Cambridge Eviction Study
CS 506 Final Project Report

Fuqing Wang, Zixiang Wei, Xiao Lu, Tian Chen

April 20, 2020

1 Introduction

1.1 Background and Motivation

Over the past two years, the City of Cambridge has been collecting and analyzing eviction data from the State of Massachusetts online court docket system. The Cambridge, MA Community Development Department wants to study eviction filings from the Massachusetts state court dockets supported by the MAPC and Census data to determine demographics and addresses.

1.2 Goals

Our goal is to distinguish what types of housing the evictions happen under and what conditions cause the evictions, a major classification is between affordable housing units and private sector housing units. We would also like to examine the cause for eviction, our current hypothesis is that money is the primary cause for evictions however we would also like to be able to predict behavioral causes such as noise complaints about the eviction filing. In the following project report, we are going to demonstrate the process and techniques we used working towards this project. The report is going to be divided into four major sections, data collection, data preparation, analysis, and presentation.

2 Data Collection

- The main body of the eviction data comes from two parts, the latest data of the most recent two years and the data from previous years. We obtain the latest data by scraping from the MA State Court website(the detail is described in Eviction Data Scraper Instructions.doc) and the data from previous years can be accessed through a accdb database which is provided in the docs. These two parts are then combined into a single csv file which we later used as the main dataset.
- As an expansion of the main dataset, we have also included AffordableUnits and Evictions Inclusionary tables from accdb database as an eviction housing type classifier.
- We have also collected Cambridge house price data from the City of Cambridge Open Data Library to add a possible price classifier.
- We have also collected Cambridge crime report data from the City of Cambridge Open Data Library to add possible classifiers.
- We have also collected Cambridge population distribution from the City of Cambridge Open data Library to add possible classifiers.

3 Data Preparation and cleaning

- The raw eviction data consist of several irrelevant properties, such as defendant, judgment type, case status, and total execution, etc. We disposed of all of these fields as well as other null value entries for the analysis.

- We added geo-coordinates to each entry. Property address is helpful for geographical classification yet barely makes sense as is, thus we used Python Geocoder to map each address to a latitude and longitude coordinate.
- We added eviction housing types to eviction data. And we did this by joining evictions, AffordableUnits and InclusionaryRentalUnits tables using regular expression matches so that we have a new type column in the raw eviction data, where 0 indicates private housing/other, 1 indicates affordable units and 3 indicates inclusionary rentals.
- We have also added a price tag with respect to each entry to the eviction data. We did this by first obtaining a price and geo coordinate relationship using Cambridge house price data from the year 2016-2020. Then we map each geo coordinate in eviction data to a price we found(If the exact coordinate is unfound, we used the one that is the closest). Eventually, we normalize the data since the housing prices are mostly big numbers.

4 Analysis

The number of eviction data entries we had is 1324 in total, which is a lot smaller than expected and is unsuitable for model training/analysis. To work around the problem of limited given entries, we instead worked on finding the relationship between eviction and other factors like house values and population distribution given that the dataset we collect from Cambridge Open data library is abundant(20000+ entries). We run several clustering algorithms(GMM, Kmeans) on the data we collect, respectively Cambridge house values, crime rates and population distribution.

We set 5 cluster centers in our models and run the model. Then for the clustering results, we count the total number of points and calculate the average house values within each cluster. The number of points and approximate range of house values are as the following table:

Cluster	*Average House Values*	*Number of Points*
1	613, 035	18278
2	15, 080, 518	2013
3	737, 759, 700	24
4	915, 527, 500	33
5	1, 070, 662, 320	75

Table 1: GMM Cluster Results

As a result, we have 5 classes and we set low value class's color to no color in our ArcGIS map, high value class's color to orange, green red and blue. Owing to small sizes of other classes, only one type of high value class can be seen in the map and we treat other high value classes as outliers.

