**“Dirty but Yummy”**

Core Message

To analyze data from the NYC Department of Health to determine if the cleanliness of restaurants correlates with their ratings.

Hypothesis

Null Hypothesis - If highly rated restaurants are as clean as all other restaurants, then 4 to 5-star restaurants in Yelp will have various inspection scores.

Alternative Hypothesis - If highly rated restaurants are clean restaurants, then 4 to 5-star restaurants in Yelp will have an inspection score of less than 15.

Questions

1. Do restaurants have low inspection scores (an inspection score of 0 means no violation) also have high customer ratings?
2. Are expensive restaurants cleaner than the less expensive ones?
3. Which neighborhood is the dirtiest? Which is the cleanest?
4. Which cuisine type is the dirtiest? Which is the cleanest?

Data Source

NYC Department of Health’s open database

* 385,386 rows of data
* 33 inspection types with a single entry for each violation
* Rolling inspection dates - Use latest inspection date data
* Filter out new restaurants not yet in business
* Final data frame – 22,700 unique CAMIS records

Yelp Fusion

* API Businesses/Phone/Search Endpoint
* Iterate through phone numbers in final NYC data frame to retrieve rating, review count, and price level data
* Convert price level in “$” to readable value
* 13,900 matches with rating, price level, and review count

Google

* Iterate through addresses in final NYC data frame to retrieve latitude and longitude data

Findings

* 1 sample t-test p value is 0.1311, which is higher than the significance level of 0.05. Therefore, we do not reject the null hypothesis.
* Based on the scatterplots, we also concluded that not all expensive restaurants are clean restaurants
* Cuisine type “Hotdogs” are the cleanest, with the lowest average inspection score.
* As shown on the Google heatmaps, dirty and clean restaurants are evenly distributed across the city.

Post Mortem and challenges

Because the NYC dataset is compiled from several large administrative data systems, it contains some illogical values and missing data that could be a result of data entry or transfer errors.

Since free Yelp Fusion API key has a 5000-daily limit, we had to generate multiple API keys with different email addresses to run through more than 22,000 phone numbers to finish the project on time.

If we have more time, it would be interesting to run analysis on the “expensive 5-star” restaurants to see if they performed better during inspection.