Building a Map Tool to Identify Stalled Construction Sites in NYC

Georgia Tech

Mackenzie Mull, Michael Anderson, Nathaniel McKeever, Zihang Yuan, and Aaron Ramirez

Introduction

- Using stalled construction site and property value data from New York City, we aimed to create a **live, interactive map** that plots report locations and ultimately reveals **interesting geographic trends** relating to the reported stalls.
- Sidewalk sheds, which are associated with city construction, have negative impacts on pedestrian health, safety, and overall socioeconomic well being.

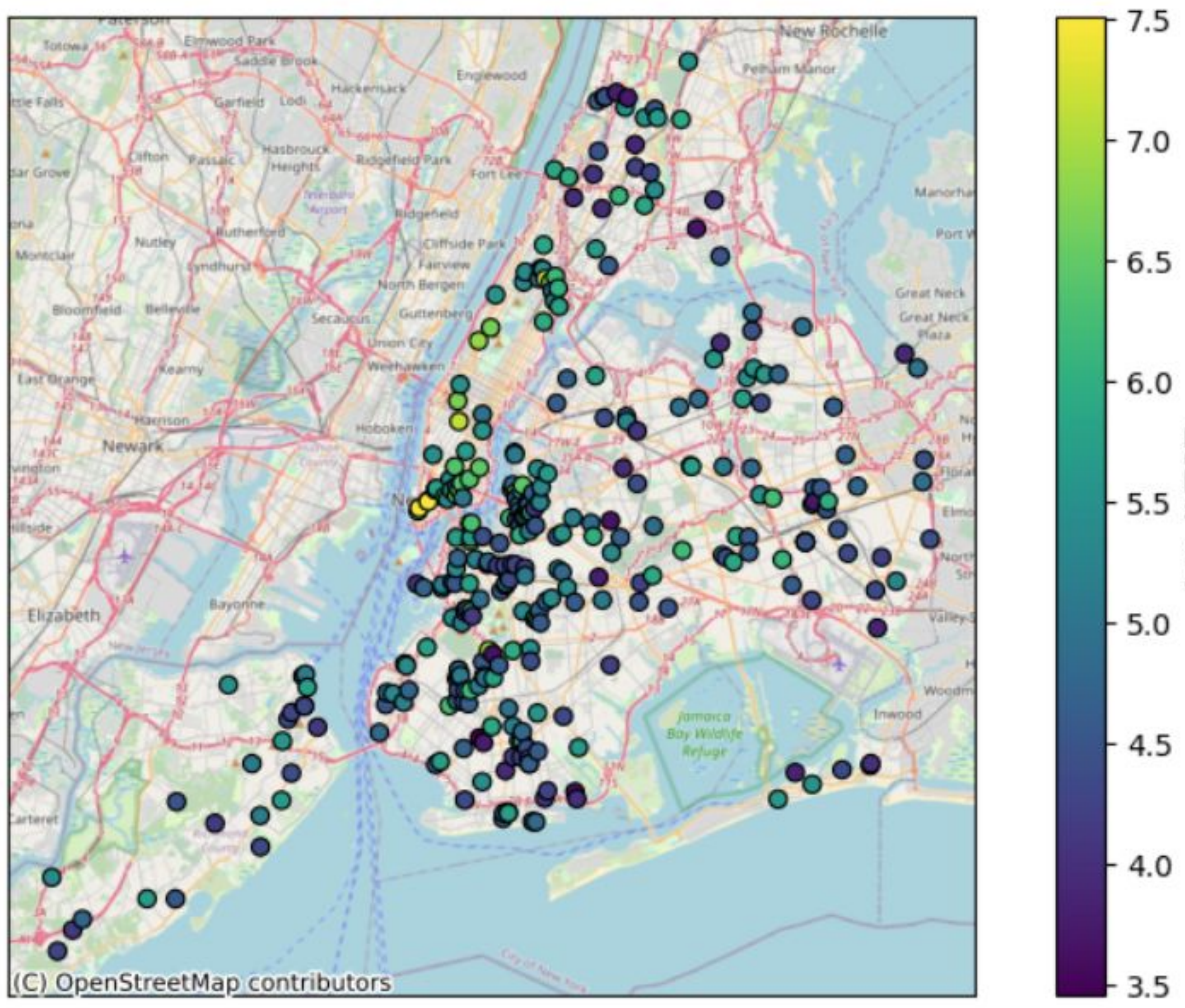
Database Curation

- The "DOB Stalled Construction Sites" and the "Property Valuation and Assessment Data" datasets were downloaded from NYC Open Data website and joined using the building identification number (BIN).
- Datapoint variables include info such as **borough, date of complaint, coordinates, and assessed value**. All original rows remained after cleaning (**>1 million data points**).
- API calls keeps data up-to-date

Innovative Approach

- Live mapping of data and report density visualization, which has never been done before for this dataset
- Intuitive user interface with color-blind friendly palettes
- Machine learning to model data (e.g. clustering) and identify trends
- **An effective visualization tool,** with features described above, **may ultimately spread awareness of stalled construction sites**

