November 7, 2014

Doug Steinberg
Sriniketh Vijayaraghavan

Data Science:
Foundations Final Project Proposal

For our foundations final project, we would like to explore Department of Health and Mental Hygiene (DOHMH) Restaurant Inspection Results. This data is available on https://nycopendata.socrata.com/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/xx67-kt59?, and provides a wide array of information on New York restaurant inspections. The open data includes the restaurants that were inspected, the location of the restaurants, the cuisine description of the restaurant (ie. bakery, coffee shop, Japanese food…), and the inspection date. In addition, the data shows the number of violations and violation descriptions per restaurant, the overall score for the restaurant, and lastly, the final grade given to the establishment.

Using this extensive library, we would like to find correlations between specific kinds of restaurants and their inspection grades. We would also like to observe any significant difference in inspection grades by specific location or borough. It is our hypothesis that bakeries, American restaurants, large-chain establishments, and delicatessen restaurants will consistently have higher grades than Chinese, Indian, Thai, and family-run fast food locations. We also predict that establishments in Manhattan, and in generally wealthy areas, will boast higher grades than equivalent restaurants in outer boroughs and in poorer areas.

Lastly, we would like to pair this data with information from Yelp. Using Yelp's price classification system, we can determine whether inspection grade has any effect on the price of the restaurant items. We predict that more expensive establishments will have higher grades, but that these higher grades are not the cause for the increased prices. If possible, we will also try to model restaurant grades with their corresponding Yelp reviews. We predict that establishments with more positive reviews will tend to have higher inspection grades that those with more negative Yelp reviews.

To investigate these hypotheses, we will utilize a combination of the models we learned in class, including: bivariate and multivariate models; logit and nearest-neighbor classifiers; Markov probability matrices; autoregressive and moving average methods; and Bayesian models. Hopefully these tools will enable us to make some sense of the notorious inspection grades we so often see at the front of our favorite restaurants.