I have been working with the 311 dataset on NYC Open Data a lot through the Data Journalism Program at Columbia. I was first looking at the rodent issue in Manhattan for my Writing with Data class. Then I noticed that most of the complaints are related to vacant lots/buildings in its location type. I was surprised to see that there are so many properties vacant in this busy city and decided to dig into those properties.

**Methodology**

The origin dataset is a nine-year 311 service requests dataset(2010-2019) from the NYC Open Data. The data is filtered by location type that contains“vacant” and by zip codes, 10026 and 10027. The 10026 and 10027 zip codes represent the Harlem neighborhoods. The origin dataset has 195 rows of data. The data is then inputted to the Jupyter Notebook and processes in Python. Pandas’ dataframe is used through the process. Incident addresses without house numbers are identified by the regular expression and dropped out of the dataset, which brings down the data size to 184.

The dataset is then randomly selected and manually checked on Google Map in order to evaluate the accuracy of the data. It finds that some of the house numbers are not the exact number of the vacant properties and the properties need to be further identified.

All the locations are visited and identified in person. Properties that are vacant, under construction and newly built, are recorded. In the process, resident interviews are conducted and pictures are taken. Since massive interviews are conducted, the author decides to make the project as profile stories of vacant properties to provide the vacancies with narratives.

After identifying all the vacant lots, the incorrect addresses are crossed out and 64 vacant properties near those incorrect locations are added to the list. All the data is inputted into the CSV file manually. The data size is 113.

The data is inputted into the Jupyter Notebook. Selenium Webdriver is used to scrape the properties’ ZoLa URLs. ZoLa is the city’s zoning and land use map system built by the Department of City Planning. The website provides property information. The attempt was to scrape the property information with Beautiful Soup. However, the information is wrapped in the JSON script in HTML codes. Therefore, the property information includes owner, land use, lot area, year built, year altered, gross floor area and residential units are manually copied and inputted to the dataset.

For the process log, please see the Jupyter Notebook posted on GitHub at https://github.com/hshirleyuhu/hshirleyuhu.github.io/tree/master/masters-project