Summary

What factors contribute to the success of Italian restaurants in Philadelphia?

Group 6 : Baiheng Chen, Ruizhen Jing, Ziang Zeng

Introduction and Background Information

Our group, using Yelp related datasets, try to uncover the determinants of success for Italian restaurants in Philadelphia.

We initially narrowed our focus to the categories with the highest frequency in the business dataset: Restaurants, Food, Shopping, Home Services, and Beauty & Spas. We decided to concentrate our efforts on the 'Restaurants' and 'Food' categories. Within this domain, we observed that Nightlife, American (Traditional), American (New), Italian, and Specialty Food categories each occur over 4,000 times, offering a robust sample size for downstream analysis. We decided to exclude Traditional and New American cuisines due to their ambiguous definitions, leading us to concentrate on Italian cuisine. After focusing on Italian restaurants, we discovered that Philadelphia has 505 Italian restaurants, the highest in the dataset, followed by Tampa with 212. That means Philadelphia can provide a rich and substantial dataset for our investigation.Hence, our study aims to answer the question: "What factors contribute to the success of Italian restaurants in Philadelphia?"

Data Selection and Preprocessing

Our analysis of the Yelp business data extracted all entries tagged with "restaurants" and "Italian" within the Philadelphia area. We get the corresponding reviews in the review data using their unique business IDs. The final dataset consisted of 505 Italian restaurants with a total 73369 reviews.

Exploratory Data Analysis

We conducted exploratory data analysis by visualizing the geographical location information of Italian restaurants and differentiating the rating distributions of high price and low price restaurants. We found that Italian restaurants in Philadelphia exhibit a concentrated distribution in the city center, forming a V-shaped distribution along two streets. This suggests that we can potentially compare the city center with non-city center areas. Additionally, high price restaurants also show distinct average ratings, indicating the need for comparison and modeling based on different price ranges.

Based on the EDA, we categorized the restaurants into four groups: downtown versus others, high price versus low price. Form the basis for our presumptive business expectations for each restaurant. In natural language processing (NLP) analysis, we tailor our recommendations to address the distinct scenarios presented by these categories.

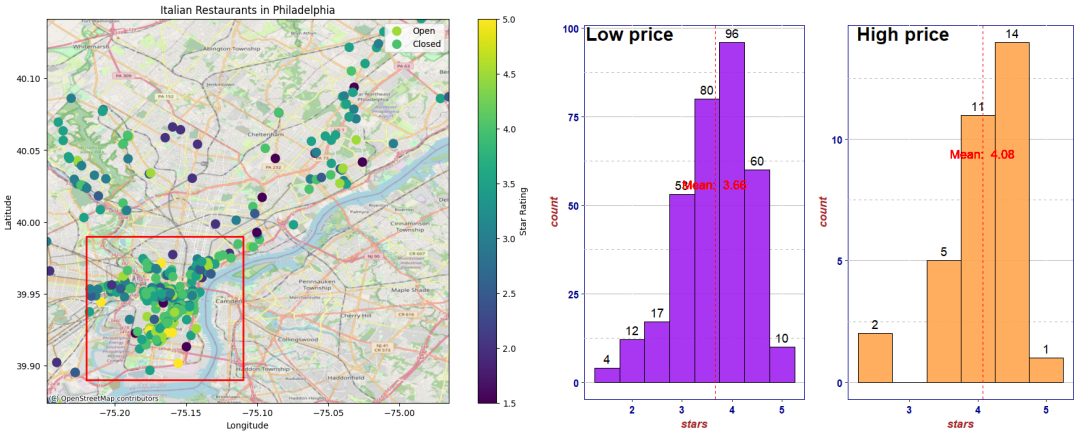

Figure 1 Location and stars of Italian restaurants in Philadelphia

Attribute Analysis

As for the attributes of the restaurants in Philadelphia, we choose the most complete 8 features, WiFi(free, no, paid), RestaurantsAttire(casual, dressy), Smoking(no, yes, outdoor) etc. We visualized their correlation by barplot and made Kruskal-Wallis test to verify if they have statistically significant differences.

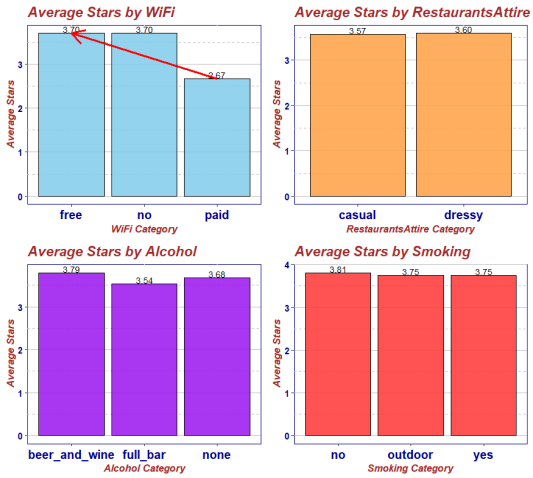

Figure 2 Bar plots of attributes

Var1	PVal1	Var2	PVal2
WiFi	0.059	**NoiseLevel**	0.041
RestaurantsAttire	0.952	**OutdoorSeating**	0.001
Alcohol	0.033	**RestaurantsDelivery**	0.937
Smoking	0.934	**GoodForKids**	0.103

