STAT 628 Data Science Practicum HW3

Zhihao Zhao, Haoran Teng, Yuanyou Yao

1 Motivation:

This project concerns about the following aspects:

1. Features for hotels to improve their star ratings. For example, should hotel provides free WiFi?

2. How to improve the performance of hotel to better serve their customers according to customers' reviews on Yelp. For example, is it important to update old facilities?

2 Background Information and Data Pre-Processing:

The data is provided by Yelp, an Internet company found in 2004 which aims to " help people find great local business" via establishing a platform for users to write reviews and rating their experiences. The data is restored in four .json files and we use Rpackages 'rjson' and 'jsonlite' to read them and convert them into .csv files.

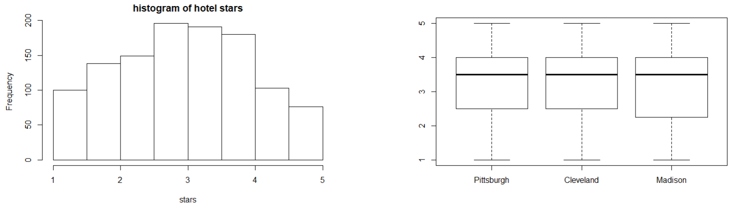
Since there are many different types of sample tag like restaurants, coffeehouses, etc, we first filter out all samples with keyword "Hotel", and then, we remove some of them which are not open according to variable "is_open". Some variables like "address", "postal_code", "latitude" and "longitude" are deleted since they have nothing to do with our analysis.

We decompose every review into words by tokenization and use indicator variables to denote the appearance of words we're interested in. Then analyze the relationship between appearance and star ratings.

3 Exploratory Data Analysis:

The average star rating for hotels is 3.18 and the standard deviation is 1.02 and the associated histogram is shown in Fig 1(a).

The average ratings of hotel in Pittsburgh, Cleveland and Madison are 3.30, 3.30 and 3.19. Although ratings in Madison seems to be a little bit lower than the others, there is no significant difference among hotel ratings in these three cities according to anova F-test (a test to compare the difference among mean of several groups) with p-value 0.602. More specifically speaking, since p-value is much larger than 0.05, we cannot reject the null hypothesis that there is no difference among the mean of hotel ratings in three cities. Besides, Tukey multiple comparisons of means confirms our previous conclusion. We also draw the boxplot for three cities in Fig 1(b), from which we observe that the 1st quarter rating in Madison is much lower than the other two. It indicates that Madison has more proportion of low-rating hotels and they downgrade the whole ratings.



(a) Histogram of hotel star ratings

(b) Boxplot for hotel ratings in three cities

Figure 1: Basic aspects of data

The review numbers according to stars from 1 to 5 are 7208, 3185, 3703, 6499, and 9307. It seems that customers with best or worst experience would more likely to share their feedback to others.

As for hotels with rating ≥ 4, it's not surprising that word "great" is mentioned most with 5644 occurrences in total among all 29902 valid comments. "Staff", "nice", "clean", "friendly", "breakfast" gets 3555, 3062, 2799, 2626 and 2251 accordingly. And for rating ≤ 2, customers tend to use some negative words like "just", "stay", "never", "didn't", "even" but don't give explicit reason. Some of them mention about "location", "bathroom", "old", "shower", "towels". So it might be a good idea for hotel owners to update their old facilities and re-decorate their bathrooms.

4 Key Findings About Hotel Marketing on Yelp

1. Findings:

a). As for noise level, the medians of categories "quiet", "average", "loud" and "very_loud" are 3.0, 2.50, 2.75, 2.50, indicating that customers prefer quiet hotel, which is reasonable. It's surprising that loud hotel gets higher ratings than hotels with average noise level. So we take a closer look at this variable and figure out the reason. There are only six people got loud-noise-hotel among all 72 valid: three of them give 2.5, one gives 3, and the other two gives 4.5 and 4.0. Most of them still give low ratings and this could be a vivid example of how extreme values influence the statistics when sample size is very small. We suggest hotel owners to provide quiet rooms to help customers get sound sleep.

b). For some other factors hotel owners may interest in, several tests are conducted to test the effect of them. The p-values for factors discussed above is shown in Table 1. P-values greater than 0.05 means there is no difference among the levels of factor since we cannot reject the null hypothesis and vice versa. We find that the existence of parking lot cannot significantly lift the star rating. Besides, customers pay little attention on whether hotel provides free WiFi, paid WiFi, or no WiFi according to anova F-test. Besides, there is no significant difference in star ratings in whether the hotel accept credit card, whether hotels allow dogs, and whether they are good for kids. But for hotels with no TV, they outperform others with TV by 1.34 in rating. Maybe TV would be the obstacle for people to get sleep and they may miss their plane or important meetings. So it might be better for hotel owners not to provide TV and it will definitely reduce their cost on decoration.

