Course: CS 109A Introduction to Data Science (Fall 2016)

Final Project: Predicting Food Inspection Outcomes in Chicago (Milestone #3)

Team Members: Calvin J Chiew, Angelo Kastroulis, Tim Hagmann

Teaching Fellow: Taylor Names

Data Exploration

Please see the following 3 ipython notebooks on GitHub for our preliminary data exploration:

GitHub URL: https://github.com/angelok1/cs109project

- Food inspections dataset: Food_Inspections.ipynb
- Business licenses, crimes and sanitation code complaints datasets: business-crime-sanitation.ipynb
- Climate dataset: Climate.ipynb

File Layout

The following directory structure is currenlty used in our github repository:

DIRECTORY	DESCRIPTION
`.`	Project files such as README
`./data/`	Data files created by scripts in `./CODE/`, or static
`./reports/`	Reports and other output are located in

Summary of Findings

Inspection type, past failure and risk category changes the probability of failing a food inspection. Failure rates also tend to increase in the mid-year months corresponding to summer and early fall when temperature is higher. There is also variation in the distribution of crimes and sanitation code complaints by location/district which may be correlated with failed outcome.

Moving Forward

We recognize that our analysis is currently fragmented and preliminary. Moving forward, we will integrate the various datasets, putting together the candidate predictors into a baseline model for inspection outcome. Specifically, we will:

- Use a broad definition of ‘food establishment’ instead of limiting our scope to restaurants, and consider facility type as a possible predictor of outcome
- Compare the performance of our models using a binary outcome (pass/fail) versus a multi-class outcome (pass/pass with conditions/fail)
- Map the food establishments into their districts/wards and calculate inspection outcome rates for each area as a baseline for comparison
- Merge the food inspection data with business licenses data by license ID, with crimes and sanitation data by location, and with climate data by date (and possibly location)
- Create baseline models using different classifiers (eg. logistic regression, LDA/QDA, random forests) and evaluate them