Visualization and Analysis on NYC Restaurant Yelp Ratings by Price Level, Cuisine Category, and Location

Hetian Bai (hb1500) Jieyu Wang (jw4937) Zhiming Guo (zg758)

Code Source: https://github.com/wangjy38/1007_Yelp_Project

***Abstract*—In this project, we use NYC restaurants data crawled from Yelp with its Fusion API to build an interactive interface on Python Jupyter Notebook, which filters restaurants based on its cuisine category, price and whether they are open at the time of searching or not. This interface also generates a pie graph that shows the rating composition of these restaurants. Besides, we visualize the data in an interactive way to reveal the relationships between rating, price level and locations of restaurants. By analyzing these plots, we find out that for French, Italian, and Japanese restaurants, higher-price ones tend to be higher rated. Conversely, for Chinese and Mexican restaurants, high priced restaurants tend to be poorly rated. In addition, for Indian and Korean restaurants, price factor seems to barely have an effect on the rating.**

I. Introduction

New York City is a world food bank. According to National Restaurant's survey conducted in 2015[1], there are 45,681 dining and drinking locations in New York City, so there is a huge amount of food options available. For example, we may choose the food spots nearby, or make our decision based on empirical data like ratings, price, and location. In addition, we may consider by food cuisines, like Chinese, Japanese, French, Italian, etc. For this reason, we are interested to know the role of food category playing in rating, price level, and location.

Problem Formulation:

1. Visualization:

a. Pin restaurant locations on Google map and briefly show restaurant profile in Jupyter Notebook using Gmap library[2].

b. Create an interactive interface in Jupyter Notebook which allows users to select restaurants by cuisines, business status (opening/closing), and price levels.

c. Visualize the relationships between NYC restaurants price levels, ratings, and food cuisines with various plot styles using Pygal[3].

2. Analysis:

Give interpretation of each graph to unfold the relationship between price levels, ratings, and food categories.

II. Methodology

1. Using Yelps Fusion API to fetch data of restaurants by certain categories and store it in a dataframe.

2. With packages including Pandas, Matplotlib and Numpy, we extract their locations and use Ipywidgets and Gmap to plot their locations on a Google map. Besides, we use Pygal to plot an interactive pie graph of these restaurants rating percentages.

3. Pygal library is utilized to visualize the relationships between NYC restaurants price levels, ratings, and food cuisines.

The data used to plot are summarized by Pandas groupby function.

III. Results and Discussion

Upper graph in Figure 1 plots Japanese restaurants whose price is $$$$ and is open at the time of searching. By clicking a restaurant's location, an information box that includes the restaurant's name, rating, price, a link to its Yelp's page, as well as a picture of its food.

With the same filtering criterion as the Figure 1's upper graph, Figure 1's lower pie graph shows the percentage of these restaurant's rating.

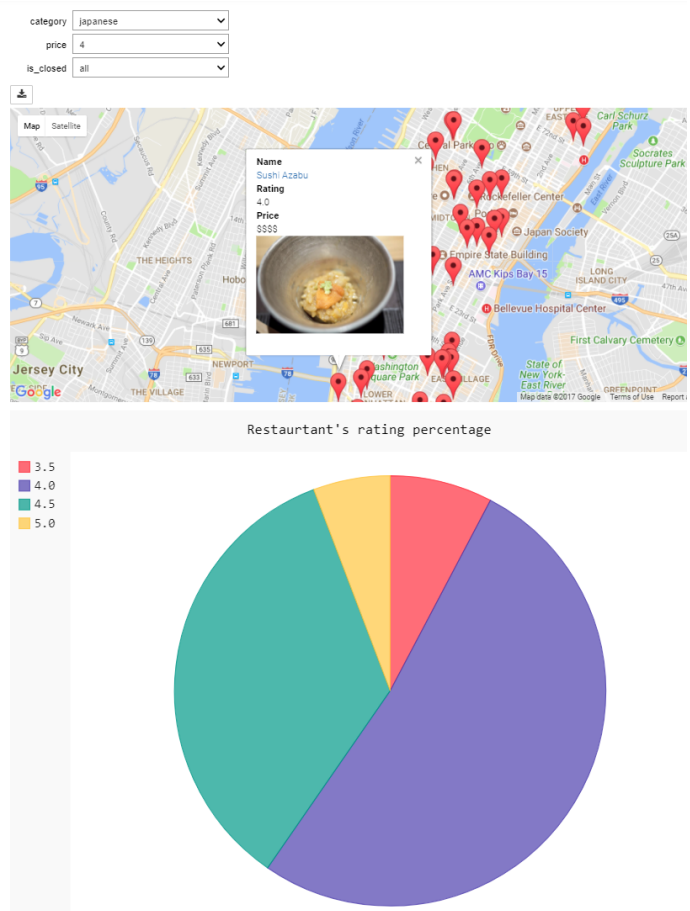

Fig. 1: Interactively Visualized Map and Restaurant's Rating Percentage Pie Chart

Figure 2 indicates that on average, Chinese restaurants have the lowest price and French restaurants have the highest. Indian restaurants are relatively poorly rated and Italian restaurants are highly-rated. Also, there is no obvious relationship between price and rating. Furthermore, Chinese and Mexican food are recommended if people want to have affordable food with a decent rating.

