Business Analysis on Cocktail Bars

Xiaowei Zhu, Yijin Guan, Tongyue Jia

December 9, 2021

1 Introduction

We decided to study the factors affecting the cocktail bars business and filtered the relevant data from Yelp. Our analysis focuses on those businesses that show 'cocktail bars' in the information of category. We mainly use the data of review text, attribute, and star to summarize some business insights and provide the owners of bars with some ways to improve the rating of their bars and attract more customers.

2 Data Pre-Processing

2.1 Data Cleaning

We first pick out data related to **Cocktail Bars**. We are interested in cocktail bars, and we select them out through the business file whose categories contain cocktail bars. After data cleaning, the amount of data has reduced significantly, details are listed as follows,

File Name	**Number of Rows**	**Cleaned Number of Rows**
Business	160, 585	1, 338
Review	8, 635, 403	328, 552
Tip	1, 162, 119	36, 535
User	2, 189, 457	220, 269

2.2 Text Cleaning

We mainly use review to give specific suggestions to the business. And for the text data in the review file, we take the following steps to clean them: remove punctuation, convert to lower case, tokenization, remove non-English words, remove stop words and Lemmatization.

For example, a text data is *'This is the best place specially during summer evenings . They do provide outer dining area , which is really beautiful . Staff is very friendly .food and drinks are good'.* After text cleaning, the text becomes *'best place specially summer provide outer dining area really beautiful staff friendly food good.'*

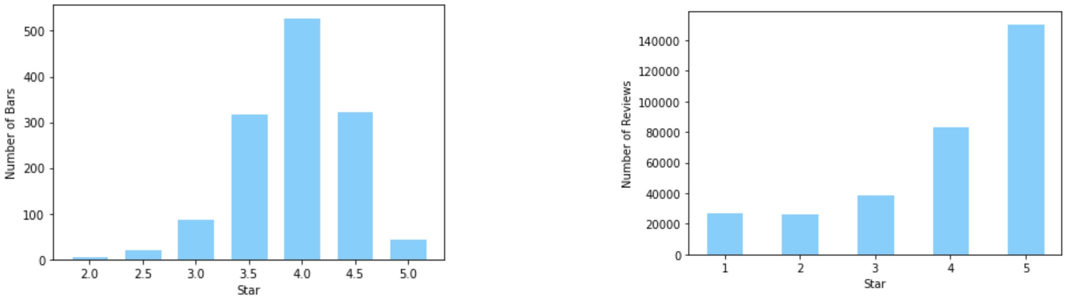
3 Exploratory Data Analysis (EDA)

3.1 Overview of Star Rating

According to Figure 1, we can conclude that the star rating of cocktail bars mostly concentrate around 4.0. Most of the reviews (nearly 140000) are 5 stars , followed by 4 stars of which the number is about 80000.

3.2 High-frequency Words from Reviews

Table 1 is the result of TF-IDF for cocktail bar reviews and we divide the words into four parts: Service, Food, Drink and Place. We chose the words with the highest frequency and filter the ones that represent the important elements or features of a cocktail bar.



(a) The distribution of bars star rating

(b) The distribution of reviews star rating

Figure 1: Star Rating

Service	**Food**	**Drink**	**Place**
Time	Chicken	Beer	Light
Staff	Cheese	Wine	Music
Bartender	Salad	Cocktail	Noisy
Waitress	Steak	Spirits	Temperature
Table	Delicious	Cheap	Live

Table 1: High-Frequency Words for TF-IDF

We selected nearly 20 keywords with high TF-IDF values, and used the following steps to make suggestions for specific bars:

1. Calculate the average rating of reviews containing specific keywords.

2. Calculate the average rating of each business that contains the keyword.

3. Perform a T-test on the two average ratings above. If the test result is significant, it is considered that the business has room for improvement in a certain aspect, and specific suggestions are given.

4 Key Findings About Cocktail Bars

Here are our key findings:

- On average, having a quiet environment increases stars.

- Having an intimate/romantic/hipster ambience increases stars.

4.1 Noise Level and Related Variables

4.1.1 $NoiseLevel$ and $HasTV$

First, we fit an ANOVA model with outcome as ratings and predictors as the non-dictionary-type attribute factors with less than 40% NA value, including $NoiseLevel$. It turns out that $NoiseLevel$ and $HasTV$ are significant predictors, with p-value 2.9e-17 and 4.4e-14 respectively.

Based on this result, we conducted the following test to see whether there is statistical difference in ratings for the different levels of these two factors. To test on the $NoiseLevel$, we suppose the null hypothesis is that mean values of ratings are the same between groups each taking different levels of $NoiseLevel$. Then we apply Tukey's multiple comparison, and we find that at 0.05 level, ratings for more quiet cocktail bars are higher than those with louder noise. For example, ratings for bars with very loud noise level are 0.8377 less than those of bars with quiet noise level (p-value=0.001); ratings for bars with very average noise level are 0.7823 less than those of bars with quiet noise level (p-value=0.001).