Table 1 Nonparametric Test and p-value

With p-values from nonparametric tests in Table 1, we found that the star ratings are significantly different among different types of Noise Level, Outdoor Seating and Alcohol type which further shows that a successful Italian restaurant should care about these important features. Also, Paid WiFi can damage star ratings in Figure 2. For instance, a good restaurant shouldn't charge for WiFi, supply beer and wine, set outdoor seating and avoid being too noisy.

Moreover, when we take the price range into consideration the test results remain the same except WiFi and smoking. As for WiFi, high priced restaurants have no paid WiFi which negatively influences the star rating which is why the test for WiFi is not significant. As for smoking, only low priced restaurants allow customers to smoke inside and for high priced restaurants the difference becomes significant between the ones who allow outdoor smoking and no smoking is allowed. This may be because customers who go to high-price restaurants expect a better environment, and allowing outdoor smoking may lead to an environment that does not meet their expectations.

We also care about other features like gun crime, income and population of the area of the restaurant, however they are weakly correlated with the star rating.

Impact of COVID-19

We conducted an analysis of the annual trends of opening and closing Italian restaurants in Philadelphia. Under normal circumstances, the number of Italian restaurants in operation and the number of closures increase steadily every year, indicating a stable market for this cuisine. However, this growth pattern was disrupted between 2020 and 2022, especially in 2021, when 110 Italian restaurants shut down. This situation partly explains the negative impact of COVID-19(during 2019-2023) on the restaurant industry just as we expected.

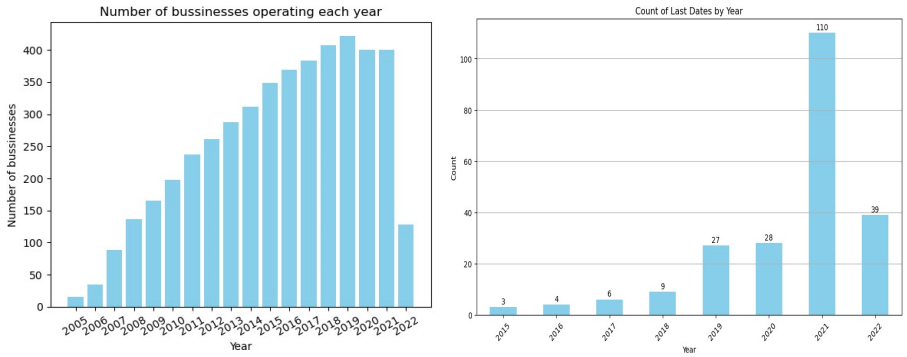

Figure 3 Operating and closed restaurants information

Natural Language Processing

For the review data, we first performed word segmentation, deleted the stop words and counted the number of occurrences of each word. We selected several words with an obvious emotional tendency to show their distribution in comments. Figure 4 shows three examples (terrible, average, delicious) :

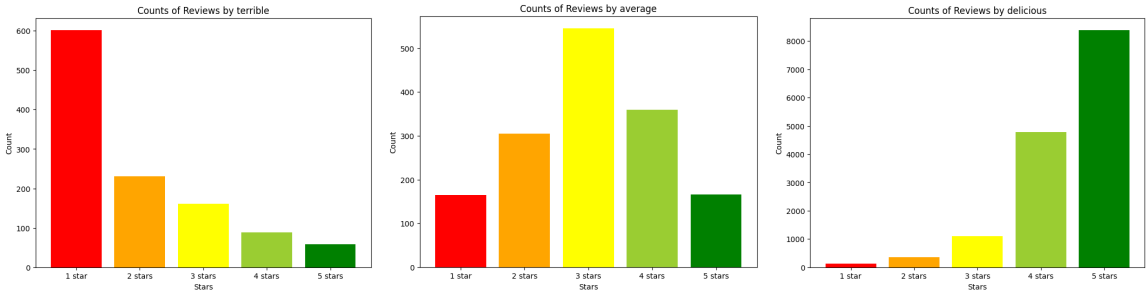

Figure 4 Word counts in review of "terrible", "average", "delicious"

It can be seen from the three bar charts that the number of corresponding words appearing in reviews is consistent with the emotional tendency, indicating that the quality of comment data can be guaranteed.