The clustering result with house values and geography location(longitude and latitude) reflects the distribution of house value in different areas and how eviction cases are related with levels of house values. Green points mean eviction points. According to Picture, we can say that eviction cases are more likely to occur where house prices are relatively lower.

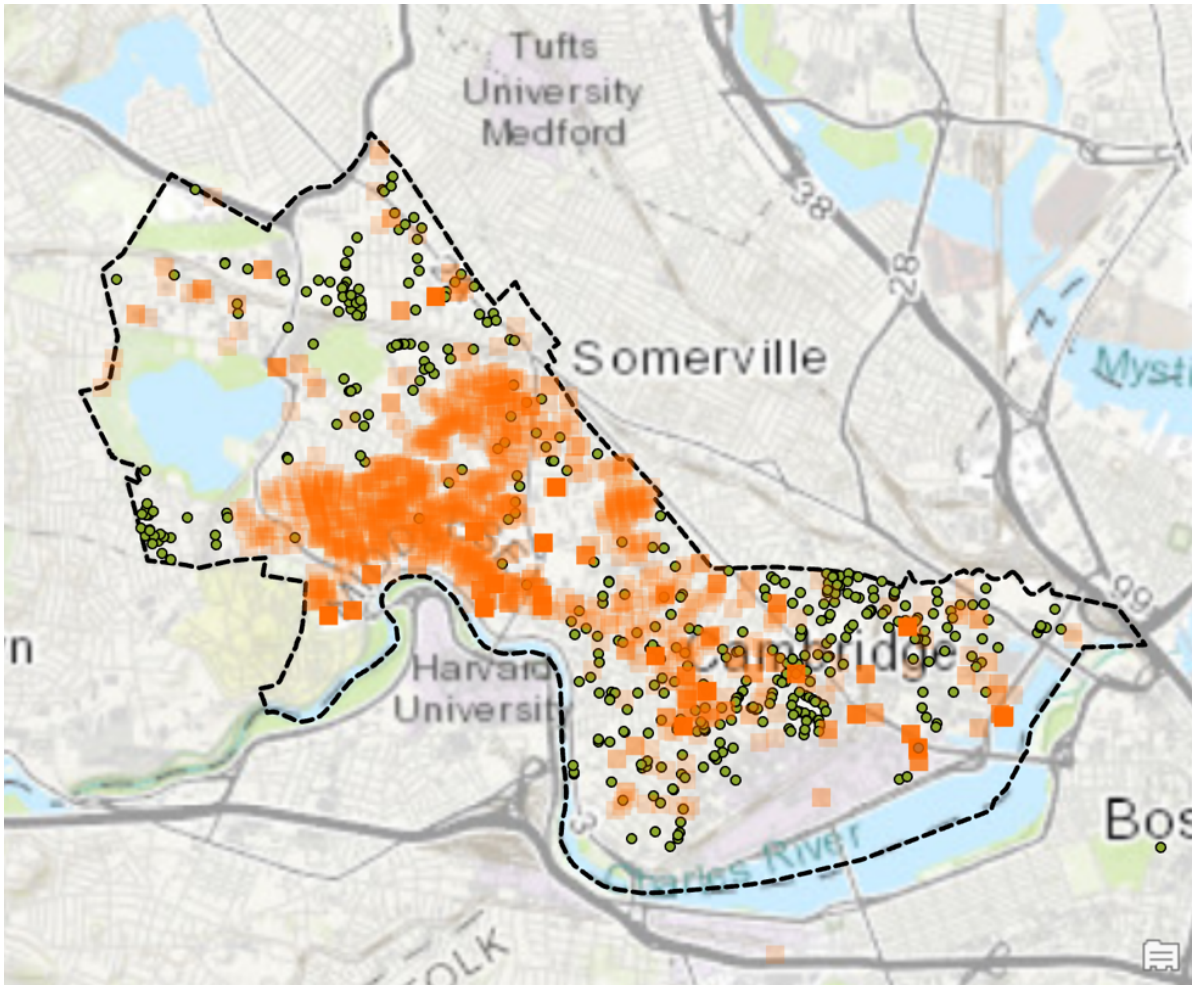

Figure 1: Clustering Visualization

5 Presentation

When we have the values of the properties in Cambridge and the eviction dataset, we convert them to CSV files and load them in ArcGIS Pro. Thereby we get eviction latlng.csv and value latlng.csv tables and we can right-click these tables and choose "Display XY data" to generate data points in the map. Given these data points, we use values of the properties to generate property values heatmap layers to see the relationship between property values and evictions. To generate the heatmap, firstly open ArcToolbox in ArcMap. Click Spatial Analyst Tools ¿ Density ¿ Point Density. Then configure the parameters in the Point Density dialog box and run it. To show the relationship, we display the property data points which are used to generate the heatmap and display the heatmap and the eviction points. The deeper color(more red) means more expensive properties. The result map should be something like the follows:

