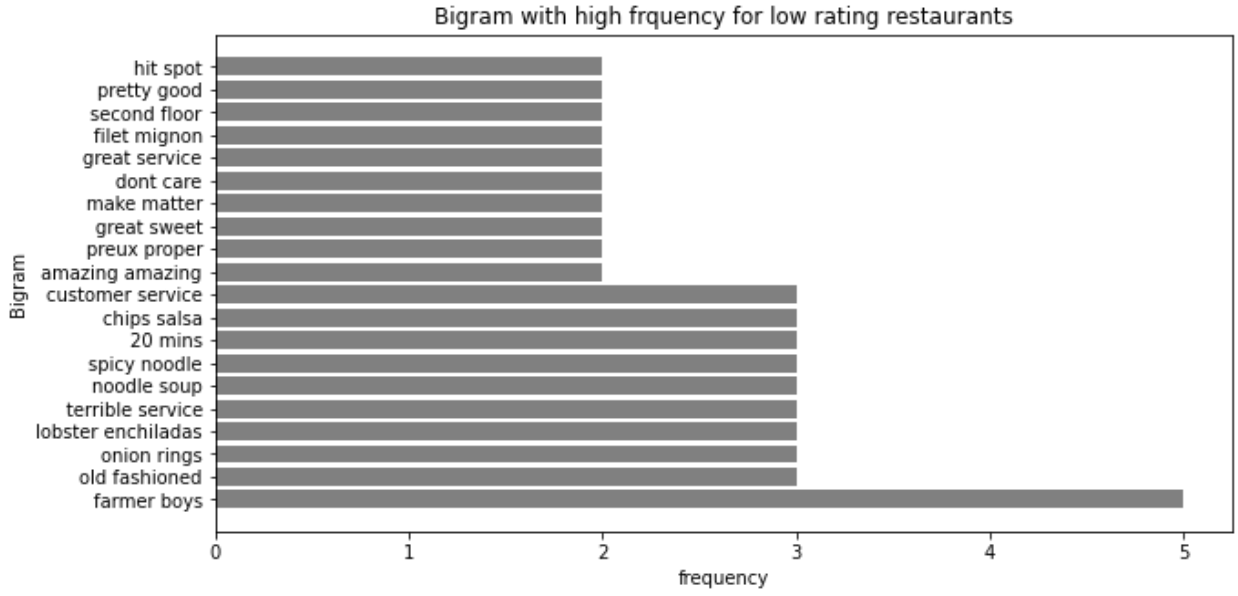
To further understand customers' attitudes, I analyzed pairs of words with their original sequence and see which of them appear most frequently.

In terms of high rating restaurants, these places provide customers not only good food, but also good location (downtown LA, art district) and friendly service. The food mentioned in these restaurants frequently is burrito, rice, fried chicken, and pork belly.

For the low rating restaurants, they have some negative bigrams, such as terrible service and old fashioned. People may not like the environment and service of these restaurants.

However, reviews on Yelp from customers are positive overall regardless of the rating of the restaurants is high or not.





(2) Are people more willing to write detailed reviews for restaurants with good reputations? The OLS regression was performed regarding this question. Through the results, both the length of the review (p = 0.000) and the picture counts of the review (p = 0.000) have a significant correlation with the rating of restaurants. Specifically, a restaurant with a longer review (I used the average review length for every restaurant), and if customers attach more pictures in their reviews, the more likely the restaurant to have a high rating.

OLS Regression Results

Dep. Variable:	Rating	**R-squared (uncentered):**	0.858
Model:	OLS	**Adj. R-squared (uncentered):**	0.858
Method:	Least Squares	**F-statistic:**	1494.
Date:	Mon, 09 May 2022	**Prob (F-statistic):**	5.94e-210
Time:	18:25:51	**Log-Likelihood:**	-931.07
No. Observations:	495	**AIC:**	1866.
Df Residuals:	493	**BIC:**	1875.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Length	0.0045	0.000	19.115	0.000	0.004	0.005
Pics_Count	0.3827	0.083	4.598	0.000	0.219	0.546

Omnibus:	47.498	**Durbin-Watson:**	1.504
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	68.801
Skew:	-0.682	**Prob(JB):**	1.15e-15
Kurtosis:	4.214	**Cond. No.**	836.

(3) Do restaurants with low prices the most popular type?

First, I used the one-way ANOVA analysis to figure out whether four different pricing level restaurants have different popularity. According to the result of the F-test ($f = 2.877$, $p = 0.036$), there might be significant differences between these three groups.