In the same way, we test the variable $HasTV$. The result is that ratings for bars with TV are 0.2696 less than those of bars without TV(p-value=0.001). Notice that when considering the upper or lower bounder of the 0.95 confidence interval of mean difference, the qualitative conclusions above still hold. So we can conclude that cocktail bars with lower noise levels and without appliances making too much noise like TV have higher ratings on average.

4.1.2 $Music$

We noticed that the dictionary variable $Music$ is also correlated with noise levels. So we also apply ANOVA to $Music$. The result is that *dj*, *background_music*, *jukebox* are significant variables. In the same way above, we apply the Tukey's multiple comparison. We find that ratings of cocktail bars with background music, jukebox, dj are 0.3759, 0.2905, 0.5631 less than those of cocktail bars without these features separately (at 0.05 level). So we can conclude that noisy music also contributes to lower ratings.

4.2 Ambience

Then we test on the dictionary variable Ambience. Same as the procedure above, we find that casual, intimate, romantic are significant at level of 0.05. After applying multiple comparison, we find that ratings for bars with intimate or romantic or hipster ambience are 0.1864, 0.1455, 0.1009 higher than those without such ambience separately (at 0.05 level). Also, bars with casual ambience have 0.1154 less ratings than those without such marks (p-value = 0.001).

5 Data-Driven Business Plan

Based on the analytics we did above, we provide business suggestions for cocktail bar owners as below.

- Try to provide a quiet environment.
- Avoid facilities making too much noise like TV.
- Avoid noisy music features like dj, jukebox.
- Try to create an intimate or romantic ambience, avoid a casual atmosphere.

5.1 Quiet environment

We apply Tukey's multiple comparison to different levels of variable $NoiseLevel$ to see if there is a difference between mean value of ratings. The result (see Table 2) shows that bars with lower noise levels tend to have higher ratings on average. For example, ratings for bars with very loud noise level are 0.8377 less than those of bars with quiet noise level (p-value=0.001); ratings for bars with very loud noise level are 0.7823 less than those of bars with average noise level (p-value=0.001). So bar owners should try to create a quiet environment to attract more customers.

Group1	Group2	Mean Difference	p-value
Average	Loud	-0.1297	0.0113
Average	Very Loud	-0.7823	0.001
Loud	Very Loud	-0.6527	0.001
Quiet	Very Loud	-0.8377	0.001

Table 2: Multiple Comparison on $NoiseLevel$.

5.2 Avoid noisy facilities

In the same way, We apply Tukey's multiple comparison to the variable *HasTV*. We find that ratings for bars with TV are 0.2696 less than those of bars without TV(p-value=0.001). It may seem strange at first sight, but it is consistent with the advice we give above. Cocktail bar owners should create a quiet environment, so they should avoid facilities making too much noise like TV.

Group1	Group2	Mean Difference	p-value
Without TV	Has TV	-0.2696	0.001

Table 3: Multiple Comparison on *HasTV*.

5.3 Avoid noisy music

Inspired by the conclusions above, we also study the effect of music type to ratings. The result of Tukey's multiple comparisons on variable *Music* show that ratings of cocktail bars with background music, jukebox, dj are 0.3759, 0.2905, 0.5631 less than those of cocktail bars without these features separately (at 0.05 level). So cocktail bar owners should avoid loud music features.

Feature	Group1	Group2	Mean Difference	p-value
Background Music	False	True	-0.3759	0.0044
DJ	False	True	-0.5631	0.001
Jukebox	False	True	-0.2905	0.0234

Table 4: Multiple Comparison on *Music*.

5.4 Create certain ambience

By applying Tukey's multiple comparison, we found that ratings for bars with intimate, romantic or hipster ambience are 0.1864, 0.1455, 0.1009 higher than those without such ambience separately (at 0.05 level). Also, bars with casual ambience have 0.1154 less ratings than those without such marks (p-value = 0.001). So owners should try to create specific ambience like intimate or romantic ambience and avoid casual atmosphere to run a successful business.

Feature	Group1	Group2	Mean Difference	p-value
Intimate	False	True	0.1864	0.0013
Romantic	False	True	0.1455	0.0332
Hipster	False	True	0.1009	0.0489
Casual	False	True	0.1154	0.001

Table 5: Multiple Comparison on *Ambience*.

6 Conclusion

Through the statistical analysis of the review texts and attributes, we found that the noise level and the ambience is the key factor affecting the star rating. In addition, we analysed the key words of each cocktail bar and rated five aspects which are generally believed to be critical to the bar business. These specific business insights are shown on the shiny app.

7 Contributions

- XW wrote the data cleaning and text cleaning part as well as review analysis of both the summary and slides. XW were also responsible for the code related to the three parts.

- YG wrote the introduction, EDA and the conclusion part of the summary. YG also wrote the code of shiny app and contributed to the creation of the slides.

- TJ wrote the code cleaning attributes data and building models, and came up with business advice. TJ also also wrote key findings and business plan part for the report and presentation.