Team Name: City of Cambridge Evictions Study, Team 2
Team Members: Tiancheng Zhu, Kevin Peters, Drew Abram, Syahrial Dahler

Deliverable 4

City of Cambridge Eviction Cases

Original Project Description (from Spark!)

Over the past few years, the City of Cambridge has been collecting and analyzing eviction data from the State of Massachusetts online court docket system. This project includes three phases of work that improve and enhance data collection and analysis procedures:

1. Update the current Python data scraping script to address both additional data collection needs and privacy concerns under discussion with the Law department.

2. Automate the data cleaning process to the extent possible.

3. Link eviction data to other data sources, such as city master address list, assessing data, American Community Survey, housing code violations, and market rate rental housing prices to gather actionable insights.

Goals of the project are twofold. We seek to update and optimize our data collection and analysis procedures. We also seek to explore research questions to obtain a better understanding of what conditions are linked to eviction and potentially where evictions might occur in the future.

Goals

This project had three main goals:

1. Optimize data collection to combine data from many different sources. This project seeks to optimize the process by using Python scripts.

2. Optimize analysis process by using some commonly used methodologies and algorithms in data science.

3. Obtain a better understanding of the links between features and evictions and where evictions might occur in the future. Specifically,

a. Are there certain feature categories (income, gender, etc.) that make a defendant more likely to be evicted by a plaintiff?
b. Do certain apartment buildings, neighborhoods, or plaintiffs have higher rates of eviction cases?
c. How does defendant representation by a lawyer impact defendant case outcomes?

Challenges

Several challenges were faced throughout the course of this project:

1. Our main customer, the City of Cambridge, suddenly abandoned the project and revoked access to the eviction data that they had previously provided.

2. The Massachusetts Court Docket forbids automatic web scraping. The court website explicitly displays a warning against using automatic scrapers and crawlers to gather data.

3. The Covid-19 pandemic created the greatest challenges for the project. Boston University suspended all in-person meetings and required students to vacate on-campus residences in the middle of the semester. Like many other BU students, our team had to adjust to these changes; in-person meetings were cancelled for the remainder of the semester and we had to re-coordinate our group schedule.

Datasets

The data used in this project was gathered from several sources.

Main dataset

The main data is gathered from Massachusetts Court Docket

Collection:

The data was retrieved manually by downloading the html file of each case to a local machine. The downloaded pages were then processed using a Python script to collect all data of interest. Manually downloading of each detailed page of the eviction case was required due to the legal warning of the website which prohibited the use of scrapers to query the website.

Eviction cases in the City of Cambridge between January 2017 and February 2020 were collected. We collaborated with BU Spark! to devise a comparable plan so that we could still analyze court data without direct interaction with the original City of Cambridge client. Secondary datasets were retrieved by downloading datasets and combining features using Python. Some secondary datasets with few categories (such as those used to determine

affordable housing unit addresses within Cambridge) were manually retrieved and used to categorize samples in Python.

Features:

Feature	Value	Type
Case Number	Court case number	alpha-numeric string
Status	Status of case - Open, Closed, or Disposed	string
File Date	mm/dd/yyyy	string
Plaintiff	Party filing for eviction	string
P-Attorney	Name of attorney representing the plaintiff	string
Defendant	Name of the defendant (person risking eviction)	string
D-Attorney	Name of attorney representing the defendant	string
Property Address	Full address the case concerns	string
Docket	Details of case proceedings	string
Judgment Date	Date of judgment decision; mm/dd/yyyy	string
Judgment Type	Result of the case	string
Judgment Method	Settled or dismissed, and compromise	string
Judgment Total	Total dollar amount to be paid, usually from defendant to plaintiff	float
Execution Total	Actual total dollar amount due (including court fees and interest)	float
Latitude, Longitude	Geographic coordinates corresponding to the address	float
Units	Number of apartment units in the building at the corresponding address	float
District	Neighborhood within Cambridge which the address belongs to	String
Address	Standardize address to use as indicator to map with other data source such as housing_stat	String
median_income	The median income based on the neighborhood	float

Secondary datasets

1. The Cambridge Neighborhood Polygon dataset was used to draw the boundaries of each district in cambridge on the plot of geographic location of datapoints.

2. Google Maps and Open Map were used to retrieve the geographic coordinates of each address. Together with neighborhood polygon data, this data was used to determine within which neighborhood each property was located.

3. The Housing Starts Map dataset provides a table of all residential building permits in Cambridge. From this data, the number of units within some buildings was able to be determined. We created a Python script to map buildings in our main data set with addresses provided in Housing Starts Map.

4. The American Community Survey dataset provides estimated median income for each neighborhood in Cambridge, drawn from years 2013 through 2017.

5. Public housing databases were used to identify affordable housing in Cambridge. Even though these are not official databases, these websites provide the best estimate of which addresses are considered to be affordable housing that we could find.

Data Preparation and Cleaning

The document entitled “stepToProduceResults” in our project folder contains a description of the methods for data preparation and cleaning.

Project folder

Folder	Description
src	Folder that contain all python files
Analysis	Folder that contain the result of analysis such as Power BI chart and HTML page of the map created by using folium
CSV	Folder that contains the csv files. Csv created from scraping and other csv files and geojson files from Cambridge open data
Data	Folder that contains the html pages that we manually downloaded from the massachusett court

Limitations and Assumption

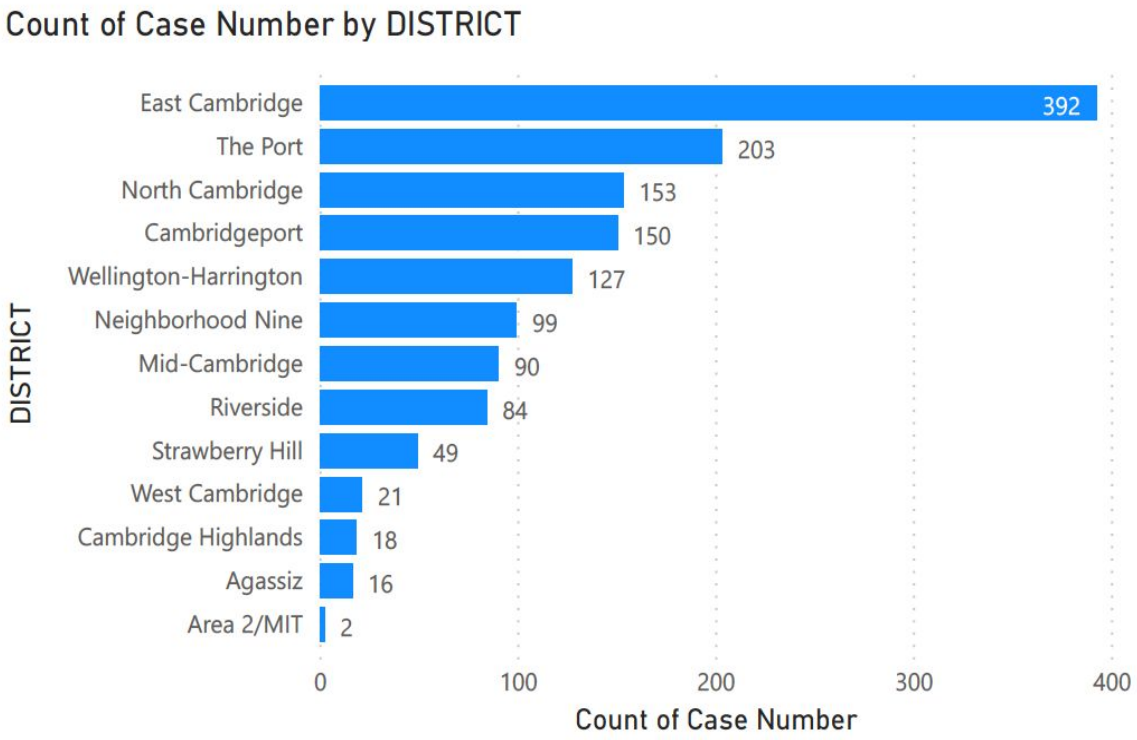
1. We assume that all cases of eviction are recorded in the Massachusetts court documents.
2. We are unable to get individual profiles of the tenants. For the income we use the median income of the district as the proxy to infer the tenant's individual income. Other defendant attributes (such as gender) were determined using a classification algorithm, typically a built-in Python package.