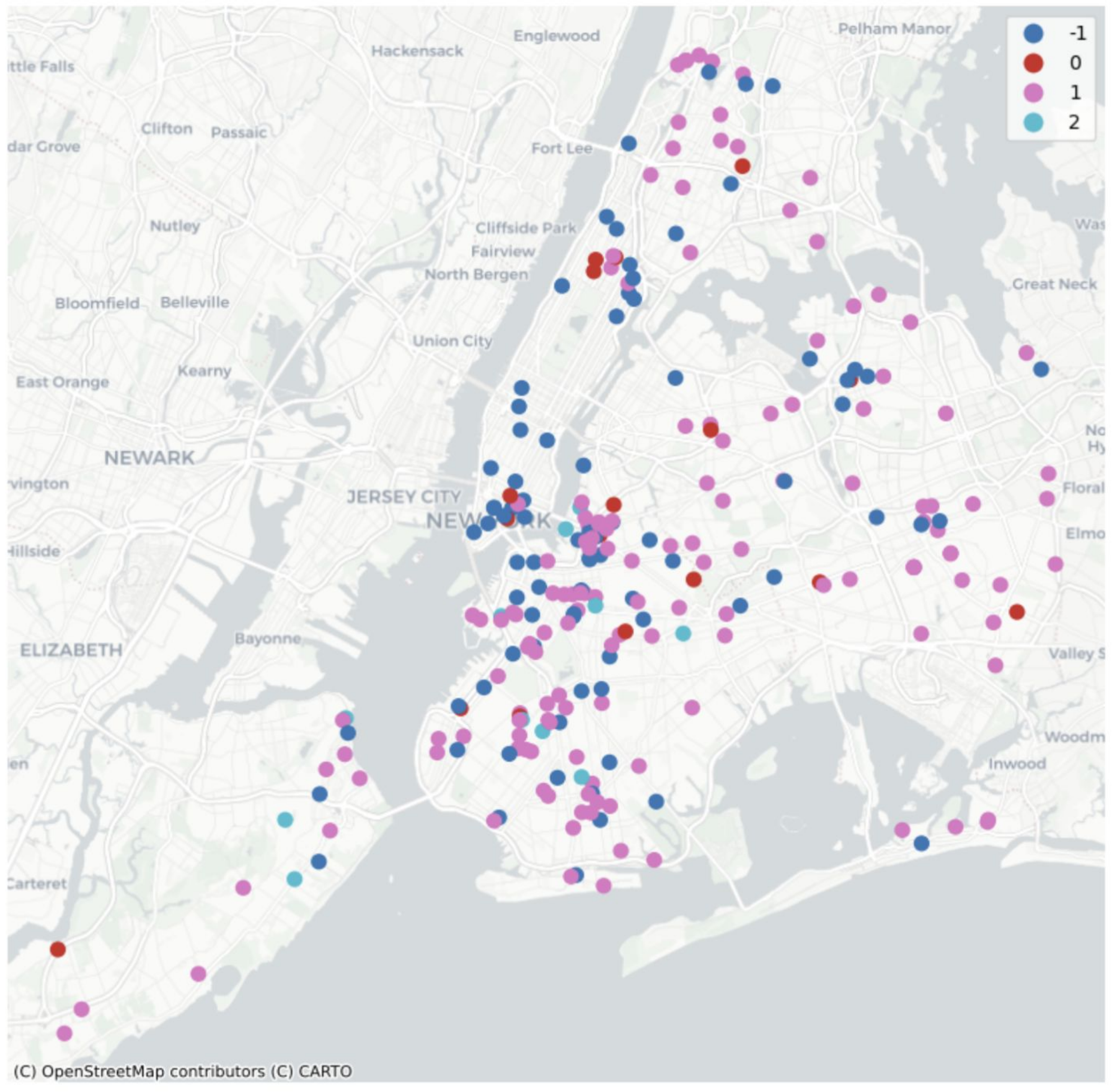
Preliminary Visualization



Before exploring ML algorithms, we wanted to better understand broad trends within the dataset. We first simplified the data by combining reports with the same latitude and longitude, which revealed a large range of report totals across the dataset when grouped by coordinate pair. The data was then log transformed to compress the range for better visual analysis (see above figure). These findings proved useful for our subsequent ML approaches and final visualizations.

