**Introduction**

Food poisoning can be a truly wretched experience. What if you could easily check for any recent health code violations at the restaurant you’re thinking of visiting? New York City provides a freely available record of health inspection results for restaurants through the Department of Health and Mental Hygiene. Connection between this database and the Foursquare database would allow consumers quick access to the “health code history” of the restaurant they are visiting, and how it compares to alternatives.

This DOHMH data may also be useful for other stakeholders. By comparing health inspection results with the ratings of the restaurants on Foursquare, we could also offer restaurants information on how health inspection performance relates to ratings from Foursquare reviewers, or on which violations are most likely to be associated with restaurant closure. Through spatial visualization of the inspection data, we may also be able to assist the DOHMH with prioritizing areas of the city to focus on.

**Data**

Together with data from the Foursquare API, I will use the DOHMH New York City Restaurant Inspection Results data set (<https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j>) to develop resources for consumers, restaurants, and health officials. These resources will include:

(1) Python code for linking the health inspection dataset with the Foursquare database using the Foursquare API.

(2) Python code allowing consumers to inspect the health inspection record of the selected restaurant.

(3) Python code for comparing the health inspection record of a selected restaurant with nearby of the same cuisine type.

(4) Statistical assessment of how health inspection performance relates to the rating of a restaurant on Foursquare.

(5) Identification of health code violations most likely to lead to restaurant closure.

(6) Spatial summaries of health inspection results designed to assist the DOHMH in setting of priorities.

**Methodology**

I will separately describe the methodology for each of the six goals presented in the **Data** section.

*(1) Python code for linking the health inspection dataset with the Foursquare database using the Foursquare API.*

Before attempting to link the DOHMH database with the Foursquare API, I conducted data cleaning on the DOHMH dataset. I then used the Name, Address, Zipcode, and Telephone of each restaurant in the DOHMH data to make a Foursquare API call, with the intent set to “match”.

*(2) Python code allowing consumers to inspect the health inspection record of the selected restaurant.*

I produced two custom python functions. The first, ‘foursquare\_inspections”, returns a summary of all the inspections for a given foursquare ID. This is accomplished using pandas. First, I select all inspection records for a restaurant. I then group these records (using the “groupby” function) by the date of inspection to get information on individual, unique inspections. I use the the pandas function “apply” on the resulting groups to extract the information about each inspection,

The second, ‘foursquare\_report”, produces a number of summary statistics and a graph showing the health inspection performance of the restaurant over time. Once again, I use pandas “groupby” and “apply” to extract information about individual inspections. Calculation of a few summary statistics follows and finally the pandas “plot” function is used to return a graph of points per inspection over time.

*(3) Python code for comparing the health inspection record of a selected restaurant with nearby of the same cuisine type.*

I produced a python function that, for a given restaurant, will return the closest five restaurants of the same cuisine type, along with their points per inspection. To calculate the five nearest neighbors, I used the KDTree algorithm from scikit-learn. I then used pandas to return the points per inspection for these restaurants.

*(4) Statistical assessment of how health inspection performance relates to the rating of a restaurant on Foursquare.*

I accessed rating information for venues can only be accessed through premium Foursquare API calls (limited to 500/day). I then used scatter plots and Pearsons correlation coefficient to examine the relationship between points per inspection and Foursquare user rating.

*(5) Identification of health code violations most likely to lead to restaurant closure.*

To identify the most important health code violations to ignore, I used two approaches for feature selection. Both relied on logistic regression. To generate the binary dependent variable, I created a new dataframe column “closed”, with 0 indicating inspections that did not result in a closure, and 1 indicating inspections resulting in a closure. I used the pandas function “crosstab” to change the categorical “violation code” column into a set of one-hot encoded variables, indicating whether a given violation code was present (“1”) or absent (“0”) from an individual inspections.

The first method I used for feature selection was Recursive Feature Elimination (or RFE, implemented using scikit-learn). I chose to retain the 10 features with the strongest relationship to restaurant closure.

The second method I used for feature selection was principal components analysis (PCA), implemented using scikit-learn. Running a logistic regression using all ~80 violation codes as predictor variables might give problematic coefficients due to multicolinearity in our predictors. Principal components are linear combinations of features that are uncorrelated with one another. I first generated a set of 10 principal components of the one-hot encoded matrix of violation codes. I then conducted logistic regression using these components. I then identified the three principal components with the strongest coefficients in the regression. Finally, to select features, I selected any features that were correlated to these components with r > 0.2 (pearson correlation coefficient).

