STAT 628: Module 3

Yinqiu Xu, Xiaofeng Wang, Milica Cvetkovic

2020-11-30

Introduction

Yelp is a powerful platform for business owners to collect information about their business and customer satisfaction. In this project, we analyzed the customer's reviews and various other attributes of restaurants and pubs in Wisconsin, trying to extract useful information and give suggestions to businesses. Among these establishments, our specific goals were to analyze the most important facilities and services that customers prefer and then provide suggestions to pubs to improve their star ratings. The rating of a restaurant is measured in stars from 0 to 5. Specifically, we asked how different facilities affected the star ratings, the affect the work hours might have on the rating and what additional services could improve the rating. We first preprocessed of the data, then we did the exploratory data analysis. Next, we built a multiple regression model by stepwise selection to predict the rating of a business. In the following sections, we cover the details of our analysis.

Preprocessing

Data and Sample Size

All the data provided by Yelp are stored in four JSON datasets. The first dataset contains information about businesses, including different facilities offered and the stars rankings. The second dataset that was useful for our analysis contained the reviews and the number of stars each reviewer gave to a business. The third dataset contained the text of the shorter reviews called tips. We mainly focused on the following features: stars, working hours and facilities/attributes of the data in business JSON file, and the content of reviews in review json file. After filtering all the open restaurants with a full bar service in Wisconsin, we got 466 pubs with 69 attributes and its stars ratings, 50569 reviews, and the corresponding business id.

Clean Attribute Variables

To obtain attributes of each pub, we separated *BusinessParking*, *GoodForMeal*, and *Ambience* into several binary variables like valet, or romantic. These features are then cleaned for redundant characters(i.e. "u'free'" is the same as "free"). To deal with the missing values of both nominal and ordinal variables, we interpolated by the mode.

Sentiment Analysis

After a thorough preprocessing of the data, we focused on further investigation into the reviews and tips. Before the sentiment analysis and developing the polarity score, we tokenized the text of the reviews, by separating lengthy sentences into single words and their counts. Next, we removed the stop words, which are the words that don't contribute to the sentiment like "the" or "he/she". To obtain the sentiments of the reviews for each pub, we created a new predictor called positive review ratio by dividing the number of positive reviews by the number of reviews of each pub. Then we counted the number of positive words and negative words in each review. Finally, we calculated the sentiment of each review by subtracting the number of negative words from the number of positive words. The sentiments of each word are defined and stored in a lexicon which is provided by the *tidytext* library. After conducting sentiment analysis of each review, we grouped the reviews by business id to get the number of positive reviews. The two separate datasets, the review dataset and the business dataset, were merged by business id column.

Exploratory Data Analysis

Part of our EDA was to plot a histogram and visually explore the star distribution in the proportion to a specific attribute. Here we will focus on a few significant findings.

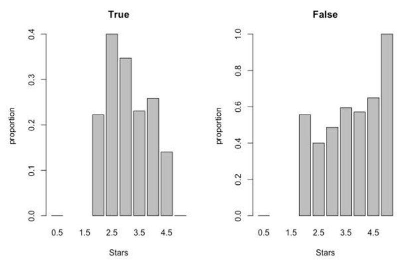


Figure 1: Delivery

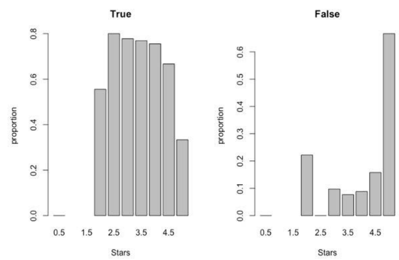


Figure 2: Takeout

In the first example (Figure 1), we tested the importance of delivery option and its effect on the stars ratings. From the plot it seems like the delivery does not affect the star ratings. In our second example (Figure 2), we tested the hypothesis if the high ratings are NOT related to takeout being offered. From the plot, it is visible that false hypothesis seemed to provide higher ratings. In other words, the pubs that offered takeout received higher ratings. This was confirmed in the Chi Squared test of the hypothesis. Then we explored