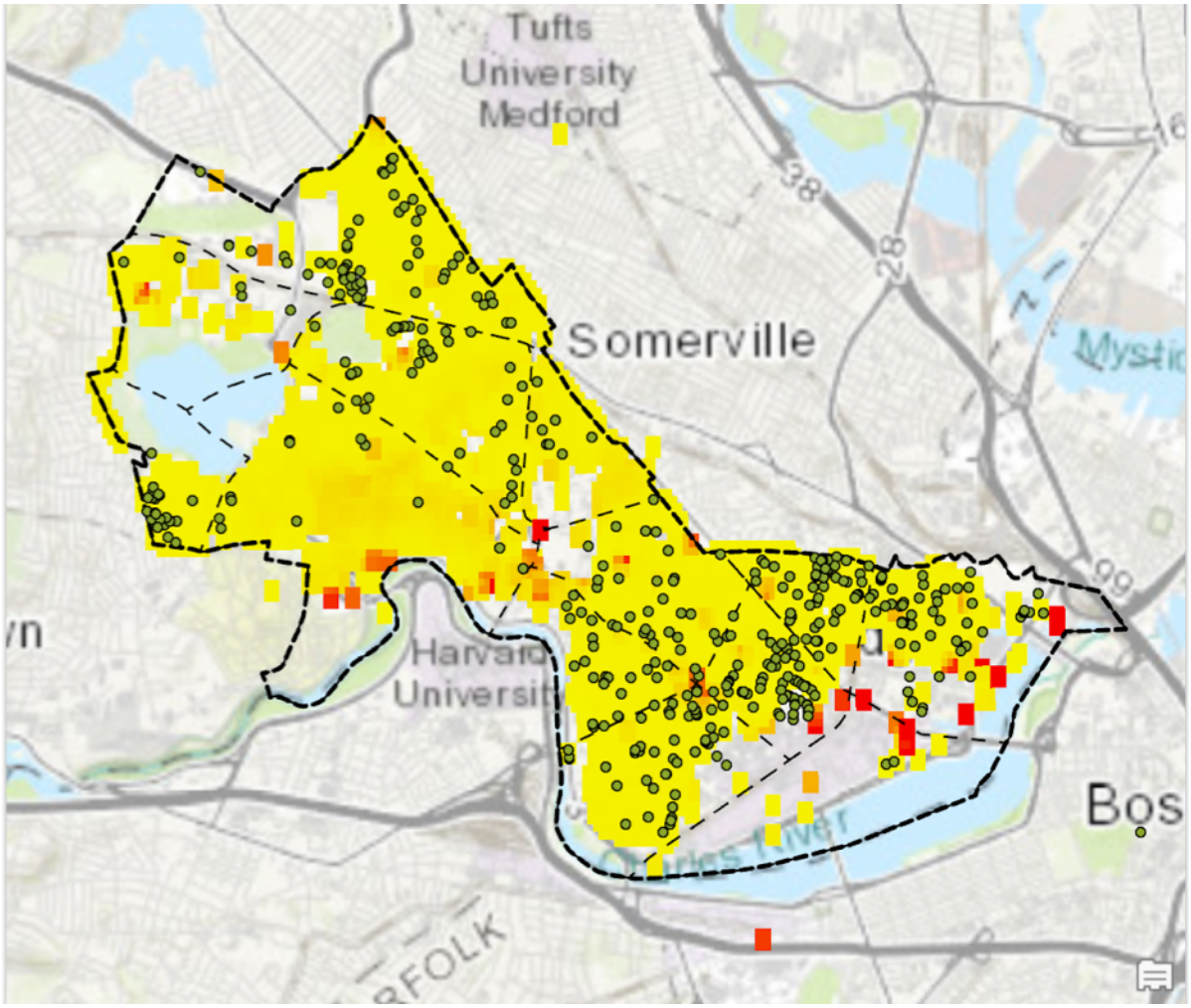

Figure 2: Heatmap Visualization

5.1 Results

According to Fig. 1, our data shows the majority of people who get evicted lives in areas close to the city edge where house prices are low. Does this indicate the evicted are mainly low income people? To address this assumption, we plotted an income-eviction relation heat map. Where the color indicates the income range(the deeper the color, the higher the income) and the dots indicate the eviction coordinate. According to Fig. 3. Eviction frequency are a lot more higher in lower income regions than in high income regions(shown by the dot density), which proves our hypothesis.

