STAT 628 Module3: Analysis on Yelp Reviews Data

Huitong Kou, Zihang Wang, Peibin Rui

Introduction

There are tens of thousands of reviews for different types of businesses. Our analysis mainly focuses on restaurants which serve steak in several US cities such as Madison and Pittsburgh. Among these restaurants, our specific goals are:

- Investigate the relationship between ratings and words in reviews from two different aspects: foods, non-food items.
- Provide useful advice not only for improving the ratings of existed steak restaurants on Yelp, but also for opening a new steak restaurant based on our analysis.

Data Cleaning

Our dataset contains a subset of million reviews from restaurants in Madison (U.S.), Cleveland (U.S.), Pittsburgh (U.S.) and Urbana-Champaign (U.S.) released by Yelp. The restaurants with at least 3 reviews older than 14 days are included and only reviews that were recommended at the time of the data collection are included

Limited by the computing power, we chose to focus on restaurants whose category contained the word “steak”. We followed the standard practice in NLP using the software package *nltk* in Python and split, clean and recreate the reviews. In order to match reviews with corresponding restaurants, we combine the tables created from *business.json* and *review.json*. To get some insights on how words in reviews are related with Yelp ratings, we created new columns by calculating the word frequencies in each review.

After filtering and combination, our primary cleaned data set *steak_cleaned.csv* has 33629 rows and 6085 columns.

Exploratory Data Analysis and Key Findings

1. Insights on different types of steak

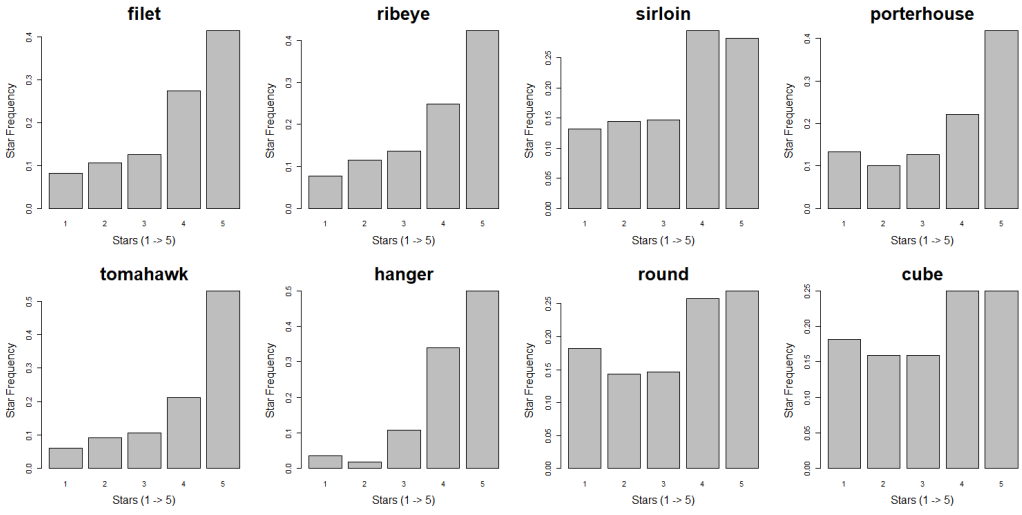

Figure 1

Firstly, we looked at the common types of steaks which appear many times in the comments and investigated the distribution of their ratings. We found that reviews mentioned filet, ribeye, porterhouse, tomahawk and hanger steaks tend to have more 5-star ratings. The portion of their 5-star comments is significantly over 45%.

However, although reviews mentioned sirloin, round, and cube steak also have a large proportion of high ratings, customers seem to be pickier since they also gave many low ratings on these steaks and the 5-star comments is roughly less than 30%.

