**Executive Summary**

TEAM MEMBERS: Hangyu Kang, Xiangyu Wang, Ruyan Zhou

**Introduction**

There are more than 40 hotels registered on Yelp in Madison area, we wondered which hotels are evaluated as good hotels and what made them obtain good reputation. Hence,for the project, we had planned to analyze yelp reviews about the hotels in Madison area. The questions for the analysis are:

1. What affects to customers’ evaluation towards hotels especially in Madison area?
2. What is the difference between high rated hotels and low rated hotels in Madison area?
3. How AC hotel (Our virtual client) can be improved?

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

We have a given 4 datasets from the Yelp. In our first step of data cleaning, we merged review\_city.json file to business\_city.json file according to the corresponding business id. Secondly, we filtered out all the rows according to the categories that contains ‘Hotels’ and then selected out the rows according to the business name that contains hotel business related words. Selecting out the row by their name is a necessary process because there are some businesses that are not related to accommodation business but are related to trip and restaurants come with the selected rows. After that, we selected out the rows correspond to Madison area and dropped all unnecessary columns. For the next step, we made word embedding dataset using the text review column. We break the review text into word, getting rid of stop words and special characters, lemmatizing the words, and making it into matrix form. Using the refined dataset, we made another word embedding dataset for the analysis. At first, we reduced the overlapping count to 1; for example, if the word ‘stayed’ are counted as 2 in a row, then we made it to 1. Secondly, we excluded unmeaningful words, by selecting the words which are more used than 3rd quantile (53.00).

**Exploratory Data Analysis (EDA)**

We sorted the words according to their frequency (Figure 1), dividing those words into four different categories (Service and furniture, Facility, Location, Atmosphere), dumping words if those are judged as unnecessary words. We used our rational to do this process.

And we produced relative frequency bar plots for each word like Figure 2, 3.

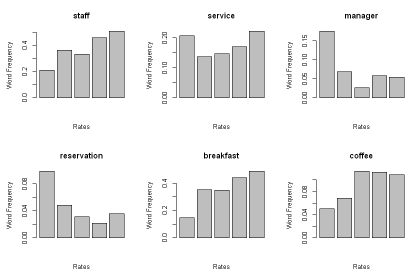
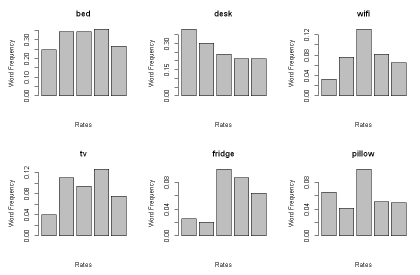
 

Figure 1 Figure 2

Based upon Figure 1 and 2, we concluded the roles of employees are important for determining the evaluation. Some plots are omitted here, but they showed that other services like reservation, provided food, breakfast, buffet, drinks like beer are played important role to determine the evaluation. However, the rates are not significantly differed by the provided furniture like bed, fridge. Price related words, price and money are also somewhat related to the rate.

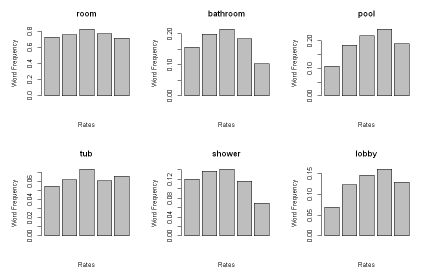
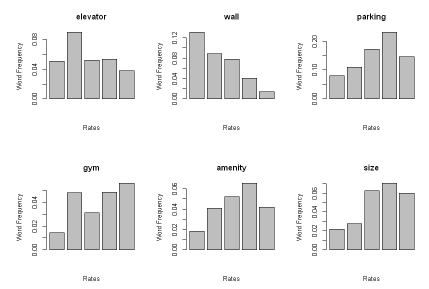
 

Figure 3 Figure 4

From Figure 3 and 4, we noticed that room itself does not play important role to the rating. However, amenities like gym, pool and parking places may affect to the rating. Interestingly, wall shows some significant difference. Some of bar plots also omitted here.