	ParkingLot	WiFi	CreditCard	AllowDogs	GoodforKids	HasTV
p-value	0.414	0.299	0.134	0.865	0.633	0.0009

Table 1: p-values for several factors

c). We fit a multiple linear regression model with outcome as star ratings and predictors as all words frequency we’re interested in like ”great”, ”clean”, ”old”, ”coffee”, ”breakfast”, etc. And we select some results as follows: the coefficient for word ”great” is 0.926, meaning that the appearance of ”great” would increase star rating by 0.926; on the contrary, word ”old” will deduct 0.679 from rating. The associated p-values of t-statistic are both $< 2e - 16$, which are much less than 0.05, so we reject the null hypothesis that the word has no linear relationship with star ratings and conclude that the linear relationship exists indeed. However, the p-values for neutral words like ”the”, ”room” are larger than 0.05, so they have little influence on star ratings.

d). The estimated intercept in regression model above is 3.04 with 95% confidence interval (2.96, 3.13), meaning that if none of our chosen words appears in the review, interval (2.96, 3.13) will contains the true rating with 95% accuracy.

2.Model Diagnosis:

Since we use linear regression, we assume that the error part is normally distributed, so we use QQ-plot as shown in Fig 2. The plot shows a trend of light-tail distribution, but not severe. We also check whether the error is homoscedastic, but the residual plot has an obvious linear pattern, which indicates lack of fit. Since the purpose of this model is just to detect the relationship between star ratings and certain words, not to explain the variation in star ratings, so we can still draw some useful conclusion from model.

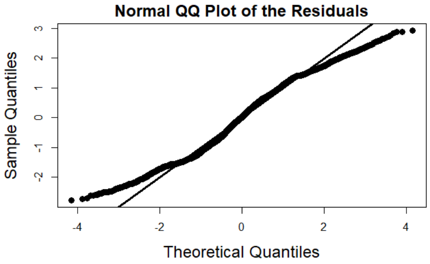


Figure 2: QQ-plot for model diagnosis

5 Recommendation for Hotel Business

1. Recommendation

a). We suggest hotel owners to provide quiet rooms for customers. (See Fig 3)

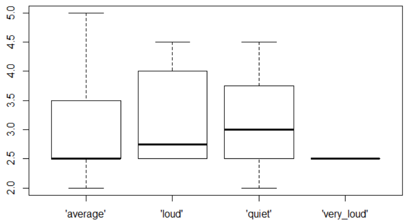


Figure 3: Boxplot for hotel ratings with noise level

b). We suggest hotel owners not to provide TV for customers: on average, no TV gains 1.34 more in star ratings. (t-test p-val=0.0009) c). Since providing free WiFi doesn't seem to influence the star ratings much on Yelp, it might be okay for owners to save money on these and pay more attention on other stuffs. (anova F-test p-val= 0.299)

d). We suggest hotel owners to offer delicious breakfast to customers for higher star ratings. The appearance of word "breakfast" would increase the star rating by 0.272. (p-val $<$ 2e-16)

e). We suggest to hire staffs with friendly and polite behavior to make customers feel better. The appearance of word "friendly" would increase the star rating by 0.567. (p-val $<$ 2e-16)

f).It worth updating old facilities, especially in bathroom: the appearance of word "old" would reduce the star rating by 0.679. (p-val $<$ 2e-16)

g).It's not worth having parking lot or having special place for kids to play in: neither has significant influence on star ratings (both t-test p-val $>$ 0.05)

2.Limitations of Analysis

a). Our Recommendations rely on some assumptions we have made about the data and the analysis like normality and homoscedasticity. But model diagnosis shows neither is perfectly matched, which brings some uncertainty.

b). Our conclusions based on variables in "attribute" part are not very reliable since most people didn't offer any information about these variables so most of them are just "NA". The conclusions are drawn only from very small sample size, which may leads to wrong results or even opposite results, and we happen to have such example when analyzing variable "NoiseLevel".

6 Conclusion

In this project, we utilize different types of analysis to help hotel owners better serve their customers and get higher star ratings on Yelp. Some results are pretty surprising like people don't care much on whether hotel would provide free WiFi. And we deeply understand that our results rely on some assumptions which are not perfectly matched. Besides, every result is associated with 5% type I error. They bring some uncertainty to our conclusion. Anyway, we hope the analysis above could help hotel owners to make reliable decision and get better and better.

7 Contribution:

Zhihao wrote and edited the project report, created and revised code 'datacleaning.Rmd', 'word.Rmd', 'Analysis.Rmd' as well as 'word-regression.Rmd', discussed the main purpose of Shiny App and helped edit it, wrote and edit .pptx for presentation in Page 1-4, 7-8, 11.

Haoran wrote and edited the Shiny app, created and revised code '628_graph.Rmd', discussed the main purpose of the project report and helped to edit it.

Yuanyou edited summary report, discussed the main purpose of presentation and edited.