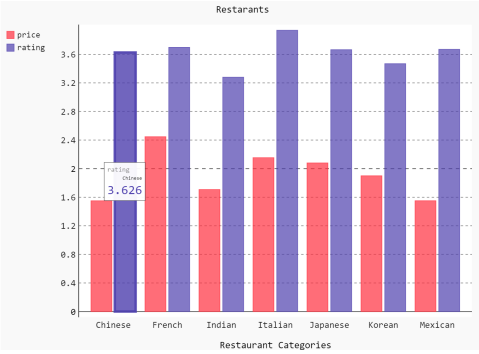

Fig. 2: Price and Rating by Restaurant Categories

Figure 3 reveals the relationship between food categories and restaurants ratings by four price levels. It is obvious that when Korean and Mexican restaurants priced at four dollars ($$$$ defines highest price level by Yelp), the ratings are comparatively low. In other words, customers generally give poor rating on these two cuisines. On the contrary, French and Japanese restaurant with four dollars price, the average of customers ratings are very high among all. Thus, this plot clearly indicates the difference customers attitudes toward cuisines and price levels.

Figure 4 is a radar graph. Color differentiate restaurant ratings. The vertical axis represents price levels. There are 7 directions (sub-axes) equaling to 7 restaurant categories. The general trend is that better rating leads to higher price. The most expensive and least expensive restaurants under all rating are summed in Table I, when a rating is below 2.5, there are not enough data for some restaurant categories. Thus, we don't make any comparisons if the rating is smaller than 2.5. By checking Table I, we find French restaurants

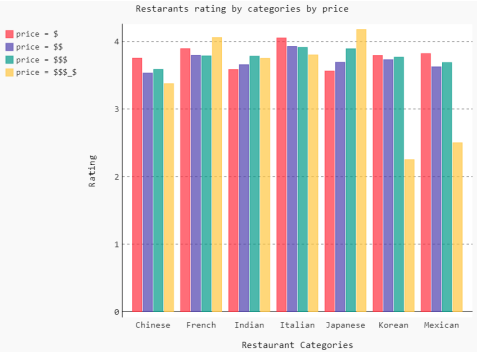

Fig. 3: Rating by Restaurant Categories by Price

TABLE I: The Most Expensive and Least Expensive Restaurant Categories by Rating

Rating	Most Expensive	Least Expensive
2.5	French	Korean
3	French	Mexican
3.5	French	Chinese
4	French	Chinese
4.5	French	Chinese
5	French	Mexican

are the most expensive across ratings, while Chinese and Mexican are the least two expensive categories.

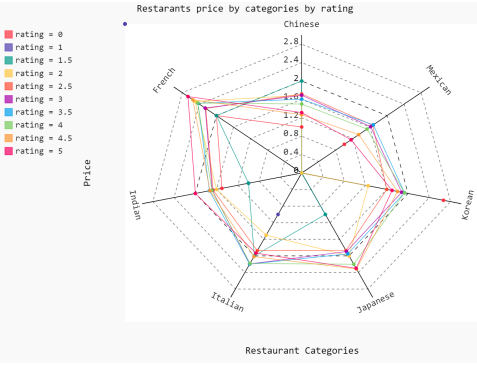

Fig. 4: Radar Plot: Restaurant Price by Categories by Rating

In Figure 5, it is clear to see that, for French, Italian, and Japanese categories, a higher priced restaurant would be higher rated. Conversely, in Chinese and Mexican categories, high priced restaurants are poorly rated. In addition, the ratings of Indian and Korean restaurants are on the medium level and keep constant across price levels.

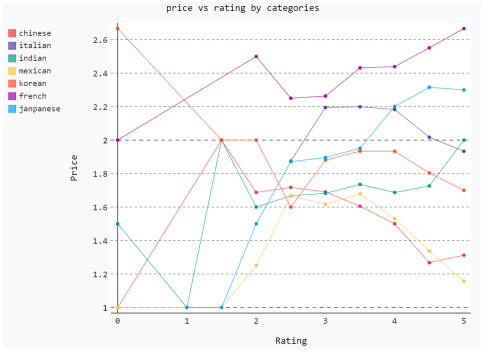

Fig. 5: Rating Level vs. Price Level by Restaurant Categories

IV. Conclusion

In Figure 5, it is clear to see that, for French, Italian, and Japanese categories, a higher priced restaurant would be higher rated. Conversely, in Chinese and Mexican categories, high priced restaurants are poorly rated. In addition, the ratings of Indian and Korean restaurants are on medium level and keep constant on price levels.

V. Limitations

There are two major limitations in our project. First, the Yelp's Fusion's search API can only return the first 1000 search results. Due to a large amount of restaurants in New York City, not all of Chinese, Italian, Mexican, and Japanese restaurants are in our dataset. This loss of data may cause our data to be not representative of all restaurants in these categories. Second, after adding filtering conditions at certain rating and price, there may be only one or two restaurants, and the small amount of data points increase our graphs' variation.

For future improvement, we can develop our own version of Yelp's search API to bypass the maximum number of search results restriction. In this way, the size of our dataset will increase and our analysis can be more accurate.

REFERENCES

[1] New York restaurant industry at a glance. (n.d.). Retrieved December 08, 2017, from http://www.restaurant.org/Downloads/PDFs/State-Statistics/

[2] Pbugnion. Pbugnion/Gmaps. GitHub, 11 Nov. 2017, github.com/pbugnion/gmaps./

[3] K. (2017, July 05). Kozea/pygal. Retrieved December 10, 2017, from https://github.com/Kozea/pygal/