Ice cream data analysis on Yelp

STAT628 Tuesday Group 2

Introduction

Based on the ice cream business and review data on Yelp, we conducted appropriate statistical models and provided useful suggestions to help ice cream business owners to improve their ratings on Yelp. In addition, we built an R shiny App for business owners to give an intuitive overview.

Data Pre-processing

First, we cleaned the data by filtering only the businesses with the category of "ice cream" in the business dataset, and then selecting the reviews only associated with those ice-cream businesses in the review dataset. Then, we tokenized the text of each review into single words and counted their frequency in each review. After eliminating some default stop words in R's package, we got nearly 5 million words in total.

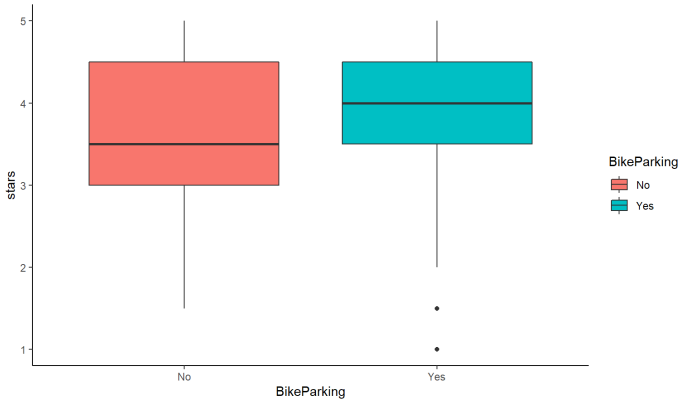
The ratings are discrete integers from 1 to 5. In order to build a more precise model, we use z scores to place them on a common scale. The formula is as follows:

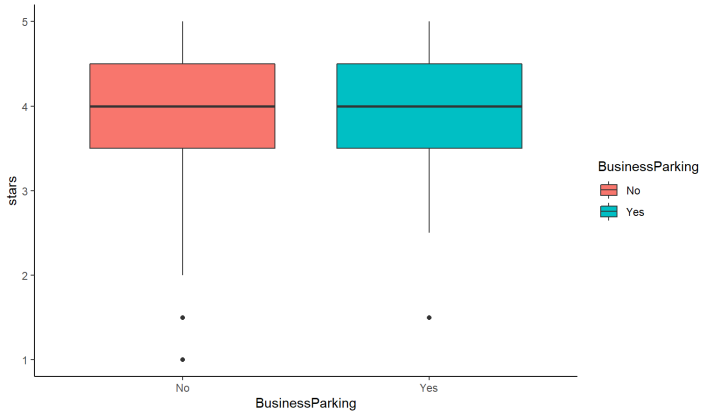
$$Z = \frac{x-\mu}{\sigma}$$, where x is the rating, μ is the mean, σ is the standard error.

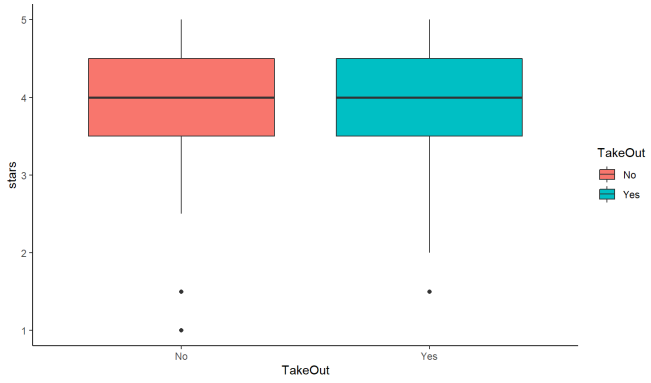
The z score is positive if the value lies above the mean and negative if it lies below the mean. After converting the ratings into continuous values by z scores, we can fit a linear model to evaluate the influence of ice cream types on the ratings.

Exploratory Data Analysis

From the business attributes, we first removed those features that contain large portions of missing values, then we had 7 attributes. In order to test whether certain attributes have effects on the ratings, the sample size of different values in one attribute must be large enough. For example, we would like to consider the influence of wheelchair access on the ratings, but there are only 7 business owners that have wheelchair access. Consequently, it is not reasonable to do an analysis on wheelchair access. Based on this criterion, we then selected 3 attributes, which are bike parking, taking out, and business parking. From the boxplot below, we can have an overview: having a bike parking place will significantly influence the ratings, and the other two attributes will not have a significant effect on the rating.







