# Group C - Yelp Data Process Book

by Chana Messinger, Liz Walters, and Michael Weisner

# Initial Project Concept

- Compare distribution of restaurant metrics (reviews, price, health violation) in New York City
- Develop neighborhood statistics, e.g. most frequent cuisine and demographic profiles
- Conduct sentiment analysis and network analysis of reviews
    - We hoped to be able to use reviewers as edges between restaurants (what restaurants are linked by the same people attending them?) or vice versa (what reviewers are linked by having reviewed the same restaurant)?

# Data

## Yelp Data

First, we explored the Yelp Academic dataset, but as it is on a national scale it did not provide a robust sample of New York City restaurants, so we turned to using the Yelp API to scrape a better sample.

We scraped approximately 4000 restaurants by running API calls to New York City, Manhattan, the Bronx, Queens, Brooklyn, and Staten Island, and then removing duplicated and non-New York restaurants. The API provided up to date address, review data, review count, and price data on open restaurants. It did not, however, provide review data, limiting the full usefulness of the data.

## Yelp Review Data

We found a curated set of curated Yelp review data by the Stonybrook University Outlier Detection Datasets (ODDS) group. While a smaller sample of restaurants compared to our Yelp API data (only 923) it still seemed like a reasonably robust sample.

## Restaurant Health Grades

NYC OpenData provides a daily updated Department of Health and Mental Hygiene (DOHMH) dataset of NYC restaurant inspection results. This data proved the most robust, as it was both longitudinal and sampled almost the entire restaurant population with over 24,000 unique restaurants.

# Demographics

We decided to use the American Community Survey for demographic estimates. Kaggle user MuonNeutrino conveniently hosted ACS 2015 estimates for New York City. This let us quickly get population estimates for use with our restaurant data.

# Joining the Data

A potential problem arose when we first investigated joining our data together. The Yelp API data and DOHMH data provided full addresses, but neither they nor their names were enough to easily merge the datasets. The names in particular were inconsistently coded. Also, problematically, the Stonybrook review data was dated, including many closed restaurants and not using Yelp's listed names.

After a great deal of data cleaning we were able to merge the Yelp data onto about 350 reviewed restaurants of the 923 total. Ideally we would be able to expand this, but time did not allow for it. Still, it did provide us with a diverse sample of restaurants to analyze their reviews in conjunction with their metrics and map them onto the city.

The question became what level of spatial unit was feasible and appropriate. Given the huge amount of datapoints we were worried that geocoding would take too long, and potentially lead to discrepancies between the different datasets, so we opted for ZIP codes, as they were included in all of the datasets already. While this area is not as robust as census tract or block, it seemed adequate for the scope of the project and allowed us to quickly get started, as we could easily join the data to an NYC ZIP code shapefile provided by NYC OpenData.

# Data Exploration

While our original goal was to compare things like health grades to metrics such as poverty and more conceptual constructs like class, the large spatial area of a zip code made this seem like a questionable statistical claim. Instead we focused on comparing general restaurant trends using ZIP codes at the unit of analysis for much of it.

## General Trends

We made mapping the Yelp API metrics and restaurant distributions in comparison to people the first priority. While perhaps unsurprising, it was interesting to understand restaurants per capita throughout the city, as well as the kinds of cuisines represented. We saw a surprisingly large range of cuisines, though a surprisingly small range of grades and average grades. Perhaps because of the "grade pending" system, most restaurants do alright on that metric (also, presumably, if they do not meet standards, they will be shut down). The Yelp data mostly focuses on the relationships between restaurant reviews, price of restaurants, and review ratings. We found some interesting attributes by zipcode, such that the higher prices and higher review counts for restaurants tend to be in Manhattan zips, and lower review ratings tend to more often be in zipcodes of Staten Island, Queens, and the Bronx. The yelp data also shows that there is a higher number of 4 star rated restaurants in all price categories, and that there