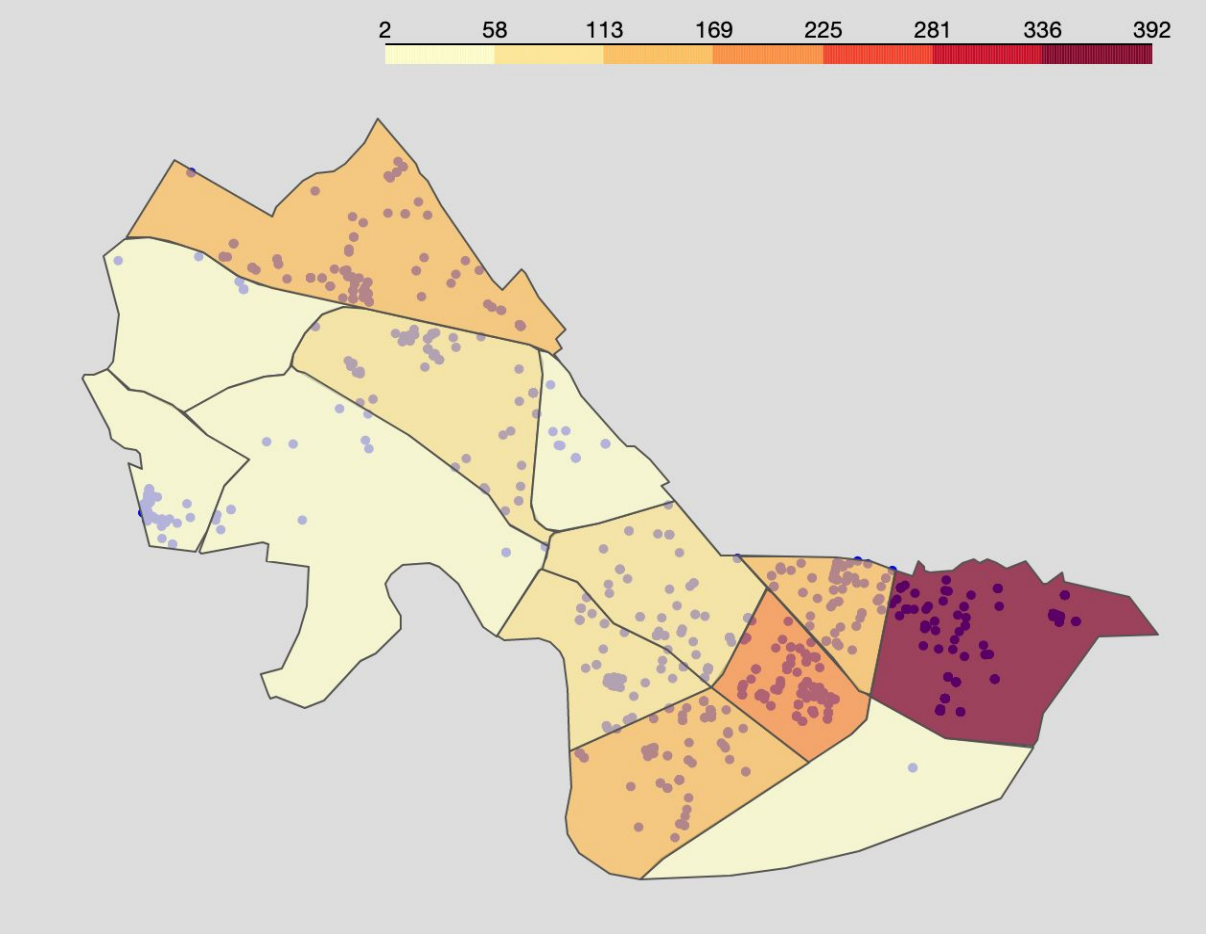
Analysis and Findings

The data contains 1404 cases of eviction which were brought to court between 2017 and 2020.

Eviction Cases By District

A majority of cases occurred in East Cambridge as depicted in the figure below. Almost a quarter of all eviction cases involved a resident of East Cambridge.



