Executive summary of findings

Group 11 Section 2 – Ananthapadmanabhan Sivasankaran, Muthuraja Palaniappan, Yash Sharma, Ibukun Phillips

# Process

We can divide the process into 3 categories: web scraping, NLP and modeling.

Web scraping Reviews of around 3,000 restaurants in San Francisco (since we wanted to test our hypothesis on one city first). We scraped from [www.tripadvisor.com](http://www.tripadvisor.com) because it has all the control variables like food rating, service rating, and value rating for individual reviews.

Natural Language Processing Using the sysnet library we identified food related keywords. We found the frequency of these keywords in the reviews to find the 5 most popular dishes. We identified all the reviews that mentions these five dishes and filtered out all the content that are not pertaining to the food (like ambience, location, delivery service, opening hours, etc). We then, assigned a sentiment score to each comment and aggregated the score at the restaurant level to find the dish sentiment for that restaurant. The sentiment analysis were done at the sentence level using NLTK package in Python.

Modeling The regression equation that we used is:

The preliminary model is built using price, food rating, value rating and service rating as those are the factors on which reviewers can rate on the website (except price). We now add the newly created factor of dish sentiment and see its impact on the overall ratings. We found through t-test that dish sentiment is highly correlated with food rating. Hence we adjusted the food rating for this dish sentiment by taking only the component that is orthogonal to dish sentiment.

# Findings

We found that

* Some dishes are significant in explaining the overall rating. Increasing customers for these highly popular dishes would increase the overall rating of the restaurant.
* Burger, taco, pasta, noodle and pizza were identified as the most commonly reviewed dishes. Even among these dishes some of them are significant in explaining the restaurant rating at 5% significance level and others at 10% significance level. This shows that not all the popular dishes have the same impact on restaurant rating.
* Some of the dishes turned out to be insignificant in explaining the increase in restaurant rating whereas food rating was significant. This shows that the food rating of other dishes explains the restaurant rating better than this dish.
* Price range of the restaurant is less significant in explaining the increase in restaurant rating. Therefore, whether the restaurant is luxury or cheap, the expectations of the customers get adjusted accordingly and hence does not impact the ratings much. That being said, price has a positive impact on the rating.
* Overall rating is proved from established studies to be highly impactful on the revenue (from a Harvard study). Hence, we take higher rating as a proxy for higher revenues.
* From our study, we conclude that small / new / niche restaurants will be better off by focusing on only few dishes and reducing the inventory cost. They can increase the quality of food and in turn the reviews. This will lead to higher revenues for them.