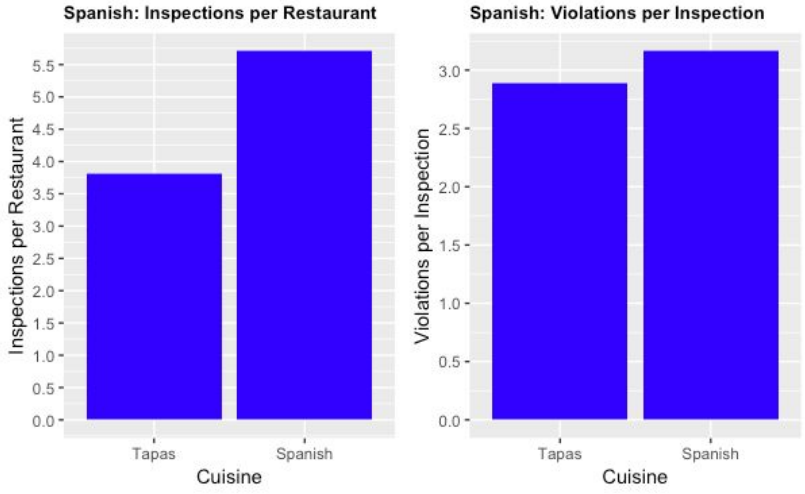
are no expensive restaurants rated below 3 stars. Possibly because expensive restaurants with low reviews would not last too long.

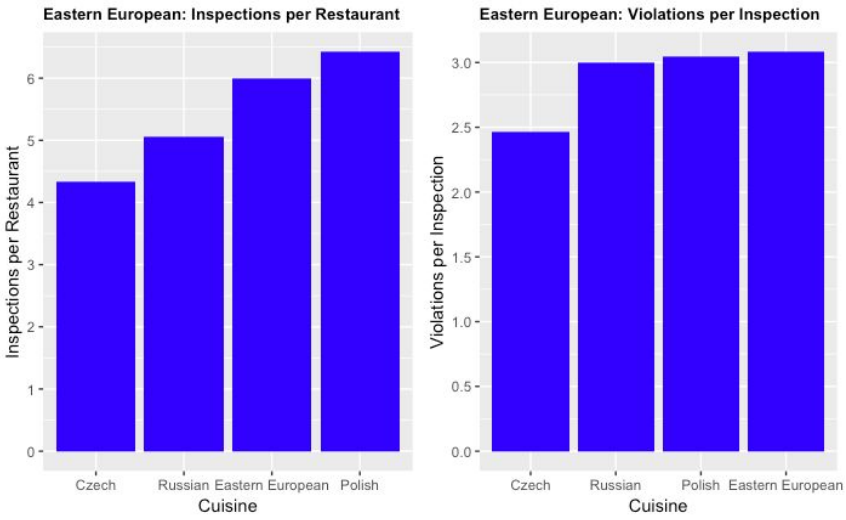
## Approach

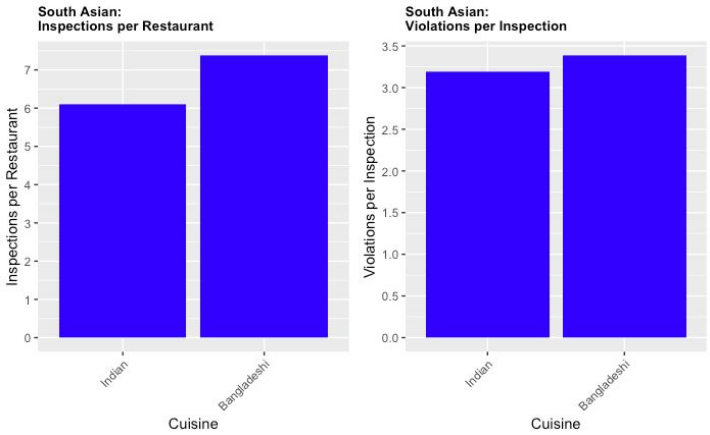
Because we could not join the datasets, we decided to explore them separately. The DOHMH data gave us overall numbers of restaurants, cuisines and health, the review data gave us reviews and sentiment and the API data gave us price and satisfaction.

## Theoretically Similar Restaurants

One idea we looked into exploring but then abandoned (and adapted) was comparing cuisines that one would expect a priori to be similar, to see if for some reason Spanish places and Tapas places had vastly different average grades, or Russian, Czech, Polish and Eastern European restaurants had highly dissimilar inspections or violations per inspection. The results of the first foray into that is below, but the differences did not tell an interesting story, so we scrapped it.






Similarly, when we first looked at chains like Starbucks and McDonalds, we were curious to see if there were vastly different grades, but the vast majority had A grades, so we switched to score per inspection as a metric.

## Text Analysis

One initial comparison we were hoping to make was to compare average ratings to average sentiment scores, but the geographical limitations of the data made this impossible. Still, there were some interesting trends discovered, such as the positive relationship of review length and price level. While not particularly robust or enlightening, it and the most common sentiment words did seem to follow logic of what we would expect to see at different restaurant price levels.