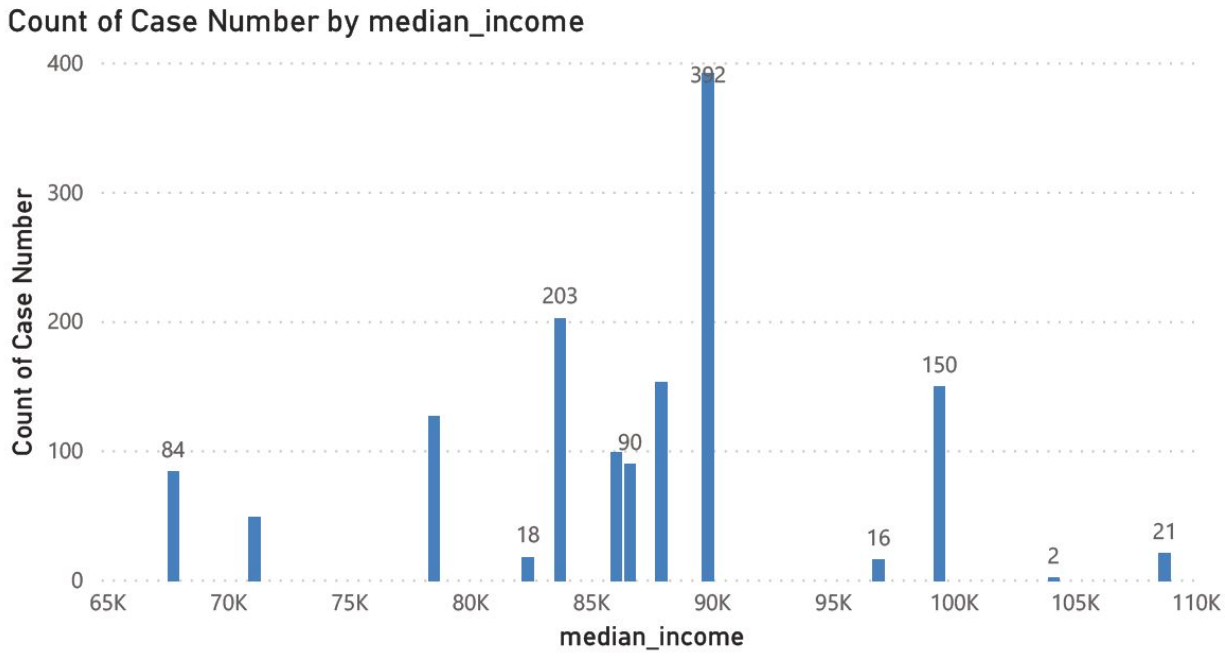


The map produced above can be found at analysis/map.html

Eviction Cases by Income

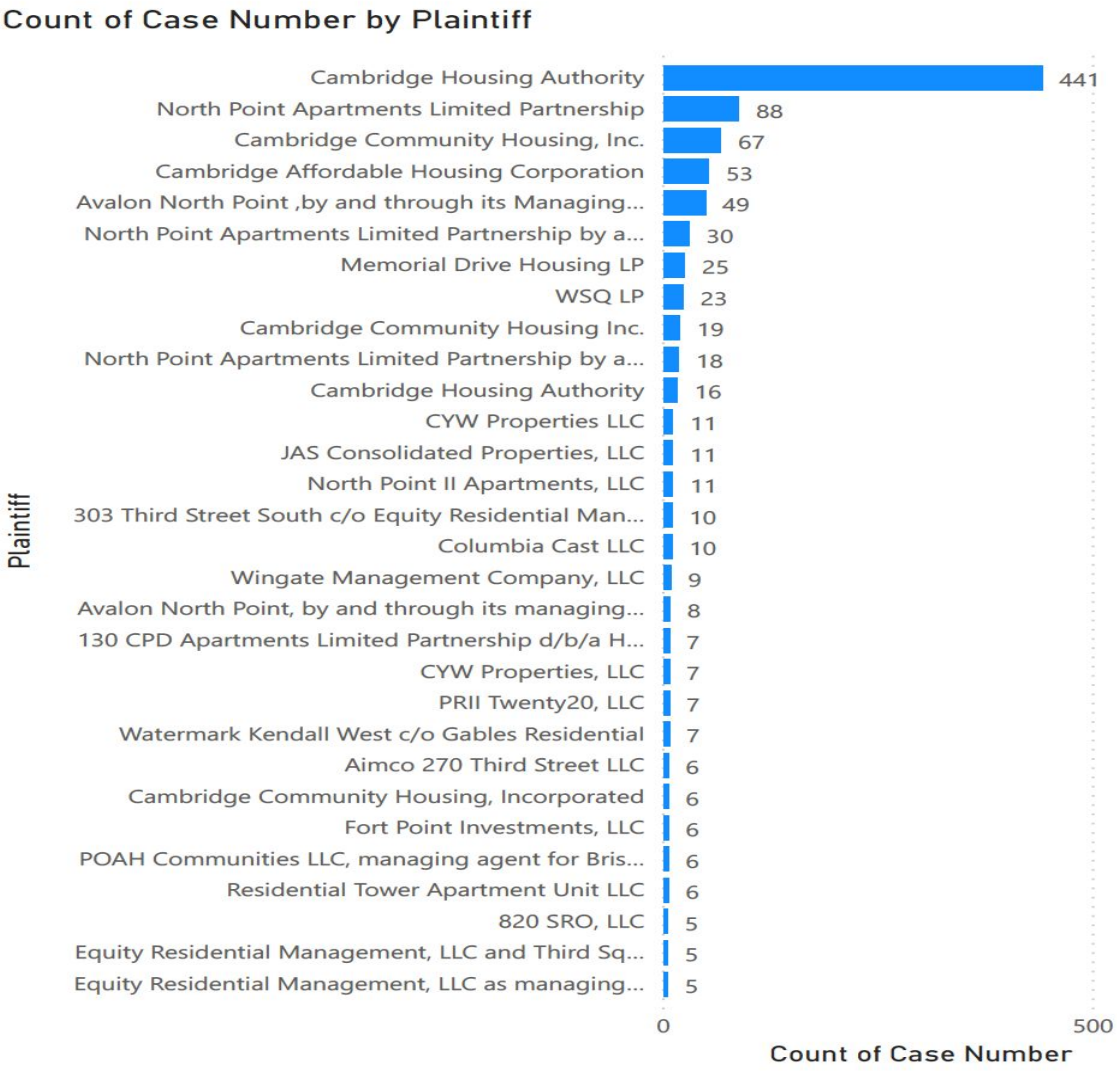
Since we are unable to get the specific income of the individual tenants (the defendants), we use median income of each neighbourhood that is published in the Cambridge open data portal to measure the cases of eviction across income brackets.

The data shows us that the majority of the cases happen to people with median income between 80,000 dollars and 95,000 dollars. One possible explanation for this is that the people in this bracket might have too high of a salary to receive subsidy or government help but still may struggle to afford housing in the Boston area.


Count of Case Number by median_income
Count of Case Number
400
300
200
100
0
392
203
150
84
90
18
16
21
2
65K
70K
75K
80K
85K
90K
95K
100K
105K
110K
median_income

Eviction Cases By Plaintiff

Most eviction cases occurred in the properties owned by Cambridge Housing Authority. This is not surprising because Cambridge Housing Authority is the organization that mainly provides housing for lower income residents.

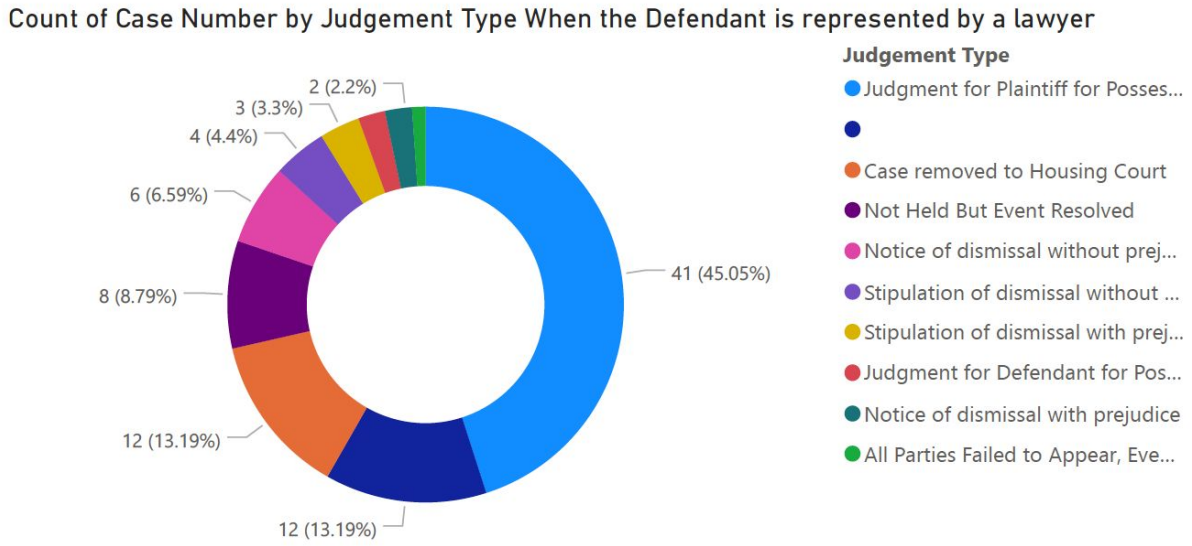


Eviction Cases by Judgement Type

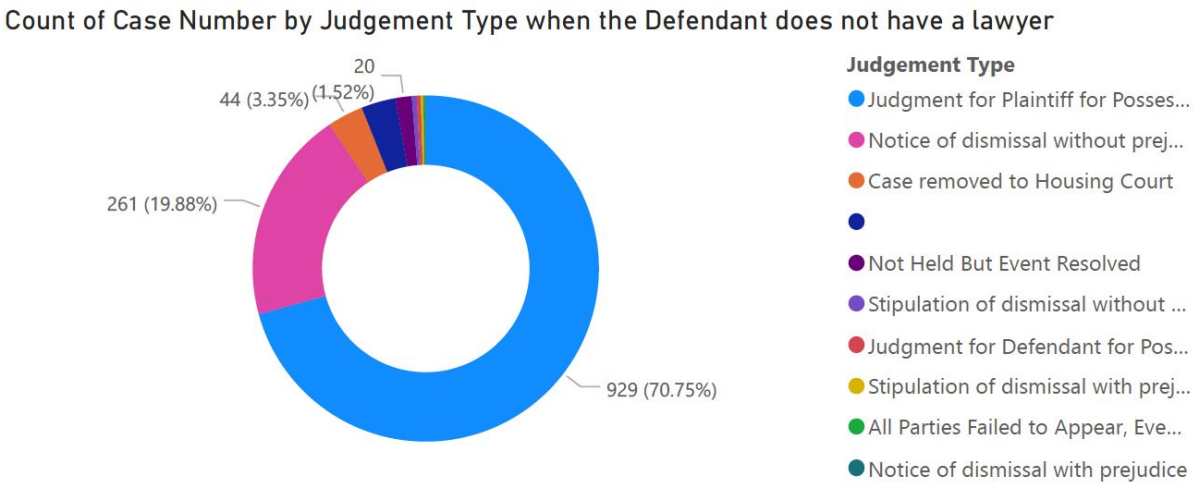
Defendant represented by attorney vs. no representation

Although there were only 91 cases out of 1404 (less than 6.5%) in which the defendant was represented by an attorney, the outcome for the defendant was better when they were represented by an attorney as opposed to when they were not represented by an attorney.

