Boston Eateries: A Statistical Analysis

Colleen Kim, Sarah M'Saad, Duy Nguyen, Kelly Zhang
Boston University

Introduction

The goal of this project is to understand the trends of Boston eateries by incorporating health code violations in order to find out the perfect place for enjoying a meal in the city of Boston. Using data from Yelp and Analyze Boston, we answer the following questions:

- Does the price matter when it comes to the quality of restaurants?
- What does the violation rate of a restaurant imply to its quality?
- Do highly dense neighbourhoods influence the average rating of restaurants?
- What are the best restaurants to visit for a set of specific preferences?

Program and Data Sets

Programming Languages and Services:

- Python
- MongoDB (community version)
- Pandas
- R (to create graphs)

Data Sets:

Analyze Boston:

- Food Establishment Inspections
- Boston Neighborhoods

Yelp Fusion API:

- Businesses Information

Algorithms and Analysis Techniques

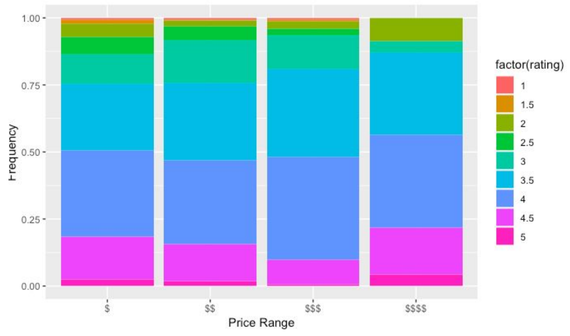
We used z3 for finding the best restaurants to try given a set of preferences. MongoDB held all the datasets and results. We utilized pandas for manipulating data, Plotly.js, D3, and R for visualizing data and Google Maps for providing reverse geocoding coordinates.

Quantifying Quality

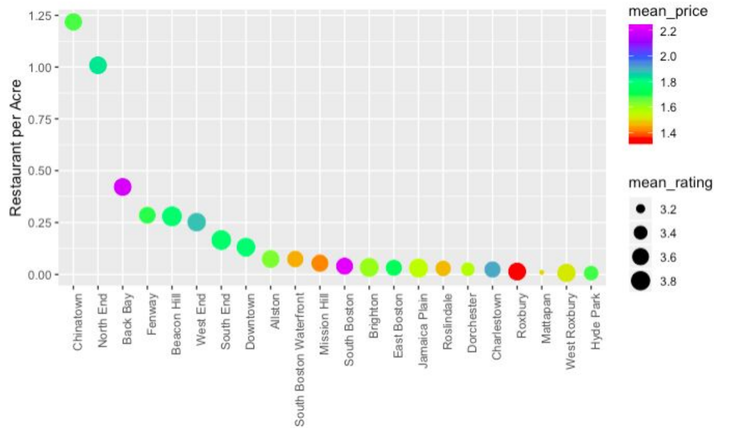
As a proxy to measure the quality of the restaurant, we use the violations recorded in the Food Establishment Inspection dataset to establish a violation rating of each restaurant in the Yelp dataset. The rating is calculated as the number of violation per day in the time period between now and the date the restaurant first inspected.

$$R_{viol} = \frac{n_{viol}}{t_{curDate} - t_{firstInsp}}$$

Analysis

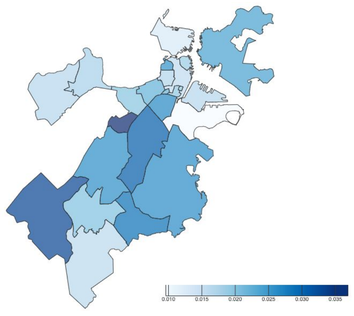


The graph above shows a positive relationship between *price range* of restaurant and the *rating*.

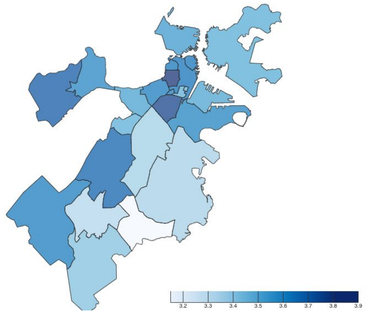


From the Graph above, we can see that Chinatown has the highest *density of restaurants*. While North End has the highest *mean rating* per restaurant. Also, Back Bay is the most expensive neighborhood for foodies.

Violation Rate by *Neighborhood*

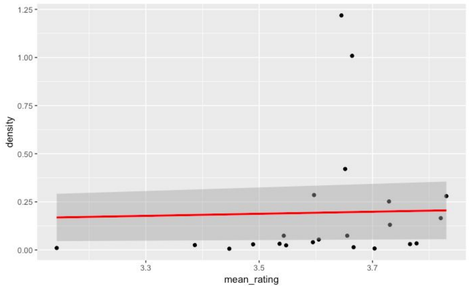


Average Rating by *Neighborhood*



Statistical Correlation

First, we conducted statistical analysis on food establishments in the Boston area. We were interested in finding a correlation between price and rating. We found out that the correlation coefficient was equal to 0.02 which denotes a non-existent correlation. However, we found a correlation between mean_rating and density by neighbourhood of 0.21.



linear model of *mean_rating* and *density*

cuisine	neighbourhood	count	mean_rating	mean_price
italian	North End	42	3.869048	2.095238
hotels	Back Bay	20	3.450000	2.750000
chinese	Chinatown	18	3.666667	1.777778
seafood	North End	18	3.722222	2.277778
coffee	Dorchester	16	3.781250	1.187500
newamerican	Back Bay	16	3.656250	2.062500
sandwiches	Back Bay	16	3.937500	1.437500
sandwiches	Fenway	16	3.531250	1.375000
cocktailbars	Back Bay	15	3.766667	2.400000
coffee	Back Bay	15	3.933333	1.400000
coffee	Fenway	15	3.533333	1.133333
breakfast_brunch	Back Bay	14	3.821429	1.857143
pizza	South Boston Waterfront	14	3.357143	1.214286
sandwiches	Beacon Hill	14	3.785714	1.357143
bars	Back Bay	13	3.653846	2.153846

Neighborhood restaurant count by *cuisine*

From the table above, we can see that North End holds up to its reputation for being the neighborhood with the most dense Italian restaurants. It is also interesting to notice the fact that Back Bay has the most hotels / sandwiches / newamerican / breakfast_brunch restaurants.

Conclusion

In conclusion, we found that price does not matter when choosing to eat a quality meal. Furthermore, we also found that areas with densely located restaurants have a higher rating (positive correlation between *density* and mean_*rating*). Nevertheless, we found that there is a negative correlation between *violation rate* and *rating;* however, it was not significant since the correlation coefficient was around 0.08. Which means that the more violations the restaurant received, the less rated it is. Finally, z3 is the best tool to find the best restaurants given a set of constraints.

Future Works

For future works, here are some things we can improve on:

- Right now, we are only looking at trends in Boston. We can potentially expand the database to observe trends of eateries in other cities and compare the results with those of Boston.
- The data provided by Yelp API contains many missing values so our findings may be inconclusive.
- Analyze Boston updates some of its datasets daily. Our current project has yet to tackle the ability of regenerating daily analysis. This would be an helpful feature as it provides valuable insight over time.
- We also notice a possible correlation between violation rate and mean household income of the neighborhood. However, this need further studies to uncover deeper potential causes.

For more on this project, scan the QR code.

BU Department of Computer Science