Group 6 Executive Summary

STAT 628

Yujie Zhao, Jiacheng Yu, Jianwei Ren

1. Introduction

In this project, we analyzed Yelp's dataset in an attempt to provide business owners with useful business insights. We chose to use the business.json dataset and our focus was on the Asian Restaurants. We analyzed the reviews counts and ratings for each restaurant in each city and visualized all restaurants on the map. The goal of our project is to make recommendations on what types of new restaurants business owners can open in each city. For example, if a person A wants to open a new Asian restaurant, we will invite him to make a city selection on our app, and then we will give him a recommendation like "If you want to open a restaurant in Madison, we would recommend you open a Japanese restaurant because…".

2. Data Pre-processing

For the data pre-processing part, we first made a filter on the category column of business.json. We selected all possible Asian restaurant categories, including Chinese, Korean, Thai.etc. We then narrowed down the categories based on the number of restaurants in each category. We tried to avoid unbalanced data, so we chose not to consider those categories with a small number of restaurants (e.g., India). We extracted around 3000 restaurants from the raw data. Finally, we grouped our data by each city and each category.

Although there are many cities, the total number of restaurant review counts in many cities is very small, which is an outlier, may be due to the incomplete collection of restaurant data. For example, the sum of Smithton's review counts is 5. Some cities with many review counts do not belong to the United States and should be outliers, possibly due to errors in statistical data. For example, Edmonton has a total of 12070 reviews, which is a Canadian city.

We believe that the review counts can reflect the popularity of the restaurant in the local area, and the stars variable can reflect the rating of the restaurant. As the owner of a restaurant, when considering the location of the restaurant, the first consideration is that opening a certain type of restaurant in a certain city will be more popular with customers, and the rating status of the restaurant. Thus, the variables we choose to study are review counts and stars to study specific types of restaurant recommendations in specific cities. As the selection of cities, for the validity of the conclusion, we selected the top 20 cities with the total number of reviews counts and incorporated some cases where the city names were not the same due to different writing habits. Example, "Saint Louis", "East Saint Louis", "St Louis", "St Louis County" and "St. Louis". As the selection of restaurant types, to achieve the same number of different levels under one factor, we finally chose Chinese food, Japanese food and Southeast Asian food (merging Thai, Singapore, etc.), three types of restaurant types.

3. Exploratory Data Analysis (EDA)

Turn the data into a tabular form suitable for ANOVA, one csv for one feature of each city, a total of 40 csv files (20 cities).

The Analysis of Variance: Firstly, we can use the plot of residuals versus fitted values. Then, we can use Bartlett Test. If the p value is less than 0.05, the null hypothesis, the assumption that the variances of each group are equal, is rejected and it means that does not satisfy the homogeneity of variances, so Welch's ANOVA is used later. If the p value is greater than 0.05, it means that the null hypothesis, the assumption that the variances of each group are equal, is accepted and the homogeneity of variances is satisfied, so the student one way ANOVA is used subsequently.

ANOVA: Single-factor ANOVA model: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases}$

describes two different situations with respect to the treatment effects τ_i

We are interested in testing the equality of the a treatment means; that is,

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_a \ \ vs \ \ H_1\colon \mu_i \neq \mu_j,\ i \neq j.$$

An equivalent way to write the above hypotheses is

$$H_0\colon \tau_1 = \tau_2 = \cdots = \tau_a = 0 \text{ vs } H_1\colon \tau_i \neq 0,\ \exists\ i$$

which means testing that the treatment effects (the τ_i) are zero.