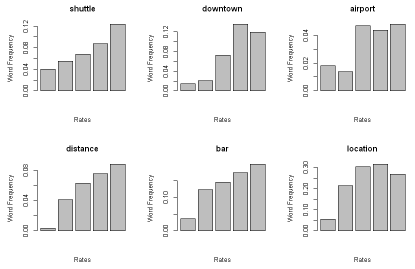
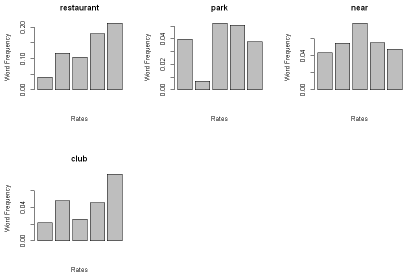
 

Figure 5 Figure 6

From Figure 5 and 6, we noticed that the words related to the location of the hotels and neighborhood restaurants, bars and clubs shows some relationship to the rates. The words associated with transportations also are relatable with the rate.

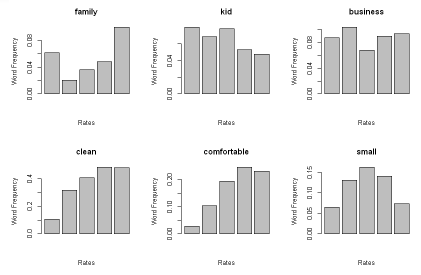
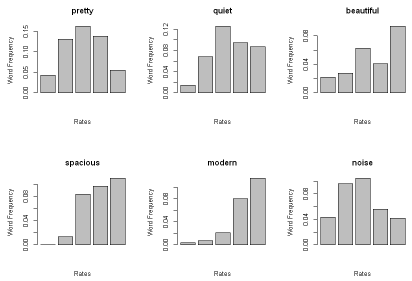
 

Figure 7 Figure 8

From Figure 7 and 8, we noticed customers expect more family-friendly and modern environments to the hotels in Madison. Also, we observed cleanness, roominess and smell of the place is also played important roles in customers’ evaluation.

For more accurate results, we performed multinomial-logistic regression and produced corresponding p-values. Also used backward AIC to select out the words that effect on customers’ rating. Multinomial-logistic regression is used for analyzing categorical variables like stars in our dataset. Basically, it uses proportion of the frequency of the word in each star rate, and see if the ratio is different from each other. And we calculated p-value of the coefficients. If the p-value of the word is less than 0.05, then we can conclude that the word frequency of the word in each star rate are statistically significantly different. In other words, it effects on the customer’s evaluation. Backward AIC is the process to see which variables effect on the outcomes (in our case, star rate) by dropping one variable for each step from the full model (containing all words). If AIC value decreased drastically comparing to others, then we can conclude the variable is important. From this process, we concluded that the words which are staff, manager, breakfast, booked, money, desk, wall, parking, restaurant, downtown, bar, location, modern, clean, pretty, comfortable, spacious, quiet, comfy, smell and dirty are statistically significant words.

**Key Findings About Businesses Hotel (Accommodation) Market on Yelp**

There are 45 hotels in Madison area. The summary of their average rating is the same as the Table2.



Table 2

We separated 45 hotels into two large groups according to the Median value (3.5) of their average rate. If their average rate is lower than 3.5, we classified those hotels as low rated hotels; otherwise, we classified those hotels as high rated hotels. There are 22 low rated hotels and 23 high rated hotels in Madison Area. Each of them obtained 567 reviews and 965 reviews, respectively. Using the selected words above, we compared high rated hotels and low rated hotels in Madison area.

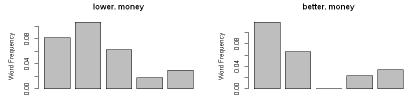
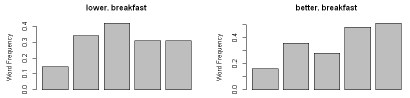
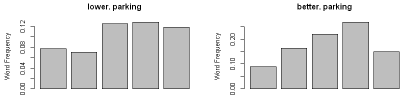
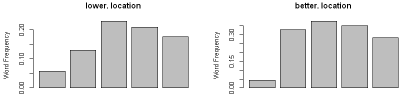
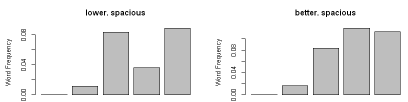
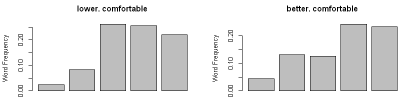
    