When the defendant was represented by an attorney, only 45% of cases resulted in Judgment for Plaintiff:

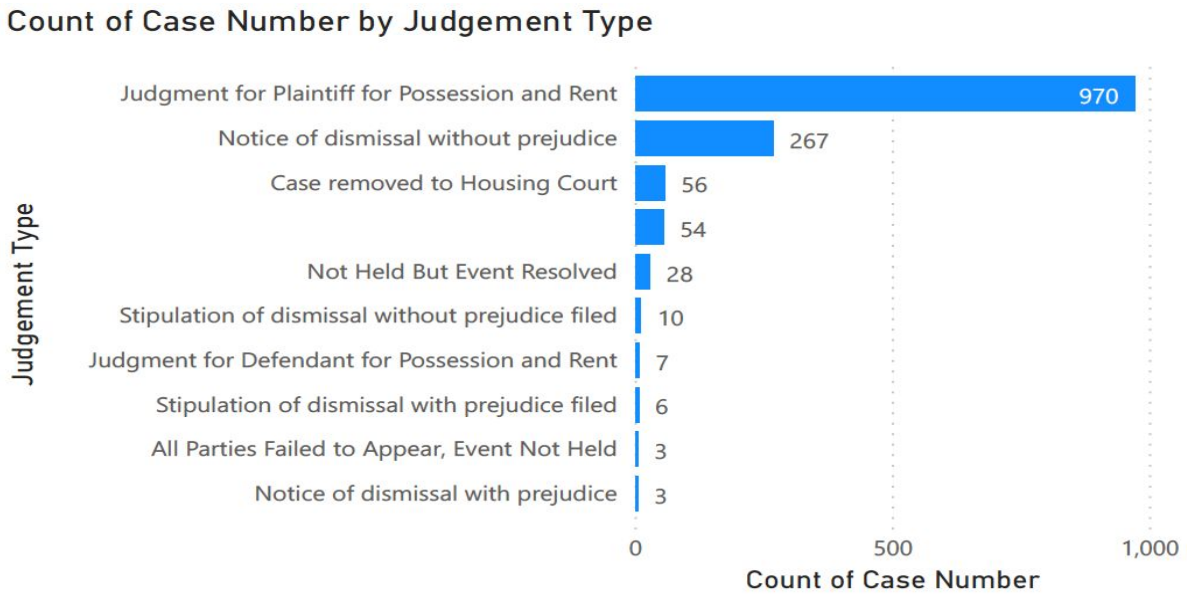


When the defendant was not represented by an attorney, the percentage of cases that resulted in Judgement for Plaintiff increased to 70% of cases:

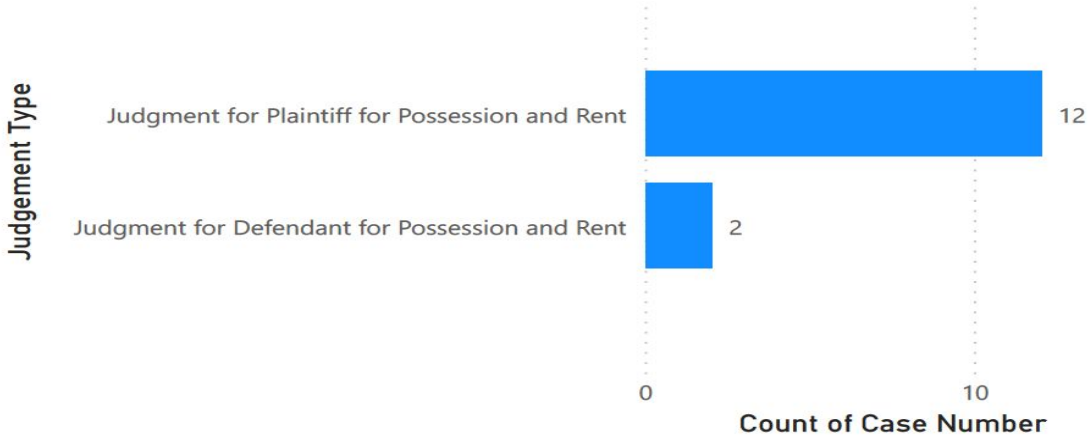


Cases that went through trial by judge vs. no trial

The vast majority of cases (970 out of 1404) resulted in Judgement for the Plaintiff; there were only 14 cases that resulted in judgement for the defendant.

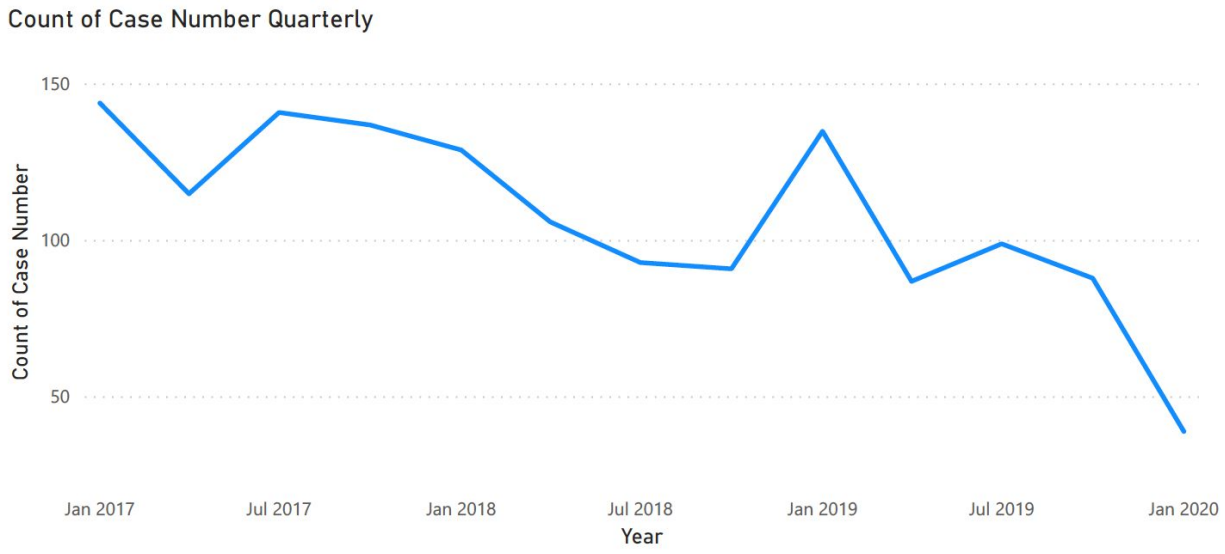


However, when we drill down to the number of cases where the verdict came after trial by judge, the odds of a positive judgement for the defendant increased to 14% (1/7) from 0.5% (7/1404) considering all cases. The figure below depicts the frequency of judgement types after trials by judge.