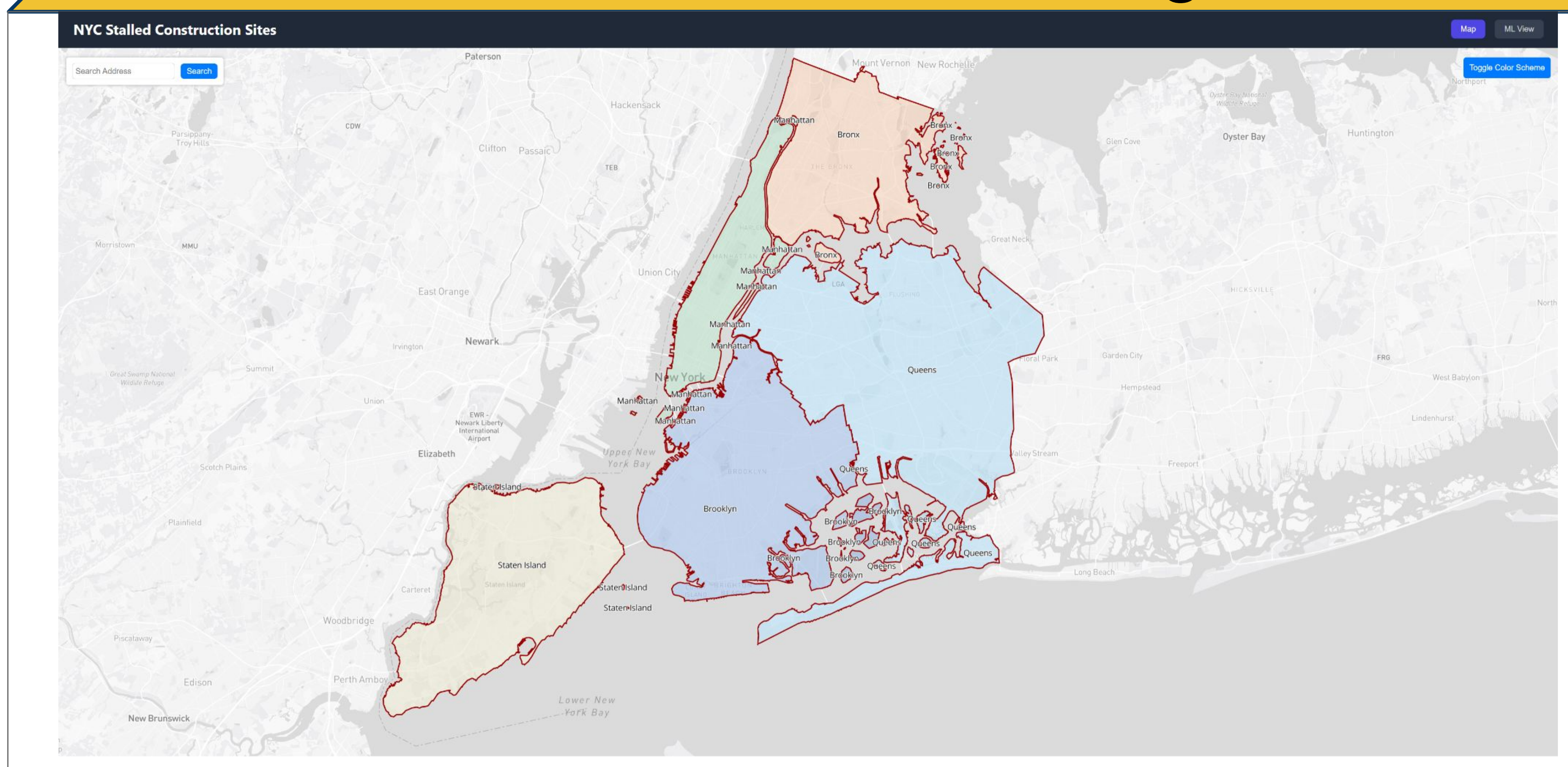
Cluster Analysis

Property Value, Days since Complaint Clusters Clusters (HDBSCAN, min_cluster_size=10, min_samples=10)



For cluster analysis, we switched to using BIN to group reports instead of coordinates. Variables examined included property value and age of complaint. K-means, Gaussian Mixture Model (GMM), and HDBSCAN clustering models were attempted, but none of these models revealed distinguished clusters and therefore these variables have no impact on geographic pattern. HDBSCAN modelling results are displayed above as we hypothesized it would best fit our data due to its effective handling of oddly shaped clusters.

User Interface Design



Above is snapshot of one of our map layers sectioned by borough. The interface features colorblind friendly shades, well defined borders, clear labels, the ability to zoom in and out, and search the map. All of these were key designs we had hoped to achieve with our maptool to ensure a user-friendly experience.

Experiments and Results

Evaluation 1: Data Loss Analysis
Assessing the amount of data lost after cleaning is a critical step in preserving data distribution and is used in all fields of data analytics. We, fortunately, had a 0% loss of data points after joining and cleaning of datasets.

Evaluation 2: Model Validation
We primarily relied on visually checking the map to see if distinguishable clusters were formed; these checks revealed no clear clusters, as described previously. For K-means and GMM clusters, we use silhouette scores to find optimal cluster number, which resulted in implementing n=5.

Evaluation 3: User Interface Survey
Unfortunately, we did not reach this part of our experiment phase due to time constraints and continued troubleshooting; however, user surveys are integral to ensuring interface success because of its insight into user experience.