We first perform a formal test of the hypothesis of no differences in treatment means, and use the test statistic

$$F_0 = \frac{SS_{Trt}/(a-1)}{SS_E/(N-a)} = \frac{MS_{Trt}}{MS_E}$$

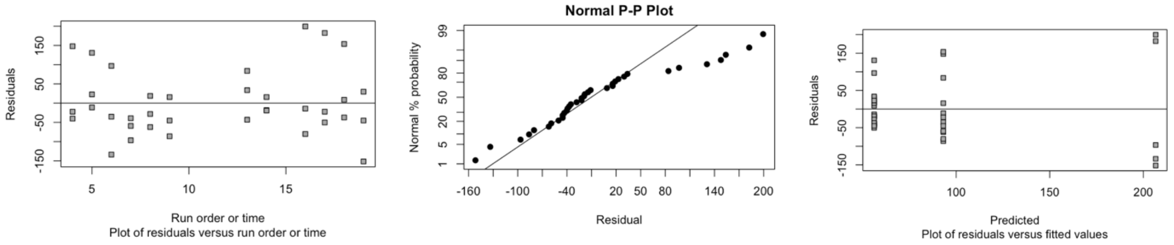
We should reject H_0 on values of the test statistic F_0 that are too large. That is, if

$$F_0 > F_{\alpha,a-1,N-a},$$

When we reject H_0, we conclude that there are differences in the treatment means and need to do the further comparisons among treatment means. Otherwise, when we receive H_0, we conclude there is no significant differences between each means.

Comparisons Among Treatment Means: Since we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons. Hypothesis:$H_0\colon \mu_i = \mu_j$ vs $H_1\colon \mu_i \neq \mu_j$, for some $i \neq j$. Example, $H_{01}\colon \mu_1 = \mu_2$; $H_{02}\colon \mu_1 = \mu_3$; $H_{03}\colon \mu_3 = \mu_2$.We use Tukey Method with assuming the type I error is at most α for any of the possible comparisons. And we make use of the distribution of $q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$, where $\bar{y}_{\max}$ and $\bar{y}_{\min}$ are the largest and smallest means out of a group of a sample means. If $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > T_\alpha$, reject the null hypothesis, where $T_\alpha = q_\alpha(a,f)\sqrt{\frac{MS_E}{n}}$ and there are significant differences between these two comparisons means.

Model Adequacy Checking: After building the model, we do the normal probability plot, the plot of these residuals versus time and the plots of the residuals versus the fitted values. Example of one of the cities, Sparks. Review counts in Sparks first use Bartlett Test. We reject the null hypothesis and means dissatisfy the homogeneity of variances, so Welch's ANOVA is used. Welch's ANOVA tell us that there was no significant difference in the number of reviews among the three category restaurants, so no further comparison needs to be done. Stars in Sparks use Bartlett Test. We accept the null hypothesis, and the variance homogeneity is satisfied, so one way student ANOVA is used. ANOVA tell us that rejects the null hypothesis that the ratings of the three types of restaurants in the city are equal, indicating that the ratings of the three types of restaurants in the city are inconsistent, so further detailed pairwise comparisons are needed. Turkey tests tell us that Chinese restaurants had significantly lower ratings than Japanese restaurants. Then, we use the model adequacy checking. The error distribution is approximately normal. The tendency of the normal probability plot to bend down slightly on the left side and upward slightly on the right side implies that the tails of the error distribution are somewhat thinner than would be anticipated in a normal distribution; that is, the largest residuals are not quite as large (in absolute value) as expected. The plot of these residuals versus time shows that there is no reason to suspect any violation of the independence or constant variance assumption. The plot of the residuals versus the fitted values shows that no unusual structure is apparent.