Then, I performed the post-hoc analysis to see the details. The result shows that pricing level1 ($) and pricing level3 ($$$) have a significant difference in popularity. Specifically, the pricing level3 restaurants are significantly more popular than pricing level1 restaurants.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

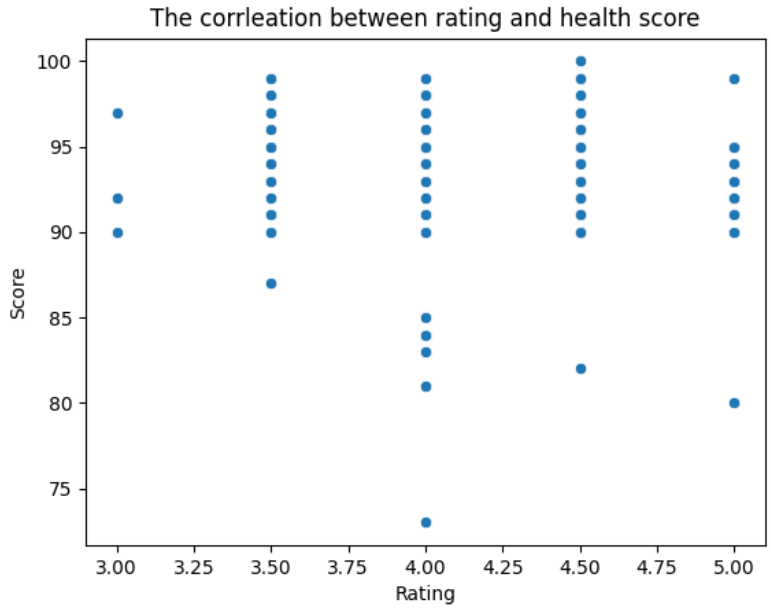
group1	group2	meandiff	p-adj	lower	upper	reject
1.0	2.0	274.7822	0.4627	-210.053	759.6174	False
1.0	3.0	855.1108	0.0206	93.8776	1616.3439	True
1.0	4.0	80.313	0.9	-1034.2828	1194.9087	False
2.0	3.0	580.3286	0.1094	-82.2081	1242.8653	False
2.0	4.0	-194.4692	0.9	-1244.1371	855.1987	False
3.0	4.0	-774.7978	0.3456	-1977.384	427.7883	False

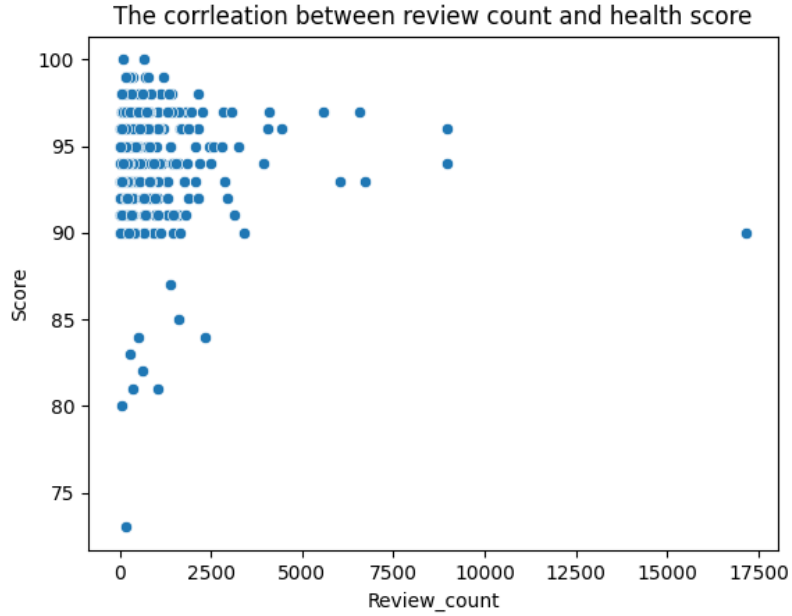
(4) Whether restaurants that are more or have a better reputation put more effort to make their place and food healthier?

The dataset Restaurant_health _score.csv was used to explain this question. Through correlation analysis between review counts rating, and health score, there is no significant negative correlation between the review count of the restaurant and its health score ($r = 0.004$, $p = 0.942$). Also, the rating and the health score have no significant relation to each other ($r = 0.058$, $p = 0.307$). Therefore, a restaurant with a high rating and popularity among customers does not mean its food is healthier.

Correlation analysis results between review count, rating, and health score

	Health Score	*P*-value
Review count	0.004	0.942
Rating	0.058	0.307





Conclusions

Through the data analysis, I make some conclusions as follows. Firstly, customers tend to write positive reviews on Yelp regardless of the restaurant is good or not. Another possibility is that the food in low-rating restaurants is good, but this kind of restaurant has bad customer service or environment. Also, restaurants with high ratings always have longer reviews and rich pictures. Customers can use these standards to choose restaurants. Thirdly, the cheapest pricing level restaurants are not as popular as the expensive ($$$) ones. Finally, the popularity or reputation of a restaurant cannot prove its food is safe enough. Customers who care about this issue should check the health inspection score of this restaurant directly.

Since the information and reviews of these restaurants change time to time, the results may change accordingly in the future.