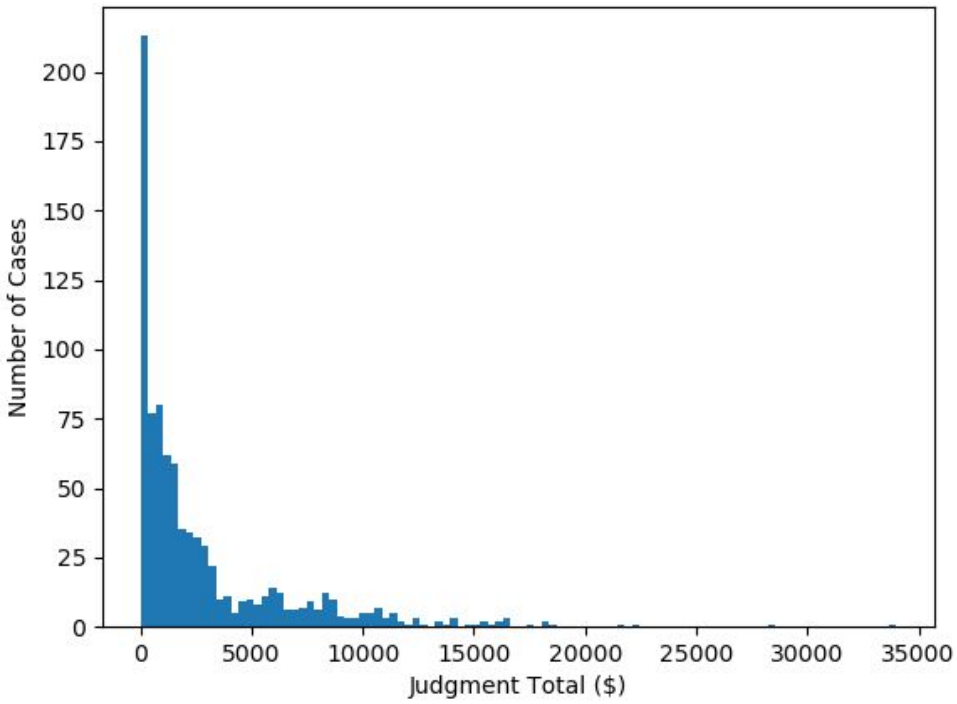
The trend of eviction cases

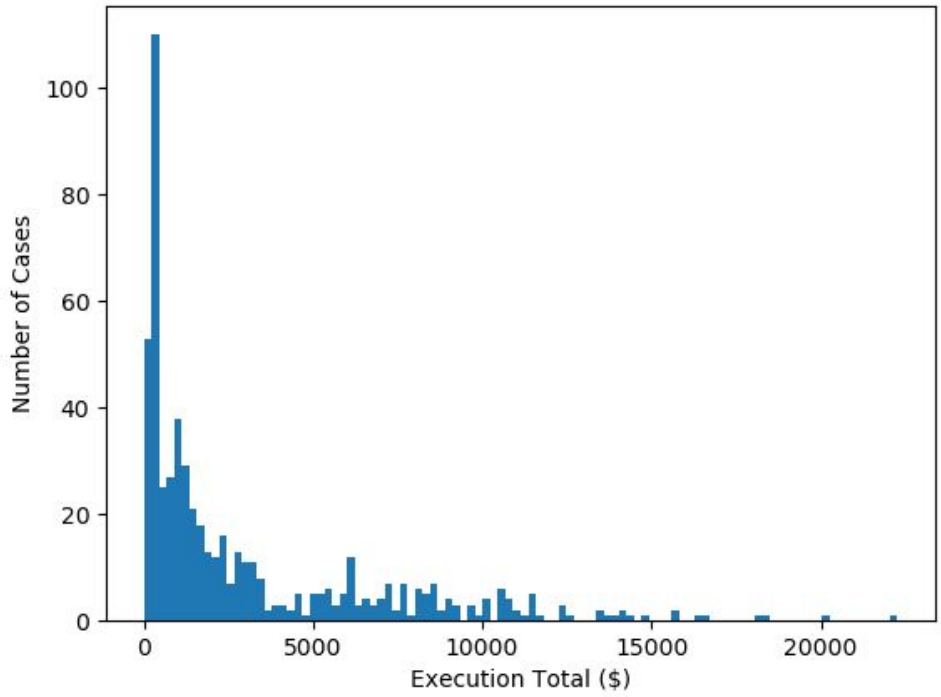
Over time, the number of eviction cases reduced. For 2020, the data is incomplete. The possible explanation for this trend is Boston's substantial economic growth over the past few years. Decreasing rates of unemployment and increasing salaries could allow for more people to afford high Boston-area rent.



Eviction Case Payment Totals

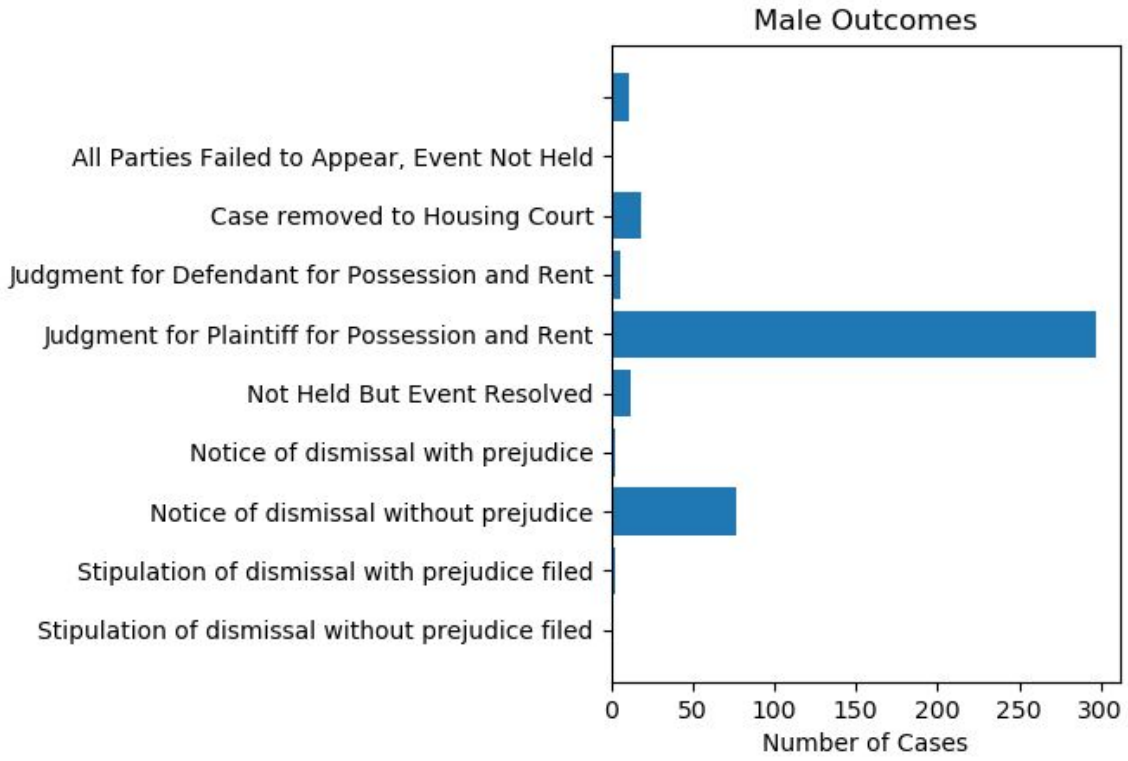
Histograms for the judgment totals and execution totals are shown below. From these plots, it is clear that the maximum execution totals were lower than the maximum original judgment totals.