*(6) Spatial summaries of health inspection results designed to assist the DOHMH in setting of priorities.*

To visualize inspection results over space, I first generated a dataframe with each row being a unique restaurant, its points per inspection average, the total number of inspections for that restaurant, and that latitude/longitude of the restaurant, once again using pandas “groupby” and “apply”. I then used the matplotlib function “hexbin” to generate a series of 2-d histograms. The first histogram displays the average ppi of all the restaurants in a given hexagonal bin. The second histogram displays the average number of inspections for all the restaurants in a given hexagonal bin. The final histogram displays the average of (ppi / number of inspections) across all restaurants in a given hexagonal bin.

**Results**

I will summarize the results separately for each of the six goals presented in the **Data** section.

*(1) Python code for linking the health inspection dataset with the Foursquare database using the Foursquare API.*

I retained 13,107 restaurants from the DOHMH data and for which a single unique foursquare ID was found. I produced a new dataframe that links foursquare IDs with health inspection results.

*(2) Python code allowing consumers to inspect the health inspection record of the selected restaurant.*

I produced two custom python functions. The first, ‘foursquare\_inspections”, returns a summary of all the inspections for a given foursquare ID. The second, ‘foursquare\_report”, produces a number of summary statistics and a graph showing the health inspection performance of the restaurant over time.

*(3) Python code for comparing the health inspection record of a selected restaurant with nearby of the same cuisine type.*

I produced a python function (cuisine\_compare) that, given a foursquare ID, will return the five closest restaurants of the same cuisine type along with their average points accrued per health inspection.

*(3) Statistical assessment of how health inspection performance relates to the rating of a restaurant on Foursquare.*

I successfully collected Foursquare user rating information for 2,345 restaurants. The observed correlation between user rating and points accrued per health inspection (ppi) was weakly negative (Pearson correlation coefficient= -0.06). Inspection of a scatter plot revealed that restaurants with ppi > 30, while rare, were nearly always rated below a 7. Following a log transformation on points per inspection, another scatter plot revealed that restaurants with among the lowest points per inspection, while rare, are nearly always rated above a 7.

*(5) Identification of health code violations most likely to lead to restaurant closure.*

Using recursive feature elimination, I found the top 10 most important violation codes for predicting restaurant closure in a logistic regression. The codes were as follows:

03E : Potable water supply inadequate. Water or ice not potable or from unapproved source. Cross connection in potable water supply system observed.

04F: Food, food preparation area, food storage area, area used by employees or patrons, contaminated by sewage or liquid waste.

05A: Sewage disposal system improper or unapproved.

05C: Food contact surface improperly constructed or located. Unacceptable material used.

05E: Toilet facility not provided for employees or for patrons when required.

05F: Insufficient or no refrigerated or hot holding equipment to keep potentially hazardous foods at required temperatures.

05H: No facilities available to wash, rinse and sanitize utensils and/or equipment.

08A: Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.

18C: Notice of the Department of Board of Health mutilated, obstructed, or removed.

18D: Failure to comply with an Order of the Board of Health, Commissioner, or Department.

I created a set of 10 principal components from the one-hot encoded violation code matrix. The first component explained 38% of the variance. The first component was the most important predictor in logistic regression (coefficient=1.87945936 ), and two other principal components had coefficients of >1 (PC-9 and PC-10). For each of these principal components, features with a correlation of > 0.2 are presented below:

*PC-1*

08A (r=0.69):Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.

04L (r=0.51):Evidence of mice or live mice present in facility's food and/or non-food areas.

04N (r=0.27):Filth flies or food/refuse/sewage-associated (FRSA) flies present in facility\x1as food and/or non-food areas. Filth flies include house flies, little house flies, blow flies, bottle flies and flesh flies. Food/refuse/sewage-associated flies include fruit flies, drain flies and Phorid flies.

*PC-9*

04M(r=0.8): Live roaches present in facility's food and/or non-food areas

08A (r=0.25:'Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.'

*PC-10*

04H (r=0.99): Raw, cooked or prepared food is adulterated, contaminated, cross-contaminated, or not discarded in accordance with HACCP plan.

*(5) Spatial summaries of health inspection results designed to assist the DOHMH in setting of priorities.*

I produced three 2d histogram plots to help visualize health inspection results across New York City. The first displays the average performance of restaurants within spatial bins. The second displays the average number of inspections performed on restaurants within spatial bins. The final plot shows the average value of (points per inspection / total number inspections) for restaurants within spatial bins.

**Discussion**

I will discuss the results separately for each of the six goals presented in the **Data** section.

*(1) Python code for linking the health inspection dataset with the Foursquare database using the Foursquare API.*

We only recovered about half of the restaurants in the DOHMH dataset. There are several ways we might improve this rate of retention. For example, I could independently geocode the addresses and pass these lat/longs to the foursquare API call. I could also try to disambiguate addresses that seem to have multiple camis IDs. The DOHMH may themselves want to collapse these records.

