Chloe Bergsma-Safar
February 16, 2022

**Classification Project Proposal**

**Business problem:**

I seek to differentiate between open New York City Department of Housing Preservation and Development (HPD) violations – more specifically, Class B (hazardous) and Class C (immediately hazardous) violations – that result in a false certification or not. A false certification means that the violation was certified as corrected by the owner on time and properly but, after a subsequent re-inspection of the condition by HPD, was found to not have been properly corrected. False certifications are costly for property owners, since they are subject to civil penalties and can also result in criminal charges and for HPD, since HPD may incur litigation expenses if it initiates actions in Housing Court. More importantly, false certifications are costly for tenants, since they mean that tenants continue to live in hazardous conditions.

If false certifications could be predicted, it would help HPD take steps to prevent them from occurring and/or take swifter actions to validate certifications through re-inspections.

**Impact Hypothesis:**

By taking into account various features of a Class B or C open building violation - such as the # of open complaints in the building; the # of open violations in the building; the # of housing litigations in which the building has been involved; the # of apartments in the building; and demographics of the building's community district, such as poverty rate and % Non-Hispanic White - HPD will be able to predict whether a violation results in a false certification.

**Data:**

- Housing Maintenance Code Violations
- Complaint Problems and Housing Maintenance Code Complaints (these have to be merged before they can be compared with violations data)
- Housing Litigations
- Community District Indicators Data
- Buildings Subject to HPD Jurisdiction (for building characteristics, such as # of legal stories in a building, # of apartments in a multiple dwelling)

**Tools:**

- Python data cleaning, EDA and modeling
- Tableau for data visualizations
- SQL for storage

**MVP:**

A minimum viable product for this project will be a base classification model (perhaps a logistic regression model) using my target variable (False Certification v Not False Certification) and one or two features, such as the # of open complaints per buildingID and # of open violations per building ID.