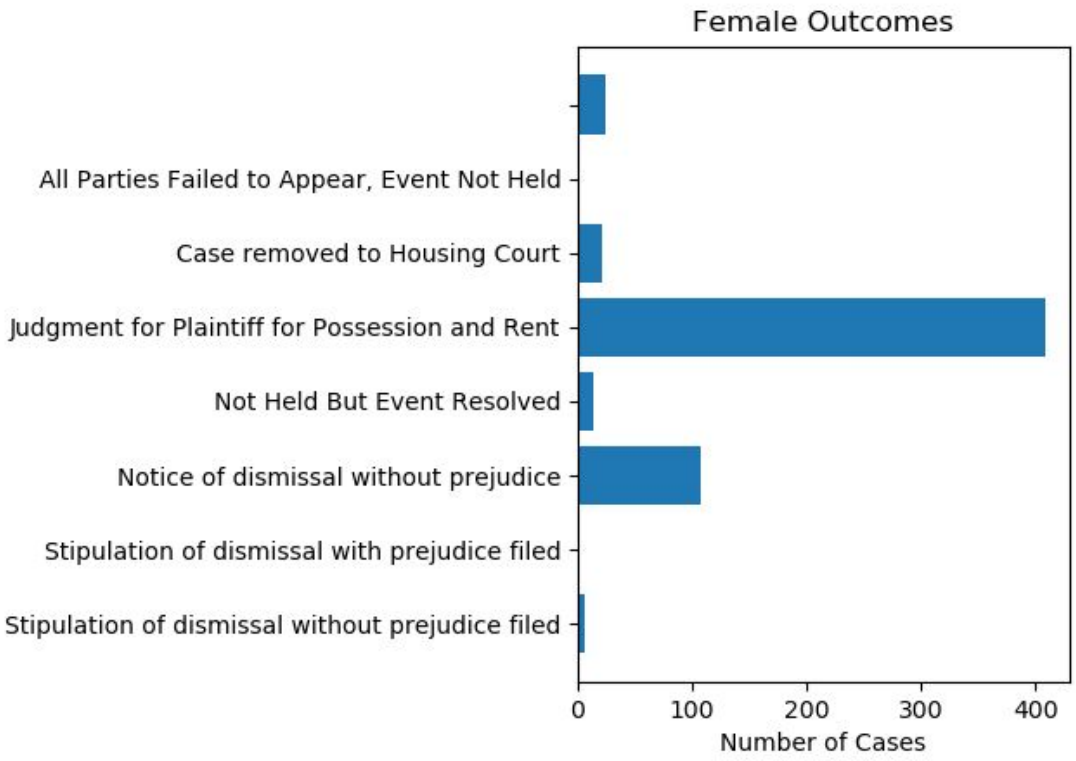

Number of Cases
200
175
150
125
100
75
50
25
0
0
5000
10000
15000
20000
25000
30000
35000
Judgment Total ($)


Number of Cases
100
80
60
40
20
0
0
5000
10000
15000
20000
Execution Total ($)

Eviction Cases and Gender

We attempted to predict gender and race using the “gender-guesser” Python package. Gender was determined for each defendant name in the database. However, gender-guesser classifies names from certain languages and cultures better than it can classify names from other languages and cultures, likely due to a limited training set. The gender-guesser algorithm guesses the gender of names that it is able to classify with high confidence, and this algorithm classifies the remaining names as “unknown.” Most of the names have been decisively classified as male or female. There are not any major differences between male and female outcomes.





Additionally, our group attempted to examine the race of each defendant using a Python package called ethnicolr, which claims to be able to predict race based on name using U.S. census data. One obstacle in using this algorithm is the potential bias in the algorithm itself. We did not know how to devise our own test to see whether the ethnicolr package was accurate or not, whereas with the gender-guesser package, we could at least examine the defendant names and, based on our own knowledge of names, check whether the name was correctly attributed to its gender. The ethnicolr package also relies heavily on Tensorflow. There were many difficulties in installing Tensorflow on a Windows PC. Perhaps in future iterations of the project, a similar package could be used and a test could be devised to examine its accuracy.

Eviction Cases and Affordable Housing

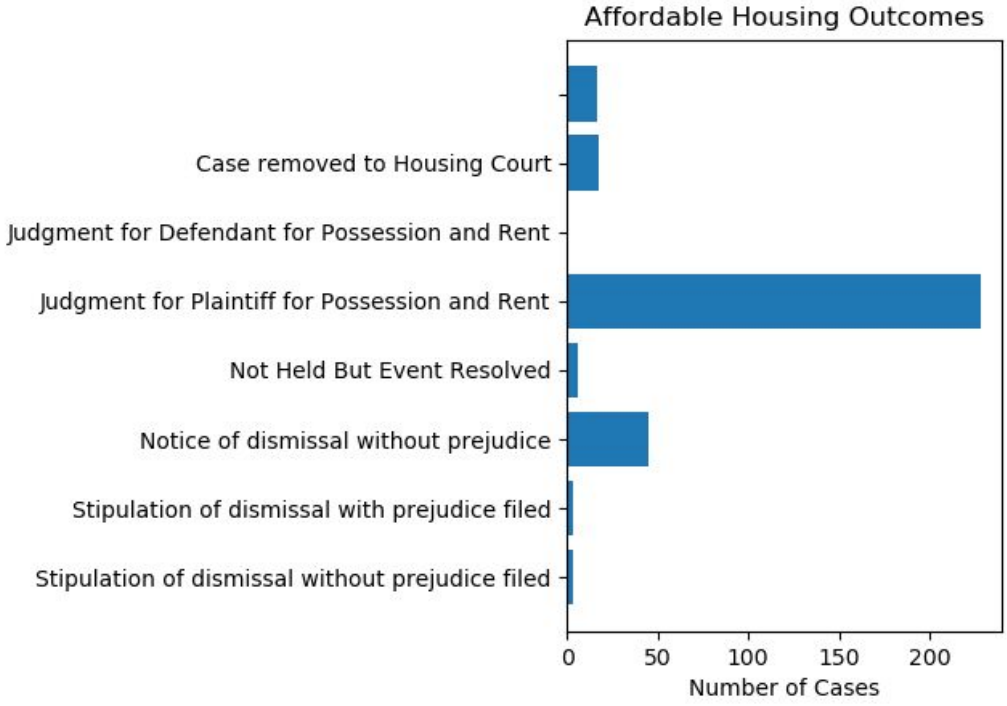
Another way of analyzing the trends in the characteristics of each defendant case is to cluster data. In order to cluster data, the features were categorized by parsing and categorizing data. Strings, such as neighborhood and judgment type, were categorized by name. Numerical data, such as median income, were categorized by binning data.