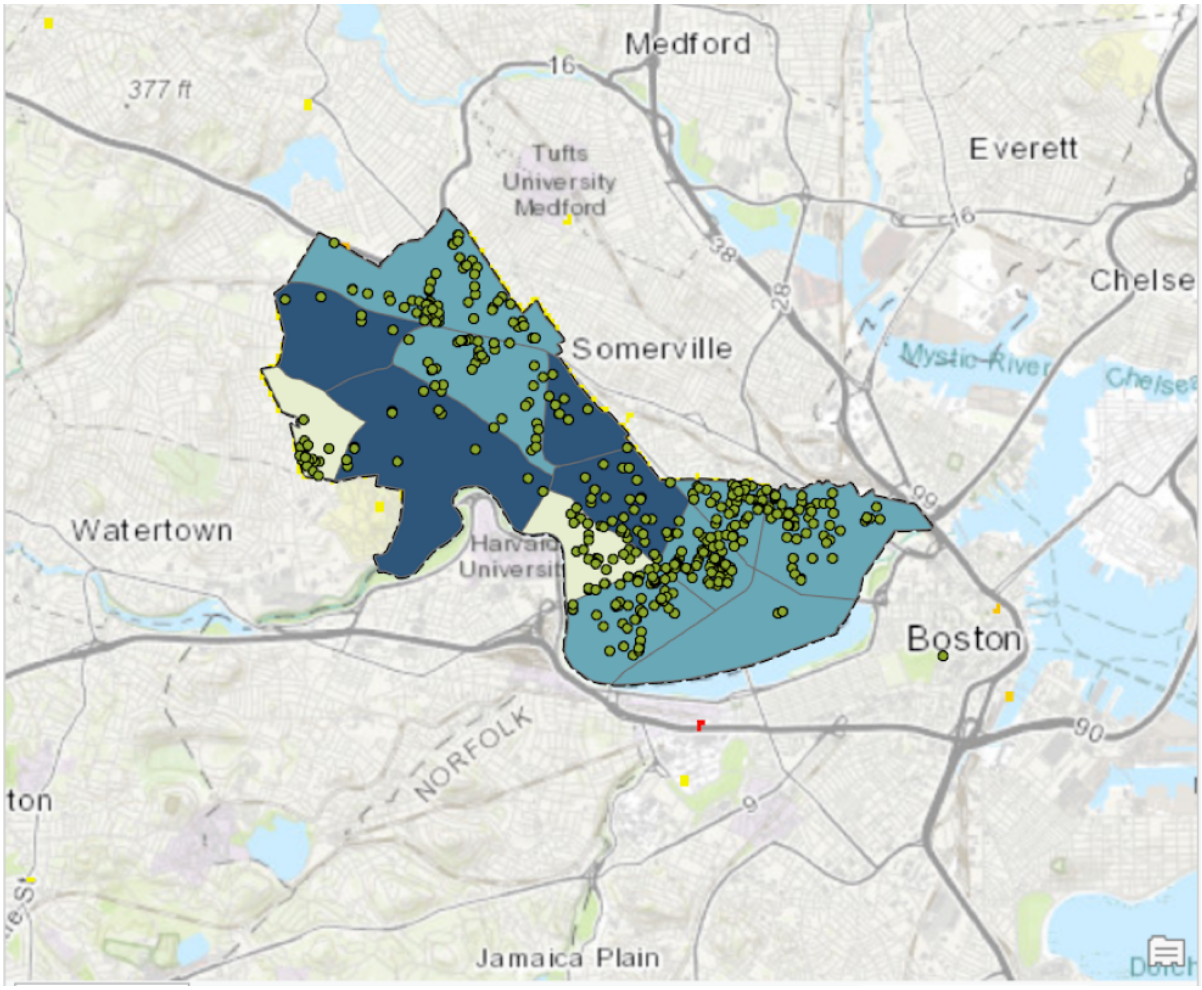

Figure 3: Income Distribution Map

5.2 Strategic Questions to be answered

- **Are there local demographic, physical housing conditions or spatial characteristics that make units more likely to be subject to eviction?**

  To address the demographic factor, we plotted a heatmap of distribution of people whose age were 18-30, the darker the region the denser the youth group is in such a region. According to Fig. 4, youth age as a demographic factor we addressed, does not have a direct relationship with eviction.

  We have not found any housing condition data thus we did not address the physical housing condition factor.

  The spatial characteristics are already mentioned in Fig. 2 and 3 - that eviction is more likely to happen on the city edge, low income areas.