Figure 9

From the Figure 9, we noticed that the relative frequency between high rated hotels and low rated hotels quite different in the words, ‘money’, ‘breakfast’, ‘parking’, ‘location’, ‘comfortable’, ‘spacious’. To see whether there are statistical differences in frequency rate of words in high evaluation range (3,4,5) between the low rated hotels and high rated hotels, we performed proportion test.

|  |  |  |  |
| --- | --- | --- | --- |
| Star | 3 | 4 | 5 |
| Money | 0.03809 | 1 | 1 |
| Breakfast | 0.06923 | 0.002298 | 0.002886 |
| Parking | 0.1259 | 0.004478 | 0.6172 |
| Location | 0.04102 | 0.008122 | 0.09249 |
| Comfortable | 0.02812 | 0.8634 | 0.9627 |
| Spacious | 1 | 0.021 | 0.7003 |

Table 3. p-value for Proportion Test

Through the proportion test, we observed that breakfast shows the largest difference. We also compared the distributions and the mean values of the star rate of corresponding words "money", "breakfast", "parking", "location", "comfortable" and "spacious" between low rated hotels and high rated hotels. For testing the differences, we produced a box plots and performed t-tests.

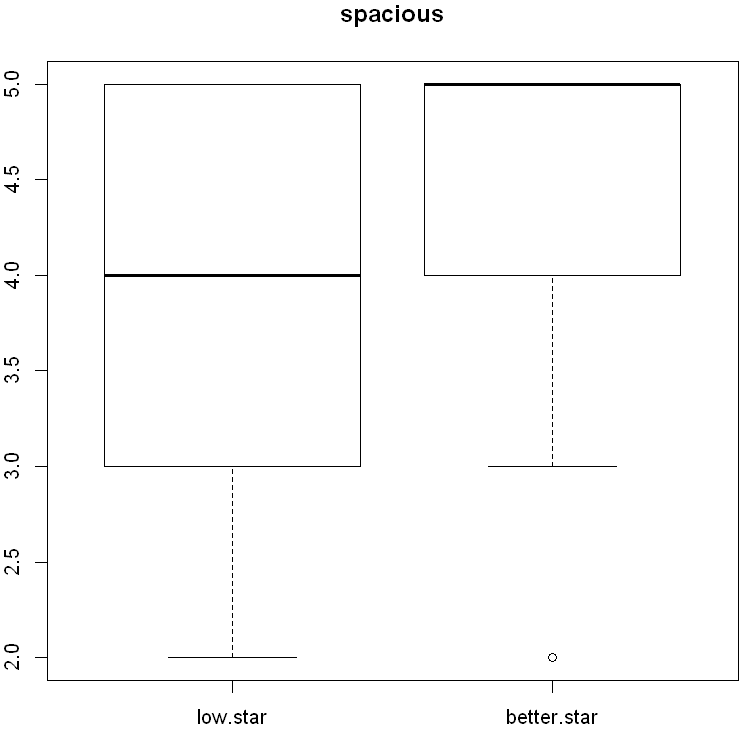
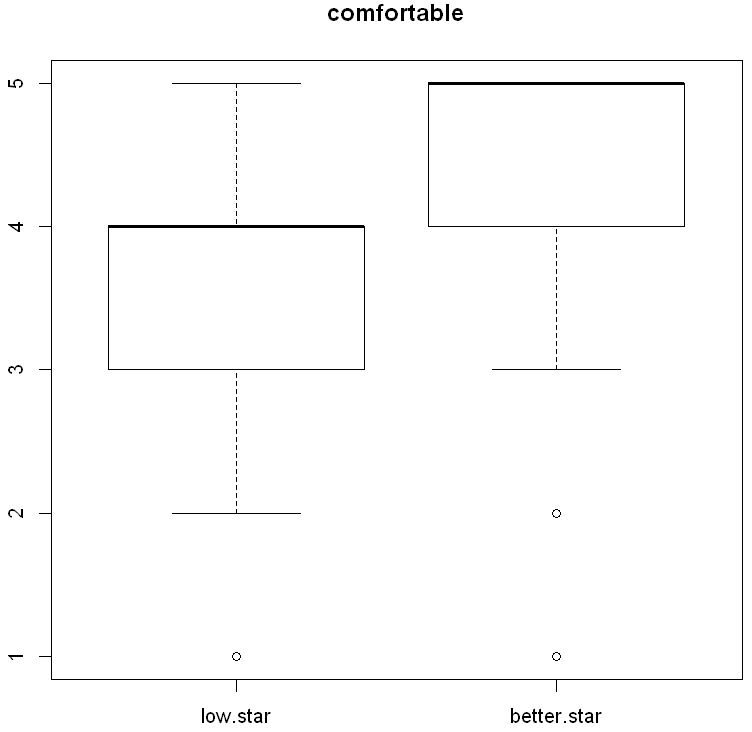
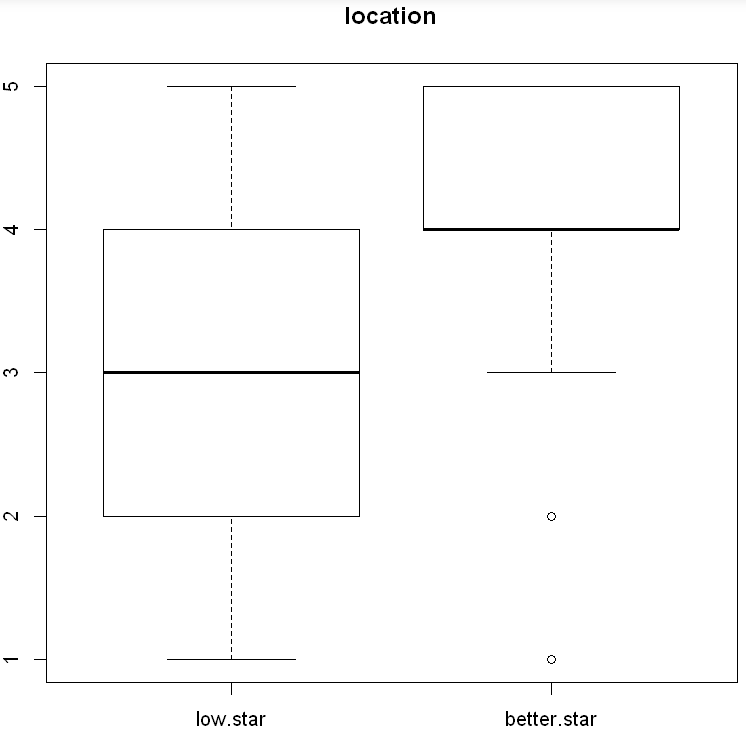
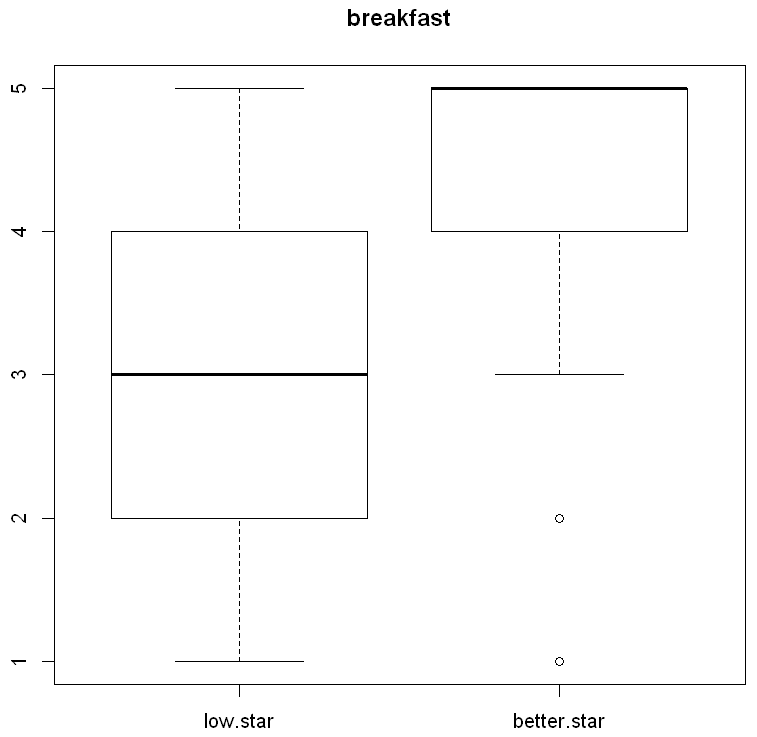


