Comparing NYC Restaurants Inspection Results with Yelp Ratings & Reviews

Summary

My proposal is to explore the correlation between the NYC inspection results and the popularity of a restaurant on Yelp. Also understanding if at all there is an improvement in health of a restaurant overtime depending on the reviews and/or ratings on Yelp. Do reviews drive the decisions taken by management at these restaurants? I do not know if it is possible to find this out though.

Description of Data Sets

NYC Open Data (https://data.cityofnewyork.us/) has published inspection results for restaurants in the New York City region provided by Department of Health and Mental Hygiene (DOHMH). Yelp has an API which allows access to ratings and reviews for restaurants across the world.

1. NYC Open Data is updated daily and is available in csv format for restaurants in the NYC region like Manhattan, Brooklyn and Queens. The current data size is about 200MB containing columns like - location, rating, violation code, violation description, critical flag etc.
2. Yelp API allows access to all possible restaurants in the NYC regions containing information like location, reviews, ratings(stars), times open, categories etc. The data size cannot be accounted for all the data format is in JSON. There are about 50 thousand restaurants in NYC.

Both datasets provide enough ground for comparison and the keys can be matched to manipulate data.

Data Manipulation

The downloaded dataset from NYC Open Data needs to be curated by combining multiple entries for the same restaurant as it contains information from numerous inspections. After normalizing this data set, the primary columns/values need to be extracted and stored into a data structure or file which can be compared/combined with data from Yelp. Yelp has about 50076 restaurants listed in NYC. There is a high probability that all the restaurants may not be available in the downloaded dataset so only the ones present will need to be stored in a data structure and/or in a file for further processing. This may also result in faster processing of data. Location, Ratings, Stars, Reviews and Names will be combined from both data sets. Using this as the base, a normalised number can be calculated to be used for the visualisation.

Visualization

An interesting visualisation could be to find the correlation between the number of ratings on Yelp for a restaurant and its health rating as per NYC Open Data. This could be represented in the form of a scatter plot with linear correlation. This could provide insight into how consistent people can be with respect to DOHMH results.

Another possibility could be to combine Yelp and NYC Open Data ratings/stars and plot a heat map based on location of the restaurants.