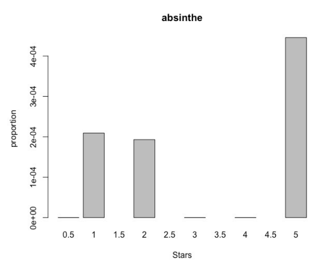


Figure 3: Absinthe

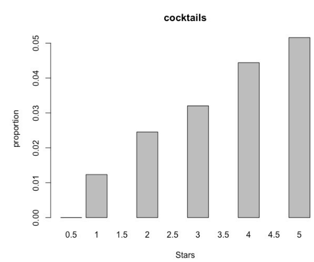


Figure 4: Cocktails

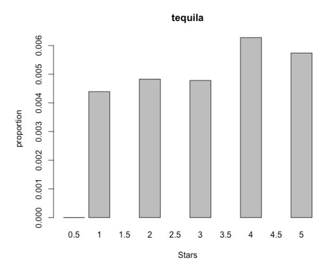


Figure 5: Tequila

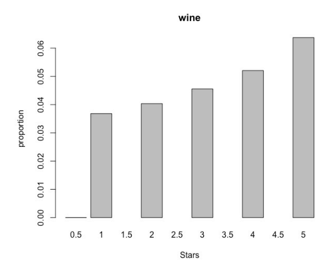
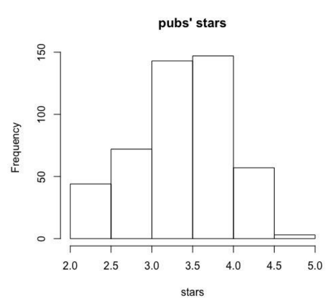
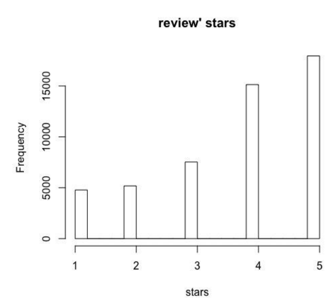


Figure 6: Wine

the plots that showed us the frequency of each word in the reviews and how it was related to the distribution of the stars. As an example, here are the graphs showing the star distribution based on the various alcohol beverages mentioned in the reviews. We used Wilcoxon test to test our hypothesis. The results showed that offering certain services like cocktails and absinthe are resulting in higher star reviews.





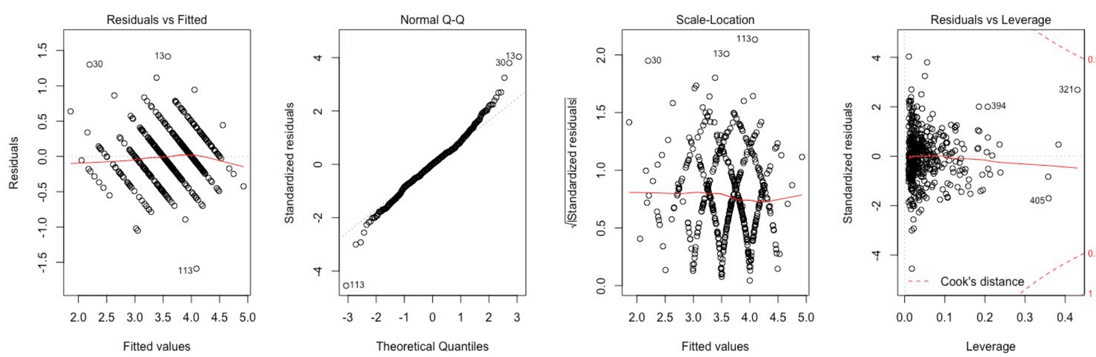
Finally, we looked into overall distribution of the star ratings among all the restaurants and bars in Wisconsin, and the distribution of the stars among each review. There are no bad ratings that are less than 2 stars, and the number of ratings that are 5 stars is very small. When we look at the number of stars each reviewer gave, it seems like in general the users tend to give more positive than the negative reviews.

Statistical Analysis and Model Diagnostics

In the table, we provide the summary of our statistical analysis. The data driven business plan is elaborated upon in the next section.