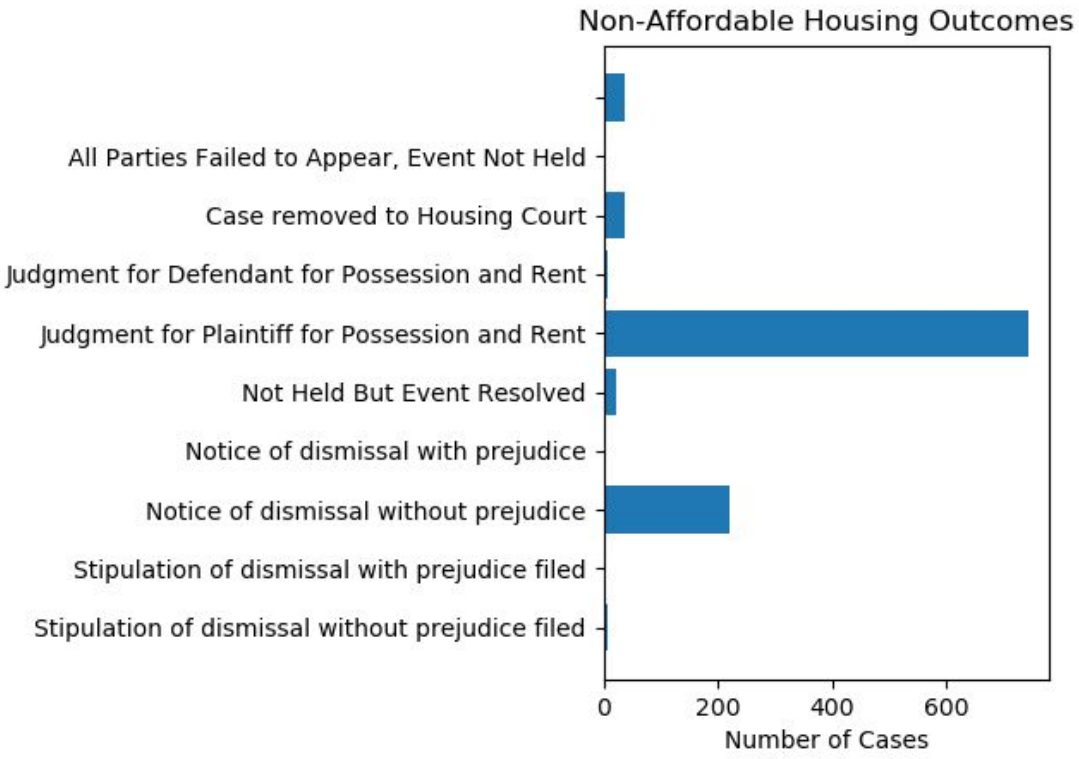
K-means is often insufficient for clustering categorical data. Instead, K-modes, a variant of K-means typically used with categorical data, was used to cluster data. K-modes works by counting the number of matching categories between data points and, from this, calculating a similarity value. For instance, for a sample A and a sample B, the K-modes algorithm calculates the number of matching values between samples. This similarity value is used in place of the distance values calculated using K-means. Another difference between K-means and K-modes

is that K-modes uses the mode as the measure of central tendency, whereas K-means uses the mean. Descriptions of K-modes come from the Python package documentation.

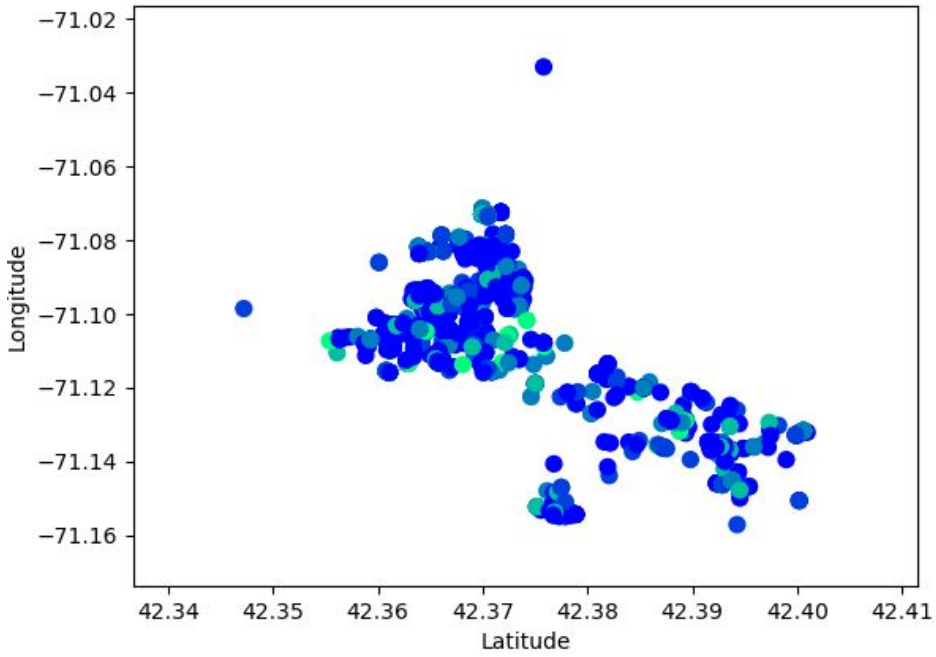
K-modes was used to examine differences between affordable housing and non-affordable housing. In order to determine the number of low-income or affordable housing units, we used various low-income housing search websites. The addresses in our database were then parsed and searched for the names and addresses of low-income housing units. 321 out of 1404 units were designated as low-income housing. Various analyses were performed, with visuals shown below, to show the relationship between low-income housing and eviction data.

There is a much higher rate of decisive evictions and dismissals for non-affordable housing units compared to affordable housing units. This could support the previously stated hypothesis that people with incomes that are too high to qualify for affordable housing (but that are still low for the Boston area) may not be able to afford rent.



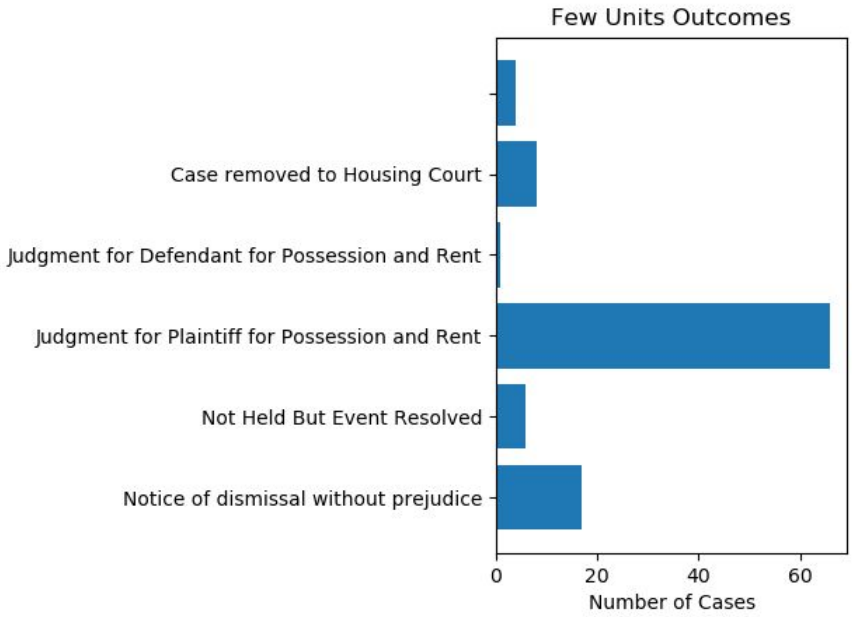


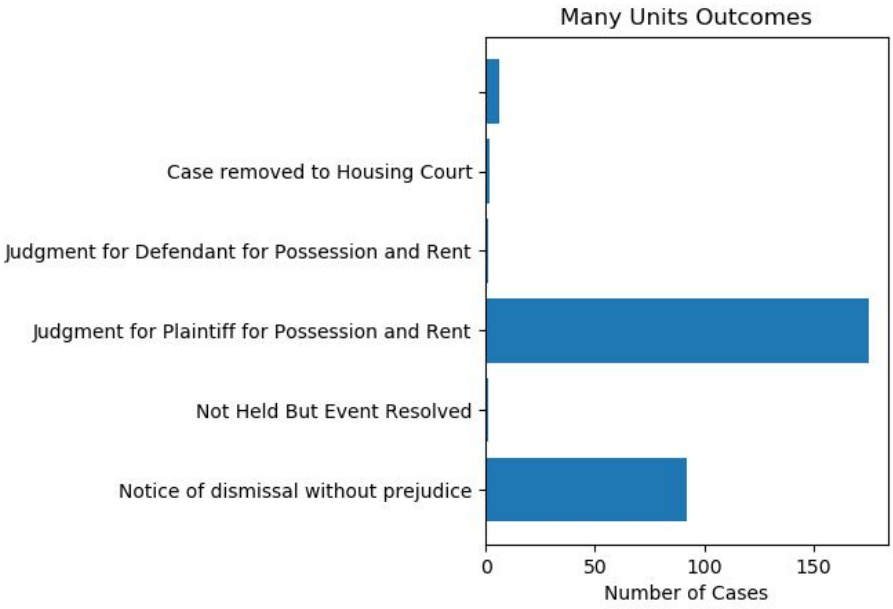
The following figure shows a K-modes analysis for affordable housing. Perhaps in future iterations of this study, the clusters from this K-modes analysis could be used to determine overarching trends for eviction cases and affordable housing units.