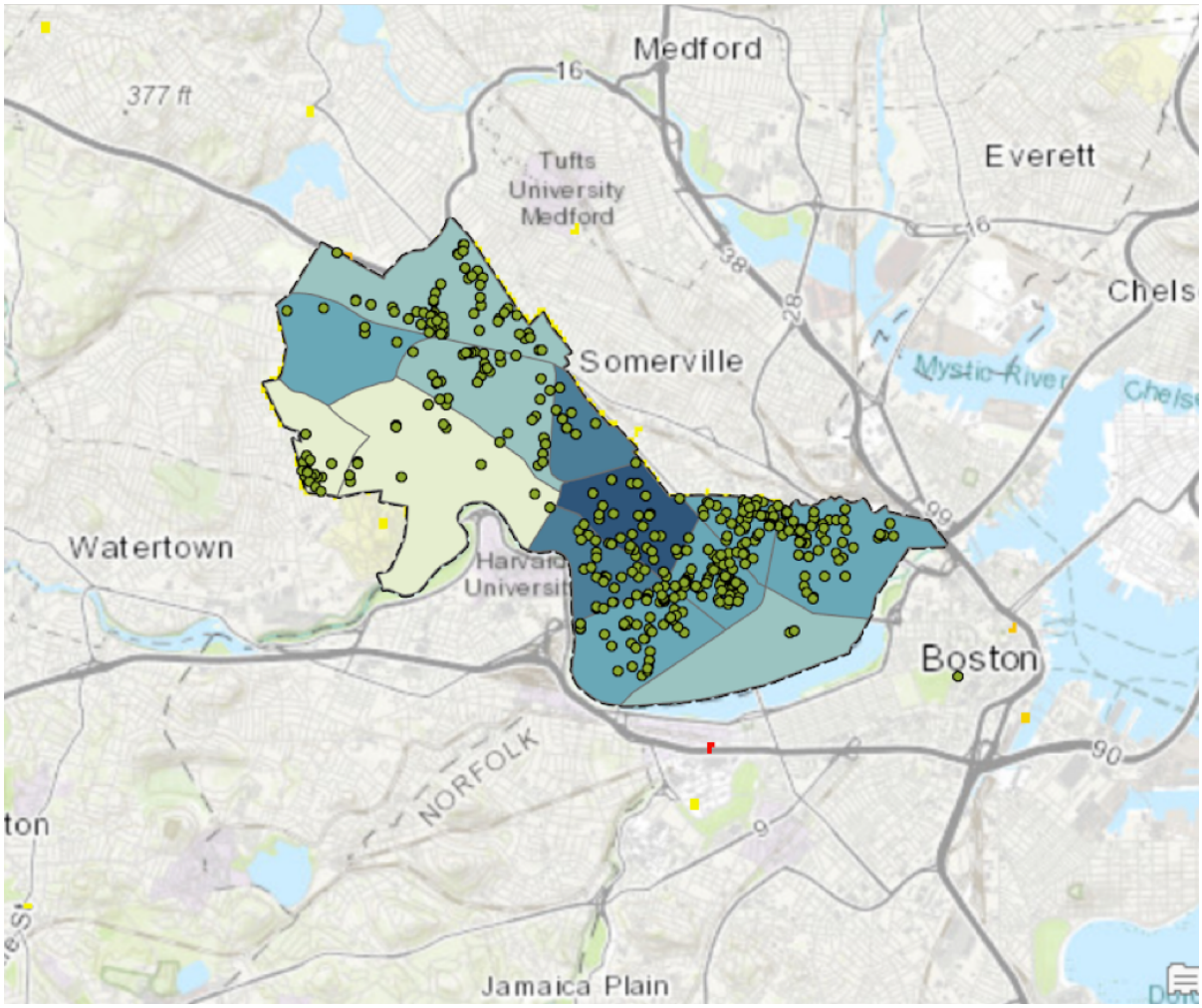

Figure 4: 18-30 Years Old Distribution Map

- **Are there distinct differences between housing that occurs in the low-rise and small scale housing stock (developments under 25 units) compared to the newer, larger buildings?**

  The map does not show distinct differences between small scale and large scale buildings. We cannot infer too much relevance between eviction cases and building scale. According to Fig. 5.