*(2) Python code allowing consumers to inspect the health inspection record of the selected restaurant.*

The code produced here offers two simple types of reports on restaurants health inspection history. This type of information could be passed to a Foursquare user through Foursquare or a third-party app. While creating this function, I noticed that many health inspections are ungraded. Basically, restaurants that do not receive an A on initial inspection have a short period of time to “clean up their act” before a re-inspection occurs and grade is assigned. See <https://www1.nyc.gov/assets/doh/downloads/pdf/rii/blue-book.pdf> for more details. To account for performance in these ungraded inspections, I created a new measure of health inspection performance, which is simply the average number of points accrued in an inspection (ppi). Although this measure would need to be explained to consumers, it should provide a better overall assessment of health inspection performance than inspection grades do.

*(3) Python code for comparing the health inspection record of a selected restaurant with nearby of the same cuisine type.*

The function produced here offers one simple way to make comparisons based on health inspections, returning the five closest restaurants of the same cuisine type and their average points accrued per health inspection.

*(4) Statistical assessment of how health inspection performance relates to the rating of a restaurant on Foursquare.*

I observed only a weak correlation between health inspection performance and user rating on Foursquare. However, it does appear that either very good or very bad performance on health inspections might predict foursquare user rating. More data is needed, but nearly all restaurants with a points per inspection of >30 had user ratings <7, and nearly all restaurants with a log(points per inspection) of <1.25 had user ratings >7. I could use a polynomial regression to try and capture this effect. It might also be useful to create bins of points per inspection, and then use a method such as Analysis of Variance (ANOVA) to see which bins have significantly different distributions of user rating. Finally, it would also be useful to examine the effect of inspection grade (perhaps the modal grade over all graded inspections) on foursquare user rating. Once again, a method like ANOVA would be appropriate.

*(5) Identification of health code violations most likely to lead to restaurant closure.*

Using recursive feature elimination and logistic regression, I selected the ten health inspection violations with the strongest relationship to restaurant closure. Unsurprisingly, many of these were serious violations, including three “Public Health Hazards”, and four “Pre-Permit Serious” violations. These violations can all result in immediate closure of a restaurant. In the case of Public Health Hazards, the restaurant can stay open if they correct the violation before the end of the inspection. See <https://www1.nyc.gov/assets/doh/downloads/pdf/rii/blue-book.pdf> for more details. Restaurants worried about closure should first make sure that they are not at risk for any of these violations.

Interestingly, three of the retained violations are not “Critical” violations. One deals with the risk potential for vermin. The final two are unscored violations, documenting either a missing DOHMH notice (which gives the public the results of the latest inspection), or “Failure to comply” with an inspector. The lesson here might be: if you don’t want your restaurant closed, respect the process, and listen to the Health Inspector!

Using principal components and logisitic regression to select important features, I found that violations pertaining to pests or the presence of pests seem to be the most important. For example, presence of mice, flies, or roaches. Surprisingly, none of these infractions are actually coded as Public Health Hazards by the DOHMH. The fact that they are linked to closure despite not carrying a penalty of “closure” is quite interesting to me. The result does make intuitive sense – if I were a health inspector, the presence of flies, mice, or roaches would certainly seem like cause for restaurant closure!

*(6) Spatial summaries of health inspection results designed to assist the DOHMH in setting of priorities.*

By visualizing inspection results and efforts over space, I revealed regions of NYC where restaurants perform particularly poorly on health inspections. The policy of the DOHMH is to revisit poorly-performing restaurants more often. With this in mind, I also visualized the average number of inspections per restaurant over space, showing where the DOHMH efforts are focused. Finally, I visualized a ratio between the average points per inspection and total number of inspections over space. This final graph may be particularly informative for the DOHMH, displaying regions where restaurants are performing poorly relative to the Departments efforts in that region. Such regions may warrant additional attention from the DOHMH.

**Conclusion**

Each of analyses presented here are still quite preliminary. For example, much more work is needed to improve the successful matching of health inspection records and Foursquare records. In a real-world application, I would need to also deal with updates to these database. Nevertheless, I will offer here some initial conclusions.

(1) The way health inspections results are displayed publically (inspection grades) may hide information from the public. Points per inspection could be a more informative metric for consumers, restaurants, and the DOHMH.

(2) It appears that there is only a very weak correlation between health inspection performance and Foursquare user rating. It may be that only very good or very bad performance on health inspections can successfully predict foursquare user rating.

(3) Aside from major problems with facility design, which should be dealt with before a permit is acquired, it appears that the presence or risk of pests may be the most important factor for predicting whether a restaurant is closed after inspection by the DOHMH.

(4) Unsurprisingly, there are spatial patterns in restaurant performance and health inspection frequency. The DOHMH could use there patterns to help prioritize future inspections.

Overall, it seems obvious that by combining the DOHMH dataset with Foursquare location and rating data, we can use data science to provide valuable insights for consumers, restaurants, the DOHMH itself.