Hypothesis	Method	p-value
High Ratings are not related with takeout offered	chisq-test	0.466
High Ratings are not related with a pub being good for groups	chisq-test	0.528
High Ratings are not related with the presence of TVs	chisq-test	0.002
High Ratings are not related with the word 'love' in reviews	wilcox-test	0.071
High Ratings are not related with offered Wifi	chisq-test	0.911
High Ratings are not related with offered delivery	chisq-test	0.007
High Ratings are not related with a pub being good for dancing	chisq-test	0.763
High Ratings are not related with the work hours on Friday	wilcox-test	0.125

After finalizing the EDA and after some statistical tests gave us the thorough insight in our data, we decided to use multiple regression model for answering our statistical questions. One of the main questions we wanted to answer was whether or not the work hours of a pub have any affect on the star rating. After standardizing work hours for each day of the week, we preceded to prepare other attributes for the regression model. These attributes were the predictors that we used in our model that we found that were significant to the star rating in our exploratory analysis. The predictors were work hours, restaurant attire, noise level, presence of WiFi, takeout and other attributes provided in the business dataset, along with the sentiment analysis of the reviews. The model is □formula goes there□. From the model diagnostics plots, we see that there is no significant deviation in QQplot and distribution trends in residual plots, thus it follows that our assumptions hold. In addition, we do ANOVA analysis and find that TV services, credit card acceptance, delivery services, and reservations requirement have significant influence to the rating of a pub with a p-value smaller than 0.001. Also, from the model we see that the percentage of positive sentiment review is really important to ratings.



Data-Driven Business Plan

Reflecting upon all of our findings, we will make recommendations for pubs in operations, service and food aspects.

Operations

- Offering the d**delivery service does not increase the ratings**. We suggest to cancel delivery services to save money.

- If a pub has TVs, they should play sports or mainstream music. If a pub does not have TVs, the **investment to install TVs should be made**. From our findings, the pubs without TV will have a tendency to get high ratings, even if the percentage is small, but it shows the trend. Then we do the chisq-test to see whether it can influence the stars, because it's p-value is 0.002, we refuse the null hypothesis, thinking TV can influence the stars. In addition, 3% of the tips with high ratings mentioned "atmosphere" as an important factor in their decision.

- To decrease spending, **do not invest in the accommodating large groups, providing WiFi or providing space to dance**. Neither of these factors don't seem to have any affect on a pub's ratings. This would be a good place to decrease spending and invest in some other meaningful areas.

- Work hours during the day do not affect ratings. This suggests that there might be an opportunity for the pubs to close earlier on some days, or close for a few hours between lunch and dinner to save some money. However, it seems like it would be good for a pub to be open late on Mondays.

- Pubs that offer late-night food service could see an increase in ratings. Whether a pub is good for meal at late night has a positive significant influence on pubs' rating with p-value equals to 0.0002.

Service

- To increase the ratings, **invest in staff training and bringing in more skilled staff**. From the reviews with low ratings, 36.8% people mentioned "service", 13.7% people mentioned "server", 14.5% people mentioned "staff", so pubs should pay attention to it. From the tips with high ratings, 7% people mentioned "service".

- To keep the high ratings, a pub should **keep encouraging staff, and potentially provide some extra benefits**. This will keep the motivation high and it will reflect on service.

Food

- To improve the ratings, a pub should **ensure the excellence in taste and quality**, with the emphasis on cheese, burgers, fish, fries and the beer selection. In addition, the pubs that are already satisfied with their ratings, in order to keep the customer satisfaction, they should safely explore and diversify their offers. To support our claim, we found that from the reviews with low ratings, 53% people mentioned "food", 21% people mentioned "drink", 17% people mentioned "price", to be the essential to a pub's rating. Among the tips with high ratings, 14% people mentioned "food" as an important factor.

- **Offer brunch**: in the reviews for the pubs that have high ratings, 2% people mentioned brunch when reviewing the pub.

Conclusion

Overall, Yelp dataset provided in depth insights in the operations of businesses through the analysis of the reviews and business profiles, and it should find its place in improving strategies and business plans.

Contributions

Yinqiu Xu contributed to the preprocessing code, the sentiment analysis, the model and the model diagnostics. She wrote a part of the summary and set up the initial presentation deck. Xiaofeng Wang contributed to the preprocessing code, the statistical analysis and to drafting the data-driven business plan. He also wrote a part of the summary and created the Shiny app. Milica Cvetkovic contributes to the preprocessing code and the sentiment analysis. She wrote and edited the summary, created and edited the presentation deck.