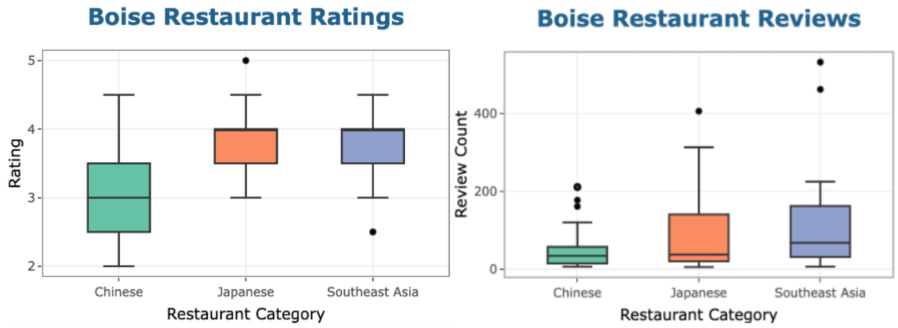


Figures 1. Model Adequacy Checking for Sparks

4. Recommendations for Businesses

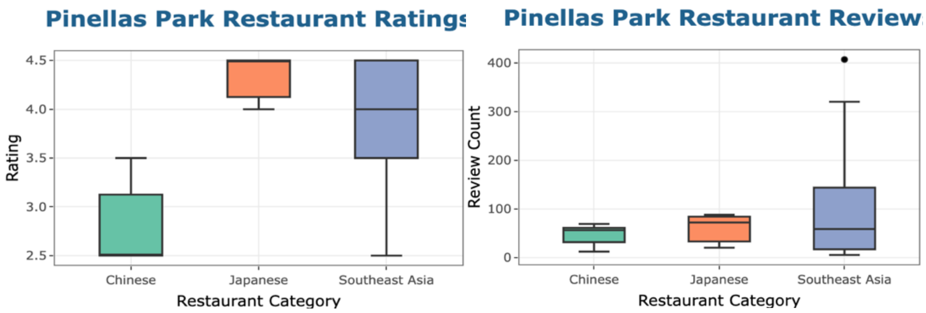
We mainly get statistics from reviews count and star boxplots as well as some tests to provide recommendations for each city.

For Boise, average reviews of Chinese restaurants are significantly lower than those of Southeast Asian restaurants. Chinese restaurants had significantly lower stars than the other two categories. The ratings and popularity of Japanese restaurants and Southeast Asian restaurants are not much different. We provide detailed information/suggestions for this city: Boise has a higher preference for Southeast Asian restaurants followed by Japan and China, which is more favourable for those looking to invest in Southeast Asian restaurants. Who wants to invest in Chinese food in Boise need to be cautious, because the consumer market for Chinese food in Boise is not large, and the ratings are low. The taste of Chinese food may not suit the local taste of Boise. For owners who want to invest in Japanese or Southeast Asian restaurants, although both are more welcome, the original restaurants have high ratings. Therefore, owners who want to open new Japanese restaurants or Southeast Asian restaurants in Boise need to pay attention to the management of restaurants such as taste and price, operating details to improve the score.



Figures 2. Boxplots of Review count and Ratings for Boise

For Pinellas Park, there was no significant difference in the number of reviews. Chinese restaurants scored significantly lower than the other two categories. Southeast Asian restaurants have larger variances in the number of reviews and ratings. Pinellas Park has same preference for all three types of restaurants. The rating of Chinese restaurants is relatively low, which means that if investors could seriously manage Chinese restaurants, the investment potential is relatively greater.



Figures 3. Boxplots of Review count and Ratings for Pinellas Park

5. Conclusion

In this project, we studied Asian food preferences in 20 American cities through data on yelp. The results show that each city has its own unique Asian food preferences. From the preferences of each city, we can offer suggestions to business owners who want to open a store. Here we used one-way ANOVA tests a lot to match the results of 20 cities. But due to the absence of yelp data, there are many cities with only a handful of Asian restaurants, which is obviously abnormal. Therefore, the results of this study, strictly speaking, cannot completely represent the dietary tendency of the whole city. At the same time, we believe that a relatively high proportion of Asian people are not inclined to make reviews on yelp, which leads to the fact that yelp reviews and ratings cannot completely represent dietary bias. Meanwhile, Asian restaurants are more popular among Asians. This makes our recommendations more limited to yelp users. At the same time, we considered that eating habits may vary from region to region within each city. However, since our data cannot provide enough data on Asian restaurants, our conclusions may be more fragile if we divide them into regions in each city.

Contribution:

1. JR worked on the recommendation and conclusion parts for both the report and the slide and reviewed and edited the whole report.
2. JY worked on the R code for EDA and Statistical analysis and wrote the EDA and Statistical analysis parts for both the report and the slide.
3. YZ wrote the introduction and data cleaning parts for both the report and the slide, and was responsible for data cleaning and Shiny App.
4. Overall, we met 4 times and spent around 10 hours on data cleaning & preprocessing, 15 hours on the Shiny App, and 20 hours on the statistical analyses and recommendation.