Tianying Zhang tz1416@nyu.edu
Annan Zhang az2345@nyu.edu
Qiang Cui qc687@nyu.edu

Big Data Project Summary

- **Project:** Spatial Profiling

- **Previous work and references:**

https://github.com/Shubham617/MapReduce-Project
https://github.com/ketanshahapure/DataScienceProject
https://github.com/dirtydupe/cisc_3140_Midterm

- **Problem description and goal:**

1. Location vs. critical violation
2. Cuisine vs. critical violation
3. Covid influence (also any improvement for restaurants over time)
4. (Optional) Water Consumption vs. violation/grade
5. (Optional) Cuisine/location vs. specific kind of violation
6. Goal: find the relationship between above entries, discover the trend of the data.

- **Date Set:**

DOHMH New York City Restaurant Inspection Results
https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

- **The method/approach you propose:**

Map Reduce, machine learning, Geospark

- **Evaluation criteria:**

R^2, p-value, … other statistical measurement

- **Week-by-week schedule with milestones for the different group members.**

Week one:

Create jupyter notebook:
https://colab.research.google.com/drive/10AMODvh3fjjnKJqcbe-333vOpZBBI6qO?usp=sharing

Do respective and discuss Tuesday:

1. check incorrect values (typos - brklyn; inconsistent zip codes or city names)
2. Data missing for certain regions

Do through a meeting:

1. Separate data set by year, group by restaurant.

Try respectively then discuss (Thinking and preparing during Thanksgiving and discussing on Monday)

1. Transfer to Geo(most important): Combine lon,lat, build model,
2. Check data quality again

Week two:

Finish Geo transfer: visualize the restaurants on map at least.

Start Analysis:

1. Prepare different methods/models to test.

Analysis and text work(construct report)

Prepare presentation