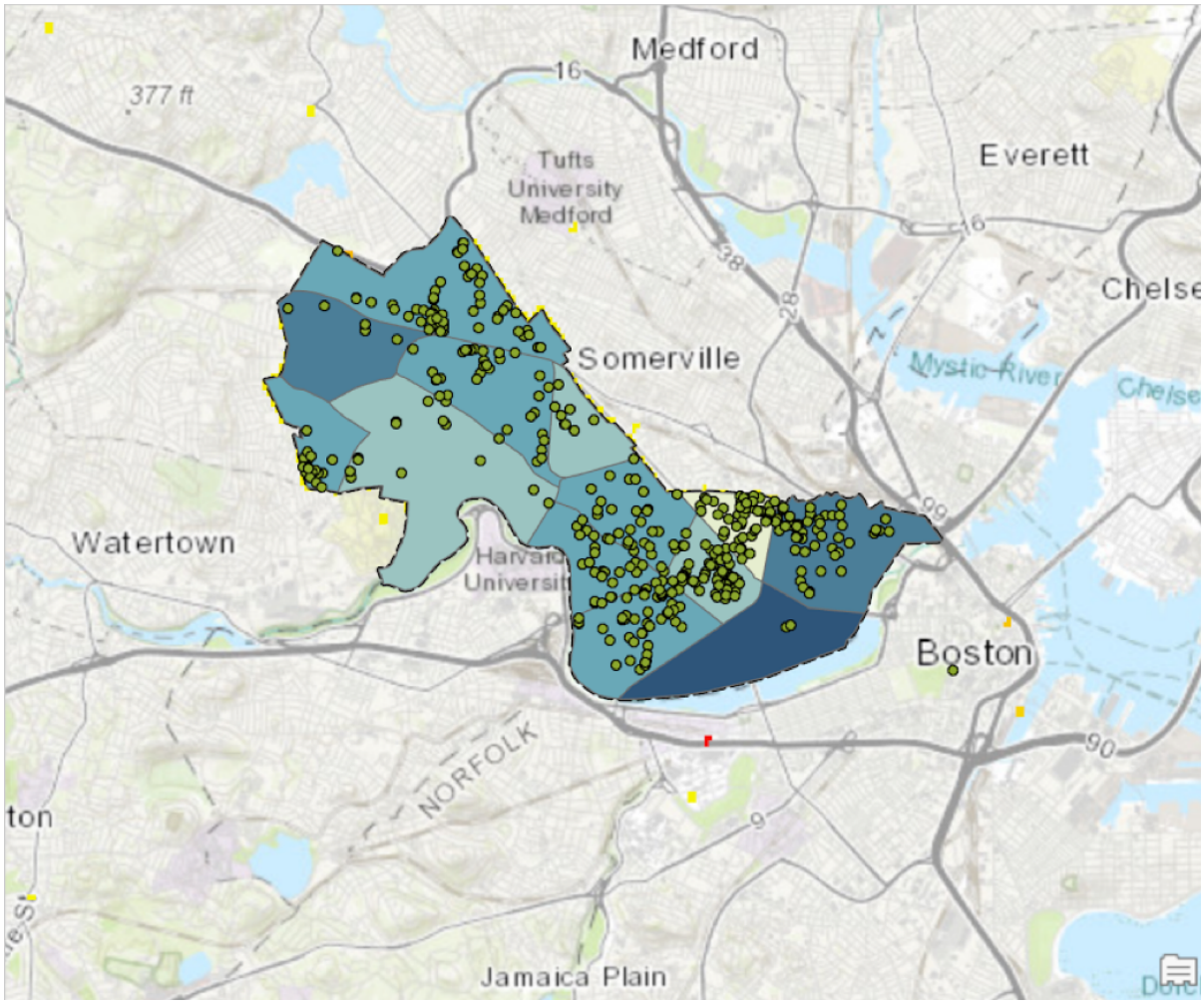

Figure 5: 25+ Units Distribution Map

- **How does tenant representation by an attorney affect outcomes?**

  The below two charts show the outcome for cases with and without attorney. According to the pie chart, cases with attorneys have a higher percentage of being closed(15.6%) compared to cases without attorneys(3.5%) after judgement. Cases with attorneys also show a higher rate of staying open(9.4%) after judgement compared to cases without attorneys(2.2%). P.S. This analysis result is limited by our sample size, cases without attorneys have 1132 cases, cases with attorneys have 97 cases.