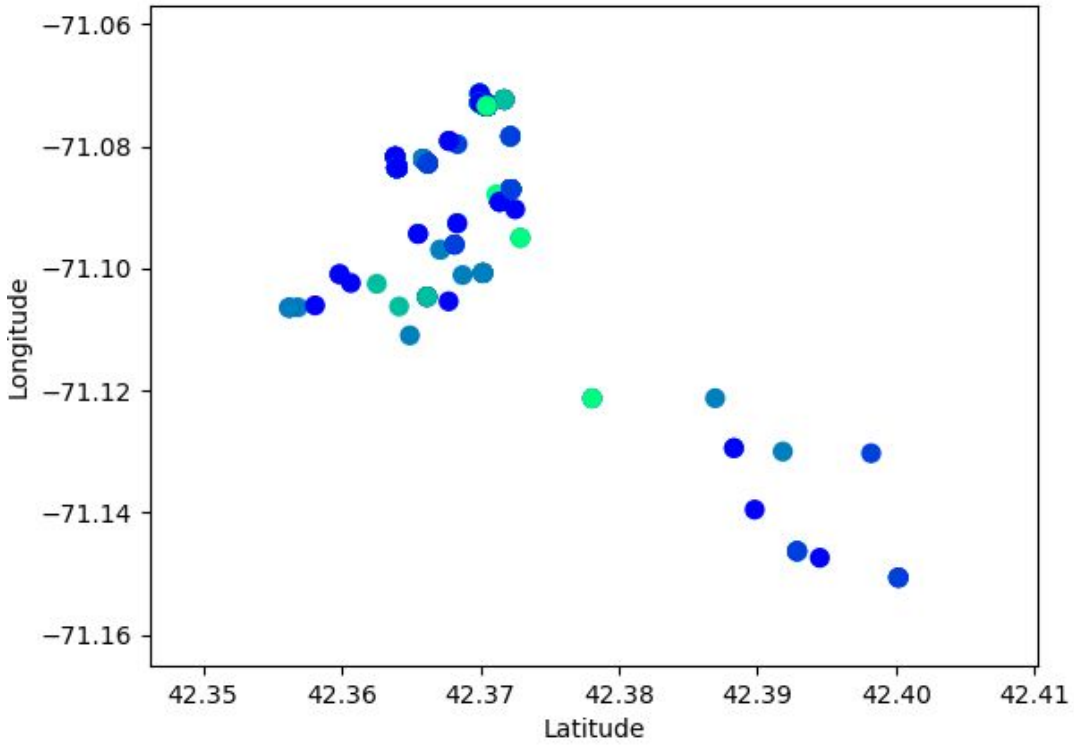
Eviction Cases by Number of Units

The data on the size of the apartment buildings is somewhat limited. We were able to retrieve the number of units at an address for 379 cases. These cases were then analyzed using K-modes as well. Visuals shown below show the relationship between number of units and eviction data. Buildings with a large number of units (>= 100) have a much higher rate of eviction and dismissal than buildings with a small number of units (< 100) compared to the other outcomes.

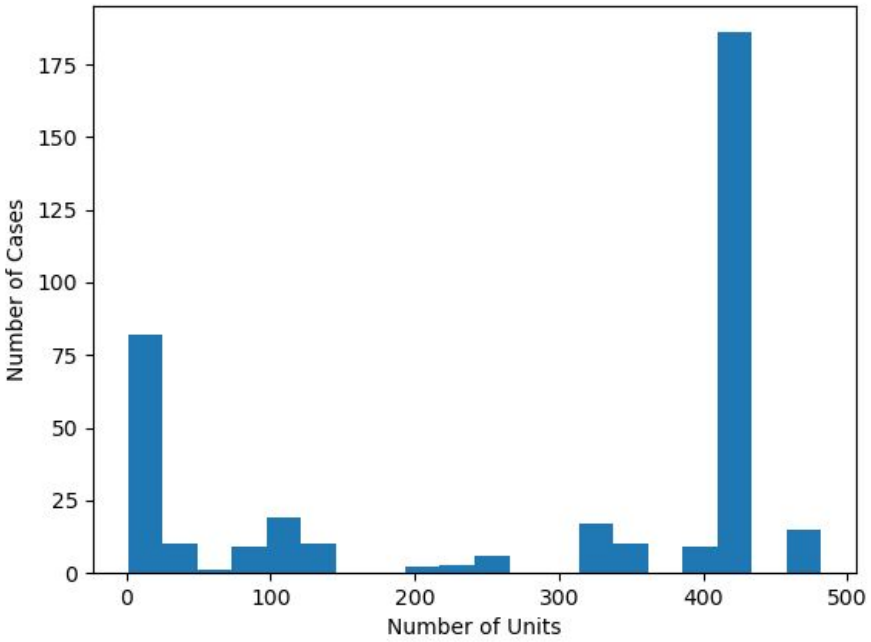




The following figure shows a preliminary K-modes analysis for the number of units. Similar to the aforementioned affordable housing K-modes clusters, perhaps these housing unit number K-modes clusters could be used to determine larger trends in Cambridge eviction cases. Eventually, similar K-modes model formats could be robust enough to apply to other cities as well.



A histogram of the number of units is shown below.



Conclusions

Through our analysis of the eviction cases in the City of Cambridge from 2017 to early 2020, our group began to develop an understanding of the eviction cases in Cambridge. We found that the factors which most influenced the likelihood of eviction were:

- Living in East Cambridge or the Port
- Living in a property managed by Cambridge Housing Authority
- Live in a neighborhood with median income between 80K and 95K
- Representation by a lawyer
- Affordable housing status
- Number of units in a building

Bias

There are several potential sources of bias in this study. One source of bias is the lack of representation for certain categories within a feature. For example, there were a limited number of cases in which there was defendant lawyer representation. These small sample sizes within categories limit the ability to draw accurate conclusions about entire populations from this study.

Additionally, there are several socioeconomic factors that could skew results in a study involving housing and the justice system. For example, people who face eviction may not be able to afford to hire a good lawyer, let alone pay rent. Additionally, there exist great disparities in whether a person is able to appear in court, due to familial, occupational, or other obligations. Some jobs and obligations may allow for time to appear in court (which could lead to a more

positive court case outcome), while other obligations may not allow a person to take time off to appear in court.

Lastly, there are inherent biases in some categorization algorithms. Although clustering and categorization algorithms may be based on thoroughly tested methods and proven mathematics, these algorithms still rely heavily on adjustable parameters such as variable weights, number of clusters, and decision boundary thresholds. Some of the aspects of our study depend heavily on extrapolation, such as the gender of defendants based on name and the attempted clustering. One example of bias that could be introduced from these methods is that the Python gender identification algorithm could be extremely biased towards the most popular names of people in the United States. In reality, people in this study could have names from many languages and cultures. In future iterations of the study, perhaps more robust methods of clustering and categorization could be implemented by using more powerful algorithms with larger training sets. Bias could also be reduced by optimizing clustering; since the client asked for conclusions based on specific features as opposed to the clustering of all categories, we decided to focus on individual feature analysis instead.

Future works

The research we conducted could be extended to build more comprehensive and detailed profiles of the tenants (defendants) and landlords (plaintiffs), which would give better explanations of the parties involved in the eviction cases. The presence of additional data sources would contribute significantly to this effort.

The study could also be extended to include records from more years in the past to get the time series trend of the eviction cases. Macro economic data such as employment rates and increase in GDP, among others, could be included to improve the analysis.