Comparison of four types of Italian restaurants based on various criteria

We divided the evaluation of the restaurant in the review into five specific indicators, Service, Location, Environment, Hygiene & Safety and Price &Value:

Service	attitude, service, staff, waiter, manager, host, friendly, waitress, attentive
Location	convenient, location, nearby, parking, central, accessible, remote, takeout
Environment	atmosphere, décor, clean, environment, ambiance, noisy, spacious, quiet, cozy
Hygiene & Safety	safe, sanitary, untidy, secure, hygienic, sterile, orderly, regulated
Price & Value	budget, worth, overpriced, expensive, cheap, price, deal, affordable, value

Table 2 Word Bags of Five Indicators

The radar chart shows that location and Hygiene & Safety are less important for most reviews. High-priced restaurants need to have good service and environment, while low-priced restaurants need to have good value for money. Low priced downtown restaurants also need to have a nice ambiance. Besides, Price & Value is more relevant for high-priced downtown restaurants than others. We evaluated each category of restaurant based on the five indicators and compared the high and low scores. This helped us identify the key factors for success for each category. Service is the most influential factor for customer satisfaction, while Environment

is the factor with the largest variation in ratings. Customers tend to be dissatisfied with the price of expensive Italian restaurants, regardless of their location in the city center or not.

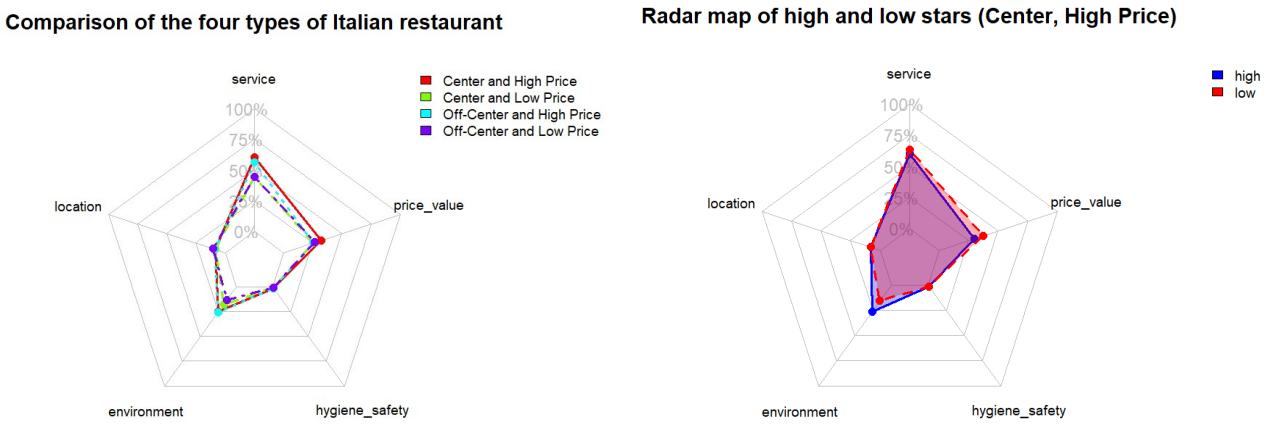

Figure 5 Radar chart based on five indicators

Food analysis

Food quality is a vital factor for a restaurant's success. We identified the most frequently mentioned dishes in the reviews, such as Pizza, Cheese, Dessert, Chicken, Salad, Pasta, Wine, Soup, Seafood and Beef. We analyzed how these dishes influenced the ratings of the four types of restaurants and plotted a line chart. The x-axis of the chart shows the dish name, the y-axis shows the rating, the position of the dot for each dish shows the average rating of the reviews that mention the dish, and the size of the dot shows the rating variance of the reviews that mention the dish. Generally, we assume that a dish with a higher average rating helps a restaurant improve its overall rating, and a dish with a lower rating variance reduces the risk of a restaurant disappointing its customers.

Conclusions

A successful Italian restaurant should offer free WiFi, serve beer and wine, have outdoor seating, and maintain a low noise level. Service and environment are the two main factors that influence customer ratings, and enhancing these areas can improve the image of restaurants. Regarding food, customers of different Italian restaurants do not share a preference for certain dishes, but salad, chicken, and especially soup are more prone to elicit negative reactions.

Contributions	Baiheng Chen	Ruizhen Jing	Ziang Zeng
Presentation 1	Slides 6,7,8 (geography and table)	Slides 2,3,4,5 (Data selection)	Slides 9, 10
Presentation 2	Slides 2, 6,7	Slides 5, 8, 9, 10	Slides 3,4
Summary	Responsible for introduction, data Selection, Categorize Restaurant and NLP	Responsible for COVID-19 Impact, Comparison of four types of Italian restaurants based on various criteria and Food analysis parts.	Responsible for restaurant attribute analysis and hypothesis testing.
Code	Data cleaning.ipynb, final_words_count.ipynb,map_PA_score.ipynb	Number of closed by year.ipynb,Operating_count.ipynb,Index_count.ipynb,Radar.R,Radar(4 region).R	attr.R, app.R
Shiny app	Introduction	Page Suggestion and plots	Interactive Map and Attribute Analysis