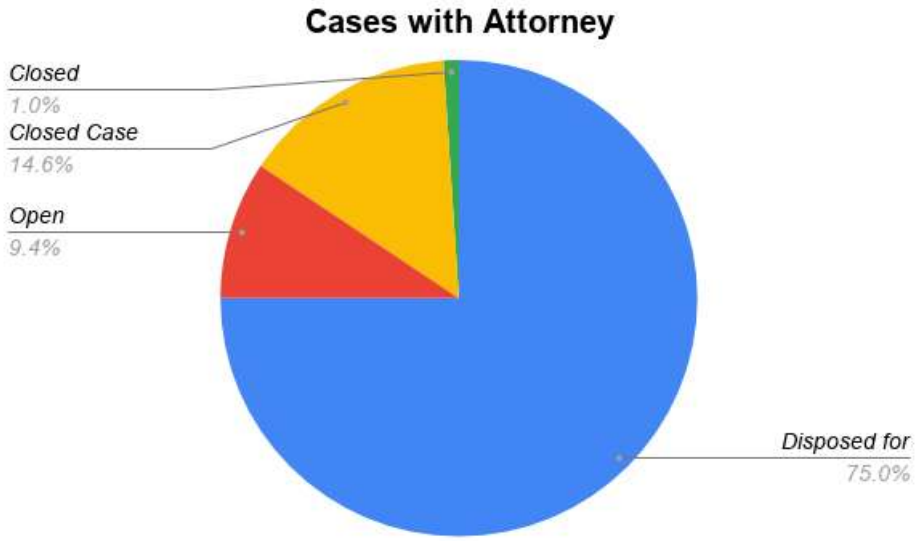

Figure 6: Case status for cases with Attorney

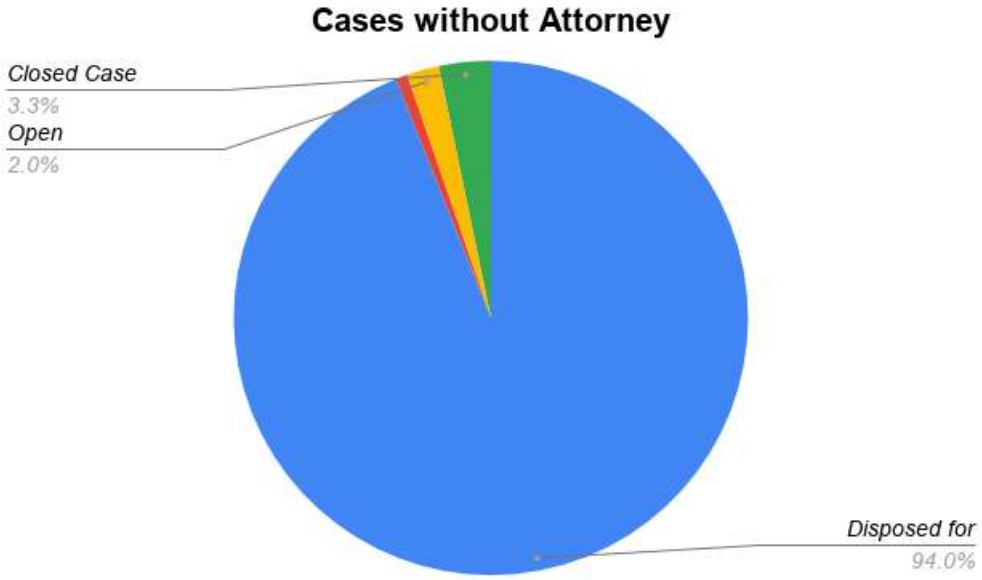

Figure 7: Case status for cases without Attorney

6 Summary

6.1 Obstacles

First, note that the court website is a subject to change. By the time we worked on this project, the web scraper given does not match the webpage. We had to modify a lot of parts of the scraper to make it work again.

Second, the eviction data we obtained is very limited and only has around 1300 entries. We had to research different data sources online and collect additional data sets to support our analysis.

6.2 Limitation

As mentioned above, the limitation is that we only have about 1,325 eviction data. We expanded our data set, but some of our analysis based on the original 1,325 entries might be biased due to limited sample size.

On a supervised learning analysis perspective, we only have eviction data, which means we do not have positive samples(people who isn't evicted). So it is unlikely for us to move towards the supervised learning analysis direction. On the other hand, the data set given is very small, unsupervised learning algorithms are also unlikely to give us a fair analysis. Thus it has been hard for us to go further on this project, such as making predictions on what kind of people are more likely to be evicted.

6.3 Conclusion

Based on our results from analysis - that the majority of people who get evicted lives in areas close to the city edge where house prices are low, we can conclude that expenses and housing prices is the primary factor that affects eviction. Also from different map plots, we can observe that eviction frequency are a lot more higher in lower income regions than in high income regions, which further proves our hypothesis, that money is the primary cause for eviction.