From the review dataset, we would like to explore the most popular and 'positive' types of ice cream, which can significantly increase the ratings of business owners. We chose 5 types of ice cream, which are cone, shake, sundae, waffle and slush. Then we tidy the dataset and made those types to be dummy variables, meaning if the type existed in this review, then it is 1.
For further evaluation, we need to conduct a formal statistical analysis of the business and review data.

Key Findings

From the statistical model below, we find that bike parking and waffle can help improve the ratings for business owners significantly.

Statistical Model

We conducted two separate statistical analyses on business attributes and types of ice cream. For the business attributes, we built an ANOVA analysis on 3 attributes, whether the business owner provides a bike parking lot, take-out service, and a business parking lot.

Attribute	P value
Bike Parking	5e-9
Takeout	0.1
Business Parking	0.01

From the p-value in the ANOVA table, we can see that bike parking plays an important role in the ratings and take-out service does not impact the ratings under the confidence degree of 0.05. Combining with the boxplot in the exploratory data analysis, we conclude that having a bike parking service will significantly increase the rating of a business owner and having a take-out service will not have an influence on the rating.

For the types of ice cream, we performed a linear regression model. Each sample is a cleaned review, containing types of ice cream in the review and the rating. The explanatory variables are 5 types of ice cream, cone, shake, sundae, waffle and slush. And the response variable is the z scores of the ratings of business owners. The model is as follows:

$$Rating(z\,score) = 0.05 * cone - 0.11 * shake + 0.05 * sundae + 0.31 * waffle - 0.28 * slush - 0.13$$

We can plug in the mean and standard error and scale the coefficients back to the model with a range of ratings from 1 to 5:

$$Rating(1-5) = 0.06 * cone - 0.14 * shake + 0.07 * sundae + 0.37 * waffle - 0.37 * slush + 3.87$$

Since the rating is scaled, the range of it is from -5 to 5. The corresponding summary table of the model is as follows:

Type	Coefficient	P value
Cone	0.05	4e-8
Shake	-0.11	0.02
Sundae	0.05	7e-6
Waffle	0.31	0.04
Slush	-0.28	<2e-16

Under the confidence degree of 0.05, we conclude that all 5 types of ice cream have significant effects on the ratings. And based on the coefficients of the factors, the waffle has the most positive influence on the ratings, the slush has the most negative influence on the ratings.

Recommendations

From the model results, we can conclude that bikeparking shows a significant influence on the ratings of the ice cream business. So It is important to pay attention to the surroundings. For

example, when they plan to open a chain store, It will be a good strategy if they choose an address near the bike parking area.
Besides, increasing the business parking will have only a slight influence on the ratings, and providing takeout service will not help the ice cream business improve its rating through the model. It is not necessary to invest funds in these two aspects.
The results are consistent with common sense, on the other hand, we can explain them with customer psychology and economic principles. Ice cream is more popular in summer than in winter. Taking out the ice cream in the hot weather is hard to prevent it from melting. And it seems no one will drive a long distance just for having ice cream, especially on a hot day.
From the linear model about the types, it might be a good choice for the business to roll out more different kinds of waffle ice cream or make it become a specialty. However, we also need more detailed information to explore the bad performance of slush.

Summary
To help ice cream business owners to improve their ratings on Yelp, we conducted statistical models and provided useful suggestions. Firstly, we filtered the business dataset with the keywords ‘ice cream’ in the category field and then filtered their reviews. After scaling the ratings with mean and standard error, we conducted two separate statistical analyses about the ratings, within which one is the ANOVA table for business attributes and the other is a linear model for the types. The results show that bike parking and waffle can help improve the ratings significantly and we recommend that the business should invest more funds in these aspects.

Contribution:

Jingyun Jia helped to discuss data preprocessing, built the two statistical models, and wrote the corresponding parts in the summary report.

Jiahao (James) Wan helped to clean and process the data used for further analysis and modeling, and then edited the data pre-processing part of the summary report and presentation slides. JW also developed, maintained and deployed the shiny app codes.

Ziao Zhang helped to discuss data preprocessing and statistical models and wrote the recommendation and summary parts in the summary report.