Figure 10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Words | Money | Breakfast | Parking | Location | Comfortable | Spacious |
| p-value | 4.202e-09 | 2.91e-08 | 0.01437 | 4.202e-09 | 2.91e-08 | 0.01437 |

Table . p-values for t.test

From Figure 10 and Table 4, we observed all the mean values of the star rate between high rated hotels and low rated hotels are significantly different.

AC Hotel is one of the hotels in Madison area, it is classified as one of low rated hotels; however, it has the highest average star rate among the low rated hotels. They obtained averagely 3.46 stars and 44 reviews from the customers. We compared this AC Hotel with the high rated hotels. For the analysis, we used all of 21 words which are concluded as significant factors. We performed t-tests to see whether there are differences in average star rates corresponding words between AC Hotel and the high rated hotels. We also produced bar plots that show the frequency proportion of each words at each star rate to show whether there are differences in their distribution. Some words like ‘booked’, ‘money’, ‘manager’, ‘spacious’, ‘pretty’ and ‘dirty’ are excluded since there is not enough number of words in AC hotel reviews.

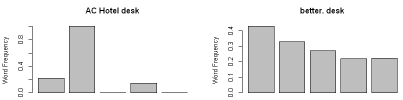
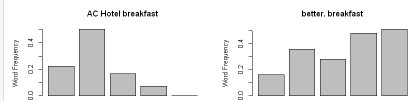


Figure 10

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | staff | breakfast | desk | wall | parking | restaurant | location | bar | downtown |
| p-value | 0.50 | 0.02 | 0.05 | 0.18 | 0.54 | 0.07 | 0.3 | 0.14 | 0.24 |
|  | comfortable | modern | clean | quiet | comfy | smell |  |  |  |
| p-value | 0.76 | 0.4 | 0.42 | 0.69 | 0.09 | 0.79 |  |  |  |

Table 5. p-values for t.test

We observed statistical difference between the star rate corresponding words, “breakfast” and “desk”.

**Recommendation for Businesses Plan**

Based upon the statistical analysis, since the average of customers’ evaluation shows difference in the six words, ‘money’, ‘breakfast’, ‘parking’, ‘location’, ‘comfortable’, ‘spacious’, we recommend the low rated hotel owners to adjust their fare properly, improve their quality of breakfast, secure the parking spots for their customers, provide transportations to overcome the locational problem and change some interior parts of the room to make customers feel comfortable and spacious.

Based up on the second part of statistical analysis, since the average of the customers’ in the two words, ‘breakfast’, and ‘desk’, between AC Hotel and high rated hotels are statistically different, we recommend the AC hotel owner to improve their quality of breakfast for customers and improve the quality of desk in the room.

**Conclusion**

We figured out there are twenty-one words that affect to customers’ evaluations. Among those twenty-one words, six words are influential to determine whether the hotel belongs to the high rated hotel or not. The highest evaluated hotel among the low rated hotels is AC hotel, and as we compared to the high rated hotels, we found their significant weakness are breakfast and the desk.

The weakness of our approach is that we could not make connection between some adjectives and nouns, which are the subjects of our analysis. For this reason, we could exactly know whether the subjects earn positive review or negative review. Instead, we judge the customers’ contentment toward the subjects only by the rated stars. There is a possibility that customers had good bad part at the same time in the same review, but we could not handle that.

**Contributions:**

HK

- Data cleaning process

- Mainly worked on the analysis part

- Wrote key finding part, recommendation for business plan part and conclusion part in executive summary

- Worked on Shiny App (Tips and part of customers’ evaluation)

- Wrote presentation slides 13-27 pages and record the voice for that parts.

XW

- I had done data cleaning process upon attribution of hotels in Madison, but sadly we could not find any useful information from it.

- Wrote introduction and data cleaning part of summary

- Main editor of Shiny App.

- Also wrote presentation slides 1-12.

RZ