2. Insights on factors other than steak

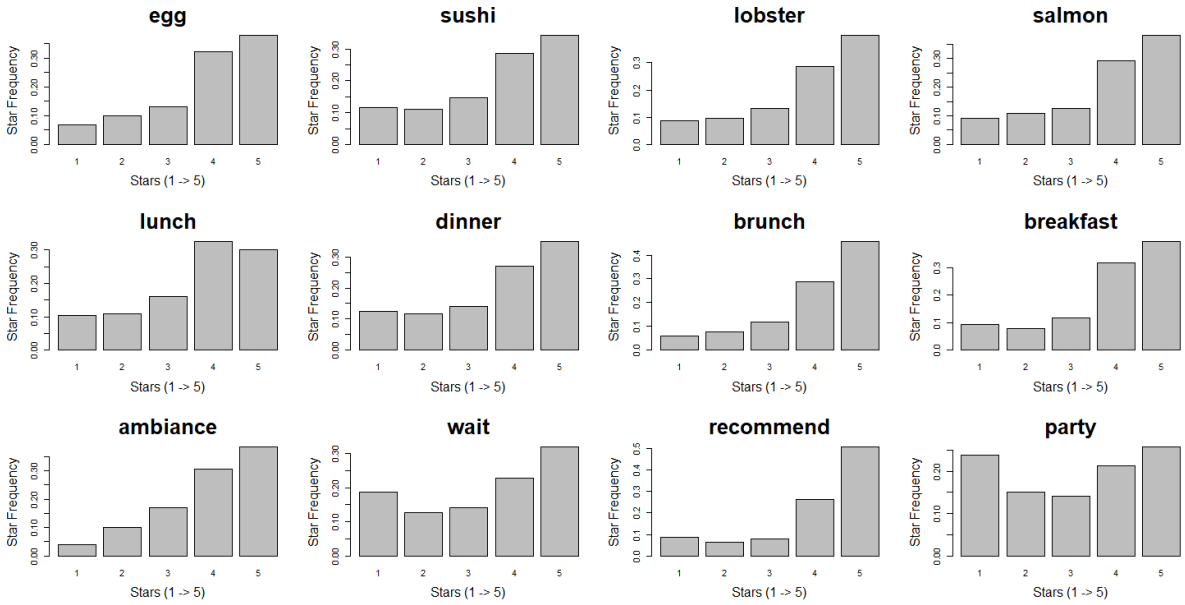

Figure 2

According to these plots of star distribution, we found that generally customers tended to give higher ratings if they mentioned some side orders other than steak. For examples, egg, sushi, lobster and salmon are more frequently mentioned as the star goes up, especially from 1-star to 4-star.

As for the meal time, all of them have a large proportion of high ratings. And brunch seems to be more popular than other meal times.

Non-food factors also play an important role in reviews' rating. Clearly, customers don't like waiting for seats and having meals when there is a party. Ambiance is one of the environmental factors that customers may pay more attention to or have higher expectation. A friendly and quiet ambiance can make people feel comfortable. And people tend to higher ratings if the steak restaurant is recommended by friends or relatives.

Statistical Analysis on Steak Restaurants

1. T-tests on Business Attributes

It's noticeable that the *business_city.json* contains many useful attributes of each restaurant. We decided to conduct t-test on some of the attributes to see whether different levels of some attributes can make a statistically significant difference on a restaurant's stars.

We actually conducted the t-tests on different subsets since the attributes each restaurant has differ from others. Checking the variance equality of each 2 subsets is the first step. Then t-tests were conducted accordingly using the *t.test* function in R.

Figure 3 shows the overall star distributions of all steak restaurants. There are seven attributes selected and the p-value of each t-test is listed in table 1. According to the results of t-tests, with significance level of 0.05, the RestaurantsReservations(True) , RestaurantsAttire(Dressy), OutdoorSeating(True) and RestaurantsDelivery(False) can statistically leads to significant higher stars. The NoiseLevel(Quiet), WiFi(Free), RestaurantsGoodForGroups(True) didn't statistically matter.

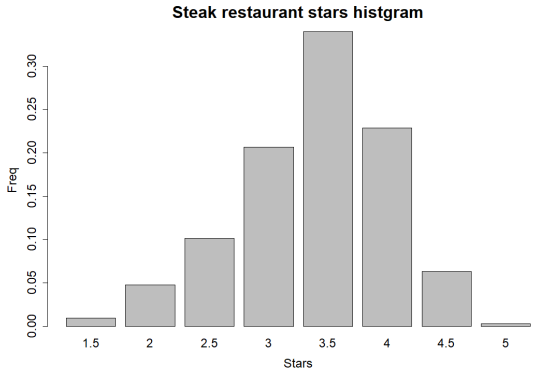

Figure 3

attribute	p.value
RestaurantsReservations	0.040536672387
NoiseLevel	0.115368255456
RestaurantsAttire	0.000007760033
WiFi	0.799610824593
OutdoorSeating	0.044275411757
RestaurantsDelivery	0.012348344280
RestaurantsGoodForGroups	0.454369986827

Table 1

2. Multiple Linear Regression Model

Data-Driven Recommendations and Actionable Plans

1. Advice on opening a new steak restaurant

For the type of steaks:

- Focus on tomahawk, hanger and poterhouse steaks at first and emphasize them on your menu.
- Filet, ribeye, sirloin and porterhouse steaks are not bad to consider.
- Make sirloin, round and cube steaks inconspicuous on your menu or avoid offering them.

For food other than steaks:

- Set a more varied range of wines and beers offered at your restaurant
- Make sure the egg and cheese served at your restaurant taste great
- Consider hiring patissiers and serve desserts
- Salad is important

For other environmental factors:

- Set up a proper reservation system is important.
- Offer well-designed attire to your waiters/waitresses.
- Look for possible locations with outdoor seating.
- Don't provide food delivery service.
- Avoid investing in WiFi or soundproof materials at the beginning.

- Allocating too much spaces for groups specially is unnecessary

2. Advice on improving an existed steak restaurant

For the kinds of steaks provided:

- Try to improve tomahawk, skirt, hanger and flank steaks if one of them has brought your restaurant many low-star comments since the customers are not so picky about them.
- Advertise your sirloin, round and cube steaks if one of them has brought your restaurant high-star comments. It is very praiseworthy to have highly rated sirloin, round and cube steaks.

For other foods provided:

- Improve the quality of wines and beers
- Cheese and salad are always focus points
- Start serving desserts may help

For other non-food factors:

- Improve or set up your reservation system.
- Pay attention to the attire of your waiters/waitresses.
- Check if it's possible to offer outdoor seating.
- Don't provide food delivery service.
- Allocating too much spaces for groups specially is unnecessary

Conclusion and Discussion

HK contributed to the coding and writing of summary outline and the food and t-test part.
ZW contributed to the coding and writing of non-food part and the slides of presentation.
PR contributed to the coding and writing of the shiny app and the slides of presentation.

Contributions

References